Inferring Natural Selection on Fine-Scale Chromatin Organization in Yeast

G. A. Babbitt* and Y. Kim*†

*Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University; and †School of Life Sciences, Arizona State University

Despite its potential role in the evolution of complex phenotypes, the detection of negative (purifying) and positive selection on noncoding regulatory sequence has been elusive because of the inherent difficulty in predicting the functional consequences of mutations on noncoding sequence. Because the functioning of regulatory sequence depends upon both chromatin configuration and cis-regulatory factor binding, we investigate the idea that the functional conservation of regulatory regions should be associated with the conservation of sequence-dependent bending properties of DNA that determine its affinity for the nucleosome. Recent advances in the computational prediction of sequence-dependent affinity to nucleosomes provide an opportunity to distinguish between neutral and nonneutral evolution of fine-scale chromatin organization. Here, a statistical test is presented for detecting evolutionary conservation and/or adaptive evolution of nucleosome affinity from interspecies comparisons of DNA sequences. Local nucleosome affinities of homologous sequences were calculated using 2 recently published methods. A randomization test was applied to sites of mutation to evaluate the similarity of DNA–nucleosome affinity between several closely related species of Saccharomyces yeast. For most of the genes we analyzed, the conservation of local nucleosome affinity was detected at a few distinct locations in the upstream noncoding region. Our results also demonstrate that different patterns of chromatin evolution have shaped DNA–nucleosome interaction at the core promoters of TATA-containing and TATA-less genes and that elevated purifying selection has maintained low affinity for nucleosome in the core promoters of the latter group. Across the entire yeast genome, DNA–nucleosome interaction was also discovered to be significantly more conserved in TATA-less genes compared with TATA-containing genes.

Introduction

Comparison of DNA sequences from different species is frequently made to infer functionally important elements in the genome (e.g., McGuire and Church 2000; Kellis et al. 2003; Boffelli et al. 2004; Hughes et al. 2005). Mutations on nucleotide sequences involved in biological functions are mostly deleterious because they are likely to destroy the optimum sequence achieved in the organism’s history of adaptation. Negative (purifying) selection against such mutations results in the “conservation” of functional DNA sequences relative to genomic regions that evolve without constraint. Population genetics and molecular evolutionary studies have advanced this principle to detect genomic elements that are under not only negative or purifying selection but also positive selection: if the rate of nucleotide substitution is estimated to be higher than expected under neutral evolution (evolution due to mutations with little functional consequence), such evolutionary change is inferred to be driven by positive Darwinian selection.

This principle of detecting negatively and positively selected sequences was most successfully applied to protein-coding sequences. Mutations on coding regions are either nonsynonymous (causing amino acid replacement) or synonymous (no replacement). Synonymous mutations are generally considered neutral. Therefore, a slow or fast nonsynonymous substitution rate relative to the synonymous rate indicates negative or positive selection on amino acid sequence, respectively. Recent genomic analysis revealed extensive negative and positive selection on protein-coding regions in many species (Smith and Eyre-Walker 2002; Bustamante et al. 2002; Fay et al. 2002; Clark et al. 2003; Nielsen et al. 2005). These surveys identified genes that are likely to cause genetic disease if mutated or to have contributed to species’ adaptation.

However, detection of selection on noncoding sequence is inherently more difficult. Although the importance of noncoding regulatory elements in the conservation and adaptation of biological processes has long been appreciated (King and Wilson 1975), there is no clear rule of functional organization that governs the regulatory sequences of arbitrary genes. Numerous probabilistic methods for predicting regulatory elements do exist (e.g., Benos et al. 2002; Boffelli et al. 2003; Dermitzakis et al. 2003; Ovcharenko et al. 2004). However, it is generally difficult to identify noncoding nucleotide sites that are expected to evolve without functional constraint and therefore can serve as a neutral reference like synonymous sites on coding regions. Conservation of transcription factor binding (TFB) motifs might be identified in the multiple alignments of intergenic sequences (Hahn 2007). However, the examination of sequence divergence on well-characterized TFB versus non-TFB sequences often fails to show nonneutral evolution of TFB sequences (Ludwig and Kreitman 1995; Ludwig 2002). Even when the expression pattern of a transcript is well conserved, rapid divergence and turnover of cis-regulatory elements (loss of old TFB sequences and gain of new TFB sequences) were observed (Ludwig et al. 2000; Wray et al. 2003). Despite this difficulty in identifying functional regulatory elements in sequence alignments, extensive functional constraint in noncoding sequence is strongly suggested. Sequence divergence on introns, untranslated transcribed regions, and intergenic regions is generally lower than that on synonymous sites in the coding region (Li 1997; Subramanian and Kumar 2003). Jointly analyzing within-species polymorphism and between-species divergence, Andolfatto (2005) suggested that a large fraction of noncoding sequences in Drosophila is involved in regulatory function and is subject to both negative and positive selection. This remarkable inconsistency, namely the prediction of the high fraction of functionally important sites,
but difficulty in finding sequence conservation, suggests that the regulatory function of noncoding DNA requires more than simple conservation of cis-regulatory motifs.

