Syst. Biol. 54(5):808–818, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500354688

An Empirical Examination of the Utility of Codon-Substitution Models in Phylogeny Reconstruction

FENGRONG REN,[1] HIROSHI TANAKA,[1] AND ZIHENG YANG[2]

[1]*Advanced Biomedical Information, Center for Information Medicine, Tokyo Medical and Dental University, Japan*
[2]*Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk*

Abstract.—Models of codon substitution have been commonly used to compare protein-coding DNA sequences and are particularly effective in detecting signals of natural selection acting on the protein. Their utility in reconstructing molecular phylogenies and in dating species divergences has not been explored. Codon models naturally accommodate synonymous and nonsynonymous substitutions, which occur at very different rates and may be informative for recent and ancient divergences, respectively. Thus codon models may be expected to make an efficient use of phylogenetic information in protein-coding DNA sequences. Here we applied codon models to 106 protein-coding genes from eight yeast species to reconstruct phylogenies using the maximum likelihood method, in comparison with nucleotide- and amino acid–based analyses. The results appeared to confirm that expectation. Nucleotide-based analysis, under simplistic substitution models, were efficient in recovering recent divergences whereas amino acid–based analysis performed better at recovering deep divergences. Codon models appeared to combine the advantages of amino acid and nucleotide data and had good performance at recovering both recent and deep divergences. Estimation of relative species divergence times using amino acid and codon models suggested that translation of gene sequences into proteins led to information loss of from 30% for deep nodes to 66% for recent nodes. Although computational burden makes codon models unfeasible for tree search in large data sets, we suggest that they may be useful for comparing candidate trees. Nucleotide models that accommodate the differences in evolutionary dynamics at the three codon positions also performed well, at much less computational cost. We discuss the relationship between a model's fit to data and its utility in phylogeny reconstruction and caution against use of overly complex substitution models. [Codon models; divergence dates; maximum likelihood; phylogenetics; phylogenetic information.]

Models of codon substitution (Goldman and Yang, 1994; Muse and Gaut, 1994) consider a codon triplet as the unit of evolution and can distinguish between synonymous (silent) and nonsynonymous (replacement) substitutions. They are widely used in analysis of protein-coding DNA sequences to detect natural selection acting on the protein, with the nonsynonymous/synonymous substitution rate ratio (ω) used as an indicator of selective pressure. The basic models have been extended in a number of ways, for example, to account for variation in selective pressure among branches on the phylogeny (Seo et al., 2004; Yang, 1998), among sites in the protein (Huelsenbeck and Dyer, 2004; Nielsen and Yang, 1998; Yang et al., 2000), or among both branches and sites (Bielawski and Yang, 2004; Forsberg and Christiansen, 2003; Guindon et al., 2004). The extended models are particularly effective in identifying genes or amino acid residues that are affected by positive Darwinian selection. However, it is unclear how useful codon models are for phylogenetic tree reconstruction. Codon models consider the genetic code and naturally accommodate synonymous and nonsynonymous substitutions. In almost all protein-coding genes, synonymous substitutions occur at high rates and are informative about recent divergences, whereas nonsynonymous substitutions occur at low rates and may be useful for resolving early divergences. By accounting for both types of substitutions, codon models may be expected to make an efficient use of information in the data, leading to high accuracies in phylogeny reconstruction. Even though nucleotide models can be formulated to accommodate differences in the evolutionary dynamics at the three codon positions (e.g., Yang, 1996), they can at best do an awkward job of describing the substitution process in protein-coding genes.

In this paper we apply codon models to 106 protein-coding genes from eight yeast species (Rokas et al., 2003) to reconstruct maximum likelihood (ML) phylogenies. Rokas et al. (2003) used the ML method to analyze the nucleotide sequences and the maximum parsimony method to analyze both the nucleotide and amino acid sequences. They emphasized the considerable phylogenetic conflicts among genes in separate analysis and the universal congruence among methods in analysis of the concatenated sequences. The same data have been reanalyzed in several recent studies (Holland et al., 2004; Phillips et al., 2004; Taylor and Piel, 2004). Taylor and Piel (2004) pointed out that Rokas et al.'s use of 70% bootstrap proportion as a cutoff for strong clade support led to an exaggerated assessment of conflicts among gene trees. Phillips et al. (2004) found that use of the minimum-evolution criterion without accommodating rate variation among sites produced a different tree for the concatenated data, suggesting that the congruence among methods was not universal even for the large concatenated data set. They suggested that differences in base compositions among species may have misled the minimum evolution criterion and cautioned that examination of model assumptions and exploration of systematic errors is important even in genome-scale data sets. Here we use the data of Rokas et al. (2003) to compare analyses at three different levels, i.e., amino acid–based, nucleotide-based, and codon-based analyses. All the 106 genes are protein-coding, and can be analyzed using either amino acid-, nucleotide-, or codon-substitution models. It is interesting to gain insights into the advantages and limitations of the different analyses of the same data. The only such study comparing analyses at all three levels we are aware of is that of Chang et al. (2002), who reconstructed ancestral protein sequences.

Those authors found that analyses using amino acid, nucleotide, and codon models produced very similar reconstructions. Here our main focus is reconstruction and test of the phylogeny. In addition, we examine the utility of the different models in estimating species divergence times under the assumption of a molecular clock. We are also interested in the relationship between the goodness of fit of a model and its utility for phylogenetic analysis.

Materials and Methods

Sequence Data

The sequence data are 106 protein-coding genes from seven *Saccharomyces* species and one outgroup yeast *Candida albicans*, published by Rokas et al. (2003). See Rokas et al. (2003) for details of the data and GenBank accession numbers. We either concatenated all genes into one "super-gene," ignoring possible differences among genes, or analyzed the 106 genes separately. Amino acid–based analysis made use of the protein sequences translated from the encoding DNA sequences. The sequence alignments are available at the *Systematic Biology* Website (www.systematicbiology.org).

Substitution Models Assumed in Likelihood Analyses

A number of continuous-time Markov-process models have been developed to model substitutions between nucleotides, amino acids, and codons (Lio and Goldman, 1998). We had two considerations in our choice of models for analysis. First we wanted to use models that are realistic for the data. Second, we wanted the models at the nucleotide, amino acid, and codon levels to be similar so that the results obtained at the three levels of analysis indicate differences in the information content rather than inadequacies of models. Table 1 lists the models that are used in this paper to analyze the yeast data sets.

In nucleotide-based analysis, we used several variations of the HKY model (Hasegawa et al., 1985), which accounts for different transition and transversion rates and unequal nucleotide frequencies. The basic HKY model involves four free parameters: the transition/transversion rate ratio κ and three equilibrium base frequencies. This model captures the major features of the substitution pattern but is highly unrealistic as it ignores differences among the three codon positions.

