Pacific Symposium on Biocomputing 5:78-89 (2000)

RELATING PHYSICOCHEMMICAL PROPERTIES OF AMINO ACIDS TO VARIABLE NUCLEOTIDE SUBSTITUTION PATTERNS AMONG SITES

ZIHENG YANG

Department of Biology (Galton Laboratory), University College London, 4 Stephenson Way, London NW1 2HE, England

Markov-process models of codon substitution were implemented that account for features of DNA sequence evolution (such as transition/transversion bias and codon usage bias) as well as heterogeneity of amino acid substitution pattern over sites. The codon (amino acid) sites are assumed to come from several classes (such as secondary structure categories), among which the rate of amino acid substitution and the effect of amino acid chemical properties vary. Parameters are estimated by the maximum likelihood method, which accounts for the phylogenetic relationship among species and corrects for multiple hits at the same site. The likelihood ratio test is used to compare models. Mitochondrial cytochrome *b* genes of 28 primate species are analyzed. The site-heterogeneity models provide much better fit to previous homogeneous models.

1 Introduction

There is an urgent need for combining phylogenetic analysis of protein or protein-coding DNA sequences with protein secondary structure prediction. The importance of accounting for the phylogenetic relationship in structure prediction by sequence comparison has been increasingly realized (Benner *et al.* 1994). For example, Goldman *et al.* (1996) and Thorne *et al.* (1996) used a hidden Markov chain approach to structure prediction. Amino acids in a protein are assumed to come from several structural classes (such as α-helix, β-sheet, turn, and coil), and a Markov chain is used to describe the transition over amino acid sites from one structural class to another. Those authors obtained relative amino acid substitution rates in each structural category from databases, and do not estimate any parameters specific to the protein being analyzed. The fine-grade classification of sites into secondary structural categories and the use of compiled substitution matrices may be important to achieve a high accuracy in structure prediction (Goldman *et al.* 1998). However, when we are interested in understanding the characteristics of the gene or protein, it may be advantageous to estimate substitution parameters from the data set. The substitution pattern, even in the same structural category, may be different among genes if the proteins perform different functions.

Phylogeny-based evolutionary analysis has largely ignored the structural differences of amino acid sites in the protein and concentrated on estimating the average substitution rates between amino acids. Early work includes the empirical amino acid substitution matrix of Dayhoff et al. (1978) and its update by Jones et al. (1992). Recently, substitution matrices were estimated for specific proteins (such as mitochondrial proteins) using more powerful likelihood methods (Adachi and Hasegawa 1996; Yang et al., 1998). Those analyses assume that the substitution pattern is homogeneous among amino acid sites.

Several attempts have been made to account for the among-site heterogeneity. For example, Yang (1994, 1995) and Felsenstein and Churchill (1996) developed models of variable substitution rates among amino acid sites. Those models account for the existence of fast and slow sites in the protein, but the relative substitution rates between amino acids are assume to be the same at all sites. Bruno (1996) suggested a model that allows each amino acid site in a protein to have a different and yet strong preference for a particular amino acid. Koshi *et al.* (1999) developed heterogeneous amino acid substitution models, in which amino acid sites come in several classes, and amino acid chemical properties affect their substitution rates in different ways among the classes. Those models ignore the mutational distance between amino acids determined by the genetic code, and the estimation procedure used needs justification as well.

