

# Evaluation of Methods for Determination of a Reconstructed History of Gene Sequence Evolution

David A. Liberles

Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

With whole-genome sequences being completed at an increasing rate, it is important to develop and assess tools to analyze them. Following annotation of the protein content of a genome, one can compare sequences with previously characterized homologous genes to detect novel functions within specific proteins in the evolution of the newly sequenced genome. One common statistical method to detect such changes is to compare the ratios of nonsynonymous ( $K_a$ ) to synonymous ( $K_s$ ) nucleotide substitution rates. Here, the effects of several parameters that can influence this calculation (sequence reconstruction method, phylogenetic tree branch length weighting, GC content, and codon bias) are examined. Also, two new alternative measures of adaptive evolution, the point accepted mutations (PAM)/neutral evolutionary distance (NED) ratio and the sequence space assessment (SSA) statistic are presented. All of these methods are compared using two sequence families: the recent divergence of leptin orthologs in primates, and the more ancient divergence of the deoxyribonucleoside kinase family. The examination of these and other measures to detect changes of gene function along branches of a phylogenetic tree will become increasingly important in the postgenomic era.

## Introduction

In the course of evolutionary history, genes and their regulatory regions undergo a variety of modifications, including nucleotide substitution, duplication, recombination, and repair events. Understanding the diversity of both life on Earth and the genomic sequences that define it requires the development of methods to accurately reconstruct the evolutionary history of these events based on an understanding of the mechanisms of molecular evolution. As new genomes are sequenced, these tools are valuable both in annotation of protein functions and in correlating organismal genotype with phenotype. Sequence analysis tools that are computationally simple and fast as well as accurate are necessary for the creation of databases that take full advantage of the power of computational genomics (Pollack et al. 2000; Liberles et al. 2001).

Coding-sequence nucleotide substitution is an important mechanism driving the differential adaptive evolution of species and has been studied most notably in hemoglobin, ribonuclease, and lysozyme (Messier and Stewart 1997; reviewed in Golding and Dean 1998). The most common approach for identifying branches of phylogenetic trees undergoing adaptive evolution is to measure the ratio of nonsynonymous ( $K_a$ ) to synonymous ( $K_s$ ) nucleotide substitution rates for the branches. Nonsynonymous mutations are mutations to DNA that result in amino acid changes in the encoded protein and are ultimately selected for or against based on their effect on the organism's ability to reproduce. Synonymous mutations are mutations to DNA that do not change the encoded amino acid and reflect the opportunity to explore sequence space in the absence of selection.  $K_a/K_s$

is a measure of accepted mutations normalized for opportunity. Because extant proteins have been selected for function for millions of years, high  $K_a/K_s$  is generally reflective of changed function and adaptive evolution (or at least reduced functional constraint).

While theoretically positive selection occurs when  $K_a \gg K_s$ , the few instances of observation of functional change along branches of phylogenetic trees indicate that it frequently occurs at  $K_a/K_s < 1$  (Crandall et al. 1999; Almgren 2001; unpublished observations). This is because substitution rates measured along branches of phylogenetic trees can average multiple evolutionary periods and because protein-folding selective constraints are different on different residues. Therefore, proteins can undergo adaptive evolution where a fraction of amino acids have  $K_a/K_s$  ratios of  $\gg 1$  for a fraction of the period reflected in a branch, and this can result in  $K_a/K_s$  ratios of  $< 1$  (see Gould and Eldredge [1993] for a possible evolutionary mechanism). Furthermore, the relaxation of selective constraints (neutral evolution) that is specific to a subset of branches on a tree can also result in changes of protein function and is important to detect when correlating molecular events with organismal phenotypes. Such changes of protein function can be adaptive at the organismal level. More benchmarking of  $K_a/K_s$  ratios with real protein families where function is known needs to be done.

Many methods are available to calculate the  $K_a/K_s$  ratio, ranging from the simpler method of Nei and Gojobori (1986) to the more computationally intensive, highly parameterized maximum-likelihood method of Yang and Nielsen (2000). In a comparison of methods (Ina 1995), the PBL method of Pamilo, Bianchi, and Li (Li, Wu, and Luo 1985; Pamilo and Bianchi 1993) performed well and remains one of the more popular methods. This method has recently been modified by Benner, Trabesinger, and Schreiber (1998) using an implementation of ancestral sequence reconstruction and is hereinafter referred to as the PBLSB method. This allows one to compare  $K_a/K_s$  ratios along branches of a phy-

Key words: nonsynonymous nucleotide substitution, synonymous nucleotide substitution, PAM distance, adaptive evolution, reconstructed ancestral sequences.

Address for correspondence and reprints: David A. Liberles, Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden. E-mail: liberles@sbc.su.se.

*Mol. Biol. Evol.* 18(11):2040–2047. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

logenetic tree instead of using a simple pairwise comparison of sequences.

