Published in final edited form as:

Immunogenetics. 2008 September ; 60(9): 495–506. doi:10.1007/s00251-008-0304-4.

Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes

Austin L. Hughes and Robert Friedman

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia, SC 29208, USA

Austin L. Hughes: austin@biol.sc.edu

Abstract

Using basic probability theory, we show that there is a substantial likelihood that even in the presence of strong purifying selection, there will be a number of codons in which the number of synonymous nucleotide substitutions per site (d_S) exceeds the number of non-synonymous nucleotide substitutions per site (d_N). In an empirical study, we examined the numbers of synonymous (b_S) and non-synonymous substitutions (b_N) along branches of the phylogenies of 69 single-copy orthologous genes from seven species of mammals. A pattern of $b_N > b_S$ was most commonly seen in the shortest branches of the tree and was associated with a high coefficient of variation in both b_N and b_S , suggesting that high stochastic error in b_N and b_S on short branches, rather than positive Darwinian selection, is the explanation of most cases where b_N is greater than b_S on a given branch. The branch-site method of Zhang et al. (Zhang, Nielsen, Yang, *Mol Biol Evol*, 22:2472–2479, 2005) identified 117 codons on 35 branches as “positively selected,” but a majority of these codons lacked synonymous substitutions, while in the others, synonymous and non-synonymous differences per site occurred in approximately equal frequencies. Thus, it was impossible to rule out the hypothesis that chance variation in the pattern of mutation across sites, rather than positive selection, accounted for the observed pattern. Our results showed that b_N/b_S was consistently elevated in immune system genes, but neither the search for branches with $b_N > b_S$ nor the branch-site method revealed this trend.

Keywords

Immune system evolution; Non-synonymous substitution; Positive Darwinian selection; Stochastic error; Synonymous substitution

Natural selection takes two major forms: (1) purifying selection, which acts to eliminate deleterious mutations, and (2) positive (Darwinian) selection, which favors genotypes conferring a fitness advantage on the organism (Hughes 1999). Positive selection in turn can be either directional (tending toward fixation of a favored allele) or balancing (maintaining a polymorphism). There is evidence that as predicted by the neutral theory of molecular evolution, purifying selection is the predominant form of natural selection at the molecular sequence level (Kimura 1977). In most protein-coding genes, the number of synonymous nucleotide substitutions per site (d_S) exceeds the number of non-synonymous nucleotide

substitutions per site (d_N), evidence that purifying selection has acted to eliminate a substantial fraction of non-synonymous mutations (Nei 1987; Hughes 1999).

Hughes and Nei (1988) used comparison of d_S and d_N to test Doherty and Zinkernagel's (1975) hypothesis that the polymorphism of classical class I major histocompatibility complex (MHC) loci is maintained by balancing selection relating to disease resistance. They reasoned that if positive selection favors amino acid diversity in the peptide-binding region (PBR) of the molecule, one might see a pattern of $d_N > d_S$ in the codons encoding the PBR. The results for the PBR matched this prediction, whereas in the remainder of the gene, $d_S > d_N$ as in most genes (Hughes and Nei 1988). Similar results were found for class II MHC loci as well (Hughes et al. 1994). Comparison of d_S and d_N was also used to test the hypothesis that recognition by the class I MHC and cytotoxic T lymphocytes (CTL) selects for escape variants in CTL epitopes within viral proteins, most notably in an experimental system involving Simian immunodeficiency virus (SIV) infection of rhesus monkeys (O'Connor et al. 2004).

As a result of these studies, it has frequently been asserted that a pattern of $d_N > d_S$ is a "signature of positive selection" and that searching sequence data sets for this pattern represents a way of discovering genes that have been subject to positive selection in the past. On the other hand, Hughes and Friedman (2005) argued that because of the stochastic nature of the mutational process, a substantial fraction of codons are likely to show $d_N > d_S$ by chance alone; thus, $d_N > d_S$ cannot accurately be called a "signature of positive selection." Moreover, it has been pointed out that both the studies of the MHC and CTL escape in SIV constituted tests of an a priori hypothesis, based on biological reasoning, rather than an undirected "fishing expedition" in search of a particular pattern of nucleotide substitution (Hughes et al. 2006; Hughes 2007). Despite these problems, studies that report searches for either branches of phylogenetic trees or individual codons with $d_N > d_S$ have become increasingly common, filling the literature with numerous claims of positive selection for which no biological basis is apparent (Hughes 2007).

In this paper, we examine aspects of synonymous and non-synonymous substitution in the light of simple probability theory and show that a substantial fraction of codons are likely to $d_N > d_S$ by chance alone, as suggested by Hughes and Friedman (2005). In addition, using analysis of a set of 69 single-copy orthologous genes, including 30 with immune system function, from each of seven mammalian species, we illustrate some problematic properties of two commonly used approaches used to test for positive Darwinian selection: (1) comparison of the number of synonymous nucleotide substitutions per synonymous site (b_S) and the number of non-synonymous substitutions per non-synonymous site (b_N) along individual branches of a phylogenetic tree and (2) so-called codon-based methods, which search for codons with $d_N > d_S$ (Yang 1997, 1998; Yang et al. 2005), in particular the so-called branch-site methods, which search for individual codons at which d_N exceeds d_S along a given branch of a phylogenetic tree (Zhang et al. 2005).

A number of studies have compared estimates of b_S and b_N along individual branches of a phylogenetic tree as a means of testing for positive selection. In some studies of individual branches, statistical methods have been used to provide a test of the hypothesis that $b_S = b_N$ along a given branch (e.g., Zhang et al. 1998). In other cases, however, the occurrence of any branch where b_N exceeds b_S has been taken as evidence of positive selection along that branch, either with no attempt to test statistically the null hypothesis that b_S and b_N are equal (e.g., Evans et al. 2006) or in the absence of statistical significance (Wang and Su 2004).

