An Empirical Codon Model for Protein Sequence Evolution

Carolin Kosiol,[1] Ian Holmes,† and Nick Goldman**

*European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom; and †Department of Bioengineering, University of California, Berkeley

In the past, 2 kinds of Markov models have been considered to describe protein sequence evolution. Codon-level models have been mechanistic with a small number of parameters designed to take into account features, such as transition–transversion bias, codon frequency bias, and synonymous–nonsynonymous amino acid substitution bias. Amino acid models have been empirical, attempting to summarize the replacement patterns observed in large quantities of data and not explicitly considering the distinct factors that shape protein evolution. We have estimated the first empirical codon model (ECM). Previous codon models assume that protein evolution proceeds only by successive single nucleotide substitutions, but our results indicate that model accuracy is significantly improved by incorporating instantaneous doublet and triplet changes. We also find that the affiliations between codons, the amino acid each encodes and the physicochemical properties of the amino acids are main factors driving the process of codon evolution. Neither multiple nucleotide changes nor the strong influence of the genetic code nor amino acids' physicochemical properties form a part of standard mechanistic models and their views of how codon evolution proceeds. We have implemented the ECM for likelihood-based phylogenetic analysis, and an assessment of its ability to describe protein evolution shows that it consistently outperforms comparable mechanistic codon models. We point out the biological interpretation of our ECM and possible consequences for studies of selection.

Introduction

Protein sequence evolution has been investigated on 2 data levels: amino acids and triplets of cDNA interpreted as codons. Amino acid sequences are popular because they evolve more slowly than DNA and are easier to align, and they are less prone to "saturation" effects that some phylogenetic inference methods handle poorly and because amino acid residue frequency biases are often less marked than DNA nucleotide frequency biases. However, DNA sequences contain more information, and studying protein evolution by modeling the evolutionary process on coding DNA is appealing because it allows us to take the genetic code into account.

There are 20 amino acids but 64 possible codons. Three amino acids—arginine, leucine and serine—are each encoded by 6 different codons, whereas another 5 can each be produced by 4 codons, which only differ in the third position. A further 9 amino acids are specified by a pair of codons which differ by a transition substitution at the third position, whereas isoleucine is produced by 3 different codons and methionine and tryptophan by only a single codon. Codon-level models are able to make distinctions between codons, which encode the same amino acid and those that do not. They also allow the study of whether there is a tendency for mutations maintaining the encoded amino acid (synonymous changes) to be accepted by selection less, equally, or more frequently than those that alter the amino acid (nonsynonymous changes). Thus, by introducing parameters describing the ratio of nonsynonymous to synonymous changes, it is possible to measure the effect of natural selection on the sequence.

Phylogenetic analyses using codon models have therefore become very popular, permitting in silico study of selective forces acting upon a protein that can be highly informative about its biological function and evolutionary history (Yang and Bielawski 2000). The interactions of proteins through their regulatory and metabolic networks are also reflected in the selection acting upon them: for example, it has been demonstrated that the more interactions a protein has with other molecules, the slower it evolves and that proteins operating in complexes (e.g., involved in translation or DNA repair) are, on average, more constrained than those with simple housekeeping functions (Aris-Brosou 2005).

Existing models that describe protein evolution at the amino acid and codon levels use Markov processes (Liò and Goldman 1998) and can be distinguished into 2 types. Empirical models do not explicitly consider biological factors that shape protein evolution but simply attempt to summarize the substitution patterns observed in large quantities of data. Typically used for amino acid level modeling, they describe substitution patterns by parameters representing the relative rates of replacements between amino acids; these parameters are an aggregated measure of all kinds of physicochemical properties of the amino acids and of their interaction with their local environment. Often empirical models have many such parameters, and these are typically estimated once from a large data set and subsequently reused with the assumption that they are applicable to a wide range of sequence data sets.

On the other hand, mechanistic models explicitly take into account features of the process of protein evolution such as selective pressures and the frequency of character states in the data (e.g., relative occurrence of different codons), allowing the testing of hypotheses related to these factors for each data set of interest. Typically, only a relatively small number of parameters is used; their values are not assumed to be widely applicable "constants" but are estimated afresh for each data set.

At the amino acid level, there is a long tradition of empirical amino acid models. Dayhoff et al. (Dayhoff and Eck 1968; Dayhoff et al. 1972, 1978) estimated the first amino acid models, resulting in the widely used point accepted mutations (PAM) matrices (see also Kosiol and Goldman 2005). Jones et al. (1992) employed much the same

[1] Present address: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York.

Key words: protein evolution, codon models, Markov models, maximum likelihood, phylogenetic inference.

E-mail: ck285@cornell.edu.

Mol. Biol. Evol. 24(7):1464–1479. 2007
doi:10.1093/molbev/msm064
Advance Access publication March 30, 2007

methods but based the estimation of the Jones-Taylor-Thornton (JTT) model on a larger sequence database; Whelan and Goldman (2001) used a maximum likelihood (ML) estimation technique to generate the Whelan and Goldman (WAG) model. The PAM, JTT, and WAG models give increasingly good descriptions of the "average" patterns and processes of evolution of large collections of sequences. Such average models can fail to describe proteins with particular functions and structures, however, and in various cases improved empirical amino acid models have been derived by estimating them from data sets representing particular functional and structural properties of the proteins (e.g., transmembrane proteins [Jones et al. 1994], different protein secondary structure contexts [Goldman et al. 1998], mitochondrially encoded proteins [Adachi and Hasegawa 1996], chloroplast-derived proteins [Adachi et al. 2000], and retroviral polymerase proteins [Dimmic et al. 2002]).

Purely mechanistic amino acid models are rare; they came much later than empirical amino acid models and were introduced to try to explain observed amino acid substitution patterns. Koshi et al. (1997) developed a mechanistic amino acid model, which incorporates the "fitness" of each of the amino acids, defined as a function of physicochemical properties of that amino acid. Their model, based on Boltzmann statistics and Metropolis kinetics (Metropolis et al. 1953), uses far fewer than the theoretical maximum of 380 adjustable parameters for a Markov process amino acid model, such that it is possible to optimize the model for each specific data set of protein sequences studied. Yang et al. (1998) reduced the mechanistic codon model M0 (see below) to a mechanistic amino acid model, enforcing the Markov property and reversibility. This "collapsed-codon" amino acid model performed significantly better when it also incorporated mechanistic parameters describing physicochemical properties.

Empirical amino acid models have also been combined with additional mechanistic parameters highly successfully. The "+F" method of Cao et al. (1994) allows the incorporation of the amino acid frequencies from a specific data set under study in place of those of the database from which the substitution matrix was estimated, and is now very widely used in phylogenetics. The inclusion of a Γ-distribution (Yang 1994b) containing a single biologically interpretable shape parameter that can accommodate varying degrees of heterogeneity of evolutionary rate among the sites of a protein has also been proven to improve the description of sequence evolution for many proteins (Goldman and Whelan 2002).

Codon models, on the other hand, are traditionally mechanistic, characterizing a Markov process using only a small number of parameters representing biologically relevant factors such as bias toward transition mutations, variability in codon frequencies, and, importantly, the tendency of mutations maintaining the encoded amino acid (synonymous changes) to be accepted by selection with a different probability from those changes that change the amino acid (nonsynonymous changes). A single parameter ω, the synonymous–nonsynonymous amino acid substitution rate ratio, is widely used to detect selection in proteins (Goldman and Yang 1994; Nielsen and Yang 1998; Yang and Bielawski 2000; Yang et al. 2000). Advanced codon models do not assume a single fixed ω but permit consideration of different ω values over sites (Yang et al. 2000; Wong et al. 2004; Massingham and Goldman 2005), lineages (Yang and Nielsen 1998), or both sites and lineages (Yang and Nielsen 2002). These models are popular for detecting proteins and individual sites in proteins undergoing positive selection (Nielsen and Yang 1998; Yang et al. 2000; Wong et al. 2004; Massingham and Goldman 2005).

All the codon models in common use make the assumption that every mutation alters just 1 nucleotide. Evolutionary change between codons varying in 2 or 3 nt are therefore necessarily interpreted as having arisen via a succession of single nucleotide changes. In contrast, Whelan and Goldman (2004) introduced a model including the same evolutionary factors as the standard mechanistic codon models but in addition allowing for instantaneous single, double, and triple nucleotide changes. Their results suggested that protein sequence evolution was better described by models that include significant proportions of double and triple changes. If this is correct, there could be important consequences for the application of codon models to detect selection—we address the question of instantaneous multiple nucleotide substitutions in detail in this paper.

The success of purely empirical models and combined mechanistic and empirical models on the amino acid level, for example, in database searches, alignment, and phylogenetic studies, suggests that empirical codon models (ECMs) could potentially be very useful for both understanding protein evolution and in phylogenetic applications. There has, however, been very little work in this area. ECMs are harder to estimate—they have a high number of parameters because they work on a 64 letter alphabet (61 if stop codons are discarded)—and application of methods analogous to those used to derive empirical amino acid models requires large amounts of protein-coding DNA sequence data not previously available in a convenient form. We know of only 1 example, by Schneider et al. (2005), in which a log-odds matrix is derived from codon sequences separated by a small evolutionary distance (time) and applied in an alignment program. However, although codon matrix of Schneider et al. is a step in the direction of an empirical model of codon sequence evolution, they only describe probabilities and log-odds values for codon substitutions for a particular set of evolutionary distances.

In this paper, we estimate an ECM from a large database of protein-coding DNA sequences. We then incorporate it in ML phylogenetic inference software to see if it gives a good description of protein evolution and may be generally useful for the phylogenetic analysis of particular proteins. We have implemented the ECM in combination with various mechanistic parameters, and our assessment of its utility for ML phylogenetics shows that it performs better than comparable existing models.

