Frequent and Widespread Parallel Evolution of Protein Sequences

Antonis Rokas and Sean B. Carroll†*

*Department of Biological Sciences, Vanderbilt University; and †Howard Hughes Medical Institute, R. M. Bock Laboratories, University of Wisconsin–Madison

Understanding the patterns and causes of protein sequence evolution is a major challenge in evolutionary biology. One of the critical unresolved issues is the relative contribution of selection and genetic drift to the fixation of amino acid sequence differences between species. Molecular homoplasy, the independent evolution of the same amino acids at orthologous sites in different taxa, is one potential signature of selection; however, relatively little is known about its prevalence in eukaryotic proteomes. To quantify the extent and type of homoplasy among evolving proteins, we used phylogenetic methodology to analyze 8 genome-scale data matrices from clades of different evolutionary depths that span the eukaryotic tree of life. We found that the frequency of homoplastic amino acid substitutions in eukaryotic proteins was more than 2-fold higher than expected under neutral models of protein evolution. The overwhelming majority of homoplastic substitutions were parallelisms that involved the most frequently exchanged amino acids with similar physicochemical properties and that could be reached by a single-mutational step. We conclude that the role of homoplasy in shaping the protein record is much larger than generally assumed, and we suggest that its high frequency can be explained by both weak positive selection for certain substitutions and purifying selection that constrains substitutions to a small number of functionally equivalent amino acids.

Introduction

Since the proposal of the neutral theory of molecular evolution (Kimura 1968; King and Jukes 1969), one of the central issues in molecular evolutionary biology has been the degree to which the divergence of protein sequences over time is governed by selection for advantageous mutations versus genetic drift. Many approaches have been taken to assess the fraction of amino acid substitutions that have been fixed by positive selection (Eyre-Walker 2006). Analyses of polymorphism and divergence in *Drosophila* proteins have led to the conclusions that approximately half (Smith and Eyre-Walker 2002) to perhaps more than 90% (Sawyer et al. 2003) of all amino acid replacements have been fixed by positive selection. However, these results include wide confidence intervals, and much lower figures have been obtained for other taxa such as humans (The Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005).

Furthermore, inferences of widespread positive selection have raised the issue of the relative magnitude of the selective effects of most individual fixed substitutions (Nei 2005). Ample biochemical evidence (reviewed in DePristo et al. [2005]) and population genetic data (Drake 2006) indicate that most substitutions at most sites are deleterious. For the remaining minority of sites where substitutions are nondeleterious, typical estimates suggest that the strength of selection fixing most substitutions is very weak, such that $1 < N_e s < 10$ (Eyre-Walker 2006; Sawyer et al. 2007), where N_e is the effective population size and s the selection coefficient, and could be regarded as nearly neutral. Therefore, the physicochemical significance of most amino acid substitutions and the role of positive selection in their fixation remain unclear.

One alternative potential signature of positive selection is the independent evolution of identical molecular character states (nucleotide or amino acid residues) in different branches of a phylogenetic tree that are not directly inherited from a common ancestor (Zhang and Kumar 1997; Zhang 2006; Jost et al. 2008). Such homoplastic character states may be derived from different ancestral states (convergent changes), from the same ancestral character state (parallel changes), or via the reversal of a derived character state to the ancestral one (reverse or back changes) (Li 1997; Page and Holmes 1998).

Homoplasy may arise due to the action of positive or balancing selection (Wells 1996). Theoretical studies have suggested that the probability of parallel evolution under natural selection is nearly twice as large as that under neutrality (Orr 2005). A variety of experimental studies have provided evidence that natural selection on the proteome can be manifested as homoplasy. For example, several experimental evolution studies examining the adaptations of viral populations on bacterial or eukaryotic hosts have uncovered striking examples of parallelism (Wichman et al. 2000; Hughes et al. 2001; Pinel-Galzi et al. 2007), convergence (Bull et al. 1997; Fares et al. 2001), and reversal (Crill et al. 2000; Depristo et al. 2007). One study which examined the adaptation of 2 closely related phages to laboratory culture conditions found that a remarkable 62% of all substitutions were parallel (Wichman et al. 2000). Great numbers of parallel and convergent homoplastic substitutions have also been uncovered in genes belonging to the major histocompatibility complex (Yeager and Hughes 1999).

Alternatively, molecular homoplasy may occur simply by chance, through the action of neutral evolutionary processes (Zhang and Kumar 1997). Because sequence evolution is a stochastic process and each site has a finite number of possible states (4 for nucleotide and 20 for amino acid residues), it is "expected" that independent evolutionary lineages will occasionally acquire the same character states independently (Zhang and Kumar 1997). Given time, homoplasy (and divergence) is expected to increase up to some level of saturation, which is determined by a variety of factors such as mutational bias (Smith JM and Smith NH 1996; Baer et al. 2007), and rates of evolution (Felsenstein 1978; Pupko and Galtier 2002).

There is, however, yet another potential cause of homoplasy that may be underappreciated, the action of purifying selection. In nucleic acids or proteins, the number of

Key words: homoplasy, positive selection, selective constraint, protein, independent evolution.

E-mail: sbcarrol@wisc.edu.

Mol. Biol. Evol. 25(9):1943–1953. 2008
doi:10.1093/molbev/msn143
Advance Access publication June 25, 2008

possible states may be frequently smaller than 4 or 20 and the frequency of homoplasy higher because certain sites may be constrained by purifying selection such that only a fraction of all possible residues are allowed, such as only those amino acids sharing the same physicochemical properties (Naylor et al. 1995). For example, it has been argued that the hydrophobic nature of most mitochondrial proteins has effectively constrained the character state space of their second codon positions to 1 of 2 states, C or T (Naylor et al. 1995). Homoplasy, then, can also be the result of substitution constraints imposed by purifying selection.

Beyond particular case studies and the expectation of homoplasy due to random chance, the actual levels of homoplasy among evolving proteins have not been well characterized. For example, in an experiment, very similar in design to the phage studies cited above, the evolution of 12 initially identical *Escherichia coli* populations for 20,000 generations under the same selective pressure produced very low levels of molecular homoplasy (Woods et al. 2006). Thus, the results of case studies may not be the best basis for predicting and assessing the levels of molecular homoplasy expected in natural populations, for reasons including but not limited to the near identity of selective environments used in experimental studies, the strength and continuity of selection (Bull et al. 1997; Hughes 2007), and the genome size and complexity of the organisms studied (Woods et al. 2006).

Although molecular homoplasy has long been appreciated in evolution studies and much effort has been invested into understanding its causes and providing corrections for them (Felsenstein 2003), relatively few studies have utilized homoplasy as a source of evidence about evolutionary hypotheses. For example, examination of patterns of amino acid variation at the *Gpdh* locus across *Drosophila* species identified 4 sites exhibiting such high levels of homoplasy that they accounted for approximately half of the substitutions "observed" (Wells 1996). We too noted surprisingly high levels of homoplasy in an analysis of phylogenetic bushes (Rokas and Carroll 2006), and studies examining noncoding and coding DNA have similarly revealed high levels of homoplasy (O'hUigin et al. 2002; Bazykin et al. 2007). For example, Bazykin et al. (2007) found an elevated rate of parallel nonsynonymous substitutions in the genomes of mammals, *Drosophila*, and yeasts. Importantly, the underlying causes of this excess homoplasy have been attributed to several different factors including differences in mutation rates (O'hUigin et al. 2002), the action of purifying selection (Wells 1996; Rokas and Carroll 2006), or the action of weak positive selection (Rokas and Carroll 2006; Bazykin et al. 2007).

