Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics is Clearly Beneficial

Quang Le Si, Olivier Gascuel

Syst. Biol. 59(3):277–287, 2010

DOI:10.1093/sysbio/syq002
Advance Access publication on March 10, 2010

Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial

SI QUANG LE[1,2] AND OLIVIER GASCUEL[1,*]

[1]*Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS–Université Montpellier II, 161 rue Ada, 34392 Montpellier Cedex 5, France; and*
[2]*Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK;*
**Correspondence to be sent to: LIRMM, 161 rue Ada, 34392, Montpellier Cedex 5, France; E-mail: gascuel@lirmm.fr.*

Received 23 January 2009; reviews returned 3 April 2009; accepted 10 December 2009
Associate Editor: Mark Holder

Abstract.—Amino acid substitution models are essential to most methods to infer phylogenies from protein data. These models represent the ways in which proteins evolve and substitutions accumulate along the course of time. It is widely accepted that the substitution processes vary depending on the structural configuration of the protein residues. However, this information is very rarely used in phylogenetic studies, though the 3-dimensional structure of dozens of thousands of proteins has been elucidated. Here, we reinvestigate the question in order to fill this gap. We use an improved estimation methodology and a very large database comprising 1471 nonredundant globular protein alignments with structural annotations to estimate new amino acid substitution models accounting for the secondary structure and solvent accessibility of the residues. These models incorporate a confidence coefficient that is estimated from the data and reflects the reliability and usefulness of structural annotations in the analyzed sequences. Our results with 300 independent test alignments show an impressive likelihood gain compared with standard models such as JTT or WAG. Moreover, the use of these models induces significant topological changes in the inferred trees, which should be of primary interest to phylogeneticists. Our data, models, and software are available for download from http://atgc.lirmm.fr/phyml-structure/. [Amino-acid substitutions; maximum likelihood; partition models; replacement rate matrices; structural annotation of proteins; topological impact.]

It is widely recognized that evolutionary divergence of protein structures occurs much less rapidly than divergence of protein sequences (e.g., Chothia and Lesk 1986). Structural constraints act on the protein sites, which in turn impact amino acid substitutions. Notably, secondary structure and solvent accessibility have been shown to have a strong influence on amino acid replacement processes (e.g., Koshi and Goldstein 1995; Thorne et al. 1996; Goldman et al. 1998). For example, buried sites are mostly hydrophobic and tend to remain hydrophobic along the course of time, meaning that substitutions in the buried parts of the proteins most often occur between hydrophobic amino acids. This type of information is only partly captured by standard models of amino acid substitution, which are based on the use of substitution rate matrices such as PAM (Dayhoff et al. 1972), JTT (Jones et al. 1992), WAG (Whelan and Goldman 2001), or LG (Le and Gascuel 2008). These matrices contain replacement rates between every amino acid pair. These rates were estimated from very large databases, without differentiating among the various site structural configurations, thus resulting in an average model for average sites and proteins. Site-dependent models were thus estimated for several structural categories based on solvent accessibility and secondary structure, showing clear distinctions among categories, most notably between exposed and buried sites (Goldman et al. 1998; Holmes and Rubin 2002). For example, our recent results (Le, Lartillot, et al. 2008) indicate that highly exposed sites evolve 3–4 times faster than buried sites, on average.

These studies on the variability of evolutionary processes across structural categories greatly improved our understanding of protein evolution. However, they are rarely used today in protein phylogenetics, despite the fact that substitution models are essential in most approaches to phylogenetic inference (e.g., in the estimation of evolutionary distances or the calculation of tree likelihoods; see textbooks: Felsenstein 2003; Bryant et al. 2005; Yang 2006). There are several reasons that these structure-based models have not been widely used. First, only a few 3-dimensional (3D) structures were available in the 1990s. Thus, these models do not use any information on the actual structure of the studied proteins. Each site is assumed to belong to each possible structural category, and the total site likelihood is the weighted average of all possibilities using a mixture approach (e.g., Le, Lartillot, et al. 2008) or a more sophisticated hidden Markov model representing the dependence of structural states along the sequences (Thorne et al. 1996; Goldman et al. 1998). Second, when the 3D structure(s) is known for one (or a few) of the studied proteins, the structural annotations for the whole protein set may not be reliable. These annotations are inferred by homology from known structures, but structure is not fully conserved along evolution. Furthermore, structural annotation procedures rely on numerical thresholds (e.g., to classify buried/exposed sites) and definitions (e.g., using psi and phi angles for the secondary structures), which are somewhat arbitrary and induce uncertainties in the assignments of sites to structural categories. Third, previously implemented structure-based models did not incorporate any gamma distribution of rates across sites (Yang 1993), which is now an important component in most sequence evolution models. All these likely explain the low

impact of these approaches in phylogenetics. Another factor might be the complexity of the calculations using software and computers of that time.