Materials and Methods

Standard Markov Models for Codon Sequence Evolution

Markov models of codon substitution were first proposed by Goldman and Yang (1994) and Muse and Gaut

(1994). We introduce these models by reference to the simple mechanistic model called M0 by Yang et al. (2000) (see also Goldman and Yang 1994). This model specifies the relative instantaneous substitution rate from codon i to codon j as:

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon or} \\ & i \rightarrow j \text{ requires} > 1 \text{ nt substitution,} \\ \pi_j & \text{if } i \rightarrow j \text{ is a synonymous transversion,} \\ \pi_j \kappa & \text{if } i \rightarrow j \text{ is a synonymous transition,} \\ \pi_j \omega_M & \text{if } i \rightarrow j \text{ is a nonsynonymous transversion,} \\ \pi_j \kappa \omega_M & \text{if } i \rightarrow j \text{ is a nonsynonymous transition.} \end{cases} \quad (1)$$

for all $i \neq j$, where parameter ω_M represents the nonsynonymous–synonymous rate ratio (the subscript M denoting the mechanistic M0 model), κ the transition–transversion rate ratio, and π_j the equilibrium frequency of codon j. Different assumptions can be made concerning π_j (Goldman and Yang 1994; Muse and Gaut 1994; Yang 1997). Here, we mostly consider the π_j as 61 parameters, independent apart from the constraint that their sum is 1 (i.e., the F61 parameterization; Yang 1997). In common with all Markov models of sequence evolution, absolute rates are found by normalizing the relative rates to a mean rate of 1 at equilibrium, that is, by enforcing $\sum_i \sum_{j \neq i} \pi_i q_{ij} = 1$ and completing the instantaneous rate matrix $Q = (q_{ij})$ by defining $q_{ii} = -\sum_{j \neq i} q_{ij}$ to give a form in which the transition probability matrix is calculated as $P(t) = e^{Qt}$ (Liò and Goldman 1998). Evolutionary times t are measured in expected numbers of nucleotide substitutions per codon.

Codon-level Markov models are typically used for ML phylogenetic inference. The model defines the likelihood for hypotheses consisting of values for all model parameters, a phylogenetic tree and its branch lengths (see, e.g., Felsenstein 1981; Goldman and Yang 1994; Liò and Goldman 1998; Felsenstein 2004), and this likelihood is then maximized over all hypotheses (parameter values) of interest. Codon models are increasingly used for estimating phylogenetic relationships, that is, the likelihood is maximized over tree shapes (Ren et al. 2005); otherwise, a good tree topology found by other means may be taken as known.

Models describing evolution at the codon level allow the estimation of measures of the selective forces acting on proteins. The ML estimate of the parameter describing the ratio of rates between nonsynonymous and synonymous substitutions, ω_M, is widely used as a direct measure of these forces. When there are few selective pressures acting, sequences are said to be evolving neutrally and the relative rates of fixation of synonymous and nonsynonymous mutations are roughly equal (ω_M is approx. 1). When a sequence has an important function, its sequence is highly conserved through evolution and ω_M takes a value substantially less than 1. Conversely, when sequences are under pressure to adapt quickly to their environment, nonsynonymous changes are strongly selected for and ω_M will take a value greater than 1.

The most advanced codon models do not assume a single fixed ω_M for all sites, but permit consideration of a distribution of values over sites. Yang et al. (2000) proposed and investigated a series of such models designated M0–M13 (the M-series). M7 is widely used, and describes among-site variation in ω_M with a β-distribution, allowing for purifying selection and neutral evolution only ($0 \leq \omega_M \leq 1$). Other models allow also for positive selection at some sites; for example, M8 contains the β-distribution of M7 and a single additional category of sites with ω_M permitted to be greater than 1. In this paper, implementations of our ECM do not attain this level of complexity, and we will concentrate on comparisons with M0 and M7 as defined in Yang et al. (2000).

Estimation of Empirical Models

Following Whelan and Goldman (2001), we use a ML approach to infer an empirical model from a data set of many multiple sequence alignments. We retain the mathematical and computational convenience that empirical models are often assumed to be reversible (Tavaré 1986; Yang 1994a; Felsenstein 2004). Under this assumption, instantaneous rates q_{ij} can be parameterized as

$$q_{ij} = \pi_j s_{ij} \text{ for all } i \neq j, \quad (2)$$

where the s_{ij}, often denoted exchangeabilities (Whelan and Goldman 2001), are symmetric ($s_{ij} = s_{ji}$) and π_j describes the equilibrium frequencies. For amino acid models, the instantaneous rate matrix can therefore be described by 208 independent terms, namely 189 exchangeabilities s_{ij} and 19 frequency parameters π_j. In general, the number of independent parameters for a reversible substitution model with N character states can be calculated as

$$\left[\frac{N^2 - N}{2} - 1\right] + [N - 1] = \frac{N(N+1)}{2} - 2, \quad (3)$$

where the 1st term in square brackets represents the exchangeabilities and the 2nd represents the state frequencies. Thus, to estimate a reversible ECM ($N = 61$), 1,889 independent parameters have to be determined.

Whelan and Goldman (2001) developed an approximate likelihood method that is based on the observation that the inference of parameters describing the evolutionary process remains stable across near-optimal tree topologies. This means that, so long as tree topologies and their branch lengths are close enough to optimal when estimating a new model, any minor inaccuracies will not influence the parameter estimates to any great extent (see also Sullivan et al. 1996; Abdo et al. 2005; Sullivan et al. 2005). Relying on this approximation, empirical model estimation proceeds by taking a large data set of many sequence alignments, each with an associated phylogenetic tree, and computing the likelihood of all these data as a function of the parameters s_{ij} and π_j. This likelihood is then maximized over the s_{ij} and π_j, taking the trees (topologies and branch lengths) as fixed.

In theory, it would be possible instead to fix only the relative branch lengths on a per-alignment basis, to reestimate all branch lengths, or even to reestimate all tree topologies and branch lengths during the estimation of the codon model. However, in practice this slows down the estimation considerably and experience from the estimation of WAG

(Whelan and Goldman 2001) shows it had little effect. Likewise, it would be possible to estimate a different set of the codon frequencies for every protein family. This would require another 60 parameters per protein family used. Again, we expect from the results of Whelan and Goldman (2001) that this would not improve the fit of the empirical model significantly.

The ML estimates, after normalization so the inferred Markov process has mean rate 1 at equilibrium, are denoted s^*_{ij} and π^*_j. We will refer to this model as ECM. Notice that in the context of codon models, we need to make no assumption that only single nucleotide changes occur. If required, this can be enforced by requiring $s^*_{ij}=0$ whenever codons i and j differ at more than 1 position.

Even using the approximation of Whelan and Goldman, an ML estimation of an ECM has previously seemed infeasible because of the computational burden of estimating 1,889 parameters and the lack of a suitable data set. The introduction of an expectation-maximization algorithm to ML training of substitution rate matrices by Holmes and Rubin (2002) has greatly speeded up the computations, now making it feasible to estimate an ECM from a database of multiple alignments and phylogenetic trees. Klosterman et al. (2006) provide a C++ implementation of this algorithm, XRATE, as part of the DART package. Robustness tests have confirmed the suitability of DART for the estimation of an ECM (Klosterman et al. 2006).

The Pandit Database

The large number of sequence alignments and phylogenies needed to estimate an ECM reliably were taken from the Pandit database of aligned protein domains (Whelan et al. 2003, 2006). Each family in Pandit includes an alignment of amino acid sequences and the corresponding alignment of the DNA sequences encoding the protein, and each alignment has an estimated phylogenetic tree associated with it (for full details, see Whelan et al. 2006).

For the estimation of an ECM only the DNA alignments and their inferred trees were utilized. Because the Pandit alignments vary in the quality of their reconstruction of homology, both within and between alignments, the profile hidden Markov model described by Whelan et al. (2006) was used to classify the columns in each alignment as being "reliable" or otherwise. All matrices were estimated using only reliable alignment columns. Further data cleaning (e.g., discarding additional codons neighboring gap regions, removing very short alignment fragments) did not noticeably change the substitution patterns of the ECMs estimated. After removing all families that could not be confidently classified as using the universal genetic code or that included any sequences with internal stop codons, we were left with 7,332 protein families from Pandit. These were used to estimate the ECM.

Pandit contains only trees based on DNA or amino acid data and not on codon data. We assumed that the DNA tree topologies were near optimal for codon-level analysis and that the branch lengths differ by just 1 scaling factor common to all alignments. This scaling factor is expected to be around 3 because there are 3 nt in a codon, and the branch lengths in the DNA trees are measured in expected number of substitutions per nucleotide site. However, the exact value of the scaling factor is irrelevant because the resulting instantaneous rate matrix is anyway normalized to mean rate 1.

For a more detailed analysis of the performance of the estimated ECM in phylogenetic analysis, a subset of 200 protein-coding DNA alignments and tree topologies was selected (see Supplementary Material online, http://www.ebi.ac.uk/goldman/ECM/ for details).

Statistical Comparison of Competing Models

We use likelihood ratio tests (LRTs) and the Akaike information criterion (AIC) to make statistical comparisons between competing codon models of protein evolution. Simply preferring the model with the highest likelihood may lead to the selection of 1 that is unnecessarily complex. For example, a more general model will always have a higher likelihood than a more restricted model nested within it. Statistical methods are required to balance model complexity against useful improvements in likelihood.

The LRT offers a very powerful way of comparing models (Silvey 1970), widely used in phylogenetics (Goldman 1993; Felsenstein 2004). It requires the formation of 2 competing hypotheses, H_0 and H_1, represented by models with different parameter constraints. The ML values ($\hat{L}$) for the competing hypotheses are compared using the LRT statistic

$$2\Delta=2\ln\left(\frac{\hat{L}_1}{\hat{L}_0}\right)=2(\ln(\hat{L}_1)-\ln(\hat{L}_0)). \qquad (4)$$

This statistic has very useful properties for significance testing (Silvey 1970). In straightforward cases, when H_0 can be formed by placing restrictions on the parameters in H_1, the hypotheses are said to be nested and for significance testing 2Δ can be compared (e.g.,) with the 95% point of a χ^2_n distribution (Felsenstein 2004), where n is the number of free parameters by which H_0 and H_1 differ (for more complex cases see Goldman 1993; Whelan and Goldman 1999; Goldman and Whelan 2000).

The AIC is an alternative method that reaches a compromise between goodness of fit and the complexity of models. It is particularly valuable when comparing multiple models and models that are not nested (Felsenstein 2004). The AIC for a hypothesis (in our application, a model) is computed by taking −2 times the maximum log-likelihood of the hypothesis and penalizing it by adding twice the number of free parameters. So, for hypothesis i with p_i free parameters,

$$\mathrm{AIC}_i=-2\ln\hat{L}_i+2p_i. \qquad (5)$$

Values of AIC_i are compared among hypotheses i with the model that has the lowest value of AIC preferred.

