Characterization of Binding Sites of Eukaryotic Transcription Factors

Jiang Qian[1]*, Jimmy Lin[1], and Donald J. Zack[1,2,3,4]

[1] *The Wilmer Institute,* [2] *Department of Molecular Biology and Genetics,* [3] *Department of Neuroscience, and* [4] *McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.*

To explore the nature of eukaryotic transcription factor (TF) binding sites and determine how they differ from surrounding DNA sequences, we examined four features associated with DNA binding sites: G+C content, pattern complexity, palindromic structure, and Markov sequence ordering. Our analysis of the regulatory motifs obtained from the TRANSFAC database, using yeast intergenic sequences as background, revealed that these four features show variable enrichment in motif sequences. For example, motif sequences were more likely to have palindromic structure than were background sequences. In addition, these features were tightly localized to the regulatory motifs, indicating that they are a property of the motif sequences themselves and are not shared by the general promoter "environment" in which the regulatory motifs reside. By breaking down the motif sequences according to the TF classes to which they bind, more specific associations were identified. Finally, we found that some correlations, such as G+C content enrichment, were species-specific, while others, such as complexity enrichment, were universal across the species examined. The quantitative analysis provided here should increase our understanding of protein-DNA interactions and also help facilitate the discovery of regulatory motifs through bioinformatics.

Key words: transcription factor, promoter, gene regulation, bioinformatics

Introduction

Deciphering transcriptional regulatory networks is crucial for the understanding of cellular processes such as growth control, differentiation, and cell death (*1*). Although post-transcriptional mechanisms often play important modulatory roles, transcriptional regulation usually is the primary determinant of which genes are expressed, when they are expressed, and in which cells they are expressed. In addition, the determination of transcription factor (TF) and target gene relationships can sometimes provide an insight into the mechanisms by which mis-regulated expression can lead to human diseases (*2*, *3*). In the future, the ability to understand and modulate gene expression may provide new avenues for therapeutic intervention (*4*).

Regulatory networks involving TFs and their target genes are being studied through both lab-based and bioinformatics approaches. Laboratory methods can identify cis-acting DNA elements and their cognate trans-acting DNA binding proteins (TFs) and can provide structural and functional information on their mechanisms of interaction. Although powerful, these methods are time-consuming, can be difficult to carry out, and thus far have yielded only partial information concerning the complex networks involved. Recently, there have been significant experimental breakthroughs in determining TF binding sites in high-throughput fashion, mostly using microarray as a platform, such as ChIP-chip (*5*, *6*) or DIP-chip (*7*). As a complement to the lab-based methods, there has been increasing interest in using bioinformatics approaches to study gene regulation, and these studies are being greatly aided by the ongoing dramatic increases in available genomic information.

Bioinformatics approaches have already been used to identify DNA regulatory regions. This work includes the establishment of binding site databases such as TRANSFAC (*8*) and JASPAR (*9*), cross-species homology comparison of putative promoter regions looking for highly conserved sequences (phy-

***Corresponding author.**
E-mail: jiang.qian@jhmi.edu

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

logenetic footprinting) (*10–12*), and comparison of promoter regions of similarly regulated genes such as those identified by high-throughput microarray studies (*5*, *13*). Bioinformatics has also been used to explore the nature of various protein-DNA interactions. For example, the concept of a "recognition code" has been proposed by a number of investigators, and hopes were raised that a simple set of rules, like the genetic code, might be able to explain the specificity of protein-DNA interactions (*14*, *15*). However, inspection of increasing numbers of crystal structures of DNA binding proteins, especially protein-DNA complexes, has indicated that this recognition is probabilistic instead of deterministic. In other words, some amino acid-base contacts (for example, Arg <=> G) are preferred, but the recognition is degenerate in both directions (*16*).

In the present study, we have utilized a statistical framework to explore the nature of protein binding DNA sequences and how they differ from surrounding non-protein binding sequences. In particular, we were interested in determining what kind(s) of DNA sequence segments are most likely to serve as the TF binding sites. The purpose of this study was two-fold: first, to identify features associated with TFs that can help us better understand protein-DNA interactions; second, to incorporate these features into motif discovery programs to improve the performance of motif prediction. When we have used motif discovery programs, such as MEME (*17*), Consensus (*18*), and AlignACE (*19*), the correct motif has often not received the highest prediction score, and in many cases the top-predicted motif was a degenerate motif such as "AAAAAA". It is our hope that with the additional information provided by the features we identified in the present study, we and others will be able to generate new and better rankings of predicted motifs and potentially move the correct motifs into the top-ranking positions.

The paper is organized as follows (Figure 1): First, we discuss the four features associated with the selected motifs and analyze the degree of dependence on these features. Then, to explore whether the features are specific to these short binding sites or are simply reflections of the surrounding promoter sequences, we examine the features of the larger promoter "environment". We also break down the binding sites according to the TF classes to which they bind (for example, homeodomain proteins) and attempt to identify features specific to TF classes. We finally examine whether the observed features are species-specific by comparing the results from yeast, murine, and human genomes.