Today, a number of 3D protein structures are available. The protein data bank (PDB; Berman et al. 2000) contains more than 50,000 proteins. Moreover, using sequence homology, we are able to infer likely structures for a number of homologous proteins. The homology-derived structures of proteins (HSSP) data bank (Schneider et al. 1997) contains more than 2 million sequences, that is, ∼35% of UniProt. In this paper, we claim that this information should be used in phylogenetics to build more accurate trees. We use a very large database (extracted from HSSP) of globular protein alignments to estimate new amino acid replacement matrices for various site structural configurations based on solvent accessibility and secondary structure. We use these matrices in simple substitution models that account for 1) site structural configurations, 2) the fact that structural annotations may not be fully reliable, and 3) the variability of rates across sites. We introduce fast methods and programs to estimate our models and infer phylogenetic trees using these models. The performance of these models and programs is assessed using 300 independent test alignments. We show that using our structural models greatly augments the likelihood values of inferred trees. Moreover, the tree topology is often different in comparison with standard models (e.g., JTT, WAG, or LG). In the following, we describe our data sets, then the models and their estimation and implementation, and finally the experiments and results. Supplementary material is provided in an online Appendix available from: http://www.sysbio.oxfordjournals.org. Our data, replacement matrices, detailed results for all test alignments, and software are downloadable from http://atgc.lirmm.fr/phyml-structure/.

Data Sets

To estimate our models, we used a large database of multiple globular protein alignments. It was extracted from HSSP, which comprises ∼50,000 alignments of protein families. Each alignment is obtained by aligning a (seed) protein with known 3D structure in the PDB to all its sequence homologues in Uniprot. The secondary structure and solvent accessibility of the seed protein are calculated using DSSP (Kabsch and Sander 1983) and are likely to be representative of the structure of all homologues in the alignment.

HSSP is highly redundant and contains a number of gaps. We thus performed an intensive cleaning of HSSP to extract independent alignments and, within each of the alignments, to select sequences and sites corresponding to well-aligned, nongapped regions. Moreover, we discarded membrane proteins (based on their presence in the membrane protein data bank Raman et al. 2006), as they show quite specific patterns of amino acid replacement (Jones et al. 1994). We obtained this way a database comprising 1771 nonredundant alignments, with an average of ∼56 sequences and ∼254 sites per alignment, ∼27 million amino acids in total and very few gaps (<0.1%). Moreover, each of these alignments contains the seed protein of the original HSSP alignment and thus (relatively) reliable annotations on the solvent accessibility and secondary structure of each of the sites. Using these annotations, we classified each site as extended (E), alpha-helix (H), or other (S, T, B, G, I, "." or "?"). Based on their relative solvent accessibility (Shrake and Rupley 1973), we also classified sites as either buried or exposed; we used a 10% threshold of relative accessibility, as in Goldman et al. (1998) and several other studies. This threshold induces nearly equally weighted buried and exposed categories. Finally, we randomly selected 300 alignments for model comparison, whereas we used the remaining 1471 to train our models.

We already used this cleaned alignment database to estimate mixture models using amino acid replacement matrices (Le, Lartillot, et al. 2008) and profiles (Le, Gascuel, et al. 2008). Readers are referred to these papers for details on the cleaning procedure. Note that the purpose of these studies was different of this one. Indeed, in mixture approaches, the protein structure is supposed to be unknown when inferring trees, whereas here, we shall use the available structural annotations to improve the accuracy of substitution models and consequently the reliability of tree reconstructions.

Models

Our models involve amino acid replacement matrices that differ depending on the structural properties of the sites (e.g., exposed/buried). We combine these matrices accounting for both the structural annotations and the fact that this information may not be fully reliable. We first describe the replacement matrices and the way we estimated them from the training data and then various models to combine these matrices.

Amino Acid Replacement Matrices

We estimated replacement rate matrices for several site partitions based on structural annotations. All these matrices comply with the general time-reversible model, which was first proposed for DNA sequences (Lanave et al. 1984; Tavaré 1986), and then applied to proteins through a number of empirical (as opposed to mechanistic) models such as PAM, JTT, WAG, or LG (see textbooks: e.g., Yang 2006; see also online Appendix 1). One matrix was estimated per site category, and 3 partitions were analyzed:

- EX is a 2-category partition corresponding to exposed/buried sites.
- EHO is a 3-category partition corresponding to extended/alpha-helix/other sites.
- EX_EHO is a 6-category partition that combines both previous criteria; sites are classified as: exposed &

extended, buried & extended, exposed & alpha-helix, buried & alpha-helix, exposed & other, or buried & other.