In this study, we revisit the questions of the prevalence and underlying causes of homoplasy using a different methodological framework. Specifically, we employed a phylogenetic approach to conduct a systematic survey of the occurrence of homoplasy across 8 clades of the tree of life. Given evolutionary trees for these 8 clades, we measured the extent of observed homoplasy on each clade and compared the observed values with the homoplasy expected based on simulation analyses of the same trees. We also measured the types of amino acid substitutions that generate homoplasy, using an index (evolutionary index [EI], Tang et al. 2004) that captures the evolutionary trends of amino acid exchangeability. We found that across these 8 clades protein sequences underwent more than twice as many homoplastic substitutions than was expected by neutral processes alone. The overwhelming majority of homoplastic amino acid substitutions were between amino acids with similar physicochemical properties. We suggest that these results are likely to be the evolutionary product of 2 different types of selection: weak positive selection for certain substitutions and purifying selection that constrains substitutions to a small number of functionally equivalent alternatives.

Materials and Methods

Data Matrix Generation

Data matrices from 8 representative clades of the eukaryotic tree of life were used to evaluate the observed and expected levels of homoplasy. Information about the clades and taxa is shown in table 1. All data matrices contained sequences from 4 taxa. The mammalian data matrix contained mitochondrial genes, whereas the other 7 data matrices contained nuclear genes. Two data matrices (*Saccharomyces* yeasts and land plants) were obtained from previously published studies (Rokas et al. 2003; Sanderson et al. 2003). The orthologous genes for the *Drosophila* data matrix were obtained from Pollard et al. (2006). Genes from the other 5 data matrices were retrieved from GenBank and genome databases. Orthologous genes were identified by using the reciprocal best Blast hit criterion (Koonin 2005), and sequences were aligned using ClustalW (Thompson et al. 1994). Areas of ambiguous alignment were removed using the Gblocks software (Castresana 2000) with the default settings. Specifically, all sites containing gaps were discarded and only blocks of alignment longer than 10 sites and that fit the following criteria of sequence conservation were retained: the "minimum number of sequences for a conserved position" and "minimum number of sequences for a flank position" parameters were set to larger than half the number of sequences in the alignment, and the "maximum number of contiguous nonconserved positions" parameter was set to 8.

Homoplasy Estimation Methodology

We estimated observed and expected homoplasy for all the data matrices using a modified version of a previously published methodology (Takezaki et al. 2004) (fig. 1). Briefly, for a set of sequences from 4 taxa, 15 different amino acid configuration patterns were possible at each site: AAAA, ABBB, BABB, BBAB, BBBA, AABB, ABAB, ABBA, AABC, ABAC, ABCA, BAAC, BACA, BCAA, and ABCD, where A, B, C, and D correspond to different amino acid residues. By counting the number of sites that conform to each pattern and multiplying them by the minimum number of substitutions required to generate that pattern, we estimated the number of observed amino acid substitutions (fig. 1*B*). For example, 34 sites exhibited

Table 1
Taxonomic Composition of Data Matrices Used in This Study

Clade	Included Taxa	Gene Number	Data Source
Saccharomyces yeasts	*Saccharomyces cerevisiae*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, and *Saccharomyces castellii*	106	(Rokas et al. 2003)
Aspergillus filamentous ascomycetes	*Aspergillus oryzae*, *Aspergillus terreus*, *Aspergillus clavatus*, and *Aspergillus fumigatus*	200	GenBank
Fungal phyla	*Neurospora crassa*, *Coprinus cinereus*, *Rhizopus oryzae*, and *Batrachochytrium dendrobatidis*	200	GenBank
Eukaryotic phyla	*Monosiga brevicollis*, *N. crassa*, *Arabidopsis thaliana*, and *Dictyostelium discoideum*	239	GenBank
Land plants	*Lotus japonicus*, *A. thaliana*, *Zea mays*, and *Oryza sativa*	39	(Sanderson et al. 2003)
Drosophila fruit flies	*Drosophila melanogaster*, *Drosophila erecta*, *Drosophila ananassae*, and *Drosophila pseudoobscura*	200	(Pollard et al. 2006)
Cetartiodactyl mammals	*Balaenoptera musculus*, *Hippopotamus amphibius*, *Bos taurus*, and *Sus scrofa*	12	GenBank
Metazoan phyla	*Homo sapiens*, *Ciona intestinalis*, *D. melanogaster*, and *Caenorhabditis elegans*	239	GenBank
Paenungulata mammals[a]	*Loxodonta africana*, *Dugong dugon*, *Procavia capensis*, and *Orycteropus afer*	12	GenBank
Vertebrates[a]	*H. sapiens*, *Latimeria chalumnae*, *Protopterus* sp., and *Danio rerio*	44	(Takezaki et al. 2004)

[a] Studies where phylogeny is unknown.

the ABCD pattern in the mammalian data matrix. Given that a minimum of 3 substitutions are required to generate the ABCD pattern, we expect a minimum of $34 \times 3 = 102$ observed substitutions to have occurred in these 34 sites (fig. 1*C*). Under parsimony, only the patterns AABB, ABAB, and ABBA contain phylogenetic information (fig. 1*B*), with the pattern supported by the largest number of sites being the most parsimonious solution to the 4-taxon phylogeny problem (fig. 1*B* and *C*).

For example, examination of the mammalian data matrix showed that pattern AABB was displayed by the largest number of amino acid sites (49) thus providing support for a grouping of whales with hippopotamuses (fig. 1*B* and *C*), a clade whose existence has been independently corroborated by multiple lines of phylogenetic evidence (Nikaido et al. 1999; Gatesy and O'Leary 2001). Importantly, a large number of sites displayed the other 2 patterns, ABAB (27 sites) and ABBA (40 sites), respectively. In the absence of lineage sorting of ancestral polymorphisms, an unlikely explanation because the length of the internode in question is thought to be greater than the 2–3 Myr required for the resolution of typical ancestral polymorphisms (Takahata 1993), these 2 alternative patterns can only be explained by homoplasy (fig. 1*B* and *C*). To reconcile the homoplastic ABBA and ABAB patterns with the mammalian phylogeny, a minimum of 2 substitutions was required; thus, the observed number of homoplastic substitutions is $(27 + 40) \times 2 = 134$ (fig. 1*B* and *C*). Dividing the number of observed homoplastic substitutions (134) by the observed number of total substitutions (1,087) gives us the percentage of observed homoplastic substitutions in the data matrix (12.3%).