The value of the PBLSB method lies in its ability to pinpoint events to specific branches of a phylogenetic tree and its computational speed, allowing exhaustive analysis of genomes and large data sets (see Liberles et al. 2001). While likelihood-based methods frequently perform better than parsimony-based methods (Zhang and Nei 1997), they are computationally slow for all but small data sets.

In studies performed with simulated data sets (e.g., Koshi and Goldstein 1996; Zhang and Nei 1997; unpublished data), ancestral sequence reconstruction is not always accurate (not even when maximum-likelihood methods are used), especially as the branch lengths under study increase. While current research with simulated data attempts to delineate when such sequences can be trusted and also to improve methodology, it is valuable to test when currently available simple methods detect changes of protein function for real data sets. The performance of different evolutionary methods with different degrees of parameterization on real data sets was analyzed here to test the importance of these methods in influencing the measurement of evolutionary events. The following new (methods 2–10) and classical (method 1) methods were tested in both a recent and a more ancient evolutionary case study: different methods for reconstructing ancestral sequences, different theoretical and empirical substitution matrices (methods 2 and 8–10), parameterization for codon bias (method 7) and GC bias (methods 5 and 6), and two new measures of adaptive evolution (methods 11 and 12).

## Materials and Methods

The ability of 12 different methods to reconstruct evolutionary histories using sequences found in families in the Master Catalog was evaluated.

### The Master Catalog

All multiple-sequence alignments and phylogenetic trees were obtained from the Master Catalog (Benner et al. 2000), a database of alignments and trees for each independently evolving protein family found in GenBank. Access can be obtained through EraGen Biosciences (Alachua, Fla.) at <http://www.eragen.com>.

The methods differed in their treatment of the parameters discussed below. Methods 2–4 represented modifications of the Fitch algorithm to calculate ancestral sequences, while methods 5–10 represented modifications of the PBL method to calculate  $K_a$  and  $K_s$ . Methods 11 and 12 are alternatives to the  $K_a/K_s$  ratio.

### Reconstruction (methods 1 and 2)

Two methods for reconstructing ancestral sequences were utilized. Both were based on the Fitch (1971) algorithm. While in its simplest form, the Fitch algorithm is not as accurate as maximum-likelihood reconstruction (Zhang and Nei 1997), it is much faster and more accessible for large data sets. The DNA reconstruction uses a

Fitch reconstruction of only DNA sequences as performed by Benner, Trabesinger, and Schreiber (1998). A new parsimony approach is introduced here where DNA and protein parsimonies are calculated simultaneously using the Fitch algorithm (DNA+protein). Then, for ambiguous positions, the intersection of the translated codon and amino acid reconstructions is taken as the reconstructed ancestral sequence. The DNA reconstruction is used when no amino acids are found in common. Probabilities are assigned to codons instead of to nucleotides (as is done by Benner, Trabesinger, and Schreiber 1998) starting from the root.

### Branch Length Weighting (methods 2–4)

When weighting is set to none, probabilities of ancestral sequences are assigned using the Fitch algorithm. Ignoring branch lengths is thought to be one of the main reasons that maximum likelihood outperforms the Fitch algorithm (Zhang and Nei 1997). Therefore, one of the modifications of the Fitch algorithm tested here involves reweighted parsimonious ambiguities based on branch length distances. Initial unweighted probabilities are calculated using the Fitch algorithm. These probabilities are then reweighted from the three connecting branches according to the probabilities of sequences at each neighboring node (using the initial unweighted probabilities) divided by the branch length distance measured in either point accepted mutations (PAM) or neutral evolutionary distance (NED) (a distance based on an approach to equilibrium calculation of synonymous two-fold-degenerate pyrimidine transition codon substitutions, as in Peltier et al. 2000). PAM distances are calculated according to standard methods.

### Substitution Matrix (methods 2 and 8–10)

A discrete form of the Grantham matrix is used in the PBLSB method to weight mutational path probabilities when multiple substitutions occur along a branch (method 2) (Grantham 1974; Li, Wu, and Luo 1985; Pamilo and Bianchi 1993). Here, the discrete Grantham matrix and three additional substitution matrices were utilized. A continuous variation of the Grantham matrix is introduced (method 8), where instead of rounding to the nearest value of 50, substitution likelihoods are divided by the Grantham value. Another discrete, but simpler theoretical matrix, the Zhang matrix (method 10) is used in a fashion analogous to that of the discrete Grantham matrix (Zhang 2000). Finally, an empirical substitution matrix, the Taylor-Jones matrix, was tested as well (method 9) (Taylor and Jones 1993).

### GC Content Correction (methods 5 and 6)

$K_s$  may be affected by a selective pressure on GC content in third positions making synonymous third-position substitution nonneutral (a selectionist model of GC content differences proposes that mutations occur and are then subject to selective pressures on local GC content to remove a subset of these mutations). These selective pressures cause an underestimation of  $K_s$ ,

which was corrected using two measures of GC selective pressure. The selective pressure of GC content bias can be inferred from the GC content (Nei 1987). Correction for global GC content (method 5) uses the value determined from third positions of all coding sequences known in a given genome (<http://www.kazusa.or.jp/codon/>). Correction for local GC content (method 6) uses the value of GC content calculated from third positions in the gene of interest.