TABLE 1. Substitution models used in this study and the number of parameters.

For amino acids	For nucleotides	For codons
WAG+F (19)	HKY (4)	F3×4 (11)
WAG+F+G (20)	HKY+G (5)	F3×4MG (11)
FromCodon (21)	HKY+C (14)	F3×4MG+M3 (16)
FromCodon+G (22)	HKY+C+G (17)	
	GTR (8)	
	GTR+G (9)	
	GTR+C (26)	
	GTR+C+G (29)	

Note.—The number of parameters (in parentheses) includes only those in the substitution model and does not include branch lengths. The three most similar models at the three levels are underlined.

TABLE 2. Base frequencies, transition/transversion rate ratios, and tree lengths at the three codon positions.

Position	π_T	π_C	π_A	π_G	κ	Tree length
1	0.24	0.15	0.31	0.30	2.18	0.889
2	0.31	0.21	0.34	0.14	1.54	0.454
3	0.34	0.20	0.28	0.19	7.95	4.584
All	0.30	0.18	0.31	0.21	3.53	1.369

Note.—There are 42,342 nucleotide sites at each codon position in the concatenated data. Tree length is the sum of branch lengths, the number of substitutions per site throughout the tree. Parameter κ and branch lengths are estimated under the HKY model on tree T_1 of Figure 1 for the three codon positions separately.

The HKY+G model combines HKY with the discrete-gamma model of variable rates among sites, with five rate categories (Yang, 1994b). The HKY+C model accounts for different transition/transversion rate ratios, different base compositions, and different substitution rates at the three codon positions, assuming proportional branch lengths among codon positions (Yang, 1996). This model involves 14 parameters: two relative rates for codon positions, three κs, and nine base frequency parameters. Basic statistics for the three codon positions are listed in Table 2, which show huge differences in the evolutionary dynamics among the codon positions. In the HKY+C+G model, the "C" component accommodates large-scale differences among codon positions, whereas a gamma distribution is used for each codon position to account for any remaining rate variation within each codon position. It has 17 parameters. We also replaced the basic substitution model HKY with the general time reversible model (GTR or REV) (Tavare, 1986; Yang, 1994a), leading to models referred to as GTR, GTR+G, GTR+C, and GTR+C+G (Table 1). Each κ parameter in HKY is replaced by five rate parameters in GTR. From previous studies (e.g., Sullivan and Swofford, 2001; Yang, 1994a), we expect the HKY and GTR models to have similar performance. We did not use models of invariable sites plus gamma (the "I+G" models), as they appear somewhat pathological due to the strong correlation between the proportion of invariable sites and the gamma shape parameter (Sullivan and Swofford, 2001; Yang, 1993).

In the codon-based analysis, the model incorporates the transition/transversion rate ratio (κ) and the nonsynonymous/synonymous rate ratio (ω), and accounts for different nucleotide compositions at the three codon positions. To make the codon model as similar to nucleotide models HKY and HKY+C as possible, we use the following specification. Suppose codon i (triplet $i_1i_2i_3$) and codon j (triplet $j_1j_2j_3$) have one difference at position k ($k = 1, 2, 3$). The relative substitution rate from codons i to j is

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at 2 or 3 position,} \\ \pi_{j_k}^{(k)}, & \text{for synonymous transversion,} \\ \kappa\pi_{j_k}^{(k)}, & \text{for synonymous transition,} \\ \omega\pi_{j_k}^{(k)}, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_{j_k}^{(k)}, & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

where $\pi_{j_k}^{(k)}$ is the frequency of the target nucleotide j_k at position k. The Markov process of codon substitution is time-reversible with equilibrium frequency π_j for codon j proportional to $\pi_{j_1}^{(1)}\pi_{j_2}^{(2)}\pi_{j_3}^{(3)}$, because the rate can be written in the form $q_{ij} = s_{ij} \times \pi_j$, with $s_{ij} = s_{ji}$, for all $i \neq j$. For example, $q_{\text{TCA}\to\text{TCG}} = \kappa\pi_G^{(3)} = (\frac{\kappa}{\pi_T^{(1)}\pi_C^{(2)}}) \times \pi_T^{(1)}\pi_C^{(2)}\pi_G^{(3)}$ and $q_{\text{TCG}\to\text{TCA}} = \kappa\pi_A^{(3)} = (\frac{\kappa}{\pi_T^{(1)}\pi_C^{(2)}}) \times \pi_T^{(1)}\pi_C^{(2)}\pi_A^{(3)}$. Equation (1) is slightly different from the F3×4 model of Goldman and Yang (1994) in that both rates q_{ij} and q_{ji} are divided by the frequencies of the two unchanged nucleotides, whereas under F3 × 4, for example, $q_{\text{TCA}\to\text{TCG}} = \kappa\pi_T^{(1)}\pi_C^{(2)}\pi_G^{(3)}$ and $q_{\text{TCG}\to\text{TCA}} = \kappa\pi_T^{(1)}\pi_C^{(2)}\pi_A^{(3)}$. Equation (1) is similar to and more general than the model of Muse and Gaut (1994), which ignores the transition/transversion rate ratio and different base compositions among codon positions. Equation (1) is referred to as F3×4MG in the PAML documentation (Yang, 1997b) and is also available in the HyPhy package (Kosakovsky Pond et al., 2005). The F3×4 and F3×4MG models have exactly the same number of parameters and are not nested within each other. We used both models F3×4 and F3×4MG, but focused on F3×4MG as it is closer to the nucleotide model HKY+C. Note that F3×4MG and HKY+C would be nearly identical (a) if there were no stop codons, and (b) if there were no rate differences among codon positions (when HKY+C became HKY and $\omega = 1$ in F3×4MG), or (b′) if all codons were fourfold degenerate. Codon models incorporate the genetic code and the rate ratio ω, naturally leading to different substitution rates and different transition/transversion rate ratios at the three codon positions. The model involves 11 free parameters: κ, ω, and nine base frequency parameters. We also use an extension of F3×4MG, in which the ω ratio varies among sites according to a discrete distribution with three categories. This is model M3 in Yang et al. (2000) and involves five additional parameters.

