Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Residue Frequencies

Aaron L. Halpern and William J. Bruno
Theoretical Biology and Biophysics, Los Alamos National Laboratory; and Santa Fe Institute

Estimation of evolutionary distances from coding sequences must take into account protein-level selection to avoid relative underestimation of longer evolutionary distances. Current modeling of selection via site-to-site rate heterogeneity generally neglects another aspect of selection, namely position-specific amino acid frequencies. These frequencies determine the maximum dissimilarity expected for highly diverged but functionally and structurally conserved sequences, and hence are crucial for estimating long distances. We introduce a codon-level model of coding sequence evolution in which position-specific amino acid frequencies are free parameters. In our implementation, these are estimated from an alignment using methods described previously. We use simulations to demonstrate the importance and feasibility of modeling such behavior; our model produces linear distance estimates over a wide range of distances, while several alternative models underestimate long distances relative to short distances. Site-to-site differences in rates, as well as synonymous/nonsynonymous and first/second/third-codon-position differences, arise as a natural consequence of the site-to-site differences in amino acid frequencies.

Introduction

The estimation of evolutionary distances from genetic sequences is important both for our understanding of evolutionary history and for our understanding of evolutionary processes, including matters such as mutation rates, population sizes, and selectional pressures. If mutation rates were constant from site to site and the process of fixation occurred purely through genetic drift (i.e., no selection), as may be the case for pseudogenes or other noncoding regions of a genome, models of evolution and methods of inferring evolutionary distances based purely on the mutation process, such as the Jukes-Cantor, Kimura two-parameter, and Felsenstein/HKY models, would be adequate; see Swofford et al. (1996) for discussion of such models. However, when evolutionary distances are calculated based on coding sequences, as is often the case, failure to take into account effects of selection may lead to substantial underestimates of evolutionary distances. The factor by which these distances are underestimated will increase with increasing distance.

That (protein-level) selection has a significant impact on sequence evolution and on the estimation of evolutionary patterns (phylogeny, patterns of synonymous and nonsynonymous substitutions, etc.) is amply clear, as illustrated by the interest in models of evolution which vary either rates of change at a site (Felsenstein and Churchill 1996; Yang 1996) or rates of change between different character states (e.g., the use of PAM matrices as in Felsenstein [1993] or biochemical measures of similarity between amino acid residues in calculating distances between proteins as in Goldman and Yang [1994] and Thorne, Goldman, and Jones [1996]).

From work on nucleic acid sequence evolution, however, it is clear that not only rates of substitution between states, but also the equilibrium frequencies of the different states are important; Felsenstein (1981) and Gojobori (1983) show that changes in equilibrium frequencies may strongly influence estimates of the degree of mutational saturation which sequences have achieved. We show here that similar effects on the protein level due to selection must be modeled in order to achieve accurate estimates of the evolutionary distances for highly divergent coding sequences.

Large-scale sequencing efforts are leading to a rapid increase in both number and size of data sets containing highly divergent but clearly homologous and functionally conserved protein-coding sequences. For coding sequence data sets of sufficient size and diversity, amino acid frequencies can be estimated directly from a multiple alignment of the sequences, and these can be incorporated into a model of protein evolution specific to that family of sequences, much as position-specific scoring matrices, profiles, and Markov models which make use of site-specific amino acid frequencies have come to be used in homology searches (Henikoff and Henikoff 1992, 1994; Baldi et al. 1994; Krogh et al. 1994; Eddy, Mitchison, and Durbin 1995; Gribskov and Veretnik 1996).

In order to enable such an approach, we provide a model of nucleic acid sequence evolution which is codon-based and employs site-specific modeling of the effects of protein-level selection. The model of selection is a simple one, in which the following four conditions hold: (1) The effects of selection are on the probabilities of fixation once a mutation has occurred in an otherwise homogeneous population, with fixation being rapid relative to mutation so that polymorphism within a population may be neglected, as modeled by Kimura (1962). (2) Selectional pressures (i.e., fixation probabilities) are assumed to be constant at a given position for all lineages in the alignment under consideration. (3) Each position is assumed to be independent of all other positions. (4) Only purifying (negative) selection is present. Since the model is codon-based, it could be used to
model codon bias, but to keep the number of parameters more manageable, we will normally assume that selection depends only on the encoded amino acid. Without denying that such issues may be biologically and analytically important, we further neglect variation in the form of the mutational and selectional matrices between lineages, other forms of selection (RNA and/or DNA structure requirements, etc.), and uncertainty of alignment.

