# Inferring Selection on Amino Acid Preference in Protein Domains

*Alan M. Moses*† and Richard Durbin**

*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; and †Department of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada

Models that explicitly account for the effect of selection on new mutations have been proposed to account for "codon bias" or the excess of "preferred" codons that results from selection for translational efficiency and/or accuracy. In principle, such models can be applied to any mutation that results in a preferred allele, but in most cases, the fitness effect of a specific mutation cannot be predicted. Here we show that it is possible to assign preferred and unpreferred states to amino acid changing mutations that occur in protein domains. We propose that mutations that lead to more common amino acids (at a given position in a domain) can be considered "preferred alleles" just as are synonymous mutations leading to codons for more abundant tRNAs. We use genome-scale polymorphism data to show that alleles for preferred amino acids in protein domains occur at higher frequencies in the population, as has been shown for preferred codons. We show that this effect is quantitative, such that there is a correlation between the shift in frequency of preferred alleles and the predicted fitness effect. As expected, we also observe a reduction in the numbers of polymorphisms and substitutions at more important positions in domains, consistent with stronger selection at those positions. We examine the derived allele frequency distribution and polymorphism to divergence ratios of preferred and unpreferred differences and find evidence for both negative and positive selections acting to maintain protein domains in the human population. Finally, we analyze a model for selection on amino acid preferences in protein domains and find that it is consistent with the quantitative effects that we observe.

## Introduction

Methods and models for analyzing natural selection on codon bias are among the most advanced (Li 1987; Bulmer 1991; Hartl et al. 1994; Akashi 1995; Akashi and Schaeffer 1997; Mcvean and Charlesworth 1999; Yang and Nielsen 2008). A major reason for this is the observation that it is possible to define a preferred set of codons for a given species (Bulmer 1991), which allows differences in codons to be classified as either preferred or unpreferred with respect to function. Taking advantage of this asymmetry has led to the development of powerful, specific tests for the action of natural selection on codon usage (Akashi 1995; Mcvean and Charlesworth 1999; Cutter and Charlesworth 2006).

Here we argue that, for a large class of amino acid replacement substitutions, namely, those occurring in conserved protein domains, it is also possible to assign a preferred and unpreferred state. Conserved protein domains are modular structural units of proteins, often associated with a very specific molecular function (Sonnhammer et al. 1998). For example, in the human genome, 67% of proteins contain domains and 38% of amino acids fall into protein domains, as identified in the protein domain database, Pfam (Finn et al. 2008). Because residues in these domains evolve under specific structural and functional constraints, the residue preferences at each position can be characterized by comparing the sequences of large numbers of domains from divergent proteins and species. We propose that mutations that lead to more common amino acids (at a given position in a domain) can be considered "preferred alleles" just as are synonymous mutations leading to codons for more abundant tRNAs. In other words, we argue that the fitness effect of a new mutation can be predicted based on the probabilities of the ancestral and derived amino acids in the protein domain model.

Key words: weakly selected variants, protein domains, polymorphism, McDonald–Kreitman test, deleterious, advantageous.

E-mail: alan.moses@utoronto.ca.

*Mol. Biol. Evol.* 26(3):527–536. 2009
doi:10.1093/molbev/msn286
Advance Access publication December 18, 2008

Although not as numerous as silent differences, large numbers of substitutions and polymorphisms in sequence coding for protein domains are found in genome-scale comparisons. Using the proposed definition of preferred and unpreferred alleles in protein domains, we show an elevation of the frequencies of preferred alleles and an excess of low-frequency–derived unpreferred alleles. In addition to affecting the frequencies of alleles, selection also affects the number of polymorphisms segregating and rate of fixation. By comparing divergence and polymorphism in protein domains, we find evidence for increased and decreased rates of fixation for preferred and unpreferred mutations, respectively. Finally, we show that our results are consistent with a quantitative model for the effects of selection.

## Materials and Methods

### Human Protein Domains

We used the nonredundant set of human proteins from TreeFam v4 (Ruan et al. 2008). To identify protein domains in this set of proteins, we used hidden Markov model (HMM)-Pfam and considered hits with $e$-value $< 0.001$ to be bona fide protein domains. We used the Pfam-ls HMM database obtained from Pfam v22.0 (Finn et al. 2008). In these HMMs, at a given position, the emission probability for each residue is a combination of the fraction of times that residue was observed in the Pfam alignment and a nine-component Dirichelet prior (Sjölander et al. 1996). In order to ensure that the HMMs included were true "domains," we considered only those for which there were at least 10 instances in the human genome. When multiple Pfam domains included a particular residue in the genome, we assigned it to the Pfam domain with the smallest $e$-value.

### SNPs in Protein Domains

To identify single nucleotide polymorphisms (SNPs) in this human protein set, we searched Ensembl v43 (Hubbard et al. 2002) using the PERL API for SNPs that were genotyped in the HapMap YRI population (The

International HapMap Consortium 2005) and fell in those regions of the protein. We chose the YRI population because it has been found to be more polymorphic and we therefore expected to have the most power to detect the effects of selection in this population. Allele frequencies were used for SNPs that were nonmonomorphic in that population and were computed by adding the frequency of the homozygote plus ½ the frequency of the heterozygote. To infer ancestral states ("orient" the SNPs), we used coding sequence alignments of the TreeFam orthologues based on t_coffee protein alignments. We considered an SNP to be ancestral if the entire codon in which it appeared was identical to the chimp codon found in the alignment. Because the Pfam alignments contain human sequences, and for any given SNP, the human residue is more likely to be ancestral than derived, the allele frequencies might not be truly independent of the residue probabilities. To rule out the possibility that this was biasing our results, we computed the preferred and derived allele frequency spectra described above excluding and SNPs that fell in human genes that were included in Pfam alignments and found similar results (supplementary fig. 3, Supplementary Material online).

To compare emission probabilities for the two alleles in protein domains, we mapped the position of the SNP to the HMMer alignment of the protein domain model to that protein and extracted the residue frequencies from the HMM matrix.

To confirm that our results were not affected by ascertainment biases in selection of SNPs for the HapMap project (e.g., see Nielsen et al. 2004), we performed the analysis of preferred allele frequencies on 337 human SNPs identified in protein domains by systematic exon resequencing (exoseq, http://www.sanger.ac.uk/humgen/exoseq/) and found similar results (data not shown).