The branch-site methods provide an alternative approach to studying positive selection along a given branch or set of branches in a phylogenetic tree, and in this case, a likelihood ratio test can be used to compare a model incorporating the supposed effects of positive selection to a

model not incorporating those effects (Zhang et al. 2005). These methods make the questionable assumption that the existence of one or more codons with $d_N > d_S$ implies positive selection.

When non-synonymous substitution is elevated in a certain set of codons or along a given branch of a tree, a biologically relevant alternative to the hypothesis of positive selection is the hypothesis of reduced functional constraint (in other words, a relaxation of purifying selection). It is particularly important that the latter hypothesis be ruled out in cases where elevated non-synonymous substitution is proposed to play a role in the evolution of some major phenotypic adaptation. For example, several recent studies have focused on loci at which mutations are known to play a role in adult brain size in humans, including *microcephalin* and *ASPM*, and elevated b_N has been taken as evidence for the adaptive evolution of increased brain size in primates (Evans et al. 2004, 2006; Kouprina et al. 2004; Pavlicek and Jurka 2006; Wang and Su 2004). Yet, there is no known biochemical or developmental mechanism by which a series of amino acid changes in the proteins encoded by these genes might have caused a gradual increase in brain size over the evolution of primates. In cases of this sort, before the claim of positive selection is made, the hypothesis of a relaxation of purifying selection must be ruled out, since if the latter explanation turns out to be correct, it may be evidence that rather than playing an important role in the evolution of the phenotype in question, the protein studied is particularly *unimportant* for that phenotypic change (Hughes 2007).

In this study, we employ an empirical approach to examine the relationship between b_S and b_N and the results of the branch-site method of Zhang et al. (2005) across a phylogeny of mammals belonging to three orders (Primates, Carnivora, and Rodentia), chosen so as to provide internal and terminal branches of a variety of different lengths. We estimated b_S and b_N for each of a sample of 69 single-copy orthologous genes from each of seven species, including the *microcephalin* and *ASPM* genes previously proposed as cases of positive selection in primates. We applied the branch-site method of Zhang et al. (2005) separately to each branch of the phylogeny for each gene. In addition, because a number of previous studies have provided evidence of an accelerated rate of amino acid sequence evolution in proteins of the vertebrate immune system (Murphy 1993; Hughes 1997), we tested for differences with respect to the pattern of b_S and b_N and positive selection inferred by the branch-site method across the phylogeny between immune system genes, genes involved in non-immune signaling functions, and other genes.

Materials and methods

Sequence retrieval

Genome-wide sequence data were retrieved from Ensembl (Hubbard et al. 2007) for the following species (Linnean name and database version in parentheses): (1) from the order Primates: human (*Homo sapiens* NCBI 36), chimp (*Pan troglodytes* PanTro 2.1), and rhesus (*Macaca mulatta* Mmul 1.0); from the order Carnivora: dog (*Canis lupus familiaris* CanFam 2.0); and from the order Rodentia: mouse (*Mus musculus* NCBI m36) and rat (*Rattus norvegicus* RGSC 3.4). Sequences for another member of Carnivora, the cat (*Felis catus*), were obtained from Genbank (Benson et al. 2007) in February 2007. Data on the human genome physical map were also retrieved from Ensembl so that overlapping and unmapped genes could be identified and removed. For all species except cat, alternate transcripts were identified by the Ensembl gene naming system and removed. The protein-coding genes of all species ($n=135,698$) were then clustered into gene families ($n=19,027$) by BLASTCLUST (Altschul et al. 1997) using default values except for a 50/70 criterion (Hughes et al. 2005a) where 50% is the minimum amino acid similarity between any two matching sequences across at least 70% of the aligned lengths. Once all pairs of similar sequences were identified, BLASTCLUST then clustered them into sets of orthologs by a single-linkage method.

Sequences with each set of orthologous genes were globally aligned at the amino acid level using ClustalW (Thompson et al. 1994), and the alignment was imposed on the DNA sequence. To check that the genes were true orthologs, an unrooted phylogenetic tree of amino acid sequences was constructed for each family by the quartet puzzling method (TreePuzzle; Schmidt et al. 2002) with the JTT model of sequence evolution (Jones et al. 1992) and rates varying among sites according to a gamma distribution, and by the neighbor-joining (NJ; Saitou and Nei 1987) method using the JTT distance.

The total number of gene families was 69 after removing those lacking a single representative from all seven species and those for which the gene tree was not congruent with the species tree (Supplementary Table S1). The 69 genes were categorized functionally based on the published literature. A gene was categorized as an immune-system gene if its primary expression is in one or more of the specialized cells of the vertebrate immune system. A gene was classified as involved in cell-cell signaling if it encodes a signaling molecule (such as a cytokine or hormone) or the receptor for a signaling molecule.

Data analysis

The number of synonymous substitutions per synonymous site (b_S) and the number of non-synonymous substitutions per non-synonymous site (b_N) along each branch of the phylogenetic trees for each of the 69 sets of orthologs (total 759 individual branches in individual genes) were estimated by the maximum likelihood method using phylogenetic analysis by maximum likelihood (PAML; Yang 1997), given the unrooted phylogenetic tree with topology as illustrated in Fig. 1. Default parameters were specified in PAML except for the following parameters which were estimated separately for each set of orthologs: ratio of transitions to transversions, the shape parameter of the gamma-distributed evolutionary rate variation among sites, and codon frequencies. Rates of evolution were allowed to vary among branches.