Application of the ECM

ECM could simply be used in the same way that the original Dayhoff, JTT, or WAG models (see above) can be used for amino acid sequences. However, for amino acid sequence evolution, past experience shows that the performance of empirical models can be significantly improved by combining them with mechanistic parameters. Existing mechanistic codon models are based on parameters describing codon frequencies π_i, transition–transversion bias κ, and nonsynonmous–synonymous bias ω. Additionally, we have seen in another study on whole-proteome data sets that codon substitution patterns vary strongly for sequences with different ω values (Kosil 2006). All this suggests that it will be beneficial to consider reintroducing mechanistic parameters π_i, κ, and ω.

Analogous to the definition of the mechanistic codon model M0 (eq. 1), we define the instantaneous rate matrix of the ECM with mechanistic parameters as

$$q_{ij}=\begin{cases}0 & \text{if } i \text{ or } j \text{ is a stop codon}\\ s^*_{ij}\pi_j\kappa(i,j) & \text{if } i\to j \text{ is a synonymous change}\\ s^*_{ij}\pi_j\kappa(i,j)\omega & \text{if } i\to j \text{ is a nonsynonymous change.}\end{cases} \quad (6)$$

where s^*_{ij} are the ECM exchangeabilities estimated from the Pandit database, π_j is the frequency of codon j estimated from each particular data set analyzed, $\kappa(i, j)$ is a term representing transition–transversion bias between codons i and j (see below), and ω represents nonsynonymous–synonymous bias. The instantaneous rate matrix $Q = (q_{ij})$ is again completed by defining $q_{ii}= -\sum_{j\neq i} q_{ij}$ and normalizing to mean rate 1. Note the use of the +F method (Cao et al. 1994) of replacing the database-wide codon frequency estimates π^*_j by a set of estimates π_j derived from each particular alignment studied (F61 model [Yang 1997]). We will denote the combined empirical and mechanistic model as ECM+F+ω+$n\kappa$, where different values of n will allow us to distinguish between model variants incorporating transition–transversion bias κ in different ways. There is no theoretical reason why the exchangeabilities s^*_{ij} should remain fixed while we reestimate the π_j for each family. However, in an alignment of 1 protein family, we often do not observe enough substitutions to infer the s_{ij} for each of the changes between codons i and j. In contrast, the reestimation of π_j is widely and successfully used in practice for nucleotide, amino acid, and codon models (see, e.g., Felsenstein 1981; Cao et al. 1994; Goldman and Yang 1994; Goldman and Whelan 2002). Note also that there is no requirement inherent in equation (6) that i and j differ at exactly 1 nucleotide position, as is required in the definition of the standard model M0 (eq. 1) and that evolutionary time is now measured in substitution events per codon.

In an ECM, the parameter ω can no longer be simply interpreted as a rate ratio. An ECM already reflects the average nonsynonymous–synonymous bias present in the proteins composing the database it was estimated from. Estimates obtained from mechanistic codon models, ω_M, and estimates from ECMs, ω, therefore cannot be compared directly: ω_M represents the absolute nonsynonymous–synonymous rate ratio, whereas ω measures the relative strength of selection with respect to an average level implicit in the Pandit database. To make a valid comparison, we need to disentangle estimated values of ω from the expected value under neutral evolution.

To do this, we take an approach that was pursued in the early mechanistic codon model of Goldman and Yang (1994). There, the ratio of the instantaneous rates per codon of nonsynonymous and synonymous nucleotide substitutions is calculated as ρ_a/ρ_s, where the nonsynonymous substitution rate is given by

$$\rho_a=\sum_i \sum_{\substack{j\neq i\\ \mathrm{aa}_j\neq \mathrm{aa}_i}} \pi_i q_{ij} \quad (7)$$

(aa_i indicates the amino acid encoded by codon i), and the synonymous rate per codon can be calculated as $\rho_s = 1 - \rho_a$ because the overall rate is normalized to 1. We also take the values $\rho_a^{\text{neutral}}=0.79$ and $\rho_s^{\text{neutral}}=0.21$, derived by Nei and Gojobori (1986) as typical values for neutrally evolving proteins. Thus the "corrected" nonsynonymous–synonymous rate ratio ω_E is given by

$$\omega_E=\frac{\rho_a\rho_s^{\text{neutral}}}{\rho_s\rho_a^{\text{neutral}}} \quad (8)$$

and can be directly compared with estimates ω_M from mechanistic models. Note that ω_E depends on ω through ρ_s and ρ_a, themselves functions of the q_{ij} (eq. 7) which depend on ω (eq. 6).

Similarly, our expression $\kappa(i, j)$ in equation (6) represents a measure of the relative strength of the transition–transversion bias with respect to the average level implicit in the Pandit database. Whereas the transition–tranversion bias is traditionally modeled by a single parameter, permitting double and triple nucleotide changes in the ECM leads to new scenarios in addition to the single transitions or single transversions inherent in single nucleotide changes. The 9 possible ways to combine transitions (ts) and transversions (tv) in multiple nucleotide changes within 1 codon are as follows:

$$\text{1 nucleotide change}: (1\text{ts}, 0\text{tv}), (0\text{ts}, 1\text{tv}); \quad (9)$$

$$\text{2 nucleotide changes}: (2\text{ts}, 0\text{tv}), (1\text{ts}, 1\text{tv}), (0\text{ts}, 2\text{tv}); \quad (10)$$

$$\text{3 nucleotide changes}: (3\text{ts}, 0\text{tv}), (2\text{ts}, 1\text{tv}), (1\text{ts}, 2\text{tv}), (0\text{ts}, 3\text{tv}). \quad (11)$$

As a consequence, transition–transversion bias may now be modeled as a function $\kappa(i, j)$ that depends on the numbers of transitions (n_{ts}) and transversions (n_{tv}) of the change from codon i to codon j.

Here, we describe the 6 formulations for $\kappa(i, j)$ that are most interesting or successful out of a larger set of relationships devised and studied without preassumptions about what might best fit real sequence data (see Supplementary Material online, http://www.ebi.ac.uk/goldman/ECM/).

- ECM+F+ω: The factor κ is set to 1 for all changes:

$$\kappa(i,j)=1$$

This model assumes that transition–transversion bias is fully accounted for by the Pandit exchangeabilities s^*_{ij} and does not vary significantly from one protein to another.

- ECM+F+ω+1κ(ts) and ECM+F+ω+1κ(tv):
- ECM+F+ω+1κ(ts) is similar to existing mechanistic codon models and considers that the biasing effect introduced by multiple transitions may be multiplicative:

$$\kappa(i,j)=\kappa^{n_{\mathrm{ts}}}.$$

In standard mechanistic codon models n_{ts} is necessarily 0 or 1 and we expect $\kappa > 1$. In our model, these constraints disappear because multiple nucleotide changes are permitted ($n_{\mathrm{ts}} = 0$, 1, 2, or 3) and κ is a measure relative to the value implicit in the s^*_{ij}.

- – ECM+F+ω+1κ(tv) is similar to ECM+F+ω+1κ(ts), except that it focuses on transversions. This is unusual, but perhaps more natural in the same way that the standard ω parameter is generally considered a "rate reducing" effect:

$$\kappa(i,j)=\kappa^{n_{\mathrm{tv}}}.$$

- ECM+F+ω+2κ: In this model, transitions and transversions are modeled with individual parameters (κ_1 for transitions and κ_2 for transversions) and the effect is seen as multiplicative in terms of the relative rates:

$$\kappa(i,j)=\kappa_1^{n_{\mathrm{ts}}}\kappa_2^{n_{\mathrm{tv}}}.$$

- ECM+F+ω+9κ: In this model, each of the 9 possible cases (listed in eqs. 9–11 above) is modeled by an individual rate-modifying parameter ($\kappa_1 - \kappa_9$). Note that because of the overall rate normalization, this model is equivalent to 1 with just 8 independent κ parameters.

Note that ECM+F+ω is nested in all the other models. The (ts) and (tv) variants of ECM+F+ω+1κ are each nested in ECM+F+ω+2κ, and all 3 of these models are nested in ECM+F+ω+9κ.

The ECMs introduced in this section were incorporated into the program codeml from release 3.14b of PAML, a software package for ML phylogenetic analysis of DNA and protein sequences written and maintained by Yang (1997). For each data set analyzed, free parameters of the models (π_j, ω, and appropriate κ parameters as described above) were estimated by ML, as were branch lengths of trees. Tree topologies from the Pandit database were assumed correct.

Results and Discussion

Empirical Codon Models Estimated from Pandit

We estimated instantaneous rate matrices from the entire collection of 7,332 protein families taken from Pandit as described above. Figure 1 illustrates ECMs in the form of "bubble plots." The areas of the bubbles represent the rates of instantaneous change $(q^*_{ij}=\pi^*_j s^*_{ij})$, with the gray bubble in the upper left corner showing the area representing an instantaneous rate of 0.5. The rate matrices are not symmetric because the codons have different frequencies. The codons are listed to the left and top, and amino acid translations are given on the bottom and right (see also Klosterman et al. 2006).

Figure 1A shows the instantaneous rate matrix permitting all single, double, and triple nucleotide changes, inferred as in Estimation of Empirical Models. For this matrix, denoted "unrest" to indicate unrestricted optimization of all exchangeability parameters, 1,889 parameters were estimated. The ML obtained was $\ln L_{\mathrm{unrest}} = -9.157731 \times 10^7$.

DART also enabled us to restrict the estimated rate matrix to single nucleotide changes only (i.e., enforcing $s^*_{ij}=0$ unless codons i and j differ by exactly 1 nucleotide). Figure 1B shows the bubble plot of the optimal instantaneous rate matrix restricted (rest) in this way. For this matrix, 322 parameters were estimated, and the ML obtained was $\ln L_{\mathrm{rest}} = -9.343274 \times 10^7$. The matrices illustrated in figure 1 are available in the Supplementary Material online (http://www.ebi.ac.uk/goldman/ECM/).

