The Effects of Selection Against Spurious Transcription Factor Binding Sites

Matthew W. Hahn, Jason E. Stajich, and Gregory A. Wray

Department of Biology, Duke University

Most genomes contain nucleotide sequences with no known function; such sequences are assumed to be free of constraints, evolving only according to the vagaries of mutation. Here we show that selection acts to remove spurious transcription factor binding site motifs throughout 52 fully sequenced genomes of Eubacteria and Archaea. Examining the sequences necessary for polymerase binding, we find that spurious binding sites are underrepresented in both coding and noncoding regions. The average proportion of spurious binding sites found relative to the expected is 80% in eubacterial genomes and 89% in archaeal genomes. We also estimate the strength of selection against spurious binding sites in the face of the constant creation of new binding sites via mutation. Under conservative assumptions, we estimate that selection is weak, with the average efficacy of selection against spurious binding sites, $N_e s$, of –0.12 for eubacterial genomes and –0.06 for archaeal genomes, similar to that of codon bias. Our results suggest that both coding and noncoding sequences are constrained by selection to avoid specific regions of sequence space.

Introduction

The idea that there are unconstrained sequences in a genome is ubiquitous in biology, largely because no function can be assigned to long stretches of DNA (Li 1997). Despite their lack of function, however, these sequences may not be freely evolving. Great differences in the frequencies of nucleotide motifs within and between genomes (Burge, Campbell, and Karlin 1992; Karlin and Burge 1995; Karlin, Campbell, and Mrazek 1998; Deschavanne et al. 1999; Hess et al. 2000) (fig. 1) suggest that selection, as well as mutation, may shape the sequence of the entire genome. This selection may be due in large part to avoidance of transcription factor binding site motifs or other important sequence motifs (such as the origin of replication) (Burge, Campbell, and Karlin 1992; Rocha, Danchin, and Viari 2001).

The biological demands on the interacting network of proteins and DNA in an organism dictate that transcription of each gene must be regulated in time, level, and, in multicellular organisms, place (Davidson 2001; Locker 2001). Transcription factors and other parts of the transcriptional machinery regulate this process in a complicated cellular environment by interacting directly in a sequence-specific manner with short stretches of DNA surrounding the target gene. These binding sites, often found in a well-defined promoter region upstream of the start of transcription, are typically 6 to 10 base pairs (bp) long (Fairall and Schwabe 2001); this short length allows for new binding sites to appear frequently in many new places in a genome (Stone and Wray 2001). Creation of new binding sites is likely to be an important way in which novel transcriptional patterns evolve. However, this frequent creation of new binding sites may also introduce noise into the efficient functioning of a cell. The binding of transcription factors to the correct set of nucleotides in inappropriate genomic locations will occur without transcription taking place (Li and Johnston 2001); this nonfunctional binding is a drain on the organism's limited pool of transcription factors and general transcription machinery and may interfere with transcriptional regulation.

Key words: comparative genomics, natural selection, motif bias, promoters.

E-mail: mwh3@duke.edu.

Mol. Biol. Evol. 20(6):901–906. 2003
DOI: 10.1093/molbev/msg096

For three main reasons, Eubacteria and Archaea are ideal systems in which to evaluate the hypothesis that selection acts against spurious binding sites. First, many, small, fully sequenced prokaryotic genomes are available for analysis. Second, these genomes are uncondensed and thus open to direct binding by transcriptional machinery (Langer et al. 1995), which presents the opportunity for selection to act against inappropriate binding of transcriptional machinery independent of chromatin condensation. Third, the sites that control transcription are well defined. In Eubacteria, transcription is controlled in large part by the RNA polymerase holoenzyme, whose contact with DNA is mediated by interactions between the σ^{70} factor and the –35 and –10 sequences (consensus sequences 5′-TTGACA-3′ and 5′-TATAAT-3′, respectively [Baumann, Qureshi, and Jackson 1995]). In Archaea, transcription resembles that in eukaryotes: an A-box motif (5′-TTTA [T/A]A-3′), centered at –27 bp from the start of transcription, is bound by TATA-binding protein independent of RNA polymerase II (Baumann, Qureshi, and Jackson 1995; Langer et al. 1995). These binding site motifs, or slight variants on them, are necessary for transcription in both groups of organisms.