Thus, a site-invariant model of codon-to-codon mutation is combined with site-specific estimates of equilibrium frequencies of each amino acid. Rates of replacement are determined for each site from these two components according to the effect of selection on the probability of fixation of mutations, following Kimura (1962) and Golding and Felsenstein (1990). A consequence of the model, rather than a feature which is explicitly modeled, is that rates of substitution are lower at positions under greater purifying selection. From this model, evolutionary distances between pairs of taxa may be calculated by the method of maximum likelihood. Likewise, data sets may be simulated according to given parameter values. By estimating parameters from real data sets and then simulating sequences from these parameter values, we may construct realistic data sets with known pairwise evolutionary distances.

We will show, through analysis of such simulated sequences, that a variety of current methods for estimating evolutionary distances which do not take site-specific character frequencies into accounts are in danger of underestimating true distances under conditions which are empirically relevant. Methods for which this may be a problem include those which model mutational biases in varying detail but do not model selection at all (from Jukes-Cantor to the general time-reversible model of mutation); those which model selection in terms of an average effect on substitution probabilities, such as the use of PAM matrices on protein sequences in PROTDIST (Felsenstein 1993); models based on biochemical measures of similarity (Goldman and Yang 1994); models including a synonymous/homosynonymous distinction (Muse and Gaut 1994); and models allowing different rates of substitution between positions (Felsenstein and Churchill 1996; Yang 1996). Recent works by Mitchison and Durbin (1995) and Thorne, Goldman, and Jones (1996) which distinguish substitution patterns for a few classes of positions—by similarity of amino acid usage in the former report and by secondary structure in the latter—go part way toward modeling selectional character biases but, in our estimation, are unlikely to be sufficiently detailed to more than partially overcome the problem. Real data sets which are currently being analyzed to determine times of divergence for early evolutionary events (e.g., Eigen and Nieselt-Struwe 1990; Miyata and Kaestne 1983; Chan et al. 1995; Wray, Levinton, and Shapiro 1996) are likely to lead to underestimations of the more remote divergence times using these methods. Furthermore, we will show that for a sufficiently large and diverse alignment, it is possible to estimate the equilibrium frequencies at each site with sufficient accuracy to obtain estimates of evolutionary distances which are linear out to very large distances.

Materials and Methods

Model

The qualitative impression one gets from working with protein alignments is that while some positions in a protein permit almost any residue (and indeed loop regions permit extensive length variation), many positions in an alignment appear to allow only a limited subset of amino acids, or at least show strong preferences for a few (sometimes only one) specific residues. Without denying that covariation and structural or functional evolution may lead proteins from different taxa to be under different selectional pressures at a given position, we will pursue a model of coding-sequence evolution in which each position in a protein is under site-specific selectional pressures which have remained constant since the common ancestor of the data set at hand.

These pressures are assumed to have a twofold effect: they influence, for each site, the relative frequencies of each residue and, in combination with a model of the mutational process, determine the rates at which substitution between given codons occurs. As we show below, for a given model of mutation and a model of fixation involving purifying selection and genetic drift, the relative frequencies of two residues at a site may be used to derive rates of substitution between codons relative to the mutation rate. Consequently, we may parameterize site-specific selection in terms of residue frequencies, which, as we shall show, may be estimated from the observed sequences in a large enough data set with sufficient accuracy to reconstruct evolutionary distances.

We model sequence evolution as a stationary Markov process in which the unit of description is the codon. The choice of the codon permits us to take into account the mutational process, the genetic code, and the selectional pressures applying at the protein level, as in the work of Goldman and Yang (1994) and Muse and Gaut (1994). A model of sequence evolution then corresponds to the specification of rates of substitution between codons. We assume that sites are independent, that reversibility applies, and that the mutation process, but not the selection process, is identical at all sites.

