Detection of Convergent and Parallel Evolution at the Amino Acid Sequence Level

Jianzhi Zhang and Sudhir Kumar

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

Adaptive evolution at the molecular level can be studied by detecting convergent and parallel evolution at the amino acid sequence level. For a set of homologous protein sequences, the ancestral amino acids at all interior nodes of the phylogenetic tree of the proteins can be statistically inferred. The amino acid sites that have experienced convergent or parallel changes on independent evolutionary lineages can then be identified by comparing the amino acids at the beginning and end of each lineage. At present, the efficiency of the methods of ancestral sequence inference in identifying convergent and parallel changes is unknown. More seriously, when we identify convergent or parallel changes, it is unclear whether these changes are attributable to random chance. For these reasons, claims of convergent and parallel evolution at the amino acid sequence level have been disputed. We have conducted computer simulations to assess the efficiencies of the parsimony and Bayesian methods of ancestral sequence inference in identifying convergent and parallel-change sites. Our results showed that the Bayesian method performs better than the parsimony method in identifying parallel changes, and both methods are inefficient in identifying convergent changes. However, the Bayesian method is recommended for estimating the number of convergent-change sites because it gives a conservative estimate. We have developed statistical tests for examining whether the observed numbers of convergent and parallel changes are due to random chance. As an example, we reanalyzed the stomach lysozyme sequences of foregut fermenters and found that parallel evolution is statistically significant, whereas convergent evolution is not well supported.

Introduction

It is important to understand adaptive evolution at the molecular level (Nei 1990). One of the approaches is to study convergent and parallel amino acid changes in protein evolution. Here, a convergent change at an amino acid site refers to changes from different ancestral amino acids to the same descendant amino acid along independent evolutionary lineages (see fig. 1*A* for examples). It is distinguished from a parallel change, in which amino acid changes along independent lineages have occurred from the same ancestral amino acid (see fig. 1*A* for examples). Both convergent and parallel evolution, if verified, suggest adaptive evolution. The reason to distinguish them is that the convergence is a stronger indication of adaptive evolution, because under purifying selection and neutral evolution, convergent changes are expected to occur more rarely than parallel changes.

The study of convergent and parallel evolution at the amino acid sequence level involves two steps. The first step is to identify the amino acid sites that have experienced convergent or parallel changes. For a given set of amino acid sequences whose phylogenetic relationships are known (or can be reconstructed), the ancestral amino acids at all interior nodes of the phylogenetic tree can be inferred. Using this information, we can tell whether there are convergent or parallel changes on particular evolutionary lineages. The parsimony method (Hartigan 1973; see also Maddison and Maddison 1992) and the Bayesian method (Yang, Kumar, and Nei 1995; Zhang and Nei 1997) are often used for inferring ancestral amino acids, but the efficiencies of both methods in inferring convergent and parallel changes are yet to be studied. The second step in the study of convergent and parallel evolution is to test whether the identified convergent and parallel changes can be attributed to random chance. This kind of test is necessary because a few convergent or parallel amino acid changes may simply arise by random chance, as protein sequence evolution is a stochastic process with at most 20 possible states at each site. Stewart, Schilling, and Wilson (1987) compared the number of uniquely shared sites (see below) for the potentially convergent or parallel sequences with the average number of uniquely shared sites for all pairs of sequences in the data and used a simple χ^2 test to see whether the former is significantly greater than the latter. Their procedure does not take into account the extent of divergence among pairs of sequences, so the test may be insensitive or liberal depending on the sequence data. Stewart, Schilling, and Wilson (1987) and Swanson, Irwin, and Wilson (1991) have invented another test, so-called the winning test, of convergent evolution. This test relies on the conflict between the true tree and the estimated tree. Because the failure of obtaining the correct tree may not be due to convergent evolution (Nei 1996) and convergent evolution may not affect the tree reconstruction (Adachi and Hasegawa 1996), the winning test is inappropriate. Because of lack of a rigorous statistical test, claims of convergent and parallel evolution have been disputed in the literature (Doolittle 1994; Kreitman and Akashi 1995 and references therein).

The purpose of this study was to investigate whether convergent and parallel changes can be correctly inferred from the present-day sequences and to develop statistical tests for examining whether the observed convergent and parallel changes can be attributed to random

Key words: convergent evolution, parallel evolution, amino acid sequence, ancestral sequence, adaptive evolution, lysozyme.

Address for correspondence and reprints: Jianzhi Zhang, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 322 Mueller Laboratory, University Park, Pennsylvania 16802. E-mail: zhang@imeg.bio.psu.edu.

Mol. Biol. Evol. 14(5):527–536. 1997

chance. As an example, we analyzed the stomach lysozyme sequences of foregut fermenters, which were thought to have undergone convergent and parallel evolution. In this work, we focused our attention on the amino acid sequences because adaptive evolutionary processes are likely to be evident at the amino acid rather than the nucleotide sequence level.

Methods

Statistical Tests for Examining Whether the Observed Numbers of Convergent-Change and Parallel-Change Sites Can Be Attributed to Random Chance

The tests of convergent and parallel evolution are similar, so we will first describe the test of convergent evolution. One may want to test the convergent evolution at each observed convergent-change site. This is not easy because the probability for a particular amino acid configuration at a given site of all sequences in the data is usually very small. Instead of conducting a statistical test at each convergent-change site, we test convergent evolution by considering all sites of the sequences. In our test, the null hypothesis is that all observed convergent changes can be explained by random chance under a certain substitution model.