## Fixed Differences in Protein Domains

Using the "clean trees" from the TreeFam database (Ruan et al. 2008), we defined 1 to 1 orthologues to be groups of genes in the tree for which there was exactly one gene from each of human, chimp, and macaque represented below an ancestral node. To identify fixed differences along the human lineage, we used the human–chimp–macaque alignments from TreeFam. We required that the chimp and macaque codons be identical. To rule out the possibility that our analysis of fixed differences along the human lineage was biased by our annotation of protein domains in the human genome, we identified protein domains as above in the inferred ancestral primate protein sequences and considered in our McDonald and Kreitman (MK) analysis only SNPs that fell in domains that were identified in both the human and inferred ancestral protein sequences. We inferred the ancestral amino acid at each position in the alignment using maximum parsimony, and where multiple amino acids were equally parsimonious we inserted an "X" into the ancestral sequence. Of the 1,630 and 2,725 amino acid changing and synonymous SNPs for which we had derived allele frequency information, this filtering left 1,617 and 2,704, respectively.

## Amino Acid Probabilities under a DNA Substitution Model

Under the assumption that the rate of DNA substitution is given by the Jukes–Cantor model, at equilibrium, each codon will have equal probability 1/64. The equilibrium probabilities of each amino acid at equilibrium, $g$, are given by the sum of the probabilities of its codons. However, because we do not consider stop codons, these probabilities must be normalized so their sum equals 1. Therefore, we have

$$g_a = \frac{\sum_{i=1}^{i=C_a} \frac{1}{64}}{\sum_b \sum_{j=1}^{j=C_b} \frac{1}{64}} = \frac{C_a}{61},$$

where $C_a$ is the number of codons for amino acid $a$.

## Predictions of MK Ratios

Under the Poisson Random Field theory (Sawyer and Hartl 1992), the equilibrium flux of fixations is given by $q\frac{2Ns}{1-e^{-2Ns}}$ or $q$, in the presence or absence of selection, respectively, where $q$ is the mutation rate, $N$ is the effective population size, and $s$ is the selection coefficient. We therefore assumed that expected amino acid replacement to synonymous rate ratio for fixed differences would be $\frac{2Ns}{1-e^{-2Ns}}$.

The total number of segregating polymorphisms is given by $H=\int_{1/2n}^{1-1/2n}\varphi(y)dy$ (Kimura 1969), where $y$ is the allele frequency, $n$ is the number of individuals, such that $^1/_{2n}$ is the frequency of a new mutation, and $\varphi(y)$ is the allele frequency density function. Under the Poisson Random Field approximation (Sawyer and Hartl 1992), $\varphi(y)=2q\frac{1-e^{-2Ns(1-y)}}{(1-e^{2Ns})y(1-y)}$ or $\varphi(y)=\frac{2q}{y}$, in the presence or absence of selection, respectively. In the presence of selection, $H=\frac{2q}{1-e^{-2Ns}}\int_{1/2n}^{1-1/2n}\frac{1-e^{-2Ns(1-y)}}{y(1-y)}dy$, whereas in the absence of selection, integrating gives $H=2q\log(2n-1)$.

Dividing the expected amino acid replacement to synonymous ratio for fixed differences by that for polymorphisms gives

$$\text{MK ratio}=\frac{\frac{2Ns}{1-e^{-2Ns}}}{\frac{\frac{2q}{1-e^{-2Ns}}\int_{1/2n}^{1-1/2n}\frac{1-e^{-2Ns(1-y)}}{y(1-y)}dy}{2q\log(2n-1)}}=\frac{2Ns\log(2n-1)}{\int_{\frac{1}{2n}}^{1-\frac{1}{2n}}\frac{1-e^{-2Ns(1-y)}}{y(1-y)}dy}.$$

To obtain the predictions in figure 5, we set $Ns$ to be the average of the upper and lower bounds of each bin and evaluated the integral numerically using the OCTAVE package setting $n = 10^6$.

Once again, we were concerned about the effects of the HapMap SNP ascertainment on our results. The calculation above uses the total polymorphism in the population; in practice, we do not have an unbiased sample of this. We therefore considered the extreme case where we compute the amino acid replacement to synonymous ratio for polymorphisms in the heterozygous sites in a single individual (the number of polymorphisms segregating in two chromosomes). This is equivalent to assuming that the SNPs were identified in only one individual.

The number of heterozygous sites per individual is approximately $H(2)=\int_0^1 2y(1-y)\varphi(y)\,dy$ (Kimura 1969), which can be integrated to give $H(2)=2q\frac{2Ns-1+e^{-2Ns}}{Ns(1-e^{-2Ns})}$ and