For position (codon) i, we define a $64 \times 64$ matrix of substitution rates $r_{ab}$ such that $r_{ab}$ is the rate at which codon $a$ is replaced by codon $b$ at position $i$. These position-specific substitution rates are determined as the product of three terms: a site-invariant probability of mutation (in one replication cycle) between codons, $p_{ab}$, a site-specific probability of fixation of a mutation, $f_{ab}$, and an arbitrary scaling constant, $k$. Thus, $r_{ab} = k \times p_{ab} \times f_{ab}$. (1)

To specify $p_{ab}$ we make use of a discrete-time model of mutation; letting $a_j$ and $b_j$ refer to the nucleotides in the $j$th positions of codons $a$ and $b$, we take...
the probability of mutation between two codons in a single replication cycle, \( p_{ab} \), to be the product of the mutation probabilities at each nucleotide position \( j \) within the codon:

\[
p_{ab} = \prod_{j=1}^{3} p_{ajbj}. \tag{3}
\]

The terms \( p_{ab} \) are specified according to an overall probability of mutation per cycle and any reversible model of mutation. In analyses presented in this paper, we make use of a model which includes nucleotide frequencies and a transition/transversion ratio, akin to the models known as HKY and F84 (Hasegawa, Kishino, and Yano 1985; Felsenstein 1993). The extension to the general reversible model is straightforward. For our implementation, we specify parameters for nucleotide frequency (\( v_A, v_C, v_G, v_T \)), the probability of a transition per replication cycle (\( p_t \)), and the ratio of transitions to transversions (\( p_d/p_v \)). Each of these parameters reflects the values that would be expected from mutation alone, in the absence of selection. For comparison, equation (4) shows the relation between \( p_{ab} p_{ba} \) and the transition/transversion rate ratio of HKY, \( \kappa \):

\[
\kappa = \frac{p_{t1}(v_A + v_C)(v_C + v_T)}{v_A v_G + v_C v_T}. \tag{4}
\]

From these specifications, we may derive the probabilities of specific mutations per replication cycle. For a transition \( x \rightarrow y \), we have

\[
p_{xy} = \frac{v_y}{2(v_A v_G + v_C v_T)} p_{t1},
\]

and for a transversion \( x \rightarrow z \), we have

\[
p_{xz} = \frac{v_z}{2(v_A + v_C)(v_C + v_T)} p_{t1}. \tag{5}
\]

The probability \( p_{xz} \) is, of course, \( 1 - \sum_{y \neq x} p_{xy} \). The resulting single-nucleotide probabilities are combined via equation (3) to obtain probabilities per replication cycle of codon-to-codon mutation.

Since the rates of mutation per cycle are low (\( p_{ab} \ll 1 \)), we may use a continuous-time approximation of the behavior of this discrete time model of mutation once our rates are defined. Indeed, replacement of our rates is taken to be small (zero) relative to the time between introduction of a mutant and its eventual fixation. Thus, given a set of parameter values specifying the mutation rates and, for each position, the equilibrium probabilities of codon frequencies as follows, with \( k \) an arbitrary constant (see below):

\[
r_{ab} = k \times p_{ab} \times \frac{\ln \left( \frac{\pi b_{P_{ba}}}{\pi a_{P_{ab}}} \right)}{1 - \pi a_{P_{ab}} / \pi b_{P_{ba}}}. \tag{10}
\]

Thus, given a set of parameter values specifying the substitution rates and, for each position, the equilibrium frequencies of each codon, we may specify position-specific substitution rates.

The resulting rates have the following properties: As required by reversibility, \( r_{ab}/r_{ba} = \pi_b/\pi_a \). However, the total amount of substitution (the “flux”) between two codons is not uniform but, rather, depends on their relative fitnesses: the greater the difference in fitness, the less overall substitution. This has an effect not only
on rates between individual pairs of codons, but also on overall rates of substitutions at different sites: the maximum rate is achieved when all codons are equally fit, and the overall rate of substitution tends to decrease with decreasing variability of a position (as measured, e.g., by entropy). Equally, pairs of synonymous codons, for which fitnesses are the same (according to the model), will have the highest possible flux, and nonsynonymous pairs will generally have lower flux. Finally, because of the greater degeneracy of third-position substitutions, third-position changes will be more frequent than first- or second-position substitutions. Although the specific assumptions regarding fitness and fixation underlying the model may be false for a given data set, the model may still be of interest for these general properties.