For amino acid sequences, we used the empirical WAG matrix of relative amino acid substitution rates (Whelan and Goldman, 2001), with the equilibrium amino acid frequencies replaced by the frequencies observed in the data (Adachi and Hasegawa, 1996). The model is referred to as WAG+F. When the substitution rate is assumed to vary among sites according to the discrete gamma model (Yang, 1994b), the model is denoted WAG+F+G. Whelan and Goldman (2001) found that the WAG matrix fitted most data sets better than other empirical models such as Dayhoff (Dayhoff et al., 1978) and JTT (Jones et al., 1992). We also used a codon-based "mechanistic" model of amino acid substitution described by Yang et al. (1998). This uses a model of codon substitution to construct an amino acid substitution model. The codon model uses the amino acid chemical distances of Miyata et al. (1979) to modify nonsynonymous substitution rates, so that the acceptance rate ω between amino acids i and j is inversely related to their chemical distance d_{ij}:

$$\omega_{ij} = \exp\{-bd_{ij}/d_{\max}\}, \quad (2)$$

where $d_{\max}$ is the maximum value, and parameter b is estimated from the data. Note that parameter a in equation 11 in Yang et al. (1998) is removed as amino acid sequences do not allow estimation of synonymous rate (see also table 3 in that paper). A reversible amino acid substitution model is constructed by collapsing states in the Markov chain, that is, by aggregating synonymous codons into the same amino acid. See Yang et al. (1998) for details. This model involves 21 parameters: κ, b and 19 amino acid frequencies. We refer to this model as FromCodon. It was also combined with a gamma distribution of rates among sites, to form the FromCodon+G model.

FromCodon for amino acids, HKY+C for nucleotides, and F3×4MG for codons are expected to be the most similar models formulated at the three levels. Thus when we compare the amount of phylogenetic information in amino acid, nucleotide and codon sequences, we focused on those three models.

TABLE 3. The number of genes out of 106 supporting each node in tree T_1 of Figure 1.

	Node 1	Node 2	Node 3	Node 4	Node 5	Whole tree
Amino acid						
WAGF	71	50	45	105	89	25
WAGF+G	68	51	42	104	91	27
FromCodon	76	54	48	105	93	28
FromCodon+G	69	52	42	104	91	27
Base						
HKY	98	81	54	104	76	53
HKY+G	92	71	45	103	78	45
HKY+C	95	79	57	105	88	55
HKY+C+G	89	74	53	102	85	52
GTR	98	82	57	104	79	56
GTR+G	95	73	46	103	77	46
GTR+C	96	82	58	105	91	56
GTR+C+G	93	81	53	104	88	52
Codon						
F3×4	88	73	59	106	90	51
F3×4MG	86	78	59	104	89	54
F3×4MG+M3	93	74	54	105	90	52

Reconstruction and Tests of Phylogenetic Trees

Most models examined in this paper are not yet implemented in efficient tree-reconstruction programs such as PAUP (Swofford, 1999) or PHYLIP (Felsenstein, 2004). The PAML package (Yang, 1997b) was thus used, with the BASEML program for nucleotide-based analysis, and CODEML for amino acid– and codon-based analyses. We used a stepwise-addition algorithm implemented in PAML as well as a strategy of evaluating candidate trees. In the latter approach, each gene or protein was analyzed using a variety of fast tree reconstruction algorithms implemented in PHYLIP, including parsimony, neighbor joining, and ML under the simple Jukes and Cantor (1969) model. Those analyses produced 51 distinct tree topologies, which were compared under more

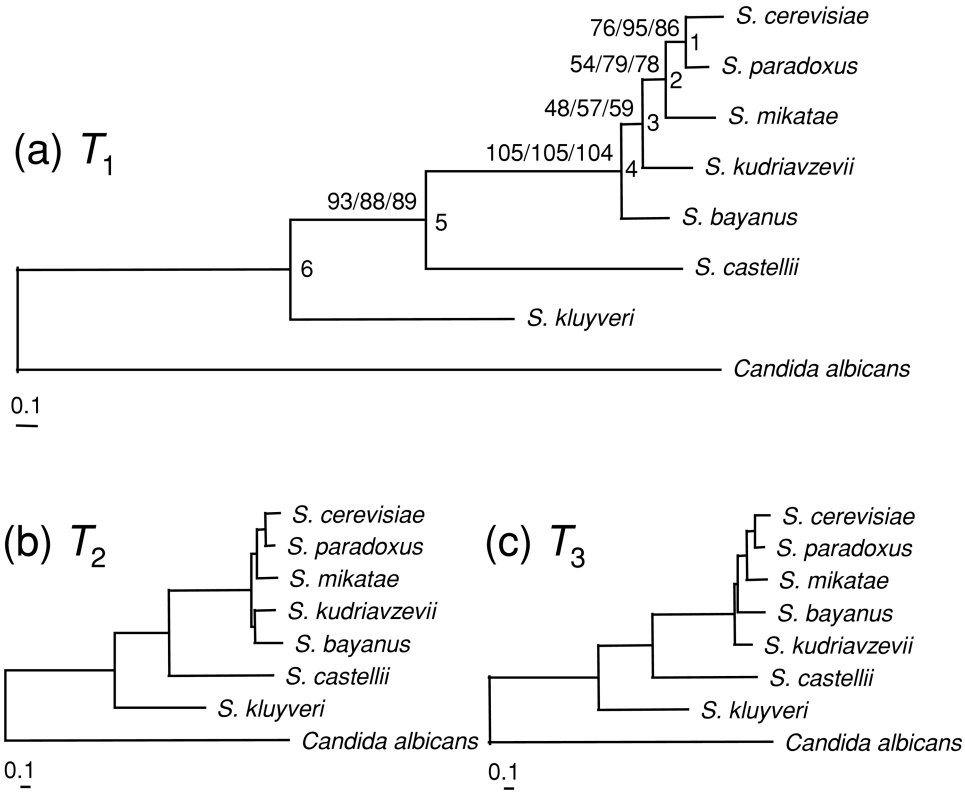

FIGURE 1. Three tree topologies (T_1, T_2, T_3) for seven *Saccharomyces* species and the outgroup yeast species *Candida albicans*. The trees are unrooted but the root is placed along the branch to the outgroup for clarity. (a) Tree T_1 is the ML tree in ML analyses of the concatenated data. (b) T_2 is the minimum evolution tree for the concatenated data inferred by Phillips, Delsuc, and Penny (2004). (c) T_3 is another tree inferred in some genes. The branch lengths, measured by the expected number of nucleotide substitutions per codon, are estimated under the codon model F3×4MG. The log-likelihood values under the model are −627,584.96, −627,951.27, and −628,006.97 for T_1, T_2, T_3, respectively. The MLEs of parameters under the codon model are $\hat{\kappa} = 2.403$ and $\hat{\omega} = 0.051$ for all three trees. The three numbers at each internal node of T_1 (a) are the numbers of genes or proteins (out of 106) that support the clade in the following three analyses: (i) analysis of the amino acid sequences under the FromCodon model; (ii) analysis of the nucleotide sequences under the HKY+C model; and (iii) analysis of the codon (nucleotide) sequences under the codon model F3×4MG. Results for other models and analyses are given in Table 3.

sophisticated models using PAML programs. Three of those trees are shown in Figure 1. For some models, both approaches were used and were found to produce identical results; that is, the ML tree from the heuristic tree search was always among the 51 candidate trees. Because the data sets are small with only eight species and most of the nodes are well resolved, tree search for those data sets were concluded quickly.