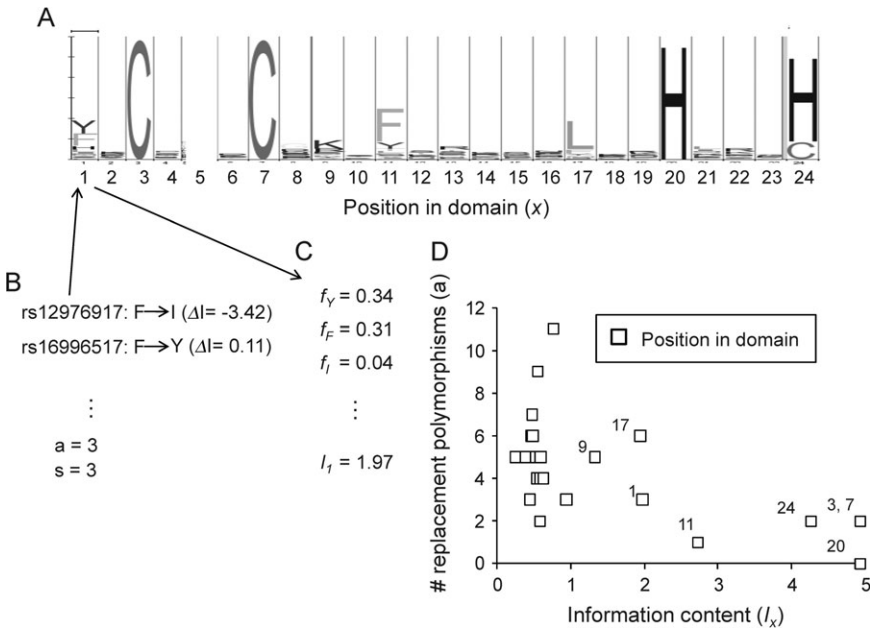


FIG. 1.—Residue probabilities and polymorphisms in C2H2 Zn fingers. (*A*) A "sequence logo" (from the Pfam database) representing the residue probabilities of the domain at each position. The heights of letters in this representation are proportional to the information content at that position (Schneider and Stephens 1990). (*B*) SNPs mapped to each position in the domain can be classified as preferred or unpreferred. At each position, total numbers of amino acid replacement (a) and synonymous (s) polymorphism can be computed by considering SNPs that occur at that position in the domain at each occurrence of that domain in the genome. (*C*) Information content (*I*) summarizes the variability in residue frequencies at each position, *x*. (*D*) Positions in the domain with greater information content show fewer amino acid polymorphisms.

$H(2)=2q$ in the presence or absence of selection, respectively. Therefore, the amino acid to synonymous ratio for polymorphisms in the heterozygous sites per individual is $\frac{2Ns-1+e^{-2Ns}}{Ns(1-e^{-2Ns})}$ and the MK ratio takes a simpler form:

$$\text{MK ratio} = \frac{\frac{2Ns}{1-e^{-2Ns}}}{\frac{(2Ns-1+e^{-2Ns})}{Ns(1-e^{-2Ns})}} = \frac{2(Ns)^2}{2Ns-1+e^{-2Ns}}.$$

We compared the predictions of this formula with that obtained by the one above (supplementary fig. 4, Supplementary Material online). Interestingly, this formula gives qualitatively similar results without the need for numerical integration or assumptions about the actual population size. That these formulas give similar results, despite the different assumptions about sampling of polymorphism supports the idea that the MK test is highly robust to the underlying assumptions.

### Statistics

Significance of correlations (using Spearman's rank correlation) and Fisher's exact tests was obtained using the R statistics package (Ihaka and Gentleman 1996).

## Results

### Identifying Preferred and Unpreferred Alleles in Protein Domains

We sought to define preferred and unpreferred amino acid changing differences in protein domains. To do so, we consider probabilistic models of protein domains (Durbin et al. 1998), HMMs, which assign each amino acid a probability, $f$ (the so-called emission probability), at each position, $x$, in the domain (fig. 1*A*). We annotated protein domains to a nonredundant set of human proteins using HMMs from Pfam (see Methods) and searched for HapMap SNPs (The International HapMap Consortium 2005) that occurred in these domains (fig. 1*B*). In all, we identified 7,108 SNPs in 5,546 domains; 95.6% of domains contained three SNPs or fewer. For each change, we can say at what position in the domain it falls and what type of substitution has occurred: 3,136 of the SNPs change the amino acid.

Positions in protein domains that are critical for the function of the domain often require specific amino acid residues. For example, in the case of C2H2 Zn fingers (fig. 1*A*), the cysteine and histidine residues coordinate a Zinc ion and are critical for the DNA-binding capability of this domain. These positions show very low variability: they have very high probabilities for specific amino acids and very low probabilities for all others. On the other hand, some positions show intermediate variability, where any of several amino acids are permitted (e.g., the first position of the C2H2, fig. 1*C*), whereas others show no obvious preference and seem to contain the background distribution of amino acids.

The variability of amino acid residues at a position in a domain can be summarized by its information content (Schneider et al. 1986), given by

$$I_x = \sum_a f_{xa} \log_2 \frac{f_{xa}}{g_a},$$

where $f_{xa}$ is the emission probability in the domain model of the amino acid $a$ at position $x$ and $g_a$ is the probability of

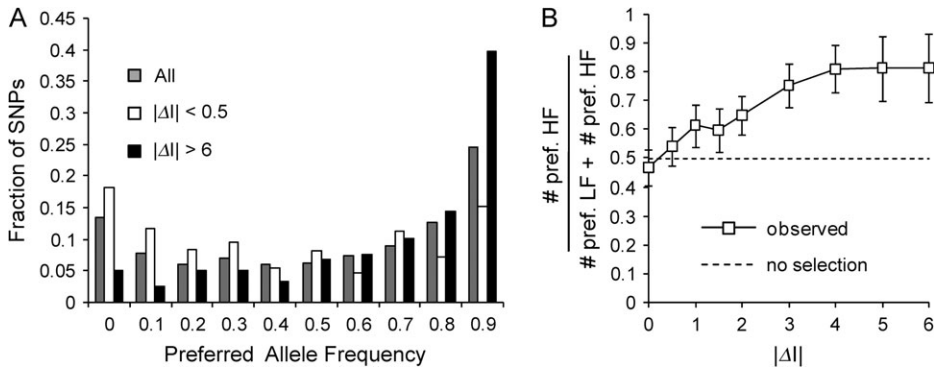


FIG. 2.—Preferred allele frequency distribution for polymorphisms in protein domains. (*A*) Preferred allele frequencies (gray bars) deviate from the symmetric distribution expected in the absence of selection. Preferred alleles with little quantitative difference ($|\Delta I| < 0.5$) in residue preferences show a nearly symmetric distribution (unfilled bars), whereas alleles predicted to be greatly preferred or unpreferred ($|\Delta I| > 6$) show a much stronger effect (black bars). Bins are labeled by the lower bound, such that "0" indicates SNPs with allele frequencies between 0 and 0.1. (*B*) The shift in preferred allele frequency increases quantitatively with increasing magnitude of predicted effect. Points are plotted by the lower bound of the bin they represent, such that the point at "2" indicates SNPs with ($2 \leq |\Delta I| < 3$). To summarize the allele frequency distribution, the number of SNPs with preferred allele frequency greater than 70% was divided by the number of SNPs with preferred allele frequency greater than 70% or less than 30%. Under the symmetric expectation in the absence of selection, this fraction equals ½ (dashed trace). Error bars represent twice the standard error of the proportion.

amino acid *a* under a background distribution. The emission probabilities *f* are precomputed based on a multiple alignment of many examples of that domain (Sonnhammer et al. 1998), and we treat these as given. For the background model, it will be mathematically convenient to choose the equilibrium probabilities of amino acids under the Jukes–Cantor model for DNA substitution (Liò and Goldman 1998); under this model, each amino acid is found proportional to its number of codons (see Methods). However, the results presented below can be obtained under a uniform background assumption as well, and these are presented as supplementary figure 1 (Supplementary Material online). Using the information content, $I_x$, as a quantitative measure of variability, we can compare the properties of SNPs at different positions in protein domains. For example, positions in C2H2 domains with high information content tend to show fewer nonsynonymous polymorphisms than SNPs at positions with low information content (fig. 1*D*).