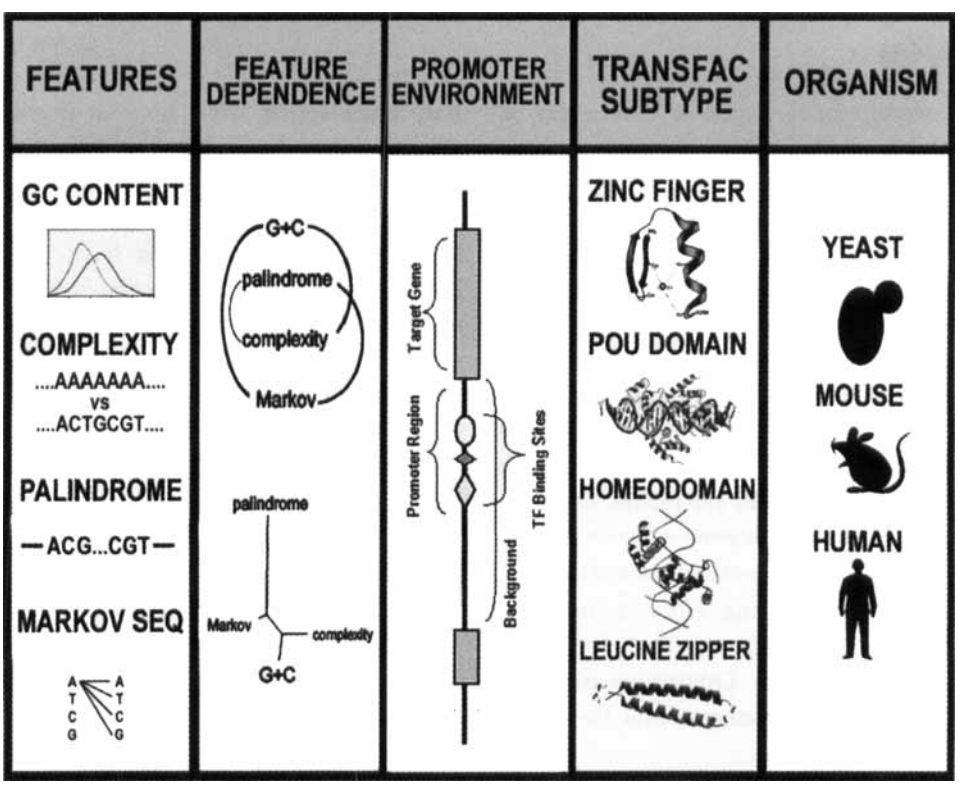

Fig. 1 A schematic overview of the work.

Results

In this study, we have investigated the features associated with the TF binding motifs and examined how they differ from the surrounding non-protein binding intergenic sequences. Odds ratios were used to show the contrast between motif sequences and background. In essence, we measured the enrichment (or depletion) of the occurrence of motifs in the binding regions versus the background occurrence of a certain feature (for example, G+C=0.8). The statistical significance of the enrichment was evaluated by bootstrap simulation, which estimated the range of random fluctuation. The four features we examined were G+C content, pattern complexity, palindromic structure, and Markov sequence ordering. The first two are related to sequence composition, while the last two are about specific sequence ordering.

Motif features

G+C content

The first feature we investigated was G+C content in motif sequences, as defined by the percentage of G and C occurring in a sequence. The G+C content for both motif sequences and intergenic background sequences was calculated for these two datasets (Figure 2A). The distribution for the motif sequences was clearly shifted to the right when compared with the distribution for the background, indicating a preference for high G+C content in motif sequences. The corresponding odds ratio is shown in Figure 2B. To assess the statistical significance of the odds ratio differences, confidence intervals (95%) from bootstrap simulations (see Materials and Methods) were calculated. The actual odds ratio curve was clearly distinct from the intervals, suggesting that the observed tendency was very unlikely to reflect a random variation in sequence distribution. This observation of high G+C content enrichment in motif sequences is not an unexpected finding, given that promoter regions in general tend to be GC-rich and are often associated with CpG islands (*20*, *21*). This observation also corroborates the previous finding that the intergenic regions between divergently transcribed gene pairs, where the regulatory motifs reside, are more (G+C)-rich (~36%) than the regions between convergently transcribed genes (~30%), where no TF binding sites are expected (*22*).

Pattern complexity

We next considered pattern complexity, which can essentially be considered the entropy of a sequence. It is defined as $-\sum_{i=ATGC}^{p_i} \log_2 p_i$, where p_i is the probability of occurrence of a particular nucleotide in the motif sequence; values range from 0 (for example, AAAAA) to 2 (for example, AATTGGGCC or ATGCTAGC). By studying this feature, we tried to understand whether TFs tend to recognize simple DNA patterns or complex patterns. Figure 2C displays the distributions of pattern complexity for motif sequences and background. Overall, both distributions were centered on the high-complexity side. Although the distributions appeared to be similar, the difference in average pattern complexity (1.71 vs. 1.66 for motif and background sequences, respectively) was statistically significant, with a t-test-derived p-value of less than 10^{-18}. The bootstrap-derived odds ratios for pattern complexity did not show as much divergence between observed and calculated data as was seen for G+C content (Figure 2D). At the high-complexity end, the curve suggests that highly complex sequences are enriched in the motif database when compared to the background. However, at the low-complexity end, as indicated by the curve being largely within the confidence intervals, there was no statistically significant difference between the motif and background sequences.

Palindromic structure

The type of DNA palindrome that is generally thought to be potentially biologically significant is a string with a 5′ to 3′ sequence that is identical to the 5′ to 3′ sequence of the reverse complimentary strand, for example, ACGT. A second type of palindrome is one in which the reverse sequence is a direct repeat of the forward sequence, for example, ATTA. We analyzed the occurrence of both types of palindromes in the set of regulatory motif and background sequences. A program was designed to detect the palindrome structure contained within a sequence. The output of the program is the type of palindrome and the repeat unit size. We allowed any length between the two repeat units. For instance, the sequence "**ACGCCGT**AAA" is an "ACGT"-type palindrome with a repeat unit size of 3, and the length between the two units (ACG and CGT) is 1. Sequences without any palindromic structure will have repeat unit sizes of 0 or 1.