There has been some debate about the existence and level of multiple nucleotide changes (Averof et al. 2000; Smith et al. 2003; Bazykin et al. 2004; Whelan and Goldman 2004). Possible biological mechanisms for changes in 2 neighboring nucleotides, for example, dipyrimidine lesions induced by ultraviolet light and template-directed mutations during DNA repair and replication, have been pointed out (Averof et al. 2000). However, their effect on evolutionary substitution patterns is likely to be small. Comparing figure 1A with 1B by eye, the existence of multiple nucleotide changes (blue and green bubbles) in the unrestricted model is quite striking. The fact that instantaneous rate matrices are normalized to mean rate 1 allows us to calculate the proportions of single, double, and triple changes (ρ_S, ρ_D, and ρ_T, respectively) in a straightforward manner. Defining S, D, and T to be the sets of codon pairs (i, j) differing by a single nucleotide change, a double change and a triple change, respectively, then we observe:

$$\rho_S=\sum_{(i,j)\in S}\pi^*_i q^*_{ij}=0.753,\ \rho_D=\sum_{(i,j)\in D}\pi^*_i q^*_{ij}=0.212,$$
$$\rho_T=\sum_{(i,j)\in T}\pi^*_i q^*_{ij}=0.035.$$

In other words, we observe 75.3% single, 21.2% double, and 3.5% triple changes.

We performed a LRT between the restricted and unrestricted ECMs to see if the addition of double and triple changes was statistically significant. Comparing the statistic $2\Delta = 2(\ln L_{\mathrm{unrest}} - \ln L_{\mathrm{rest}}) = 3.71 \times 10^6$ (eq. 4) with a χ^2_{1567} distribution, we see this is highly significant; the P-value is too small to be calculated reliably. This means that the codon substitution patterns in the Pandit data set are overwhelmingly better explained by a model that allows for multiple nucleotide changes to occur instantaneously, rather than only via successive single changes.

We also estimated rate matrices restricted to single and double, or single and triple, changes only. The ML calculated for an instantaneous rate matrix restricted to single and double changes is $\ln L = -9.167463 \times 10^7$ (75.3% single and 24.7% double changes) and that for a matrix restricted

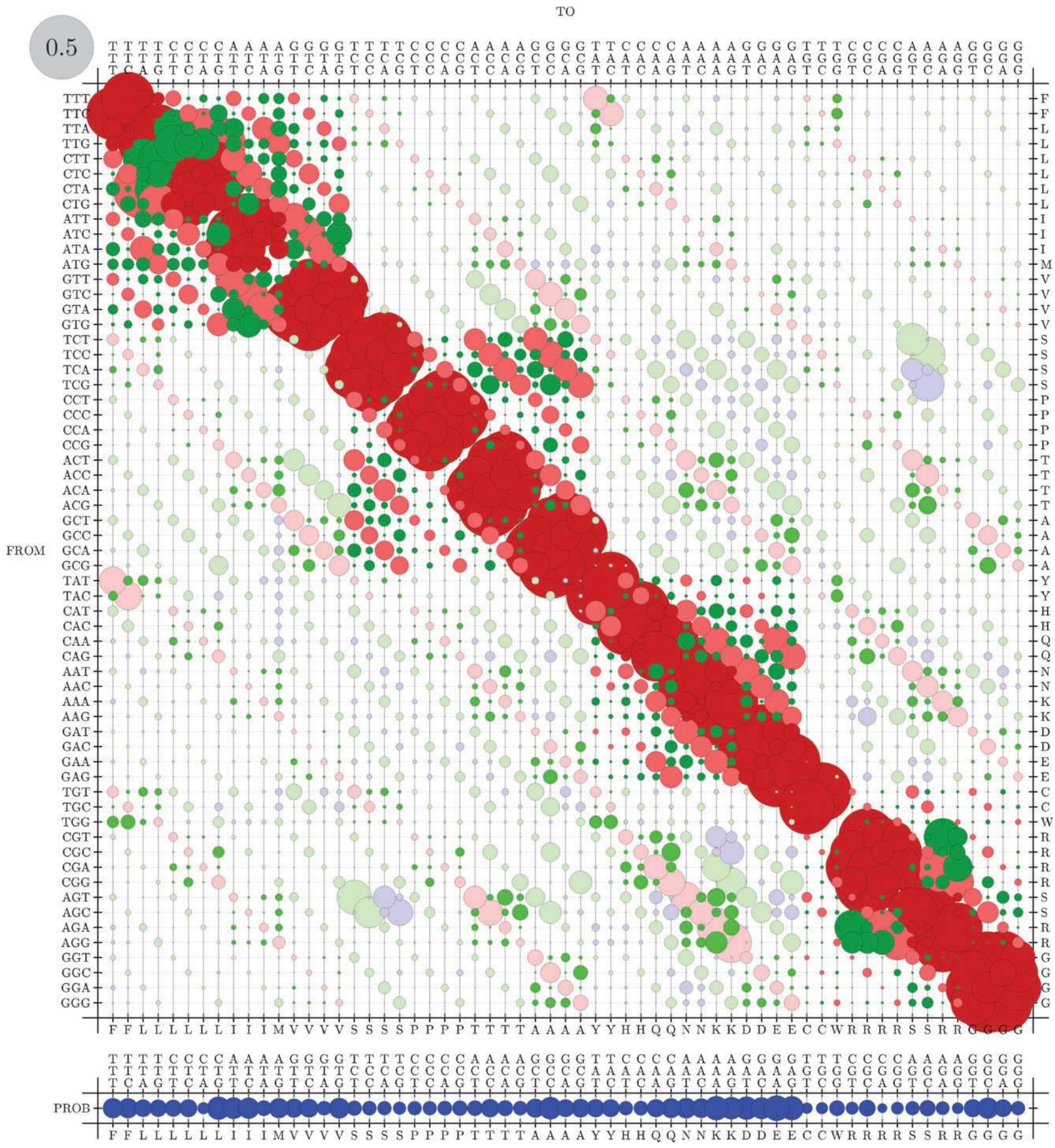

Fig. 1A.—Bubble plots of ECMs for the Pandit data set. Codons are ordered according to Urbina et al. (2006).

to single and triple changes is ln $L = -9.195009 \times 10^7$ (88.3% single and 11.7% triple changes). Appropriate LRTs indicate that the introduction of either double or triple changes to the restricted model permitting single changes only is a significant improvement, as is the subsequent addition of triple or double, as appropriate changes. In brief, our statistical tests confirm that both double and triple changes are making a significant contribution to the fit of the ECM to the evolution of the proteins represented in the Pandit data sets.

A further illustration of the importance of double and triple nucleotide changes is given in figure 2. Here, we present histograms of the magnitudes of the instantaneous rates q^*_{ij} from the ECM for all double and triple nucleotide changes $i \rightarrow j$. These are compared with corresponding histograms from a simulation study in which data conforming to M0, that is, with no double or triple changes, were analyzed using the same methods (see Supplementary Material online for further details). Whereas DART was able to recover M0 well (note that very few nonzero rates

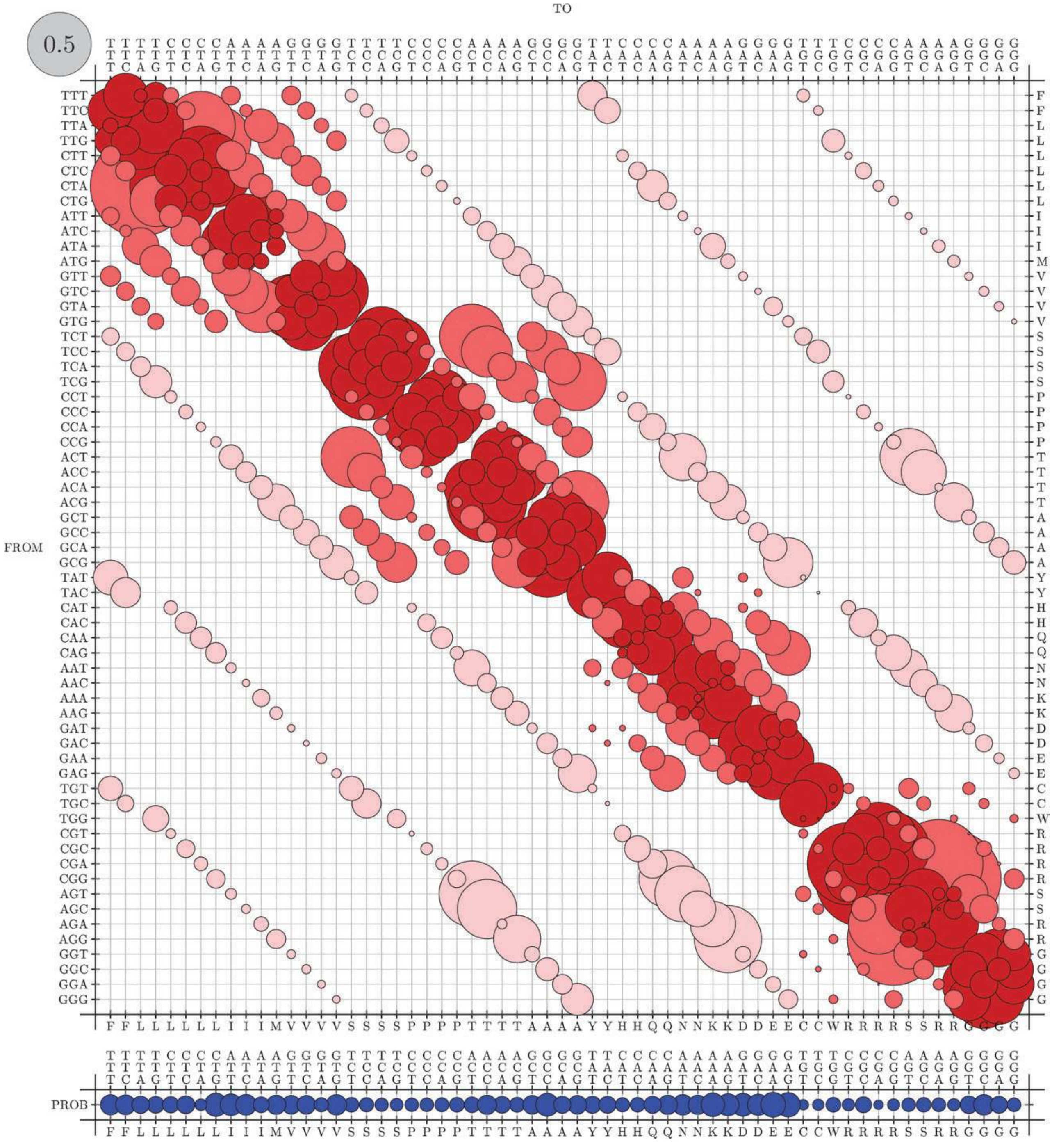

FIG. 1B. (Continued).

were estimated for double changes and virtually none for triple changes), the majority of the double and triple nucleotide changes estimated from the Pandit data sets are well above these estimation errors. This confirms that our methodology and the DART software can accurately recover zero rates when these do exist; therefore, we can trust the small but nonzero rates observed for multiple nucleotide changes in real data (e.g., in fig. 1A) to be genuine and not an artifact.

