Apparent Trends of Amino Acid Gain and Loss in Protein Evolution Due to Nearly Neutral Variation

John H. McDonald
Department of Biological Sciences, University of Delaware

It has recently been claimed that certain amino acids have been increasing in frequency in all living organisms for most of the history of life on earth, while other amino acids have been decreasing in frequency. Three lines of evidence have been offered for this assertion, but each has a more plausible alternative interpretation. Here I show that unequal patterns of gains and losses for particular pairs of amino acids (such as more leucine → phenylalanine than phenylalanine → leucine substitutions in humans and chimpanzees since they split from a common ancestor) are consistent with a simple neutral model at equilibrium amino acid frequencies. Unequal numbers of gains and losses for particular amino acids (such as more gains than losses of cysteine) are shown by simulations to be explainable by the nearly neutral model as well. In a comparison of protein sequences from four strains of *Escherichia coli*, polarized by one outgroup strain of *Salmonella*, the disparity in number of gains and losses for particular amino acids is strong in terminal branches but weaker or nonexistent in internal branches, which is inconsistent with the universal trend model but as expected under the nearly neutral model.

Introduction

Jordan et al. (2005) compared protein sequences from pairs of closely related bacteria, yeast, and mammals, inferring the ancestral states by comparison with a more distantly related outgroup for each pair of species. They also examined human protein polymorphism data, with the ancestral allele at each polymorphic site inferred from comparison with chimpanzee sequences. Their surprising conclusion was that certain amino acids, such as cysteine, methionine, histidine, serine, and phenylalanine, are more likely to be gained than lost in protein evolution, while other amino acids, such as proline, alanine, glutamic acid, and glycine, are more likely to be lost than gained. Furthermore, they concluded that these trends are universal in all living organisms and have been ongoing for billions of years. They speculated that these trends result because some amino acids, having been added to the genetic code relatively recently, are still increasing toward their equilibrium frequency. Earlier, Zuckerkandl, Derancourt, and Vogel (1971) performed a similar analysis of a much smaller data set (globins and cytochrome c in several vertebrates) and reached similar conclusions. Here I examine the evidence provided by Jordan et al. (2005) and suggest alternate explanations that do not involve universal, directional changes in amino acid composition.

Substitutional Asymmetry of Pairs of Amino Acids

Jordan et al. (2005) offered three lines of evidence for universal trends of change in amino acid frequencies. The first is that for many pairs of amino acids, there are unequal numbers of substitutions in one direction compared with those in the opposite direction. For example, in the comparison of humans with chimpanzees, using mice (*Mus musculus*) as the outgroup, there are 102 sites where phenylalanine has been gained and leucine lost, while only 57 sites have the opposite pattern (leucine gained and phenylalanine lost). Jordan et al. (2005) conclude that amino acid frequencies are not stationary but are instead changing in frequency.

In a two-allele system at equilibrium, there must be equal number of substitutions in each direction; a significant difference in the number of substitutions would indeed indicate that the allele frequencies were changing. However, there are 20 possible amino acids at any site. For a site with more than two possible alleles, equal numbers of substitutions in each direction would only be expected if the evolutionary process were a time-reversible Markov process with stationary amino acid frequencies (Li and Goldman 1998). In a reversible process, by definition the number of A → B substitutions per unit time is equal to the number of B → A substitutions, so that the process would look the same whether time was running forward or backward. Reversibility is usually assumed when inferring a matrix of substitution rates from pairwise comparisons of sequences (Dayhoff, Schwartz, and Orcutt 1978; Jones, Taylor, and Thornton 1992; S. Henikoff and J. G. Henikoff 2000; Müller and Vingron 2000; Goldman and Whelan 2002; Veerassamy, Smith, and Tillier 2003; Kosiol and Goldman 2005). However, this assumption is made purely for mathematical convenience; there is no biological evidence for reversibility (Li and Goldman 1998).

With more than two alleles, it is easy to create neutral models of protein evolution that are not reversible. For example, consider a three-allele model in which the substitution rate from B to C ($P_{BC}$) is $4 \times 10^{-5}$ substitutions per generation and $P_{AB} = P_{AC} = P_{BA} = P_{CA} = P_{CB} = 1 \times 10^{-5}$. Using this rate matrix in a Markov chain model until the amino acid frequencies become stationary, the frequencies of A, B, and C are $f_A = 0.333$, $f_B = 0.167$, and $f_C = 0.500$. Because there are twice as many A as B amino acids and $P_{AB} = P_{BA}$, there will be twice as many A → B as B → A substitutions per unit time. Clearly, differences in the numbers of forward and reverse substitutions for pairs of amino acids are consistent with a simple neutral model at equilibrium and do not necessarily indicate changing amino acid frequencies.
Unequal Numbers of Gains and Losses of Amino Acids