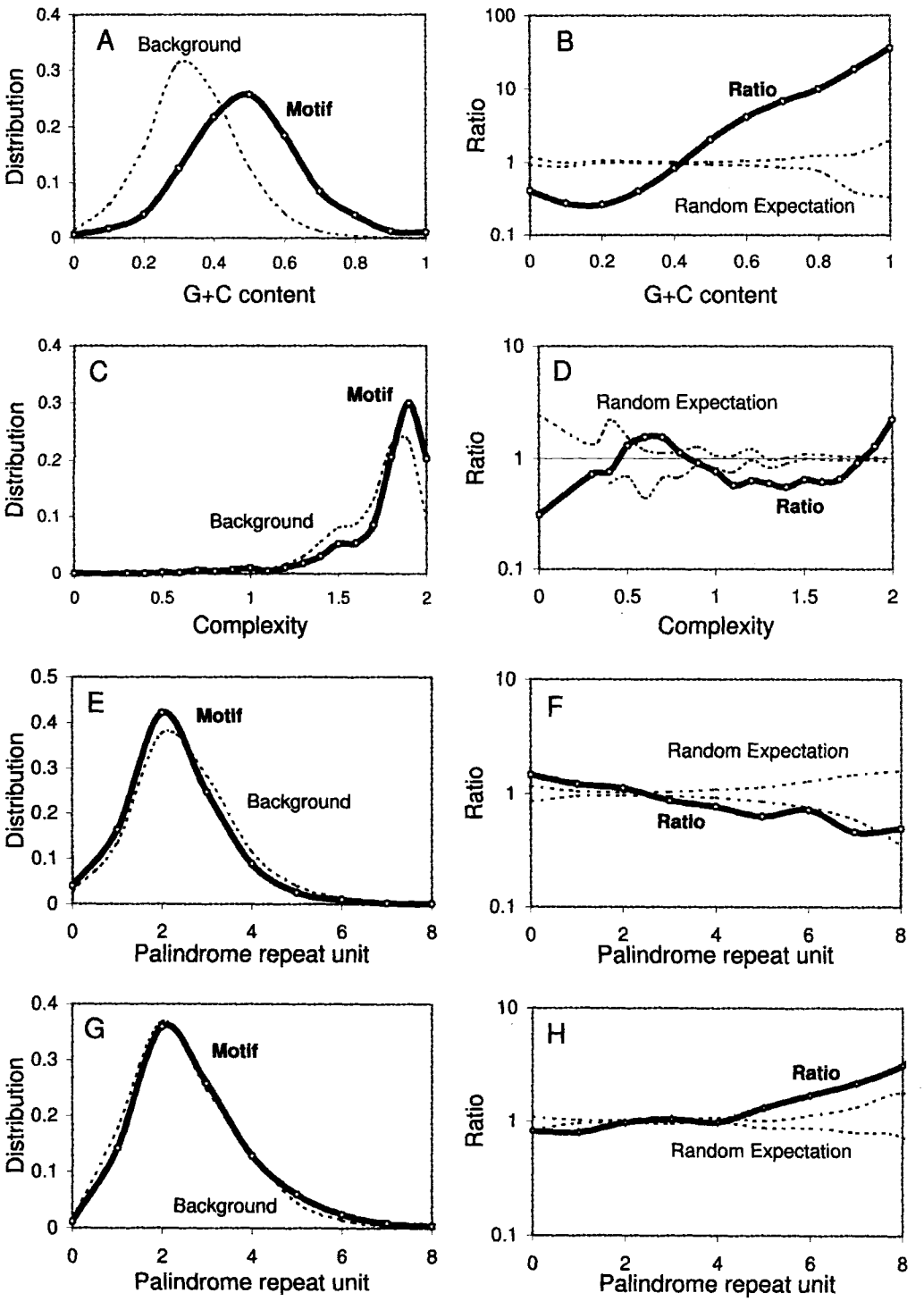

Fig. 2 Features associated with motif sequences. **A**. Distribution of G+C content in motif sequences (solid line) and yeast background sequences (dashed line), showing the frequency of occurrence of sequences with various G+C contents. **B**. The odds ratio for G+C content, obtained by dividing the distributions of motif sequences (solid line in A) by that of the background (dashed line in A); two thin lines indicate the 95% confidence intervals from a bootstrap simulation (see Materials and Methods). **C** and **D** are the distribution and ratio for pattern complexity. **E** and **F** are the distribution and ratio for "ATTA"-type palindromes. **G** and **H** are the distribution and ratio for "ACGT"-type palindromes.

Figure 2E shows the odds ratios between motifs and background for "ATTA"-type palindromes. As indicated by the confidence intervals from bootstrap simulation, the curve is not distinguishable from a random sequence (Figure 2F). This indicates that, on average, "ATTA"-type palindromes are not enriched in the motif sequences. On the other hand, the odds ratio curve for "ACGT"-type palindromes (Figure 2G and H) appeared to be different from the "ATTA"-type, with enrichment for "ACGT"-type palindromes monotonically increasing with increasing repeat unit size. The enrichment for palindrome sequences can be rationalized by the fact that many TFs bind to the promoter regions as homo-dimers.

Markov sequence ordering

The fourth feature we analyzed was Markov sequence ordering in the motif sequences. In other words, we were interested in the probability distribution of each nucleotide following a particular prior nucleotide. A transition probability table was generated for this purpose.

The transition probability for a sequence was calculated according to the equation:

$$P(i \to j) = \frac{N(i \to j)}{\sum_{j'=ATGC} N(i \to j')}$$

where i and j represent two adjacent nucleotides in a sequence and $N(i \to j)$ is the occurrence of dimers of ij in the sequence. The denominator sums over all possible j's. $P(i \to j)$ was calculated for both motif sequences and intergenic background sequences.

Table 1 shows the odds ratio for the transition probabilities, where the first row contains the values for which an "A" is followed by each of the four bases, and so on for the other rows. Positive values indicate that certain transitions (for example, "A" followed by "G") happen more often in motif sequences than in the background, while negative values indicate the opposite. From the table it is clear that there are indeed certain preferences in terms of nucleotide ordering. For example, G following A is more likely than T following A in the motif sequences. Note that these results demonstrate the contrast between motif sequences and background. When we checked the Markov preference in general intergenic sequences (background), there was a tendency to avoid NpG and NpC.

The data in Table 1 indicate that within the motif sequences, G and C are in general more likely to follow any nucleotide than are A or T. This apparent preference could arise mainly from the fact that, as noted above, motif sequences are GC-rich. One might therefore wonder if Markov sequence ordering contains the same information as sequence composition bias (for example, G+C content), viewed from a different perspective. To determine whether there was additional information to be obtained from the Markov sequence ordering, we re-calculated the odds ratios for transition probabilities by comparing motif versus permuted motif sequences, instead of motif versus background. Since permutation does not change the composition of a sequence, the permuted motif sequences maintain the same G+C content. If the observed Markov sequence ordering totally stems from sequence composition bias, the new odds ratios should be around 0. Most of the re-calculated values were indeed around 0 (Table 2), indicating that the observed Markov sequence ordering arises mainly from sequence composition bias. However, there were also several points that deviated from 0. For instance, it was unlikely for a G to follow a C, considering the fact that the motif sequences are usually (G+C)-rich. Thus, Markov sequence ordering does apparently provide some additional information and is not simply a reflection of composition bias.