Here we propose that residues more commonly observed at a given position in a protein domain are preferred (have greater fitness) than those that are less commonly observed at that position. In order to quantify the preference of one residue for another in a protein domain, we define

$$\Delta I_{xab} = \log_2 \frac{f_{xb}}{g_b} - \log_2 \frac{f_{xa}}{g_a},$$

associated with a change from residue *a* to residue *b*, where, once again, *x* is the position in the domain, *g* are background probabilities, and *f* are the probabilities in the Pfam HMM model. Inclusion of the background probabilities here accounts for the intuition that residue *a* should not be considered preferred over residue *b* in the domain, unless this preference exceeds that which is observed in the absence of selection. Thus, if the ratio of the probability of residue *b* at position *x* to the background is greater than the ratio of the probability of residue *a* at position *x* to the background, $\Delta I > 0$, and we consider residue *b* to be preferred. Similarly, if residue *a* is more probable relative to the background, $\Delta I < 0$, and we consider residue *a* to be preferred. Finally, if the probability ratios are equal, $\Delta I = 0$, and there is no preference between the two alleles. We note, once again, that the results presented below can be obtained without accounting for the background probabilities using the simpler definition $\Delta I_{xab} = \log_2 f_{xb} - \log_2 f_{xa}$ (supplementary fig. 1, Supplementary Material online).

## Frequencies of Preferred Alleles

To test the hypothesis that we can classify mutations in protein domains as preferred or unpreferred, we computed allele frequencies (see Methods) for the 1,879 amino acid changing SNPs in protein domains that were polymorphic in the YRI HapMap population (The International HapMap Consortium 2005). In the absence of selection, the frequencies of preferred alleles are expected to follow a symmetrical, U-shaped distribution (Mcvean and Charlesworth 1999). We found that the frequency of preferred polymorphisms in protein domains is not symmetrical, indicating the action of selection (fig. 2*A*, gray bars). Taking advantage of the quantitative metric for preference of an allele described above, we binned the polymorphisms by the magnitude of the change in $\Delta I$ ($|\Delta I|$). SNPs that were preferred but showed a small magnitude in change, that is, the preferred allele is not greatly preferred ($0 < |\Delta I| < 0.5$), showed a nearly symmetric distribution (fig. 2*A*, unfilled bars). On the other hand, SNPs that have large differences in preference ($|\Delta I| > 6$) show a greatly skewed distribution (fig. 2*A*, black bars). To test if this effect was quantitative, we compared the number of SNPs with preferred allele frequency greater than 70% with the number of SNPs with preferred allele frequency greater than 70% or less than 30%. If the preferred allele frequency distribution were symmetric (as expected in the absence of selection), this ratio is expected to be ½. However, we found that this ratio differed significantly from ½ for $|\Delta I| > 0.5$ (fig. 2*B*) and that this ratio was positively correlated with $|\Delta I|$ ($R^2 = 0.90$, $P < 0.001$), such that polymorphisms with

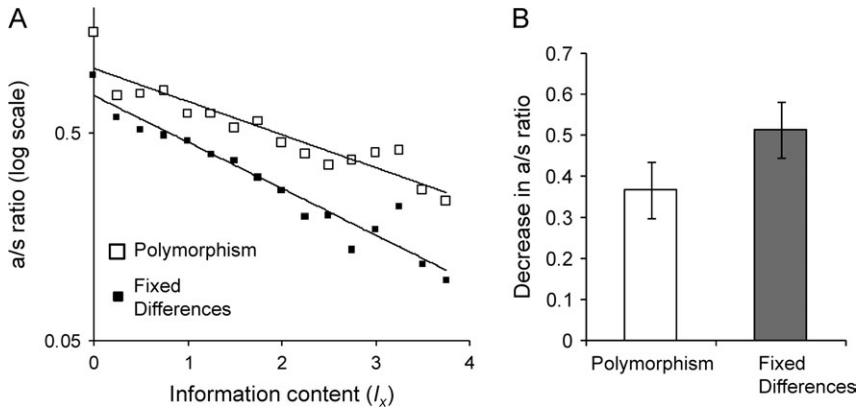


FIG. 3.—The a/s ratios are correlated with predicted effect of mutations in protein domains. (*A*) For both human polymorphism (unfilled symbols) and fixed differences along the human linage (filled symbols), the a/s ratio decreases at positions in domains with higher information content. Points are plotted by the lower bound of the bin of information content they represent. (*B*) The rate of decrease is greater for fixed differences than for polymorphism. Error bars represent twice the standard error of the slope in the regression of log of a/s ratio on information.

greater predicted effects on the protein domain show greater skew in the preferred allele frequency.

## Position-Specific Divergence and Diversity in Protein Domains

Selection is expected to affect not only the frequencies of polymorphisms but also the abundance. Positions in protein domains with strong residue preferences can be viewed as more important for the domain function and mutations at these positions will have on average large negative values of $\Delta I$. Such changes are expected to be removed from the population by purifying selection; these will be less abundant in the population and less likely to be fixed by genetic drift.

To compare the number of polymorphisms at different positions in protein domains, we computed the ratio of amino acid changing variants to synonymous variants (a/s ratio) as a function of the information content of the domain position ($I_x$) at which they occur. We find a strong negative correlation ($R^2 = 0.87$, fig. 3*A*, unfilled symbols) between the logarithm of the a/s ratio and $I_x$, consistent with the action of purifying selection removing more amino acid changing mutations at positions with higher information content. We also computed the a/s ratio of fixed differences (see Methods) along the human lineage since divergence with chimp and found a similar, strong negative correlation ($R^2 = 0.93$, fig. 3*A*, filled symbols) with $I_x$. Interestingly, we note that the rate of decrease of a/s ratio is faster for the substitutions than for polymorphism (fig. 3*B*), consistent with the prediction that the effects of natural selection are felt more strongly on fixed differences than on polymorphisms (e.g., Kimura 1984). Thus, these results indicate that natural selection is affecting the number of fixed differences as well as the number of polymorphisms in protein domains.