The second line of evidence offered by Jordan et al. (2005) is that certain amino acids are gained more than they are lost, while others are lost more than they are gained. For example, in the comparison of humans with chimpanzees there are 137 sites where cysteine has been substituted for a different ancestral amino acid (“gains” of cysteine) and only 72 sites where cysteine has been lost; similar excess gains of cysteine are seen in rodents, yeast, and 12 taxa of bacteria. This could be explained by a trend of increasing frequencies of cysteine, but there are other explanations that do not involve changing amino acid composition.

One problem with using parsimony to infer gains and losses of amino acids is that the inferences are sometimes incorrect; when an outgroup and one ingroup have amino acid B and the other ingroup has A, there may have been two A → B substitutions, not one B → A. This is particularly likely if A is more common than B; there will be more inferred common → rare than rare → common substitutions, even when the actual number of substitutions in each direction is equal (Collins, Wimberger, and Naylor 1994; Perna and Kocher 1995). Jordan et al. (2005) attempted to avoid this problem by using closely related species, but there can be substantially more inferred common → rare substitutions despite fairly small amounts of divergence (Eyre-Walker 1998).

Sites which fit the nearly neutral model of protein evolution (Ohta 1992) may be particularly likely to show more inferred substitutions in one direction than the other. Consider a site at which amino acid A is favored by natural selection. When a mutation to amino acid B occurs, it is selected against. If the selection against B is weak enough, B may remain present for a while before being replaced by A again. Under this model, there is a chance that when comparing sequences from an outgroup and two ingroup species, one of the ingroups will have the mildly deleterious B, while the other two species have the preferred A. This would be interpreted as a gain of B. On the other hand, an apparent loss of B, in which the outgroup and one of the ingroup species would both have the deleterious B while only one of the ingroup species has the preferred A, would be quite unlikely; it would require either two independent substitutions of the deleterious B, one in an outgroup and one in an ingroup, or a deleterious B that survived in two lineages since the common ancestor of all three species. Under this nearly neutral model, there could be many more apparent gains of B than losses, even if the ancestral state at sites that differ among the taxa is always inferred correctly. This may seem paradoxical, but the “missing” losses of B would occur at sites where a common ancestor of two species happened to have the mildly deleterious B, which was replaced by the favored A in both daughter lineages (fig. 1).

To illustrate how nearly neutral evolution could produce a difference in the number of gains and losses, I wrote a computer program to simulate mutation, drift, and selection for a two-allele locus, with allele A being preferred and allele B being deleterious. The population size was 25 diploid individuals, and the fitnesses of genotypes AA, AB, and BB were 1, 1 − s, and 1 − 2s, respectively. The A → B and B → A mutation rates, μA and μB, were 0.0001 per generation. For each replicate, a single population was started with the initial state either fixed for A or fixed for B. The initial probability of being fixed for B, Pinitial, was determined by setting the number of A → B substitutions equal to the number of B → A substitutions, Pinitial = Pfinal, solving for Pfinal, Pfinal = μA/μB, and then using equation (10) of Kimura (1962) for μA and μB, the probabilities of fixation of A and B. Evolution in this initial population was first simulated for 1,000 generations to produce equilibrium levels of polymorphism. The population was then duplicated into an outgroup and an ingroup lineage, and evolution was simulated for 1,000 more generations. The ingroup was then duplicated, and evolution was simulated in the outgroup and the two ingroups for 1,000 more generations. One allele was then sampled at random from each of the three species. If the two ingroup alleles were different, the allele that was in the outgroup was counted as lost in one of the ingroups, while the allele that was not in the outgroup was counted as gained. The simulations were replicated 10,000 times for each selection coefficient.

When 2Ns = 0 (the neutral model), the equilibrium frequency of B is 0.5; as the selection coefficient against B increases, the equilibrium frequency of B declines (fig. 2). The final frequency is the same as the initial frequency, indicating that there is no trend of changing allele frequencies in
these simulations. The number of gains of B initially increases and then declines as the selection coefficient against B increases (fig. 3). The increase is presumably due to the increasing frequency of A (and thus increasing frequency of sites where B can be gained). The number of losses of B declines more rapidly; as selection against B intensifies and the average frequency of B decreases, it becomes unlikely that both the outgroup and one of the ingroups would have B at the end of the simulated generations (the pattern that is interpreted as a loss of B). As a result, nearly neutral evolution produces many more apparent gains than losses of a mildly deleterious allele. Although this simple two-allele model could be made more elaborate and realistic, the results seem sufficient to demonstrate that unequal numbers of gains and losses do not necessarily indicate changing allele frequencies.