For simplicity, let us assume that the data set consists of five aligned present-day amino acid sequences (1–5), whose phylogenetic relationships are given in figure 1*B*. The ancestral sequences at the interior nodes are sequences 6, 7, and 8, which can be statistically inferred (reviewed in Zhang and Nei 1997). First, we have to choose two (or more) lineages (called focused lineages) along which we are going to study the convergent evolution. Generally, a focused lineage can be one branch of the tree or several head-to-tail-connected branches. However, a lineage of only one branch is recommended since it enables us to know more specifically when the convergent or parallel evolution occurred. The focused lineage must begin at an interior node and end at either an interior or an exterior node. The direction of evolution on each focused lineage must be known so that one end of the lineage represents the ancestral state, and the other represents the descendant state. We will consider the amino acid change on the lineage by comparing these two states irrespective of any intermediate state. This is because, in practice, the intermediate state is often unknown and a focused lineage usually only consists of one branch. It is obvious that the tree root is not allowed to be on any of the focused lineages. It is also required that the focused lineages be independent, i.e., (1) there is no shared tree branch between different focused lineages, and (2) there is no shared point between different focused lineages except that they can have the same starting point. For simplicity, we choose the branch from node 6 to node 1 and the branch from node 8 to node 3 (fig. 1*B*) as two focused lineages for further explanations. Then nodes 6 and 8 are the ancestral nodes and nodes 1 and 3 are the descendant nodes of the two focused lineages, respectively. For a given site, let x_k be the amino acid in sequence k. As mentioned earlier, a site is called a convergent-change site under the following conditions (fig. 1*B*): the amino acids at the descendant nodes are identical with each other ($x_1 = x_3$) and different from their respective ancestral amino acids ($x_1 \neq x_6$ and $x_3 \neq x_8$), and these ancestral amino acids are different ($x_6 \neq x_8$). Note that whether a site is a convergent-change site depends on the focused lineages we choose.

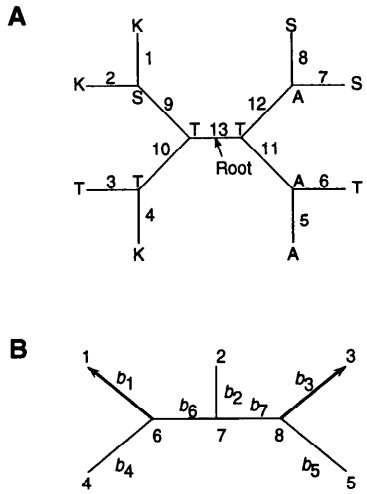


FIG. 1.—Examples of convergent and parallel changes. *A*, Convergent changes, parallel changes, and uniquely shared sites. There are convergent changes on branches 8 and 9 (A→S and T→S, respectively). There are parallel changes on branches 7 and 8 (A→S). When a focused lineage consists of several head-to-tail-connected branches, the amino acid change is determined by comparing the beginning and end of the lineage. For example, a T→K change is considered on the lineage consisting of branches 9 and 1. A uniquely shared site may not be a convergent- or parallel-change site and vice versa. For instance, when branches 1 and 2 are chosen as the focused lineages, there are parallel changes (S→K) on the lineages, but K is not uniquely shared by the descendants of the focused lineages. When branches 3 and 6 are chosen as the focused lineages, T is uniquely shared by the descendants of the focused lineages, but there is no parallel change on the focused lineages. Note that the tree topology is predetermined and is not inferred just from the amino acids shown in the figure. *B*, A model tree used to explain the statistical tests. The thick lines show the focused lineages on which the convergent and parallel evolution is studied. The arrows show the direction of evolution on the focused lineages. The b_i's are the branch lengths.

When an amino acid substitution model is given, the probability that an amino acid i changes to j along a branch with length b, $P_{ij}(b)$, can be computed (e.g., see Dayhoff, Schwartz, and Orcutt 1978; Yang and Kumar 1996). We denote the configuration of a site by $x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, where x_k is the amino acid of sequence k and can take any of the 20 amino acid states. The probability (p_x) that a site has the configuration x is computed using the following equation.

$$p_x = \pi_{x_2} P_{x_2 x_7}(b_2) P_{x_7 x_6}(b_6) P_{x_6 x_1}(b_1) P_{x_6 x_4}(b_4) P_{x_7 x_8}(b_7) \cdot P_{x_8 x_3}(b_3) P_{x_8 x_5}(b_5), \quad (1)$$

where π_{x_k} is the observed frequency of amino acid x_k in the five present-day sequences. In the above formulation, the tree root was arbitrarily assumed to be at node 2 (fig. 1*B*). However, this does not affect the computation if a time-reversible substitution model is used.

The probability that a site is a convergent-change site (f_c) is the sum of probability of occurrence of all site configurations satisfying the condition $\{x_1 = x_3;\ x_1 \neq x_6$ and $x_3 \neq x_8$; and $x_6 \neq x_8\}$. Therefore,

$$f_c = \sum_{x_1}\sum_{x_2}\sum_{x_3=x_1}\sum_{x_4}\sum_{x_5}\sum_{x_6\neq x_1}\sum_{x_7}\sum_{x_8\neq x_3,x_6} p_x. \quad (2)$$

Since only the amino acids at nodes 1, 3, 6, and 8 and the branch lengths b_1, b_3, b_6, and b_7 affect the f_c, equation (2) can be simplified to

$$f_c = \sum_{x_1}\sum_{x_3=x_1}\sum_{x_6\neq x_1}\sum_{x_8\neq x_3,x_6} \pi_{x_1} P_{x_1x_6}(b_1) \cdot P_{x_6x_8}(b_6 + b_7) P_{x_8x_3}(b_3). \quad (3)$$

If the sequences used are m amino acids long and all sites evolve according to the same substitution model used, the observed number of convergent-change sites (n_c) follows a binomial distribution with the mean and variance equal to mf_c and $mf_c(1 - f_c)$, respectively. So, ϕ, the probability of observing n_c or more convergent-change sites by chance, is given by

$$\phi = \sum_{i=n_c}^{m} \frac{m!}{i!(m-i)!} f_C^i (1 - f_C)^{m-i} = 1 - \sum_{i=0}^{n_c-1} \frac{m!}{i!(m-i)!} f_C^i (1 - f_C)^{m-i}. \quad (4)$$

Equation (4) is applicable when n_c is equal to or greater than 1. When n_c is 0, ϕ is obviously 1. When m is large (e.g., > 30) and mf_c is small (e.g., < 7), this binomial distribution can be approximated by a Poisson distribution with mean and variance both equal to mf_c. Therefore,

$$\phi \approx \sum_{i=n_c}^{m} \frac{e^{-mf_C}(mf_C)^i}{i!} = 1 - \sum_{i=0}^{n_c-1} \frac{e^{-mf_C}(mf_C)^i}{i!}. \quad (5)$$

Similarly, equation (5) is applicable when n_c is equal to or greater than 1. When n_c is 0, ϕ is 1. Thus, if ϕ is smaller than 0.01, we can reject our null hypothesis that the observed convergent changes are simply due to random chance at the 1% significance level.