Models of codon substitution make it possible to separate mutational biases in the DNA from selective constraints on the protein, and offer a great advantage over amino acid models for understanding the evolutionary process of proteins and protein-coding DNA sequences. An important biological parameter in codon-based analysis is the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$), also known as the acceptance rate by Miyata *et al.* (1979). This parameter measures the selective constraint in the protein. Simply, a nonsynonymous mutation is neutral if $\omega = 1$, advantageous if $\omega > 1$, or deleterious if $\omega < 1$. A low substitution rate between two amino acids can either be due to a large mutational distance between the two amino acids or a small acceptance rate, which may be caused by a large physico-chemical distance. Chemical properties of amino acids should be used to modify acceptance rates and not amino acid substitution rates. Yang *et al*. (1998) described an approach to constructing an amino acid substitution model from a codon substitution model, and examined the relationship between amino acid chemical properties and acceptance rates. That relationship was also examined by Xia and Li (1998), who reconstructed ancestral DNA sequences to count changes along the phylogeny. Both studies confirmed early suggestions (e.g., Zuckerkandl and Pauling 1965) that similar amino acids tend to replace each other more often than dissimilar ones. The relationship between the chemical distance and the acceptance rate is not simple, however, and one reason suggested was the dependence of acceptance rate on the structural context of the protein. Nielsen and

Yang (1998) constructed likelihood models that allow the ω ratio to vary among sites.

In this paper, the acceptance rate ω is assumed to be influenced by the chemical properties of the amino acids interchanged, and the effect of the chemical properties is assumed to differ among site classes. The model extends previous codon-substitution models of Goldman and Yang (1994), Yang *et al.* (1998), and Nielsen and Yang (1998). It also appears more plausible biologically than the amino acid substitution models of Bruno (1996) and Koshi *et al.* (1999). The codon-based model accounts for features of nucleotide substitution such as transition/transversion rate bias and codon usage bias. It also takes into account biological processes such as translation of the DNA into protein according to the genetic code and acceptance or rejection of the resulting amino acid under selective constraints on the protein. The estimation is achieved using maximum likelihood (ML), which naturally accounts for the phylogenetic relationship and corrects for multiple hits at the same site. A data set of mitochondrial cytochrome b genes from 28 primate species is analyzed to compare different models.

2 Theory

2.1 Markov Model of Codon Substitution

A simple codon substitution model is described first with further complications introduced later. The basic model specifies instantaneous substitution rate from codon *u* to codon *v* as

$$q_{uv} = \begin{cases} 0, & \text{if } u \text{ and } v \text{ differ at two or three codon positions,} \\ \pi_v, & \text{if } u \text{ and } v \text{ differ by a synonymous transversion,} \\ \kappa\pi_v, & \text{if } u \text{ and } v \text{ differ by a synonymous transition,} \\ \omega\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transition,} \end{cases} \quad (1)$$

(Goldman and Yang 1994). Parameter κ is the transition/transversion rate ratio, with $\kappa = 1$ meaning no transition bias. The equilibrium frequency of codon *v* (π_v) are calculated using the nucleotide frequencies at the three codon positions, with 9 [$= 3 \times (4 - 1)$] free parameters used.

The model specified by equation 1 assumes that different nonsynonymous mutations are fixed at the same rate and ignores the fact that some amino acids are similar in chemical properties so that changes between them are less disruptive to the structure and function of the protein than changes between dissimilar amino acids. Empirical observations suggest that amino acids with similar properties tend

to exchange more often than dissimilar amino acids (e.g., Zuckerkandl and Pauling 1965). Thus, ω in equation 1 is replaced by ω_{ij}, where $i = aa_u$ and $j = aa_v$ are the two amino acids involved. To account for the dependence of the acceptance rate on the chemical distance, a geometric relationship is used:

$$\omega_{ij} = a \exp\{- b d_{ij}/d_{\max}\}, \quad a \geq 0. \tag{2}$$

Yang et al. (1998) found that the distance of Miyata *et al.* (1979) provides the best fit to data among several distances examined. This distance measure is used in this paper, which is based on two chemical properties, polarity (p) and volume (v):

$$d_{ij} = \sqrt{(p_i - p_j)^2 / \sigma_{\Delta p}^2 + (v_i - v_j)^2 / \sigma_{\Delta v}^2}\ , \tag{3}$$

where σ_{Δ_p} and σ_{Δ_v} are the standard deviations of $|p_i - p_j|$ and $|v_i - v_j|$, respectively. The distance ranges from 0.06 for Pro – Ala to 5.13 for Gly – Trp.

