A MODEL FOR PHYLOGENETIC INFERENCE USING STRUCTURAL AND CHEMICAL COVARIATES

SIMON TAVARÉ
Departments of Biological Sciences, Mathematics and Preventative Medicine,
University of Southern California,
Los Angeles, CA 90089, USA

DEAN C. ADAMS, OLIVIER FEDRIGO, GAVIN J. P. NAYLOR
Department of Zoology and Genetics,
Iowa State University,
Ames, IA 50011, USA

We investigated whether or not evolutionary change in DNA sequence data was homogeneous across different classes of base pairs. DNA sequences for eight protein-coding mitochrondrial genes were obtained for 38 vertebrate taxa from GenBank. Each nucleotide site in the alignment was classified according to a number of covariates, including its codon position, genetic code degeneracy, and hydrophobicity. The evolutionary transition matrix for each base was estimated by tracing implied character changes under parsimony on a known phylogenetic tree. Canonical variates analyses of the inferred transition matrices were performed for each gene to determine whether or not different classes of bases behaved similarly. We found five distinct clusters of transition matrices that could be roughly defined by combinations of codon position and degeneracy. This pattern was consistent among all genes. A stochastic model of rate variation based on the interaction of the covariates was developed to assess the statistical significance of the clusters. The five-group classification was found to explain significantly more sequence variation than did a codon only classification, a codon degeneracy classification, or a codon *and* degeneracy classification. The same five-group classification was found for all genes tested, suggesting a common process underlying the molecular evolution of the mitochondrial genome. These results confirm that there are classes of base pairs that evolve differently, and suggest that models of sequence evolution that incorporate covariate information may be useful in developing nucleotide substitution models that more accurately reflect evolutionary history.

1. Introduction

Estimating the phylogenetic relationships of organisms is the first step of many evolutionary studies. Typically, DNA sequence data are collected for the taxa of interest, and a phylogenetic relationship is estimated using a particular model of evolutionary change. The model is based on assumptions about the evolution of DNA. For example, maximum parsimony is based on the assumption that evolutionary changes are rare, and therefore minimizes the number of substitutions along the resulting phylogeny[1]. By contrast, maximum likelihood (ML) methods identify the phylogeny from which the observed sequence data is most likely to have evolved given a particular model of change[2]. In this approach parameters in the substitution model, the tree topology and branch lengths can be estimated from

the data. Several different models are used. Most use a 4×4 transition matrix describing the changes from A $\rightarrow$ C, A $\rightarrow$ G and so on. The Jukes-Cantor model is the simplest, in that it assumes that all transformations among nucleotides are equally likely. Other models, such as K2P[3], F81[4], TKF[5], HKY[6], variously account for differences in base composition and the ratio of transitions to transversions. The General Time Reversible model (GTR) allows for 9 of the 12 possible parameters of the nucleotide transition matrix[7].

Among the critical assumptions in this likelihood approach are (i) that the different base positions along the DNA sequence evolve according to the same stochastic model, represented by a single transition matrix; and (ii) that substitutions at different sites occur independently. It has long been known that these assumptions can be unrealistic[4,8], and several authors have proposed alternatives. For example, a rate parameter having the gamma distribution is sometimes incorporated in the models to account for rate variation across sites[9,10,11], and dependence among sites has been addressed as well[12,13,14]. The reviews of Yang[15] and Liò and Goldman[16] provide a good introduction to these issues.

More recently, structural and functional information have been included in models for DNA or protein sequence evolution. For example, Goldman and colleagues[17,18,19], working at the level of amino acid replacements, have incorporated secondary structure into phylogenetic estimation. Naylor and Brown[20] made allowance for covariates such as hydrophobicity, charge and size. In this paper, we introduce another approach to rate variation that can be used in nucleotide-based substitution models. Motivated by methods for analyzing contingency tables using log-linear models (cf. Dobson[21], Chapter 9), we parameterize the rates at each site in terms of main effects and interactions of a set of covariates. The method appears to be useful for the systematic identification of parsimonious classes of sites that have similar mutation rates.