Unlike coding DNA, which acts as the medium of information needed for protein structure and function, regulatory noncoding DNA physically interacts with other cellular components to modulate gene expression. One may thus argue that conservation of certain physical properties of noncoding DNA is more important than the conservation of base sequence itself. In addition to binding affinity to a collection of transcription factors, there are other physical requirements for regulatory sequence, including its ability to wrap around histone octamers, bend to make loops for transcription factor complexes, and make transition between double- and single-stranded DNA (Turner 2001). Many of these properties of DNA are likely to be determined by base sequence in a context-dependent manner, namely, the correlated occurrence of certain sets of nucleotides at certain positions and order along the sequence would achieve desired properties of bending, twisting, and unwinding (Trifonov and Sussman 1980; Widom 1996; Trifonov 1998; Herzl et al. 1999; Fukushima et al. 2002; Dalal et al. 2005). It is believed that this periodicity in base sequences influences the bending property of DNA double strand. In all, 10 or 11 bp correspond approximately to one helical turn of DNA strand. Therefore, adjacent base pairs that are tilted in the same direction should be placed at this periodicity to achieve the overall bending of DNA double strand within one geometric plane. This bending property is believed to be crucial for the ability of DNA to form nucleosomes (Ioshikhes et al. 1996; Stein and Bina 1999; Levitsky et al. 2001; Widom 2002). Each nucleosome contains a 147-bp stretch of DNA, which is tightly wrapped around a histone protein octamer. Within this stretch, sharp bending occurs repeatedly when the major groove of DNA double helix faces toward the histone octamer. Ioshikhes et al. (1996) compiled a large number of nucleosome-associated sequences from various organisms in the literature and discovered a roughly 10-bp periodic pattern of dinucleotides AA/TT. From this work, Ioshikhes et al. (2006) developed a method of predicting nucleosome positioning by calculating correlation of a given pattern of AA/TT dinucleotides and the periodic pattern of AA and TT positional bias obtained from 204 published sequences of nucleosome-bound DNA. In another study, Segal et al. (2006) experimentally isolated a large number of yeast DNA segments that are wrapped stably in nucleosomes and discovered a clear pattern of ~10-bp periodic AA/TT/TA dinucleotides, which is out of phase with similar period in GC. Disruption of this periodic pattern greatly reduced the binding affinity to nucleosomes in vitro. This discovery led to a probabilistic model of sequence-dependent nucleosome affinity that was used to predict stable nucleosome positions along the yeast genome. Unlike the method of Segal et al. (2006), AA and TT dinucleotides in the method of Ioshikhes are not considered functionally equivalent but counted separately, exhibiting opposite asymmetric distribution around the center of 139-bp nucleosomal DNA. Ioshikhes et al. (2006) claim that this definition predicts nucleosomal DNA better than the method of Segal et al. (2006) when the result was compared with experimentally determined positions of nucleosomes in Saccharomyces cerevisiae (Yuan et al. 2005). However, nucleosome positions predicted by both methods of Ioshikhes et al. (2006) and Segal et al. (2006) significantly correlated with known positions in the literature. Therefore, it is clearly suggested that nucleosome positioning in yeast can be inferred, though not perfectly, by calculating the affinity score along sequence. Levitsky et al. (2001) also recognized similar periodicities on nucleosomal DNA and proposed different statistical approaches to predict nucleosome positions.

The nucleosome is the unit of DNA packaging that creates a dynamic chromatin structure. It is generally assumed that nucleosomes influence gene regulation as repressors by hindering the access to regulatory elements (Paranjape et al. 1994; Gregory and Horz 1998; Workman and Kingston 1998). Therefore, the positions of nucleosomes relative to the location of cis-acting elements are critical to gene regulation and even appear to define broad functional classes of genes (Ercan and Lieb 2006; Ioshikhes et al. 2006). In agreement with the repressive role of nucleosomes, the upstream sequences of tissue-specific genes showed higher nucleosome formation potential than housekeeping genes in humans (Levitsky et al. 2001; Ganapathi et al. 2005). However, a yeast experiment that induces genome-wide depletion of nucleosomes resulted in both increase and decrease of gene expressions, indicating that the role of nucleosomes in gene regulation is more than a simple repressive one (Wyrick et al. 1999).

We reason that, if the fine-scale chromatin organization is crucial for biological functions that are under natural selection, the conservation or adaptive change in local nucleosome affinity should be observed through the conservation or change in the predicted affinity of DNA for nucleosomes between species. Detecting natural selection on nucleosome affinity is not, however, a simple task. Similarity in affinity for 2 homologous sequences can be due to the evolutionary conservation of affinity or simply due to the corresponding similarity of the base sequences. Here we propose the first test for detection of negative selection against, or positive selection for, mutations affecting local nucleosome affinity for arbitrary pairs of homologous sequences. Our method is fundamentally an application of the sampled randomization test (Sokal and Rolph 1995, section 18.3, p. 803–820) to DNA sequence and forms the basis for a statistical comparison of the observed divergence in the nucleosome affinity of DNA to the expected divergence under a neutral evolution model in which mutations occur in random positions. This test could theoretically be applied to any computable property of a DNA sequence. But its application here provides a unique opportunity to survey a genome-wide pattern of sequence evolution associated with the role of chromatin in gene regulation in Saccharomyces yeasts.
Materials and Methods
Source of Nucleosome Formation Metrics

In this study, we use 2 methods of predicting DNA sequence’s affinity to nucleosomes: the nucleosome positioning sequence (NPS) correlation of Ioshikhes et al. (2006) and the (nucleosomal) “free energy” calculation of Segal et al. (2006). We abbreviate these methods as the Ioshikhes–Pugh (IP) method and Segal–Widom (SW) method using the initials of the first and last authors of the papers describing them. Each computation is described in more detail in Appendix A. Data for the NPS pattern (figure 1a in Ioshikhes et al. [2006]) for IP method were kindly provided by Dr Ilya Ioshikhes, Department of Bio-medical Informatics, The Ohio State University. The experimental data on which the SW method is based are offered freely by Dr Eran Segal at http://genie.weizmann.ac.il/pubs/nucleosomes06/segal06_data.html.

Summary of the Computational Method

The computational method for inferring purifying or positive selection upon nucleosome affinity is employed within a sliding window and can be summarized as a 4-step process as follows. More detailed description of the procedure is given in Appendix A.

1. Homologous sequences from 2 species (A and B) are aligned. The unit of analysis is a sliding window (denoted R-window) on the alignment that covers ~174 (using IP method) or 176 (using SW method) bases in species A. Homologous sequences included in this window are termed sequence A and B.

2. Using either IP or SW method, a nucleosome affinity score (σ) is determined over a 139-bp (IP) or 141-bp (SW) long sequence (Appendix A). Of 35 (=174 – 139; IP) values of σ that can be calculated within sequence A, we take the “average” or maximum. We define this average (or maximum), σ_A^{avg} (σ_A^{max}), as the local affinity of sequence A. The local affinity of sequence B, σ_B^{avg} is similarly determined. Consequently, |σ_A^{avg} – σ_B^{avg}| is the local measure of the absolute difference in nucleosome affinity that exists between 2 species and is a benchmark for which the effect of randomization (step 3) can be compared.