To estimate the corresponding $(2 + 3 + 6 = 11)$ replacement matrices, we applied the procedure detailed in Le and Gascuel (2008) and Le, Lartillot, et al. (2008) to each site category. This procedure is based on the maximum-likelihood (ML) principle and uses XRate (Klosterman et al. 2006) and a new version of PhyML (Guindon and Gascuel 2003), which implements our multiple-matrix models (see below). A gamma distribution of rates across sites (Yang 1993) is incorporated in both the tree inference and the matrix estimation. PhyML is used to estimate the phylogenies of each of the 1471 training alignments using all the sites. In the next step, we treat the phylogenies as fixed. We preprocess the data to account for rate heterogeneity and then use XRate to estimate the replacement matrix of each structural category separately. Preprocessing involves assigning each site to the most likely rate category and rescaling the phylogeny associated to this site accordingly (Le and Gascuel 2008). This procedure is iterated: When a first set of matrices corresponding to the site partition has been estimated, the phylogenies are reinferred using this new model, the replacement matrices are reestimated based on the new phylogenies, and so on until convergence. All together, this estimation procedure improves considerably over standard counting approaches that were used to estimate PAM and JTT matrices as well as all matrices studied by Thorne et al. (1996) and Goldman et al. (1998). It also outperforms the ML procedure of Whelan and Goldman (2001), thanks to the use of rates across sites in the matrix estimation (Le and Gascuel 2008).

All these matrices are available from our Web site and are compared with the (average) LG matrix. EX and EHO matrices have already been analyzed in a mixture context (Le, Lartillot, et al. 2008). One of the main observations (see also Goldman et al. 1998) is that the global average substitution rate is much higher for exposed than for buried sites (ratio = 2.45). This is an expected result as buried sites are subject to strong structural constraints, whereas exposed sites are often in the flexible and variable parts of the proteins. On the other hand, the global average rate does not change much among secondary structure categories, some sites within each category being with strong structural constraints and the others not so. Moreover, each matrix defines an equilibrium distribution of amino acids. As expected, amino acid distributions are quite different among site categories. For example, the buried category mostly contains hydrophobic amino acids, whereas the helix category contains a large proportion of alanines but very few prolines (commonly called "helix-breakers"). Finally, the rates in the matrices (once normalized, i.e., divided by the global average rate) also differ among categories but to a lesser extent than amino acid compositions. Again, the exposed/buried partition induces higher contrasts among site categories than the extended/helix/other partition and thus is expected to bring more information on the substitution processes.

The secondary structure does not impose homogeneous structural constraints within each category. For example, the "other" category contains some deeply buried residues situated in the core of the proteins as well as residues in the flexible regions. In the same way, it is well known (e.g., Branden and Tooze 1999) that a number of alpha-helices are amphiphilic, with a side on the protein surface and the other side in the protein core with strong structural constraints. The 6-category partition EX_EHO makes the secondary structure categorization effective. Indeed, there is a high contrast among EX_EHO categories. The slowest category is "buried & other" (global rate = 0.535), the fastest "exposed & helix" (global rate = 2.064), and the 2 buried/exposed helix categories clearly differ (rate ratio $\sim$ 2.8). Matrix entries are also more distant from average LG matrix than are the entries in uncombined EX and EHO matrices. We thus expect good performance of this crossed site partition in modeling amino acid substitutions.

Models to Account for Site Structural Annotations

Amino acid replacement matrices are used to compute the likelihood of the data, given a tree with branch lengths, and the ML principle involves maximizing this likelihood to search for the optimal tree (in most cases, one must be satisfied with near optimal trees for computing time reasons). Let D be the analyzed alignment and D_i the data at site i. When all sites are assumed to belong to a single average category with replacement matrix $\mathbf{Q}$ (e.g., JTT or WAG), the tree likelihood is expressed using the independence assumption as

$$L(T, \mathbf{Q}, \Theta|D) = \prod_i L(T, \mathbf{Q}, \Theta|D_i), \qquad (1)$$

where the product runs over all the sites, T is the tree (including branch lengths), Θ the set of additional model parameters (typically the gamma distribution parameter), and $L(T, \mathbf{Q}, \Theta|D_i)$ the likelihood of the data at site i.

When the sites are classified into several known categories, the standard approach, called "partition model," uses this knowledge in a natural way. The partition model is commonly used to account for different site categories depending on codon positions or genes in concatenated alignments (e.g., Rannala and Yang 2008). Here, we consider structural categories, but the mathematical formulation remains the same. Let C denote any given category and $\mathbf{Q} = \{\mathbf{Q}_C\}$ be the set of replacement matrices corresponding to the studied categories; for example, $\mathbf{Q}$ may contain the exposed and buried matrices for the EX model. Moreover, let C_i be the category of site i (e.g., "buried" or "exposed") and $\{C_i\}$ the set of assignments for each site. Then, the tree likelihood is equal to

$$L(T, \mathbf{Q}, \{C_i\}, \Theta|D.) = \prod_i L(T, \mathbf{Q}_{C_i}, \Theta|D_i). \qquad (2)$$

In other words, for each site, we simply use its associated replacement matrix. No free parameter is added in comparison with single matrix–based Equation (1), and the computing time remains basically unchanged (at least when no extra parameters are added in Θ, see below). Even though this approach is simple and natural, we have not been able to find any paper relating its performance with structural categories. Part of the explanation derives from the fact that structural annotations may be poor or noninformative for some alignments. In these cases, some sites are analyzed with inappropriate replacement matrices and the results may be worse than using a single average matrix as in Equation (1).