We applied the branch-site method of Zhang et al. (2005) to each branch in each gene. In this method, a certain branch or set of branches (foreground) are chosen as those to which a positive selection model is to be applied, while a null model is applied to the other branches in the tree (background). In the present case, we chose each individual branch as the foreground, with the other branches as background. The null model assumes that, in both background and foreground, a proportion (p_0) of codons have a ratio d_N/d_S equal to ω_0 , where $0 < \omega_0 < 1$; and that a proportion (p_1) of codons have a ratio d_N/d_S equal to $\omega_1 = 1$. In the selection model, there are two additional classes of codons: (1) in $(1 - p_0 - p_1)p_0/(p_0 + p_1)$ of codons, d_N/d_S equals ω_0 in the background but ω_2 in the foreground, where $\omega_2 > 1$, and (2) in $(1 - p_0 - p_1)p_1/(p_0 + p_1)$, d_N/d_S equals ω_1 in the background but ω_2 in the foreground. In the likelihood ratio test comparing the two models, we used 2.71 as the 5% significance level and 5.41 as the 1% significance level following Zhang et al. (2005). We used the Bayes empirical Bayes (BEB) method to identify individual codons inferred to be subject to positive selection (Yang et al. 2005).

To provide a measure of how divergent a given value of b_S or b_N was in comparison to corresponding values for the same gene, we computed standardized scores (normalized scores or Z scores) for each value of b_S and b_N . The standardized score was computed by the formula $(x_i - \bar{x})/SD$, where x_i is the i th individual value of b_S or b_N , $\bar{x}$ is the mean value for of b_S or b_N for the gene, and SD is the standard deviation of b_S or b_N for the gene.

We used Nei and Gojobori's (1986) method to estimate the proportion of synonymous substitutions per synonymous site (p_S) and the proportion of non-synonymous substitutions per non-synonymous site (p_N) at individual codons identified as "positively selected" by the branch-site method. We used this simple counting method because the assumptions of more

complex methods are not met in the case of individual codons. The ancestral sequence at these codons was reconstructed by the maximum parsimony method (Swofford 1999).

In certain statistical analyses, the individual branch ($N=759$) was the unit of analysis, since substitutions along each individual branch are expected to be independent of those on other individual branches. In other analyses, the branch of the tree ($N=11$) was the unit of the analysis; in the latter case, the means of certain quantities were computed for all 69 individual branches corresponding to the 69 individual genes. Because the distributions of b_S and b_N were non-normal (see “Results”), we used nonparametric methods in most analyses. Statistical analyses were conducted using the Minitab statistical package, release 13 (<http://www.minitab.com>).

Results

Theoretical study

Consider a given codon, either in a pairwise comparison of two sequences or across a phylogeny of a number of sequences. Let x be the number of synonymous substitutions per synonymous site occurring at that codon. Whatever method is used to estimate synonymous and non-synonymous substitutions, x will be a continuous variable taking values from 0 to ∞ . If $P_{CS}(x)$ is the probability of x synonymous substitutions per synonymous site at a codon, we can define the cumulative distribution function

$$f_S(x) = \int_0^x P_{CS}(y) dy \quad (1)$$

Likewise, if $P_{CN}(x)$ is the probability of x non-synonymous differences per non-synonymous site at a codon, we can define the cumulative distribution function

$$f_N(x) = \int_0^x P_{CN}(y) dy. \quad (2)$$

The so-called codon-based tests of positive selection depend on the assumption that the existence of one or more codons with $d_N > d_S$ implies positive selection. This assumption is false if there exists in the universe even a single codon for which, in the absence of positive selection,

$$f_S(x)[1 - f_N(x)] > 0. \quad (3)$$

However, it is likely that inequality 3 holds in many cases.

To see this, let us consider a simple example involving discrete distributions. In practice, codon-based methods generally identify as “positively selected” any codon at which there are no synonymous differences and one or more non-synonymous differences (see “Results” below). Considering only cases where the number of synonymous differences can take positive integral values, let $Q_{CS}(i)$ be the probability of exactly i synonymous differences at a given codon ($i=0, 1, 2, \dots, \infty$). Then, the cumulative distribution function, which gives the probability that there are n or fewer synonymous differences at a codon, is

$$F_S(n) = \sum_{i=0}^n Q_{CS}(i). \quad (4)$$

If $Q_{CN}(i)$ be the probability of exactly non-synonymous differences at a given codon ($i=0, 1, 2, \dots, \infty$), the cumulative distribution function, which gives the probability that there are n or fewer non-synonymous differences at a codon, is

$$F_N(n) = \sum_{i=0}^n Q_{CN}(i). \quad (5)$$

Codons with no synonymous differences and one or more non-synonymous differences will occur in the absence of positive selection as long as

$$F_S(0)[1 - F_N(0)] > 0. \quad (6)$$

To show that this condition will often be met, we assumed a simple model by which synonymous and non-synonymous substitution follow Poisson distributions with different means. The effect of purifying selection was modeled by a lower mean number of differences per codon for non-synonymous substitutions than for synonymous substitutions. We modeled cases where the observed proportion of non-synonymous substitutions per non-synonymous site (p_N) is some fraction (0.1, 0.2, or 0.5) of the observed proportion of synonymous substitutions per synonymous site (p_S ; Fig. 2). To translate p_S and p_N into the expected numbers of synonymous and non-synonymous substitutions per codon (designated, respectively, μ_S and μ_N), we multiplied by the average number of synonymous sites per codon (0.745) and the average number of non-synonymous sites per codon (2.255) estimated by Hughes and Friedman (2005) for nearly two million fungal codons using the modified Nei–Gojobori method (Zhang et al. 1998) with a transition/transversion ratio of 4.5.