3. All mutation events observed between sequence A and B are randomly repositioned within sequence A, creating a “randomized” sequence (termed sequence RA). The level of sequence divergence between A and RA is therefore the same as between A and B. The same randomization procedure is independently performed on sequence B producing RB. If regular periodicity of certain dinucleotides has been maintained by selection, this randomization process should break the pattern and allow the detection of selection. Five hundred replicates of RA and RB sequences (250 each) are produced (see justification of number of replications in Appendix B), and the local affinity is calculated for each sequence RA and RB.

4. The absolute difference in the local nucleosome affinity between a sequence from one species and its randomized sequence (i.e., |σ_A^{avg} – σ_{RA}^{avg}| or |σ_B^{avg} – σ_{RB}^{avg}|) is computed and compared with the existing absolute difference between the 2 species, |σ_A^{avg} – σ_B^{avg}|. If nucleosome affinity is governed by neutral evolution (substitutions of mutations occur at random positions) in the evolutionary lineages leading to species A and B, then we expect no difference on average between the observed difference and the randomized differences in nucleosome affinity (fig. 1). However, under purifying selection (functional conservation of the local affinity), the former is expected to be smaller than the latter. Therefore, to detect purifying selection P values are determined by counting the frequency of randomizations for a given sample size which create

\[ |σ_A^{avg} – σ_{RA}^{avg}| > |σ_A^{avg} – σ_B^{avg}| \text{ or } |σ_B^{avg} – σ_{RB}^{avg}| > |σ_A^{avg} – σ_B^{avg}|. \]

Similarly, the P value for positive directional selection is computed (above inequalities reversed; fig. 1). For a given sliding window, a P value < 0.025 rejects the null hypothesis of neutral evolution in favor of the corresponding type of selection in that region.

Simulation of Neutrally Evolving Sequences

We simulate the evolution of DNA sequences upon which we evaluate the performance of our sampled randomization test. A random DNA sequence of 100 kb (putative ancestral sequence) was generated using the same dinucleotide composition of noncoding regions in the S. cerevisiae genome. This sequence is diverged in 2 directions by adding single base substitutions and insertion/deletions at random locations. The substitution to indel ratio is arbitrarily set to 9:1. Mutations are introduced at differing extent to simulate differences in mutation rate or divergence time. Indel size is exponentially distributed with mean and variance of 2. Two substitution models, Jukes–Cantor (JC69; Jukes and Cantor 1969) model and Hasegawa–Kishino–Yano (HKY85; Hasegawa et al. 1985) model, are used for sequence evolution. Under the simulations using the HKY substitution model, a 5 to 1 transition to transversion rate with 30% G/C content is used. These parameters are intended for examining the performance of our test under an extreme transition/transversion ratio and high A/T richness (as observed in yeast noncoding regions).

Source of Yeast Alignments

The yeast sequence alignment data used are also offered freely by Dr Manolis Kellis at http://www.broad.mit.edu/annotation/fungi/comp_yeasts/. Each alignment file is consisted of multiple alignments of sequences (ClustalW) covering one open reading frame (ORF) and variable lengths of upstream and downstream noncoding regions. Visual inspection of sequential files along chromosomes revealed considerable overlap between noncoding regions of adjacent files, indicating good coverage by the alignments of the noncoding portion of the yeast genome.
Purifying Selection

Neutral Drift

Directional Selection

Fig. 1.—Three possible outcomes of the sampled randomization test (i.e., random relocation of mutations between homologous sequences) for detecting selection on DNA–nucleosome affinity. In each case, the absolute difference in affinity between 2 observed homologous sequences \[|\sigma_{12}-\sigma_{21}|\] are shown in comparison to the distribution of absolute differences between real and randomized sequences \[|\sigma_{wk}-\sigma_{kj}|\] for arbitrary \(j, k = 1, \ldots, n\). Regions of CYC1 gene corresponding to each outcome were used to generate the distributions in the figure.

Results

Simulation of Neutral Evolution on Chromatin Organization

We evaluated the performance of our sampled randomization test for detecting selection on nucleosome affinity, as measured by both the SW and IP methods, by using sequences that evolved neutrally with respect to nucleosome affinity starting from a random sequence. The rate of detecting selection is defined as the proportion of R-windows (out of all tested R-windows that partially overlap) from which the corresponding type of selection was detected. When the test was applied to these simulated sequences, the frequency of the detection of either form of selection, thus false positive rate, was close to the significance level of the randomization test (\(\alpha = 0.025\)) for both SW and IP methods (table 1). False positive results for both types (purifying and positive) of selection averaged to 2.7%. The level of sequence divergence did not affect the level of false positive results in neutral simulations. There were also no significant differences dependent upon which model of substitution, JC69 or HKY85, was used.

Performance of Randomization Test Using Different Methods

To explore the most effective way for detecting selection on nucleosome positioning, we first examined the performance of our randomization test using different combinations of methods: IP versus SW methods and taking average versus maximum for determining local affinity. Both IP and SW methods rely upon a 10- to 11-bp periodicity of position-dependent probabilities of the dinucleotides AA and TT. However, in the SW method, which includes information about all other dinucleotides, the periodicity of the dinucleotide TA is also a major determinant of nucleosome affinity. As we would expect, there is a highly significant but weak correlation between the NPS patterns of AA and TT (IP method) and the position-dependent probabilities of AA and TT in the SW method (for AA; \(r = 0.276, P = 0.003\) and for TT; \(r = 0.339, P = 0.004\)). This weak correlation suggests that the inferences of nucleosome positioning using 2 methods, thus the results of randomization tests, can be discordant.

We scanned all genes reported by Kellis et al. (2003) on chromosome 8 (=chromosome H) with our randomization test and determined the rate of detecting purifying selection in core promoter regions. The comparison of the overall detection of conservation of each method is given in table 2. We found that the IP method detected more purifying selection than the SW method, especially near the core promoters (−200 bp to +100 bp relative to the start codon). Further comparison of the IP method at various genetic distances comparing S. cerevisiae to 2 other congeners, Saccharomyces mikatae and Saccharomyces bayanus, revealed that detection of purifying selection was highest when comparing S. cerevisiae to S. mikatae (table 2).

Another advantage of the IP method is its relative simplicity and faster calculation. We therefore use this method, with local average affinity, for the following data analyses. We also note that the locations of purifying selection appear more clustered over small regions of many genes, which may indicate less spurious detection, when using the IP method.