To calculate the homoplasy expected under neutral conditions, the best-fit model of amino acid evolution for each of the 8 data matrices was selected using ProtTest (Abascal et al. 2005). Given the amino acid sequence alignment for each data matrix, model selection was performed by evaluating the fit of 12 different alternative models of amino acid evolution according to the Akaike information criterion (Abascal et al. 2005). Models of amino acid evolution are typically derived by counting observed amino acid substitutions in large sequence databases, making allowance for multiple substitutions and the phylogeny of the species used (Whelan et al. 2001). Because the history of the proteins in the sequence databases used to construct models have been shaped by the effects of both selection and mutational biases (Thorne 2007), the simulation conditions may not be strictly neutral. Thus, levels of expected homoplasy may actually be overestimated with our approach.

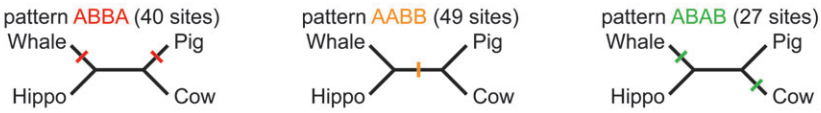
Using the parameters obtained from ProtTest, the maximum likelihood phylogeny was generated using Phyml (Guindon and Gascuel 2003). Importantly, the phylogenetic relationships for all data matrices used in this study are unambiguous and our analyses confirm the results of several previous studies. To test whether accurate knowledge is a prerequisite for our analyses, we also examined 2 data matrices in which the true topology is ambiguous, 1 from vertebrates (Takezaki et al. 2004) and 1 from Paenungulata mammals (Nishihara et al. 2005), selecting as the "correct" topology the one that minimized the number of inferred homoplastic substitutions.

To calculate the expected average values for the 15 amino acid patterns under simulation, we generated 100 data sets of equal size as the original set using the amino acid evolution parameters and the maximum likelihood tree as inputs into the simulation software Evolver (part of the PAML software package, Yang 1997). By calculating the number of expected homoplastic substitutions and dividing it by the number of expected total substitutions, we estimated the expected homoplasy for all the data matrices under study. For example, the expected homoplasy calculated from simulation analysis of the mammalian data matrix was 5% (fig. 1*C* and *D*). Thus, the fold difference between the observed homoplasy (12.3%) and that expected from simulation analysis (5%) was 2.5 (12.3/5.0%).

A

```
          1        2        3        4        5        6        7        8
1........0.........0.........0.........0.........0.........0.........0.........0...
Whale TGFRNKTKMSLAHLLPQGTPTFLIPMLVIIETISLFIQPVALAVRLTANITAGHLLMHLIGETTLVLMSTSLFTAIITFT...
Hippo MGFRNKTKISLAHFLPQGTPTPLIPMLVIIETISLFIQPMALAVRLTANITAGHLLMHLIGGATLALMNISMTTALITFI...
Cow   TGFRNKTKASLAHFLPQGTPTPLIPMLVIIETISLFIQPMALAVRLTANITAGHLLIHLIGGATLALMSISTTTALITFT...
Pig   TGFRYKTKTSLAHFLPQGTPALLIPMLVIIETISLFIQPVALAVRLTANITAGHLLIHLIGGATLALLNINTMTAFITFT...
```

B



C

Pattern	Observed sites	Observed substitutions (observed sites x substitutions)	Expected sites	Expected substitutions (expected sites x substitutions)
ABBB, BABB, BBAB, BBBA	460	460 x 1 = 460	456	456 x 1 = 456
AABB	49	49 x 1 = 49	37	37 x 1 = 37
ABAB	27	27 x 2 = 54	12	12 x 2 = 24
ABBA	40	40 x 2 = 80	13.5	13.5 x 2 = 27
AABC, ABAC, ABCA, BAAC, BACA, BCAA	171	171 x 2 = 342	185.5	185.5 x 2= 371
ABCD	34	34 x 3 = 102	40	40 x 3 = 120

D

$$\textit{Observed Homoplasy} = \frac{\text{ABAB+ ABBA observed substitutions}}{\text{Total number of observed substitutions}} = \frac{(54+80)}{(1087)} = 12.3\%$$

$$\textit{Expected Homoplasy} = \frac{\text{ABAB+ABBA expected substitutions}}{\text{Total number of expected substitutions}} = \frac{(24+27)}{(1035)} = 5.0\%$$

$$\textit{Excess Homoplasy} = \textit{Observed Homoplasy} - \textit{Expected Homoplasy} = 12.3\% - 5.0\% = 7.3\%$$

or 2.5-fold difference

FIG. 1.—Estimation of excess homoplasy in Cetartiodactyl mitochondrial proteomes. (*A*) Part of the alignment of amino acid sequences from the mitochondrial genes of whale, hippopotamus, cow, and pig. Parsimony-informative sites are in color (sites with the ABBA pattern are shown in red, sites with the AABB pattern are shown in orange, whereas sites with the ABAB pattern are shown in green). (*B*) The phylogenetic relationship among the 4 mammalian taxa and most parsimonious scenarios for the generation of patterns ABBA, AABB, and ABAB. Note that patterns ABBA and ABAB must have been generated by at least 2 independent, homoplastic substitutions. (*C*) Distribution of amino acid patterns and substitutions for the Cetartiodactyl mitochondrial genome alignment (observed data) as well as those expected from 100 simulation analyses using the amino acid evolution parameters inferred from the observed data. The patterns resulting from homoplastic substitutions (ABBA and ABAB) are shown in blue. (*D*) Calculation of observed, expected, and excess homoplasy values.

Assessing the Relative Contribution of Convergence, Parallelism, and Reversal to Homoplasy

Parallelisms, reversals, and convergences cannot be identified in 4-taxon data matrices. Thus, to assess the relative contribution of parallel, convergent, and reverse substitutions to homoplasy, we expanded 3 of the 10 four-taxon data matrices used in this study (Cetartiodactyl mammals, *Saccharomyces* yeasts, and *Aspergillus* filamentous ascomycetes) for which the species phylogenetic relationships are well supported (Nikaido et al. 1999; Rokas et al. 2003; Rokas and Galagan 2008) by adding sequence data from several additional species. Using the parsimony criterion, we counted all substitutions occurring in all parsimony-informative sites that generated homoplasy via parallelisms, convergences, and reversals in these enlarged data matrices. All substitutions for which their state in the most immediate ancestor was ambiguous were ignored. In the rare cases where the same substitution contributed to 2 homoplastic types (e.g., a reversal and a parallelism), the substitution was counted as an equal split between the 2 types of events.

Any potential differences in the relative contribution of parallel, convergent, and reverse substitutions to homoplasy across the 3 clades could be the consequence of differences in the shape of each topology or, alternatively, due to bias of the excess homoplastic events toward specific types of homoplasy. To discriminate between the 2 alternatives, we conducted simulation analyses on the 3 expanded data matrices. We counted all substitutions occurring in the first 1,000 parsimony-informative sites from each simulated data matrix that generated homoplasy via parallelisms, convergences, and reversals. For both simulation calculations as well as for the counting of types of homoplasy, we used the same methodologies as above.