Table 1 Markov Sequence Ordering: Odds Ratio for Transition Probabilities (Motif vs. Background)

	A	T	G	C
A	−0.23	−0.13	0.31	0.31
T	−0.29	−0.32	0.43	0.42
G	−0.20	−0.28	0.49	0.10
C	−0.15	−0.36	0.26	0.43

Table 2 Markov Sequence Ordering: Odds Ratio for Transition Probabilities (Motif vs. Permuted Motif)

	A	T	G	C
A	0.050	0.067	−0.058	−0.085
T	−0.200*	0.030	0.095	0.069
G	0.031	−0.037	0.056	−0.064
C	0.140*	−0.064	−0.170*	0.060

*The cells with large absolute values.

Dependence of the features

As shown above, the features of Markov sequence ordering and G+C content are not totally independent. In order to explore this relationship further, we attempted to clarify whether the four analyzed features are confounded with each other. The scatter plots for various features are shown in Figure 3A–C. Each spot in the figure represents a motif sequence, and its location is determined by its feature values. Figure 3A shows the relationship between G+C content and Markov score. These two features displayed a linear relationship, with a correlation coefficient of 0.81. If all the spots in Figure 3A were perfectly on one line, these two features would be considered essentially exchangeable, that is, one could exactly determine a feature value from the value of the other. Thus, the deviation from a fitting line shows the degree of independence of the two features. The relationship between G+C content and complexity also showed some degree of dependence (Figure 3B); however, it was not linear. The palindromic feature was not strongly related with the other features. Figure 3C shows the palindrome unit as a function of G+C content.

To quantitatively evaluate the degree of independence of these features, we calculated the residue after a LOWESS fitting between all possible pairs of features (Figure 3A–C). The residue measures the deviation from a perfect mathematical relationship and thus the degree of independence between two features. It is defined as the root mean-of-square of the difference between the actual and the fitting values. The distance (the degree of independence) between each pair of features is described by the value of the residue. Using these distances, a tree was constructed to show the overall relationship between these features (Figure 3D). G+C content was closest to the Markov sequence ordering, while the palindromic feature was far from the other three features.

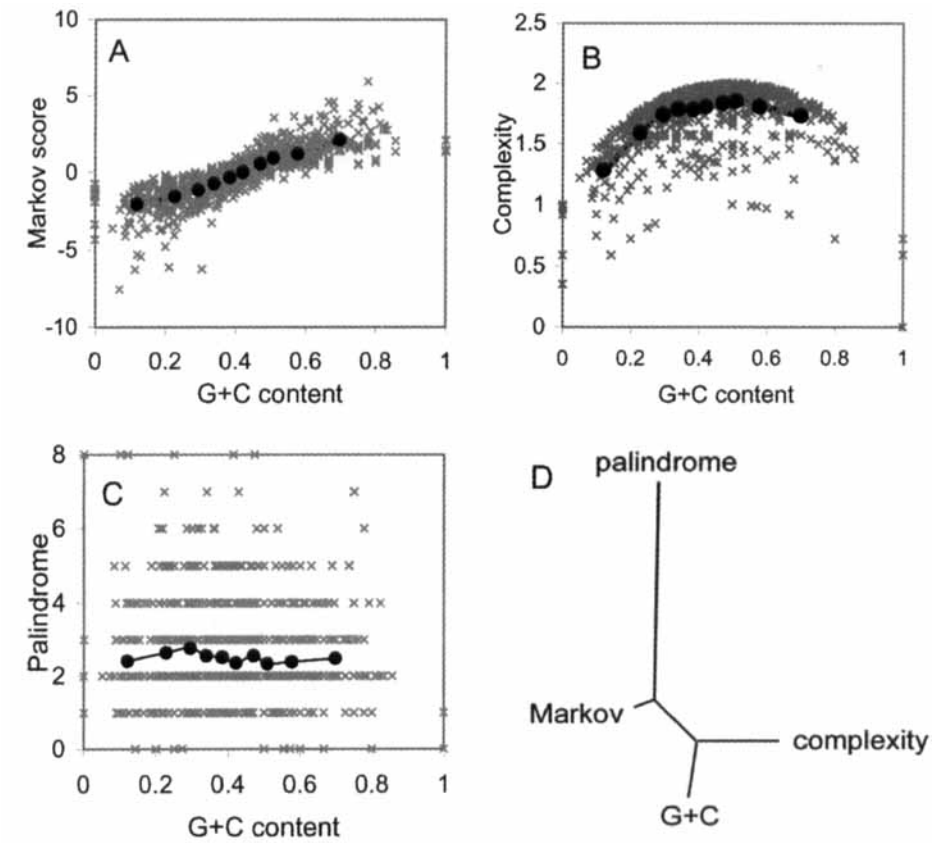

Fig. 3 The relationships between the four features. **A**. Scatter plot for G+C content and Markov score. Markov score is defined as $\sum_{i=1,M} P(i \rightarrow j)$. The summation is over the motif sequence. Each point represents a sequence. The thick dots show the LOWESS fitting. The difference between these spots and the fitting dots is considered the distance between two features. **B**. Scatter plot for G+C content and complexity. **C**. Scatter plot for G+C content and palindrome unit. **D**. Tree of these features.

Analysis of promoter "environment"

The analyses described above have compared motif sequences to a "background" consisting of intergenic sequences. In order to explore whether the observed differences indicated characteristics that are specific to the motif sequences themselves or are simply reflections of the general nature of the surrounding promoter sequence within which the motifs are embedded, we performed a similar analysis comparing overall promoter sequences to the intergenic background. The promoter sequences were obtained from the *Saccharomyces* Genome Database (http://www.yeastgenome.org) and were defined as the sequences upstream of the predicted translational start codon. Since the length of upstream sequence that defines a "promoter" is somewhat arbitrary, we tested three sets of promoter sequences with different lengths: 100, 200, and 500 bp. Short sequences were randomly selected from the various promoter sets such that they had the same length distribution as the binding sites from TRANSFAC (*8*).