Single Gene Examples

We demonstrate the result of tests applied to the upstream regions of 4 well-known genes in comparisons between S. cerevisiae and 2 sister species (Saccharomyces paradoxus and S. mikatae). Figures 2–4 show the P value of the randomization test along with the profile of nucleosome positioning signals calculated by the IP method. These genes are 1) CYC1, a gene coding for iso-1-cytochrome c and induced by heme protein in the presence of O2 (fig. 2); 2) GAL1, a gene coding for a galactokinase
induced in the presence of galactose (fig. 3); and 3) PHO5, a gene induced under low phosphate conditions and undergoes extensive chromatin remodeling during activation (fig. 4). (References for the locations and functioning of regulatory elements are given in the figure captions.)

In core promoter regions of all genes listed above, nucleosome affinity fluctuates along the sequence. There is no general correlation between the level of local nucleosome affinity and the detection of selection. However, we notice that the areas of distinctly high or low nucleosome affinity often correspond to the position where purifying selection is detected. For example, around the position −500 bp of CYC1 gene, nucleosome affinity is very high in all 3 species of yeast and the randomized sequences consistently yield lower affinity, thus producing a strong signal of functional conservation (fig. 2). In GAL1 gene, a narrow region of a very low nucleosome affinity exists near the transcription start site. Purifying selection that maintains this low affinity is observed in all 3 pairs of species comparison (fig. 3). In PHO5, which is a classic textbook example of a chromatin-repressed gene, we find significant functional conservation of high nucleosome affinity in the regions immediately flanking the 4 nucleosomes that are known to be removed during gene activation (Turner 2001, p. 120) (fig. 4). The level of nucleosome affinity throughout remodeled region fluctuates near zero (average affinity) and shows no statistically significant conservation.

Whole-Genome Analysis in Yeast

We sequentially analyzed all yeast ORF alignment files in Kellis et al. (2003) using the randomization test. We limited our analysis to the region covering from 600 bp upstream to 600 bp downstream of the translation start site of each gene and to the alignment of S. cerevisiae and S. mikatae. Genes were selected if alignment of promoter region had less than 30% gap. In this whole-genome scan, we are particularly interested in comparing the profile of natural selection in TATA-containing (or TATA(+) versus TATA-less (or TATA(−)) genes, as classified by Basehoar et al. (2004), because clear division in the functional organization of nucleosome positioning between these 2 classes of genes were reported (Ioshikhes et al. 2006). In total, 882 TATA(+) and 3,759 TATA(−) genes were analyzed. The results of individual gene scans were aligned to the start codon position and averaged in each gene class (TATA(+) or TATA(−)) to produce figures 5 and 6. (The upstream noncoding region in each alignment file may occasionally be shorter than 600 bp. Therefore, the results in figs. 5 and 6 are the averages among genes in which the sequences at the corresponding position exist.) We averaged 1) the local nucleosome affinity of S. cerevisiae across all gene regions and 2) this local affinity conditioning upon the detection of conservation by the randomization test. Then, we recorded the rate of detecting purifying selection (fig. 5) and positive selection (fig. 6) for each location.

First, interesting patterns emerged from the statistical comparisons of detection rates of functional conservation between TATA(+) versus TATA(−) genes (fig. 5). The overall percentage of functional conservation across all coding and noncoding regions was fairly low but differed significantly for TATA(+) and TATA(−) genes (TATA(+) = 2.9% and TATA(−) = 6.3%, t = 20.67, P < 0.0001). The detection rate peaked to 5.1% for

---

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% Neutral</td>
<td>% (−)a</td>
</tr>
<tr>
<td>JC69</td>
<td>5</td>
<td>94.9</td>
<td>2.5</td>
</tr>
<tr>
<td>JC69</td>
<td>10</td>
<td>94.6</td>
<td>2.7</td>
</tr>
<tr>
<td>JC69</td>
<td>5</td>
<td>94.7</td>
<td>2.6</td>
</tr>
<tr>
<td>HKY85</td>
<td>5</td>
<td>94.8</td>
<td>2.6</td>
</tr>
<tr>
<td>HKY85</td>
<td>10</td>
<td>94.9</td>
<td>2.6</td>
</tr>
<tr>
<td>HKY85</td>
<td>15</td>
<td>94.9</td>
<td>2.6</td>
</tr>
</tbody>
</table>

*NOTE.—Results of 2 versions of randomization test on randomly diverged 10-kb random DNA sequences. Nucleosome affinity was calculated both as the free energy score in Segal et al. (2006) and the NPS correlation in Ioshikhes et al. (2006). All simulations use JC or HKY model for simulating substitution events. In HKY model, the transition/transversion ratio was 5 to 1. n = 250.

a Negative selection or conservation.

b Positive selection.

---

**Table 2**

<table>
<thead>
<tr>
<th>Species Divergence</th>
<th>Smoothing Method</th>
<th>IP Method</th>
<th>SW Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>Average</td>
<td>0.055 (0.057)</td>
<td>0.037 (0.034)</td>
</tr>
<tr>
<td>S. paradosus (0.22)</td>
<td>Maximum</td>
<td>0.046 (0.045)</td>
<td>0.025 (0.026)</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Average</td>
<td>0.102 (0.059)</td>
<td></td>
</tr>
<tr>
<td>S. mikatae (0.30)</td>
<td>Average</td>
<td>0.044 (0.044)</td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Average</td>
<td>0.044 (0.044)</td>
<td></td>
</tr>
</tbody>
</table>

*NOTE.—Comparison of mean frequency of detection of purifying selection upon chromatin organization in both core promoters measured as −200 to +100 bp relative to start codon (and total region −600 to +600 bp) of all genes on yeast chromosome 8 (B) using randomization tests based upon different genetic distances and different methods of calculating and smoothing the nucleosome affinity scores. The 2 methods of calculating affinity score are those of Segal et al. (2006) or SW method and Ioshikhes et al. (2006) or IP method. The 2 methods of smoothing the calculations during scanning of sequences were a 20 base (m = 20) local average score and local maximum score. Synonymous nucleotide divergence (Ks) for whole-genome comparisons (Gao and Innan 2004) is given in left margin.
TATA(+) and 8.5% for TATA(−) at approximately +100 bp and −150 bp (relative to the start codon), respectively (fig. 5). Results also indicate that purifying selection on nucleosome affinity occurs at rates consistently higher than expected under neutrality (P < 0.025) across all regions of TATA(−) genes. In these genes, the levels of purifying selection in the promoter and coding regions are similar outside of the main peak occurring between the start codon and 200 bp upstream. On the other hand, in TATA(+) genes, the only region exhibiting significant functional conservation (detection rate >0.025) is the short region immediately at and after the start codon.