Classification of Homoplastic Amino Acid Substitutions according to the EI

Of the 190 possible interchanges among the 20 amino acids, only 75 can be attained via a single-nucleotide substitution. The remaining 115 substitutions require 2 or 3

Table 2
Total, Variable, and Parsimony-Informative Numbers of Sites

Clade	Total Number of Sites	Variable Sites	Parsimony-Informative Sites
Saccharomyces yeasts	42,342	10,318 (24%)	687 (1.6%)
Aspergillus filamentous ascomycetes	99,204	22,785 (23%)	4,127 (4.2%)
Fungal phyla	86,658	51,966 (60%)	5,048 (5.8%)
Eukaryotic phyla	68,407	42,558 (62%)	4,397 (6.4%)
Land plants	8,173	1,165 (14%)	448 (0.5%)
Drosophila fruit flies	68,121	17,237 (25%)	2,649 (3.9%)
Cetartiodactyl mammals	3,574	781 (22%)	116 (3.2%)
Metazoan phyla	68,407	34,510 (50%)	3,700 (5.4%)
Paenungulata mammals[a]	3,568	1,144 (32%)	127 (3.6%)
Vertebrates[a]	10,404	3,694 (36%)	294 (2.8%)
Averages	**45,886**	**18,616 (41%)**	**2,159 (4.7%)**

[a] Studies where phylogeny is unknown.

single-nucleotide mutational steps. Among these 190 amino acid interchanges, some can be achieved much more easily than others, the ease of substitution being largely dependent on the mutational distance (determined by the genetic code and mutational biases) and the physicochemical distance between amino acids (Yang et al. 1998; Tang et al. 2004). To better understand which amino acid substitutions most commonly contribute to homoplasy, we employed the EI devised by Tang et al. (2004) to classify all parsimony-informative patterns (AABB, ABAB, and ABBA) into 4 categories: 1) sites exhibiting the 12 most frequent single-mutational step amino acid substitutions (top12), 2) sites exhibiting the middle 51 most frequent single-mutational step substitutions (middle51), 3) sites exhibiting the 12 least frequent single-mutational step substitutions (bottom12), and 4) sites exhibiting the 115 amino acid substitutions requiring 2 or 3 mutational steps (multiple115). The EI was chosen because it has been shown to perform better than other measures such as PAM and Grantham's distance, and because its predictions hold well across genes and organisms (Tang et al. 2004).

Results

Data Matrices

The 8 data matrices that we assembled to measure the extent of homoplasy in molecular data sets included a wide range of gene numbers and amino acid sites (table 2). The smallest set, the mitochondrial mammalian data matrix, was composed of just 12 genes and slightly more than 3,500 amino acid sites, whereas the largest sets were the metazoan and eukaryotic phyla data matrices containing 239 genes each and the 200-gene *Aspergillus* data matrix containing 99,204 amino acid sites (table 2). The percentages of variable and parsimony-informative sites also varied, ranging from 14% to 62% for variable sites and from 0.5% to 6.4% for parsimony-informative sites (table 2). The maximum likelihood phylogenetic tree for each clade is shown in the supplementary figure S1 (Supplementary Material online) and is in agreement with the published literature (Rokas et al. 2003; Sanderson et al. 2003; Cracraft and Donoghue 2004; Takezaki et al. 2004; Nishihara et al. 2005; Pollard et al. 2006; Rokas and Galagan 2008).

Levels of Observed and Excess Homoplasy

The observed, expected, and excess homoplasy values for the 8 data matrices are shown in table 3. Values of observed homoplasy ranged from 2.3% (for the land plant data matrix) to 12.3% (for the Cetartiodactyl mammalian data matrix), whereas the levels of expected homoplasy ranged from 0.9% to 5.0% for the same data matrices resulting in an excess homoplasy of 1.3–7.4% (table 3). Thus, the observed homoplasy among all data sets was consistently 1.9- to 3.2-fold greater than expected from the simulation analyses (table 3). Similar fold differences in homoplasy were observed in the data sets from clades whose phylogeny is ambiguous (Paenungulata mammals and vertebrates) (table 3).

Importantly, the excess homoplasy was not the resulter of a larger number of substitutions but the result of a specific increase of homoplastic substitutions. Comparison of the parsimony-informative substitutions that support the correct topology (i.e., examination of just the AABB sites in fig. 1) from both observed and expected sites across all data matrices reveals that the number of parsimony-informative substitutions

Table 3
Observed, Expected, and Excess Levels of Homoplasy (homoplastic substitutions/total substitutions) under Parsimony

Clade	Observed Homoplasy	Expected Homoplasy[a]	Fold Difference	Excess Homoplasy (%)
Saccharomyces yeasts	4.7% (572/12,246)	2.0% (230/11,666)	2.4	2.7
Aspergillus filamentous ascomycetes	6.9% (2,082/30,089)	2.5% (719/29,012)	2.8	4.4
Fungal phyla	6.7% (5,804/86,825)	3.5% (2,908/83,820)	1.9	3.2
Eukaryotic phyla	7.5% (5,422/72,322)	3.5% (2,415/68,514)	2.1	4.0
Land plants	2.3% (32/1,365)	0.9% (12/1,363)	2.7	1.4
Drosophila fruit flies	2.6% (592/23,050)	1.3% (288/22,273)	2.0	1.3
Cetartiodactyl mammals	12.3% (134/1,087)	5.0% (51/1,035)	2.5	7.4
Metazoan phyla	7.4% (4,040/54,493)	3.3% (1,712/52,035)	2.3	4.1
Paenungulata mammals[b]	10.4% (178/1,705)	4.9% (82/1,656)	2.1	5.5
Vertebrates[b]	10.2% (376/3,694)	3.2% (113/3,543)	3.2	7.0
Averages	**7.1%**	**3.0%**	**2.4**	**4.1**

[a] Expected homoplasy values were obtained by simulation analysis.
[b] Studies where phylogeny is unknown.

Table 4
Observed, Expected, and Excess Levels of Parsimony-Informative Substitutions (PISs) (PISs/total substitutions) That Support the Correct Topology

Clade	Observed PISs	Expected PISs[a]	Excess PISs (%)
Saccharomyces yeasts	3.3% (401/12,246)	3.3% (382/11,666)	0.0
Aspergillus filamentous ascomycetes	10.3% (3,086/30,089)	9.8% (2,842/29,012)	0.5
Fungal phyla	2.5% (2,146/86,825)	2.2% (1,851/83,820)	0.3
Eukaryotic phyla	2.3% (1,686/72,322)	2.1% (1,419/68,514)	0.3
Land plants	31.6% (432/1,365)	31.7% (432/1,363)	0.0
Drosophila fruit flies	10.2% (2,353/23,050)	10.1% (2,243/22,273)	0.1
Cetartiodactyl mammals	4.5% (49/1,087)	3.6% (37/1,035)	0.9
Metazoan phyla	3.1% (1,680/54,493)	2.8% (1,434/52,035)	0.3
Paenungulata mammals[b]	5.4% (92/1,705)	2.4% (40/1,656)	3.0
Vertebrates[b]	2.9% (106/36,94)	2.4% (85/3,543)	0.5
Averages	**7.6%**	**7.0%**	**0.6**

[a] Expected PISs values were obtained by simulation analysis.
[b] Studies where phylogeny is unknown.

in both cases are very similar (table 4). Similarly, comparison of the observed and expected sites for the remaining 12 nonparsimony-informative patterns revealed very small disagreements (data not shown). Only in the case of the Paenungulata, data matrix was the excess of substitutions very large, whereas in the case of the land plant data matrix, the expected number of parsimony-informative substitutions was actually slightly larger than the observed number (table 4). Examination of data sets from clades whose phylogeny is ambiguous revealed similar results (tables 1–5).