To assess uncertainties in the estimated phylogenies, the bootstrap method (Felsenstein, 1985) was used. Under nucleotide models with the "C" component, sites at the three codon positions do not follow the same distribution, and stratified sampling is used; that is, the same number of sites are always sampled at each position. When comparing the 51 candidate trees, we applied an approximate bootstrap method, which used maximum likelihood estimates (MLEs) of parameters from the original data set to evaluate the log likelihoods of the trees rather than re-estimating parameters from each bootstrap pseudo-sample. The method, known as the RELL bootstrap (for Resampling Estimated Log Likelihoods) (Kishino et al., 1990), appears to approximate the bootstrap method of Felsenstein (1985) very well (Hasegawa and Kishino, 1994). Although controversies exist concerning interpretation of the bootstrap (and its RELL approximation), we suggest that the measure is adequate for our purpose of evaluating relative support levels for the tree and the amount of phylogenetic information when the data are analyzed at the three different levels. For example, if the amino acid– and codon-based analyses produce the same ML tree but the codon-based analysis provides stronger clade support, the result may be interpreted as the codon sequences having more phylogenetic information.

RESULTS

Analyses of the Concatenated Data

The 106 genes (or proteins) were concatenated as one super-gene (or super-protein) and analyzed under the models of Table 1. The stepwise addition algorithm produced tree T_1 of Figure 1 as the ML tree in every analysis/model. The bootstrap method (Felsenstein, 1985) was used to calculate support values for clades, with 100

bootstrap replicates for the nucleotide- and amino acid–based analyses, and 50 replicates for the codon-based analyses. Every clade in tree T_1 received 100% bootstrap support. Those results are similar to the findings of Rokas et al. (2003).

The approach of evaluating candidate trees was also applied to the concatenated data sets. As mentioned in Materials Methods, application of fast tree reconstruction methods to analyze the 106 genes or proteins produced 51 candidate trees, which are assumed to include all reasonable trees that are likely to be correct. Comparison of those candidate trees using the concatenated data sets under the models of Table 1 produced tree T_1 as the ML tree in every analysis, with 100% support for every node in tree T_1 according to the RELL bootstrap (Kishino et al., 1990). The results are consistent with those obtained using the heuristic tree search in combination with the proper bootstrap, discussed above.

Analyses of Separate Genes

Each gene or protein was analyzed using the models listed in Table 1 to compare the 51 candidate trees, and the RELL bootstrap method was used to calculate their support values. First we summarize the ML trees obtained in the separate analysis. As T_1 is most likely the correct tree for those genes (Phillips et al., 2004; Rokas et al., 2003), the number of genes or proteins in which T_1 is recovered as the ML tree may be used to indicate the performance of the model/analysis. The amino acid models estimated T_1 as the ML tree in 25 to 28 proteins (Table 3). In nucleotide-based analysis, the "G" models recovered T_1 as the ML tree in 45 to 46 genes, whereas all other nucleotide models performed better and recovered T_1 in 52 to 56 genes, similar to codon-based analysis, in which the number is 51 to 54. The poor performance of the amino acid models is due to loss of phylogenetic information when the DNA sequences are translated into proteins. This interpretation receives more support when we consider the numbers of genes that support individual clades (defined by nodes 1 to 5) in tree T_1 of Figure 1, shown in Table 3. Clade 4 is defined by a long branch and is recovered in almost all analyses and models. Nodes 1, 2, and 3 represent recent divergences. The amino acid models were far poorer at recovering these nodes than the nucleotide or codon models. For example, node 1 was recovered in only 68 to 76 proteins (out of 106) in amino acid–based analyses, whereas it is recovered in 89 to 98 or 86 to 93 genes in nucleotide- or codon-based analyses. HKY and GTR were slightly more efficient than the more complex "C" or "G" models at recovering this node, consistent with previous simulation studies in which simplistic models were found to perform better than more complex and realistic models in recovering recent divergences (e.g., Tateno et al., 1994). In contrast, node 5 represents a deep divergence. It was recovered in 89 to 93 proteins by amino acid models and in 89 to 90 genes by codon models. In nucleotide-based analysis, the simplistic models HKY and GTR recovered this node in only 76 and 79 genes, much worse than the amino acid and codon models. However, the "C" models recovered node 5 in 88 and 91 genes and performed as well as the amino acid and codon models. The gamma models ("G") were much poorer than the "C" models in recovering this node. In sum, the results suggest that the poor performance of amino acid models in recovering the whole tree T_1, discussed above, was due to their poor performance in recovering recent nodes. The results support the intuitive expectation that amino acid or nonsynonymous substitutions are informative concerning deep divergences whereas nucleotide or silent substitutions are informative for recent divergences. The codon models performed well and appeared to make a good use of both kinds of information. The "C" nucleotide models also performed well, at much less computational cost.

Next we counted the number of genes or proteins (out of 106) in which two analyses produced the same ML tree (Table 4). The larger this number is, the more similar the two analyses are. In this analysis only HKY is used because GTR performed similarly and because HKY is expected to be closer to the codon models. If we consider amino acid–, nucleotide-, and codon-based models as three different types of models, the most conspicuous pattern in Table 4 is that models of the same type are much more similar to each other than they are to models of different types. The only exception to this pattern is the

TABLE 4. The number of genes out of 106 in which two analyses produced the same best tree in comparison of the 51 candidate trees.

	WAGF	WAGF+G	FromCodon	FromCodon+G	HKY	HKY+C	HKY+C+G	F3×4	F3×4MG
Amino acid									
WAGF									
WAGF+G	59 (19)								
FromCodon	67 (22)	69 (24)							
FromCodon+G	61 (19)	103 (27)	69 (24)						
Base									
HKY	18 (14)	17 (15)	20 (17)	18 (15)					
HKY+C	29 (19)	30 (22)	36 (23)	30 (22)	43 (36)				
HKY+C+G	27 (19)	32 (22)	32 (23)	32 (22)	43 (34)	83 (47)			
Codon									
F3×4	37 (20)	40 (22)	49 (25)	40 (22)	38 (34)	63 (40)	56 (39)		
F3×4MG	36 (21)	40 (22)	47 (25)	41 (22)	42 (36)	64 (42)	59 (41)	86 (48)	
F3×4MG+M3	33 (20)	35 (20)	36 (22)	36 (20)	37 (31)	59 (41)	51 (36)	66 (41)	72 (45)

Note.—The number in parentheses is the number of genes in which T_1 is the best tree for both methods/analyses.

nucleotide model HKY, which appears to be the least realistic model considered here and which was most different from all other models (Table 4). Models FromCodon for amino acids, HKY+C for nucleotides and F3×4MG for codons, are expected to be the most similar at the three levels, and they did show the greatest similarity between model types. For example, HKY+C was much more similar to the codon or amino acid models than HKY was. Overall, the differences between the analyses appeared rather large.