Similar statistics can be applied to the observed number of parallel-change sites (n_P). In this example, a parallel-change site is a site that satisfies the following conditions: $x_1 = x_3$, $x_1 \neq x_6$, $x_3 \neq x_8$, and $x_6 = x_8$ (see fig. 1*B*).

Above, we discussed only the case where only two lineages are studied and both lineages end at exterior nodes. Our statistical tests apply to more lineages as well as lineages that end at interior nodes.

The computation of the probability that a site is a convergent-change site (f_c) and the probability that a site is a parallel-change site (f_P) requires the information of the branch lengths of the tree and the amino acid substitution patterns. The branch lengths can be estimated by various methods. In this paper, the pairwise gamma distances (with the shape parameter = 2.4; see Zhang and Nei 1997) among the amino acid sequences were computed (Ota and Nei, 1994), and the branch lengths of the tree were estimated by the least-squares method with the restriction that all branches are nonnegative (Felsenstein 1995). Since the estimation of f_c and f_P depends on the substitution model used, we have used three different models. The first model we used was the equal-input model, which assumes that the probability of the substitution from amino acid i to amino acid j is proportional to the frequency of j in the data. The second was the JTT-f model, which was modified from the JTT model (Jones, Tayor, and Thornton 1992) to make the equilibrium frequencies of amino acids equal to the observed frequencies in the data and was shown to be quite good in approximating the evolution of protein sequences (e.g., Cao et al. 1994). The original JTT model is an update of the Dayhoff model (Dayhoff, Schwartz, and Orcutt 1978), which was derived from many protein sequences and can be regarded as an average substitution pattern of all proteins. The third model we used was a general reversal model whose parameters were estimated specifically for the protein sequences used (Yang and Kumar 1996). We refer to this model as the data-specific model in this paper.

Computer Simulations

The efficiencies of the parsimony and Bayesian methods of ancestral sequence inference in identifying convergent and parallel changes were investigated by computer simulations. The parsimony method is the simplest method for inferring ancestral sequences. In this method, each amino acid site is considered separately, and the amino acid at each interior node of the tree is determined so as to make the total number of changes at the site smallest. The pattern of amino acid substitution and the tree branch lengths are not considered in this method. In the Bayesian method of ancestral sequence inference, first the tree branches are estimated, and then the posterior probability of each assignment of amino acids at ancestral nodes is computed at every site by using the Bayesian approach. At each node, the amino acid that has the highest posterior probability is chosen as the ancestral amino acid. There are two versions of the Bayesian method. The difference between them is that in the Yang, Kumar, and Nei (1995) version (also called the maximum-likelihood method), the branch lengths of the tree are estimated by the likelihood method, whereas in the Zhang and Nei (1997) version (also called the distance method), the branch lengths are estimated by the ordinary least-squares method. Computer simulations (Zhang and Nei 1997) showed that these two versions almost always give the same ancestral amino acids. However, the computational time required by the distance method is considerably smaller than that required by the maximum-likelihood method. In this study, the efficiencies of both versions of the Bayesian method were examined.

Computer simulations were conducted by using the model tree given in figure 2*A*. Two different levels of sequence divergence were used with the largest pairwise distances (d_{max}) among the present-day sequences being 0.6 and 1.2 amino acid substitutions per site. The exterior branches leading to sequences 1 and 7 were chosen as the focused lineages. The simulation scheme was

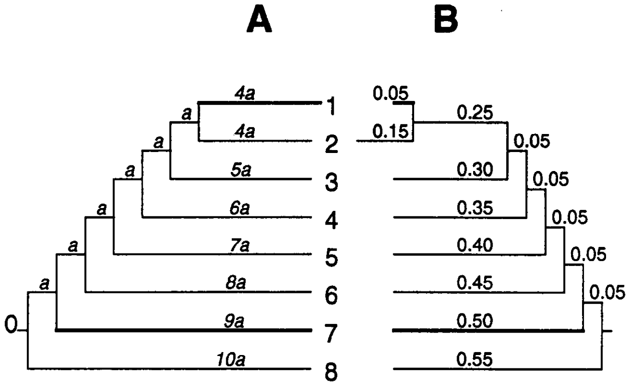


FIG. 2.—Model trees used for studying the efficiencies of the Bayesian and parsimony methods in identifying convergent and parallel changes and for studying the frequencies of convergent-change, parallel-change, uniquely shared, and binary-unique sites. Thick lines show the focused lineages. *A,* The model tree. The *a* values used are 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 amino acid substitutions per site for six different levels of sequence divergence, respectively. The largest pairwise distance (d_{max}) among the present-day sequences is equal to 20*a*. *B,* An example in which the number of uniquely shared sites is larger than the sum of the numbers of parallel-change and convergent-change sites. The numbers of uniquely shared, parallel-change, and convergent-change sites are 0.307, 0.185, and 0.042 per 100 amino acid residues, respectively. Branch lengths are shown above the branches.

as follows: First, a random sequence of 200 amino acids was generated at the tree root, with the expected amino acid frequencies equal to the equilibrium frequencies given in the JTT model. Second, the sequence evolved according to the branching pattern of the tree. Random amino acid substitutions were introduced following the JTT model. The expected number of substitutions per amino acid site for a branch was equal to the branch length in the model tree. Thus, the ancestral amino acid sequences at all interior nodes and the present-day sequences at all exterior nodes were generated and recorded. Third, the ancestral amino acids for all interior nodes were inferred by the parsimony and Bayesian methods, and their efficiencies were assessed. In the parsimony method, there were often multiple reconstructions requiring the same number of amino acid changes at a site. In this case, the fraction of these reconstructions indicating a convergent (or parallel) change on the focused lineages was counted as the number of inferred convergent- (or parallel-) change sites at this site. The total number of convergent- (or parallel-) change sites of the sequence was the summation over all sites. In the Bayesian method, the number of inferred convergent- (or parallel-) change sites was defined as the number of sites at which the reconstructed ancestral amino acids indicated a convergent (or parallel) change on our focused lineages. The simulation was replicated 5,000 times in the case of d_{max} = 0.6 and 1,000 times in the case of d_{max} = 1.2. Note that, in carrying out the statistical tests of convergent and parallel evolution, we were mainly interested in the numbers of convergent-change and parallel-change sites rather than the inferred ancestral amino acids themselves. Therefore, we investigated the correctness of the substitution type (parallel, convergent, or other) instead of the accuracy of inference of the ancestral amino acids, which has been studied by Zhang and Nei (1997).