#### Codon Bias Correction (method 7)

Correction for codon bias (the nonrandom utilization of different codons encoding the same amino acid) is performed by multiplying the substitution probabilities in the PBL method in each case by the normalized codon usage value (CB), where

$$CB = CD \times CU / \left( \sum_{1 \leq i \leq CD} CU_i \right) \quad (1)$$

where CD is codon degeneracy (the number of codons that encode an amino acid), CU is codon usage for the encoded codon, and  $\sum_{1 \leq i \leq CD} CU_i$  is the sum of usages of all codons that encode the same amino acid. CU values are calculated from reported values (<http://www.kazusa.or.jp/codon/>). This correction is also made to  $K_s$  for each synonymous codon substitution.

#### PAM/NED Ratio (method 11)

The PAM/NED ratio is calculated simply by dividing the PAM distance of a branch by the NED distance along the same branch.

#### Sequence Space Assessment Statistic (method 12)

The sequence space assessment (SSA) statistic is a measure of the difference between the number of amino acid sites that undergo mutation along a branch (pairwise between branch termini) and the number of those that are found to be variant throughout the sequence in at least one protein in the tree, normalized for the number of taxa. The normalization equation is taken from Tajima's  $D$  statistic, for which pairwise variation is compared with the number of polymorphic sites in an interbreeding population. Here,

$$SSA = (AA_{pw} - AA_{clade}/a1) \div (e1 \times AA_{clade} + e2 \times AA_{clade} \times (AA_{clade} - 1))^{0.5}, \quad (2)$$

where  $AA_{pw}$  is the number of pairwise amino acid substitutions along a branch,  $AA_{clade}$  is the total number of sites at which amino acid substitution has occurred at least once, and  $a1$ ,  $e1$ , and  $e2$  are defined according to Tajima (1989) to correct for the size of a clade. Chemically, what this value represents the proportion of amino acid positions that undergo substitution along a specific branch that are available to the protein for mutation dictated by the specific folding requirements of that protein. In practice, SSA gives trends similar to those from

simply dividing  $AA_{pw}$  by  $AA_{clade}$  (data not shown). This statistic is ideal for proteins with folds that are evolving in a stationary gamma process rather than in a covarion process (see Miyamoto and Fitch [1995] and Chelvanayagam et al. [1997] for a discussion of this issue).

#### Robustness and NED tests

The same robustness test and the same NED test used in the Master Catalog (Benner et al. 2000) were applied here. The robustness test eliminates very short branch lengths with fractional mutations. The NED test eliminates long branches with equilibrated third-position NED codons. The NED test involves a test of saturation or equilibration of twofold-degenerate pyrimidine transition codons along a branch (Smith and Smith 1996; Peltier et al. 2000; Liberles et al. 2001). Due to the inaccuracies of ancestral sequence reconstruction, along-branch NED distances may be underestimated on extremely long branches, as indicated by preliminary simulations (including some of those in fig. 2) (unpublished data). The implications of this are discussed below.

## Results and Discussion

The Fitch (1971) algorithm for generating ancestral sequences from a phylogeny has been modified by Benner and combined with the PBL method for calculating the ratio of nonsynonymous to synonymous nucleotide substitutions (known here as the PBLSB method) to produce evolutionary histories for every family found in GenBank (Benner et al. 2000; Liberles et al. 2001). This computationally simple and fast approach is an extremely valuable starting point for the exploration of the roles of specific proteins in the adaptive evolution of species.

The Benner application of the Fitch method utilizes nucleotide sequences as the basis for the reconstruction of ancestral sequences. A new method is presented here that simultaneously reconstructs nucleotide and amino acid sequences. The nucleotide sequences are translated in ambiguous positions, and the intersection of the two sets of sequences from the two simultaneous reconstructions is adopted. If there is no intersection, then the nucleotide-sequence-based reconstruction is used. Probabilities are assigned from the root, with equal probabilities assigned to codons instead of to nucleotides. This approach is expected to more accurately measure evolutionary intermediates that are tolerated by natural selection and are more parsimonious with respect to protein sequence. Variations of the Fitch algorithm were also tested that use various measures of branch length distances based on the initially assigned Fitch probabilities. The absence of branch length weighting in the original implementation of the Fitch algorithm may account for its poor performance (Zhang and Nei 1997). PAM distance as a measure of protein distance and NED as a measure of synonymous nucleotide distance were used for branch length weighting. Branch length weighting is expected to remove some of the biases associated with parsimony reconstructions.