We note that the rates of substitution may be scaled arbitrarily to choose the units of evolutionary distance as we please. By calculating the expected overall rate of change at a position under no selectional pressure, we may make our units the expected number of “hits” (codon substitutions) separating sequences at a site under no selectional pressure:

\[ k = \sum_{a,b} \pi'_{a} r'_{a \rightarrow b} \]  

\[ \pi'_{a} = \nu_{a1} \nu_{a2} \nu_{a3} \]  

\[ r'_{ab} = \begin{cases} 
\frac{\log \frac{\pi'_{b}}{\pi'_{a}} p_{ba}}{\pi'_{a} p_{ab}} & \text{for } \pi'_{b} \neq \pi'_{a} \, p_{ba} \\
1 - \frac{\pi'_{a} p_{ba}}{\pi'_{b} p_{ba}} & \text{for } \pi'_{a} p_{ba} = \pi'_{b} p_{ba} 
\end{cases} \]

This has the appeal that it should not depend on the strength of the selectional pressures on the region of interest, allowing direct comparison of distances calculated from different regions. Assuming that most hits on a site involve single-point mutations, we may approximate the number of hits per nucleotide at an unselected site as one third the number of hits per unselected codon.

Given a model of (codon) substitution rates at each position, the maximum likelihood estimate (MLE) of evolutionary distance between two sequences may be obtained using standard methods. Briefly, the matrix of substitution rates at each position, \( r' \), is multiplied by \( t \) and exponentiated to give a matrix, \( P(t) \), which has elements \( p_{iab}(t) \), the likelihood of codon \( a \) at site \( i \) in one sequence corresponding to that of codon \( b \) in another as a function of evolutionary distance, \( t \). (See Swofford et al. [1996] for an introduction to maximum-likelihood distance and phylogeny estimation.) From this, the likelihood of two observed sequences being separated by distance \( t \) can be determined. The value of \( t \) which maximizes this likelihood is \( \hat{t} \), our estimate of the distance between the two sequences.

Moreover, given such a model, it is possible to generate a simulated family of related proteins. For a given phylogenetic tree (topology and branch lengths), and a specification of the mutation and selection parameters, we may generate an initial random sequence according to the equilibrium frequencies, associate it with some arbitrarily chosen node in the tree, and then “evolve” this sequence along the branches of the tree according to the model.

Frequency Estimation

We turn now to the issue of estimating the parameters required by the model for the calculation of distances. In principle, all parameters could be simultaneously estimated from a large enough data set, given sufficient computing power. We note that the approach of Lewis (1998), involving the use of genetic algorithms to simultaneously estimate a phylogeny and various model parameters, may make this more feasible for our model. At present, we have chosen to implement our model in a program which requires as input specification of the mutational and selectional parameters. In this section, we briefly discuss how such specifications may be determined using existing methods.

Regarding the mutational parameters, nucleotide frequencies may be estimated by the observed frequencies in the data set. The transition/transversion ratio, as well as additional parameters if one considers a more complicated mutational model than discussed above, could be estimated using methods described elsewhere (Yang 1995; Bruno and Arvestad 1997). We note, however, that application of these methods to coding sequences is not ideal, since it may result in estimates of values reflecting selection, rather than mutation alone. For instance, the estimated ratio of transitions to transversions might be higher than the true ratio of such mutations because of (among other things) the bias in third positions: selection tends to favor silent changes over amino-acid-replacing changes, and third-position transitions tend to be silent, in contrast to third-position transversions. Estimation of mutational parameters from noncoding sequences such as pseudogenes, if available, might be preferred. In the simulations discussed below, since we are mostly concerned with illustrating the importance and feasibility of modeling site-specific frequency effects, we will simply set the nucleotide frequencies and transition/transversion ratio to the values used in simulating the sequences.