We refined the partition model to account for this uncertainty in structural category assignment. A confidence coefficient, denoted as χ, measures the reliability of structural annotations. χ represents the fraction of sites with reliable annotations in the alignment. When χ is equal to 1, all structural annotations are used, just as in Equation (2). When χ is less than 1, some annotations are not reliable or not useful in a phylogenetic context. Given an alignment, all the sites are analyzed using a single χ value that is a free parameter estimated by ML from the data. Moreover, we assume that no additional knowledge is available to detect the sites with poor annotations (this hypothesis will be further discussed). For each site, we thus envisage 2 possibilities (reliable/not reliable), and the site likelihood is the weighted average of both corresponding likelihoods, following the law of total probability. We studied 2 approaches to deal with nonreliable annotations. In the simplest one, we use a single standard average replacement matrix denoted as $\mathbf{Q}$ (here LG), and the tree likelihood is equal to

$$L(T,\mathbf{Q},\{C_i\},\Theta|D) = \prod_i [\chi L(T,\mathbf{Q}_{C_i},\Theta|D_i) + (1-\chi)L(T,\mathbf{Q},\Theta|D_i)]. \quad (3)$$

In other words, we make a weighted sum of Equations (1) and (2), assuming that $\mathbf{Q}$ is well suited to model the substitutions of sites with poor structural annotations. When χ is 0, structural annotations are not used and the sites are analyzed with $\mathbf{Q}$ only.

In the second approach, we replace the single $\mathbf{Q}$ matrix in Equation (3) by a mixture of matrices (models) corresponding to all categories in the site partition. Mixtures are the standard approach when we assume that the heterogeneity of substitution processes derives from various site categories corresponding to different evolutionary constraints but ignore the classification of the sites into these categories (Gascuel and Guindon 2007). Mixtures have been proposed as part of several substitution models (e.g., Pagel and Meade 2005), including with structural site categories where they show a clear improvement over single-matrix models (Le, Lartillot, et al. 2008). Let P_C be the probability for any given site to belong to category C. Using (again) the law of total probability, the tree likelihood is given by

$$L(T,\mathbf{Q},\Theta|D) = \prod_i \left[\sum_C P_C L(T,\mathbf{Q}_C,\Theta|D_i)\right], \quad (4)$$

where the total site likelihood is the weighted sum of the site likelihoods for all possible categories. This approach performs well when the site partition is relevant for the analyzed data set. In this case, the likelihood of each site with the proper (unknown) category is much larger than the likelihood with the average model, and the weighted sum in Equation (4) tends to be larger than the single term in Equation (1). As such a case is encountered with most data sets (Le, Lartillot, et al. 2008), we combine Equation (2) with Equation (4) to obtain

$$L(T,\mathbf{Q},\{C_i\},\Theta|D) = \prod_i \left[\chi L(T,\mathbf{Q}_{C_i},\Theta|D_i) + (1-\chi)\sum_C P_C L(T,\mathbf{Q}_C,\Theta|D_i)\right]. \quad (5)$$

When χ is 1 (annotations are fully reliable), this becomes identical to the partition model (2), whereas when χ is 0 (annotations are just random), we use the mixture (4). With mixtures, P_C represents the probability of category C. Structural category proportions are not the same for all proteins, some only have alpha-helices and some beta sheets only, whereas some have both (e.g., Branden and Tooze 1999). P_C proportions thus are free parameters to be estimated for each analyzed data set. With our compound model (5), the interpretation of P_C proportions is slightly different, as they represent the category proportions among sites with nonreliable annotations. Again, these proportions are free parameters estimated from the data for each alignment separately.

As mentioned above, Θ contains additional model parameters. In our case, Θ contains the shape parameter of the gamma distribution, which defines the variability of rates across sites (Yang 1993). This is used within every structural category, including in Equation (1) where all sites are analyzed with a unique replacement matrix. As usual, the gamma distribution parameter is free and estimated from the analyzed data. More sophisticated models are conceivable, with different gamma distributions among structural categories, but preliminary experiments (not shown) did not reveal any significant improvement, and a single gamma distribution of rates is used in the following.