**Nearly Neutral Human Polymorphisms**

The third line of evidence for a universal trend offered by Jordan et al. (2005) is that for human protein polymorphisms (with chimpanzee as the outgroup), the ratio of gains to losses for most amino acids is significantly different from one. However, the nearly neutral model could also explain this result. Consider a site where A is favored by selection and B is mildly deleterious. If the selection against B is weak, it is possible for B to appear as a polymorphism for a while. There will thus be some polymorphisms where A is the ancestral allele and B has been gained. But because B is mildly deleterious, it is unlikely to occur in both the outgroup and as a polymorphism in the ingroup. Therefore, there will be fewer sites where B is the ancestral allele and A is a derived polymorphism.

To test whether nearly neutral evolution could produce more gains than losses for polymorphism data, I simulated evolution using the same model described above. The original population evolved for 1,000 generations, then it was duplicated and the outgroup and ingroup species evolved for 1,000 more generations. If the ingroup was polymorphic, one allele was sampled at random from the outgroup to infer which allele was gained and which was lost. The simulations were replicated 10,000 times for each selection coefficient.

The simulations show that for a broad range of selection coefficients, nearly neutral evolution results in a large number of differences between the number of gains and losses in polymorphism data (fig. 4). The results demonstrate that the unequal numbers of gains and losses in human polymorphism data found by Jordan et al. (2005) do not necessarily indicate changing allele frequencies but may merely add to the evidence that many protein polymorphisms in humans are mildly deleterious, an interpretation that is simply supported by evidence from mitochondrial (Nachman et al. 1996; Hasegawa, Cao, and Yang 1998; Moilanen and Majamaa 2003; Elson, Turnbull, and Howell 2004; Ho et al. 2005) and nuclear genes (Cargill et al. 1999; Fay, Wyckoff, and Wu 2001; Sunyaev et al. 2001, 2003; Hughes et al. 2003; Williamson et al. 2005).

**Nearly Neutral Polymorphisms in* Escherichia coli***

One way to distinguish between the universal trend model and the nearly neutral model is to examine the pattern of gains and losses on a phylogeny with more than three taxa. If the long-term directional trend postulated by Jordan et al. (2005) is correct, the ratio of gains to losses should be similar in all parts of a phylogeny. Under the nearly neutral model, however, the apparent excess gains will be confined mostly to the tips of a phylogeny, as mildly deleterious substitutions have a relatively short “lifetime” and are unlikely to survive long enough to appear in more than one taxon. To test this, I used the program Blastall (Altschul et al. 1997) to identify matching protein sequences in one outgroup, *Salmonella enterica* Typhi CT18 (Parkhill et al. 2001) and four completely sequenced strains of *Escherichia coli*: *E. coli* K12 (Blattner et al. 1997), *E. coli* O157:H7 (Peña et al. 2001), *E. coli* CFT073 (Welch et al. 2002), and “*Shigella flexneri*” 2a301 (Jin et al. 2002). Sequences with the matching portion reported by Blastall shorter than 50 amino acids, sets of sequences for which not all *E. coli* strains had a match in *S. enterica*, and sets of sequences in which the proportion of pairwise differences between the *S. enterica* and *E. coli* sequences varied significantly (2 × 4 G-test, *P* < 0.05) were deleted. The protein sequences were then aligned using ClustalW (Thompson, Higgins, and Gibson 1994). Ambiguously aligned sites adjacent to gaps were omitted, with the omitted sites extending from the gap to the nearest pair of adjacent sites that were both identical in all five sequences. Different genes had different estimated phylogenies, presumably because of recombination (Guttman and Dykuizen 1994; McGraw et al. 1999), so no attempt was made to estimate an evolutionary
tree and assign substitutions to particular branches of the tree. Instead, gains and losses were simply divided into two classes, “unique” (the derived amino acid present in only one strain of *E. coli*) or “shared” (the derived amino acid present in more than one strain).