Here we test the hypothesis that selection acts to remove spurious transcription factor binding sites throughout 52 genomes of Eubacteria and Archaea. We use the consensus binding site from Eubacteria and the two main variants from Archaea as our focal sequences and test for underrepresentation of these sequences in both coding and noncoding regions of the genome. A model for the loss and gain of binding sites is also introduced; this model allows us to estimate the average strength of selection against spurious binding sites across genomes.

Materials and Methods

Motif Counts

Motif counts were generated with a sliding window approach applied to all 52 (table 1) completely sequenced prokaryote genomes as of October 30, 2001 (completed eubacterial and archaeal genomes available from NCBI at ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria) using the EMBOSS program compseq (Rice, Longden, and Bleasby 2000). The same qualitative results were obtained from genomes of multiple strains of the same species. The

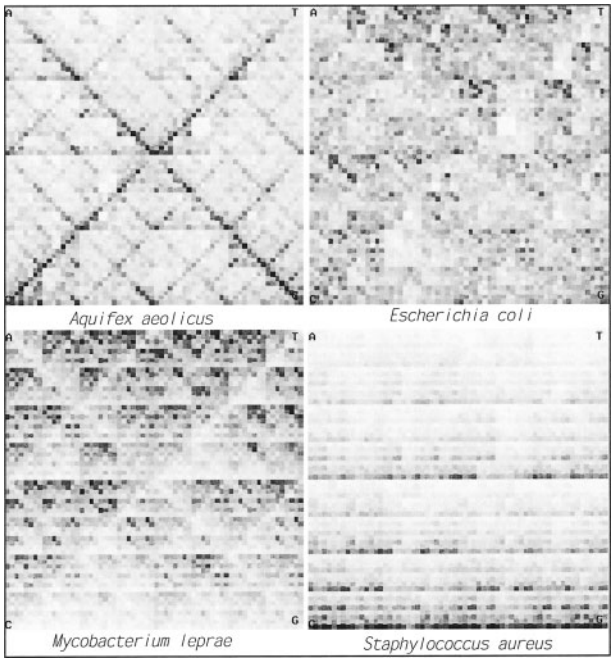

FIG. 1.—Motif bias in four representative genomes. All 4,096 possible six-nucleotide motifs are plotted in their own bin according to their frequency in a Chaos Game Representation (Deschavanne et al.1999) constructed by the program CGRmotif (see *Materials and Methods*). Motif frequency in a genome is represented in each bin on a scale from white (indicating the most rare motifs) to black (indicating the most common motifs).

expected number of motifs in each genome, $E(W)$, was calculated using the Markov maximal order model (Karlin, Burge, and Campbell 1992): $E(W) = N(w_1w_2 \ldots w_{m-1})N(w_2w_3 \ldots w_m)/N(w_2w_3 \ldots w_{m-1})$, where the number of expected occurrences of motif W made up of m nucleotides $w_1w_2 \ldots w_m$ is estimated using the observed number of occurrences, $N(W)$, of subsequences. Assignment of overrepresentation or underrepresentation of motifs in a genome was done by comparing expected to observed. The expected value was calculated separately for all partitions of the genomic sequences. Significance of under/over comparisons across genomes was carried out by a G-test of independence with Williams' correction.

Counting Promoter Sites

The number of true binding sites in the promoter regions was calculated by first parsing gene locations from GenBank annotations of each genome using an already available Perl module from the Bioperl Project (Stajich et al. 2002). A database of promoter sequences upstream from each ORF was then created. The criterion for promoter sequence was the smaller of either the first 1,000 bases upstream or until reaching a gene on either strand. Each sequence in the database was tested for the presence of one copy of the consensus promoter binding sites specific to either Eubacteria or Archaea in the correct orientation. The number of total sequences containing consensus binding sites in the promoter database was subtracted from the total seen for the genome. All expected values were then recalculated without the identified true binding sites.