Lack of Correlation between Homoplasy and the Amount of Evolution

It is widely held that levels of homoplasy are positively correlated with the total amount of evolution (Kallersjo et al. 1999; Rogozin et al. 2008). In the case of evolutionary trees, the amount of evolution can be approximated by the tree's evolutionary distance (the sum of branch lengths), which is the product of the evolutionary time elapsed and the rates of substitution across the genes examined. Plotting the ratio of observed over excess homoplasy as a function of total tree length did not reveal a significant positive correlation (fig. 2*A*). Although total tree length may not be a good proxy, several phylogenetic studies have shown that a critical parameter in the generation of homoplasy is the ratio of the stem's length over that of the external branches (Felsenstein 1978; Gaut and Lewis 1995). However, plotting the ratio of observed/expected homoplasy as a function of the stem/external branch length ratio also did not reveal a significant correlation (fig. 2*B*). Similarly, excess homoplasy was not correlated with the stem/external branch length ratio (fig. 2*C*), whereas observed homoplasy showed a very weak negative correlation with the stem/external branch length ratio (fig. 2*D*), which disappeared after the removal of a single outlier datum (data not shown). Therefore, the excess homoplasy we observed was independent from the amount of evolution or from the stem/external branch length ratio.

The Relative Contribution of Convergence, Parallelism, and Reversal to Homoplasy

The large number of homoplastic sites in these 4-taxon data sets presented the opportunity to assess which types of mutational events contributed to homoplasy. Because it was impossible to distinguish parallelisms from reversals and convergences in 4-taxon trees, we expanded 3 of the 10 data matrices by adding taxa. Previous experimental and simulation studies have indicated that parallelisms and reversals are much more frequent than convergences (Wells 1996; Zhang and Kumar 1997). Examination of approximately 5,000 parsimony-informative sites across these 3 expanded and well-supported phylogenies revealed that the overwhelming majority (91%) of homoplastic substitutions were generated via parallel substitutions (fig. 3), with the remaining 9% generated by reversals. The contribution of convergent substitutions to homoplasy in these 3 data matrices was extremely small (10 out of a total of 2,316 events).

We noted that the relative contribution of each type of homoplastic event varied widely across the 3 data sets. For example, whereas 41% of homoplastic substitutions in the *Saccharomyces* yeasts were due to reversals, only 1% were due to reversals in *Aspergillus* filamentous ascomycetes (fig. 3). Examination of the fraction of parallelisms, reversals, and convergences observed with those expected from the simulation analyses for each of the 3 data matrices did not reveal any major differences, with the possible exception of the smaller than expected fraction of reversals and the larger than expected faction of parallelisms in the *Aspergillus* data matrix (fig. 3). Similarly, the expected fractions of homoplastic substitutions classified to different classes according to the EI (Tang et al. 2004) were very similar to the observed fractions (fig. 3). These results suggest that the relative contributions of each type of event to the overall levels of homoplasy largely depended on the shape of the topology and branch lengths of the clade studied.

Excess Homoplasy Is Largely due to Substitutions between Frequently Exchangeable Amino Acids

To determine if certain amino acid substitutions contribute more often to homoplasy, we first classified all sites that exhibited parsimony-informative patterns into 1 of 4 substitution groups according to the EI (Tang et al. 2004) (table 5). We found that, on average, 43.4% of parsimony-informative sites exhibited substitutions belonging to the frequently exchangeable top12 category and 76% of