The results show that a certain number of codons with no synonymous differences and one or more non-synonymous differences are expected to occur even under strong purifying selection. Even when $p_S=0.10$ and $p_N=0.01$, it is expected that about 2% of codons will have this property (Fig. 2). As purifying selection becomes relaxed, the proportion of such codons increases (Fig. 2). Even though this model is oversimplified, it is worth noting that it provides a good fit to the data of Hughes and Friedman (2005) who conducted a codon-by-codon comparison of orthologous genes between two closely related species of fungi. Using the observed p_S (0.317) and p_N (0.040) in that data set and the same assumptions used in generating Fig. 2, we predict that about 6.7% of codons should have no synonymous differences and one or more non-synonymous differences. The observed frequency of codons with $p_N > p_S$ in that data set was 6.5% (Hughes and Friedman 2005).

It should be obvious that similar reasoning can apply to the estimation of the number of synonymous substitutions per synonymous site (b_S) and the number of non-synonymous substitutions per non-synonymous site (b_N) along branches of a phylogenetic tree. When branch lengths are very short, the probability that b_S is equal to zero or nearly zero becomes large. Under those circumstances, there will be a nonzero probability that b_N will exceed b_S by chance alone, even under strong purifying selection.

Differences among branches

The number of synonymous substitutions per synonymous site (b_S) and the number of non-synonymous substitutions per non-synonymous site (b_N) were estimated separately for 69 genes along each branch of the phylogenetic tree of seven mammalian species (Fig. 1). Both mean b_S and mean b_N varied substantially among the 11 branches of the tree (Table 1). The largest mean b_S (that for branch 11, ancestral to the common ancestor of mouse and rat) was more than 75 times the smallest mean b_S (that for branch 1, ancestral to human; Table 1).

Median b_S and median b_N for the 11 branches showed a similar pattern to that shown by the mean values, but median values were consistently slightly lower than the corresponding mean values, indicating a modest positive skew in the distributions of b_S and b_N .

When all 759 branches for the 69 genes were analyzed separately, b_S was greater than b_N in the case of 685 branches (90.2%); b_N was greater than b_S in the case of 72 branches (9.5%); and b_S and b_N were equal in the case of two branches (0.3%). The predominance of branches with $b_S > b_N$ was highly significant (sign test; $P < 0.001$), as expected given the prevalence of purifying selection on protein-coding genes (Nei 1987). However, the occurrence of genes with $b_N > b_S$ was not uniform throughout the tree; rather, cases of $b_N > b_S$ were significantly more likely to occur on certain branches than on others ($\chi^2 = 49.7$; 10 df; $P < 0.001$; Fig. 3a). The percentages of genes with $b_N > b_S$ was highest on the terminal branches ancestral to human (branch 1) and chimp (branch 2) and on the internal branch leading to the human/chimp ancestor (branch 3), while no genes showed $b_N > b_S$ on the internal branch leading to the mouse/rat ancestor (branch 11; Fig. 3a).

Using mean b_S along a branch as a measure of the evolutionary time along that branch, we tested the hypothesis that b_N was more likely to exceed b_S on shorter branches, in other words, those corresponding to shorter evolutionary time spans (Fig. 3b). There was a strong negative rank correlation coefficient ($r_s = -0.817$; $P = 0.002$) between mean b_S and the percentage of genes with $b_N > b_S$ (Fig. 3b).

The coefficient of variation (CV; the standard deviation expressed as a percent of the mean) provides a scale-free measure of variability. For all 11 branches, CV in b_N was greater than CV in b_S for the same branch (sign test; $P = 0.001$; Fig. 3c). CV in both b_S and b_N was significantly negatively correlated with mean b_S (Fig. 3c). Thus, both b_S and b_N showed greater levels of variability in shorter branches.

Functional categories

When genes were categorized by immune system function and function in cell–cell signaling, there was no difference among categories with respect to median b_S (Fig. 4a). However, median b_N differed significantly among categories, with the highest median b_N occurring in genes with immune system function but not signaling function (Fig. 4b). In individual comparisons, median b_N for each of the other categories differed significantly from that for genes with immune system function but not signaling function (Fig. 4b). The same pattern was seen when the analysis was applied to b_N/b_S (Fig. 4c). In the latter case, a total of ten individual branches were excluded because b_N/b_S was undefined (Fig. 4c). Overall, the results showed increased average b_N in genes encoding immune system proteins, particularly immune system proteins not directly involved in cell–cell signaling.

To examine how the distinct pattern of evolution in immune system genes was related to branch length, we plotted the ratio of mean b_N in immune system genes to that in other genes against mean b_S for the 11 branches (Fig. 5). There was a strong positive correlation between these two quantities ($r_s = 0.745$; $P = 0.008$; Fig. 5a). Branch 1 (ancestral to human) was an influential point in this relationship; branch 1 had both the lowest mean b_S and the lowest ratio of mean b_N in immune system genes to that in other genes (Fig. 5). However, even when this point was removed, there was still a positive correlation between mean b_S and the ratio of mean b_N in immune system genes to that in other genes ($r_s = 0.673$; $P = 0.033$).

We classified each individual branch based on whether or not $b_N > b_S$ and whether or not the gene had immune system function, and we examined the relationship between these two classificatory variables separately for branches 1–4 of the phylogenetic tree and branches 5–11 of the phylogenetic tree. These separate analyses were conducted because branches 1–4

were the shortest branches and included a substantial majority (50/72) of individual branches with $b_N > b_S$ (Fig. 1). In the case of branches 1–4, there was not a significant association between $b_N > b_S$ and immune function ($\chi^2=1.80$; 1 *df*; n.s.). In the case of branches 1–4, 24 of 156 (15.4%) individual branches from genes without immune system function had $b_N > b_S$, while 26 of 120 (21.7%) individual branches from genes with immune system function had $b_N > b_S$. On the other hand, in the case of branches 5–11, there was a significant association between $b_N > b_S$ and immune function ($\chi^2=8.02$; 1 *df*; $P=0.005$). In the case of branches 5–11, only six of 273 (2.2%) individual branches from genes without immune system function had $b_N > b_S$, whereas 16 of 210 (7.6%) individual branches from genes with immune system function had $b_N > b_S$.