When using a single gamma distribution, the partition model (2) does not add any free parameter to standard model (1). Our first model (3) to account for uncertainty adds a single free parameter (χ) to be estimated from the analyzed data. Let c be the number of site categories; the mixture approach (4) adds $c-1$ free parameters (category proportions) compared with the standard model, whereas our second approach to

accommodate for uncertainty (5) adds c free parameters (χ and category proportions). Because of the nested relationships of these models, the likelihood value using the confidence-based model (3) is guaranteed to be higher than (or equal to) the likelihood values of the standard (1) and partition (2) models. In the same way, the confidence-based model (5) is guaranteed to obtain better (or equal) likelihood values than (to) the partition (2) and mixture (4) models, at least when the same tree topology is analyzed. However, these gains in likelihood could be counterbalanced by the larger number of free parameters, necessitating a penalized likelihood criterion such as Akaike information criterion (AIC; Akaike 1974).

These models involve variable computing times, which are basically proportional to the number of categories used to analyze each of the sites (Bryant et al. 2005). The partition model (2) thus requires nearly the same time as the standard model (1), whereas the confidence-based model (3) is 2 times slower than the standard model, and the mixture (4) and confidence-based (5) models are c times slower. The same ratios apply to memory consumption. However, this analysis is based on tree likelihood calculation only, whereas model parameter estimation also induces significant computational cost increase compared with the standard model (see illustrative examples below).

All these models are implemented in a new version of PhyML (available from http://www.atgc-montpellier.fr/phyml/structure/). An initial ML tree is first inferred with LG in the usual manner, and then the chosen model is used to refine this first tree, in terms of both topology and branch lengths, with the model parameters (χ value, category proportions, and gamma shape parameter) being adjusted along the way.

Results and Discussion

In this section, we present the results obtained with the 300 HSSP test alignments: 1) we show the likelihood gains provided by our models compared with standard approaches, 2) we analyze the values of our confidence coefficient (χ) regarding the various structural categories, and 3) we discuss the impact on the topologies of inferred trees.

Likelihood Gains

We used the 300 HSSP test alignments to compare standard single-matrix models (JTT, WAG, and LG) to our new models based on EX (solvent accessibility), EHO (secondary structure), and EX_EHO (both solvent accessibility and secondary structure) site partitions. For each of the multimatrix models, we used the 4 combinations detailed in the previous section:

- MIX is the mixture approach, expressed by Equation (4), where the site categories are supposed to be unknown.
- PART is the partition approach, expressed by Equation (2), where one uses the known HSSP site categories.
- CONF/LG is our first confidence-based approach, expressed by Equation (3), where sites with unreliable annotations are analyzed using LG (HSSP version, see below).
- CONF/MIX is our second confidence-based approach, expressed by Equation (5), where sites with unreliable annotations are analyzed using a mixture.

These abbreviations are combined with the name of the studied site partition; for example, EX_PART is the partition model with exposed and buried categories. All models were run using the new version of PhyML (http://atgc.lirmm.fr/phyml-structure/) with 4 gamma rate categories ($\Gamma 4$), BioNJ (Gascuel 1997) starting tree, and subtree pruning and regrafting (SPR)–based tree topology search (Hordijk and Gascuel 2005; Guindon et al. 2010).

For all models, we measured the AIC (Akaike 1974) on each of the test alignments:

$$\mathrm{AIC}(M, D) = 2\#\ \text{parameters}(M) - 2\ln L(M, T|D),$$

where $\#\ \text{parameters}(M)$ is the number of free parameters of model M and $\ln L(M, T|D)$ is the log-likelihood value of alignment D given model M and inferred tree T. This criterion has to be minimized; best scores are given to models with low number of free parameters and high likelihood value. We also computed the average AIC per site of model M for all test alignments, which is simply

$$\mathrm{AIC/site}(M) = \sum_D \mathrm{AIC}(M, D) \Big/ \sum_D s, \qquad (6)$$

where s is the number of sites in D. All models were compared with LG using criterion (6) by computing the difference $\mathrm{AIC/site(LG)} - \mathrm{AIC/site}(M)$, which is positive when M has a better fit than LG. To complete this global average result, we also compared all models with LG by counting the number of alignments where $\mathrm{AIC}(M, D)$ is better/worse than $\mathrm{AIC(LG}, D)$. Moreover, to assess the statistical significance of the observed difference between M and LG for any given alignment, we used a Kishino–Hasegawa (KH; 1989) test with $P < 0.01$. As the number of free parameters may differ between M and LG, we used AIC-corrected log-likelihood values instead of simple log-likelihood values (see Shimodaira [1997] for explanations and justifications of this procedure).

Average AIC results using criterion (6) are displayed in Figure 1, whereas Figure 2 provides the number of alignments where each model is (significantly) better/worse than LG. Main observations and conclusions are as follows.