Second, the PBL method for calculating  $K_d/K_s$  ratios was examined. The usefulness of this ratio hinges

on  $K_s$  being a truly neutral measure of biological evolution (which  $K_s$  may not be in many cases). When  $K_s$  is subject to selective pressures, it no longer accurately reflects the amount of substitution available to a gene. Examination of codon usage tables in mammals indicates that not all codons are used equally, and this may result in a bias in nucleotide substitution (<http://www.kazusa.or.jp/codon/>). Normalized nucleotide substitution matrices were multiplied by the path probability in the PBL method as a correction for  $K_a$ .  $K_s$  was also corrected directly for selective pressure due to codon bias.

GC content bias may also have an effect on the determination of  $K_s$ . GC percentage globally within a genome may be dictated by the physiological temperature of an organism. Locally, the isochore structure of a gene may exert a selective pressure on  $K_s$  (Matassi, Sharp, and Gautier 1999). Here, the selective force of GC content bias (Nei 1987), measured either globally throughout the genome or locally within the homologous gene set, was used to correct  $K_s$ .

Codon bias and GC content bias are well known, both theoretically and from simulations, to influence the calculation of  $K_a/K_s$  ratios (see, e.g., Smith 1994). The prevalence of these effects on genes in various species is not known. Their effects can be ascertained only if the calculation of  $K_a/K_s$  ratios is influenced by parameterization to correct for such factors.

Finally, two additional measures of adaptive evolution were explored. Because  $K_a$  does not measure the relative neutrality of a mutation, PAM was examined as a measure of protein distance. Furthermore, because NED has been reported to be more clocklike than  $K_s$ , it was used to normalize PAM distances (see, e.g., Peltier et al. 2000). Thus, the PAM/NED ratio was explored as an alternative to the  $K_a/K_s$  ratio.

Analogous to  $K_s$ , the number of amino acid sites found to vary throughout a clade normalized for sample size can be a measure of the number of sites at which mutation can be tolerated by the physical chemistry (e.g., some interior residues are crucial for proper folding of the protein) of the specific protein structure. The number of those sites that vary along a specific branch, like  $K_a$ , shows how much of this potential for mutation is utilized. Using a mathematical framework similar to that developed for comparisons of polymorphism versus pairwise or along-branch divergence, the SSA statistic is also presented as a measure of adaptive evolution, where neutral evolution is expected to have a value of zero (Tajima 1989).

These measures of adaptive evolution were compared on a recent example of high  $K_a/K_s$  rate ratios in primates, that of the leptin protein family, and on a more ancient protein family that spans eubacterial and eukaryotic life, that of the deoxyribonucleoside kinases. The value of a recent example involving closely related species is that genes in these cases frequently share genomic localizations, and closely related species frequently have similar GC contents and codon preferences. This case is ideal for testing parameters correcting such genomic variables that are likely to be stable over the tree. Alternatively, the value of a more ancient and

broad example is that the more extensive amino acid substitution is ideal for testing models that account for this differentially.

Leptin has recently become a gene of pharmaceutical interest, having a role in obesity (Friedman and Halaas 1998). Two previous evolutionary analyses have been performed on the leptin protein in mammals (Benner, Trabesinger, and Schreiber 1998; Benner et al. 2000). Both studies demonstrated episodes of rapid sequence evolution between primates and rodents, in which the leptin gene has been implicated in obesity. However, the two studies utilized different methods for assigning ancestral sequences and different methods for calculating the  $K_a/K_s$  ratio. From the different methods, different branches representing different periods of evolutionary history were identified as adaptive in the two studies. In the first study, adaptive branches were identified in lineages leading to Hominidae, orangutans, and gorillas and a potentially adaptive event in the lineage leading to rhesus monkeys (Benner, Trabesinger, and Schreiber 1998). In the second study, adaptive events were assigned to the lineages leading to orangutans, rhesus monkeys, and Hominidae (Benner et al. 2000). (It should be noted that fig. 2 in Benner et al. [2000] differs from the representation of the leptin phylogenetic tree in the database described in the paper). This was the starting point for the examination of different methodologies presented in table 1, based on the phylogenetic tree in figure 1. This tree was not calculated here but is taken from an accepted biological tree for these species (Arnason, Gullberg, and Graur 1996).

In this study of recent evolutionary events, reconstruction using only DNA sequences seems to overestimate  $K_a/K_s$  ratios as compared with the more parsimonious method of performing a combined DNA and protein sequence reconstruction (method 1 vs. method 2). Correcting the data set for the codon bias found in the human genome (<http://www.kazusa.or.jp/codon/>) seems to have only minimal effects here and in other examples, indicating that codon bias does not seriously affect the calculation of  $K_a/K_s$  ratios in mammalian data sets (method 7). If there is a selective pressure on GC content, either at the global level related to melting temperature or at the local level related to isochore structure, then this will cause  $K_s$  to be underestimated. Correcting for this obviously reduces  $K_a/K_s$ , as seen in table 1 (methods 5 and 6). Matassi, Sharp, and Gautier (1999) have shown that neighboring genes are more likely to have similar  $K_s$  rates but that this does not correlate with GC content at the third positions. The selective reason and underlying mechanism for this neighboring gene effect are not known, so it is not clear whether this GC correction is necessary. In fact, Knight, Freeland, and Landweber (2001) have proposed that variation in genome GC content is largely due to mutational biases rather than any selectional effects on mutations that have occurred.