The *ASPM* and *microcephalin* loci, previously identified as candidates for positive selection relating to brain size increase in primates (Evans et al. 2004, 2006; Kouprina et al. 2004; Pavlicek and Jurka 2006; Wang and Su 2004) were among the 69 genes analyzed here. In the present analysis, no branch of the *ASPM* tree showed $b_N > b_S$. In the *microcephalin* tree, only one branch (branch 3, ancestral to human and chimp, Fig. 1) showed $b_N > b_S$, and on that branch, both b_S (0.0069) and b_N (0.0139) were quite low.

Branch-site method

We applied the branch site method of Zhang et al. (2005), in turn setting each of the 759 branches in turn as the “foreground.” The likelihood ratio test yielded a significant result at the 5% level in the case of 82 branches (10.8% of total). The likelihood ratio test was significant in the case of only seven of the 72 branches with $b_N > b_S$ (9.7%). A similar proportion of the remaining branches (75/687 or 10.9%) showed significant likelihood ratio tests. There was not a significant difference between branches with $b_N > b_S$ and other branches with respect to the proportion of branches with significant likelihood ratio tests ($\chi^2=0.10$; 1 *df*; n.s.). Likewise, there was not a significant difference with respect to frequency of a significant likelihood ratio test between branches from genes with immune system function (39/330 or 11.8%) and those without immune system function (43/429 or 10.0%; $\chi^2=0.62$; 1 *df*; n.s.). The likelihood ratio test was not significant for any branch in the case of either the *ASPM* gene or the *microcephalin* gene.

A more stringent criterion for cases where the branch-site method infers “positive selection” would be to include only cases with both a significant ($P<0.05$) likelihood ratio test and one or more codons with 95% or greater probability of $\omega > 1$ by the BEB method. A total of 35 branches met this criterion. There was not a significant difference in the proportion of branches deemed significant by this criterion between those with $b_N > b_S$ (3/72 or 4.2%) and other branches (32/687 or 4.7%; $\chi^2=0.04$; 1 *df*; n.s.). Likewise, there was not a significant difference in the proportion of branches deemed significant by this criterion between branches from genes with immune system function (17/330 or 5.2%) and those without immune system function (18/429 or 4.2%; $\chi^2=0.38$; 1 *df*; n.s.).

A total of 117 codons from 28 genes were identified as “positively selected” by a significant likelihood ratio test and 95% probability of $\omega > 1$. These codons occurred on all 11 branches of the tree except branch 2 (ancestral to chimpanzee; Fig. 1). We used partial rank correlation to analyze the relationships between standardized scores of b_S and b_N and the number of codons identified as “positively selected.” We computed second-order partial rank correlations among these three variables simultaneously controlling for the other two variables. There was a significant positive partial rank correlation between the standardized score of b_S and that of b_N (0.813; $P<0.001$). There was also a significant positive partial rank correlation between the standardized score of b_N and the number of “positively selected” codons (0.219; $P<0.001$). On the other hand, there was a significant negative partial rank correlation between the standardized score of b_S and the number of “positively selected” codons (−0.128; $P<0.001$). Similarly, we used standardized scores of b_S and that of b_N as predictor variables in a logistic

regression predicting “positive selection” on a branch. There were significant effects ($P < 0.001$) of standardized scores of both b_S and that of b_N ; but the coefficient for the former was negative (-0.993), while that for the latter was positive ($+1.573$). Thus, “positive selection” was in general associated with unusually low b_S on a branch but with unusually high b_N .

Using Nei and Gojobori's (1986) method, we estimated the proportions of synonymous (p_S) and non-synonymous (p_N) differences at these codons along the “positively selected” branch. At 67 of 117 codons (57.3%), there were no synonymous differences (Fig. 6). At the remaining 50 codons, the distributions of p_S and p_N were very similar (Fig. 6). In the case of these 50 codons, mean p_S was 1.074 ± 0.071 S.E. (median=1.000), mean p_N was 0.985 ± 0.018 (median=1.000), and mean p_N/p_S was 1.083 ± 0.066 (median=1.000).

The 67 “positively selected” codons at which there were no synonymous differences were found on 27 of the 759 total branches. None of these branches showed a value of b_S equal to zero. In fact, median b_S for these 27 branches (0.1142) was slightly but significantly higher than that (0.0847) for the remaining 732 branches ($P = 0.037$; Kruskal–Wallis test). These results suggest that these 67 codons represented codons at which synonymous substitution was unusually low in comparison to that in other codons in the gene along the same branch.

Somewhat surprisingly, the 117 “positively selected” codons included five at which no amino acid difference was observed. Rather, at these five codons, a change occurred between the two sets of serine codons. This was reconstructed as involving non-synonymous substitutions because no single-step pathway exists between the two sets of serine codons that involves only synonymous substitutions.

Discussion

Using basic probability theory, we showed that there is a substantial likelihood that even in the presence of purifying selection, there will be a number of codons in which the number of synonymous nucleotide substitutions per site (d_S) exceeds the number of non-synonymous nucleotide substitutions per site (d_N). A simple model showed that probability that such codons increase as d_N increases relative to d_S ; yet, even under very strong purifying selection, there is a substantial probability that such codons will occur. Although the model we examined made simplifying assumptions, the results provided a very good fit to data based on the codon-by-codon comparison of two closely related fungal genomes (Hughes and Friedman 2005).

In an empirical study, we examined the relationship between the number of synonymous substitutions per synonymous site (b_S) and the number of non-synonymous substitutions per non-synonymous site (b_N) along branches of the phylogenies of 69 single-copy orthologous genes from seven species of mammals belonging to three orders (primates, rodents, and carnivores). A pattern of $b_N > b_S$ was most commonly seen in the shortest branches of the tree and was associated with a high coefficient of variation in both b_N and b_S . This in turn suggests that high stochastic error in b_N and b_S on short branches, rather than positive Darwinian selection, is the explanation of most cases where b_N is greater than b_S on a given branch.