As for the estimation of site-specific codon frequencies, given a sufficiently large data set, individual codon frequencies could in principle be estimated, allowing the modeling of codon bias (see also Goldman and Yang 1994), but the large number of parameters to be estimated would require unrealistically large data sets. To make the problem more tractable, we make one simplification which considerably reduces the estimation problem, albeit at the loss of some of the true complexity of the evolutionary process: we ignore the possibility of site-specific nucleic-acid-level selection effects (e.g., codon bias or RNA secondary structure) and calculate the frequency of a given codon as a function of the frequency of the corresponding amino acid and the nucleotide frequencies. Concretely, we let the equilibrium frequency of codon \( i \) be estimated as follows, where
$aa(i)$ is the amino acid coded for by codon $i$, and $\Pi_{aa(i)}$ is the frequency of $aa(i)$:

$$\pi_i = \prod_{aa(i)} \frac{v_{ij} v_{ij}}{v_{ij} v_{ij}}.$$  \hspace{1cm} (14)

We thus reduce the codon frequency estimation problem to the estimation of overall nucleotide frequencies (for which the observed frequencies in the alignment of interest may be used), and the estimation of equilibrium frequencies of each amino acid at a given site. This latter problem is a topic which has received considerable attention in the literature on detection of sequence homologs and sequence alignment (Baldi et al. 1994; Eddy 1996; Henikoff and Henikoff 1996; Sjölander et al. 1996; and references cited therein). We discuss here a specific method for the estimation of equilibrium frequencies which may be used to set the selectional parameters required by our model for distance calculations (or sequence simulation). We show below that for simulated but relatively realistic data, the method gives useful results. We wish to emphasize, however, that the choice of a specific method for estimating amino acid frequencies is separate from the use of these estimates in the model described above, so the specific method we are about to describe could be replaced by other methods.

The estimation involves two steps, the estimation of the number of times each amino acid has been introduced at a given position over the evolution of the sequences in the data set, and the addition of pseudocounts to these counts. Both of these procedures are described in detail elsewhere. Briefly, the effect of phylogenetic structure on the estimates is taken into account using the RIND program described by Bruno (1996), which iteratively estimates the evolutionary history and amino acid frequencies under a model of sequence evolution which ignores the genetic code. One result is a set of estimated “counts” of the number of times each amino acid is introduced at a given position. The final estimate of site-specific frequencies is obtained by adding pseudocounts to these counts and taking the fraction of the total at a given position of all counts and pseudocounts which is contributed by a given amino acid to be the frequency of that amino acid, as described by Karplus (1995a, 1995b) and Henikoff and Henikoff (1996). We find, empirically, that adding a uniform pseudocount in the range of 0.5 to 1.0 counts to each category results in good distance estimates (see below). We expect that since estimating amino acid frequencies is of considerable interest for homology search methods and other sorts of sequence analyses, better methods will continue to be developed; see Sjölander et al. (1996) for an ambitious approach.

Sequence Simulation

In order to illustrate the potential effects of site-specific frequencies on the estimation of evolutionary distances, we have constructed simulated sequences based on parameters estimated from a real alignment of sequences of the major capsid protein, L1, of the papillomaviruses. An alignment of 74 full-length L1 sequences was obtained from the Papillomavirus Database (Myers et al. 1996) and slightly modified by hand; all positions containing gaps were removed, leaving an amino acid alignment of 430 residues and the corresponding nucleotide alignment of 1,320 positions. The sequences in the resulting nucleotide alignment vary in dissimilarity (Hamming distance) from 6% to 48%.

A phylogeny was estimated from the nucleotide alignment using DNADIST (Felsenstein 1993; Kimura two-parameter model with default settings) and a version of NEIGHBOR which returns only positive branch lengths, NNNEIGHBOR (Bruno 1995). In order to permit investigation of long evolutionary distances, the lengths of all branches in the tree were multiplied by a factor of 6.0; we note that since the distances between the more divergent sequences are likely to have been substantially underestimated by DNADIST, this scaling does not necessarily result in unrealistically diverged sequences. From the protein alignment, amino acid frequencies at each position were estimated as described above using RIND; a small pseudocount (0.2) was added to each category on the assumption that the counts underestimated the true variability in the original data set.

Simulated sequences were generated from the scaled tree according to the model described above using the amino acid frequency estimates derived from the real sequence set, with mutational parameters specified as follows: the probability of a transition at a given nucleotide in a single replication cycle, $p_{\text{trans}}$, was given as 0.000001 (10^{-8}), the transition-transversion ratio $(p_t/p_v)$ as 2.0, and the nucleotide frequencies $v_A$, $v_C$, $v_G$, $v_T$ as 0.3, 0.2, 0.2, and 0.3, respectively.