We assume that amino acids come in several categories, among which the substitution pattern reflected in parameters a and b of equation 2 is different. One motivation for such models is the existence of secondary structure categories in the protein. Suppose that there are K site classes. Parameters in the model will include the proportions (subject to the constraint that the sum is one) and parameters a and b in equation 2 for each site class:

Category	1	2	…	K
Proportion	p_1	p_2	…	p_K
Parameters	a_1, b_1	a_2, b_2	…	a_K, b_K

Use of more categories will increase the fit of the model, but the data may not contain much information to allow estimation of many parameters. So only a few categories may be used in practice.

Another model used for comparison in this paper ignores the chemical distances between amino acids but assume that the ω ratio varies among amino acid sites (Nielsen and Yang 1998). The model parameters are

Category	1	2	…	K
Proportion	p_1	p_2	…	p_K
Parameters	ω_1	ω_2	…	ω_K

2.2 Maximum Likelihood Calculation on a Phylogeny

Given the substitution rate matrix $Q = \{q_{uv}\}$, the matrix of transition probabilities

over time t can be calculated as $P(t) = \{p_{uv}(t)\} = e^{Qt}$, where $p_{uv}(t)$ is the probability that codon u changes into v after time t. Time or branch length t is measured by the expected number of nucleotide substitutions per codon, averaged over the site classes. Note that the probability that codon u changes into codon v over any time interval t is positive; that is, $p_{uv}(t) > 0$ for any $t > 0$, even if the two codons are separated by two or three differences. A standard numerical algorithm is used to calculate the eigenvalues and eigenvectors of Q to calculate $P(t)$. Likelihood calculation under the heterogeneous models is described by Nielsen and Yang (1998). Let n be the number of sites (codons) in the sequence and the data at site h be x_h ($h = 1, 2, ..., n$); x_h is a vector of codons from different sequences at that codon site. Let y_h denote the class that site h belongs to; y_h takes a value from 1, 2, ..., K. The conditional probability, $f(x_h|y_h)$, of the data at site h given that site h is from class y_h, can be calculated for a given phylogenetic tree and branch lengths using Felsenstein's (1981) pruning algorithm (see also Goldman and Yang 1994; Muse and Gaut 1994). We assume that each site belongs to one of the K classes, but no information is available about which class each site is from. The probability of the data at the site is then an average of the conditional probability over the distribution of y_h.

$$f(x_h) = \sum_{y_h=1}^{K} p_k f(x_h \mid y_h). \tag{4}$$

The log likelihood is a sum over all n sites in the sequence

$$\ell = \sum_{h=1}^{n} \log\{f(x_h)\} \tag{5}$$

We assume independence of the substitution process among codon (amino acid) sites, although Markov dependence can easily be introduced through a hidden Markov chain model (see, e.g., Yang 1995; Felsenstein and Churchill 1996; Goldman *et al*. 1996). A numerical optimization algorithm is used to obtain ML estimates of parameters.

After parameter estimates are obtained, an empirical Bayes approach can be used to infer which class the site most likely belongs to. The posterior probability that a site with data x_h is from site class k is

$$\Pi(y_h = k \mid x_h) = \frac{p_k f(x_h \mid y_h = k)}{f(x_h)} = \frac{p_k f(x_h \mid y_h = k)}{\sum_{j=1}^{K} p_j f(x_h \mid y_h = j)} \tag{6}$$

The class k that maximizes the posterior probability is the most likely class for the site. The posterior probability provides a measure of accuracy for that inference.