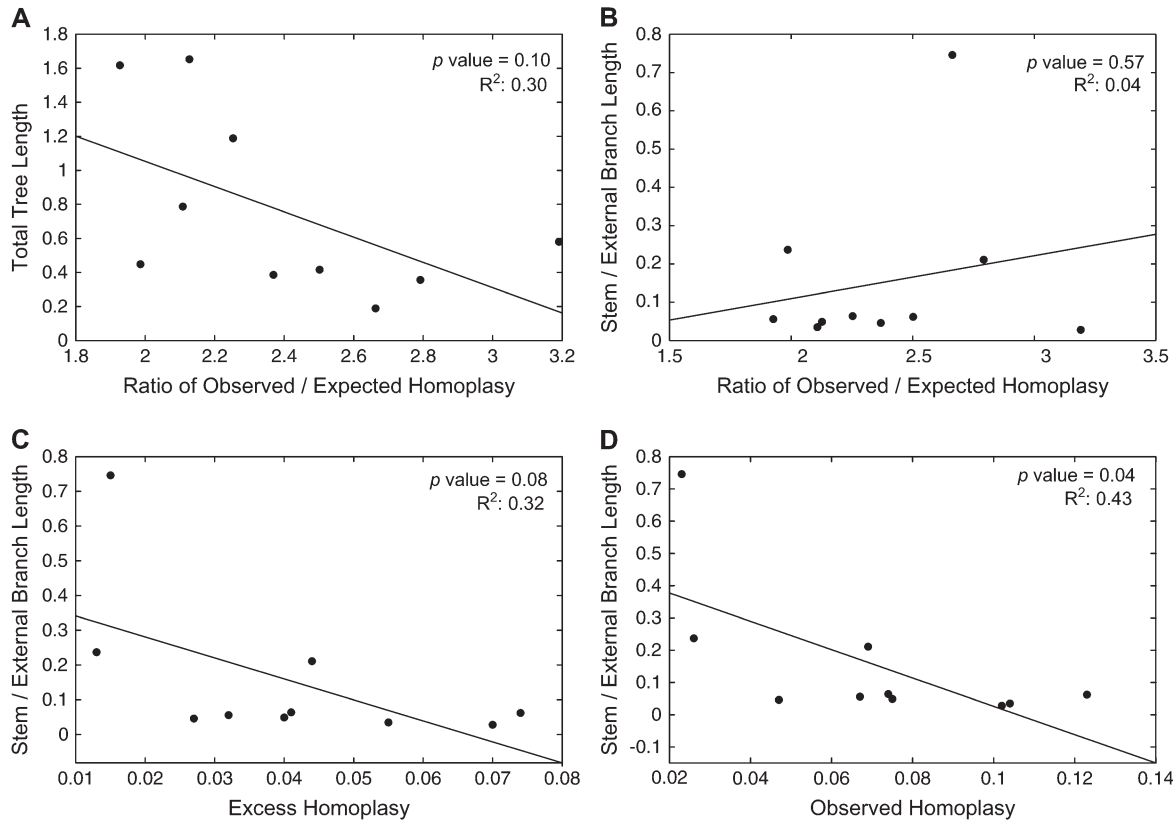


FIG. 2.—Excess homoplasy is a general feature of the eukaryotic proteome. The lack of correlation between homoplasy and amount of evolution measurements indicate that excess homoplasy is a general feature of the eukaryotic proteome and is not dependent on the topology of the lineages examined or their age. (*A*) Total tree length plotted against the ratio of observed/expected homoplasy for the 10 data matrices, (*B*) the ratio of stem to external branch lengths plotted against the ratio of observed/expected homoplasy, (*C*) the ratio of stem/external branch lengths plotted against excess homoplasy, and (*D*) the ratio of stem/external branch lengths plotted against observed homoplasy. All graphs have been fitted with linear regression trend lines, but other regression types (e.g., power or logarithmic) give similar fits and results.

sites involved substitutions from the top12 and middle51 categories combined, whereas just 0.6% of sites exhibited substitutions between the rarely exchangeable amino acids in the bottom12 category. The remaining 22.8% of sites exhibited substitutions between amino acids 2 or 3 mutational steps away that belong to the multiple115 category.

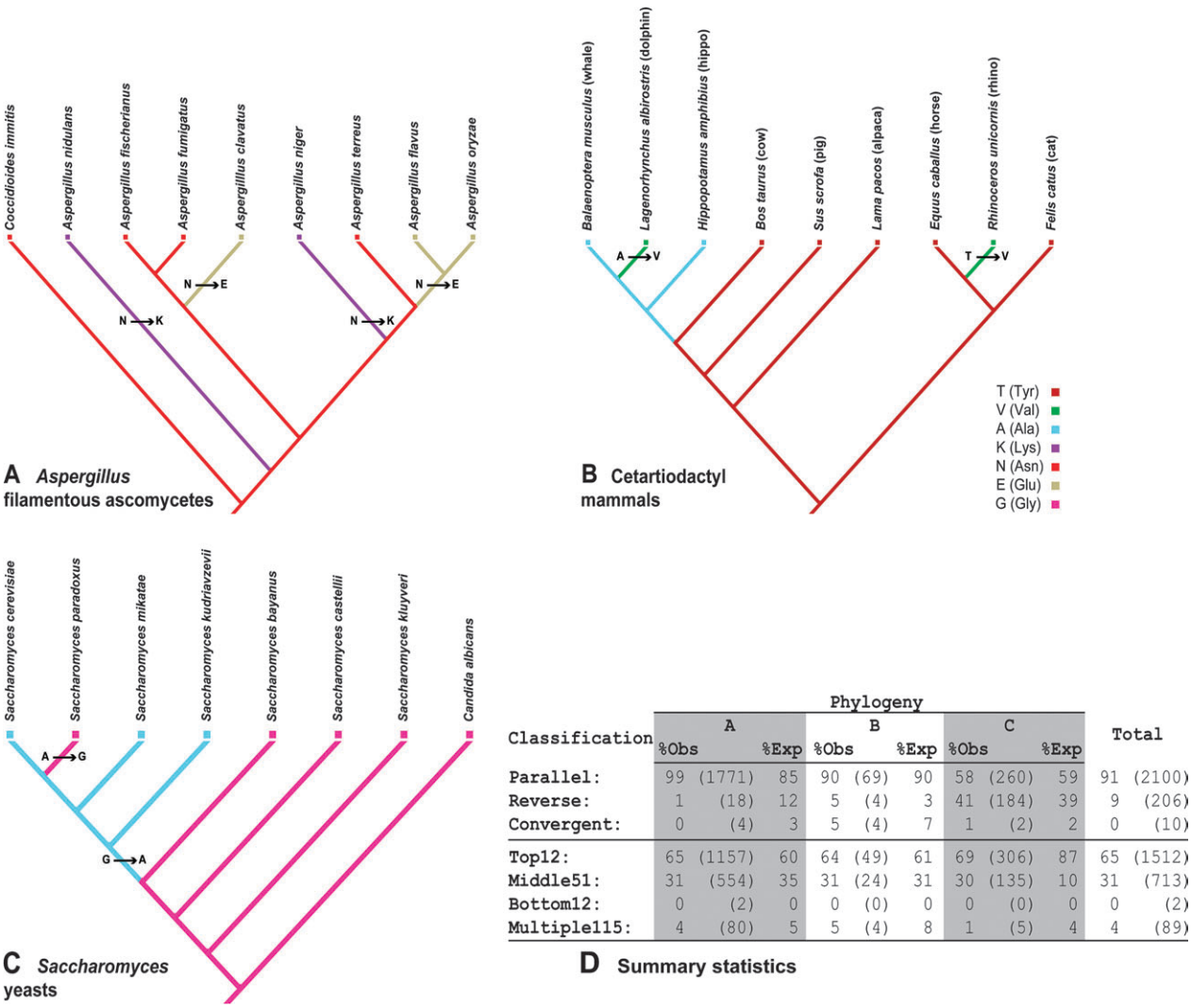
We then examined the most common homoplastic amino acid substitutions in the 3 data matrices shown in figure 3. Strikingly, 65% of homoplasies involved substitutions from just the top12 category and 96% of all homoplastic substitutions belonged to the top12 and middle51 categories (fig. 3). The finding that substitutions in the top12 category were more numerous than substitutions in all other categories combined demonstrates that a very large fraction of all homoplastic substitutions are between amino acids with very similar physicochemical properties that can be reached via a single-mutational step.

Discussion

Understanding the extent and causes of homoplasy is important for understanding the processes that have sculpted the protein record. We used phylogenetic methodology to quantify the extent and type of homoplasy present on the eukaryotic proteome. We found that the frequency of homoplastic amino acid substitutions in eukaryotic proteins was on average 2.4-fold higher than would be expected under widely accepted models of protein evolution. Remarkably, this ratio is relatively stable across clades that differ by more than an order of magnitude in time of origin. In light of the diversity of proteins and taxa sampled, this consistency suggests that the levels of homoplasy observed reflect fundamental and general aspects of protein evolution. Indeed, we found that the majority of these homoplastic substitutions were between frequently exchangeable amino acids that are only one mutational step away and that only an extremely small fraction of substitutions involved rarely exchangeable amino acids. These results bear on our understanding of the role of selection in shaping protein evolution and the biological significance of amino acid sequence differences between species.