We calculated the RELL bootstrap support values for the 51 candidate trees under each model of Table 1. In Figure 2, the bootstrap proportions (P_1) for T_1 were compared across several analyses. We used P_1 as a measure

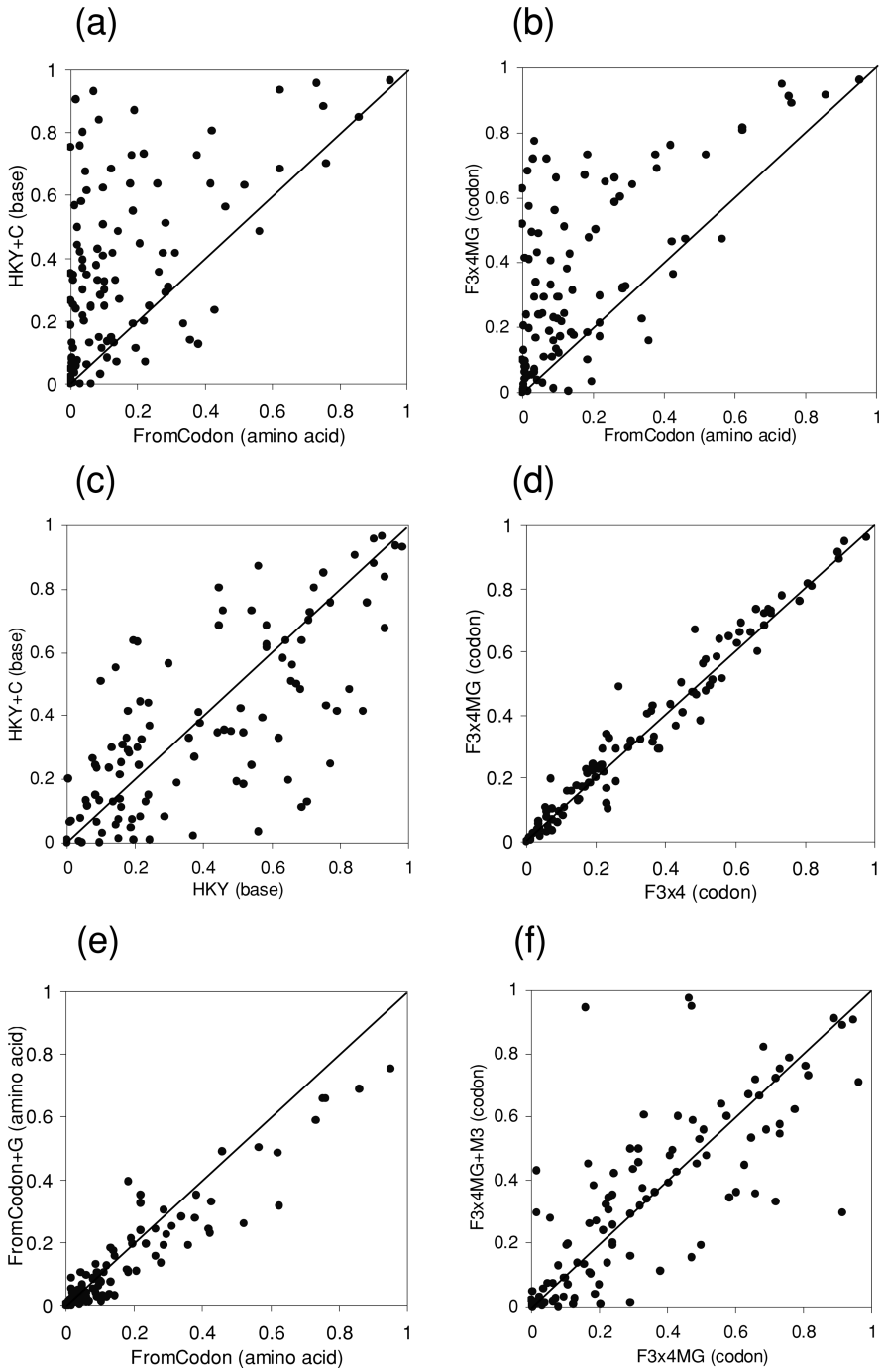

FIGURE 2. Comparison of RELL bootstrap support values for tree T_1 calculated from different models/analyses in separate analyses of the 106 genes or proteins. The RELL method (Kishino et al., 1990) was used to compare 51 candidate trees, and the approximate bootstrap proportion P_1 for T_1 of Figure 1 is used for the plots. See Table 1 and the text for definitions of the models.

of information content, with higher P_1 indicating greater amount of information. The following observations can be made. First, P_1 calculated from nucleotide or codon models tend to be higher than under amino acid models, indicating that the nucleotide (or codon) sequences contain more phylogenetic information than the amino acid sequences. For example, the nucleotide model HKY+C and the codon model F3×4MG produced higher P_1 values than the amino acid model FromCodon in 88 and 91 genes, respectively (Fig. 2a and b). Second, in most genes, P_1 was not very extreme (for example, not greater than 95%) in most analyses. Thus the substantial differences in the ML trees estimated in the separate analyses, as described above and by Rokas et al. (2003), reflect more a lack of information to resolve the phylogeny with confidence than genuine conflicts among gene trees. As pointed out by Taylor and Piel (2004), Rokas et al. (2003) overstated phylogenetic conflict among genes as they considered bootstrap support values of >70% as indicating significant clades. Third, the correlation in P_1 between analyses or models was not very strong. For example, even though applied to the same data, the two nucleotide models HKY and HKY+C showed considerable differences (Fig. 2c). The two codon models F3×4 and F3×4MG are nevertheless very similar (Fig. 2d). Fourth, incorporation of the gamma model for rates at sites did not seem to improve the accuracy of phylogeny reconstruction for those data. For example, P_1 in most proteins decreased when the gamma model is added to the amino acid model FromCodon (Fig. 2e), even though it improved the model's fit enormously (see below). The poorer performance of the gamma models, relative to both the "C" models and the plain HKY and REV models, is somewhat surprising and may be particular to the data analyzed here. For other data sets not partitioned by codon positions, use of the gamma model may well improve phylogenetic accuracy, as demonstrated in numerous previous studies (e.g., Takezaki and Gojobori, 1999).