Branches with $b_N > b_S$ were much more frequently observed in the portion of the tree (branches 1–4, Fig. 1) that corresponded to the primate phylogeny. If every case of $b_N > b_S$ was taken as a case of positive selection, one might conclude on this basis that positive selection is more frequent in the primates than in carnivores or rodents. However, the shortest branches in the phylogeny were those involving primates. Human and chimpanzee are more closely related than any other pair of species analyzed, and the internal branch between the common ancestor of human, chimp, and rhesus and the common ancestor of human and chimp (branch 3, Fig. 1) was the shortest internal branch in the phylogeny. Thus, the high incidence of $b_N > b_S$ in the

primate portion of the tree appears to be an artifactual effect of short branches rather than evidence of an unusually high incidence of positive selection in primates.

ASPM and *microcephalin* are two loci at which mutations in human are known to affect brain size and at which previous studies had suggested accelerated rates of non-synonymous substitution in primates (Evans et al. 2004, 2006; Kouprina et al. 2004; Pavlicek and Jurka 2006; Wang and Su 2004). In the present analyses, no branches in the *ASPM* tree showed $b_N > b_S$. In the *microcephalin* tree, only one branch (branch 3, ancestral to human and chimp; Fig. 1) showed $b_N > b_S$, but on that branch, both b_S and b_N were rather low. Furthermore, neither gene showed a significant result in the branch-site test for positive selection.

The branch-site test for positive selection yielded results that were not concordant with the results of comparing b_S and b_N . While the pattern of $b_N > b_S$ was most often seen on short branches, the branch-site method tended to infer positive selection on branches with unusually high levels of non-synonymous divergence but low levels of synonymous divergence in comparison with other branches in a given phylogenetic tree. This pattern is exactly what we would expect if the cases of “positive selection” identified by the branch-site test in fact represent cases where of $d_N > d_S$ occurs by chance, since elevation of non-synonymous substitution relative to synonymous substitution is predicted to increase the frequency of such codons by chance alone (Fig. 2).

In addition, there were a number of anomalous aspects to the results of the branch-site tests that suggested that few, if any, of these cases involved genuine positive selection. First, a majority of the 117 codons identified as subject to positive selection were codons at which non-synonymous but not synonymous differences occurred. Since all of these codons were on branches on which at least some synonymous substitutions occurred, this finding implies that synonymous substitution was unusually low at these codons. Therefore, it is not possible to rule out the hypothesis that the high ω inferred at these codons was due in many cases to a chance reduction in synonymous substitution rather than to the elevation of non-synonymous substitution by positive selection. At the remaining codons, by contrast, the proportion of synonymous and non-synonymous substitutions were actually very similar, contrary to what is expected under positive selection. Thus, the branch-site method proved incapable of deciding between the hypothesis that positive selection is acting at these sites and the alternative hypothesis that the observed patterns are due to the random variation in ω across codons (Hughes and Friedman 2005).

An additional troubling aspect of the branch-site method was that at five of the codons at which positive selection was inferred, no amino acid change occurred; rather, at these five codons, the inferred positive selection involved the use of the different sets of serine codons. It is not well understood how the use of the different sets of serine codons arises in evolution. It is possible that in some cases, a slightly deleterious non-synonymous mutation may occur, replacing serine with some other amino acid. If this slightly deleterious mutation becomes fixed during a population bottleneck (Ohta 2002), a subsequent mutation may occur that reintroduces serine at that site but that uses a different codon set from the one used ancestrally. The latter mutation might then be fixed as a result of positive selection. In this scenario, positive selection would be involved in the substitution of a codon from one serine codon set for a codon from the other set, but positive selection would not be the only process involved. Furthermore, since no net increase in adaptation would result, this process would not amount to Darwinian evolution as usually understood.

On the other hand, it is possible that complex mutational processes (Ripley 1999) occur that allow replacement of one serine codon set by the other in a single mutational step, in which case, positive selection would not be involved at all. There is some evidence in support of the

hypothesis that mutations between the two serine codon sets occur much more frequently than would be expected if mutation proceeds in single-substitution steps. For example, a survey of 17,502 alignments of vertebrate orthologs showed that changes between the two serine codon sets occurred more frequently than any amino acid replacement at a serine codon requiring two or three mutational steps (Schneider et al. 2005). In addition, repeated arrays of serine codons often include a mixture of codons from both sets (Huntley and Golding 2004).

An additional problem with the branch-site method is that the set of codons inferred to be subject to positive selection may have included some where the alignment was uncertain. In this study, we conducted alignments by an automated process, and individual alignments were not improved by eye. Such a procedure—commonly used in studies that apply codon-based methods at a genome-wide scale—increases the danger of incorrectly inferring positive selection. Because of the focus on a few non-synonymous differences at one or more individual codons, the codon-based methods are extraordinarily sensitive to the effects of sequencing error or misalignment (Wong et al. 2008).

Consistent with previous data showing an unusually high rate of amino acid replacement in proteins having immune system functions (Murphy 1993; Hughes 1997), our results showed consistently elevated b_N and b_N/b_S in immune system genes. Neither the search for branches with $b_N > b_S$ nor the branch-site method uncovered this aspect of the data. Unlike the majority of cases with $b_N > b_S$, which occurred on short branches, the enhanced rate of non-synonymous substitution was most apparent in the longer branches of the tree. Thus, the results support a reduction of the strength of purifying selection in immune system genes in comparison to other genes.