2. Methods

2.1 Canonical variates analysis

DNA sequences for eight protein-coding mitochondrial genes were obtained for 38 vertebrate taxa from GenBank. For each gene we counted the number of changes occurring at each site using parsimony on the known phylogenetic tree (Figure 1). This yielded an evolutionary transition matrix for each site, which was used as input data for exploratory multivariate analyses. Each base was then assigned to one of three codon positions, one of three degeneracy classes (2, 4, 6) based on the redundancy of the genetic code at the amino acid level, one of two hydrophobicity classes (hydrophobic/hydrophilic), and for Cytochrome b, one of three secondary structure classes (helix, sheet, or neither). All of these classifications were

determined using the modal value in the alignment. The estimated transition matrix for each site was then used in a canonical variates analysis (CVA) to determine whether natural classes of sites with distinct modes of evolutionary change existed. Separate CVA analyses were performed for each gene, and their results were compared to one another.

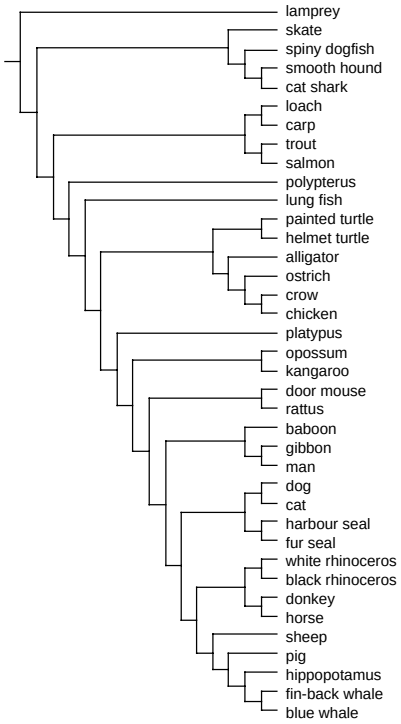

Figure 1. Phylogenetic tree used in this analysis. Tree was derived from both extant and fossil comparative anatomical data.

2.2 Statistical approach

To evaluate the statistical significance of the rate classes found by the exploratory CVA analyses, we use the following likelihood-based approach. We assume initially that the substitution matrix at site i has the form $m_i Q$, where the rate matrix Q is common to each site, and $m_i \geq 0$ specifies the relative rate at site i. Suppose for ease of exposition that each site can be classified according to a number of categorical covariates, such as codon position and degeneracy. A given site is therefore classified into one of $J \times K \times \cdots \times L$ cells. For a main effects model, the

rate parameter m of a site with covariate values $j, k, \ldots, l$ satisfying $1 \le j \le J, \ldots, 1 \le l \le L$ is given by

$$\log m = \mu + \alpha_j + \beta_k + \cdots + \lambda_l . \qquad (1)$$

Some care must be taken to ensure that the parameters are estimable. For the rooted trees discussed below, we chose to fix the tree height at 1 (say), and we set $\alpha_1 = \beta_1 = \cdots = \lambda_1 = 0$. In this formulation, e^{μ} represents the baseline rate for a site with covariate values 1,1,…,1 and other covariate values affect this rate in a multiplicative fashion. We note that (i) interactions among the covariates can be modeled by adding appropriate terms like $(\alpha\beta)_{jk}$, $(\alpha\beta\gamma)_{jkl}$ on the right of (1); (ii) a quantitative covariate x can be included by adding terms such as δx to the right of (1); (iii) the link function log on the left of (1) can be replaced by linear or other functions; (iv) the rate matrix Q can vary with the covariate class, and (v) the method generalizes naturally to the simultaneous study of multiple genes.

In our application, we used the TKF[5] model that allows for a transition-transversion parameter $\kappa > 0$ (assumed to be the same at each site), and fixed base frequencies as estimated from the data. Likelihoods, denoted by L, were obtained for the model with no covariates, corresponding to constant rate across sites. Covariates (codon position, codon degeneracy, hydrophobicity, and secondary protein structure) were then added sequentially to the model, and the corresponding likelihoods were found. The covariate providing the largest increase in the likelihood was determined. This factor was added to the 'null' model, and the remaining factors were added incrementally in a step-wise regression fashion. The differences in $-2 \log L$ values were calculated for each additional factor and compared to a χ^2 distribution to assess significance[22].