We compared the promoter sequences with intergenic sequences to determine whether they shared the same features as the regulatory motifs. Figure 4A shows the G+C enrichment for the sets of sequences from the promoter sets. For comparison, the curve for motifs is also shown in this figure. The G+C content of the promoter regions was only slightly enriched with respect to the intergenic background, with markedly less enrichment than was observed for the motif sequences. Figure 4B and C show the analogous results for complexity and the "ACGT"-type palindrome, respectively, indicating that there was minimal, if any, enrichment for these features among the three sets of promoter regions. Figure 4D displays the log odds ratio for the Markov sequence transition probabilities. The log odds ratios for promoters showed the same tendency as those from the motif sequences, but the magnitudes of the deviations were much smaller in the promoter sequence sets. These results indicate that the identified features that distinguish motif sequences from intergenic sequences are localized to the regulatory motifs themselves and are not simply a reflection of differences in the surrounding promoter environment.

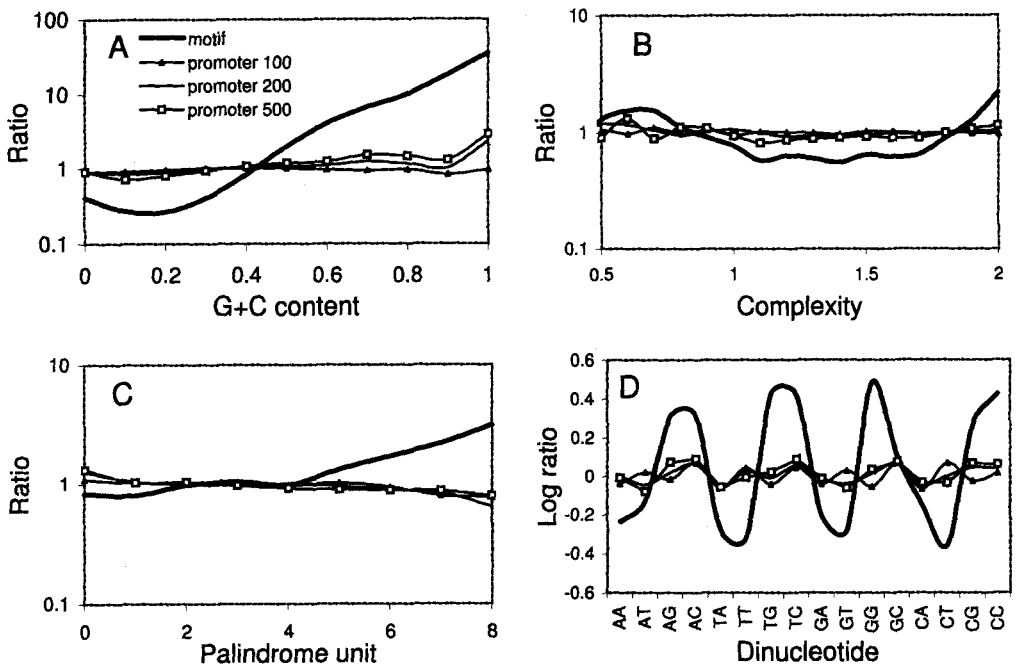

Fig. 4 Localization effect of the features. **A.** Odds ratio for G+C content between the promoters and background. Three curves correspond to promoters with lengths of 100, 200, and 500 bp. The odds ratio for motifs versus background is also shown for comparison. It is clear that the enrichment for high G+C content in motifs was much stronger than in promoters, indicating that high G+C content is specific to the short motif sequences and not shared by surrounding promoter sequences. **B.** Ratio for complexity between the promoters and background. **C.** Ratio for "ACGT"-type palindromes between the promoters and background. **D.** Log ratio for transition probabilities between the promoters and background.

Variation in motif features based on TF subclass

The analysis up to this point has considered all motif sequences as a single group. Since different TF classes have different binding properties, we next examined motif features as a function of the class of their presumptive corresponding binding proteins. The motif sequences from TRANSFAC were partitioned into groups according to the TF classes to which they are reported to bind. The major classes, defined as those with more than 400 binding sites (Table 3), were then analyzed for motif features in a manner analogous to that described above for the full motif dataset. Classes with fewer than 400 listed binding sites, such as helix-turn-helix proteins, were not used for this analysis because of the concern that the lower numbers would make it less likely that statistically significant results could be obtained.

The average G+C content clearly varied among the segregated motif groups (Left column in Figure 5). Most notably, the binding sites of zinc finger proteins had the highest G+C content (average of 0.59). In contrast, the binding sites of POU domain proteins had the lowest G+C content (0.33), which was lower than that of background (0.34). The observed differences were statistically significant, as indicated by an ANOVA derived p-value of 10^{-102}.

For pattern complexity (Left column in Figure 5), the difference among TF classes was not as dramatic as for G+C content. Leucine zipper factors tended to bind to sequences with high complexity, whereas zinc finger proteins have the opposite preference. Although the degree of difference between the TF classes was small, it was still highly statistically significant (ANOVA p-value of 10^{-31}).

As we mentioned previously, the "ACGT"-type palindrome showed significant enrichment in all TF classes combined (Figure 2H). The middle column in Figure 5 shows that the binding sites of zinc twist proteins were most enriched for palindromic structure, while for POU domains and homeobox proteins there was no enrichment. This finding could be rationalized by the symmetry, or lack thereof, of the three-dimensional structure of TF. For example, the DNA binding site of zinc twist is located between and anchored by two zinc atom complexes (*23*).

We also calculated the value of $\log\left(R(i \rightarrow j)\right)$ for the various TF classes. The right column in Figure 5 displays two representative classes and, for comparison, the average values for all classes combined. The correlation coefficients of $\log\left(R(i \rightarrow j)\right)$ were calculated between all possible pairs of groups. The distribution of the correlation coefficients for these groups was very broad, ranging from -0.26 (between zinc twist and homeodomain) to 0.75 (between zinc twist and leucine zipper). Interestingly, some values of $\log\left(R(i \rightarrow j)\right)$ deviated significantly from the general tendency, such as $R(C \rightarrow G)$ for homeodomain proteins.