## Separating the Effects of Positive and Negative Selection

The elevated frequency of preferred alleles and reduction in a/s ratio at positions with high information content supports the model that selection is acting to preserve the residues at each position in the domain. However, this could be accomplished either by purifying selection reducing the frequencies and removing new nonpreferred alleles or by positive selection increasing the frequencies of and fixing new preferred alleles or by some combination of both.

In order to test whether both these processes were involved, we used chimp sequence to infer the ancestral allele for each SNP (see Methods). For the 1,630 SNPs where this was possible, we compared the derived allele frequencies in bins of $\Delta I$ with that for the 2,725 SNPs at in synonymous positions in the protein domains (see Methods, fig. 4*A*). Consistent with the action of purifying selection removing unpreferred alleles in the human population, we find a leftward skew in the allele frequencies of SNPs predicted to be strongly unpreferred $\Delta I < -4$ (fig. 4*A*, green bars). However, Although we observed a weak positive shift relative to synonymous sites for alleles predicted to be preferred, $\Delta I > 2$ (fig. 4*A*, red bars), this difference was not significant. Interestingly, we again found a quantitative relationship between the allele frequency spectrum and the predicted effect of the mutation, such that mutations predicted to be deleterious showed a lower fraction of high-frequency alleles (defined as derived allele frequency >30%), and mutations predicted to be beneficial showed a higher fraction of high-frequency alleles ($R^2 = 0.64$, $P = 0.03$, excluding $\Delta I = 4$, $R^2 = 0.53$, $P = 0.09$, fig. 4*B*). However, the fraction of high-frequency–derived alleles in mutations predicted to be beneficial did not significantly exceed that of synonymous sites (fig. 4*B*, dotted trace). Thus, the derived allele frequencies in the current human population provide some evidence for negative selection acting to preserve the residues in protein domains.

It is possible that the number of beneficial alleles in the current population is too small to detect the effects of positive selection. We therefore sought evidence of historical fixations of beneficial alleles in protein domains by comparing the numbers of preferred and unpreferred alleles in segregating polymorphism to that in fixed differences between human and chimp. We find a significant ($P < 10^{-6}$, Fisher's exact Test) difference in the distribution of $\Delta I$ in these two classes (fig. 5*A*), indicating the effects of selection. To

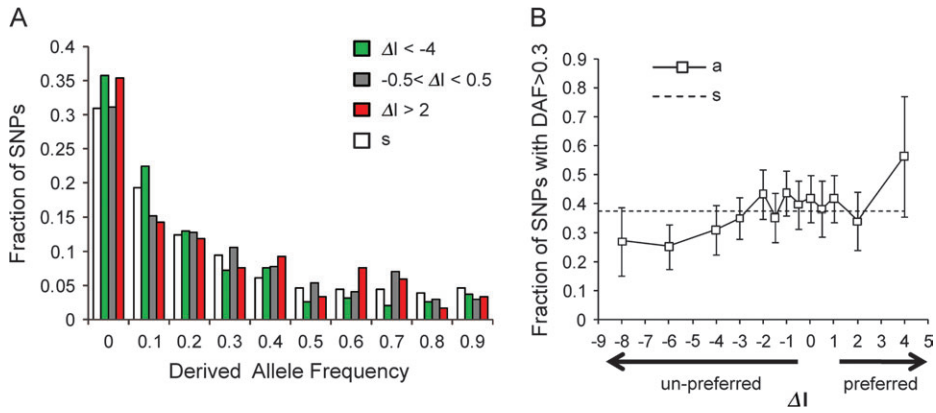


FIG. 4.—Derived allele frequencies of preferred and unpreferred polymorphisms. (*A*) Polymorphisms that are predicted to have a strong deleterious effect (green bars) show a skewed derived allele frequency distribution relative to synonymous sites (unfilled bars). Polymorphisms that are predicted to have little effect (gray bars) or a strong positive effect (red bars) show no difference from synonymous sites. Bins are labeled by the lower bound, such that "0" indicates SNPs with allele frequencies between 0 and 0.1. (*B*) The proportion of high-frequency–derived alleles (frequency >30%) is correlated with the predicted effect of the polymorphism. Points are plotted by the lower bound of the bin of $\Delta I$ they represent, such that "−6" indicates the SNPs with $(-6 \leq \Delta I < -4)$. Error bars represent twice the standard error of the proportion.

distinguish the effects of positive and negative selection, we compared the ratio of amino acid polymorphism and fixed differences for preferred and unpreferred alleles with synonymous polymorphism and fixed differences. This approach was suggested by McDonald and Kreitman (1991) and was extended to preferred and unpreferred synonymous alleles soon after (Akashi 1995). To extend this framework to polymorphism and divergence in protein domains, we defined the MK test as follows. We counted the number of fixed differences and polymorphisms (see Methods) in a given interval of $\Delta I$. We then divided by the number of synonymous differences or polymorphisms in the protein