In addition to this difference in the overall level of purifying selection between TATA(+) and TATA(−) genes, there are general differences in the nature (direction) of functional conservation. TATA(+) genes exhibit functionally conserved average affinity for the nucleosome in the narrow region just after the start codon, whereas TATA(−) genes exhibit functional conservation of very low affinity for roughly 200 bp upstream of the start codon (fig. 5). In TATA(−) genes, this low affinity region is flanked by 2 regions of high affinity for nucleosome that are apparently not under as high a rate of purifying selection. This pattern of conserved nucleosome affinity in the promoter of TATA(−) genes would serve to impede nucleosome formation in the region of transcription initiation.

We similarly analyzed the positive directional selection on nucleosome affinity in the yeast genome (fig. 6). In all regions of TATA(−) genes and most regions of TATA(+) genes, the rate of detecting positive selection was near or below the significance level (0.025). Interestingly, the overall detection rate was higher in TATA(+) genes (2.8%) than in TATA(−) genes (1.5%) (t = 11.50, P < 0.0001). In TATA(+) genes, the detection rate for positive selection peaked (5.0%) in the 150-bp region immediately following the start codon.

We also compared the average rates of detecting both purifying and positive selection on chromatin across all 16 yeast chromosomes. Chromosomes 3, 8, 9, and 13–15 exhibited elevated levels of purifying selection (F = 2.59, P = 0.001 for TATA(+) genes and F = 61.44, P < 0.001 for TATA(−) genes), whereas chromosomes 1, 5, 6, and 10–12 demonstrated higher levels of positive selection (F = 4.35, P < 0.001 for TATA(+) genes and F = 27.98, P < 0.001 for TATA(−) genes).

There was no overall correlation between sequence divergence, as measured by local relative entropy, and the rates for detecting purifying or positive selection nor was there any general correlation between nucleosome affinity and degree of selection. Therefore, selection appears able to act on any level of nucleosome affinity by affecting critical bases (at 10-bp periodicities) without strongly affecting
sequence divergence patterns governed largely by 3-bp periodicities in coding regions.

**Discussion**

The main objective of this study is to examine whether the fine-scale chromatin structure, which is an essential component of eukaryotic gene regulation, plays a role in DNA sequence evolution. We proposed a statistical test of detecting natural selection on nucleosome positioning based on recently developed computational predictions of nucleosome positioning. This method detected both purifying and positive selection throughout yeast genome, thus supporting the idea that natural selection on gene regulation should result in conservation or adaptive change in nucleosome positioning. However, although the overall rate of detecting purifying selection is higher than expected under the null model of neutral evolution, this overall rate is rather low (figs. 3 and 4). This raises concerns that there might be no selection on nucleosome positioning but our result might be due to an unknown feature of yeast gene sequences that generates false positives in the test or that there may be universal selection on nucleosome positioning but our method had low statistical power. However, in the following, we offer several potential explanations for why our method should yield this level of detection rate throughout in yeast genome. Most importantly, we will argue that the observed pattern of purifying selection, which shows clear nonrandom correlations to the features of functional organization in yeast genes, strongly suggests that our result is not a statistical artifact but truly reflects evolutionary events in nature.

There are several possible biological reasons why selection on nucleosome positioning should be detected within narrow and intermittently spaced regions in the genome. First of all, accurate nucleosome positioning or exclusion may be required only at specific functional sites in promoters, and these sites usually comprise only a small fraction of the total promoter region. This tendency is strongly suggested by the evolutionarily conserved patterns of high nucleosome affinity we observe in the upstream noncoding region of the CYC1 gene and also the region of nucleosome exclusion near the TATA box in the GAL1 gene. Second, in regions subject to chromatin remodeling, the selective pressure for maintaining nucleosome affinity might be very weak. In genes like PHO5, which is extensively remodeled during its activation in low phosphate environments, the activity of chromatin remodeling proteins is specified by histone modification rather than by direct DNA–histone interactions. Thus, the evolution of chromatin organization may often occur at the level protein modification and/or interaction and therefore cannot be observed at the level of basic DNA–histone interaction for which our test is designed. Last
and most importantly, precise positioning of consecutive nucleosomes in a region can be achieved by the physical positioning of only a few nucleosomes. The idea that periodic regular nucleosome positioning throughout a promoter region can be achieved by just a few intermittently spaced unmovable or strongly “positioned” nucleosomes was theorized by Kiyama and Trifonov (2002) as the partial positioning case of the “parking lot” model of nucleosome positioning. This is strongly suggested by our test result for the PHO5 promoter, where the 4 nucleosomes known to be removed during remodeling occupy a region of intermediate DNA–nucleosome affinity but might be themselves precisely positioned by 2 adjacent nucleosomes with functionally conserved high affinity for DNA which act as “bookends” for the remodeled region.