Species specificity of motif features

An important question is whether the above observations made in the yeast genome can apply to mammalian genomes. As we used odds ratio to characterize each feature, there were two factors that contribute to species specificity: the difference from motif sequences of various genomes, and the background sequence difference in genomes. However, since TRANSFAC contains only about 480 motif sequences from the yeast genome, it is difficult to make a statistically reliable comparison. In this study, we concentrated only on the difference attributed to the backgrounds from various genomes. The same method was employed to extract two sets of background sequences from human and murine intergenic regions, respectively (see Materials and Methods).

To explore this issue further, we performed a similar analysis using murine and human intergenic sequences as background. In terms of G+C content, the enrichment varied between species: the yeast binding

Table 3 TF Classes

ID	No. of sites associated with factor	Factor description
C0001	1,309	zinc finger
C0002	815	zinc finger; zinc twist
C0006	948	zinc homeodomain; homeobox protein
C0007	468	zinc POU domain
C0008	1,202	zinc basic region+leucine zipper

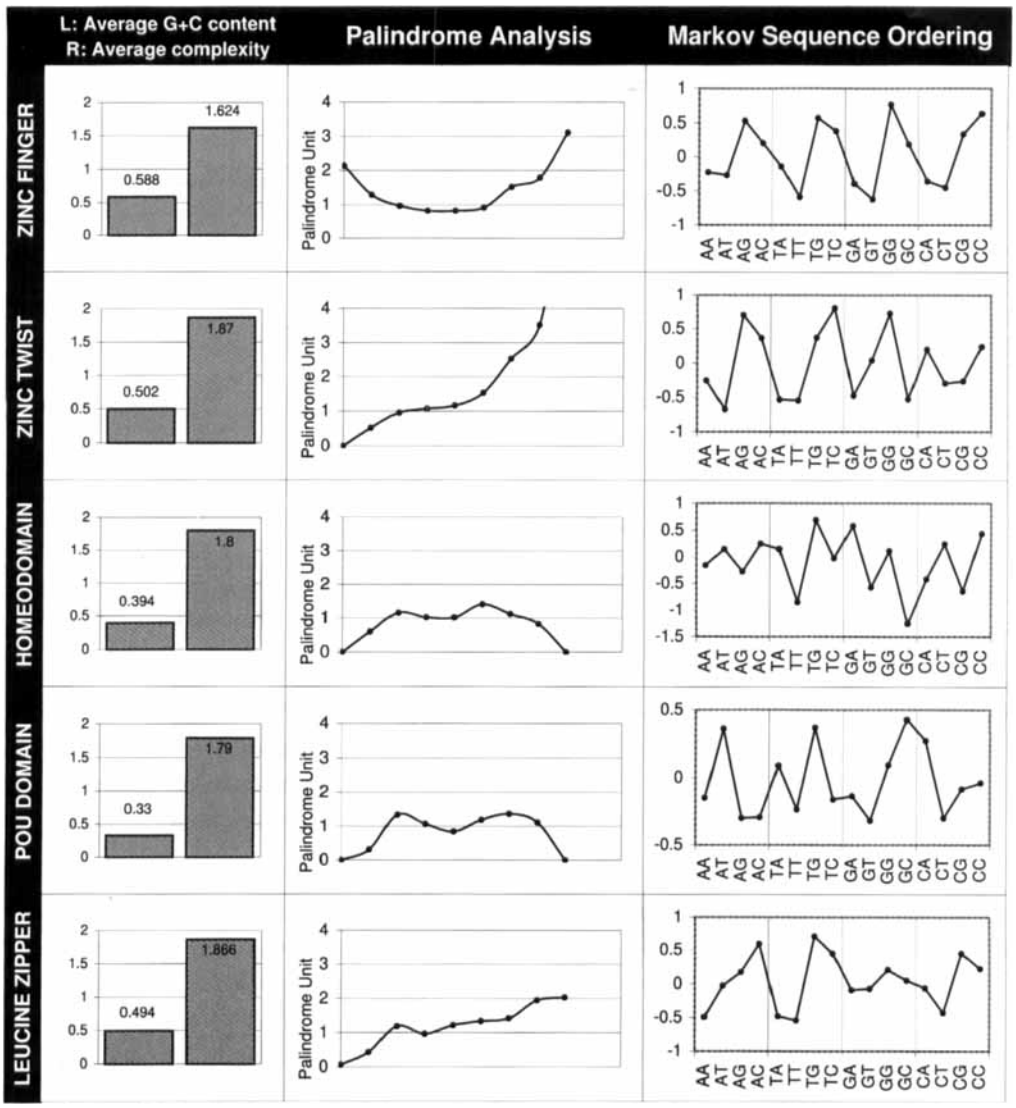

Fig. 5 TF effect. **A**. Average G+C content of the binding sites for different TF classes. Zinc finger proteins tended to bind the sequences with very high G+C content, whereas POU domain proteins bound to sequences with a G+C content comparable to background. **B**. Average pattern complexity of the binding sites for different TF classes. **C**. Odds ratio for "ACGT"-type palindrome between motif sequences and background. Zinc twist proteins bind to sequences having a palindromic structure with a 7- or 8-bp unit, whereas POU domains bind to sequences without palindromic structures. **D**. Log odds ratios of transition probability for three representative TF classes.

sequences had the highest enrichment, while the enrichment of the human and murine genomes was almost identical (Figure 6A). This pattern appears to reflect a difference in background characteristics, since the relative G+C contents of background sequences in yeast, murine, and human coding regions are 35%, 41%, and 40%, respectively. Interestingly, these numbers are slightly different from the overall genomic G+C content (38%, 42%, and 41%, respectively), with the G+C content distribution along the human genome being the broadest (*24*).

The pattern complexity distributions for the murine and human genomes were very similar to that for yeast (Figure 6B). For "ACGT"-type palindromes, the murine (especially) and human genomes showed a higher enrichment for palindromes with unit sizes of 6–8 (Figure 6C). The results for smaller palindromes in the three genomes were almost identical.