The factors causing the relaxation of purifying selection in immune system genes remain poorly understood (Hughes et al. 2005b). Many immune system proteins are involved in cell–cell signaling, and one hypothesis might be that the relaxation of purifying selection is a characteristic of proteins involved in such signaling in general and not of immune system proteins in particular. However, in the present data, the relaxation of purifying selection was strongest in immune system proteins other than cytokines and their receptors. In the present data set, the immune system proteins other than cytokines and their receptors included a number of Toll-like receptors and leukocyte surface proteins such as CD2, CD8 α , CD8 β , and CD28 (Supplementary Table S1). Our results suggested that on average, cytokines and their receptors are subject to greater functional constraint than other immune system proteins. Such functional constraint might arise from the interaction between receptor and signal (Bagley et al. 2001).

In addition to a simple relaxation of constraint, it is possible that the accelerated rate of non-synonymous substitution in immune system proteins represents positive selection, at least in some cases. Positive selection on the genes of the vertebrate major histocompatibility complex, which encodes proteins that function in immune recognition of foreign peptides, is well documented (Hughes and Nei 1988; Hughes et al. 1994). The present data include Toll-like receptors, which are also involved in pathogen recognition (Dunne and O'Neill 2005), and thus may conceivably be subject to a co-evolutionary process with pathogens.

However, many proteins of the vertebrate immune system are not involved directly in pathogen recognition; for example, co-receptor molecules such as the CD8 $\alpha\beta$ heterodimer (Lustgarten et al. 1991), both chains of which were included in the present data set (Supplementary Table S1). Therefore, if positive selection acts on the latter genes, this selection may have some other basis. One hypothesis is that such selection arises as a result of adaptations on the part of pathogens—particularly viruses—that disrupt immune system function or that make use of immune system cell-surface proteins to gain entrance into host cells (Murphy 1993).

In summary, an accelerated rate of non-synonymous substitution in immune system genes constituted a major signal in the present data set, and positive Darwinian selection may have played a role in producing this pattern. Ironically, however, this signal was completely missed by the two most commonly used methods of testing for positive selection on branches of phylogenetic trees, namely, the comparison of b_S and b_N and the branch-site method. The inability of these methods to detect such a strong and biologically relevant signal provides further support for the conclusion that these methods focus largely on statistical artifacts and do not reveal any meaningful aspects of sequence evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grant GM43940 from the National Institutes of Health.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
- Bagley CJ, Woodcock JM, Guthridge MA, Stomski FC, Lopez AF. Structural and functional hot spots in cytokine receptors. *Int J Hematol* 2001;73:299–307. [PubMed: 11345195]
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. Genbank. *Nucleic Acids Res* 2007;35:D21–D25. [PubMed: 17202161]
- Doherty PC, Zinkernagel RM. Enhanced immunologic surveillance in mice heterozygous at the *H-2* complex. *Nature* 1975;256:50–52. [PubMed: 1079575]
- Dunn OJ. Multiple comparisons using rank sums. *Technometrics* 1964;6:241–252.
- Dunne A, O'Neill LA. Adaptor usage and Toll-like receptor signaling specificity. *FEBS Lett* 2005;579:3330–3335. [PubMed: 15876435]
- Evans PD, Anderson JR, Vallender EJ, Choi SS, Lahn BT. Reconstructing the evolutionary history of *microcephalin*, a gene controlling human brain size. *Human Mol Genet* 2004;13:1139–1145. [PubMed: 15056607]
- Evans PD, Vallender EJ, Lahn BT. Molecular evolution of the brain size regulator genes *CDK5RAP2* and *CENPJ*. *Gene* 2006;375:75–79. [PubMed: 16631324]
- Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. *Nucleic Acids Res* 2007;35:D610–D617. [PubMed: 17148474]
- Hughes AL. Rapid evolution of immunoglobulin superfamily domains expressed in immune system cells. *Mol Biol Evol* 1997;14:1–5. [PubMed: 9000748]
- Hughes AL. Adaptive evolution of genes and genomes. New York: Oxford University Press; 1999.
- Hughes AL. Searching for Darwin in all the wrong places: the misguided quest for positive selection at the molecular level. *Heredity* 2007;99:364–373. [PubMed: 17622265]
- Hughes AL, Friedman R. Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. *Mol Biol Evol* 2005;22:1320–1324. [PubMed: 15746015]
- Hughes AL, Nei M. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature* 1988;335:167–170. [PubMed: 3412472]