Physicochemical Interpretation of ECM

Apart from the observation of the existence of multiple nucleotide changes, it is quite difficult to extract biologically relevant information from all 61 × 61 matrix elements at once. The almost invariant sets (AIS) algorithm (Kosiol et al. 2004) is a method to summarize the information of Markov substitution models by analyzing their instantaneous rate matrices. It is a grouping method that identifies

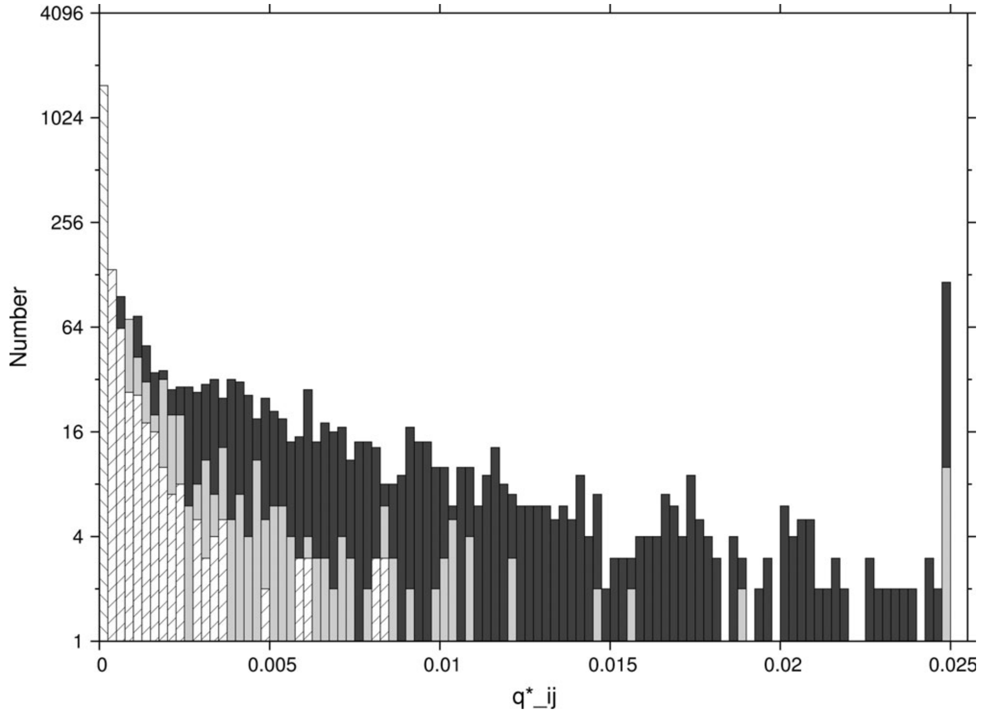

FIG. 2.—Histogram comparing instantaneous rates estimated from the Pandit data and from simulated M0 data. Note the logarithmic scale on the *y* axis. For the ECM estimated from the Pandit database the dark gray bars show the distribution of values of instantaneous rates of double nucleotide changes and light gray bars represent the rates of triple changes. For the model estimated from M0 simulated data, upward stripes (/) indicate double changes and downward stripes (\) triple changes, respectively.

disjoint sets with high rates of change between elements of each set but small rates of change between elements of different sets. This gives a quantitative method of identifying subsets of the states of models within which interchanges occur readily but between which interchanges are relatively uncommon. Table 1 shows the results of applying AIS to the unrestricted ECM derived in ECM Estimated from Pandit and, for comparative purposes, to the mechanistic codon model M0 and the WAG amino acid model.

For the ECM, a natural grouping to consider is the division into 20 subsets. This perfectly separates the 61 codons according to the amino acids they encode, that is, in

Table 1
Application of the AIS Algorithm to the ECM, the M0, and the WAG Amino Acid Model

Empirical Codon Model (ECM)		Mechanistic Codon Model (M0)		Empirical AA Model (WAG)
20 subsets	7 subsets	20 subsets	7 subsets	7 subsets
{W}	**{W}**	**{W}**	**{W}**	**{W}**
{YY}		**{YY}**		
{FF}	**{YY FF}**	**{FF**(TTY) **LL**(CTY)**}**	**{FF LLLLLL}**	**{Y F}**
{LLLLLL}		**{LL**(CTR) **LL**(TTR)**}**		
{M}	**{LLLLLL M**	**{M}**	**{M III VVVV**	**{L M I}**
{III}	**II VVVV}**	**{III}**	**EE DD QQ KK}**	
{VVVV}		**{VVVV}**		
{CC}	**{CC}**	**{CC}**		**{V C}**
{TTTT}		**{TTTT}**	**{CC TTTT**	
{SSSSSS}		**{SSSS**(TCN)**}**	**SS**(AGY)	
{AAAA}	**{TTTT**	**{SS**(AGY) **RR**(AGR)**}**	**AAAA NN**	**{T**
{EE}	**SSSSSS**	**{AAAA}**	**RR**(AGR)	**S**
{DD}	**AAAA EE DD**	**{EE**(GAY) **DD**(GAR)**}**	**GGGG}**	**A E D**
{NN}	**NN QQ KK**	**{NN}**		**N Q K**
{QQ}	**RRRRRR**	**{QQ}**		**R**
{KK}	**HH}**	**{KK}**		**H}**
{RRRRRR}		**{RRRR**(CGN)**}**	**{RRRR**(CGN)**}**	
{HH}		**{HH}**	**{HH YY}**	
{GGGG}	**{GGGG}**	**{GGGG}**		**{G}**
{PPPP}	**{PPPP}**	**{PPPP}**	**{PPPP SSSS**(TCN)**}**	**{P}**

NOTE.—For clarity the codons are generally represented by the amino acid they encode. Where informative, codons are also given, with R = purine, Y = pyrimidine, N = any base. Boldface distinguishes amino acids from codons.

perfect agreement with the genetic code (table 1, ECM, 20 subsets). This recovery of the genetic code is in itself a remarkable result and shows that amino acid identity is highly relevant to codon substitution patterns.

A division into 7 subsets is also interesting as it is easily compared with results from studies on amino acid models (Kosiol et al. 2004). This leads to a result very similar to the corresponding grouping of the (empirical) WAG amino acid replacement matrix (table 1, ECM, 7 subsets cf. WAG, 7 subsets). This similarity is particularly striking as the 2 models were estimated from very different data sets (see Whelan and Goldman [2001]; Whelan et al. [2006]) and with 1 data set interpreted at the amino acid level and the other at the codon level. The grouping derived from the ECM has the following, biochemically reasonable, interpretation. The codons encoding hydrophilic and basic amino acids (T, S, A, E, D, N, Q, K, R, H) are grouped together as are the codons encoding the aromatics (Y, F). Four amino acids (W, C, G, P) each have a group consisting of only their codons; these singletons appear to be the most conserved amino acids. All codons of the aliphatics (L, M, I, V) form 1 group. In the grouping derived from the WAG model, the only difference is that valine (V) is removed from the aliphatic group and placed instead with cysteine (C).

We have investigated whether the alignment algorithms underlying the Pandit data sets could have added bias toward these results. Pandit alignments are performed on the proteins' amino acid sequences, and we wondered whether amino acid sequence alignments could be biased toward aligning nonhomologous residues because of chance amino acid identity or physicochemical similarity. If so, we would expect this effect to be strongest in hard to align regions. Our results using stricter criteria for removing uncertain alignment regions (see above) show no significant differences, however. Additionally, in a study of proteomic data sets, we have compared results from sequences aligned on the amino acid level and on the DNA level, and again no significant differences were observed (Kosiol and Goldman, in preparation).

Although instantaneous rate matrices estimated from DNA alignments might suffer from different artifacts, they should not suffer from the same alignment artifacts as matrices estimated from amino acid alignments. Thus, the observation that both matrices show strong influence of the genetic code and physicochemical properties indicates that these observed substitution patterns are not artifacts of the alignment program used.

Applying the AIS algorithm to an instantaneous rate matrix defined by the M0 model (see Supplementary Material online) reveals quite different groups (table 1, M0). In particular, transition–transversion differences seem to play an overly important role with too little importance placed on the identity or physicochemical properties of encoded amino acids. In the grouping into 20 subsets, for example, codons encoding phenylalanine (F) share a group with some of the leucine (L) codons. Likewise, the codons of serine (S) and arginine (R) are each split over 2 groups. For the grouping of M0 into 7 subsets, the groups contain codons coding for mixtures of amino acids with very different physicochemical properties (e.g., {M, I, V, E, D, Q, K}), and the codons encoding serine and arginine remain separated. In particular, we note that the serine codons AGY are grouped with threonine (T; ACN) and alanine (A; GCN), but the TCN serine codons (only differing by 1 nt from threonine and alanine) are not. Instead, these are placed with proline (P; CCN) that is also only separated by 1-nt substitution, but is physicochemically quite different. Because the AIS grouping is purely based on replacement rates and not amino acid properties, the discrepancies observed between groupings and physicochemical properties can be interpreted as a failure of M0 to reflect evolutionary pressures. In contrast to ECM, the M0 results are difficult to interpret in a biologically meaningful manner. Note that these patterns are not fully dictated by inferred evolutionary dynamics but are to a large degree influenced by the parametric form enforced in this model (eq. 1).

In contrast, the "rediscovery" of the genetic code and the detection of biologically meaningful groupings based on amino acids' physicochemical properties, both found from purely evolutionary patterns in the ECM, indicate that these are highly significant in determining the dynamics of evolutionary change in protein sequences. These factors are at best poorly incorporated in existing mechanistic codon models. Although physicochemical properties were introduced in early codon models by Goldman and Yang (1994), based on the Grantham matrix (Grantham 1974), they were subsequently omitted from further developments of these models (e.g., Nielsen and Yang 1998; Yang et al. 2000). Massingham (2002) used large quantities of data to estimate empirical exchangeability parameters, finding that different amino acid pairs have different tendencies to replace one another over evolutionary time and that using these parameters in an evolutionary model gave significant improvements for many data sets.