The Underlying Causes of Molecular Homoplasy

Two major explanations that have been proposed to account for the elevated levels of homoplasy are mutational rate differences (O'hUigin et al. 2002) and selection (Wells 1996; Bazykin et al. 2007). For example, an examination of 51 primate loci revealed a weak positive correlation between the rate of substitution at individual genes and the



D Summary statistics

Classification	Phylogeny A %Obs	A	%Exp	B %Obs	B	%Exp	C %Obs	C	%Exp	Total	
Parallel:	99	(1771)	85	90	(69)	90	58	(260)	59	91	(2100)
Reverse:	1	(18)	12	5	(4)	3	41	(184)	39	9	(206)
Convergent:	0	(4)	3	5	(4)	7	1	(2)	2	0	(10)
Top12:	65	(1157)	60	64	(49)	61	69	(306)	87	65	(1512)
Middle51:	31	(554)	35	31	(24)	31	30	(135)	10	31	(713)
Bottom12:	0	(2)	0	0	(0)	0	0	(0)	0	0	(2)
Multiple115:	4	(80)	5	5	(4)	8	1	(5)	4	4	(89)

FIG. 3.—Parallelism, reversal, and convergence in eukaryotic proteins. (*A*) An example of 4 homoplastic substitutions giving rise to 2 parallelisms in the *Aspergillus* filamentous ascomycetes. (*B*) An example of 2 homoplastic substitutions giving rise to a convergence in the Cetartiodactyl mammals. (*C*) An example of 2 homoplastic substitutions giving rise to a reversal in the *Saccharomyces* yeasts. Each phylogeny depicts the evolutionary history of an amino acid site, with branch colors corresponding to the amino acid residues present at each branch of the tree. (*D*) Classification of all homoplastic substitutions identified in the examination of the 3 clades shown in panels A–C according to phylogenetic criteria (parallel, reverse, or convergent) and their relative placement along the EI (Tang et al. 2004) (in the top12, middle51, bottom12, or in the multiple115 category). The observed ("%Obs" column) and expected ("%Exp" column) percentage of each type of homoplastic event is shown. The observed number of each type of homoplasy is also shown in parentheses.

degree of homoplasy, which argues for a mutation-based explanation of homoplasy (O'hUigin et al. 2002). However, this trend did not hold for all loci (e.g., slowly evolving genes did not fit the pattern) and the inference rests on the assumption that substitution rates are a good proxy for mutation rates (O'hUigin et al. 2002). Examination of our results from several different eukaryote clades revealed the opposite trend, with homoplasy actually decreasing slightly as the degree of substitution increased (fig. 2). The remarkably similar levels of excess homoplasy, examined in the light of orders-of-magnitude differences in mutation rates across the genes, and organisms used in this study (Baer et al. 2007) make it highly unlikely that mutational rate differences are the principal explanation of excess homoplasy.

There are substantial grounds for considering the role of 2 modes of selection, both "positive" and "purifying" selection, in the generation of homoplasy. Theoretical work suggests that the probability of parallel evolution approximately doubles under positive selection, relative to neutral expectations (Orr 2005), and experimental work has identified several genetic loci in which positive selection has resulted in the parallel evolution of identical amino acid residues in different lineages (Bull et al. 1997; Yeager and Hughes 1999; Jost et al. 2008). Support for the role of weak positive selection in generating excess homoplasy was obtained in a recent study of coding sequences from mammals, yeasts, and *Drosophila* (Bazykin et al. 2007). Comparison of the rate of nonsynonymous (*d*N) to synonymous (*d*S) substitutions with that of nonsynonymous

Table 5
Classification of PISs according to the EI

		Top12		Middle51		Bottom12		Mutliple115	
Clade	No. of Sites	*N*	%	*N*	%	*N*	%	*N*	%
Saccharomyces yeasts	687	297	43.2	210	30.6	4	0.6	176	25.6
Aspergillus filamentous ascomycetes	4,127	1,857	45.0	1,459	35.4	31	0.8	780	18.9
Fungal phyla	5,048	2,234	44.3	1,331	26.4	16	0.3	1,467	29.1
Eukaryotic phyla	4,397	1,898	43.2	1,190	27.1	17	0.4	1,292	29.4
Land plants	448	137	30.6	221	49.3	7	1.6	83	18.5
Drosophila fruit flies	2,649	1,092	41.2	1,119	42.2	35	1.3	403	15.2
Cetartiodactyl mammals	116	49	42.2	43	37.1	0	0.0	24	20.7
Metazoan phyla	3,700	1,699	45.9	751	20.3	12	0.3	1,238	33.5
Paenungulata mammals[a]	127	62	48.8	47	37.0	0	0.0	18	14.2
Vertebrates[a]	300	148	49.3	81	27.0	3	1.0	68	22.7
Averages	**2,160**	**947.3**	**43.4**	**645**	**33.2**	**13**	**0.6**	**555**	**22.8**

[a] Studies where phylogeny is unknown.

parallel (*d*NP) to synonymous parallel (*d*SP) substitutions, revealed that the *d*NP/*d*SP ratio was approximately 5-fold higher than the *d*N/*d*S ratio (Bazykin et al. 2007). By assuming that *d*SP and *d*S, the rate of synonymous parallel and synonymous substitutions, respectively, were good proxies for the rates of selectively neutral substitutions in the lineages examined, the elevated rates observed for *d*NP relative to *d*N could be attributed in part to the action of weak positive selection (Bazykin et al. 2007).

But positive selection is not the only selective explanation for excess homoplasy. An alternative, but not mutually exclusive, explanation is raised by considering the effect of purifying selection. Purifying selection constrains the amino acid residues permitted at variable sites in protein sequences (Kimura 1983; Wells 1996; Naylor and Brown 1997; Bazykin et al. 2007). Some of the most common parallel substitutions in our data matrices involve amino acids with similar physicochemical properties, for example, valine and isoleucine (both hydrophobic and aliphatic) or aspartate and glutamate (both negatively charged and polar). For sites in protein sequences which can only accept amino acids with specific physicochemical properties, the substitution of one amino acid for its equivalent may be functionally neutral, whereas substitutions for all other, nonequivalent, amino acids will be deleterious. Thus, purifying selection on such sites makes the fixation of homoplastic substitutions more likely and frequent relative to other sites that tolerate a wider range of substitutions.

A large body of biophysical studies of protein structure and activity suggest that such constrained sites constitute a substantial fraction of all residues within and are widely dispersed throughout most proteins (reviewed in Pakula and Sauer [1989]). Most phenotypically defective missense mutants do not affect protein activity directly but do so indirectly (Pakula and Sauer 1989). Systematic replacement of amino acids within a variety of proteins has revealed that many mutations at many positions outside of active sites affect properties such as protein folding, stability, and aggregation (Pakula and Sauer 1989; DePristo et al. 2005). These functional studies, and our observation of the bias for parallel replacement of physicochemically similar amino acids, suggest that strong selective constraints operate not only upon the most conserved parts of proteins but also upon their nonconserved parts as well.