Table 1
List of Genomes Used in This Study

Genome	Accession Number
Archaea	
Aeropyrum pernix	NC_000854
Archaeoglobus fulgidus	NC_000917
Halobacterium sp. NRC-1	NC_002607
Methanococcus jannaschii	NC_000909
Methanothermobacter thermoautotrophicus	NC_000916
Pyrococcus abyssi	NC_000868
Pyrococcus horikoshii	NC_000961
Sulfolobus solfataricus	NC_002754
Sulfolobus tokodaii	NC_003106
Thermoplasma acidophilum	NC_002578
Thermoplasma volcanium	NC_002689
Eubacteria	
Agrobacterium tumefaciens	NC_003062 and NC_003063
Aquifex aeolicus	NC_000918
Bacillus halodurans	NC_002570
Bacillus subtilis	NC_000964
Borrelia burgdorferi	NC_001318
Buchnera sp. APS	NC_002528
Campylobacter jejuni	NC_002163
Caulobacter crescentus	NC_002696
Chlamydia muridarum	NC_002620
Chlamydia trachomatis	NC_000117
Chlamydophila pneumoniae J138	NC_002491
Clostridium acetobutylicum	NC_003030
Deinococcus radiodurans	NC_001263
Escherichia coli K12	NC_000913
Haemophilus influenzae Rd	NC_000907
Helicobacter pylori J99	NC_000921
Lactococcus lactis	NC_002662
Mesorhizobium loti	NC_002678
Mycobacterium leprae	NC_002677
Mycobacterium tuberculosis CDC1551	NC_002755
Mycoplasma genitalium	NC_000908
Mycoplasma pneumoniae	NC_000912
Mycoplasma pulmonis	NC_002771
Neisseria meningitidis Z2491	NC_003116
Pasteurella multocida	NC_002663
Pseudomonas aeruginosa	NC_002516
Rickettsia conorii	NC_003103
Rickettsia prowazekii	NC_000963
Salmonella enterica ser. Typhi	NC_003198
Salmonella typhimurium LT2	NC_003197
Sinorhizobium meliloti	NC_003047
Staphylococcus aureus N315	NC_002745
Streptococcus pneumoniae TIGR4	NC_003028
Streptococcus pyogenes M1	NC_002737
Synechocystis sp. PCC6803	NC_000911
Thermotoga maritima	NC_000853
Treponema pallidum	NC_000919
Ureaplasma urealyticum	NC_002162
Vibrio cholerae	NC_002505 and NC_002506
Xylella fastidiosa	NC_002488
Yersinia pestis	NC_003143

Chaos Game Representation of Genomes

Counts obtained from above were graphed according to a chaos game representation (CGR) algorithm (Deschavanne et al. 1999). The program made by the authors to

generate these figures, CGRmotif, is available at http://www.duke.edu/~jes12/cgr. Whereas most methods used to find binding sites attempt to identify sites that are overrepresented in upstream regions (Wagner 1998; Berman et al. 2002), our program can be used to identify previously unknown transcription factor binding sites by looking for deficiencies of sites in a genome (Stajich, Hahn, and Wray, unpublished data).

Results and Discussion

Underrepresentation of Binding Sites

We counted the number of our focal binding sites present in each of 52 whole, unicellular genomes (41 eubacterial and 11 archaeal) (table 1) using a sliding window and examining both strands. In order to estimate the number of binding sites expected in each genome, we used a Markov model approach (Karlin, Burge, and Campbell 1992). This method takes into account motif bias of all subsequences found within our target hexamer (e.g., overrepresentation of the dinucleotide motif TA or underrepresentation of the stop codon TGA, in any frame, will not bias our results). Because we account for genome-specific nucleotide and motif frequencies, and hence for differences in mutation spectra among genomes, selection is the only plausible explanation for any observed patterns. For 32 out of 41 whole eubacterial genomes, TTGACA sequences were underrepresented ($P < 0.01$; table 2). For 29 out of 41 eubacterial genomes, TATAAT sequences were underrepresented ($P < 0.05$; table 2). In comparison, the two archaeal A-box motifs that are not used as binding sites for transcription in Eubacteria are not significantly underrepresented (21 of 41 for TTTAAA and 16 of 41 for TTTATA). In Archaea, seven of 11 genomes are underrepresented for both TTTAAA and TTTATA (table 2), sequences which act as binding sites in this group.

These findings are inherently conservative because our counts of binding sites do not distinguish between spurious binding sites and those that are actively used in transcriptional regulation. Therefore, we are likely to overestimate the number of spurious binding sites by including binding sites maintained by selection. Nonetheless, we still detect significant underrepresentation of these sequences. If we examine just coding regions, we should eliminate the overcounting due to true binding sites because transcriptionally active binding sites should be absent in these regions. Table 2 shows that examining only open reading frames (ORFs) increases the number of underrepresented genomes to 32 of 41 for TTGACA and 37 of 41 for TATAAT in Eubacteria and 9 of 11 for TTTAAA and 11 out of 11 for TTTATA in Archaea. In contrast, noncoding regions show little if any pattern towards underrepresentation when examined separately and, in fact, show a slight tendency towards overrepresentation; this is because the binding sites that are maintained by selection are found in these regions (table 2).