The Fit of Models to Data and Their Utility in Phylogeny Reconstruction

We use three likelihood-based criteria to evaluate the fit of models to data. The likelihood-ratio test (LRT) can be used to compare two nested models. The AIC (Akaike, 1974) and the BIC (Schwaz, 1978) criteria penalize parameter-rich models, applying the parsimony principle of model building. AIC considers one extra parameter as being worthy of 1 log-likelihood unit. BIC penalizes parameter-rich models more severely, and one parameter is considered to be worthy of $\frac{1}{2}\ln(n)$ log-likelihood units, or 3 to 4 units, as the sequence length ranges from $n = 390$ to 2994 sites among the 106 genes. We used tree T_1 for the likelihood calculation, although use of other reasonable trees led to the same conclusions. This is because model assumptions considered here had far greater impact on the log likelihood than the tree topologies (e.g., Yang et al., 1995).

For both the empirical amino acid substitution model WAG+F and the mechanistic model FromCodon, incorporation of the gamma model of rates among sites leads to significant improvement in the model's fit. The log-likelihood difference $\Delta\ell$ ranges from 14.9 to 237.4 among the proteins for WAG+F and from 12.8 to 266.3 for FromCodon. Substitution rates are clearly variable among amino acid sites. Such rate variation has been well documented and was expected. Also WAG+F was found to fit the data much better than FromCodon in every protein, with the log-likelihood difference $\Delta\ell$ ranging from 24.3 to 308.6. The two models are not nested so we cannot use the χ^2 approximation to the LRT. However, because FromCodon involves more parameters than WAG+F (Table 1), use of both AIC and BIC led to rejection of FromCodon in every protein.

TABLE 5. Range of log-likelihood differences across 106 genes between nucleotide models.

Model 1	Model 2	Range of ℓ_2–ℓ_1	Difference in P
HKY	HKY+G	70.2–586.8	1
HKY	HKY+C	161.9–1316.2	10
HKY+G	HKY+C	90.1–766.7	9
HKY+C	HKY+C+G	5.1–86.1	3
HKY	GTR	2.3–159.6	4
HKY+G	GTR+G	2.6–93.6	4
HKY+C	GTR+C	12.7–167.8	12
HKY+C+G	GTR+C+G	11.6–160.3	12
GTR	GTR+G	67.9–544.8	1
GTR	GTR+C	172.6–1342.4	18
GTR+G	GTR+C	99.7–812.1	17
GTR+C	GTR+C+G	6.3–167.0	3

The ranges (across the 106 genes) of log likelihood differences between the nucleotide models are listed in Table 5. By all three criteria, substitution rates are highly variable among sites and adding the "G" and "C" components improved the fit of both HKY and GTR models in every gene. Both AIC and BIC preferred the "C" models to the "G" models, reflecting the huge differences in the substitution patterns among the three codon positions (Table 2). The "C+G" models are the best according to all three criteria. The GTR models are preferred to the corresponding HKY models in most genes, but they produced very similar results in the analyses, as mentioned earlier.

The relative fit of the two codon models, F3×4 and F3×4MG, depends on the gene, with F3×4MG being better in 79 genes and F3×4 being better in 27 genes. Estimates of parameters κ and ω were almost identical between the two models. The performance of the two models in phylogeny reconstruction was also highly similar, as noted earlier. Thus, despite our initial concern that F3×4MG should be closer to the nucleotide model HKY+C than is F3×4, the differences between the two codon models appeared rather insignificant. The codon and nucleotide models are not nested, so we cannot use the χ^2 distribution for the LRT. Both codon models F3×4 and F3×4MG fitted the data better than the nucleotide model HKY+C in every gene. The log-likelihood difference ($\Delta\ell$) ranged from 51.1 to 614.6 among genes for F3×4 and from 40.3 to 594.8 for F3×4MG. As HKY+C involve more parameters (Table 1), both AIC and BIC favor

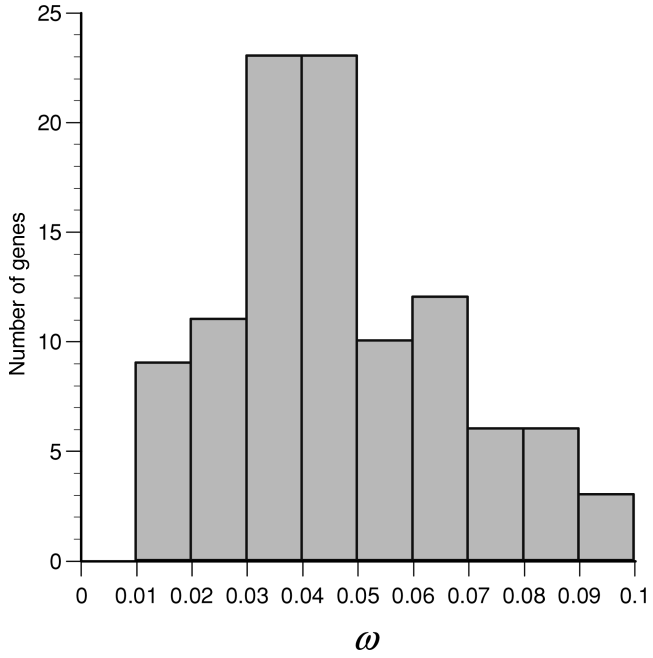

FIGURE 3. Histogram of MLEs of the ω ratio under the codon model F3×4MG across 106 genes. Tree T_1 of Figure 1 was used in the ML analysis.

the codon models. By incorporating the genetic code, the codon models fitted the data much better than the nucleotide model HKY+C, with three fewer parameters. The single nonsynonymous/synonymous rate ratio ω in the codon models was able to account for the large differences among codon positions in the substitution rate and in the transition/transversion rate ratio. The MLEs of ω calculated using tree T_1 of Figure 1 are shown in Figure 3. The estimates ranged from 0.010 to 0.136, with more than 10-fold differences among genes. The very small ω ratios indicate that all the 106 genes are highly constrained.