The Functional and Biological Significance of Molecular Homoplasy

What is the functional meaning of this abundance of homoplasy in eukaryotic proteomes? Although several cases of conservative amino acid substitutions resulting in parallel adaptation have been identified (Stewart et al. 1987; Deeb et al. 2003; Zhang 2006), it is highly unlikely that the majority of parallel amino acid substitutions observed in the protein record has been driven by large selection coefficients. Several recent population genetic studies of model organisms (mainly *Drosophila*) have estimated that a significant fraction of amino acid substitutions in protein-coding genes (Smith and Eyre-Walker 2002; Sawyer et al. 2003; Begun et al. 2007) has been driven by positive selection but that the magnitude of their selective effects is nearly neutral (Sawyer et al. 2007). This statistical inference is supported by some lines of experimental evidence. For example, *Saccharomyces cerevisiae* strains differing by a handful to even scores of amino acid substitutions in a variety of proteins exhibit no significant differences in fitness in the most sensitive assays devised to date (Williams BL, Carroll SB, in preparation). Thus, our finding that a large fraction of homoplastic substitutions are conservative is consistent with emerging statistical and experimental data and suggests that many of these substitutions are either functionally equivalent or have been driven by very small selection coefficients and are thus unlikely to contribute to adaptation.

Excess homoplasy and the contribution of different modes of selection to its generation bear important implications for studies in phylogenetics and molecular evolution. On the one hand, excess homoplasy raises novel statistical challenges with the analysis of molecular data because a general lack of correspondence between an underlying model and actual evolutionary processes can lead to the failure of a statistical methodology (Naylor and Brown 1998; Rokas and Carroll 2006). On the other hand, the finding that the majority of homoplastic substitutions are conservative in nature signifies that, although statistically

important, most of these homoplastic substitutions are not biologically meaningful in terms of shaping molecular function or organismal diversity (Nei 2005).

Supplementary Material

Supplementary figure S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Barry L. Williams for insightful comments on the manuscript. Research in AR's laboratory is supported by the Searle Scholars Program and Vanderbilt University. SBC is an investigator of the Howard Hughes Medical Institute.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet. 8:619–631.

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. Biol Direct. 2:20.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5:e310.

Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997. Exceptional convergent evolution in a virus. Genetics. 147:1497–1507.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Cracraft J, Donoghue MJ. 2004. Assembling the tree of life. Oxford: Oxford University Press. p. 576.

Crill WD, Wichman HA, Bull JJ. 2000. Evolutionary reversals during viral adaptation to alternating hosts. Genetics. 154:27–37.

Deeb SS, Wakefield MJ, Tada T, Marotte L, Yokoyama S, Marshall Graves JA. 2003. The cone visual pigments of an Australian marsupial, the tammar wallaby (*Macropus eugenii*): sequence, spectral tuning, and evolution. Mol Biol Evol. 20:1642–1649.

Depristo MA, Hartl DL, Weinreich DM. 2007. Mutational reversions during adaptive protein evolution. Mol Biol Evol. 24:1608–1610.

DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 6:678–687.

Drake JW. 2006. Chaos and order in spontaneous mutation. Genetics. 173:1–8.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol. 21:569–575.

Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol Biol Evol. 18:10–21.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer.

Gatesy J, O'Leary MA. 2001. Deciphering whale origins with molecules and fossils. Trends Ecol Evol. 16:562–570.

Gaut BS, Lewis PO. 1995. Success of maximum likelihood phylogeny inference in the 4-taxon case. Mol Biol Evol. 12:152–162.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 99:364–373.

Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI. 2001. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J Virol. 75:7966–7972.

Jost MC, Hillis DM, Lu Y, Kyle JW, Fozzard HA, Zakon HH. 2008. Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. Mol Biol Evol. 25:1016–1024.

Kallersjo M, Albert VA, Farris JS. 1999. Homoplasy increases phylogenetic structure. Cladistics. 15:91–93.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217:624–626.

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

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science. 164:788–798.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39:309–338.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer.

Naylor GJP, Brown WM. 1997. Structural biology and phylogenetic estimation. Nature. 388:527–528.

Naylor GJP, Brown WM. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst Biol. 47:61–76.

Naylor GJP, Collins TM, Brown WM. 1995. Hydrophobicity and phylogeny. Nature. 373:565–566.

Nei M. 2005. Selectionism and neutralism in molecular evolution. Mol Biol Evol. 22:2318–2342.

Nikaido M, Rooney AP, Okada N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci USA. 96p. 10261–10266.

Nishihara H, Satta Y, Nikaido M, Thewissen JGM, Stanhope MJ, Okada N. 2005. A retroposon analysis of Afrotherian phylogeny. Mol Biol Evol. 22:1823–1833.

O'hUigin C, Satta Y, Takahata N, Klein J. 2002. Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. Mol Biol Evol. 19:1501–1513.

Orr HA. 2005. The probability of parallel evolution. Evol Int J Org Evol. 59:216–220.

Page RDM, Holmes EC. 1998. Molecular evolution: a phylogenetic approach. Oxford (UK): Blackwell Science.

Pakula AA, Sauer RT. 1989. Genetic analysis of protein stability and function. Annu Rev Genet. 23:289–310.

Pinel-Galzi A, Rakotomalala M, Sangu E, et al. (14 co-authors). 2007. Theme and variations in the evolutionary pathways to virulence of an RNA plant virus species. PLoS Pathog. 3:e180.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc R Soc Lond Ser B Biol Sci. 269:1313–1316.

Rogozin IB, Thomson K, Csuros M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct. 3:7.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Rokas A, Galagan JE. 2008. The *Aspergillus nidulans* genome and a comparative analysis of genome evolution in *Aspergillus*. In: Osmani SA, Goldman GH, editors. The aspergilli: genomics, medical applications, biotechnology, and research methods. Boca Raton (FL): CRC Press. p. 43–55.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol Biol Evol. 20: 1036–1042.

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J Mol Evol. 57(Suppl 1):S154–S164.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci USA. 104: 6504–6510.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is "saturation"? Genetics. 142:1033–1036.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature. 415:1022–1024.

Stewart CB, Schilling JW, Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature. 330:401–404.

Takahata N. 1993. Allelic genealogy and human evolution. Mol Biol Evol. 10:2–22.

Takezaki N, Figueroa F, Zaleska-Rutczynska Z, Takahata N, Klein J. 2004. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of 44 nuclear genes. Mol Biol Evol. 21:1512–1524.

Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. Mol Biol Evol. 21:1548–1556.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thorne JL. 2007. Protein evolution constraints and model-based techniques to study them. Curr Opin Struct Biol. 17:337–341.

Wells RS. 1996. Excessive homoplasy in an evolutionarily constrained protein. Proc R Soc Lond Ser B Biol Sci. 263:393–400.

Whelan S, Lio P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 17:262–272.

Wichman HA, Scott LA, Yarber CD, Bull JJ. 2000. Experimental evolution recapitulates natural evolution. Philos Trans R Soc Lond Ser B Biol Sci. 355:1677–1684.

Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. Proc Natl Acad Sci USA. 103:9107–9112.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 15:1600–1611.

Yeager M, Hughes AL. 1999. Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunol Rev. 167:45–58.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 38:819–823.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol. 14:527–536.

Zhang L, Li WH. 2005. Human SNPs reveal no evidence of frequent positive selection. Mol Biol Evol. 22:2504–2507.

Kenneth Wolfe, Associate Editor

Accepted June 19, 2008