As an alternative way to correct for counting transcriptionally relevant binding sites in noncoding regions, we scanned promoter regions 5′ of coding DNA, excluding a single binding site motif, if any were found, that was in the correct orientation (see *Materials and Methods* for details). This method is needed for two reasons. First, not every gene has its own promoter sequence: many genes are cotranscribed in polycistronic operons in both groups of organisms. Second, not every gene uses the strongly binding consensus sites. Genes that are not transcribed at high levels may use nonconsensus, weaker binding sites as a regulating mechanism (Kobayashi, Nagata, and Ishihama 1990; Xu, McCabe, and Koudelka 2001). Taking into account the active binding sites, we find that spurious binding sites are underrepresented in noncoding regions in almost every archaeal and eubacterial genome (table 2, ''Noncoding-x''). This indicates that there is consistently stronger selection against spurious binding sites in noncoding regions. A concern is that when examining any motif, regardless of identity or function, our attempt to correct for selectively maintained motifs may lead to some amount of underrepresentation. However, we are still confident in our results because both across whole genomes and in coding regions, where controlling for true binding sites does not depend on correctly identifying such sites, we still see a significant underrepresentation of spurious transcription factor binding sites.

Table 2
Underrepresentation of Binding Sites Across Genomes

	Total Genome	Genome-x[a]	ORFs[b]	Noncoding	Noncoding-x
Eubacteria					
TATAAT	29:12	40:1	37:4	14:27	41:0
TTGACA	32:9	33:8	32:9	20:21	33:8
Archaea					
TTTAAA	7:4	10:1	9:2	4:7	11:0
TTTATA	7:4	11:0	11:0	2:9	10:1

NOTE.—For every data set, the number of genomes in which the binding site is underrepresented is given compared with the number of genomes in which it is overrepresented (under:over). Underrepresentation and overrepresentation are determined by comparing expected with observed for each genome (see *Materials and Methods* for details).

[a] ''x'' represents the number of estimated binding sites that are actively involved in transcription.

[b] Open reading frames.

Our results suggest two different selective mechanisms acting against spurious binding sites. Selection most likely acts throughout the genome to reduce random binding, thus enhancing transcriptional efficiency. In promoter regions, selection may also act to remove spurious sites to avoid steric hindrance of transcription factors or to avoid gene silencing via ectopic transcription and RNA interference. The avoidance of steric hindrance in promoter regions may explain the conservation of sequences with no known binding affinities in-between binding sites in promoters: any motif that possibly binds a transcription factor is deleterious, thus further constraining sequence space.

Strength of Selection

Because little is known about the strength of the selective forces that constrain the genome as a whole, we also estimated the strength of selection against spurious binding sites. Selection acting on any single gene for translational efficiency is revealed by the nonrandom use

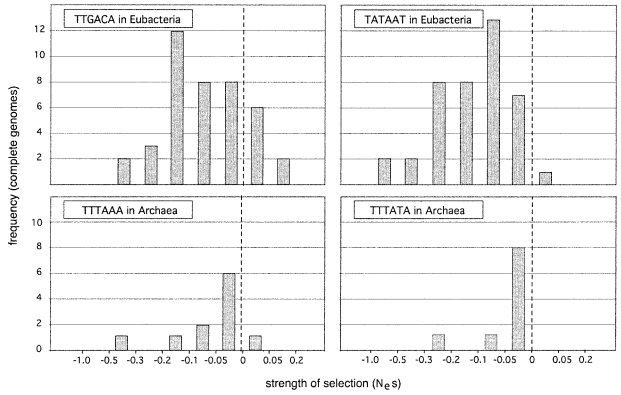

FIG. 2.—Distribution of N_es values in prokaryotic genomes. The average value of N_es for spurious binding sites is plotted for 41 eubacterial and 11 archaeal genomes. TTGACA and TATAAT are binding sites in Eubacteria; TTTATA and TTTAAA are binding sites in Archaea.