Correcting parsimony by branch length does make a difference in the branches leading to rhesus monkeys and Hominidae (methods 1 and 2 vs. methods 3 and 4). This appears to be the classical problem of a long branch

**Table 1**  
**Various Measures of Nucleotide Substitution and Adaptive Evolution Applied to Leptin Primate Sequences with  $K_a/K_s$  > 1 (bold) from Master Catalog Family 9614 (Benner et al. 2000)**

Method/Test	$K_a/K_s$	$K_a/K_s$	$K_a/K_s$	$K_a/K_s$	$K_a/K_s$	$K_a/K_s$	$K_a/K_s$	Robust?	NED?
Reconstruction	DNA	DNA + P	DNA + P	DNA + P	DNA + P	DNA + P	DNA + P		
Weighting	None	None	PAM	NED	None	None	None		
Matrix	Dis-Gran	Dis-Gran	Dis-Gran	Dis-Gran	Dis-Gran	Dis-Gran	Dis-Gran		
GC	None	None	None	None	Global	Local	None		
Codon bias	None	None	None	None	None	None	Human		
Method no.	1	2	3	4	5	6	7		
Orangutan 1	<b>1.06</b>	<b>1.02</b>	<b>1.02</b>	<b>1.02</b>	0.86	0.67	<b>1.02</b>	No	Yes
Rhesus 2	<b>1.17</b>	<b>1.25</b>	<b>1.09</b>	<b>1.10</b>	<b>1.05</b>	0.82	<b>1.20</b>	Yes	Yes
Hominidae 3	<b>1.67</b>	<b>1.09</b>	<b>1.22</b>	<b>1.18</b>	0.91	0.71	<b>1.08</b>	Yes	Yes

NOTE.—The reconstruction method, branch length weighting, substitution matrix, and parameterization are shown. The robustness test is a test for insignificantly short branch lengths, and the neutral evolutionary distance (NED) test is a test for the equilibration of silent codons applied in Benner et al. (2000). The actual adaptivity of the different proteins is not known. Numbers in column 1 correspond to branches in figure 1. DNA + P refers to ancestral sequence reconstruction based upon DNA + protein.

dominating a parsimony reconstruction in methods 1 and 2, which is corrected in methods 3 and 4. As NED is more neutral than PAM, method 4 is the preferred method for this data set.

Overall, the branches in the tree that appear to be adaptive using these methods are the Hominidae and, possibly, the rhesus monkey branches, but not the orangutan branch. The branch leading to orangutans was not considered adaptive in the Liberles et al. (2001) study with the application of a robustness test to eliminate overly short branch lengths despite the overestimation of  $K_a/K_s$  using only the method that was applied in that study. The adaptive branch in the lineage leading to Hominidae has been used to explain differences in the behavior of leptin between mice and humans and its applicability as a drug candidate (Benner, Trabesinger, and Schreiber 1998; Benner et al. 2000).

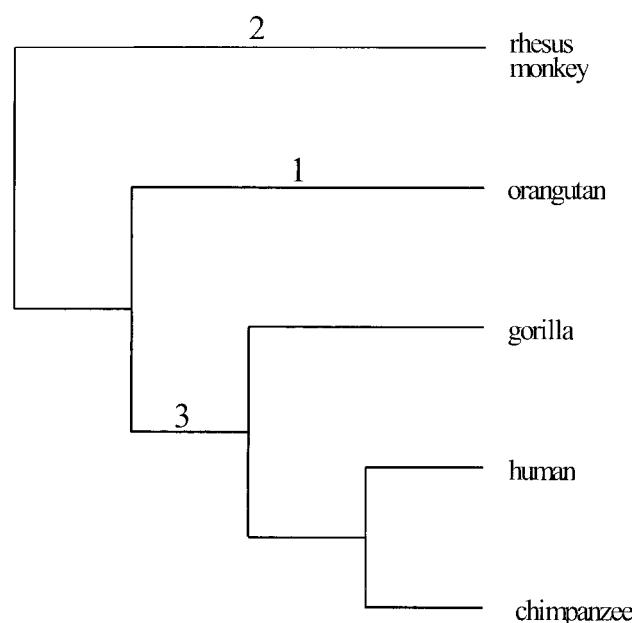


FIG. 1.—This tree reflects a common biological view of the relationship of these primate species (Arnason, Gullberg, and Graur 1996). Sequences from Master Catalog family 9614 are analyzed in table 1 using this tree (Benner et al. 2000).