The data of Jordan et al. (2005) and the unique substitutions (those seen in two or three strains of *E. coli*) are shown for comparison. D, the number of gains minus losses divided by the total number of substitutions. For 18 amino acids, the shared substitutions have a larger difference for unique substitutions than for the data of Jordan et al. (2005) (table 1). However, there is a marked difference between unique substitutions and shared substitutions in the patterns of gain and loss. For 18 amino acids, the shared substitutions have a normalized difference between gains and losses that is either smaller than or in the opposite direction to those seen at unique sites. Only 8 amino acids have a significant difference between gains and losses for shared substitutions, while 19 amino acids have a significant difference for unique substitutions. The stronger bias seen for unique gains and losses is inconsistent with the universal trend postulated by Jordan et al. (2005), which would predict the same amount of bias on all branches of a phylogeny, but it is expected under the nearly neutral model. This adds to the evidence that many protein polymorphisms in *E. coli* are mildly deleterious (Sawyer, Dykhuizen, and Hartl 1987; Hughes 2005). The bias for shared gains and losses seen for some amino acids may result from universal trends in amino acid composition, but they may also represent mildly deleterious alleles that are so nearly neutral that they have survived in more than one strain. The bias in shared gains and losses may also result from directional changes in amino acid composition due to positive selection, as has been peatedly shown for related species of prokaryotes living in different environments (Haney et al. 1999; McDonald, Grasso, and Rejo 1999; McDonald 2001; Nishio et al. 2003; Di Giulio 2005; Methé et al. 2005).

### Table 1

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Unique Gains</th>
<th>Unique Losses</th>
<th>Shared Gains</th>
<th>Shared Losses</th>
<th>Jordan et al. (2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys</td>
<td>177</td>
<td>31</td>
<td>16</td>
<td>11</td>
<td>0.19</td>
</tr>
<tr>
<td>His</td>
<td>326</td>
<td>137</td>
<td>109</td>
<td>81</td>
<td>0.15</td>
</tr>
<tr>
<td>Ser</td>
<td>638</td>
<td>285</td>
<td>179</td>
<td>159</td>
<td>0.06</td>
</tr>
<tr>
<td>Met</td>
<td>194</td>
<td>103</td>
<td>65</td>
<td>41</td>
<td>0.22</td>
</tr>
<tr>
<td>Tyr</td>
<td>114</td>
<td>65</td>
<td>30</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Phe</td>
<td>144</td>
<td>95</td>
<td>35</td>
<td>31</td>
<td>0.06</td>
</tr>
<tr>
<td>Thr</td>
<td>504</td>
<td>376</td>
<td>169</td>
<td>159</td>
<td>0.03</td>
</tr>
<tr>
<td>Lys</td>
<td>202</td>
<td>162</td>
<td>64</td>
<td>89</td>
<td>-0.16</td>
</tr>
<tr>
<td>Val</td>
<td>551</td>
<td>510</td>
<td>220</td>
<td>191</td>
<td>0.07</td>
</tr>
<tr>
<td>Leu</td>
<td>360</td>
<td>349</td>
<td>96</td>
<td>128</td>
<td>-0.14</td>
</tr>
<tr>
<td>Glu</td>
<td>262</td>
<td>326</td>
<td>148</td>
<td>157</td>
<td>-0.03</td>
</tr>
<tr>
<td>Asp</td>
<td>268</td>
<td>395</td>
<td>155</td>
<td>149</td>
<td>0.02</td>
</tr>
<tr>
<td>Gln</td>
<td>181</td>
<td>285</td>
<td>84</td>
<td>83</td>
<td>0.01</td>
</tr>
<tr>
<td>Arg</td>
<td>230</td>
<td>398</td>
<td>65</td>
<td>88</td>
<td>-0.15</td>
</tr>
<tr>
<td>Gly</td>
<td>184</td>
<td>365</td>
<td>56</td>
<td>54</td>
<td>0.02</td>
</tr>
<tr>
<td>Trp</td>
<td>13</td>
<td>31</td>
<td>4</td>
<td>3</td>
<td>0.14</td>
</tr>
<tr>
<td>Ala</td>
<td>347</td>
<td>1033</td>
<td>210</td>
<td>269</td>
<td>-0.12</td>
</tr>
<tr>
<td>Pro</td>
<td>90</td>
<td>310</td>
<td>26</td>
<td>54</td>
<td>-0.35</td>
</tr>
</tbody>
</table>

Note.—For each amino acid, the numbers of gains and losses are shown for unique substitutions (those seen in only one strain of *E. coli*) and shared substitutions (those seen in two or three strains of *E. coli*). Data on *E. coli* from Jordan et al. (2005) are shown for comparison. D, the number of gains minus losses divided by the total number of substitutions. *Values of D that are significantly different from 0 (exact binomial test, P < 0.05).*

### Acknowledgments

I thank H. Akashi, B. C. Verrelli, and B. J. Wolpert for helpful comments on the manuscript.

### Literature Cited


William Martin, Associate Editor

Accepted September 19, 2005