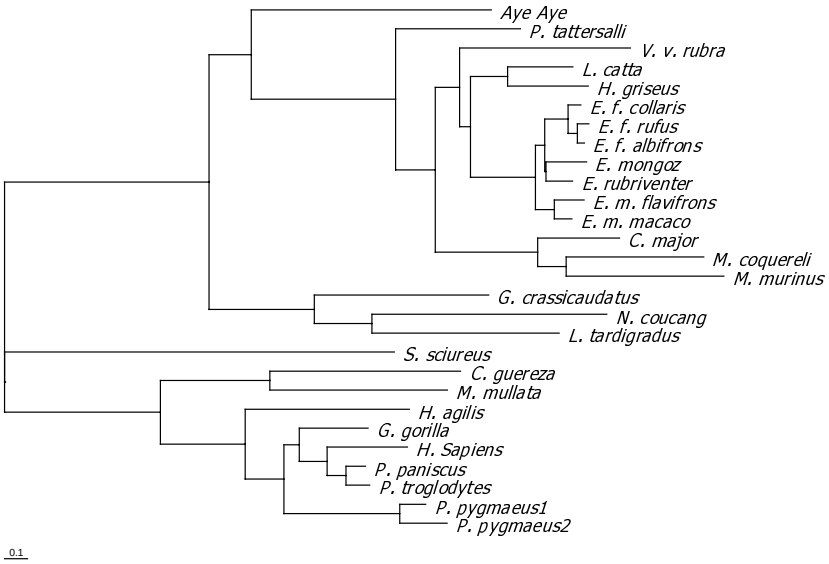

Fig. 1. *The phylogenetic tree of the 28 primate species analyzed in this paper. Branch lengths, measured by the number of nucleotide substitutions per codon, are estimated by ML under the model of equation 1 (Goldman and Yang 1994).*

3 Application to Mitochondrial Cytochrome *b* Genes of Primates

3.1 Sequence Data

The data are mitochondrial cytochrome *b* genes from 28 species of primates. The species include 15 Lemuriformes: *Lemur catta* (GenBank accession number U53575), *Hapalemur griseus* (U53574), *Eulemur fulvus collaris* (U53576), *Eulemur fulvus rufus* (U53577), *Eulemur fulvus albifrons* (AF081048), *Eulemur macaco macaco* (AF081049), *Eulemur macaco flavifrons* (AF081050), *Eulemur mongoz* (AF081051), *Eulemur rubriventer* (AF081052), *Varecia variegata rubra* (U53578), *Cheirogaleus major* (U53570), *Mirza coquereli* (U53571), *Microcebus murinus* (U53572), *Propithecus tattersalli* (U53573), and *Daubentonia madagascariensis* (U53569); 3 Lorisiformes: *Galago crassicaudatus* (U53579), *Loris tardigradus* (U53581), and *Nycticebus coucang* (U53580); and 10 Anthropoidea: *Saimiri sciureus* (U53582), *Colobus guereza* (U38264), *Macaca mulatta* (U38272), *Hylobates agilis* (U38263), *Pongo pygmaeus* (U38274), *Pongo pygmaeus*

(D38115), *Pan paniscus* (D38116), *Pan troglodytes* (D38113), *Gorilla gorilla* (D38114), *Homo sapiens* (J01415). The alignment is from Yang and Yoder (1999). Two codons involve undetermined nucleotides and are removed, with 377 codons (1131 nucleotides) in the sequence. The phylogeny of those species is largely resolved, and one of the most likely phylogenies is shown in figure 1. Some analyses were performed using several candidate tree topologies, and the parameter estimates are virtually identical; results obtained from using the tree of figure 1 only are presented in this paper.

3.2 Comparison of Models and Estimation of Parameters

Some parameters are common to all models considered in this paper, which are estimated for each model but are not presented. These include 53 branch lengths in the tree (figure 1), which are estimated by ML, and the base frequency parameters at the three codon positions (9 free parameters), which are estimated by the observed frequencies. All models also involve the transition/transversion rate ratio parameter κ.