A second, more ancient, protein family of evolutionary interest is the dexoyribonucleotide kinase superfamily C-terminal independently evolving unit (Benner et al. 2000; Almgren 2001). Here, both the gene function and the nucleotide specificity are undergoing adaptive evolution (see Almgren 2001 for details). In more ancient evolutionary events, the substitution model can affect the calculation of  $K_a$  significantly. In table 2, the five occurrences with  $K_a/K_s$  ratios of >0.6 from the PBLSB method were considered along with two control points that were least likely to be adaptive, representing the branches separating human and mouse deoxycytosine kinase. Some branches that obviously reflect changes of protein function are not detected with the 0.6 cutoff. There are several possible reasons for the failure to detect these change-of-function events (e.g., long-branch averaging, neutral evolution). A corresponding tree for this gene family derived from the Master Catalog (Benner et al. 2000) can be found in figure 2.

Again, the DNA-only reconstruction seems to overestimate  $K_a/K_s$  ratios relative to the DNA+protein reconstruction (method 1 vs. method 2). Several different substitution matrices were compared as well. Zhang (2000) has recently developed a discrete physicochemical substitution pattern table that was converted into a matrix and applied instead of the discrete Grantham matrix in the PBL method, resulting in only small differences (method 10). A continuous version of the Grantham matrix was applied instead of dividing the Grantham matrix into categories as is done in the PBL method (method 8) (Grantham 1974). This method underestimates  $K_a/K_s$  ratios relative to the other methods. A second continuous substitution matrix, the empirical Taylor-Jones matrix, was applied (method 9) (Taylor and Jones 1993). Use of this empirical matrix instead of theoretical matrices also resulted in only small differences. It appears as if the discrete Grantham, Taylor-Jones, and Zhang matrices function equally well. PAM and NED distance weighting do result in less averaging of  $K_a/K_s$  ratios from nodes at which there is branch length disparity, such as the rhesus monkey/Hominidae split in primate leptin (methods 3 and 4). NED as a neutral measure of evolution is conceptually pref-

**Table 2**  
**Various Methods Applied to the Evolution of Deoxyribonucleotide Kinase Sequences with  $K_a/K_s > 0.6$  (bold) from Master Catalog Family 2587 (Benner et al. 2000; Almgren 2001)**

Method/test	$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		$K_a/K_s$		Robust?	NED?	
	DNA	None	DNA	None	DNA	None	DNA	None	DNA	None	DNA	None	DNA	None	DNA	None	DNA	None			
Reconstruction	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran			
Weighting	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran			
Matrix	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran	None	Dis-Gran			
GC	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None			
Codon bias	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None			
Method no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
<i>C. elegans</i> 1	<b>0.61</b>	<b>0.62</b>	<b>0.65</b>	<b>0.80</b>	0.47	0.31	0.54	0.41	<b>0.61</b>	0.54	0.61	1.11	0.54	0.36	0.59	0.37	0.43	0.04	0.11	Yes	Yes
<i>Eubacteria</i> 2	<b>0.68</b>	0.40	0.39	0.41	0.31	0.40	0.40	0.48	0.35	0.36	0.36	-0.09	0.36	0.36	0.39	0.39	0.41	0.04	0.11	Yes	Yes
Metazoa 3	<b>0.75</b>	<b>0.64</b>	0.59	<b>0.68</b>	0.54	0.40	0.40	0.48	<b>0.65</b>	0.59	0.65	1.48	0.59	0.39	0.39	0.39	0.41	0.04	0.11	Yes	Yes
Metazoa 4	<b>0.74</b>	0.40	0.39	0.40	0.35	0.40	0.40	0.48	0.39	0.37	0.37	0.16	0.37	0.39	0.39	0.39	0.41	0.04	0.11	Yes	Yes
Metazoa 5	<b>0.62</b>	0.45	0.43	0.48	0.41	0.48	0.48	0.48	0.43	0.43	0.43	-0.25	0.43	0.43	0.43	0.43	0.41	0.04	0.11	Yes	Yes
Mouse 6	0.03	0.04	0.04	0.05	0.04	0.05	0.05	0.05	0.04	0.04	0.04	-3.71	0.04	0.04	0.04	0.04	0.04	0.04	0.11	Yes	Yes
Human 7	0.09	0.13	0.13	0.13	0.10	0.13	0.13	0.13	0.10	0.11	0.11	-3.31	0.11	0.11	0.11	0.11	0.11	0.11	0.11	—	—

NOTE.—Various  $K_a/K_s$  measurements and two additional measures, the point accepted mutations (PAM)/neutral evolutionary distance (NED) ratio and the sequence space assessment (SSA) statistic, are indicated. The NED-weighted reconstruction is based on ancestral sequences instead of pairwise comparisons. The phenotypic events corresponding to each branch are as follows: (1) *Caenorhabditis elegans* NADH dehydrogenase 42-kDa subunit from mammalian NADH dehydrogenase 42-kDa subunit; (2) eubacterial purine kinase speciation; (3) metazoan NADH dehydrogenase 42-kDa subunit from eubacterial kinases; (4) metazoan dG, dC kinases from metazoan tk kinase subfamily; (5) metazoan kinases from eubacterial kinases and metazoan NADH dehydrogenase 42-kDa subunit; (6) mouse dC kinase from human dC kinase; (7) human dC kinase from mouse dC kinase. Numbers in column 1 correspond to branches in figure 2. The substitution matrices used were discrete Grantham (Dis-Gran), continuous Grantham (Con-Gran), continuous Taylor-Jones (Con-TJ), and discrete Zhang (Dis-Zhang).