Recently, Higgs et al. (2007) developed a mechanistic codon model that incorporates distances reflecting amino acid properties and allows for multiple nucleotide changes. They found that variants that do not include double and triple substitutions perform worse. Our empirical codon matrix gives further evidence that a much finer distinction than simply considering whether evolving codons are synonymous or nonsynonymous is important to accurate modeling of protein evolution. A major application of codon models is the detection of selection, and it is likely that these findings will also have consequences for selection studies.

ML Performance Analysis

We next consider whether our implementation of the ECM, in combination with mechanistic parameters as described in Application of the ECM, performs well in phylogenetic analysis of individual protein-coding DNA alignments.

A small preliminary study showed that among our $\kappa(i, j)$-model variants, the likelihood score of the ECM+F+ω+9κ was always best, but the improvement it gave in likelihood values over any of the less parameter-rich κ-models was never significant. This clearly indicates that ECM+F+ω+9κ is overparameterized and, consequently, the ML analyses we present focus on 0κ-, 1κ-, and 2κ-models. We compare these to each other and to the

Table 2
Log-Likelihood Values for 4 Protein Families under Different Mechanistic Models and ECMs

	Family (Pandit ID)			
Model	PF01226	PF01229	PF01231	PF01233
M0	−5659.72	−6718.81	−5430.65	−2400.04
M7	−5656.22	−6682.72	−5386.97	−2375.59
ECM	−5604.26	−6680.39	−5369.42	−2340.62
ECM+F	−5521.26	−6618.70	−5291.26	−2335.20
Improvement over ECM[a]	83.00**	61.69**	78.16**	5.42
ECM+F+ω	−5499.90	−6604.24	−5291.25	−2285.63
Improvement over M0[b]	159.82	114.57	139.40	114.41
Improvement over M7[b]	156.32	78.48	89.96	54.99
Improvement over ECM+F[a]	21.36**	14.46**	0.01	49.57**
ECM+F+ω+1κ(ts)	−5499.58	−6601.98	−5289.41	−2285.54
Improvement over M0[b]	160.14	116.83	141.24	114.50
Improvement over M7[b]	156.64	80.47	97.56	90.05
Improvement over ECM+F+ω[a]	0.32	2.26*	1.84	0.09
ECM+F+ω+1κ(tv)	−5499.56	−6596.51	−5287.64	−2285.23
Improvement over M0[b]	160.16	122.30	143.01	114.81
Improvement over M7[b]	156.66	86.21	99.33	90.36
Improvement over ECM + F + ω[a]	0.34	7.73**	3.61**	0.40
Improvement over ECM + F + ω + 1κ(ts)[b]	0.02	5.47	1.77	0.31
ECM+F+ω+2κ	−5499.53	−6595.48	−5287.55	−2285.13
Improvement over M0[b]	160.19	123.33	143.10	114.91
Improvement over M7[b]	156.69	87.24	99.42	90.46
Improvement over ECM+F+ω[a]	0.37	8.76**	3.70*	0.50
Improvement over ECM+F+ω+1κ(ts)[a]	0.05	6.05**	1.86	0.41
Improvement over ECM+F+ω+1κ(tv)[a]	0.03	1.03	0.09	0.10

[a] For nested models, asterisks indicate statistically significant increases in likelihood (*$P < 0.05$, $\chi^2_{1,0.05}=3.84$, $\chi^2_{2,0.05}=5.99$, and $\chi^2_{60,0.05}=79.08$ and **$P < 0.01$, $\chi^2_{1,0.01}=6.63$, $\chi^2_{2,0.01}=9.21$, and $\chi^2_{60,0.05}=88.38$).

[b] For nonnested models, the AIC prefers the model with higher likelihood in all cases shown.

mechanistic models M0, M7 (Yang et al. 2000), and single doublet triplet (SDT) model (Whelan and Goldman 2004; see also Comparison of ECM Variants).

We calculated the MLs for 200 protein family cDNA alignments under different variants of ECM and also under M0, M7, and SDT. Table 2 shows the results for 4 representative families, and table 3 summarizes the results of the full test set of 200 families. A brief note on the use of LRT and AIC in this context is in order: the exchangeability parameters s^*_{ij} are interpreted as fixed although they have in fact been estimated from 7,332 protein families, 1 of which is the protein family under investigation. One way to avoid this problem would be to reestimate another 200 ECMs, each time removing the test family from the database of 7,332 protein families. However, this would be impractically time-consuming, and it is highly unlikely that any 1 of the protein families could influence the overall estimation of the ECM enough to create a detectable bias.

Comparison of ECM Variants

First, we assess the performance of the unmodified ECM and of ECM+F for 200 protein families. For ECM+F, the 61 codon frequencies can be described by 60 additional free parameters because of the constraint $\sum_j \pi_j = 1$. Using the LRT described in Statistical Comparison of Competing Models, we test for significance using a χ^2_{60} distribution. Table 2 illustrates this LRT for 4 test data sets and shows the improvement of ECM+F over ECM to be significant in 3 cases at the 0.01 significance level. In

Table 3
Comparison of Codon Models over 200 Protein-Coding DNA Data Sets

	ECM				
	+F	+F+ω	+F+ω+1κ(ts)	+F+ω+1κ(tv)	+F+ω+2κ
M0	200 (n/a)	200 (n/a)	200 (n/a)	200 (n/a)	200 (n/a)
M7	197 (n/a)	200 (n/a)	200 (n/a)	200 (n/a)	200 (n/a)
ECM	70 (111)	123 (152)	125 (156)	131 (158)	132 (159)
ECM+F		184 (181)	191 (186)	195 (194)	196 (188)
ECM+F+ω			84 (62)	140 (109)	134 (117)
ECM+F+ω+1κ(ts)				200 (n/a)	143 (121)
ECM+F+ω+1κ(tv)					89 (73)

NOTE.—The table gives the number of protein families for which the model indicated by the column labels (hypothesis H_1 in LRTs) is significantly better than the model given by the row labels (H_0 in LRTs). The upper number given for each model comparison corresponds to AIC results; for nested models, results of LRTs are below, in parentheses (otherwise the LRT was not applicable [n/a]).

table 3, we confirm that for the majority of the test cases (111 out of the 200) a per-data set estimation of π_i improves the fit of the ECM significantly ($P < 0.05$). Because the +F modeling of frequencies is often good and following its almost universal acceptance in DNA, amino acid, and codon models, we adopt its use throughout the rest of this paper.

We then investigated the value of introducing the mechanistic parameters ω and κ(i, j) (eq. 6). To confirm the value of ω, a suitable test is to compare (hypothesis H_0) ECM+F+ω with (H_1) ECM+F, by which we mean the same model but with the additional constraint ω = 1. This, in effect, removes ω from equation (6) and assumes that the effects of natural selection are adequately described for all proteins by the exchangeabilities s^*_{ij} estimated from the 7,738 Pandit data sets. Table 2 illustrates this LRT for 4 test data sets and shows the introduction of ω to be significant ($P < 0.01$) in 3 cases. Furthermore, we found in 181 out of 200 test cases (see table 3) a significant improvement, confirming that per-data set estimation of ω is highly valuable in the ECM. All applications of the ECM discussed from now on include the parameter ω.

The relative success of the different transition–transversion bias models was also assessed by likelihood-based tests. Here, results are less clear. Table 2 illustrates cases where ECM+F+ω seems to have adequately captured the transition–transversion bias (PF01226, PF01233), where ECM+F+ω+1κ(tv) is clearly preferred (data set PF01231) and where all +1κ- and +2κ-variants appear to perform well (PF01229). The results from all 200 test data sets confirm this pattern (table 3). There is no clear-cut leader among our κ-models, although it is interesting to note that of the +1κ-models, the (tv) variant is always preferred to the (ts) variant that is more similar to the formulation used in existing mechanistic codon models.

These results suggest that much of the transition–transversion bias effect is common to many proteins studied and is quite well modeled by the bias already implicitly captured by the parameters s^*_{ij}. The small observed residual effect (i.e., some variation in preferred κ-model over data sets) suggests that maybe some slight extra transition–transversion effect was detected, which is varying between data sets and is possibly not very well modeled by our κ-models. We investigated whether the small effect measured by the κ-models could be capturing some other variation as transition–transversion bias varies both at the level of organisms and genes (e.g., mitochondrially encoded proteins are known to have elevated levels of bias [Brown et al. 1982]). For families that had unusually improved likelihoods under some κ-models, we checked the Pfam annotation (Bateman et al. 2004) for any unusual features but could identify no relationships between the organisms or genes and likelihood performance.

Comparison of ECM with M0 and M7

Having confirmed the ECM with mechanistic parameters ω and κ(i, j) introduced (eq. 6) worthy of further consideration, our main aim is to see how the ECM fares in comparison with comparable existing mechanistic codon models. Table 2 illustrates that the log-likelihoods of M0 and M7 were lower than under any of the ECM+F+ω+nκ variants of the ECM, significantly so according to the AIC test. This result was confirmed across each of the 200 test data sets (table 3).

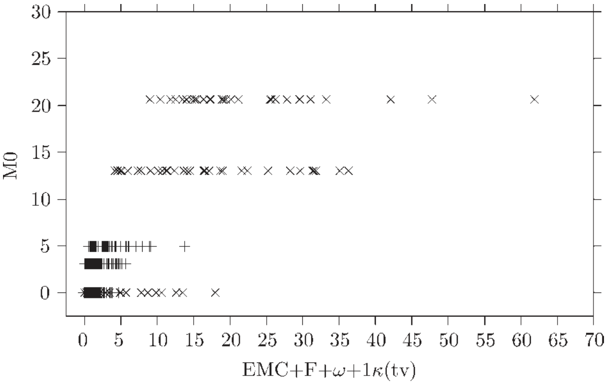

FIG. 3.—Instantaneous rates, adjusted for codon frequencies, from ECM+F+ω+1κ(tv) and M0 estimated for protein family PF01231. These are calculated as q_{ij}/π_j from equations (1) (M0) and (6) ECM+F+ω+1κ(tv). Rates of nonsynonymous changes are represented by +, rates of synonymous changes by ×.