The site-homogeneous model of Goldman and Yang (1994), specified by equation 1, is applied to the data set (table 1). The log-likelihood value under this model is –12285.15. The estimate $\omega = 0.041$ suggests that on average, nonsynonymous mutations are fixed at a rate only 4% that of the synonymous mutations, indicating that cytochrome *b* is a highly conserved protein. Results obtained under models of variable ω ratios among sites (Nielsen and Yang 1998) are listed in table 1. These are the A*k* models (table 1), where *k* is the number of site classes in the model. Allowing for heterogeneous ω ratios among (codon) sites increases the model's fit greatly. For example, the model with two site classes (A2) involves only two more parameters than the homogeneous model with one site class (Model A1), but the log-likelihood difference is $\Delta\ell = (-11892.55) - (-12285.15) = 392.60$. This is much greater than $\frac{1}{2}\chi^2_{1\%} = 4.65$. There is no doubt that the selective constraint reflected in the ω ratio differs among sites. The estimates under model A2 suggests that a large proportion of sites (>70%) are highly conserved with $\omega = 0.007$, while the remaining sites are moderately conserved ($\omega = 0.137$). The gain in log likelihood upon adding more site classes quickly becomes small. The difference between model A4 (4 site classes) and model A3 (3 site classes) is marginally significant, with the log-likelihood difference $\Delta\ell = 3.44$.

Table 1. Models of variable acceptance rates ($\omega = d_N/d_S$) among sites

Model	ℓ	κ	Parameters for site classes	d_N/d_S	S
A1	−12285.15	6.51	$\omega = 0.041$	0.041	19.3
A2	−11892.55	6.83	p: 0.731 0.269 ω: 0.007 0.137	0.042	20.4
A3	−11849.46	7.17	p: 0.618 0.262 0.119 ω: 0.003 0.058 0.225	0.044	21.4
A4	−11846.02	7.20	p: 0.584 0.243 0.121 0.052 ω: 0.002 0.043 0.138 0.306	0.044	21.5

Note.—Tree length (S) is the expected number of nucleotide substitutions per codon along the tree.

Table 2. Models incorporating amino acid chemical properties

Model	ℓ	κ	Parameters for site classes	d_N/d_S	S
B1	−12200.51	7.07	a: 0.086 b: 2.832	0.042	20.0
B2	−11805.70	7.76	p: 0.752 0.248 a: 0.036 0.277 b: 6.462 2.498	0.042	21.9
B3	−11758.40	8.19	p: 0.607 0.277 0.116 a: 0.011 0.172 0.382 b: 5.612 4.646 1.970	0.044	23.0
B4	−11747.95	8.50	p: 0.102 0.248 0.548 0.103 a: 0.817 0.169 0.007 0.417 b: 31.105 3.982 3.871 2.086	0.043	23.5

Note.—Tree length (S) is the expected number of nucleotide substitutions per codon along the tree.

Likelihood values and ML parameter estimates obtained under models accounting for amino acid chemical properties are listed in table 2. Those are the B*k* models (table 2), where *k* is the number of site classes. Parameters *a* and *b* are defined in equation 2. Models A1 and B1 are both homogeneous models (with one site class), but model B1 uses amino acid distances of Miyata *et al.* (1979) to modify the acceptance rate ω. The log-likelihood difference between the two models is $\Delta\ell$ = 84.64, and model B1 is significantly better than A1. Use of amino acid chemical properties improves the fit of the model significantly, and the acceptance rate is negatively correlated with the chemical distance.

Comparison of models of table 2 again indicates a huge amount of among-site

heterogeneity. For example, model B2 (with 2 classes) involves two more parameters than the homogeneous model B1 (with 1 site class). The log-likelihood difference, $\Delta\ell = 394.81$, is extremely significant (compared with a χ^2 distribution with d.f. = 2). Similar to results of table 1, the gain upon adding more site classes becomes minor. The fit of the model measured by the log-likelihood value is plotted in figure 2 as a function of the model complexity (the number of parameters in the model). Model B4 with 4 site classes fits the data significantly better than model 3B with 3 site classes, but the log-likelihood difference ($\Delta\ell = 10.45$) is not very large.