Having an (artificial) data set corresponding to a known phylogeny permits comparison of the true pairwise distances (those corresponding to the distances in the tree used to generate the sequences) with those estimated by various methods. Figure 1 compares results from DNADIST with all sites identical; DNARates and fastDNAml, allowing site-to-site rate variation; the model described here, using the input parameters (that is, the model used to generate the sequences); and the model described here with parameters estimated from the simulated data set. Details are given in the legend to figure 1.

Programs for distance estimation and sequence simulation implementing the methods presented above are available from the authors.

Results and Discussion

Figure 1A shows that site-invariant models may seriously underestimate longer distances if sequences evolve under selection that results in site-specific residue frequencies. Figure 1B shows that this underestimation is also seen with models allowing site-to-site rate variation but not modeling site-specific frequencies. Figure 1C shows, not surprisingly, that when distances are estimated based on the correct model, they remain linear out to greater distances. Figure 1D shows that it is possible to estimate amino acid frequencies from the (sim-
Distances were estimated under the same model as in DNADIST (Felsenstein 1993) with parameter settings corresponding to the mutational parameters used to simulate the sequences (frequencies of A, C, G, T = 0.3, 0.2, 0.2, 0.3, respectively; transition/transversion ratio 2.0, corresponding to setting the T option of the program to 1.5833), with rates at all sites equal. A, Distances were estimated under the same model as in A, but allowing site-to-site rate variation. Positions were assigned to one of 9 rate categories by DNArates (Olsen et al. 1994), using an initial topology estimated with DNADIST (as above) and NEIGHBOR. Pairwise distances between sequences were estimated using these rate assignments using fastDNAm (Olsen et al. 1994). B, Distances were estimated using the same parameter values as were used to simulate the sequences. D, Distances were estimated with the model described here, with parameters estimated from the simulated data set using RIND counts (Bruno 1996), adding 0.65 pseudocounts to each category. In all cases, the line indicates the best linear fit through the origin for points with input distance \( \leq 1.5 \).

Three points should be made regarding these observations. First, the different distance measures return results in different units; a slope of one in figure 1 is thus not necessarily expected. However, the different measures would nonetheless be linearly related to the original distances if the methods were giving unbiased estimates of the evolutionary distances between sequences.

Second, the conditions under which the simpler models failed to give linear distance estimates can occur in real data sets. Although the model introduced here and the parameters used to generate the sequences under discussion involve many assumptions that are only approximately correct (e.g., constancy of evolutionary pressures on all lineages, or site-to-site independence), we believe that they are a reasonable approximation of real evolutionary processes. The amino acid frequencies estimated with the model described here, using the same parameter values used to simulate the sequences (frequencies of A, C, G, T = 0.3, 0.2, 0.2, 0.3, respectively; transition/transversion ratio 2.0, corresponding to setting the T option of the program to 1.5833), with rates at all sites equal. A, Distances were estimated under the same model as in A, but allowing site-to-site rate variation. Positions were assigned to one of 9 rate categories by DNArates (Olsen et al. 1994), using an initial topology estimated with DNADIST (as above) and NEIGHBOR. Pairwise distances between sequences were estimated using these rate assignments using fastDNAm (Olsen et al. 1994). B, Distances were estimated using the same parameter values as were used to simulate the sequences. D, Distances were estimated with the model described here, with parameters estimated from the simulated data set using RIND counts (Bruno 1996), adding 0.65 pseudocounts to each category. In all cases, the line indicates the best linear fit through the origin for points with input distance \( \leq 1.5 \).

Three points should be made regarding these observations. First, the different distance measures return results in different units; a slope of one in figure 1 is thus not necessarily expected. However, the different measures would nonetheless be linearly related to the original distances if the methods were giving unbiased estimates of the evolutionary distances between sequences.

Second, the conditions under which the simpler models failed to give linear distance estimates can occur in real data sets. Although the model introduced here and the parameters used to generate the sequences under discussion involve many assumptions that are only approximately correct (e.g., constancy of evolutionary pressures on all lineages, or site-to-site independence), we believe that they are a reasonable approximation of real evolutionary processes. The amino acid frequencies used in the simulation were estimated from a real data set, and the branch lengths likewise reflect an attainable degree of divergence.