These results indicate that the ECM gives a very much more accurate description of the observed patterns of protein-coding DNA sequence evolution than do the models M0 and M7. Figure 3 illustrates a comparison of instantaneous rates, adjusted for codon frequencies, from M0 and ECM. Although M0 gives only 5 values (0, for multiple nucleotide substitutions, and 4 other values arising from its mechanistic transition–transversion bias and nonsynonymous–synonymous bias parameters), ECM takes many different values, over a wider range, reflecting much finer distinctions being made (including differences in nonsynonymous changes originating from amino acid properties).

The improvement of ECM+F over M7 in 197 out of 200 cases is particularly impressive because that M7 permits variation of nonsynonymous–synonymous bias among sites, whereas ECM+F does not even have a family-specific parameter ω. Given the existing success of M7 and variants of it for phylogenetic inference and, particularly, analysis of natural selection, our results argue very strongly in favor of the use of the ECM and its future development.

Comparison of ECM with SDT

We also compared the ECM to the mechanistic SDT model (Whelan and Goldman 2004). The SDT model describes protein-coding sequence evolution at the codon level, allowing for single, double, and triple substitutions both within codons and spanning codon boundaries. The SDT model's parameters, estimated on a per-data set basis, describe the proportions of single, double, and triple changes, transition–transversion bias on the nucleotide level, nonsynoymous–synonymous substitution biases and codon frequencies (for full details, see Whelan and Goldman 2004).

To make a fair comparison with SDT, we need to change the method used to parameterize codon frequencies within the ECM. The SDT model, in common with the

Table 4
Log-Likelihood Values for Protein Families from Pandit under Different Mechanistic Models and ECMs, using the F1×4MG Parameterization of Codon Frequencies

	Family (Pandit ID)			
Model	PF01056	PF01226	PF01229	PF01231
M0	−5483.54	−5853.10	−6865.90	−5567.11
SDT	−5360.42	−5771.16	−6818.44	−5508.95
Improvement over M0[a]	123.12	81.94	47.46	58.16
ECM+F+ω	−5397.32	−5697.21	−6770.95	−5451.76
Improvement over SDT[a]	−36.90	73.95	47.49	57.19
ECM+F+ω+1κ(ts)	−5392.67	−5697.19	−6765.33	−5449.78
Improvement over SDT[a]	−32.25	73.97	53.11	59.17
ECM+F+ω+1κ(tv)	−5373.78	−5696.95	−6753.29	−5445.28
Improvement over SDT[a]	−13.36	74.21	65.15	63.67
ECM+F+ω+2κ	−5367.29	−5696.79	−6750.14	−5543.90
Improvement over SDT[a]	−6.87	74.37	68.30	65.05

[a] For nonnested models, the AIC prefers the model with higher likelihood in all cases shown.

model of Muse and Gaut (1994), assumes that the instantaneous rates of change are proportional to the frequency of the replacement nucleotides and not the replacement codon. This parameterization for codon frequencies, referred to as F1×4MG (Yang 1997), was implemented in our ECM for comparisons with SDT. It is already available in M0 in the codeml program (Yang 1997).

Comparison with the SDT model was restricted to a total of 15 families, corresponding to those analyzed by Whelan and Goldman (2004) and whose DNA sequences remain available in the current version of Pandit (see Supplementary Material online for full details). Results for 4 typical protein families are shown in table 4. Of all 15 protein families studied, PF01056 is the only 1 for which SDT is preferred to the ECMs according to the AIC. For all other protein families the ECMs perform better, as illustrated for PF01226, PF01229, and PF01231 in table 4.

In all 15 comparisons, the SDT model is always better than M0 (illustrated in table 4), suggesting that SDT, with its inclusions of single, double, and triple nucleotide substitutions, was a good attempt at modeling a real effect (see also Whelan and Goldman 2004). However, the general superiority of all variants of the empirical codon in this study suggests that these have successfully captured more information on typical patterns of codon substitutions.

LRT comparisons between F61 (table 2) and F1×4MG (table 4) variants of the ECM for protein families PF01226, PF01229, and PF01231 show that the F61 variants perform significantly better. The overall picture among the κ-models remains inconclusive (table 4; see also Conclusions).

Comparison of Estimates of Nonsynonymous–Synonymous Bias

For the ECM estimated from Pandit, we find $\omega_E = 0.192$ (eq. 8). For applications of ECM to other data sets, this value will vary, obviously greatly affected by estimates of ω and also depending (less strongly) on family-specific estimates of π_j and any κ-parameters. We have calculated ω_E values from ECM+F+ω+2κ for all alignments in our test set of 200 proteins, and we compare them with corresponding estimates of ω_M from M0 in figure 4. The ω_M and ω_E values are largely similar as the inset plot of figure 4 shows.

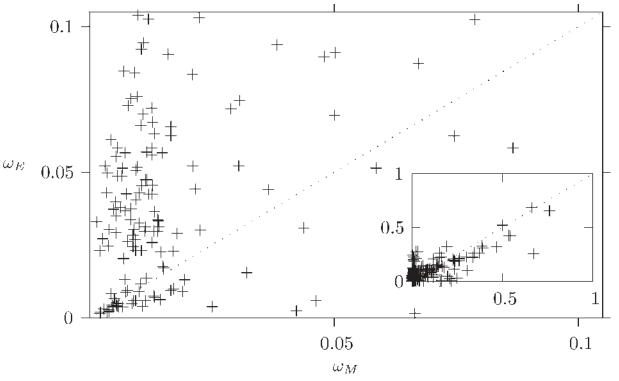

FIG. 4.—Nonsynonymous–synonymous rate ratios for 200 protein families estimated using ECM+F+ω+2κ(ω_E) and the mechanistic model M0 (ω_M). The dotted line indicates $\omega_E = \omega_M$. Note that the inset plot shows all nonsynonymous–synonymous rate ratios estimated, whereas the larger plot is an expanded version of the region $0 \leq \omega_E, \omega_M \leq 0.1$.

However, there is some interesting variation and, in particular, we note that the cases with strongest purifying selection (e.g., $\omega_M < 0.1$) are often assessed as less extreme under the empirical model ($\omega_E > \omega_M$). Conversely, proteins experiencing weaker purifying selection are generally assessed as having more constraints under ECM ($\omega_E < \omega_M$). Under strong purifying selection most observed changes will be synonymous. In ECM, however, there is not only a probability that synonymous change occurs via single synonymous substitutions, but also a nonzero probability via nonsynonymous double and triple nucleotide changes, thus, decreasing the inferred strength of purifying selection. For genes under weaker purifying selection, more nonsynonymous changes are observed; ECM allows for a nonzero probability that these nonsynonymous changes happened via multiple nucleotide substitutions to synonymous intermediates, resulting in the estimation of lower ω_E values. The changeover value for these competing effects lies at approximately $\omega_E = \omega_M = 0.15$ for our test data set of 200 protein families. In the future, it will require further investigation into what the 2 parameters ω_E and ω_M are measuring and which is most useful.

Conclusions

We have estimated an ECM, from alignments in the Pandit database, using a ML method embodied in the DART software. Analyzing the substitution patterns represented by ECM allows us to draw conclusions about the biological pressures and processes acting during codon sequence evolution. Existing codon models generally only allow for single nucleotide changes. However, our results indicate that modeling can be significantly improved by allowing for single, double, and triple nucleotide changes. Groupings of the 61 sense codons into subsets with high probability of change among codons of each group but small probability of change between groups shows that the affiliation between a triplet of DNA and the amino acid it encodes is a main factor driving the process of codon evolution. Relationships between different amino acids based on their physicochemical properties also have a strong influence.

The observations of multiple nucleotide change and the strong influence of physicochemical properties are not reflected in existing mechanistic models such as the widely used "M-series" of standard codon models (Yang et al. 2000). The importance of the genetic code may also be underestimated in existing models. In M0 (eq. 1) and M7, for example, it is only incorporated through the placement of the parameter ω_M and is entirely confounded with the strength of selection. In future, it may be important to give further consideration to how we should weight the evidence for natural selection given by multiple nucleotide replacements, nonsynonymous replacements between biochemically similar amino acids, and nonsynonymous replacements between biochemically different amino acids. Our analysis of estimates of parameters representing the strength of purifying selection derived from existing models and from our ECM suggests a complex relationship that requires further investigation before we fully understand what effects our new model may have on methods for detecting positively selected proteins and proteins sites.

The existence of simultaneous multiple nucleotide changes is controversial: Averof et al. (2000) find evidence for simultaneous multiple changes in residues coding for serines, and results from the use of the SDT model (Whelan and Goldman 2004) imply that multiple nucleotides changes occur. However, Bazykin et al. (2004) argue for successive single compensatory changes instead.

Some of our findings suggest that on the mutation level only single nucleotide changes occur. In particular, the relatively common occurrence of double changes in the 1st and 3rd positions of a codon (e.g., CGT (R) ↔ AGA (R); GTG (V) ↔ ATC (I); TTG (L) ↔ CTA (L); TTA (L) ↔ CTT (L)—see fig. 1) suggests a process of compensatory change: we do not know of any biological mechanism affecting noncontiguous nucleotides, and the relatively lower frequency of triple nucleotide substitutions means that an explanation by triple mutations that by chance have matching 2nd positions is highly unlikely.

A highly significant component of our findings is, however, that codon-level sequence evolution is better modeled when we include simultaneous multiple nucleotide substitutions. How, then, can we reconcile these 2 aspects of our findings? Arguing on the population level, realistic rates of mutation per generation (e.g., Neuhauser 2003) mean that the probability of multiple independent mutations in 1 individual is far too low to explain the proportions of double and triple changes observed in our ECM. Likewise, recombination events (Nordborg 2003) are not a plausible explanation for the observed effect: the probability of an individual having a mutation at 1 site, another individual a mutation at a neighboring site, and those 2 mating and the crossover placing the 2 mutations onto 1 genome is too low, particularly because the crossovers would require a break exactly between the 2 neighboring sites.

Positive selection favoring the compensation for a deleterious mutation by a mutation at another, epistatically interacting, site in the genome, seems to be the most likely mechanism to explain the multiple changes observed. Such a process will be dependent on often unknown population genetic factors such as population size, allowing for various scenarios. Multiple nucleotide changes could be the result of neutral mutations spreading in a population by genetic drift (Neuhauser 2003) and then an advantageous mutation occurring which is positively selected for. In large populations, mildly deleterious mutations can also be sustained in a subpopulation (Excoffier 2003); if a compensatory mutation then occurs, it will be positively selected and may spread through the whole population and be fixed. On the other hand, small populations are more susceptible to even deleterious mutations becoming fixed in the population (Neuhauser 2003). These mutations may then be followed by compensatory mutations that become fixed too: this mechanism could give a plausible mechanism for serine switches (AGY (S) ↔ TGY (C) or ACY (T) ↔ TCY (S)), where the substitution to the intermediate amino acid is believed to be very deleterious in general (Averof et al. 2000).