Although it seemed easy to decide which of the models fitted the data better, it is much less straightforward to decide which of them was better at reconstructing phylogenetic trees. The unconventional nature of the estimation problem (Yang et al., 1995) makes it possible for a simple and incorrect model to outperform the more complex true model, as discovered in computer simulation studies (Yang, 1997a). The goodness of fit of a model was known not to translate directly into good performance in tree reconstruction (Gaut and Lewis, 1995). Overall, simple models tend to perform better when the tree is easy to recover, whereas more complex and realistic models are critical for recovering difficult trees, for example, for avoiding long-branch attraction (Felsenstein, 1978; Huelsenbeck, 1997). In the yeast data, the gamma model of rates for sites, in the amino acid models WAG+F+G and FromCodon+G and in the nucleotide model HKY+G and HKY+C+G, did not improve the accuracy of tree reconstruction although it significantly improved the model's fit to data. We note that in the literature, simple-minded use of LRT and AIC for model selection (Posada and Crandall, 1998) almost invariably led to overly complex models such as GTR+I+G. We warn against such a practice, as such parameter-rich models may not produce more reliable phylogenies. Besides the fit of the model to data, one should also consider the biological interpretations of the models and the robustness of the analysis to model assumptions. Similar points were made by Minin et al. (2003), who chose models to achieve best performance in estimating branch lengths.

Estimation of Relative Divergence Times

Here we studied a more conventional estimation problem in phylogenetics, i.e., molecular clock dating of species divergences, to characterize the information loss when DNA sequences are translated into proteins. We used the molecular clock to estimate the relative node ages, defined as the ratio of the age of the internal node to the age of the root in the rooted tree T_1 of Figure 1. If the evolutionary process is clock-like, both amino acid substitutions and codon substitutions will accumulate at constant rates, and the relative node ages will have the same biological definitions under amino acid and codon models. Tree T_1 has six non-root internal nodes, so we estimated six relative node ages. The analysis was achieved by treating the root as a "fossil" calibration node, fixing its age at 1. Results obtained from the concatenated data are presented in Table 6. Model assumptions are known to be much more important for estimation of divergence dates than for tree topology reconstruction (e.g., Yoder and Yang, 2000). Indeed, considerable differences exist among node ages estimated under different models. In particular, HKY is highly unrealistic as it fails to account for differences among codon positions or for variable rates among sites and produced serious underestimation of large branch lengths around the root. As nodes 1 to 6 are all younger than the root (the calibration node), all their ages are overestimated by the model (Yang, 1996). The differences in estimates between models with and without gamma rates for sites (Table 6) can be explained in the same way. The GTR models were used in the analysis as well, and produced very similar results to those obtained under the corresponding HKY models (results not shown).

We compared the codon model F3×4MG and the amino acid model FromCodon to assess the information loss when the gene sequences are translated into proteins. The MLEs of relative node ages were similar between the two models, although the recent divergences were slightly younger under the amino acid model (Table 6). We calculated the asymptotic variances of the age estimates by inverting the Hessian matrix, the matrix of the second derivatives of the log likelihood with respect to the parameters (Efron and Hinkley, 1978), and defined V_c/V_{aa} as the efficiency of amino acid-based analysis relative to the codon-based analysis. The proportion of information loss is then measured by $1 - V_c/V_{aa}$. This was calculated to be 0.30, 0.36, 0.48, 0.42, 0.49, and 0.66 for nodes 6 to 1, respectively (see Fig. 1). Thus, translation of the DNA sequences into proteins caused an information

TABLE 6. MLEs of relative divergence times in tree T_1 of Figure 1 under different models using concatenated data.

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Amino acid						
WAGF	0.0386 ± 0.0011	0.0653 ± 0.0013	0.0897 ± 0.0015	0.1136 ± 0.0018	0.4117 ± 0.0042	0.5469 ± 0.0049
WAGF+G	0.0291 ± 0.0008	0.0495 ± 0.0011	0.0684 ± 0.0012	0.0847 ± 0.0015	0.3441 ± 0.0041	0.4481 ± 0.0049
FromCodon	0.0314 ± 0.0009	0.0529 ± 0.0011	0.0735 ± 0.0012	0.0943 ± 0.0015	0.3768 ± 0.0036	0.5087 ± 0.0044
FromCodon+G	0.0162 ± 0.0005	0.0276 ± 0.0006	0.0390 ± 0.0008	0.0482 ± 0.0010	0.2520 ± 0.0036	0.3307 ± 0.0044
Nucleotide						
HKY	0.1177 ± 0.0012	0.1880 ± 0.0015	0.2478 ± 0.0017	0.2889 ± 0.0019	0.6072 ± 0.0034	0.7227 ± 0.0039
HKY+C	0.0631 ± 0.0008	0.1026 ± 0.0011	0.1351 ± 0.0014	0.1555 ± 0.0015	0.4623 ± 0.0034	0.5777 ± 0.0040
HKY+G	0.0455 ± 0.0008	0.0745 ± 0.0011	0.0995 ± 0.0014	0.1114 ± 0.0016	0.3602 ± 0.0041	0.4281 ± 0.0048
HKY+C+G	0.0380 ± 0.0006	0.0648 ± 0.0009	0.0882 ± 0.0012	0.1011 ± 0.0013	0.3721 ± 0.0037	0.4500 ± 0.0044
Codon						
F3×4	0.0374 ± 0.0006	0.0613 ± 0.0009	0.0807 ± 0.0011	0.0925 ± 0.0012	0.3723 ± 0.0034	0.5011 ± 0.0041
F3×4MG	0.0419 ± 0.0006	0.0687 ± 0.0009	0.0897 ± 0.0012	0.1021 ± 0.0013	0.3720 ± 0.0034	0.4993 ± 0.0041
F3×4MG+M3	0.0231 ± 0.0005	0.0380 ± 0.0007	0.0501 ± 0.0009	0.0555 ± 0.0010	0.2468 ± 0.0033	0.3201 ± 0.0041

loss of from 30% for the deep node to 66% for the most recent node. Similar patterns were observed in separate analyses of individual genes or proteins, although there was considerable variation among genes, especially for the recent nodes (results not shown).

DISCUSSION

Information Content in Amino Acid and Codon Sequences

We note that several factors may complicate our effort to quantify the amount of phylogenetic information in amino acid, nucleotide and codon sequences. First, the yeast genes are real data, for which none of the models considered in this article can be expected to be true. Thus differences among the analyses or models may reflect not only different precisions but also systematic biases due to model inadequacies. Ideally we would like to use the "true" model at each level of the data, so that the phylogenetic precision is accurately measured and the differences in the results reflect the information content at each level. In that case, the nucleotide- and codon-based analyses should produce identical results since the data are the same. Usually computer simulation can be used to avoid the problem of not knowing the truth, as in a simulation, the model is under the control of the investigator and the same model can be used both to simulate and to analyze the data. However, it is difficult or impossible to construct amino acid or nucleotide models to accurately describe codon evolution in protein-coding genes, so that this "ideal" situation is not achievable even in simulations. Nevertheless, the similarity of results among models used at each level of the data suggests that this problem may not be too serious. Admittedly, HKY and GTR are very unrealistic for coding sequences. However, the "C" and "C+G" models appear to capture, even though in ad hoc ways, the major features of the evolutionary process in coding sequences. The similarities between amino acid models WAG+F and FromCodon, and between codon models F3×4 and F3×4MG, were even greater, implying that our conclusions about the utility of codon models, and about the phylogenetic information content in amino acid and nucleotide sequences, etc., are unlikely to be affected by minor changes to the assumed models.