3. Results

3.1 Exploratory analysis

For all genes, CVA revealed highly significant differences in evolutionary transition matrices for codon position and genetic code degeneracy groups, a slight effect for hydrophobicity, and negligible effects for secondary structure for Cytochrome b. Five distinct clusters were recognized in the ordination plot. When bases were assigned to one of these five groups using the canonical discriminant function, approximately 45% were correctly classified. Inspection of plots of canonical scores revealed that several position/codon degeneracy groups that were consistently close to one another, implying that their evolutionary transition matrices were similar (Figure 2). All second codon positions formed a distinct

cluster, regardless of their degeneracy groupings. First codon positions from 2-fold and 4-fold groups were similar, as were third codon positions from 4-fold and 6-fold groups. These were therefore considered as two larger clusters rather than four individual groups. Thus, the nine (3×3) possible combinations of codon position and codon degeneracy collapsed to five distinct modes of evolutionary change; a pattern which was consistent among genes. We calculated classification rates for these five larger clusters and found that $\approx$ 65% of the sites were now correctly classified, implying that many of the original misclassifications were sites assigned to position-degeneracy groups that were part of a larger group.

3.2 Likelihood results

Results for the main effects models, based on a separate analysis of each gene, are shown in Table 1. These results correspond to statistical tests for the exploratory CVA analyses, and reveal a similar pattern to that described above. Likelihood scores associated with different decompositions of covariate structure for codon position and degeneracy are shown in Table 2. It is clear that codon position and degeneracy make an important contribution to the fit, whereas hydrophobicity and structure are less important. The five-group codon position/degeneracy model suggested by the exploratory analysis includes interaction effects, and its likelihood was also computed. A comparison of likelihoods shows that the five-group model is better than the additive main effects model using degeneracy and codon position for all genes (Table 2). We conclude that there is a significant interaction effect between degeneracy and codon position in determining the rates at these sites. The five-group model seems to capture much of the true biological signal in sequence variation for all eight protein-coding genes evaluated.