Results

Efficiencies of the Parsimony and Bayesian Methods in Identifying Convergent and Parallel Changes

The numbers of actual and inferred parallel-change and convergent-change sites per 100,000 sites from the computer simulation are shown in figure 3. Since the two versions of the Bayesian method give virtually the same result, we present the result from the distance-based Bayesian method only. In the Bayesian method, the numbers of inferred parallel-change sites are about 103% (391/381) and 114% (1,032/905) of those of actual parallel-change sites when d_{max} is 0.6 and 1.2, respectively, which means that the estimates are quite accurate. Unfortunately, some non-parallel-change sites were erroneously inferred as parallel and vice versa. For example, when d_{max} = 0.6, 14% (1 − 336/391) of the inferred parallel-change sites have not experienced parallel changes. Some of these sites are actually convergent-change sites, but many are neither convergent nor parallel. The probability of an actual parallel-change site

Bayesian method

d_{max} = 0.6

Actual	Inferred: Parallel	Convergent	Neither	All
Parallel	336	1	44	381
Convergent	11	8	2	21
Neither	44	3		
All	391	12		

d_{max} = 1.2

Actual	Inferred: Parallel	Convergent	Neither	All
Parallel	734	13	158	905
Convergent	78	27	42	147
Neither	220	11		
All	1032	51		

Parsimony method

d_{max} = 0.6

Actual	Inferred: Parallel	Convergent	Neither	All
Parallel	242	37	102	381
Convergent	3	12	6	21
Neither	14	12		
All	259	61		

d_{max} = 1.2

Actual	Inferred: Parallel	Convergent	Neither	All
Parallel	384	144	377	905
Convergent	21	56	70	147
Neither	66	68		
All	471	268		

FIG. 3.—Efficiencies of the Bayesian and parsimony methods in inferring parallel and convergent changes. The rows show the numbers of parallel- and convergent-change sites observed per 100,000 simulated sites, and the columns show the numbers estimated by the inference of ancestral sequences (see fig. 2A for the model tree used). The category "Neither" consists of those sites that are neither parallel nor convergent.

being inferred correctly is about 89% (336/381) and 81% (734/905) when d_{max} is 0.6 and 1.2, respectively.

The numbers of convergent-change sites inferred by the Bayesian method are about 57% (12/21) and 35% (51/147) of the actual numbers for d_{max} of 0.6 and 1.2, respectively, suggesting that the number of convergent-change sites is largely underestimated with this method. Furthermore, even when d_{max} was 0.6, 8% (1/12) of the inferred convergent-change sites were in fact parallel, and 25% (3/12) were neither parallel nor convergent. The probability of a convergent-change site being correctly inferred is only 38% (8/21).

In practice, the pattern of amino acid substitution for a given protein is generally unknown, and a simple substitution model is often used in the analysis. Such applications are known to decrease the accuracy of the Bayesian method (Zhang and Nei 1997). We investigated the efficiency of the Bayesian method in inferring parallel- and convergent-change sites when a simple model is used. For this purpose, we simulated sequence evolution by using the JTT model, but inferred ancestral amino acids according to the Poisson (equal probability for any amino acid substitution) model. The results show that the numbers of inferred and actual parallel-change sites are quite similar, but the efficiency of identification of convergent-change sites becomes even lower.

In the case of the parsimony method, the numbers of inferred parallel-change sites are about 68% (259/381) and 52% (471/905) of the actual numbers when d_{max} is 0.6 and 1.2, respectively. By contrast, the numbers of inferred convergent-change sites are about 290% (61/21) and 182% (268/147) of the actual numbers for the two levels of divergence. These results indicate that the parsimony method largely underestimates the number of parallel-change sites but substantially overestimates the number of convergent-change sites.

These results suggest that ancestral sequence inference by the parsimony method may not be appropriate for estimating the numbers of parallel-change and convergent-change sites. The Bayesian method appears to be useful in estimating the number of parallel-change sites, but it underestimates the number of convergent-change sites. For conducting the statistical test of convergent evolution, use of the Bayesian method is more appropriate than use of the parsimony method, because the test becomes conservative rather than liberal.

Uniquely Shared and Binary-Unique Sites

Without distinguishing between convergent and parallel changes, some authors have assumed that uniquely shared sites have experienced either convergent or parallel changes (e.g., Setewart, Schilling, and Wilson 1987). A site is said to be uniquely shared when the potentially convergent or parallel (present-day) sequences share an amino acid that is not found in other present-day sequences in the data (e.g., see fig. 1*A*). Clearly, whether a site is a uniquely shared site depends largely on the number of sequences in the data. More seriously, the unique share of amino acids is neither sufficient nor necessary for convergence or parallelism (see fig. 1*A* for examples). Therefore, the utility of the uniquely shared sites in studying convergent and parallel evolution needs to be explored.

Goldman (1993) developed an algorithm for identifying parallel-change sites. If we consider the situation where all focused lineages end at exterior nodes, his parallel-change sites are uniquely shared sites where all exterior nodes other than the descendant nodes of the focused lineages share the same amino acid. Since there are only two states at each of these sites, we will call them the binary-unique sites (e.g., a site with $x_1 = x_3 =$ A, $x_2 = x_4 = x_5 =$ S in fig. 1*B*). Although binary-unique sites are mostly parallel-change sites, the reverse is often not true. Furthermore, the binary-unique sites cannot be used for identifying convergent-change sites. The reason is that a binary-unique site has only two different states among the present-day sequences, whereas a convergent-change site usually requires at least three different states, and therefore they are mutually exclusive.

To examine the relationships of the numbers of convergent-change, parallel-change, uniquely shared, and binary-unique sites, we conducted a computer simulation by using the model tree of figure 2*A*. Six different levels of sequence divergence were used, with d_{max} equal to 0.2, 0.4, 0.6, 0.8, 1.0, and 1.2. In this simulation, two exterior branches leading to sequences 1 and 7 were chosen to be the focused lineages, and the JTT model of amino acid substitution was used.