of synonymous codons (codon bias) to match tRNA species abundance (Ikemura 1985; Moriyama and Powell 1997). Selection for codon usage has been estimated to be approximately $-2 < N_es < 1$ in both bacteria (Hartl, Moriyama, and Sawyer 1994) and Drosophila (Akashi 1995) (''N_e'' represents the effective population size of the species, and ''s'' represents the selection coefficient for each variant). In order to measure the average strength of selection against spurious binding sites, we modeled the number of spurious binding sites in a genome as a balance between mutation and selection. In the neutral case without selection against binding sites, the equilibrium number of spurious binding sites expected in a genome is described by

$$dn_0/dt = (N - n_0)\alpha - n_0\beta = 0 \qquad (1)$$

where N equals the number of possible binding sites in a genome and n_0 equals the number of occurrences of a given binding site motif in a genome. The coefficients α and β encompass the mutation rates and fixation probabilities of neutral mutants for creation of binding sites (α) and the loss of binding sites (β). The first term of the equation describes the origin of new binding sites, and the second term describes the loss. If we add selection, both to maintain proper binding sites in the genome and against spurious binding sites, this equation becomes

$$dn/dt = [N - (n + x)]\alpha\gamma - n\beta = 0 \qquad (2)$$

where x equals the number of true binding sites, n equals the number of spurious binding sites, and γ represents the effect of selection on the fixation probability of any binding site relative to the neutral case; this was approximated by Kimura (1983) as

$$\gamma = 4N_es/1 - e^{-4Nes} \qquad (3)$$

That is, γ equals 1 for a neutral mutant ($N_es = 0$), is greater than 1 for a mutant under positive selection ($N_es > 0$), and is less than 1 for a mutant under negative selection ($N_es < 1$), as is expected for spurious binding sites.

Solving for n_0 and n in equations 1 and 2 and dividing gives the result $n/n_0 = \gamma$. Here we assume that β is much larger than α (that mutation away from any particular sequence is much more likely than mutation to that sequence) and that x, the number of true binding sites, is negligible compared with the number of possible binding sites in a genome, N (where N = size of the genome in base pairs − size of binding motif in base pairs + 1).

Intuitively, this result makes sense: the deficiency of binding sites in the genome (the ratio of observed number of binding sites to expected) is due to the lower probability of fixation of mutants that are selected against (the selection parameter, γ). Equation 3 allows us to numerically solve for the average efficacy of selection (as measured by N_es) on spurious binding sites. These values are plotted for the Eubacteria and Archaea in figure 2.

The average value of N_es for mutations to TTGACA is –0.09 ($-0.32 < N_es < 0.11$) and for mutations to TATAAT is –0.15 ($-0.56 < N_es < 0.04$) (table 3). These values are similar to those found for codon bias (Hartl, Moriyama, and Sawyer 1994; Akashi 1995) and fall within the range of ''nearly neutral'' mutations ($|N_es| < 1$) (Ohta 1992). Mutations in this range are strongly affected by population bottlenecks, reduced recombination, and other deficits in effective population size. It should be noted that the values of N_es are averages of many mutations with an unknown distribution of fitness effects. In addition, these estimates of selection against any one mutation are averages across different regions of the genome; table 3 shows that, as expected from above, the estimates of selection differ in coding and noncoding regions, with greater selection against spurious binding sites in noncoding regions.

Table 3
Average Values for N_es Across Genomes

	Genome-x[a]	ORFs[b]	Noncoding-x
Eubacteria			
TATAAT	−0.15(0.13)	−0.12(0.13)	−0.36(0.47)
TTGACA	−0.09(0.09)	−0.07(0.09)	−0.18(0.23)
Archaea			
TTTAAA	−0.06(0.10)	−0.05(0.08)	−0.11(0.11)
TTTATA	−0.05(0.06)	−0.05(0.04)	−0.06(0.08)

NOTE.—Standard deviations are given in parentheses.

[a] ''x'' represents the number of estimated binding sites that are actively involved in transcription.

[b] Open reading frames.

The Effects of Selection Against Binding Sites

We have demonstrated here a form of natural selection that is only evident at the level of the whole genome. Selection against spurious transcription factor binding sites cannot be detected at a single locus, but requires a whole genome, or multiple whole genomes, to be sequenced. The effects of this selection may go a long way toward explaining both motif bias within genomes (where binding sites are underrepresented relative to nonbinding sites) and between genomes (where different suites of transcription factors are used in different organisms) (fig. 1). Across the genomes examined in this study, the effects of selection are obvious. Correcting for the binding sites used in

transcriptional regulation, 33 of 41 and 40 of 41 eubacterial genomes are underrepresented for TTGACA and TATAAT, respectively (table 2, ''Genome-x''). In Archaea, 10 of 11 and 11 of 11 genomes are underrepresented for TTTAAA and TTTATA, respectively (table 2, ''Genome-x''). It is important to note that the consensus binding sites used here are often only identified biochemically in one or a very few species of Eubacteria or Archaea. For this reason, the underrepresented/overrepresented approach is conservative within groups. The fact that binding sites are not underrepresented in those genomes in which they are not used is not evidence against our hypothesis; in fact, it may be used to form new hypotheses about the sequences of binding sites used in organisms for which experimental evidence is lacking.