The tree length, that is, the sum of branch lengths, measured by the total number of nucleotide substitutions per codon along the phylogenetic tree, is greater for more complex models than for simple models. This pattern is the same as observed in nucleotide-based analysis, since simple models do not correct for multiple hits properly and tend to underestimate branch lengths. For similar reasons, simple models also tend to produce smaller estimates of the transition/transversion rate ratio κ. However, the differences in those estimates are small, probably because nonsynonymous rates are quite low in the data. For other data sets with higher nonsynonymous divergences, the differences among models may be much larger.

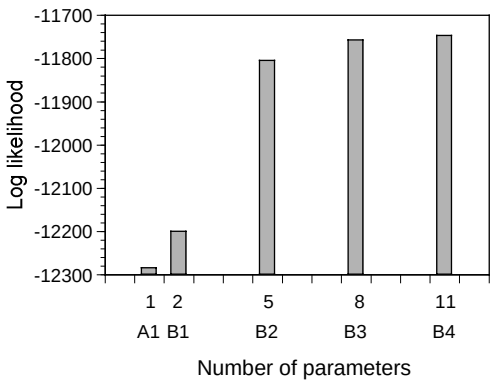

Fig. 2. *The fit of the model as a function of the complexity of the model*

3.3 Inference of Site Classes

Equation 6 is used to calculate the posterior probabilities of site classes for each site in the protein under the model of four site classes (see model B4 in table 2). The functional relationships between d and ω are shown in figure 3 for the four site classes C_1, C_2, C_3, and C_4. Class C_1 includes highly conserved sites with ω close to 0,

while class C_4 includes more variable sites, at which nonsynonymous mutations are tolerated more frequently (see model B4 in table 2). Posterior probabilities for the first 100 sites are plotted in figure 4.

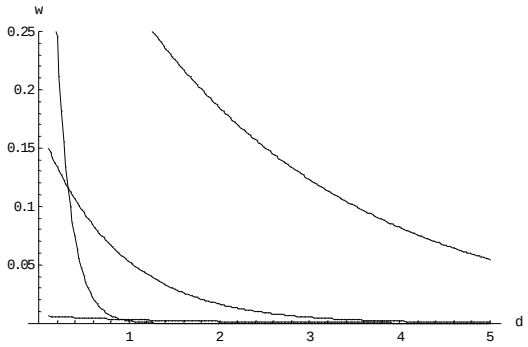

Fig. 3. *The four functional relationships estimated from the data (from bottom to top: C_1, C_2, C_3, C_4)*

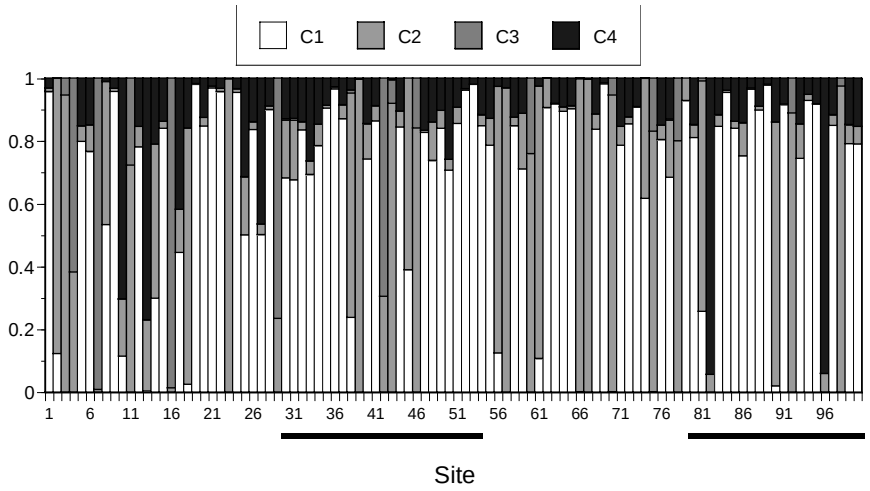

Fig. 4. *Posterior probabilities for site classes for the first 100 sites in the protein. The first two of the eight transmembrame regions are indicted by bars.*

Discussions

Both biological considerations and the analysis of the cytochrome *b* data sets suggest the importance of the heterogeneity of amino acid substitution patterns among sites.