The numbers of convergent-change, parallel-change, uniquely-shared, and binary-unique sites observed per 100 sites are shown in figure 4. The random chance occurrence of convergent and parallel changes increases with the extent of sequence divergence. In general, however, the frequencies of the convergent- and parallel-change sites, particularly the former, are quite low. The number of uniquely shared sites is close to the sum of the numbers of parallel- and convergent-change sites only when the sequence divergence is relatively low ($d_{max} < 0.4$). This means that under this condition, the former is a good estimate of the latter. When the sequence divergence is higher, the number of uniquely shared sites tends to be an underestimate of the total number of parallel- and convergent-change sites. Therefore, many convergent- and parallel-change sites will remain unexplored if only the uniquely shared sites are studied. In fact, when $d_{max} = 1.2$, only 74% of the convergent-change sites and 64% of the parallel-change sites are uniquely shared. Moreover, as mentioned earlier, some uniquely shared sites are neither convergent- nor parallel-change sites (18% in the case of $d_{max} = 1.2$). Note that in this computer simulation, we have used only one model tree, in which evolutionary rates are constant among different lineages. In fact, the number of uniquely shared sites may be greater than the number of parallel- and convergent-change sites. One such example is given in figure 2*B*. At any rate, the number of uniquely-shared sites is expected to be close to the sum of the numbers of parallel- and convergent-change sites only when closely related sequences are studied. As for the number of binary-unique sites, figure 4 shows that

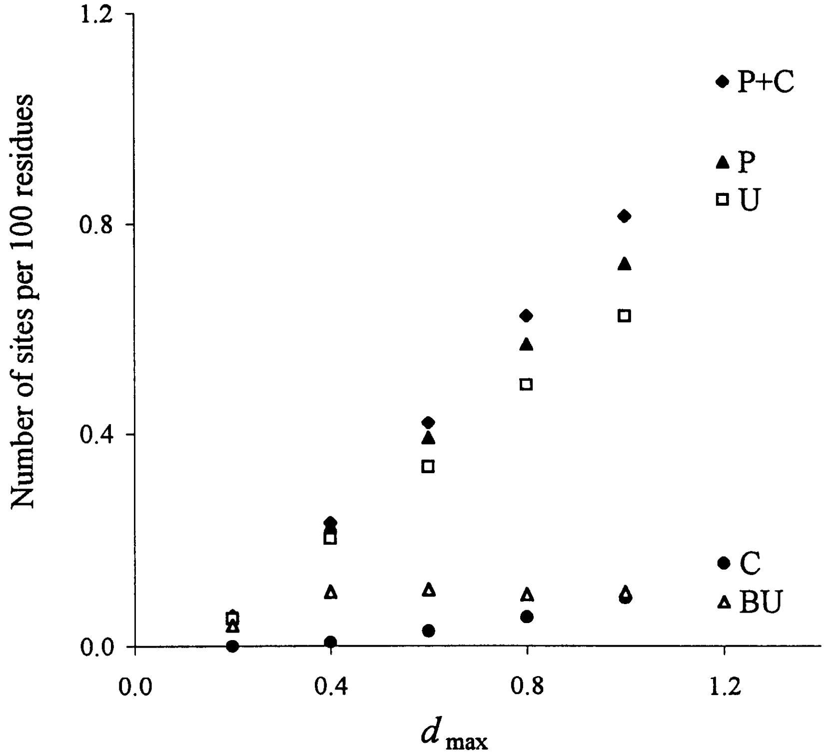


FIG. 4.—Numbers of uniquely shared, parallel-change, convergent-change, and binary-unique sites per 100 amino acid sites observed from a simulation of 100,000 sites (see fig. 2*A* for the model tree used). BU, binary-unique sites; C, convergent-change sites; P, parallel-change sites; P + C, parallel- or convergent-change sites; U, uniquely-shared sites. The d_{max} is the largest pairwise distance (see fig. 2*A*).

it is substantially smaller than that of the parallel-change sites.

Evolution of Stomach Lysozyme Sequences of the Foregut Fermenters: A Case Study

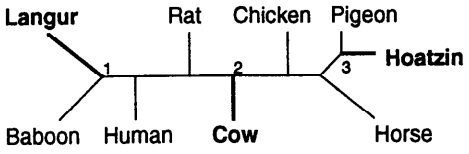
Background Information

The lysozyme of higher vertebrates is normally expressed in macrophages, tears, saliva, avian egg white, and mammalian milk to fight invading bacteria. But in foregut-fermenting organisms such as the ruminants, colobine monkeys, and hoatzins (an avian species), lysozymes have been recruited independently in stomachs to prevent the loss of nutrient assimilated by bacteria that pass through the guts. These stomach lysozymes have similar biochemical properties and functions (Dobson, Prager, and Wilson 1984). Previous studies suggested that the stomach lysozymes have evolved to the same biological function through convergent and parallel evolution at certain amino acid sites (Stewart, Schilling, and Wilson 1987; Kornegay, Schilling, and Wilson 1994). To determine if this is the case, we obtained all the lysozyme *c* sequences available at the time of this study (ENTREZ, release 18) and reconstructed a phylogenetic tree of these sequences by the neighbor-joining method (Saitou and Nei 1987). We found that the tree topology was not stable and might change with the number of sequences used (see also Adachi and Hasegawa 1996). We then selected stomach lysozyme sequences of the langur (*Presbytis entellus*), cow (*Bos taurus*), and hoatzin (*Opisthocomus hoatzin*) and nonstomach lysozyme sequences of the human (*Homo sapiens*), baboon (*Papio cynocephalus*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), pigeon (*Columba livia*), and horse (*Equus caballus*) for primary analysis because previous studies of convergent and parallel evolution have been based on the analyses of these sequences (Stewart, Schilling, and Wilson 1987; Kornegay, Schilling, and Wilson 1994). Our phylogenetic analysis suggested that these lysozyme genes are not orthologous, but this does not affect the statistical tests as long as the gene tree is correct. The phylogenetic tree of the nine selected lysozyme sequences is given in figure 5*A*, which is derived from our phylogenetic analysis of all available lysozyme *c* sequences. Note that this tree is similar in topology to those used in previous studies (Stewart, Schilling, and Wilson 1987; Kornegay, Schilling, and Wilson 1994; Adachi and Hasegawa 1996). We removed all sites containing alignment gaps, and the final sequence length was 124 amino acids.