The effects of selection within each genome are also quite large. The average numbers of binding sites in a genome, relative to the expected, are 91% for TTGACA and 89% for TATAAT in Eubacteria and 99% for TTTAAA and 98% for TTTATA in Archaea (for uncorrected genomes); in coding regions alone the numbers are 87% for TTGACA, 81% for TATAAT, 91% for TTTAAA, and 90% for TTTATA. If we correct for true binding sites across the genome, the average numbers of spurious binding sites, relative to the expected, are 85% for TTGACA, 75% for TATAAT, 89% for TTTAAA, and 90% for TTTATA. Once again, because we assume that the consensus binding sites are the same across each taxonomic group, these numbers are conservative estimates of the effects of selection.

The method used to estimate the expected number of motifs in a genome (Karlin, Burge, and Campbell 1992) corrects for motif bias in all subsequences of our focal motifs. This means that simple codon bias, in or out of frame, does not affect our estimates. Unfortunately, this method does not take into account any effects of di-codon bias: the nonrandom distribution of neighboring codon pairs (Gutman and Hatfield 1989). However, we have good reason to think that this effect is minimal or nonexistent. Codon bias across these diverse sets of organisms, which have many hundreds of millions of years separating them even within Eubacteria or Archaea, is extremely varied (Nakamura, Gojobori, and Ikemura 2000). The different genomes differ in the synonymous codons that are used, in the GC content of coding regions, and in the amino acids that are used (Singer and Hickey 2000). In addition, it has been shown that di-codon bias differs among the species of Eubacteria and Archaea (Badger and Olsen 1999; McVean and Hurst 2000) and so should not explain the patterns we see across these groups. Finally, this bias only has effects on one DNA strand, whereas our results use both strands. It should be noted, however, that none of the reasons stated above argues against the contention that di-codon bias may be caused *by* selection against spurious binding sites in any single genome.

The next step in the study of selection against binding site motifs will be to examine eukaryotic genomes, where this form of selection may introduce a low level of background selection (Charlesworth, Morgan, and Charlesworth 1993) throughout the genome. It will be interesting to learn whether the motif bias observed here is as evenly distributed in eukaryotes, where heterochromatin, gene-rich regions, and different rates of recombination introduce a greater degree of spatial heterogeneity across a genome. In regions of heterochromatin, where DNA may not be open to spurious binding by transcription factors, selection against spurious binding site motifs would be unnecessary. In euchromatin, areas that are gene rich, and hence transcriptionally active much of the time, may show the strongest effects of this selection. On top of both of these conditions, rates of recombination along a chromosome show an effect on the efficacy of selection (Kliman and Hey 1993; Comeron and Kreitman 2002) and may add to the spatial heterogeneity in underrepresentation. Finally, even though the transcriptional machinery of the Archaea shares many similarities with that of eukaryotes (Baumann, Qureshi, and Jackson 1995; Langer et al. 1995), the complexity of multicellular regulatory regions is unnecessary in prokaryotic genomes. In eukaryotes, there are often multiply-represented transcription factor binding sites in any one promoter region, as opposed to the single site necessary to initiate transcription. In these cases calculating underrepresentation only in coding regions may be preferred to avoid the inclusion of multiple binding sites maintained by selection.

The pattern of binding site motif underrepresentation presented here clearly supports the action of selection in constraining sequences throughout the genome, regardless of function. In fact, selection is almost certainly constraining sequences *without* biologically relevant function, as well as coding and regulatory sequences, to a specific region of sequence space. Although we have demonstrated selection only on the consensus sequences in the focal binding sites, this method may be used to estimate the strength of selection against variants of the consensus and other transcription factor binding sites. The use of population genetics models and theory along with the tools of comparative genomics has allowed new insight into the effects of natural selection at the level of the whole genome (e.g., Comeron and Kreitman 2002; Lynch 2002). Here we have extended this approach by connecting the effects of purifying selection on genomic sequences with stabilizing selection at the level of transcriptional output.