Third, there remains one “free parameter” in our estimation of amino acid frequencies which must currently be specified by the user: the number of pseudocounts which are added (\( \zeta \)) to regularize the estimates. Figure 1D shows the best results out of several attempts using different values of \( \zeta \). The results in figure 1D show that it is possible to choose a value of \( \zeta \) which gives linear estimates out to quite long evolutionary distances. It should, in principle, be possible to provide a method for deciding on this value for a given data set without knowing the distances a priori. As a point of departure, we note that methods based on urn models of substitution, involving either uniform pseudocounts (“zero-offsets”) or more complicated prior distributions such as Dirichlet mixtures estimated from protein alignments (see Sjølander et al. 1996 and references therein), appear to give underestimates of the amount of regularization needed, as has been discussed by Henikoff and Henikoff (1996) in connection with defining profiles for homology searches. We believe the effect is due to the genetic code: substitutions do not correspond to draws from an urn, since the choice of the new state is dependent on the old state; failure to recognize this dependence leads to an overestimation of the amount of evidence, which can be (partially) compensated for by adding more pseudocounts. If the lack of independence of counts can be dealt with, it would certainly be worthwhile to consider more complicated methods of determining pseudocounts, such as Dirichlet mixtures (Sjölander et al. 1996) or substitution matrix methods (Henikoff and Henikoff 1996).

Conclusion

Perhaps the key contribution of the current work is the modeling of selection in terms of its effect on equilibrium frequencies of residues. We have shown that failing to model such effects on the evolution of coding regions may lead to substantial underestimation of longer distances. We have shown that it is possible to model these effects using standard models of mutation and population genetics, deriving rates of substitution which would lead to the estimated equilibrium frequencies at each position. Finally, we have shown that it is possible to estimate equilibrium frequencies from data sets of realistic size and divergence with sufficient accuracy to allow estimation of distances between highly divergent sequences.

By separating the mutational process from the fixation process, we may capture differences in codon-to-codon rates due to the genetic code that are relevant at small degrees of divergence, as well as differences in amino acid frequencies that are key to estimation of larger distances. We also have a separation of site-independent parameters (mutational) and site-specific parameters (selectional), each of which may be estimated by previous methods. This contrasts with the rates implied by, e.g., PAM matrices, which fail to distinguish
effects of the code from effects of protein-level selection. An additional feature of the model that we find appealing is that it results in site-to-site rate variation without including a set of site-specific rate parameters in the model; the model thus accounts for heterogeneity of rates in a novel, and probably biologically meaningful, way. Without denying the existence of rate variation due to positive selection (e.g., as result of immune pressure on an antigenic region of a pathogen) or purifying selection at levels other than the amino acid level (e.g., RNA secondary structure, same-residue codon bias, etc.), it would be interesting to see the extent to which observed rate heterogeneity in coding regions could be reduced to variation in the strength of residue preferences as indicated by their frequencies.

In the implementation of distance estimation presented here, we have assumed that mutational and selectional parameters would be estimated independently, using established methods. However, separating the estimation of these parameters from the estimation of distances is clearly not optimal. Ideally, these parameters and a phylogeny would be estimated simultaneously. The likelihood functions described here make that possible in principle, although the computation may not currently be practical.

We feel that the method presented here will be of use in assessing relative evolutionary distances for data sets containing a mixture of highly divergent and less divergent pairs of sequences, whereas previous methods are likely to underestimate longer distances and thus underestimate differences between more and less divergent pairs. We also hope that the method will find use in evaluating the degree to which the more divergent clades of a tree have approached saturation, which should permit us to consider questions regarding the likelihood of increased divergence over time. Finally, the method may be helpful in assessing whether uncertainty regarding the inner branchings of a tree reflects a period of explosive diversification or loss of evolutionary signal in the noise of time.

Acknowledgments

We thank Stanley Sawyer and an anonymous referee for editorial input. Funding for A.L.H. was provided by the Sexually-Transmitted Diseases Branch of the National Institute of Allergy and Infectious Diseases and Family Health International; W.J.B. was supported by a grant from the Department of Energy Office of Biological and Environmental Research.

LITERATURE CITED


MYERS, G., A. HALPERN, C. BAKER, A. McBRIDE, C. WHEELER, and J. DOORBAR. 1996. Human papillomaviruses 1996, a compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, N.M.


STANLEY SAWYER, reviewing editor

Accepted April 8, 1998