The Markov sequence ordering results for the murine and human genomes were similar, with a correlation coefficient of 0.98 (Figure 6D). However, they were quite different from the results for yeast, with

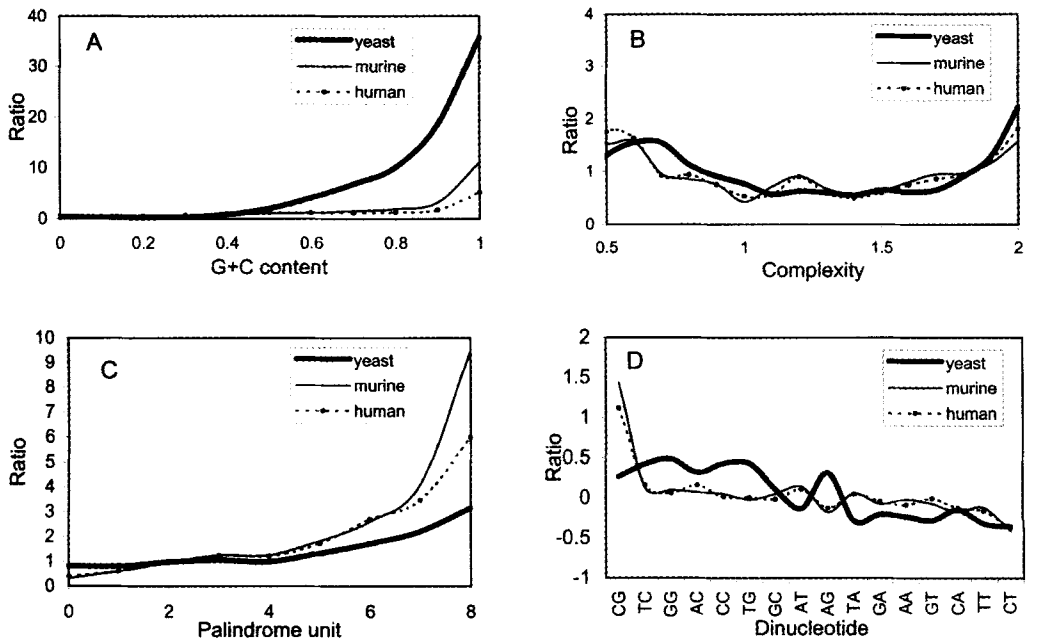

Fig. 6 Species specificity. **A.** Odds ratio for G+C content for human, murine, and yeast coding sequences. It is clear that the high G+C content feature of yeast motifs was much more significant than that in the human and murine samples. **B.** Odds ratio for pattern complexity. This feature was almost indistinguishable among the three species. **C.** Odds ratio for "ACGT"-type palindromes. For this feature, the motifs in the human and murine genomes tended to have more palindromic structures than did those in yeast. **D.** Odds ratios for transition probability.

correlation coefficients 0.26 and 0.17 between human and yeast and between murine and yeast, respectively. The Markov sequence orderings for human and murine samples were not as strong as those for yeast, as indicated by smaller oscillations in the curve. One exception was that the $\log\left(R(C \rightarrow G)\right)$ values were much higher for human (1.11) and murine (1.43) samples than for yeast (0.27).

In summary, the results of the feature analysis were somewhat species-dependent, although the differences were largely quantitative rather than qualitative. The species specificity seemed to arise mainly from the variation in the properties of the intergenic regions from the different genomes.

Discussion

Recently, there has been increasing interest in using bioinformatics approaches both to identify cis-acting DNA regulatory regions within the genome and to provide new insights into the nature of protein-DNA interactions (*25*, *26*). A number of programs have been developed that search for conservation either across species and/or between similarly regulated genes within a species (*10*, *19*, *27*, *28*). Although useful, such programs are challenged by the small size of DNA binding sites, their degeneracy, and the huge size of the surrounding background (non-regulatory DNA). As a result, the output of these programs often contains significant numbers of "false-positive" sequences that do not reflect biologically meaningful regulatory regions (*29*). In an effort to develop complementary approaches to aid in the identification of "true" regulatory regions, we have been exploring the inherent characteristics of known binding motifs.

In the present study, we investigated four sequence features associated with TF binding sites. Some features (for example, G+C content) were strongly enriched in binding sites when compared to "control" DNA, while others (for example, pattern complexity) were only weakly enriched. In terms of specific classes of protein-DNA interaction, we made several additional observations. For instance, zinc finger proteins tend to recognize sequences with high G+C content, while POU domains prefer relatively low G+C content. Also, some characteristics such as G+C content show significant species specificity, while some, like preference for complexity, tend to be more universal.

It should be noted that the four sequence features we analyzed are not totally independent of each other.

For example, G+C content and pattern complexity are correlated; sequences with extremely high (or extremely low) G+C content cannot have maximal pattern complexity. Also, Markov sequence ordering is partially a function of sequence composition. For example, high G+C content results in high $R(i \rightarrow G)$ and $R(i \rightarrow C)$. As noted, when the Markov sequence ordering results were corrected for G+C content, the differences between motif and background sequences were not so large.

As unpublished results, we attempted to predict TF binding sites based on these features using support vector machine (SVM) approach. We could predict regulatory motifs with an accuracy of 67% purely based on the information of these features. We believe that the features alone are not sufficient to allow us to reliably distinguish regulatory motifs from background. However, by integrating these features with more conventional methods, such as sequence alignment-based approaches that compare similarly regulated genes (*18*, *28*, *30*), it should be possible to improve the success of motif discovery algorithms and to develop more robust systems to detect, both accurately and with high-throughput, biologically relevant regulatory regions. A greater molecular understanding of the basis of protein-DNA interactions should also be helpful to this endeavor. For example, understanding why zinc finger proteins tend to recognize sequences with high G+C content, while POU domains prefer relatively low G+C content sequences, may aid in the design of "smarter" algorithms that are better at identifying biologically relevant binding sites. Ongoing x-ray crystallographic and other structural studies are clearly providing important information in this direction. The integration of such knowledge of protein-DNA interaction could transform the current motif discovery efforts from an unsupervised to a supervised approach. This change, together with an increasing availability of large datasets of co-regulated genes from microarray experiments and of sequence information from genome projects, will likely lead to substantial advancements in our understanding of the complex networks of regulatory regions that control gene expression.