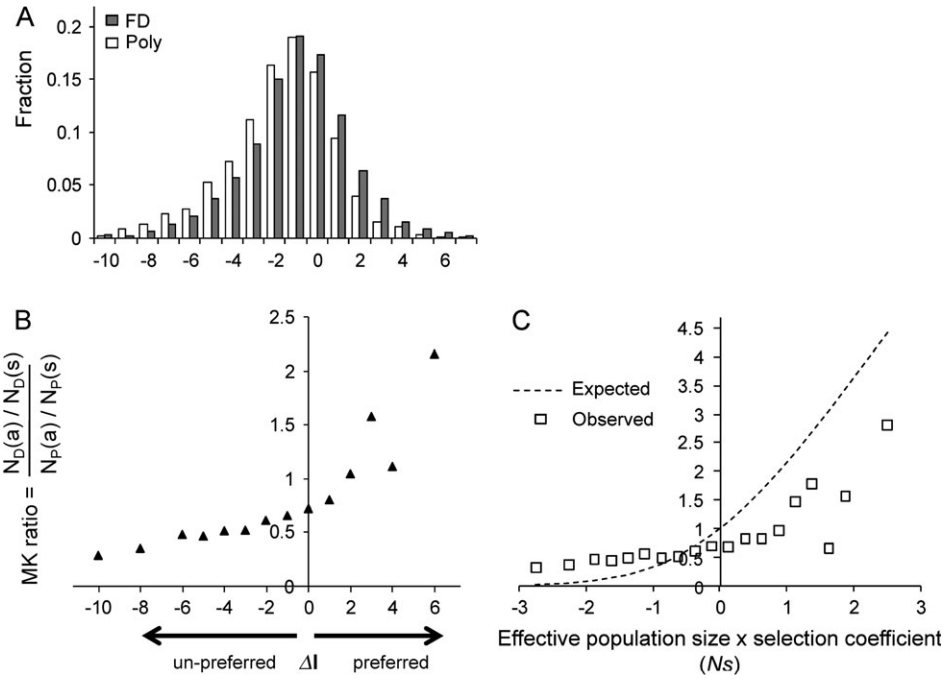


FIG. 5.—Divergence and polymorphism in protein domains. (*A*) Distribution of $\Delta I$ in polymorphism (unfilled bars) and divergence (gray bars). Bins are labeled by the lower bound, such that "0" indicates SNPs with $0 \leq \Delta I < 1$. (*B*) MK ratios comparing the amino acid replacement to synonymous ratio for divergence (D) to polymorphism (P), as a function of the predicted effect (triangles). That this ratio is less than one for unpreferred changes, and greater than one for preferred changes indicates the action of both negative and positive selection on unpreferred polymorphism and preferred fixed differences respectively. Points are plotted by the lower bound of the bin of $\Delta I$ they represent, such that "−6" indicates the SNPs with $-6 \leq \Delta I < -5$. (*C*) Compares a theoretical prediction for the dependence of the MK ratio on $Ns$ to what is observed in proteins domains. For details, see text. Points are plotted by the average of the bounds of the bin of $Ns$, such that "0.5" indicates SNPs with $0 \leq Ns < 1$.

**Table 1**
**MK Tests Provide Evidence for Positive and Negative Selection on Fixed Amino Acid Substitutions in Protein Domains**

| Site Class | Synonymous Changes in Domains | Amino Acid Changes in Domains | | | | |
|---|---|---|---|---|---|---|
| | | $\Delta I < -3$ | $-3$ c $\Delta I < 0$ | $0 < \Delta I < 1$ | $1 < \Delta I < 3$ | $\Delta I > 3$ |
| SNPs | 2,704 | 343 | 752 | 253 | 216 | 53 |
| Fixed differences | 10,339 | 592 | 1,740 | 699 | 724 | 288 |
| MK ratio | | 0.44 | 0.65 | 0.72 | 0.88 | 1.42 |
| *P* value | | $<10^{-15}$ | $<10^{-15}$ | $<10^{-4}$ | 0.11 | 0.02 |

domains to obtain an a/s ratio for divergence and polymorphism for each bin. We defined the MK ratio as the ratio of these ratios, such that

$$\text{MK ratio}_{i<\Delta I\leq j} = \frac{\frac{N_D(i<\Delta I\leq j)}{N_D(s)}}{\frac{N_P(i<\Delta I\leq j)}{N_P(s)}},$$

where $N_D$ and $N_P$ are the numbers of fixed differences or polymorphic sites with $\Delta I$ between $i$ and $j$ or at synonymous sites ($s$) in all the protein domains. In the absence of selection (and under certain assumptions, see McDonald and Kreitman 1991), this ratio is expected to be one. Significance can be tested using a Fisher's exact test, under the assumption that sites are i.i.d.

The results of this analysis are shown in figure 5*B* and table 1. Consistent with the action of purifying selection removing unpreferred mutations, we find MK ratios significantly less than one for substitutions with negative values of $\Delta I$. We also find MK ratios significantly greater than one for preferred substitutions, suggesting that a portion of these have been fixed by positive selection.

In general, we were concerned about the effect of CpG hypermutable sites on our results concerning divergence and polymorphism.We therefore performed the MK analysis described above excluding C/T differences that are followed by G or A/G differences preceded by a C. The results (table 2) were similar to those found using all SNPs indicating that CpG bias does not significantly impact our results.

We noted that the shape of the divergence to polymorphism ratio as a function of $\Delta I$ (fig. 5*B*) shows striking qualitative similarity to the theoretical ratio as a function of the product of effective population size and selection coefficient, $Ns$ (see Methods, fig. 5*C*, dotted trace). We therefore investigate the relationship between these quantities below.

## The Halpern–Bruno Model for Selection on Amino Acid Changes in Protein Domains

The emission probabilities, $f$, at each position in protein domains can be regarded as the equilibrium distribution of a mutation and selection process that has occurred over long evolutionary time. Under certain assumptions, we can relate the observed equilibrium probabilities to the selection coefficients (Halpern and Bruno 1998). For a given type of substitution at a particular position in the domain, the rate of evolution, $R$, can be written as

$$R_{xab} = Q_{ab}F_{xab},$$

where $Q$ is a mutation matrix giving the spontaneous rate of mutation of residue $a$ to residue $b$ and $F_{xab}$ is the probability of fixation of that particular type of mutation at position $x$ in the domain. This probability of fixation depends on the selection coefficient $s_{xab}$ associated with the mutation from $a$ to $b$ at position $x$:

$$F_{xab} \cong \frac{2s_{xab}}{1 - e^{-2Ns_{xab}}},$$

where $N$ is the effective population size (Kimura 1962). Given reversibility, the ratio of equilibrium probabilities at each position is determined by the ratio of substitution rates,

$$\frac{f_{xb}}{f_{xa}} = \frac{R_{xab}}{R_{xba}} = \frac{Q_{ab}F_{xab}}{Q_{ba}F_{xba}},$$

which implies that $\frac{F_{xab}}{F_{xba}} = \frac{Q_{ba}f_{xb}}{Q_{ab}f_{xa}}$. Assuming that the reverse mutation from $b$ to $a$ at position $x$ is associated with selection coefficient $-s_{xab}$, Halpern and Bruno noted that