Recent work by Ioshikhes et al. (2006) has demonstrated a marked contrast in nucleosome affinity profiles of TATA(+) genes, which are usually induced in response to change in environment, and TATA(−) genes, which are often constitutively expressed. Although both groups exhibit regions of low nucleosome occupancy directly upstream from coding sequence, the latter group has a much larger nucleosome-free region that probably allows continuous access for binding of transcription factors. The primary result of our analysis of the yeast genome demonstrates clear statistical evidence of a major difference in the selective forces maintaining chromatin organization over the core promoter and initial coding regions of TATA(−) and TATA(+) genes. In TATA(−) genes, purifying selection upon low DNA–histone affinity acts to maintain a low nucleosome occupancy in the core promoter (∼150 bp upstream of translation), presumably to enable rapid and efficient binding of transcription factors. This region of elevated purifying selection also coincides directly with the region of low bendability in TATA(−) genes reported by Tirosh et al. (2007) further supporting this functional interpretation. This region was also recently discovered to be bordered by 2 H2A.Z variant nucleosomes, containing a histone modification that was recently discovered to be important in gene activation and/or antagonizing gene silencing in yeast (Raisner and Madhani 2006). Our results imply that purifying selection actively maintains the spacing of H2A.Z nucleosomes in TATA(−) genes. In TATA(+) genes, purifying selection on chromatin organization appears less important, except in the initial coding region where selection maintains high affinity of DNA for nucleosome (∼100 bp downstream of translation). This might suggest that the evolution of function in inducible genes is less dependent upon initial chromatin organization (prior to remodeling) and is governed to a larger degree by histone modification that is not able to be detected using our method. No evidence of positive (directional) selection was
detected in TATA(−) genes. However, in TATA(+) genes, we found significant directional selection corresponding to the position of the second H2A.Z variant nucleosome that incorporates the initial coding region (∼100 bp downstream of translation). This detection of positive selection in TATA(+) genes may be due to the fact that the expression of many TATA(+) genes is induced by environmental changes: such genes will be more likely to be involved in adaptive evolution in changing environments. We believe that the overall patterns of selection on fine-scale chromatin organization, which are highly compatible with the nucleosome affinity profiles and other general characteristics of TATA(+) and TATA(−) genes, clearly demonstrates the biological relevance of our statistical test even though our overall rate of detecting natural selection across the yeast genome is rather low. These results also demonstrate the potential of our method to refine our knowledge of chromatin organization affecting gene regulation.

Our proposed test for detecting selection on nucleosome positioning was designed to be applied to arbitrary homologous sequences in yeast and potentially other eukaryotes. The limited availability of such a general test for selection on regulatory sequences, unlike that for proteincoding regions where nonsynonymous and synonymous substitutions are routinely compared, has been a major problem in evolutionary genetics. Castillo-Davis et al. (2004) proposed to measure the evolutionary distance of upstream regulatory sequences using the fraction of sequences occupied by motifs shared between species. This method can be applied to any pair of homologues without prior molecular biological characterization of regulatory sequences. However, there is no clear functional interpretation of this distance, and tests of negative and positive selection are difficult to build upon it. The method of comparing the rate of nucleotide substitutions in noncoding regions relative to synonymous sites in the coding region (e.g., Wong and Nielsen 2004; Andolfatto 2005) can be applied to arbitrary homologues. However, this approach is critically dependent on the accurate alignment of noncoding sequences, which is considerably harder than that of coding sequences. Our method largely avoids these difficulties. Frequent indels and inaccurate alignments are not a major problem because our test is not based on the nucleotide-to-nucleotide comparison of homologous sequences. The accuracy of alignment may influence our test because the number and size of indels determine the degree of randomization. However, because the purpose of the randomization is to break dinucleotide periodicity while maintaining the observed level of sequence divergence, the success of our test is not likely to depend on accurate alignment.

Conclusion

Function in noncoding DNA is ultimately determined by the interaction of several different cellular components...
(transcription factors, nucleosomes, and chromatin remodeling proteins) with several quantifiable aspects of DNA sequence (binding motifs, bendability, and nucleosome affinity). Therefore, it is unlikely that a single feature of regulatory noncoding DNA can be relied upon to assess the total cumulative affect of mutation upon gene regulation. Therefore, a single test for the functional conservation of regulatory features of genes may eventually prove unrealistic. However, in the future, evolutionary inferences about multiple aspects of regulatory function in noncoding DNA, including the organization of chromatin as well as the binding of various transcription and remodeling factors, may provide the best possible approach for investigating the evolution of noncoding DNA. Here, we offer a method to accomplish the former goal: the detection of selection upon chromatin organization. This represents a new direction for methods designed to infer natural selection acting upon the physical properties of DNA sequences, one that does not depend upon either the genetic code nor the matching of pattern motifs but only the universal interaction of DNA with the proteins that package it.

Supplementary Material

Supplementary figure A is available at Molecular Biology Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We acknowledge Dr Ilya Ioshikhes, Ohio State University, and Dr Jonathan Widom, Northwestern University, for helpful discussions regarding this work. We also thank Dr Arndt von Haeseler and 2 anonymous reviewers for their helpful comments on the manuscript. This research was supported by National Science Foundation grant DEB-0449581 and by Arizona State University.

Appendix A: Detailed Computational Methods
Calculating Nucleosome Affinity (SW Method)

We describe the method to predict the ability of an arbitrary stretch of DNA to wrap nucleosome. The method is an extension of the experimental results of Segal et al. (2006) and utilizes their positional dinucleotide probabilities of nucleosome-bound DNA in yeast in vivo (table kindly provided by Dr Widom, available at http://genie.weizmann.ac.il/pubs/nucleosomes06/segal06_data.html) in the calculations. The probability of observing a sequence, $S$, of length 141 bp given that it is nucleosome bound is computed as:

$$P(S) = P_i(S_i) \prod_{i=2}^{141} P_i(S_i|S_{i-1}),$$

where $P_i(S_i)$ is the probability of observing nucleotide $S_i$ at position $i$ in the 141-bp sequence and $P_i(S_i|S_{i-1})$ is the experimentally determined probability of $S_i$ at position $i$ conditional on $S_{i-1}$ at position $i-1$. This probability is then normalized against a background model ($P_B(\cdot)$) using a log ratio, and the score $\sigma$ is determined as:

$$\sigma = \log \frac{P_i(S)}{P_B(S)} = \log \frac{P_i(S_1) \prod_{i=2}^{141} P_i(S_i|S_{i-1})}{P_B(S_1) \prod_{i=2}^{141} P_B(S_i|S_{i-1})}$$

$$= \log P_i(S_1) + \sum_{i=2}^{141} \log \frac{P_i(S_i|S_{i-1})}{P_B(S_i|S_{i-1})}.$$

Equation 2 was reported as a free energy score in Segal et al. (2006) but can also more intuitively be considered as DNA sequence affinity for the nucleosome. In this paper, this value will often be referred to as a “nucleosome affinity score.” We do not use the same background models reported by these authors but rather use the dinucleotide probability at the midpoint of the experimentally derived sequence, which approximates the average dinucleotide probability across positions (i.e., $P_B(S_i|S_{i-1}) = P_i(S_7|S_{70})$ for all $i$). This causes $\sigma$ to be completely determined by the comparison of position-dependent dinucleotide frequencies to position-independent frequency, and, therefore, $\sigma$ is independent of base or dinucleotide composition of the sequence being scanned.