Materials and Methods

Datasets

The binding motifs were extracted from the TRANSFAC database (*8*). In total, 11,700 motif sequences were used for this study. It is obvious that more than one binding site can be associated with one TF, indicating that our analysis takes into account the nature of degenerated sequences of TF binding sites. To avoid the redundancy, we excluded the consensus sequences that already contain degenerated positions. We also constructed a set of random sequences as background. The background dataset was generated from yeast intergenic sequences, which are defined as the regions between two adjacent genes. The intergenic sequences were obtained from the *Saccharomyces* Genome Database (*31*). Short sequence segments were "cut" from the intergenic sequences at randomly selected locations. The lengths of the short segments were chosen such that this set of short sequences had the same length distribution as the motif dataset. Overall, this background set contained 1,170,000 short sequence segments. Since the background consisted of a large amount of intergenic segments, we expected the heterogeneities from various intergenic sequences, for example, the composition difference from "divergent promoters" and "convergent terminators" (*22*) to be "averaged" out. For the murine and human genomes, the coding sequences were obtained from twinscan track in UCSC genome browser (*32*). To select a set of short sequences from murine and human intergenic regions as background, the same procedure was applied as for the yeast genome.

Odds ratio

The odds ratio, defined as the enrichment of a particular feature in the motif database with respect to the random expectation for the occurrence of that feature, was calculated by the equation:

$$R = \frac{P(feature|motif)}{P(feature)}$$

where $P(feature|motif)$ is the probability of occurrence of a motif with a certain feature (for example, GC=0.8) in the motif database, and $P(feature)$ is the probability of occurrence of a random sequence segment from the genomic intergenic sequences having the same feature. Since regulatory elements represent only a very small portion of all the intergenic sequences, this probability is essentially the same as that for non-motifs. Therefore, the odds ratio is essentially used to compare the probability of a certain event for two groups (binding motifs versus random sequences).

Bootstrap simulation

A bootstrap technique was employed to test the statistical significance of the observations. Instead of comparing motif dataset versus background, a set of random sequences was compared with background. This set of random sequences was extracted from intergenic sequences, and the dataset consisted of the same number of sequences as in the motif dataset. Since the background also consisted of random sequences, the bootstrap simulation was essentially a comparison of "random" versus "random", which provided the range of random fluctuation around the expected value of 1. The procedure was repeated 1,000 times, and each time a new set of random sequences was extracted and compared with the background. The 1,000 odds ratios obtained were sorted, and the 5^{th} and 95^{th} percentiles of the values were considered to define the range of random fluctuations (confidence intervals). If the feature values fell into the confidence intervals, the observations were considered random events and not significant.

Acknowledgements

We thank Dongmei Liu and Dr. Giovanni Parmigiani for stimulating discussions. The research is supported in part by grants from the National Eye Institute and the Foundation Fighting Blindness, and by a generous gift from Mr. Robert Smith and Mrs. Clarice Smith. DJZ is the Guerrieri Professor of Genetic Engineering and Molecular Ophthalmology and the recipient of a Research to Prevent Blindness Senior Investigator Award.

Authors' contributions

JQ conceived the study, performed the analysis, and wrote the paper. JL helped to analyze the data and write manuscript. DJZ provided the biological consultation and wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Alberts, B., *et al.* 1994. *Molecular Biology of the Cell* (third edition). Garland Publishing, New York, USA.
2. Tupler, R., *et al.* 1999. Profound misregulation of muscle-specific gene expression in facioscapulohumeral muscular dystrophy. *Proc. Natl. Acad. Sci. USA* 96: 12650-12654.
3. Ly, D.H., *et al.* 2000. Mitotic misregulation and human aging. *Science* 287: 2486-2492.
4. Sioud, M. 2004. Therapeutic siRNAs. *Trends Pharmacol. Sci.* 25: 22-28.
5. Lee, T.I., *et al.* 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
6. Horak, C.E., *et al.* 2002. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* 16: 3017-3033.
7. Liu, X., *et al.* 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* 15: 421-427.
8. Wingender, E., *et al.* 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* 29: 281-283.
9. Sandelin, A., *et al.* 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32: D91-94.
10. Wasserman, W.W., *et al.* 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26: 225-228.
11. Kellis, M., *et al.* 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241-254.
12. Cliften, P., *et al.* 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71-76.
13. Tavazoie, S., *et al.* 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22: 281-285.
14. Benos, P.V., *et al.* 2002. Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.* 323: 701-727.
15. Suzuki, M. 1994. A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure* 2: 317-326.
16. Benos, P.V., *et al.* 2002. Is there a code for protein-DNA recognition? Probab(ilistical)ly... *Bioessays* 24: 466-475.
17. Grundy, W.N., *et al.* 1997. Meta-MEME: motif-based hidden Markov models of protein families. *Comput. Appl. Biosci.* 13: 397-406.
18. Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant

alignments of multiple sequences. *Bioinformatics* 15: 563-577.

19. Hughes, J.D., *et al.* 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296: 1205-1214.
20. Yamada, Y., *et al.* 2004. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res.* 14: 247-266.
21. Antequera, F. and Bird, A. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* 90: 11995-11999.
22. Dujon, B., *et al.* 1994. Complete DNA sequence of yeast chromosome XI. *Nature* 369: 371-378.
23. Vallee, B.L., *et al.* 1991. Zinc fingers, zinc clusters, and zinc twists in DNA-binding protein domains. *Proc. Natl. Acad. Sci. USA* 88: 999-1003.
24. Waterston, R.H., *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.
25. Arnosti, D.N. 2003. Analysis and function of transcriptional regulatory elements: insights from *Drosophila*. *Annu. Rev. Entomol.* 48: 579-602.
26. Halfon, M.S. and Michelson, A.M. 2002. Exploring genetic regulatory networks in metazoan development: methods and models. *Physiol. Genomics* 10: 131-143.
27. Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 28-36.
28. Liu, X.S., *et al.* 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* 20: 835-839.
29. Tompa, M., *et al.* 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23: 137-144.
30. Roth, F.P., *et al.* 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16: 939-945.
31. Cherry, J.M., *et al.* 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* 26: 73-79.
32. Kent, W.J., *et al.* 2002. The human genome browser at UCSC. *Genome Res.* 12: 996-1006.