A second complication is the unusual nature of the phylogeny if viewed as a parameter to be estimated, and the resulting difficulty of defining the "variance" or some other measure of the sampling error for the estimated tree. When the amino acid and codon models infer different ML trees, it is not straightforward to measure their difference. This difficulty is conceptual and appears to arise in every comparison of tree reconstruction methods (Yang et al., 1995). In this study, we used the RELL bootstrap proportion for tree T_1 of Figure 1 as a measure of phylogenetic information, so that both failure to infer T_1 as the ML tree and a lower support for T_1 will be considered poor performance. Our results appeared to be consistent across the different analyses and also with previous simulation studies. For example, several analyses we conducted serve to demonstrate the loss of information when gene sequences were translated into proteins: (a) amino acid models recovered tree T_1 in fewer genes or proteins than codon models or nucleotide models HKY+C and GTR+C, mainly due to poor performance of amino acid models in recovering recent nodes; (b) the RELL bootstrap support values for T_1 were lower in amino acid-based analyses than in codon- or nucleotide-based analyses; and (c) estimates of divergence times had larger sampling errors under amino acid models, especially for recent nodes.

The approach we took in this study is the same as those of Cummings et al. (1995), Russo et al. (1996), and Zardoya and Meyer (1996), who evaluated the performance of mitochondrial protein-coding genes and of different tree reconstruction methods to recover mammalian or vertebrate phylogenies. Even though such empirical studies suffer from limitations including those discussed above, we suggest that they are highly valuable for furthering our understanding of advantages and limitations of various models in solving real-world phylogenetic problems.

Utility of Codon Models in Phylogeny Reconstruction

Our analysis of the yeast genes in this study suggests that codon models may be useful for phylogenetic tree

reconstruction. However, codon models involve much heavier computation than nucleotide or amino acid-based models, because of the expanded character state space: the number of characters or the number of states in the Markov process under the nucleotide, amino acid, and codon models is 4, 20, and 61, respectively. The computational burden means that codon models will not be feasible for heuristic tree search in large data sets. Clearly, the utility of codon models will be greatly enhanced by implementing efficient tree search algorithms under such models. However, we suggest that they may be used to evaluate candidate phylogenies collected using other faster methods, as in this study. Likelihood calculation under codon models on fixed trees is currently feasible for data sets of a few hundred sequences (Yang, 2000). Second, for distantly related species, the base compositions at the third codon positions may differ among sequences, indicating that the substitution process is not homogeneous, and thus an assumption of the codon models is violated. Previous studies on nucleotide-based analysis suggest that nonhomogeneous base compositions may mislead tree reconstruction methods (e.g., Galtier and Gouy, 1998). Similar problems may be expected under codon models. In such cases, analysis of amino acid sequences may be advantageous (Kishino et al., 1990). Some of the 106 yeast genes appear to exhibit nonhomogeneous base compositions at the third codon positions. However, results of this study suggest that the problem is not serious enough to mislead ML analysis under the nucleotide and codon models as long as the heterogeneity among the three codon positions is accommodated. Deep nodes 4 and 5 (Fig. 1a) might be expected to be affected by nonhomogenous base compositions, but they were recovered in most genes by the codon or nucleotide "C" models (Table 3). The poorer performance in recovering recent nodes 2 and 3 (Table 3) appeared to be due to the short internal branch lengths or lack of phylogenetic resolution, as base compositions among the recent species are quite homogeneous. Instead, loss of information in the amino acid sequences appears to be a more serious concern for those data. It is worthwhile to analyze more data sets, either real or simulated, to examine the robustness of phylogenetic analysis under codon models to such violations of assumptions.

ACKNOWLEDGMENTS

We thank Tim Collins, Jeff Thorne, and an anonymous referee for critical comments. This study is supported by grants from the Biotechnological and Biological Sciences Research Council and the Human Frontier Science Programme to Z.Y. and a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, and Technology of Japan to F.R. and H.T.

REFERENCES

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

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. AC 19:716–723.

Bielawski, J. P., and Z. Yang. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59:121–132.

Chang, B. S., K. Jonsson, M. A. Kazmi, M. J. Donoghue, and T. P. Sakmar. 2002. Recreating a functional ancestral archosaur visual pigment. Mol. Biol. Evol. 19:1483–1489.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12:814–822.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pages 345–352 *in* Atlas of protein sequence and structure, Volume 5, Supplement 3, National Biomedical Research Foundation, Washington DC.

Efron, B., and D. V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed and expected information. Biometrika 65:457–487.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Forsberg, R., and F. B. Christiansen. 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol. Biol. Evol. 20:1252–1259.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

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

Guindon, S., A. G. Rodrigo, K. A. Dyer, and J. P. Huelsenbeck. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl. Acad. Sci. USA. 101:12957–12962.

Hasegawa, M., and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. Mol. Biol. Evol. 11:142–145.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Holland, B. R., K. T. Huber, V. Moulton, and P. J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. 21:1459–1461.

Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

Huelsenbeck, J. P., and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. J. Mol. Evol. 58:661–672.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Kosakovsky Pond, S. L., S. D. W. Frost, and S. V. Muse. 2005. HyPhy: Hypothesis testing using phylogenies. BioInformatics 21:676–679.

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

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12:219–236.

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

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

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Posada, D., and K. A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Russo, C. A., N. Takezaki, and M. Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13:525–536.

Schwaz, G. 1978. Estimating the dimension of a model. Ann. Statist. 6:461–464.

Seo, T. K., H. Kishino, and J. L. Thorne. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Mol. Biol. Evol. 21:1201–1213.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

Swofford, D. L. 1999. PAUP*: Phylogenetic analysis by parsimony, Version 4. Sinauer Associates, Sunderland, Massachusetts.

Takezaki, N., and T. Gojobori. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. Mol. Biol. Evol. 16:590–601.

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11:261–277.

Tavare, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics in the Life Sciences 17:57–86.

Taylor, D. J., and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21:1534–1537.

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

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

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. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Yang, Z. 1997a. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 14:105–108.

Yang, Z. 1997b. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556 (http://abacus.gene.ucl.ac.uk/software/paml.html).

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

Yang, Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51:423–432.

Yang, Z., N. Goldman, and A. E. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

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

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

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

Zardoya, R., and A. Meyer. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol Biol Evol 13:933–942.

First submitted 19 January 2005; reviews returned 25 March 2005; final acceptance 24 May 2005
Associate Editor: Tim Collins