The improvement from JTT to WAG and then LG is quite significant. The AIC gain per site of LG compared with JTT is 0.71, meaning that with 300 sites (standard

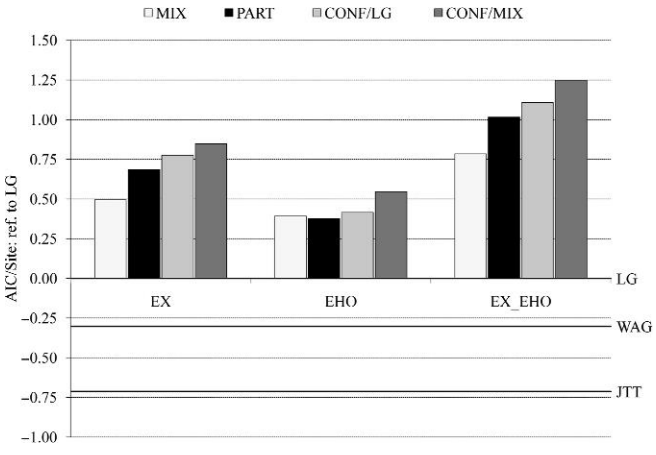

FIGURE 1. AIC gain per site compared with LG (and WAG and JTT). All models are compared with LG (Le and Gascuel 2008) using the average AIC/site criterion (Equation (6)), estimated with the 300 HSSP test alignments. Both WAG (Whelan and Goldman 2001) and JTT (Jones et al. 1992) are worse than LG, with a negative difference of −0.30 and −0.71, respectively. All multiple-matrix models are better than LG with positive AIC/site difference. MIX = mixture model; PART = partition model; CONF/LG = confidence-based model using LG for nonreliable sites; CONF/MIX = confidence-based model using mixtures for nonreliable sites; EX = solvent accessibility–based site partition; EHO = secondary structure–based site partition; EX_EHO = combination of EX and EHO site partitions.

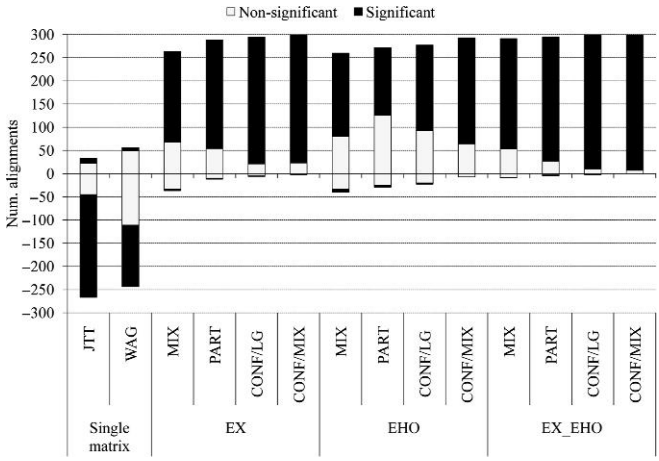

FIGURE 2. Number of alignments with better/worse likelihood values than LG. Number of alignments (among the 300 HSSP test alignments) where each model provides a better (positive side) and a worse (negative side) likelihood value than LG. The black bars correspond to the numbers of significant differences using the KH test on AIC values with $P < 0.01$. MIX = mixture model; PART = partition model; CONF/LG = confidence-based model using LG for nonreliable sites; CONF/MIX = confidence-based model using mixtures for nonreliable sites; EX = solvent accessibility–based site partition; EHO = secondary structure–based site partition; EX_EHO = combination of EX and EHO site partitions.

length of protein alignments) the gain is as high as ∼200 AIC points, that is, ∼100 log-likelihood points. Moreover, LG is often significantly better than JTT and WAG, but rarely worse.

We also tested a new version of LG, obtained from HSSP instead of Pfam (Bateman et al. 2002), but using the same estimation procedure as for original LG (Le and Gascuel 2008). The aim was to check that the improvements provided by our new models do not derive from HSSP training alignments, but from the models themselves. Indeed, this HSSP version is only slightly better than original LG with the 300 HSSP test alignments (AIC gain per site ∼0.02). Both LG versions are actually very close (the correlation of log entries is equal to 0.98), which shows that LG matrices are relatively stable when estimated from different general databases.

All our multiple-matrix models improve a lot compared with single-matrix models, including LG. For example, the gain between our best model (EX_EHO_CONF/MIX; both solvent accessibility and secondary structure site partitions, nonreliable sites are analyzed using a mixture) and JTT is 1.96 AIC points per site, meaning that with 300 sites the gain will be close to 600 AIC points. This gain is of the same magnitude as that provided by the use of a gamma distribution of rates across sites. For example, the gain between LG with and without gamma distribution equals 2.90 with HSSP test alignments, and similar values are found for JTT and WAG. We thus believe that multiple-matrix models should become standard in the near future, just as the gamma distribution is standard today and used in most phylogenetic analyses.