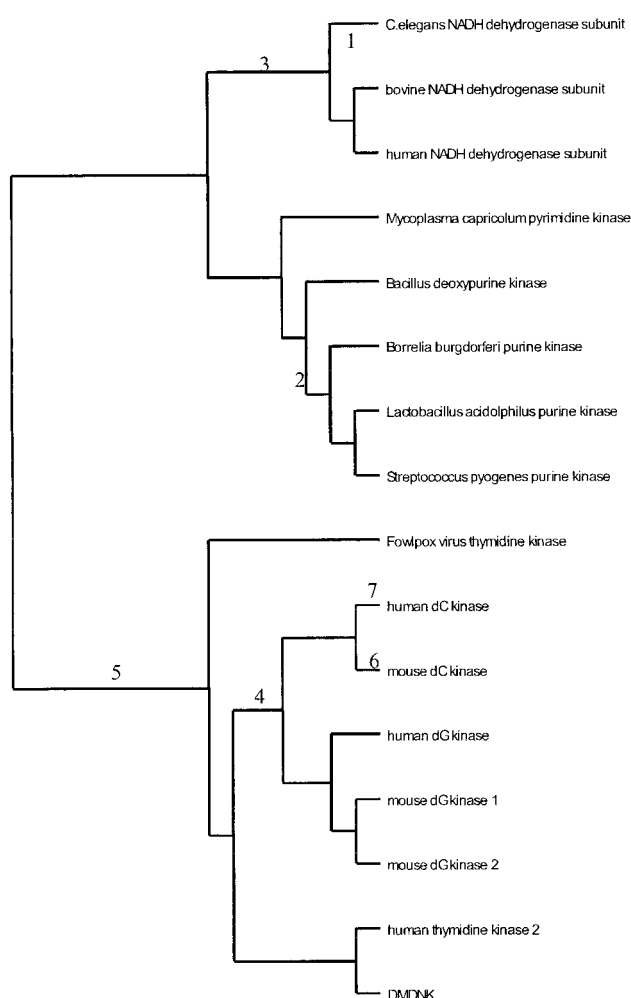


FIG. 2.—This tree is derived from Master Catalog family 2587. The data calculated in table 2 are based on this tree (Benner et al. 2000).

erable to PAM as a weighting tool. However, pairwise NED distance weighting has the disadvantage of equilibration of codon substitution with time (Smith and Smith 1996). It should be noted that calculating NED rates along branches of a tree allows one to access more distant dates than does simply comparing pairwise NED rates between extant genes (see Koshi and Goldstein 1996).

This advantage of calculating NED distances along a branch is dependent on the validity of reconstruction along sufficiently short branches throughout the tree. This is not the case for all of the branches in figure 2. Some changes of protein function are likely at high PAM distances, even when  $K_a/K_s$  is overestimated due to saturation. Detecting such changes of function is not informative about the evolutionary mechanism (selection vs. neutrality). Further simulation can indicate exactly when ancestral sequence reconstruction can be trusted in the calculation of NED distances or  $K_s$  rates for such molecular-evolution studies. However, the problem of inaccurate ancestral sequence reconstruction along long branches will diminish as more genes and

genomes are sequenced and gene trees become better articulated.

The PAM/NED ratio (method 11) seems to follow a pattern similar to that of the  $K_a/K_s$  ratios. Branch 3 from table 2 most obviously contains a change of protein function. This is best discriminated by SSA (method 12). Branch 4 may or may not represent a change of function because it is not obvious what the enzymatic specificity was at either end of the branch. The same ambiguity about functions also characterizes branch 1. During the time represented in branches 2 and 5, major changes of function of the protein are less clear and, furthermore, may average multiple events (adaptive, nonadaptive, and conservative) over long branch lengths. Values for these four branches (branches 1, 2, 4, and 5) are intermediate by the SSA measure but not negatively discriminated from branch 1 by  $K_a/K_s$  and PAM/NED. In all cases, values for branches 1–5 are still significantly higher than those for control branches 6 and 7, which are clearly undergoing negative selection. The SSA statistic is interesting as a different parameter that is more closely tied to the specific protein structure and fold. The PAM/NED ratio (or a Grantham/NED ratio) considers more information about the chemical degree of change (it is based on a substitutional model of evolution) per biological time, and PAM/NED ratios appear to correlate with  $K_a/K_s$  throughout the tree. As more examples of adaptivity are documented in the literature, it will be interesting to benchmark SSA, PAM/NED ratios, and  $K_a/K_s$  ratios against real evolutionary events. Interestingly,  $K_a/K_s$  ratios appear to correlate at some level with some putatively adaptive events in Eubacteria. Given the large degrees of codon bias and GC bias that make  $K_s$  nonneutral in Eubacteria, this is surprising, although the effect may be similar to that of  $K_s$  underestimation on long branches where  $K_a$  is sufficiently large to accurately reflect changes of protein function.