$$\frac{F_{xab}}{F_{xba}} = \frac{e^{2Ns_{xab}} - 1}{1 - e^{-2Ns_{xab}}} = e^{2Ns_{xab}},$$

which implies that $\frac{Q_{ba}f_{xb}}{Q_{ab}f_{xa}} = e^{2Ns_{xab}}$ Because we have assumed an average over all domains and long evolutionary time (Pfam alignments include sequences from all kingdoms of life), there was no obvious choice for GC content or transition–transversion rate ratio. We therefore used the Jukes-Cantor model for the DNA evolution. Because this model is also reversible, the equilibrium probabilities, $g$, are given by the ratio of mutation rates: $\frac{Q_{ba}}{Q_{ab}} = \frac{g_a}{g_b}$. This means that we can write $\frac{g_a f_{xb}}{g_b f_{xa}} = e^{2Ns_{xab}}$ or

$$\log\frac{f_{xb}}{g_b} - \log\frac{f_{xa}}{g_a} = 2Ns_{xab},$$

where log is the natural logarithm. Therefore, from the definition of $\Delta I$, we have

$$\Delta I_{xab}\log 2 = 2Ns_{xab}.$$

Thus, under this model and choice of background distribution, the $\Delta I$ associated with a mutation is directly proportional to the selection coefficient.

To test the predictions of this model, we computed $Ns_{xab} = \frac{1}{2}\Delta I_{xab}\log 2$ for each of our SNPs and fixed differences in Pfam domains and computed the MK ratio as a function of the predicted $Ns$. We compared these data with

**Table 2**
**Control for CpG Effects**

| Site Class | Synonymous Changes in Domains | Amino Acid Changes in Domains | | | | |
|---|---|---|---|---|---|---|
| | | $\Delta I < -3$ | $-3 < \Delta I < 0$ | $0 < \Delta I < 1$ | $1 < \Delta I < 3$ | $\Delta I > 3$ |
| SNPs | 1,472 | 237 | 494 | 160 | 124 | 31 |
| Fixed differences | 5,684 | 384 | 1,182 | 451 | 465 | 195 |
| MK ratio | | 0.42 | 0.62 | 0.73 | 0.97 | 1.63 |
| *P* value | | $<10^{-15}$ | $<10^{-13}$ | 0.001 | 0.79 | 0.01 |

NOTE.—Calculations in table 1 repeated excluding polymorphism and substitutions where C/T difference is followed by a G or an A/G difference is preceded by a C.

the theoretical dependence of the MK ratio on the selection coefficient (see Methods):

$$\text{MK ratio} = \frac{2Ns\log(2n-1)}{\int_{\frac{1}{2n}}^{1-\frac{1}{2n}} \frac{1-e^{-2Ns(1-y)}}{y(1-y)}\,dy},$$

where $s$ is the selection coefficient for differences in a particular bin, $y$ is the frequency of a mutant allele, $N$ is the effective population size, and $n$ is the actual population size so that $^1/_{2n}$ is the frequency of a new mutation. Comparison of these predictions and our observations (fig. 5*C*) indicates that despite qualitative agreement, our model overestimates the strength of selection by a constant factor of about two (supplementary fig. 2, Supplementary Material online). Nevertheless, these results support the model that both weakly advantageous and deleterious substitutions can be identified in protein domains and suggest that analysis of these differences represents a means to test quantitative models of selection (see Discussion).

## Discussion

### The Analogy between Protein Domains and Synonymous Sites

In attempting to define preferred and unpreferred changes in conserved protein domains, we have suggested that mutations that lead to more probable residues in domain models are analogous to mutations to codons for more abundant tRNAs. However, there are important differences between these two situations. First, in protein coding regions, we must consider the effects of the genetic code on the substitution process: this leads to nonuniform probabilities of amino acids even in the absence of selection. We have accounted for this by comparing the probabilities of residues in the protein domain to those expected at equilibrium under the Jukes–Cantor DNA substitution model. Under this model, the equilibrium probabilities of amino acids are proportional to the number of codons that code for each amino acid, which explains much of the variance in the substitution probabilities of amino acids in real proteins (Ohta and Kimura 1971).

Another important difference between protein domains and synonymous sites is that residues in protein domains may be under more complex constraints that are not reflected in the probability distribution in the protein domain model. For example, in C2H2 Zn fingers (fig. 1), the central residues are responsible for the DNA-binding specificity. Because this is a property of individual genes and not the C2H2 domain family as a whole, this functional constraint is not captured in the domain model. Therefore, mutations with $\Delta I = 0$ might still be deleterious or advantageous. It may be more appropriate to consider only the relative differences between classes of sites in protein domains, as opposed to comparisons with synonymous sites.

### A Model for Evolution of Protein Domains

The Halpern–Bruno model applied above assumes no population structure as well as constant ploidy, population size, and selection coefficients. Because the Pfam alignments include domains from diverse proteins from diverse species, these assumptions cannot possibly hold. Therefore, the qualitative agreement of the predictions with our data should be taken as evidence that the models are quite robust, and the predictions represent some sort of average, both over long evolutionary time as well as over population genetic parameters. Understanding the effects of this averaging is an area for further research. In addition, this model assumes that mutations affect protein domains independently of the other residues in the domain; this is also unlikely to be true. However, it may be possible to account for this by using models that consider the position of the entire domain's sequence on a fitness landscape (Berg et al. 2004; Choi et al. 2008).

Despite these simplifying assumptions, we were able to compare the quantitative relationship between the predicted selection coefficient and the MK ratio to theoretical expectations (Sawyer and Hartl 1992). Our observations lend empirical support to the quantitative picture of nearly neutral alleles obtained using classical methods (Kimura 1984) and are consistent with models where mutation–selection balance preserves molecular function.

### Evidence for Weakly Deleterious Polymorphism and Weakly Advantageous Fixed Differences

Our analysis of polymorphism in protein domains provides evidence for large numbers of slightly deleterious alleles. The amino acid replacement to synonymous ratios for changes predicted to be deleterious ($Ns < 0$) were 0.23 and 0.40 for divergence and polymorphism, respectively. This suggests that 44% ($1 - 0.22/0.40$) of the 1,095 segregating polymorphisms predicted to be deleterious will eventually be removed by natural selection (Smith and Eyre-Walker 2002). These results are consistent with several lines of evidence suggesting that large numbers of slightly deleterious amino acid polymorphisms segregate in the human population (Charlesworth and Eyre-Walker 2007).