Acknowledgments

Many thanks to M. Rausher for comments and criticism, especially on the model, and to M. Rockman for helping to shape the questions. We also thank J. Balhoff, F. Dietrich, D. Hartl, S. Miller, L. Moyle, F. Nijhout, and M. Uyenoyama for comments on the manuscript. D. Rand and two anonymous reviewers helped to improve the paper.

Literature Cited

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in Drosophila DNA. Genetics **139**:1067–1076.

Badger, J., and G. Olsen. 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol. Biol. Evol. **16**:512–524.

Baumann, P., S. A. Qureshi, and S. P. Jackson. 1995. Transcrip-

tion: new insights from studies on Archaea. Trends Genet. **11**:279–283.

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA **99**:757–762.

Burge, C., A. M. Campbell, and S. Karlin. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**:1358–1362.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

Comeron, J. M., and M. Kreitman. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics **161**:389–410.

Davidson, E. H. 2001. Genomic regulatory systems: development and evolution. Academic Press, San Diego.

Deschavanne, P. J., A. Giron, J. Vilain, G. Fagot, and B. Fertil. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. **16**:1391–1399.

Fairall, L., and J. W. R. Schwabe. 2001. DNA binding by transcription factors. Pp. 65–84 *in* J. Locker, eds. Transcription factors. Academic Press, San Diego.

Gutman, G. A., and G. W. Hatfield. 1989. Nonrandom utilization of codon pairs in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **86**:3699–3703.

Hartl, D. L., E. N. Moriyama, and S. A. Sawyer. 1994. Selection intensity for codon bias. Genetics **138**:227–234.

Hess, C. M., J. Gasper, H. E. Hoekstra, C. E. Hill, and S. V. Edwards. 2000. MHC class II pseudogene and genomic signature of a 32-kb cosmid in the house finch (*Carpodacus mexicanus*). Genome Res. **10**:613–623.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Karlin, S., and C. Burge. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11**:283–290.

Karlin, S., C. Burge, and A. M. Campbell. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. **20**:1363–1370.

Karlin, S., A. M. Campbell, and J. Mrazek. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**: 185–225.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kliman, R. and J. Hey. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10**:1239–1258.

Kobayashi, M., K. Nagata, and A. Ishihama. 1990. Promoter selectivity of *Escherichia coli* RNA polymerase: effect of base substitutions in the promoter –35 region on promoter strength. Nucleic Acids Res. **18**:7367–7372.

Langer, D., J. Hain, P. Thuriaux, and W. Zillig. 1995. Transcription in Archaea: similarity to that in Eucarya. Proc. Natl. Acad. Sci. USA **92**:5768–5772.

Li, Q. M., and S. A. Johnston. 2001. Are all DNA binding and transcription regulation by an activator physiologically relevant? Mol. Cell. Biol. **21**:2467–2474.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Locker, J. 2001. Transcription factors. Academic Press, San Diego, Calif.

Lynch, M. 2002. Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. USA. **99**:6118–6123.

McVean, G. A. T., and G. D. D. Hurst. 2000. Evolutionary lability of context-dependent codon bias in bacteria. J. Mol. Evol. **50**:264–275.

Moriyama, E. N., and J. R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45**: 514–523.

Nakamura, Y., T. Gojobori, and T. Ikemura. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. **28**:292.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. Ann. Rev. Ecol. Syst. **23**:263–286.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet. **16**:276–277.

Rocha, E. P. C., A. Danchin, and A. Viari. 2001. Evolutionary role of restriction modification systems as revealed by comparative genome analysis. Genome Res. **11**:946–958.

Singer, G. A. C., and D. A. Hickey. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. **17**:1581–1588.

Stajich, J. E., D. Block, K. Boulez et al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. **12**:1611–1618.

Stone, J. R., and G. A. Wray. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. **18**:1764–1770.

Wagner, A. 1998. Distribution of transcription factor binding sites in the yeast genome suggests abundance of coordinately regulated genes. Genomics **50**:293–295.

Xu, J., B. C. McCabe, and G. B. Koudelka. 2001. Function-based selection and characterization of base-pair polymorphisms in a promoter of *Escherichia coli* RNA polymerase sigma(70). J. Bacteriol. **183**:2866–2873.

David Rand, Associate Editor

Accepted January 24, 2003