However, using a gamma distribution of rates across sites is still required with multiple-matrix models, though they involve site categories with variable global rates. For example, EX_EHO_CONF/MIX without gamma distribution is not any better than LG + Γ4 (AIC gain per site < 0.01). Another example is PASSML (Lio et al. 1998), which is clearly worse than JTT + Γ4 due to its lack of gamma distribution (AIC difference per site ∼0.5; see online Appendix 2 for comparisons of PASSML with our models). These results show that sites within structural categories are not subject to the same constraints and do not evolve at the same rate. However, part of the rate heterogeneity is accounted for by the structural categories and their variable global rates; the estimated value of the gamma distribution parameter tends to be higher (i.e., the rate variability is lower) with our multimatrix models (∼0.9 on average) than with single-matrix models (∼0.65 on average).

Overall, the site partition based on solvent accessibility (EX) has higher gains than the secondary structure (EHO) partition, whereas the combination of both (EX_EHO) is clearly best. As discussed earlier, this illustrates that secondary structures are not homogeneous, with highly conserved, hydrophobic sites, and other sites situated in the exposed and variable regions. However, EHO models clearly improve over single matrices as they still account for a part of site heterogeneity.

The partition approach (PART) performs better than mixtures (MIX), except with EHO where both obtain similar average performance (see online Appendix 3 for significance results). This is explained by the fact that secondary structure annotation is somewhat arbitrary

and not fully reliable; for example, the extremities of alpha-helices are typically difficult to define. A quantitative measurement of this uncertainty is provided below. With reliable annotations, PART benefits this information and outperforms MIX.

Our confidence-based models (CONF/LG, which uses LG for unreliable sites, and CONF/MIX using a mixture instead of LG) clearly improve over both PART and MIX. Due to nested relationships (see above), CONF/MIX always has higher likelihood values than MIX and PART. Figure 3 shows that the AIC differences are significant in a large number of cases: CONF/MIX is significantly better than MIX with >200 alignments and significantly better than PART with 55–136 alignments. The AIC differences between CONF/LG, MIX, and PART are positive in most cases, but less often significant (see online Appendix 3). All these indicate that some alignments (some sites within some alignments) are poorly annotated from a phylogenetic perspective and that our simple confidence-based models offer efficient solutions to deal with this uncertainty. CONF/MIX has better AIC values than CONF/LG because unreliable sites are analyzed with a mixture that outperforms LG in most cases. However, CONF/LG is clearly of interest with EX_EHO site partition where it is about 3 times faster than CONF/MIX with similar likelihood performance. Actually, EX_EHO_CONF/LG is even faster than EX_CONF/MIX and EHO_CONF/MIX, although providing much better likelihood values. Its interest would be even higher with a larger number of categories, for example, combining 3 solvent accessibility categories (Le, Lartillot, et al. 2008) or 4 secondary structure ones (Goldman et al. 1998), with a computing time still nearly equals to twice that required by standard single-matrix models.

The computing time required by our best (and slowest) model, EX_EHO_CONF/MIX, is ∼9 times the computing time required by LG with the same program options (Γ4, SPR, etc.). The average running time with the test alignments (∼56 sequences and ∼254 sites on average) equals 40 and 4.5 min per alignment using a standard PC (Intel(R) Xeon(R) 1.86 GHz) for EX_EHO_CONF/MIX and LG, respectively. The slowest alignment (1o98, 92 taxa, and 352 sites) requires 270 and 30 min for EX_EHO_CONF/MIX and LG, respectively, but a number of alignments are dealt with much faster. Other models require running times in between these 2 extreme models. The observed ratio of ∼9 between EX_EHO_CONF/MIX and LG is significantly higher than 6, as expected based on tree-likelihood calculation (see above). The difference is basically explained by the estimation of model parameters (i.e., χ, 5 category proportions, and the gamma distribution parameter) that are highly correlated.

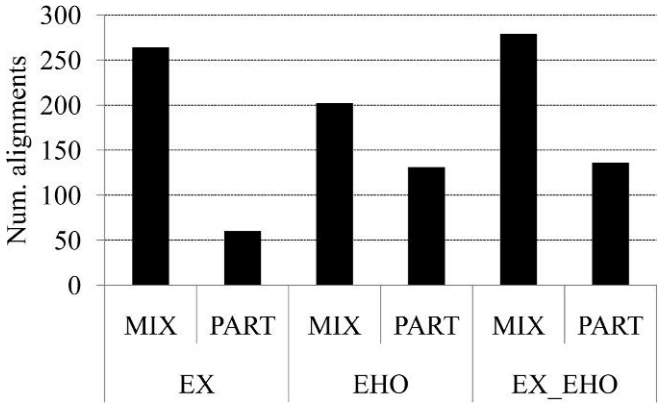

FIGURE 3. Comparison of CONF/MIX, PART, and MIX. Number of alignments where CONF/MIX (confidence-based model using mixtures for nonreliable sites) is significantly better than MIX (mixture model) and PART (partition model), using the KH test on AIC values with $P < 0.01$. EX = solvent accessibility–based site partition; EHO = secondary structure–based site partition; EX_EHO = combination of EX and EHO site partitions.