Bioinformatics and functional genomics approaches to detecting the genes responsible for the adaptive evolution of species are important tools as more metazoan (and mammalian) genomes are sequenced. The evaluation of tools described here is necessary for application of the appropriate tools to uncover the evolutionary history of species.

### Acknowledgments

This work was supported by a center grant from the Swedish Foundation for Strategic Research. I am grateful to Mike Hendy, Rob Knight, Steve Benner, David Schreiber, Steve Chamberlin, Sridhar Govindarajan, Arne Elofsson, and Jens Lagergren for helpful discussions. I am grateful to Malin Almgren and Bengt Sennblad for careful reading of this manuscript. I am also grateful to EraGen Biosciences, Inc. (Alachua, Fla.) for supplying multiple sequence alignments and phylogenetic trees from the Master Catalog, to the Computational Biochemistry Research Group at the Swiss Federal Institute of Technology-Zurich for supplying a copy of Darwin, and to David Schreiber for supplying a copy of his program encoding the PBLSB method.

### LITERATURE CITED

- ALMGREN, M. A. E. 2001. Genomic approaches to understanding protein function. M.Sc. thesis, Stockholm University, Stockholm, Sweden.
- ARNASON, U., X. XU, A. GULLBERG, and D. GRAUR. 1996. The "Phoca Standard": an external molecular reference for calibrating recent evolutionary divergences. *J. Mol. Evol.* **43**: 41–45.
- BENNER, S. A., S. G. CHAMBERLIN, D. A. LIBERLES, S. GOVINDARAJAN, and L. KNECHT. 2000. Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**:97–106.
- BENNER, S. A., N. TRABESINGER, and D. SCHREIBER. 1998. Post genomic science: converting primary structure into physiological function. *Adv. Enzyme Regul.* **38**:155–180.
- CHELVANAYAGAM, G., A. EGGENSWILER, L. KNECHT, G. H. GONNET, and S. A. BENNER. 1997. An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**:307–316.
- CRANDALL, K. A., C. R. KELSEY, H. IMAMICHI, H. C. LANE, and N. P. SALZMAN. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16**: 372–382.
- FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**:406–416.
- FRIEDMAN, J. M., and J. L. HALAAS. 1998. Leptin and the regulation of body weight in mammals. *Nature* **395**:763–770.
- GOLDING, G. B., and A. M. DEAN. 1998. The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**:355–369.
- GOULD, S. J., and N. ELDREDGE. 1993. Punctuated equilibrium comes of age. *Nature* **366**:223–227.
- GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**:862–864.
- INA, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**:190–226.
- KNIGHT, R. D., S. J. FREELAND, and L. F. LANDWEBER. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**:research0010.1–research0010.13.
- KOSHI, J. M., and R. A. GOLDSTEIN. 1996. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* **42**: 313–320.
- LI, W. H., C. I. WU, and C. C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**:150–174.
- LIBERLES, D. A., D. R. SCHREIBER, S. GOVINDARAJAN, S. G. CHAMBERLIN, and S. A. BENNER. 2001. The adaptive evolution database (TAED). *Genome Biol.* **2**:research0028.1–0028.6.
- MATASSI, G., P. M. SHARP, and C. GAUTIER. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**:786–791.
- MESSIER, W., and C. B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* **385**:151–154.
- MIYAMOTO, M. M., and W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**:503–513.
- NEI, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.

- NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- PAMILO, P., and N. O. BIANCHI. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.* **10**:271–281.
- PELTIER, M. R., L. C. RALEY, D. A. LIBERLES, S. A. BENNER, and P. J. HANSEN. 2000. Evolutionary history of the uterine serpins. *J. Exp. Zool.* **288**:165–174.
- POLLACK, D. D., J. A. EISEN, N. A. DOGGETT, and M. P. CUMMINGS. 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* **17**:1776–1788.
- SMITH, J. M. 1994. Estimating selection by comparing synonymous and substitutional changes. *J. Mol. Evol.* **39**:123–128.
- SMITH, J. M., and N. H. SMITH. 1996. Synonymous nucleotide divergence: what is “saturation”? *Genetics* **142**:1033–1036.
- TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- TAYLOR, W. R., and D. T. JONES. 1993. Deriving an amino acid distance matrix. *J. Theor. Biol.* **164**:65–83.
- YANG, Z., and R. NIELSEN. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**:32–43.
- ZHANG, J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* **50**:56–68.
- ZHANG, J., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44**:S139–S146.

MICHAEL HENDY, reviewing editor

Accepted July 13, 2001