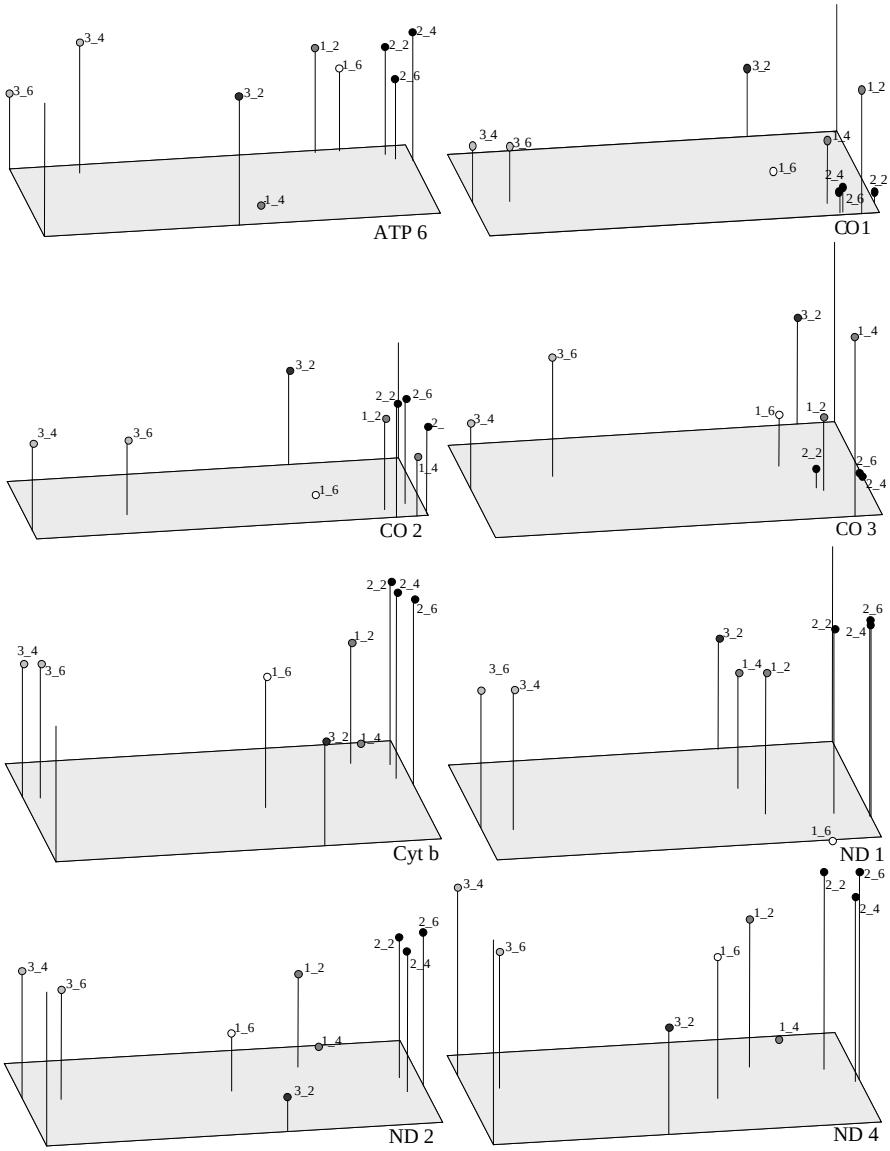

Figure 2. Canonical variates plots of all possible codon position-degeneracy combinations for eight mitochondrial protein-coding genes. Only group means are shown. Labels for group means designate codon position and degree of fold degeneracy (gray scale included to emphasize five-group classification).

Table 1. Likelihood scores for main effects covariate models for the eight protein-coding genes. Model parameters are: CR: constant rate, CD: codon, DG: degeneracy, HD: hydrophobicity, ST: structure.

Model	# Par	-2log*L*	Diff	Df	P
ATP6					
CR	38	40317.6			
+ CD	40	37806.6	2511.0	2	***
+ DG	42	37757.8	48.8	2	**
+ HD	43	37756.9	0.9	1	ns
CO 1					
CR	38	65621.5			
+ CD	40	56364.1	9257.4	2	***
+ DG	42	53213.5	3150.6	2	***
+ HD	43	53102.3	111.2	1	*
CO 2					
CR	38	29645.2			
+ CD	40	26420.1	3225.1	2	***
+ DG	42	26171.1	249.0	2	***
+ HD	43	26090.4	80.7	1	*
CO 3					
CR	38	39559.4			
+ CD	40	35538.9	4020.5	2	***
+ DG	42	34834.8	704.1	2	***
+ HD	43	34832.5	2.3	1	ns

Model	# Par	-2log*L*	Diff	Df	P
Cyt b					
CR	38	52983.1			
+ CD	40	48470.9	3912.2	2	***
+ DG	42	47159.5	1311.4	2	***
+ HD	43	47125.6	33.9	1	*
ND 1					
CR	38	53128.6			
+ CD	40	49215.5	3913.1	2	***
+ DG	42	48936.5	279.0	2	***
+ HD	43	48918.0	18.5	1	*
ND 2					
CR	38	51776.9			
+ CD	40	48652.9	3124.0	2	***
+ DG	42	48389.0	263.9	2	***
+ HD	43	48375.4	13.6	1	*
ND 4					
CR	38	74304.0			
+ CD	40	69235.3	5068.7	2	***
+ DG	42	68922.6	312.7	2	***
+ HD	43	68918.7	3.9	1	ns

Table 2. Likelihood scores for different decompositions of covariate structure. Model parameters are: CR: constant rate, C+D: codon plus degeneracy (additive), 5G:five codon/degeneracy group model.

Model	# Par	-2log*L*
ATP6		
CR	38	40317.6
C+D	42	37757.8
5G	42	37552.7
CO 3		
CR	38	39559.4
C+D	42	34834.8
5G	42	34596.5
ND 2		
CR	38	51776.9
C+D	42	48389.0
5G	42	48100.2

Model	# Par	-2log*L*
CO 1		
CR	38	65621.5
C+D	42	53213.5
5G	42	52570.7
Cyt b		
CR	38	52983.1
C+D	42	47159.5
5G	42	46844.0
ND 4		
CR	38	74304.0
C+D	42	68922.6
5G	42	68503.8

Model	# Par	-2log*L*
CO 2		
CR	38	29645.2
C+D	42	26171.1
5G	42	25772.4
ND 1		
CR	38	53128.6
C+D	42	48936.5
5G	42	48562.3

4. Discussion

The fact that the multidimensional ordination yielded five distinct clusters of transition matrices, and that these clusters are statistically significant from one another indicates that sequence evolution is far from homogeneous across sites. Furthermore, the variation in rates is more complicated than that predicted by considering codon position alone. A covariate model fitting separate parameters for each class of sites is more likely to capture the diversity of molecular dynamics than is a 'one-model-fits-all' approach. While we acknowledge that complicated and parameter-rich homogeneous models can outperform the covariate approach advocated herein, such models have less biological explanatory power.