Statistical Tests of Convergent and Parallel Evolution

The evolution of the new function of the stomach lysozymes is thought to have occurred independently in the langur, cow, and hoatzin lineages. Therefore, we focused our attention on the three lineages: from node 1 to langur, from node 2 to cow, and from node 3 to hoatzin (fig. 5*A*). In the nine lysozyme sequences analyzed, there were two parallel-change sites (sites 75 and 87) but no convergent-change sites identified by the Bayesian method (fig. 5*B*). Our statistical test shows that the

A



B

Site	14	21	41	75	76	83	87	126
Human	R	R	R	N	A	A	D	Q
Langur	**K**	**K**	**E**	**D**	**A**	**A**	**N**	**K**
Node 1	r	r	q	n	a	a	d	q
Baboon	R	R	Q	N	A	A	D	Q
Rat	R	Y	Q	N	A	A	D	R
Cow	**K**	**K**	**K**	**D**	**G**	**E**	**N**	**E**
Node 2	r	r	q	n	a	a	d	r
Chicken	R	R	Q	N	L	A	S	R
Pigeon	R	V	T	N	A	K	D	R
Hoatzin	**E**	**E**	**E**	**D**	**G**	**E**	**N**	**K**
Node 3	r	e	q	n	a	k	d	r
Horse	A	G	R	N	A	K	E	A

FIG. 5.—Parallel and convergent amino acid substitutions in the stomach lysozymes. *A*, The phylogenetic relationships of the nine vertebrate lysozymes. The focused lineages are shown by thick lines. Note that the gene tree is not identical to the species tree because some of the genes are paralogous. The branch lengths of the tree are not proportional to the extent of sequence divergence. *B*, The identified parallel-change and convergent-change sites of the stomach lysozymes in the two-lineage and three-lineage comparisons. Site positions are according to the human lysozyme sequence. Amino acid residues are indicated by single-letter symbols, where uppercase letters denote present-day sequences and lowercase letters denote ancestral sequences inferred by the Bayesian method. The ancestral amino acids with probabilities lower than 80% are underlined. The foregut fermenters and their stomach lysozyme sequences are shown in bold type.

observed number of parallel-change sites is significantly greater than the random chance expectation (table 1).

We also tested parallel and convergent evolution in every pair of the three stomach lysozyme sequences mentioned above. In each pair comparison, the third stomach lysozyme sequence was removed from the analysis. Table 1 shows that the number of parallel-change sites is significantly greater than the expectation in the cow–langur comparison and the langur–hoatzin comparison. In the cow–hoatzin comparison, the significance level is marginal ($\phi = 0.04$). There is a convergent-change site identified in the cow–hoatzin comparison and another in the langur–hoatzin comparison, but the former is not significantly greater than the expectation and the latter is only marginally significant ($\phi = 0.05$).

In our analysis, we considered a convergent- or parallel-change site only when the descendant amino acids in independent lineages were identical. Although some amino acid residues may be functionally equivalent because of similar biochemical properties, and convergence and parallelism should not be restricted to sites with the same descendant states, it is usually difficult to know which two amino acids at a particular site would be functionally similar without extensive experimental studies.

In the statistical tests, we have used three different models of amino acid substitution: the equal-input, JTT-f, and data-specific models. Since all these models have some similar assumptions, such as time reversibility and constant amino acid frequencies in evolution, it is necessary to inspect whether these models fit the data at least with respect to the numbers of parallel-change and convergent-change sites. As a negative control, we examined the numbers of these two types of sites by comparing lysozyme sequences that are not likely to be under parallel or convergent evolution (e.g., cow–baboon, pigeon–langur, pigeon–rat pairs). We found that

Table 1
Tests of Convergent and Parallel Evolution of Stomach Lysozyme Sequences of the Cow, Langur, and Hoatzin

	Expected Number of Sites			Observed Number[a]	Probability ϕ		
Model Used	EI[c]	JTT-f	DS[d]	(Site Positions[b])	EI	JTT-f	DS
Cow, langur, and hoatzin comparison							
Parallel-change	0.000[e]	0.011	0.019	2 (75, 87)	<0.001	<0.001	<0.001
Convergent-change	0.000	0.000	0.000	0	1	1	1
Cow and langur comparison							
Parallel-change	0.110	0.299	0.374	4 (14, 21, 75, 87)	<0.001	<0.001	<0.001
Convergent-change	0.006	0.007	0.010	0	1	1	1
Cow and hoatzin comparison							
Parallel-change	0.280	0.716	0.755	3 (75, 76, 87)	0.003	0.036	0.041
Convergent-change	0.104	0.109	0.148	1 (83)	0.099	0.103	0.138
Langur and hoatzin comparison							
Parallel-change	0.072	0.191	0.204	3 (41, 75, 87)	<0.001	0.001	0.001
Convergent-change	0.037	0.039	0.051	1 (126)	0.036	0.038	0.050

[a] Estimated by comparing the present-day sequences and the ancestral sequences inferred by the Bayesian method.
[b] Site positions are according to the human lysozyme sequence.
[c] Equal-input model.
[d] Data-specific model.
[e] Smaller than 0.0005.

when the JTT-f or data-specific model is used, there is no case where the observed number of parallel- or convergent-change sites is significantly greater than the corresponding expectation. This result suggested that the use of the JTT-f and data-specific models is appropriate.

Parallel Evolution of Stomach Lysozyme Sequences of Foregut Fermenters

An examination of the lysozyme sequences of the nine species revealed that two sites (sites 75 and 87) have experienced parallel substitutions in the evolutionary lineages of the foregut fermenters (cow, langur, and hoatzin). Statistical tests suggested that these two sites have evolved with unusual substitution patterns which could be due to positive selection. The lysozyme *c* sequences of many mammal and bird species are known. Particularly, stomach lysozyme sequences of the advanced ruminants goat (*Capra hircus*), sheep (*Ovis aries*), and deer (*Cervus axis*) have been determined (Jollès et al. 1990; Irwin and Wilson 1990). (Recently, multiple copies of stomach lysozyme genes were found in the advanced ruminants and hoatzin [Jollès et al. 1990; Irwin and Wilson 1990; Kornegay 1996]. However, these genes are subject to concerted evolution, so the stomach lysozyme sequences within species are very similar.) If sites 75 and 87 have actually undergone adaptive evolution, we expect to see the same amino acids at these sites in all stomach lysozymes, but different amino acids in any other lysozyme. We find that at site 75, there is a D (Asp) in all of the stomach lysozymes, but an N (Asn) in all of the 34 nonstomach lysozymes examined. At site 87, all stomach lysozymes have an N; other lysozymes have an A (Ala), D, E (Glu), or S (Ser). Amino acid residue A is neutral and hydrophobic, D and E are acidic, and N and S are neutral and polar. It seems that site 87 can have a variety of amino acids with different physiochemical properties in nonstomach lysozymes, but in the stomach lysozymes, this site seems to be invariant, because only the N, one of the seven neutral and polar amino acids, is observed. This suggests that the variability of a site can change when the function of the protein shifts. From these analyses, it appears that sites 75 and 87 of the stomach lysozymes have evolved adaptively under positive selection.