- Hughes AL, Hughes MK, Howell CY, Nei M. Natural selection at the class II major histocompatibility complex loci of mammals. *Philos Trans R Soc Lond B* 1994;346:359–367. [PubMed: 7708830]
- Hughes AL, Ekollu V, Friedman R, Rose JR. Gene family content-based phylogeny of prokaryotes: the effect of search criteria. *Syst Biol* 2005a;54:268–276. [PubMed: 16012097]
- Hughes AL, Packer B, Welsch R, Chanock SJ, Yeager M. High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. *Immunogenetics* 2005b; 57:821–827. [PubMed: 16261383]
- Hughes AL, Friedman R, Glenn NL. The future of data analysis in evolutionary genomics. *Curr Genomics* 2006;7:227–234.
- Huntley MA, Golding GB. Neurological proteins are not enriched for repetitive sequences. *Genetics* 2004;166:1141–1154. [PubMed: 15082536]
- Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;8:275–282. [PubMed: 1633570]
- Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 1977;267:275–276. [PubMed: 865622]
- Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, Yoon Y-H, Collura R, Ruvolo M, Barrett JC, Woods CG, Walsh CA, Jurka J, Larionov V. Accelerated evolution of the *ASPM* gene controlling brain size begins prior to human brain expansion. *PloS Biol* 2004;2:0653–0663.
- Lustgarten J, Waks T, Eshhar Z. CD4 and CD8 accessory molecules function through interactions with major histocompatibility complex molecules which are not directly associated with the T cell receptor–antigen complex. *Eur J Immunol* 1991;21:2507–2515. [PubMed: 1915555]
- Murphy PM. Molecular mimicry and the generation of host defense protein diversity. *Cell* 1993;72:823–826. [PubMed: 8458078]
- Nei, M. *Molecular evolutionary genetics*. New York: Columbia University Press; 1987.
- Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986;3:418–426. [PubMed: 3444411]
- O'Connor DH, McDermott AB, Krebs AC, Dodds EJ, Miller JE, Gonzalez EJ, Jacoby TJ, Yant L, Piontkivska H, Pantophlet R, Burton DR, Rehauer WR, Wilson N, Hughes AL, Watkins DI. A dominant role for CD8+T-lymphocyte selection in simian immunodeficiency virus sequence variation. *J Virol* 2004;78:14012–14022. [PubMed: 15564508]
- Ohta T. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A* 2002;99:16134–16137. [PubMed: 12461171]
- Pavlicek A, Jurka J. Positive selection on the nonhomologous end-joining factor Cernunnos-XLF in the human lineage. *Biol Direct* 2006;15.
- Ripley LS. Predictability of mutant sequences: relationships between mutational mechanisms and mutant specificity. *Ann NY Acad Sci* 1999;870:159–170. [PubMed: 10415481]
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–425. [PubMed: 3447015]
- Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 2002;18:502–504. [PubMed: 11934758]
- Schneider A, Cannarozzi G, Gonnet GH. Empirical codon substitution matrix. *BMC Bioinformatics* 2005;6:134. [PubMed: 15927081]
- Swofford, DL. *PAUP*: Phylogenetic analysis using parsimony (*and other methods)*. Sunderland MA: Sinauer; 1999.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
- Wang Y, Su B. Molecular evolution of *microcephalin*, a gene determining human brain size. *Human Mol Genet* 2004;13:1131–1137. [PubMed: 15056608]
- Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science* 2008;319:473–476. [PubMed: 18218900]

- Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997;13:555–556. [PubMed: 9367129]
- Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 1998;15:568–573. [PubMed: 9580986]
- Yang Z, Wong WS, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 2005;22:1107–1118. [PubMed: 15689528]
- Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* 1998;98:3708–3713. [PubMed: 9520431]
- Zhang J, Nielsen R, Yang Z. Evaluation of an improved branchsite likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 2005;22:2472–2479. [PubMed: 16107592]

Fig. 1.
NJ tree of BRCA2 amino acid sequences illustrating the numbering scheme for branches used in this paper

Fig. 2. Probability of a codon with no synonymous differences and one or more non-synonymous differences, as a function of p_S . Curves are plotted for three different p_N/p_S ratios: 0.1, 0.2, and 0.5. A Poisson distribution is assumed and the expected numbers of synonymous and non-synonymous sites per codon (0.745 and 2.255, respectively) are based on the data of Hughes and Friedman (2005)

Fig. 3.
a Numbers of individual branches (in trees for individual genes) with $b_N > b_S$, illustrated for each branch of the tree. **b** Percent of branches with $b_N > b_S$ plotted against mean b_S ($r_S = -0.817$; $P = 0.002$); the points corresponding to each branch of the tree are numbered as in Fig. 1. **c** Plot of coefficients of variation (CV) in b_N (solid diamonds; $r_S = -0.773$; $P = 0.005$) and b_S (open circles; $r_S = -0.745$; $P = 0.008$) plotted against mean b_S for each branch of the tree

Fig. 4.

Median values of **a** b_S ; **b** b_N ; and **c** b_N/b_S for individual branches categorized according to the function presence ($Im+$) or absence ($Im-$) of immune system function and the presence ($Sig+$) or absence ($Sig-$) of signaling function. In each case, there was a significance difference among the four categories (Kruskal–Wallis test; $P < 0.001$). Individual comparisons (Dunn 1964) between the $Im+/Sig-$ category and other categories: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. Numbers of individual branches in each category are indicated in **a** and **c**

Fig. 5. Plot of the ratio of mean b_N in immune system genes to mean b_N in other genes vs. mean b_S ($r_S=0.745$; $P=0.008$); the points corresponding to each branch of the tree is numbered as in Fig. 1

Fig. 6.

Histograms of p_S (cross-hatched bars) and p_N (solid bars) between 117 codons identified as positively selected by the branch-site method and their immediate ancestors (reconstructed by maximum parsimony)

Table 1

Numbers of synonymous substitutions per synonymous site (b_S) and numbers of non-synonymous substitutions per non-synonymous site (b_N) in 69 genes along each branch of a mammalian phylogeny

Branch		b_S	b_N		
Number ^a	Ancestral to	Mean±SE	Median	Mean±SE	Median
1	Human	0.0054±0.0007	0.0046	0.0017±0.0004	0.0007
2	Chimp	0.0075±0.0009	0.0069	0.0025±0.0005	0.0017
3	Human/Chimp	0.0273±0.0028	0.0235	0.0097±0.0014	0.0070
4	Rhesus	0.0447±0.0041	0.0395	0.0139±0.0014	0.0122
5	Primates	0.1375±0.0096	0.1210	0.0452±0.0034	0.0405
6	Cat	0.1131±0.0080	0.1190	0.0413±0.0041	0.0298
7	Dog	0.1456±0.0088	0.1331	0.0446±0.0040	0.0322
8	Carnivores	0.1123±0.0064	0.1074	0.0389±0.0031	0.0343
9	Mouse	0.1185±0.0079	0.1025	0.0362±0.0041	0.0277
10	Rat	0.1085±0.0059	0.1055	0.0304±0.0026	0.0241
11	Rodents	0.4180±0.0178	0.3735	0.1018±0.0074	0.0902

^aBranches are numbered as in Fig. 1.