The five groupings we identified correspond closely to the combinatorial freedom to vary that might be predicted from a careful study of the genetic code. We found that all second position sites formed a distinct cluster: changes at those sites involve a one-to-one mapping between nucleotide and amino-acid sequence, because a second position change necessarily results in a change in amino acid regardless of the amino acids' degeneracy. Similarly, changes at third codon positions for both 4-fold and 6-fold degenerate amino acids involve a 4-to-1 mapping from the nucleotide sequence to the amino acid sequence at the third codon position, and thus might be expected to have similar degrees of evolutionary room to maneuver. Finally, changes at first position sites for all but 6-fold degenerate amino acids involve a one-to-one mapping, as they result in a change in amino acid. It might seem surprising that these first position sites, and second position sites, fall out in separate clusters, given that they both involve a one-to-one mapping. The explanation for this is that changes at first position sites generally involve transformations among amino acids with similar properties, whereas those at second positions do not. Thus, from a functional perspective, first position changes have more freedom to vary than do second position sites, even though both have similar degrees of freedom from a purely combinatorial perspective.

We feel that this preliminary survey of the use of covariate models shows promise for the systematic assessment of the role of chemical and structural features in improving our understanding of molecular evolution and phylogenetic estimation from sequence data. Although maximum likelihood models that incorporate rate classes have previously been used to accommodate rate variation among genes[23], the approach taken here exploits their use to account for rate heterogeneity among classes of sites *within* genes. We have not addressed the issue of goodness-of-fit here, but we note that the fact that different classes of sites have significantly different rates might reflect different underlying substitution models. To explore this possibility we estimated substitution models for several genes and found that indeed, different models existed for each of the five classes of sites (Table 3). These results imply that molecular data require approaches that incorporate not only

different rate classes, but also different substitution models for the different classes of sites.

Table 3. Substitution models for the five classes are shown below for three representative genes. The best fitting GTR model with Γ rate variation estimated from the data was used to obtain branch lengths for Figure 1, and the substitution models below were calculated from implied changes on that tree. Model classes are: Gp 1: 1_2, 1_4; Gp 2: 1_6; Gp 3: 2_2, 2_4, 2_6; Gp 4: 3_2; and Gp 5: 3_4, 3_6.

	CO 3					Cyt b					ND 1			
Gp1	A	C	G	T	Gp1	A	C	G	T	Gp1	A	C	G	T
A	-	.057	.172	.059	A	-	.089	.212	.132	A	-	.092	.238	.146
C	.050	-	.065	.021	C	.029	-	.111	.020	C	.067	-	.100	.049
G	.765	.600	-	.609	G	.783	.444	-	.595	G	.626	.333	-	.815
T	.090	.052	.103	-	T	.200	.055	.103	-	T	.166	.071	.212	-
Gp2	A	C	G	T	Gp2	A	C	G	T	Gp2	A	C	G	T
A	-	.028	.045	.031	A	-	.083	.109	.080	A	-	.040	.041	.037
C	.132	-	.258	.185	C	.132	-	.232	.155	C	.123	-	.200	.215
G	.052	.000	-	.261	G	.060	.056	-	.143	G	.107	.083	-	.074
T	.097	.017	.172	-	T	.081	.054	.448	-	T	.179	.042	.333	-
Gp3	A	C	G	T	Gp3	A	C	G	T	Gp3	A	C	G	T
A	-	.016	.020	.042	A	-	.015	.031	.042	A	-	.023	.043	.062
C	.036	-	.355	.058	C	.026	-	.162	.065	C	.038	-	.311	.078
G	.061	.333	-	.130	G	.051	.389	-	.095	G	.053	.361	-	.074
T	.200	.156	.207	-	T	.222	.349	.293	-	T	.172	.303	.212	-
Gp4	A	C	G	T	Gp4	A	C	G	T	Gp4	A	C	G	T
A	-	.054	.279	.059	A	-	.050	.215	.034	A	-	0.053	.278	.066
C	.064	-	.032	.450	C	.175	-	.091	.484	C	.177	-	.167	.377
G	.061	.000	-	.000	G	.030	.037	-	.024	G	.097	.028	-	.000
T	.124	.454	.138	-	T	.133	.363	.034	-	T	.199	.449	.121	-
Gp5	A	C	G	T	Gp5	A	C	G	T	Gp5	A	C	G	T
A	-	.845	.485	.809	A	-	.763	.434	.712	A	-	.793	.400	.689
C	.719	-	.290	.287	C	.637	-	.404	.275	C	.595	-	.222	.282
G	.061	.067	-	.000	G	.077	.074	-	.143	G	.117	.194	-	.037
T	.490	.320	.379	-	T	.363	.179	.121	-	T	.285	.134	.121	-

It should be noted that all of the models used in this study were implemented for rooted tree topologies and a molecular clock. We are currently extending and generalizing these methods to accommodate more relaxed assumptions and more complex biologically-based covariate structures. We have also implemented a covariate approach for gamma rate variation[24] in which the mean rate at a site is parameterized as in (1). It should not escape notice that all covariates used in this

study were scored as the modal value in the alignment. This is clearly a first approximation and methods that allow time-varying covariate assignments for individual sequences need to be developed.