In addition to sites 75 and 87, other parallel-change sites in pairwise comparisons of the three stomach lysozymes are observed (table 1). Are they also adaptive? To address this question, we first removed sites 75 and 87 from the sequences and then tested whether parallel evolution was still significant in pairwise comparisons. We found that now the null hypothesis could not be rejected at the 1% level. So, positive selection is not necessary to explain the evolution of stomach lysozymes except for sites 75 and 87. At some sites, the parallel or convergent changes probably occurred just by chance. For example, parallel changes from R (Arg) to K (Lys) were inferred in the cow and langur lineages at site 14. But this parallel substitution does not seem to have resulted from adaptive evolution, since the rabbit lysozyme also has K (Ito et al. 1990), whereas the deer stomach lysozyme has an E (Irwin and Wilson 1990). Since amino acid R changes to K with a relatively high probability (Jones, Taylor, and Thornton 1992), it is possible that this parallel change occurred in the cow and langur lineages just by chance.

Nevertheless, even after we removed sites 75 and 87, the numbers of parallel-change sites are still larger than the random chance expectations in all pairwise comparisons (although not significantly so). It is possible that some of these sites are subject to positive selection. For example, site 76 is reconstructed as a parallel-change site in the cow–hoatzin comparison. This site may have undergone adaptive evolution, since only in the ruminant and hoatzin stomach lysozymes does it have a G (Gly). In nonstomach lysozymes, there is an A, L (Lys), or V (Val) at this site, all of which are hydrophobic, whereas G is polar. But the foregut fermenter langur also has an A at this site, as many nonforegut fermenters do. Probably, the stomach lysozymes of the ruminants and hoatzin have some common properties that the langur lysozyme does not possess. It is possible that the functions of the stomach lysozymes in the three groups of foregut fermenters are not exactly the same.

In the analysis, we have assumed that the ancestral amino acids inferred by the Bayesian method are correct. In fact, the reliability of some of the inferred amino acids at the convergent- and parallel-change sites is not very high (<80%). More present-day sequences seem to be necessary to increase the accuracy of the inference of the ancestral amino acids. Also note that the number of convergent-change sites identified in this analysis is a conservative estimate due to the properties of the Bayesian method of ancestral sequence inference.

Discussion

Difficulties in Identifying Parallel and Convergent Changes

Our simulation results suggest that the Bayesian method of ancestral amino acid inference is generally accurate in estimating the number of parallel-change sites. For convergent-change sites, neither the Bayesian nor the parsimony method is efficient: the Bayesian method tends to underestimate whereas the parsimony method tends to overestimate the number of convergent-change sites. In this case, the Bayesian method is recommended because it will make the statistical test of convergent evolution conservative. Another advantage of the Bayesian method is that the reliability of the inferred ancestral amino acids can be evaluated.

If sequence convergence is the basis of functional convergence, the sites that have experienced convergent evolution are likely to be crucial for the protein function. Convergence implies that there is more than one possible amino acid state at a site before the occurrence of convergent evolution, but only one state after the convergence. In other words, the variability of the site changes when the protein function changes. Although possible, we think that the variability of a site rarely changes unless the sequences are highly diverged (e.g.,

Miyamoto and Fitch 1995). Parallel evolution, however, occurs more easily because the site is conservative before as well as after the parallel changes have occurred. Interestingly, convergent evolution was claimed more often than parallel evolution in previous studies, even though both were rare. This is probably due to use of the parsimony method, which overestimates the number of convergent-change sites and underestimates the number of parallel-change sites; claims of convergent evolution without distinguishing it from parallel evolution are another possible reason.

In our computer simulations, all sites in an amino acid sequence are assumed to have the same substitution pattern. In reality, the substitution pattern of a site undergoing adaptive evolution is likely to be different from that of other sites. This makes the identification of convergent-change sites more difficult.

Factors Affecting the Performance of the Statistical Tests

The statistical tests of convergent and parallel evolution depend on the estimates of the probability of a site being a convergent-change site (f_c) and the probability of a site being a parallel-change site (f_P), which are affected by the amino acid substitution model used. We have used the equal-input, JTT-f, and data-specific models in computing f_c and f_P for the lysozyme sequences. The substitution model adopted affected these estimates substantially (2–3 times for two-lineage comparisons and 2–10 times for the three-lineage comparison, see table 1). The estimates of f_c and f_P under the data-specific model are similar to those under the JTT-f model. This suggests that the JTT-f model is sufficient for estimating f_c and f_P, especially when we consider large errors of the data-specific model. The estimates under the equal-input model are much smaller, indicating that the statistical tests may become liberal if these estimates are used. Another factor that affects the computation of f_c and f_P is the evolutionary rate variation among sites. Under the JTT-f model, the evolutionary rate at a site depends only on its amino acid state; however, in reality, different sites with identical amino acid residue may evolve with different rates. This is expected to increase the f_c and f_P. The gamma distribution has been used to approximate the rate variation among sites (see Yang 1996 for a review). Zhang and Nei (1997) found that the number of amino acid substitutions per site between two sequences which have evolved under the JTT model can be approximated by a gamma distance with the shape parameter equal to 2.4, which is a little smaller than that for the lysozyme sequences (2.8; estimated by Yang and Kumar's [1996] method). Therefore, the use of the JTT-f model for the lysozyme sequences is unlikely to underestimate the f_c and f_P. But when the gamma shape parameter is very small (i.e., large rate variation among sites), the JTT-f-gamma model will be better. In this case, however, the $P_{ij}(b)$ of equation (1) is hard to obtain analytically, but a simulation approach as described in the *Methods* section can be used to obtain the f_c and f_P values directly. In a small-scale simulation study using the tree in figure 2A, we found that use of the JTT-f-gamma model with the shape parameter equal to 0.5 resulted in about a two-fold increase in f_c and f_P, as compared to the case of the simple JTT-f model. Moreover, ignoring rate variation among sites leads to underestimation of the tree branch lengths, which may largely decrease the estimated values of f_c and f_P. In the analysis of the lysozyme sequences, the tree branch lengths were estimated by using the gamma distances with the shape parameter equal to 2.4, so the branch lengths are unlikely to be underestimated for the lysozyme tree. In practice, one has to be cautious about the influence of the rate variation among sites on the statistical tests.