Identifying weakly deleterious variants in the human population is of great interest as they may be associated with human disease (Ng and Henikoff 2006). Indeed, several methods have recently been proposed that use protein domains to predict the functional consequences of mutations (Clifford et al. 2004; Han et al. 2006; Worth et al. 2007). Our results, as well as recent work predicting selection coefficients from protein structure and protein domain models (Thorne et al. 2007; Choi et al. 2008), represent an important step in relating those methods to models of selection and population genetics.

In addition to the weakly deleterious mutations described above, it has been observed that the nearly neutral theory of molecular evolution must posit weakly advantageous mutations as well. Although there has been abundant evidence for weakly deleterious mutations, the evidence for weakly advantageous mutations is scarcer.

Weakly advantageous alleles are thought to include slightly beneficial "back mutations" (Charlesworth and Eyre-Walker 2007). These back mutations may be fixed by positive selection but represent the correction of previous slightly deleterious substitutions; they therefore do not represent bona fide adaptation of an organism to its environment. Because the structural and functional constraints on protein domains are likely to be conserved over evolution, the excess of fixed preferred mutations that we observed may indicate the correction of previous weakly deleterious substitutions. Mutations to preferred alleles in protein domains could represent an important example of weakly advantageous back mutations. Thus, our results suggest that assigning preferred and unpreferred states to mutations may facilitate detection of advantageous back mutations and may provide an approach to distinguish these from truly "adaptive" changes.

### The Relationship between Bioinformatic Information and Natural Selection

As large data sets of SNPs become increasingly available, the ability to assign quantitatively preferred and unpreferred states to a large number of amino acid differences will allow tests of detailed models of selection. In fact, the availability of large numbers of protein sequences from many species means that probabilistic models specifying the preferences for each residue at each position will soon be available for all genes (see e.g., TreeFam, Li et al. 2006). This implies that the methods proposed here will eventually be applicable to nearly all protein coding polymorphisms and fixed differences.

That natural selection can increase and preserve information against the entropic force of mutation is of considerable interest (Kimura 1961; Schneider 2000). We believe that models relating skewed residue preferences and information content in bioinformatics models to evolutionary quantities, such as allele frequency distributions and a/s ratios will yield insight. Although there has been recent progress in this area (Halpern and Bruno 1998; Moses et al. 2003; Tang et al. 2004; Berg et al. 2004; Gojobori et al. 2007; Choi et al. 2008), a more complete theoretical framework that yields simple testable predictions will be of great utility.

## Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

We thank Dr Avril Coghlan for helpful discussions and comments on the manuscript. We thank Drs Casey Bergman, Michael Lässig, Paul Flicek, Stephen Wright, and Asher Cutter for useful discussions. We thank Drs Bill Bruno, Jeffrey Thorne, and an anonymous reviewer for helpful critique of the manuscript. R.D., A.M.M., and the Wellcome Trust Sanger Institute are supported by the Wellcome Trust.

## Literature Cited

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics. 139(2):1067–1076.

Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics. 146(1):295–307.

Berg J, Willmann S, Lässig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4(1):42.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129(3):897–907.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci USA. 104(43):16992–16997.

Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. 2004. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics. 20(7):1006–1014.

Choi SC, Redelings BD, Thorne JL. 2008. Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. Philos Trans R Soc Lond B Biol Sci. doi: U026W7637GVT1296.

Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. Curr Biol. 16(20):2053–2057.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis. Cambridge: Cambridge University Press.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H, et al. (11 co-authors). 2008. The Pfam protein families database. Nucleic Acids Res. 36(Suppl_1):D281–D288.

Gojobori J, Tang H, Akey JM, Wu C. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. Proc Natl Acad Sci USA. 104(10):3907–3912.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 15(7):910–917.

Han A, Kang HJ, Cho Y, Lee S, Kim YJ, Gong S. 2006. SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. Nucleic Acids Res. 34(Web Server issue):W642–W644.

Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. Genetics. 138(1):227–234.

Hubbard T, Barker D, Birney E, et al. (35 co-authors). 2002. The Ensembl genome database project. Nucleic Acids Res. 30(1):38–41.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat. 5(3):299–314.
Kimura M. 1961. Natural selection as the process of accumulating genetic information in adaptive evolution. Genet Res. 2:127–140.
Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics. 47:713–719.
Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. 61(4):893–903.
Kimura M. 1984. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Li H, Coghlan A, Ruan J, et al. (15 co-authors). 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 34(Database issue):D572–D580.
Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 24(4):337–345.
Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8(12):1233–1244.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351(6328):652–654.
Mcvean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res. 74(02):145–158.
Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evolutionary Biology. 3(19): doi: 10.1186/1471-2148-3-19.
Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 168(4):2373–2382.
Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 7:61–80.
Ohta T, Kimura M. 1971. Amino acid composition of proteins as a product of molecular evolution. Science. 174(5):150–153.
Ruan J, Li H, Chen Z, et al. (19 co-authors). 2008. TreeFam: 2008 Update. Nucleic Acids Res. 36(Database issue): D735–D40.
Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics. 132(4):1161–1176.
Schneider TD. 2000. Evolution of biological information. Nucleic Acids Res. 28(14):2794–2799.
Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18(20): 6097–6100.
Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. J Mol Biol. 188(3):415–431.
Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci. 12(4):327–345.
Sonnhammer E, Eddy S, Birney E, Bateman A, Durbin R. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 26(1):320–322.
Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415(6875):1022–1024.
Tang H, Wyckoff GJ, Lu J, Wu C. 2004. A universal evolutionary index for amino acid changes. Mol Biol Evol. 21(8):1548–1556.
The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature. 437(7063):1299–1320.
Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. 2007. Population genetics without intraspecific data. Mol Biol Evol. 24(8):1667–1677.
Worth CL, Bickerton GRJ, Schreyer A, Forman JR, Cheng TMK, Lee S, Gong S, Burke DF, Blundell TL. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. J Bioinform Comput Biol. 5(6):1297–1318.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25(3):568–579.

Jeffrey Thorne, Associate Editor

Accepted December 9, 2008