The likelihood improvement when site heterogeneity is introduced into the model is tremendous. However, the relationship between amino acid chemical properties and the acceptance rates may not be so simple (Yang *et al.* 1998). For example, the size of an amino acid may be more important if the amino acid is buried inside than if it is exposed. Use of a common distance formula of Miyata *et al.* (1979) for all site classes does not catch this complexity of the substitution process. It may be more realistic to consider individual properties in each site class.

References

1. J. Adachi and M. Hasegawa, "Model of amino acid substitution in proteins encoded by mitochondrial DNA" *J. Mol. Evol.* **42**, 459–468 (1996)
2. S. A. Benner, I. Badcoe, M. A. Cohen and D. L. Gerloff, "*Bona fide* prediction of aspects of protein conformation. assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences" *J. Mol. Biol*. **235**, 926-958 (1994)
3. W. J. Bruno, "Modeling residue usage in aligned protein sequences via maximum likelihood" *Mol. Biol. Evol.* **13**, 1368-1374 (1996)
4. M. O. Dayhoff, R. M. Schwartz and B. C. Orcutt, Pp. 345–352 in *Atlas of protein sequence and structure, Vol 5, Suppl. 3*, "A model of evolutionary change in proteins" (National Biomedical Research Foundation, Washington D. C. 1978)
5. J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach" *J. Mol. Evol.* **17**, 368-376 (1981)
6. J. Felsenstein and G. A. Churchill, "A hidden Markov model approach to variation among sites in rate of evolution" *Mol. Biol. Evol.* **13**, 93-104 (1996)
7. N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences" *Mol. Biol. Evol.* **11**, 725 (1994)
8. N. Goldman, J. L. Thorne and D. T. Jones, "Using evolutionary trees in protein secondary structure prediction and other comparative sequence analysis" *J. Mol. Biol.* **263**, 196-208 (1996)
9. N. Goldman, J. L. Thorne and D. T. Jones, "Assessing the impact of secondary structure and solvent assessibility on protein evolution" *Genetics* **149**, 445-458 (1998)
10. D. T. Jones, W. R. Taylor and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences" *Comp. Appl. Biosci.* **8**, 275–282 (1992)
11. J. M. Koshi, D. P. Mindell and R. A. Goldstein, "Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes" *Mol. Biol. Evol.* **16**, 173-179 (1999)

12. T. Miyata, S. Miyazawa and T. Yasunaga, "Two types of amino acid substitutions in protein evolution" *J. Mol. Evol.* **12**:219–236 (1979)
13. S. V. Muse and B. S. Gaut, "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome" *Mol. Biol. Evol.* **11**, 715-724 (1994)
14. M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions" *Mol. Biol. Evol.* **3**, 418–426 (1986)
15. R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene" *Genetics* **148**, 929-936 (1998)
16. J. L. Thorne, N. Goldman and D. T. Jones, "Combining protein evolution and secondary structure" *Mol. Biol. Evol.* **13**, 666–673 (1996)
17. X.-H. Xia and W.-H. Li, "What amino acid properties affect protein evolution" *J. Mol. Evol*. **47**, 557-564 (1998)
18. Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods" *J. Mol. Evol.* **39**, 306- 314 (1994)
19. Z. Yang, "A space-time process model for the evolution of DNA sequences" *Genetics* **139**, 993-1005 (1995)
20. Z. Yang and A. Yoder, "Estimation of the transition/transversion rate bias and species sampling" *J. Mol. Evol.* **48**, 274-283 (1999)
21. Z. Yang, R. Nielsen and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution" *Mol. Biol. Evol.* **15**, 1600-1611 (1998)
22. E. Zuckerkandl and L. Pauling, Pp. 97–116 in *Evolving genes and proteins*, "Evolutionary divergence and convergence in proteins". Eds. V. Bryson and H. J. Vogel (Academic Press, New York, 1965)