Parallel and Convergent Amino Acid Changes and the Role of Positive Darwinian Selection

In the computer simulations and in the analysis of the lysozyme data, we found that the expected numbers of parallel- and convergent-change sites are rather small. This means that the statistical tests of parallel and convergent evolution are expected to be quite powerful because the presence of a few parallel- or convergent-change sites could lead to the rejection of the null hypothesis of random chance. However, the rejection of the null hypothesis does not necessarily indicate positive selection at these convergent- or parallel-change sites. An excess of convergent and parallel-change sites may be observed because of the use of an inadequate substitution model, as discussed before. For instance, if either amino acid Asn or Asp is necessary at a certain site to keep the protein functional, this site is likely to be identified as a parallel-change site even if the substitutions between the Asn and Asp are neutral to the function of the protein. This is because parallel substitutions from one amino acid to the other easily occur if there are only two possible states at the site. However, if this is the case, we are also likely to see this kind of parallel change on all lineages, rather than just on the focused lineages (e.g., see Well 1996). By using other lineages as a negative control, we may be able to identify such sites and remove them from the analysis. We analyzed the lysozyme sequences using this strategy and did not find any such site.

However, it is generally difficult to know which substitution model is appropriate for every site of a given protein and the use of a simple model is common in practice. This may underestimate f_c and f_P, as discussed earlier. Therefore, to be more conservative, we recommend that the 1% (instead of 5%) significance level be used to reject the null hypothesis in the statistical tests. Also, the inferred ancestral amino acids with low probabilities should be used with caution.

Since the expected numbers of parallel-change or convergent-change sites are usually less than 1 per 100 amino acids, the presence of a few such sites in a not-very-long protein sequence would indicate at least a different substitution pattern at these sites from that assumed in the analysis, and probably imply real parallel and convergent evolution under positive selection. Although direct evidence of the fitness change corresponding to the amino acid substitution is necessary to finally

demonstrate positive selection, the identified parallel and convergent changes provide potential examples of positive selection and adaptive evolution, which can be tested in the future.

Program Availability

A program for computing f_c and f_P is available on request.

Acknowledgments

We are grateful to M. Nei for stimulating discussions and to M. Nei, T. Sitnikova, X. Gu, Y. Ina, and two anonymous reviewers for their comments. This work was supported by NIH and NSF research grants to M. Nei.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. Computer science monographs, no. 28: MOLPHY: programs for molecular phylogenetics based on maximum likelihood. Version 2.3. Institute of Statistical Mathematics, Tokyo.

CAO, Y., J. ADACHI, A. JANKE, S. PÄÄBO, and M. HASEGAWA. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39**:519–527.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC.

DOBSON, D. E., E. M. PRAGER, and A. C. WILSON. 1984. Stomach lysozymes of ruminants. I. Distribution and catalytic properties. J. Biol. Chem. **259**:11607–11616.

DOOLITTLE, R. F. 1994. Convergent evolution: the need to be explicit. Trends Biochem. Sci. **19**:15–18.

FELSENSTEIN, J. 1995. PHYLIP: phylogeny inference package. Version 3.57c. University of Washington, Seattle.

GOLDMAN, N. 1993. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. **37**:650–661.

HARTIGAN, J. A. 1973. Minimum evolution fits to a given tree. Biometrics **29**:53–65.

IRWIN, D. W., and A. C. WILSON. 1990. Concerted evolution of ruminant lysozymes: characterization of lysozyme cDNA clones from sheep and deer. J. Biol. Chem. **265**:4944–4952.

ITO, Y., H. YAMADA, S. NAKAMURA, and T. IMOTO. 1990. Purification, amino acid sequence, and some properties of rabbit kidney lysozyme. J. Biochem. **107**:236–241.

JOLLÈS, J., E. M. PRAGER, E. S. ALNEMRI, P. JOLLÈS, I. M. IBRAHIMI, and A. C. WILSON. 1990. Amino acid sequences of stomach and nonstomach lysozymes of ruminants. J. Mol. Evol. **30**:370–382.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

KORNEGAY, J. R. 1996. Molecular genetics and evolution of stomach and nonstomach lysozymes in the hoatzin. J. Mol. Evol. **42**:676–684.

KORNEGAY, J. R., J. W. SCHILLING, and A. C. WILSON. 1994. Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. Mol. Biol. Evol. **11**:921–928.

KREITMAN, M., and H. AKASHI. 1995. Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26**:403–422.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade: analysis of phylogeny and character evolution. Version 3. Sinauer, Sunderland, Mass.

MIYAMOTO, M. M., and W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. **12**:503–513.

NEI, M. 1990. DNA polymorphism and adaptive evolution. Pp. 128–142 *in* A. H. D. BROWN, T. CLEGG, A. L. KAHLER, and B. S. WEIR, eds. Plant population genetics, breeding, and genetic resources. Sinauer, Sunderland, Mass.

———. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. **30**:371–403.

OTA, T., and M. NEI. 1994. Estimating the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. **38**:642–643.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for econstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

STEWART, C.-B., J. W. SCHILLING, and A. C. WILSON. 1987. Adaptive evolution in the lysozymes of foregut fermenters. Nature **330**:401–404.

SWANSON, K. W., D. M. IRWIN, and A. C. WILSON. 1991. Stomach lysozyme gene of the langur monkey: tests for convergence and positive selection. J. Mol. Evol. **33**:418–425.

WELL, R. S. 1996. Excessive homoplasy in an evolutionarily constrained protein. Proc. R. Soc. Lond. B **263**:393–400.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. **11**:367–372.

YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**: 650–659.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

ZHANG, J., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. (in press).

SHOZO YOKOYAMA, reviewing editor

Accepted January 29, 1997