Acknowledgments

This work was supported in part by National Science Foundation Grant DBI-9504393 (to ST), National Science Foundation Grant DEB-9707145 (to GJPN), and by a National Science Foundation Postdoctoral Fellowship in Biological Informatics DBI-9974207 (to DCA).

References

1. J.S. Farris, in *Advances in Cladistics*, Vol. 2 "The logical basis of phylogenetic analysis" Eds. N.A. Platnick and V.A. Funk (Columbia University Press, New York, 1983).
2. J. Felsenstein, "Cases in which parsimony or compatability methods will be positively misleading" *Syst. Zool.* **27**, 401 (1978).
3. M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences" *J. Mol. Evol.* **6**, 111 (1980).
4. J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach" *J. Mol. Biol.* **17**, 368 (1981).
5. J. L. Thorne H. Kishino and J. Felsenstein, "Inching towards reality: An improved likelihood model of sequence evolution" *J. Mol. Evol.* **34**, 3 (1992).
6. M. Hasegawa H. Kishino and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA" *J. Mol. Evol.* **22**, 160 (1985).
7. S. Tavaré, in *Lectures in Mathematics in the Life Sciences*, Vol. 17 "Some probabilistic and statistical problems in the analysis of DNA sequences" Ed. R.M. Muira (American Mathematical Society, Providence, 1986).
8. J. Neyman, in *Statistical Decision Theory and Related Topics*, "Molecular studies of evolution: A source of novel statistical problems" Eds. S.S. Gupta and J. Tackel (Academic Press, New York, 1971).
9. Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites" *Mol. Biol. Evol.* **10,** 1396 (1993).
10. Z. Yang N. Goldman and A. Friday, "Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation" *Mol. Biol. Evol.***11**, 316 (1994).
11. J. Sullivan K.E. Holsinger and C. Simon, "Among-site rate variation and phylogenetic analysis of 12s rRNA data in sigmodontine rodents.

Mol. Biol. Evol. **12**, 988 (1995).
12. S. Tavaré and Y. Feng, in *Proceedings of Phylogeny Workshop,* Technical Report 95-48 "Reconstructing phylogenetic trees when sites are dependent" (DIMACS, Rutgers University, 1995).
13. Z. Yang, "A space-time process model for the evolution of DNA sequences" *Genetics* **139**, 993 (1995).
14. J. Felsenstein and G.A. Churchill, "A hidden markov model approach to variation among sites in rate of evolution" *Mol. Biol. Evol.* **13**, 93 (1996).
15. Z. Yang, "Among-site rate variation and its impact on phylogenetic analysis" TREE **11,** 367 (1996).
16. P. Liò and N. Goldman, "Models of molecular evolution and phylogeny" *Gen. Res.* **8**, 1233 (1998).
17. N. Goldman J.L. Thorne and D.T. Jones, "Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses" *J. Mol. Biol.* **263,** 196 (1996).
18. N. Goldman J.L. Thorne and D.T. Jones, "Assessing the impact of secondary structure and solvent accessibility on protein evolution" *Genetics* **149**, 445 (1998).
19. P. Liò N. Goldman J. L. Thorne and D. T. Jones, "PASSML: Combining evolutionary inference and protein secondary structure prediction" *Bioinformatics* **14,** 726 (1998).
20. G. J. P. Naylor and W. M. Brown, "Structural biology and phylogenetic estimation" *Nature* **388,** 527 (1997).
21. A. J. Dobson, *An Introduction to Generalized Linear Models* (Chapman and Hall, London, 1990).
22. N. Goldman, "Statistical tests of models of DNA substitution" *J. Mol. Evol.* **36,** 182 (1993).
23. S. V. Muse and B. S. Gaut. "Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test" *Genetics* **146,** 393 (1997).
24. Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Biol.* **39**, 306 (1994).