In summary, ECM suggests the existence of double and triple nucleotide changes, but the study of the patterns suggests that only single changes occur instantaneously. The explanation of this apparent discrepancy is that the multiple changes are in fact successive single changes occurring on a much faster timescale. This is expected from our explanation as positive selection will act to fix compensatory mutations at a much higher rate than neutral or mildly deleterious mutations. The phylogenetic application of ECM is successful because phylogenetic data represent evolution over long timescales and cannot discriminate the short timescales over which compensatory changes occur.

Similar arguments have been used to explain pairs of changes in sequences encoding functional RNA structures. Here, mutations that change a single base in a stem region of an RNA molecule are rare because there is strong selection to maintain complementary base pairing. Replacement of paired bases by different complementary pairs does, however, occur regularly in stem regions. This process has also been successfully modeled as an instantaneous change of multiple nucleotides (Higgs 1998; Savill et al. 2001). However, this topic requires further study, for example, by combining comparative analysis with large-scale polymorphism data (e.g., HapMap (The International HapMap Consortium 2003) and the Trace Archive (2006)).

We also tested ECM for utility in phylogenetic analyses. Past experience suggested that it would be beneficial to consider combining some mechanistic parameters with the pure ECM, and our choice of parameters was oriented toward those used in existing mechanistic codon models used for the detection of selection: codon frequencies, transition–transversion bias, and nonsynonymous–synonymous bias were used and combined models successfully implemented in PAML. Various parameterizations of the transition–transversion $\kappa(i, j)$ (eq. 6) were investigated, inspired by new scenarios which arise because instantaneous single, double, and triple nucleotide changes are permitted in the ECM. Compared with the simplest model, the more complex transition–transversion bias models can further improve likelihoods significantly in many, but clearly not all, cases. We recommend consideration of four κ-models

(ECM+F+ω, ECM+F+ω+1κ(ts), ECM+F+ω+1κ(tv), and ECM+F+ω+2κ) with choice among them determined using LRTs on a per-data set basis.

Overwhelmingly, the empirical models outperform the mechanistic models M0 and M7 and these results argue very strongly in favor of reconsidering codon models which do not treat all nonsynonymous changes equally (Massingham 2002). However, the original Goldman and Yang model which incorporated amino acid properties based on the Grantham matrix is known to perform worse than M0 (Yang et al. 1998). We therefore focus further comparisons to mechanistic models allowing for multiple nucleotide changes, and we show that ECM outperforms the SDT model in most cases. This proves that our ECM is suitable for use in phylogenetic analysis. Because codon models are becoming an option in phylogenetic reconstruction, despite their computational burden (Ren et al. 2005), we hope that our ECMs will be used for this purpose.

The mechanistic models M0 and M7 form the basis of current methods for detecting the footprints of positive selection acting on protein evolution (Yang et al. 2000). Great advances in the power to detect selection have been achieved by adapting M0-type models to allow for heterogeneity of nonsynonymous–synonymous biases among protein sites: for example, M7 uses a β-distribution of ω values and M8 adds the possibility of codons evolving with $\omega > 1$. It is remarkable that our ECM, which assumes a homogeneous pattern of evolutionary change at all sites, consistently outperforms M7 in our test set of 200 alignments. We have indicated how our per-data set estimates of the parameter ω can be used to compute a measure that is, in effect, the protein-wide average synonymous–nonsynonymous bias. This gives values comparable to those obtained using the mechanistic M0 model. In the future, we plan to adapt our ECM to incorporate site-specific synonymous–nonsynonymous biases and investigate to the consequences for studies aimed at determining the existence and location of selective effects.

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Simon Whelan for many helpful discussions of codon models and access to SDT model results, Ziheng Yang and Tim Massingham for assistance with modifications to PAML code, and Robert Bradley for further developing our bubble plot code and help with figure 1. C.K. was supported in part by a Wellcome Trust Prize Studentship and was a member of Wolfson College, University of Cambridge. N.G. was supported in part by a Wellcome Trust Senior Fellowship in Basic Biomedical Research.

Literature Cited

Abdo Z, Minin V, Joyce P, Sullivan J. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol Biol Evol. 22:691–703.

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 42:459–468.

Adachi J, Waddell P, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol. 50:348–358.

Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol. 22:200–209.

Averof M, Rokas A, Wolfe K, Sharp P. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science. 287:1283–1286.

Bateman A, Coin L, Durbin R, et al. (13 co-authors). 2004. The Pfam protein families database. Nucleic Acids Res. 32:D138–D141.

Bazykin G, Kondrashov F, Ogurtsov A, Sunyaev S, Kondrashov A. 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature. 429:558–562.

Brown W, Prager E, Wang A, Wilson A. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol. 18:225–239.

Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M. 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, Eck R. 1968. A model of evolutionary change in proteins. In: Dayhoff M, Eck R, editors. Atlas of protein sequence and structure 1967–68. Washington (DC): National Biomedical Research Foundation. p. 33–41.

Dayhoff M, Eck R, Park C. 1972. A model of evolutionary change in proteins. In: Dayhoff M, editor. Atlas of protein sequence and structure. Vol. 5. Washington (DC): Biomedical Research Foundation. p. 89–99.

Dayhoff M, Schwarz R, Orcutt B. 1978. A model of evolutionary change in proteins. In: Dayhoff M, editor. Atlas of protein sequence and structure. Vol. 5(suppl 3). Washington (DC): National Biomedical Research Foundation. p. 345–352.

Dimmic M, Rest J, Mindell D, Goldstein R. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol. 55:65–73.

Excoffier L. 2003. Analysis of population subdivision. In: Balding D, Bishop M, Cannings C, editors. Handbook of statistical genetics. 2nd ed. Vol. 2. Chichester (UK): Wiley. p. 713–745.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Goldman N. 1993. Statistical tests of models of DNA substitution. J Mol Evol. 36:182–198.

Goldman N, Thorne J, Jones D. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

Goldman N, Whelan S. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol. 17:975–978.

Goldman N, Whelan S. 2002. A novel use of equilibrium frequencies in models of sequence evolution. Mol Biol Evol. 19:1821–1831.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862–864.

Higgs P. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica. 102–103:91–101.

Higgs P, Hao W, Golding B. 2007. Identification of conflicting selective effects on highly expressed genes. Evol Bioinform. 2:1–13.

Holmes I, Rubin G. 2002. An expectation maximization algorithm for training hidden substitution models. J Mol Biol. 317: 753–764.

Jones D, Taylor W, Thornton J. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Jones D, Taylor W, Thornton J. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339:269–275.

Klosterman P, Uzilov A, Bendana Y, Bradley R, Chao S, Kosiol C, Goldman N, Holmes I. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics. 7:428.

Koshi J, Mindell D, Goldstein R. 1997. Beyond mutation matrices: physical-chemistry based evolutionary models. Genome Inform. 8:80–89.

Kosiol C. 2006. Markov Models for Protein Sequence Evolution. [Ph.D. thesis]. EMBL-European Bioinformatics Institute. Cambridge: University of Cambridge.

Kosiol C, Goldman N. 2005. Different versions of the Dayhoff rate matrix. Mol Biol Evol. 22:193–199.

Kosiol C, Goldman N, Buttimore N. 2004. A new criterion and method for amino acid classification. J Theor Biol. 228:97–106.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8:1233–1244.

Massingham T. 2002. Detecting positive selection in proteins: models of evolution and statistical tests. [Ph.D. thesis]. Cambridge: University of Cambridge.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics. 169:1753–1762.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculation for fast computing machines. J Chem Phys. 21:1087–1092.

Muse S, Gaut B. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Neuhauser C. 2003. Mathematical models in population genetics. In: Balding D, Bishop M, Cannings C, editors. Handbook of statistical genetics. 2nd ed. Vol. 2. Chichester (UK): Wiley. p. 577–599.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Nordborg M. 2003. Coalescent theory. In: Balding D, Bishop M, Cannings C, editors. Handbook of statistical genetics. 2nd ed. Vol. 2. Chichester (UK): Wiley. p. 602–631.

Ren F, Tanaka H, Yang Z. 2005. An empirical examination of the utility of codon substitution models in phylogenetic reconstruction. Syst Biol. 54:808–818.

Savill N, Hoyle D, Higgs P. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics. 157:399–411.

Schneider A, Cannarozzi G, Gonnet G. 2005. Empirical codon substitution matrix. BMC Bioinformatics. 6:134.

Silvey S. 1970. Statistical inference. London: Chapman and Hall.

Smith N, Webster M, Ellegren H. 2003. A low rate of simultaneous double-nucleotide mutations in primates. Mol Biol Evol. 20:47–53.

Sullivan J, Abdo Z, Joyce P, Swofford D. 2005. Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. Mol Biol Evol. 22:1386–1392.

Sullivan J, Holsinger K, Simon C. 1996. The effect of topology on estimates of among-site rate variation. J Mol Evol. 42:308–312.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura R, editor. Lectures on mathematics in the life sciences. Providence (RI): American Mathematical Society. p. 57–86.

The International HapMap Consortium. 2003. The international HapMap project. Nature. 426:789–796.

Trace Archive V4.1::NCBI/NLM/NIH. 2007. Available from: http://www.ncbi.nlm.nih.gov/Traces. Accessed 17 May 2007.

Urbina D, Tang B, Higgs P. 2006. The response of amino acid frequencies to directional mutational pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. J Mol Evol. 62:340–361.

Whelan S, de Bakker P, Goldman N. 2003. Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics. 19:1556–1563.

Whelan S, de Bakker P, Quevillon E, Rodriguez N, Goldman N. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res. 34:D327–D331.

Whelan S, Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol. 16:1292–1299.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics. 167: 2027–2043.

Wong W, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J Mol Evol. 39:105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–314.

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

Yang Z, Bielawski J. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Yang Z, Nielsen R, Goldman N, Pedersen A-M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

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

Arndt von Haeseler, Associate Editor

Accepted March 19, 2007