Calculating Nucleosome Affinity (IP Method)

The NPS correlation is based upon positional bias of 2 dinucleotides AA and TT throughout sequences associated with 204 sequences associated with nucleosome formation in the literature (Ioshikhes et al. 1996). This average pattern of dinucleotide occurrences over 139-bp long sequence is plotted in figure 1a of Ioshikhes et al. (2006) for each dinucleotide and is denoted as $P(\text{AA}_i)$ and $P(\text{TT}_i)$ ($i = 1, \ldots, 138$). The NPS correlation is the correlation of AA and TT frequency in a sequence under consideration, $S(\text{AA}_i)$ and $S(\text{TT}_i)$, to the average pattern of AA and TT bias, or $P(\text{AA})$ and $P(\text{TT})$, observed in nucleosome-bound sequences. Within each scanned point (using a 139-bp sliding window), the occurrence of AA and TT dinucleotide patterns, $P(\text{AA})$, and $P(\text{TT})$, are determined as $S(\text{AA}_i) = 1$ if AA occurs at position $i$ or $S(\text{AA}_i) = 0$ if AA does not. Similarly, $S(\text{TT}_i) = 1$ if TT occurs at position $i$ or $S(\text{TT}_i) = 0$ if TT does not. This positional dinucleotide pattern is normalized to the local AA and TT content and denoted $M_{\text{AA}} = \sum S(\text{AA}_i) / 138$. Similarly, $M_{\text{TT}} = \sum S(\text{TT}_i) / 138$. The normalized $S(\text{AA}_i) = S(\text{AA}_i) - M_{\text{AA}}$ and the normalized $S(\text{TT}_i) = S(\text{TT}_i) - M_{\text{TT}}$. The NPS correlation is:

$$\sigma = \sum [S(\text{AA}_i) \cdot P(\text{AA})_i + S(\text{TT}_i) \cdot P(\text{TT})_i].$$

Smoothing the Nucleosome Affinity Scores Calculated along a Sliding Window

Either of the above calculations assumes that the 139- or 141-bp sequence (=138 and 140 dinucleotides,
the segment. (For convenience, in the remainder of this section, the average or maximum of 35 (or 37 for SW method) values obtained by sliding S-window over a defined area. The local maximum and local average scores are, in fact, highly correlated (e.g., $r = 0.767$, $P < 0.001$ for IP method), so we assume that they contain nearly the same information. In the randomization test below, we calculate the local nucleosome affinity of mainly 174-bp (or 176 bp for SW method) long segment of DNA. The local affinity of the segment is thus given by the average or maximum of 35 ($= 174 - 139$; IP method) or 37 ($= 176 - 141$; SW method) $\sigma$ values calculated within the segment. (For convenience, in the remainder of this Appendix, we describe our algorithm using only $\sigma_{avg}$). For the procedures using the maximum as a smoothing method, $\sigma_{avg}$ can be simply replaced by $\sigma_{max}$.) $\sigma_{avg}$ is then standardized using the mean and standard deviation obtained from the application of this sliding-window calculation to a 100-kb length of random DNA matching the base composition of noncoding sequence in yeast. This standardization results in a distribution of $\sigma_{avg}$ that is normal with mean of zero and deviation of 1. Local affinity scores mentioned in the text refer this standardized $\sigma_{avg}$ unless specified otherwise.

A Sampled Randomization Test for Selection on Nucleosome Affinity

We first align homologous sequences from 2 species (A and B). To scan a chromosome to find the evidence of purifying or positive selection on nucleosome affinity, we analyze a sliding window (denoted R-window) on the alignment that covers $139 + m$ bp (IP method) sequence of species A, which is defined as sequence A. The $m$ equals 35 unless a boundary of the initial R-window intersects an insertion or deletion (gap in the alignment), in which case, the boundary of R-window, thus $m$, is altered so as not to divide the indel (see below). The sequence from species B that is included in this R-window is defined as sequence B. (It should be noted that the length of sequence B is $139 + m'$, where $m'$ can be larger or smaller than $m$ due to insertions/deletions between the 2 sequences.) The local nucleosome affinities for sequence A and B, $\sigma_{avg}^A$ and $\sigma_{avg}^B$, respectively, are calculated as described above. Then, the procedure proposed below tests whether the difference in nucleosome affinity between species, $|\sigma_{avg}^A - \sigma_{avg}^B|$, is too small or too large given the between-species divergence in DNA base sequences within this R-window.

The evolutionary changes (single nucleotide substitutions, insertions, and deletions) between species are inferred from the alignment of 2 sequences. Directional changes from species A to B in an R-window are tabulated, assuming a hypothetical evolution that starts with sequence A and ends in sequence B. Then, a new sequence that is divergent from sequence A to the same degree that sequence B is from sequence A is created by letting sequence A undergo the same evolutionary changes that are observed above but in randomly chosen positions. This procedure of randomly repositioning changes is performed in the order of deletion, substitution, and insertion steps. First, if a deletion of $x$ bp sequence from sequence A is observed, an $x$-mer at a random position on sequence A is chosen and deleted. This step is repeated for the remaining deletions, progressively modifying sequence A. Second, if a single nucleotide change at position $k$, from base $S_1$ on sequence A to $S_2$ on sequence B, is observed, a new position $l$ is chosen randomly among all sites carrying $S_1$ in the current sequence (the product of all changes applied to sequence A so far), and $S_1$ is replaced by $S_2$ at this position (no change in position $k$). This procedure is equivalent to exchanging $S_1$ and $S_2$ between positions corresponding to $k$ and $l$, respectively, on sequence B. This step is repeated until all single nucleotide changes from sequence A were repositioned. Finally, insertions to sequence A are repositioned: random positions are chosen on the current sequence, and the same bases as observed are inserted. The final product of these modifications is a sequence that is of the same length as sequence B and, when aligned with sequence A, produces the same number and kinds of mismatches. The base composition of this sequence is also close to that of sequence B (the small difference originates when random bases are eliminated in the deletion step). This sequence is referred to as randomized sequence throughout. This procedure starting from sequence A is repeated to generate $n$ replicates of randomized sequences, denoted $R_{A1}, R_{A2}, \ldots, R_{An}$. Then, the procedure is repeated again starting from sequence B (randomizing the positions of mutations on sequence B), producing sequences $R_{B1}, R_{B2}, \ldots, R_{Bn}$.

The local nucleosome affinity of each randomized sequence is obtained by sliding the S-window within each randomized sequence and finding the average $\sigma$. Thus, $\sigma_{avg}^A$ to $\sigma_{avg}^{A_n}$ and $\sigma_{avg}^B$ to $\sigma_{avg}^{B_n}$ for $2n$ randomized sequences are obtained. If evolutionary changes between sequence A and B occurred randomly regardless of their consequences on the nucleosome affinity, the repositioning of these changes would not affect the nucleosome affinity, and $|\sigma_{avg}^A - \sigma_{avg}^B|$ would be expected to be similar to $|\sigma_{avg}^{A_j} - \sigma_{avg}^{A_k}|$ or $|\sigma_{avg}^{B_j} - \sigma_{avg}^{B_k}|$ for arbitrary $j, k$ ($= 1, \ldots, n$) (fig. 1). However, if negative selection occurred against mutations that change the nucleosome affinity, the observed changes between sequence A and B may represent those mutations allowed at positions that minimize the effect on the nucleosome affinity. In this case, $|\sigma_{avg}^A - \sigma_{avg}^B|$ is expected to be smaller than $|\sigma_{avg}^{A_j} - \sigma_{avg}^{A_k}|$ or $|\sigma_{avg}^{B_j} - \sigma_{avg}^{B_k}|$ for most $j, k$ ($= 1, \ldots, n$). Therefore, to detect purifying (stabilizing) selection that maintains a small $|\sigma_{avg}^A - \sigma_{avg}^B|$, we calculate the proportion of randomizations that results...
in $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| > |\sigma_{A}^{avg} - \sigma_{B}^{avg}|$ or $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| > |\sigma_{avg}^{B} - \sigma_{avg}^{A}|$ ($j, k = 1, \ldots, n$). This proportion is defined as the $P$ value for purifying selection. Conversely, the proportion of randomizations that results in $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| < |\sigma_{A}^{avg} - \sigma_{B}^{avg}|$ or $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| < |\sigma_{avg}^{B} - \sigma_{avg}^{A}|$ ($j, k = 1, \ldots, n$) is defined as the $P$ value for positive directional selection. (For a given R-window, $P$ values for purifying and positive selection may not add up to one because some randomizations may produce $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| = |\sigma_{A}^{avg} - \sigma_{B}^{avg}|$ or $|\sigma_{avg}^{A} - \sigma_{avg}^{B}| = |\sigma_{avg}^{B} - \sigma_{avg}^{A}|$.) If either $P$ value is smaller than the critical level $\alpha$, then the null hypothesis of neutral evolution (substitution of mutations occurs at random positions) is rejected, and the alternative hypothesis (the corresponding type of selection) is accepted. It should be noted that purifying selection can operate on all values of $\sigma_{A}^{avg}$ (or $\sigma_{B}^{avg}$); throughout this study, the inferred conservation of nucleosome affinity does not necessarily suggest the maintenance of high affinity in the tested region. It could also indicate the maintenance of low affinity as well.

The scan for a given chromosomal region for detecting selection is conducted by moving R-window along the sequence and performing the randomization test for each R-window. An R-window is initially chosen over a 174-bp sequence ($m = 35$) starting with position $i$ in species $A$. However, if either boundary of the window is placed on a site involved in insertion/deletion, the boundary is shifted left until it moves out of the indel segment. The next R-window was chosen to start from position $i + d$ (i.e., R-window slides by $d$ bases). In this study, $d = 35$ unless specified otherwise. Again, the R-window boundaries are adjusted to avoid splitting indels.

Appendix B: Determination of Adequate Replication Size for Sampled Randomization Test

The unit replicated during randomization of the location of mutations within the sliding window is the difference in nucleosome affinity score before and after one randomization treatment ($\Delta \sigma = \sigma_{0} - \sigma_{rand}$). Therefore, the expected distribution of $\Delta \sigma$ within a single stationary window is similar to that of the sampling distribution of variance (student’s $t$-distribution) and is a close approximation to the normal distribution when sampling is large. As the window moves along the homologous DNA sequences, the sequence divergence, or mutation count, fluctuates causing the potential magnitude of the difference between nucleosome affinity score before and after randomization ($\Delta \sigma$) to fluctuate as well. The distribution of $\Delta \sigma$ while the window is sliding is therefore closely approximated by the Laplace or double exponential distribution which represents a mixture of normal distributions with varying parameters. Because the $\Delta \sigma$ within a stationary window only approximates normal, the distribution of $\Delta \sigma$ along the sliding window only approximates the Laplace model and is actually log-Laplace with very short tails. Multimodel inference (Akaike information criterion [AIC]) confirms the log-Laplace and Laplace as better fit than normal and lognormal distribution of $\Delta \sigma$ generated while moving along yeast chromosome 1 ($\Delta$ AIC log-Laplace = 0; $\Delta$ AIC Laplace = 1621; $\Delta$ AIC lognormal = 2714; $\Delta$ AIC normal = 5341). Maximum likelihood estimation of the best fitting log-Laplace distribution resulted in the following parameters, scale ($\delta = 0.266$) and right and left tails ($\alpha = \beta = 62.74$). Note that large tail parameters denote short tails. Convergence to the mean under the best fitting log-Laplace model is only slightly slower when compared with a normal distribution with identical variance (supplementary fig. A, Supplementary Material online). A sampled randomization test with 500 randomizations per step was determined to be sufficiently accurate without being overly expensive in terms of computation time. Log-likelihood functions, AIC, and maximum likelihood estimators for the Laplace and log-Laplace models were obtained from Kotz et al. (2001) and Kozubowski and Podgorski (2002), and all computations were implemented in Mathcad 2001 Professional edition. See Burnham and Anderson (1998) for explanation of methods of multimodel inference.

**Literature Cited**


Arndt von Haeseler, Associate Editor

Accepted May 20, 2008