All together, these results strongly support the use in phylogenetics of solvent accessibility and secondary structure annotations. Using this information in the most appropriate way (EX_EHO_CONF/LG and EX_EHO_CONF/MIX models) provides very high AIC gains compared with the best known single matrices (LG, WAG, or JTT), and these gains are significant for most alignments (∼290/300 in our experiments). Moreover, computing times are still acceptable for most data sets using standard computers.

Reliability and Usefulness of Structural Annotations

Our confidence coefficient χ in Equations (3) and (5) is estimated separately for each of the alignments and reflects the reliability of the annotations in the given alignment. Actually, χ combines several factors: 1) precision of the annotation that may be somewhat arbitrary, for example, at both alpha-helix extremities; 2) reliability of the annotation, as some errors may be induced by the complex computations to infer them from crystallographic or nuclear magnetic resonance data (3D structure elucidation, multiple alignment, DSSP, etc.; 3) conservation of the annotation, as the structures of proteins in the multiple alignment may have evolved and become slightly different to the structure of the seed protein; and 4) usefulness of the annotation, which can be more or less appropriate to analyze site evolution in a phylogenetic context.

Figure 4 plots the distribution of χ depending on the site partition with CONF/MIX model. The solvent accessibility–based partition (EX) is associated with the highest χ values (i.e., the most reliable annotations), whereas EHO corresponds to the lowest ones. This is an expected result, as solvent accessibility is well defined and associated with high-likelihood gains (Fig. 1), whereas secondary structure annotations are not fully precise and useful, as seen from Figure 1 where the partition approach (PART) is not any better than the mixture (MIX). When combining these 2 site partitions to obtain EX_EHO, the χ values are intermediary. This reflects both the uncertainty in EX_EHO annotations,

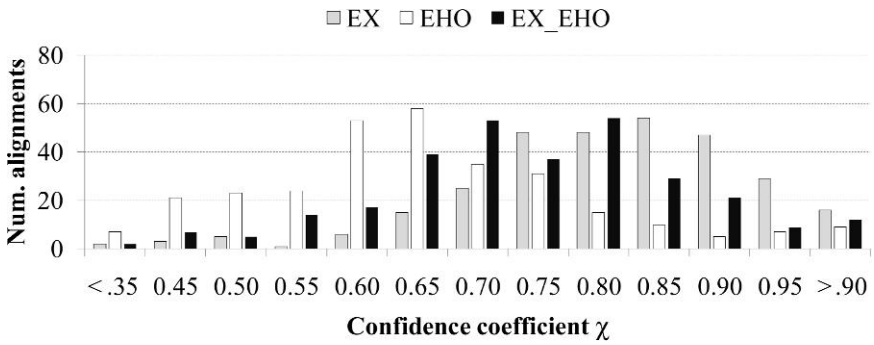

FIGURE 4. Distribution of the confidence coefficient χ depending on the site partition. This graph was obtained with CONF/MIX (confidence-based model using mixtures for nonreliable sites) and the 300 HSSP test alignments. When χ equals 1, annotations are fully reliable, whereas $\chi = 0$ means that available annotations are not any better than random guessing. EX = solvent accessibility–based site partition; EHO = secondary structure–based site partition; EX_EHO = combination of EX and EHO site partitions.

which cannot be more reliable and precise than EHO annotations, and the fact that the 6 EX_EHO categories are specially relevant and useful in a phylogenetic context. Overall, we thus see that χ combines several different aspects. Moreover, Figure 4 shows that the χ values are highly variable, ranging from $\sim$0 to $\sim$1. Further investigation would be needed to characterize the sites and alignments with poor annotation and build new models, for example, with χ values modulated along the sequences (typically close to 0 at alpha-helix extremities) and among proteins (e.g., disregarding secondary structure with small proteins dominated by disulfide bridges).

Topological Impact

The true tree is usually unknown with real data (as opposed to simulated data), and thus, it is hard to assess the topological accuracy induced by any tree building approach in a realistic setting. Here, we studied the topological impact of our models, that is, whether using these models, we frequently infer trees that differ from those inferred with standard models.

Using the 300 HSSP test alignments and previous experimental conditions (Γ4, SPR tree search, etc.), we compared the topologies obtained with our multiple-matrix models with the LG topologies. We counted the number of alignments where the tree built using any given model M is not the same as the tree inferred with LG. Both M and LG trees were also compared using the Robinson and Foulds (1979) distance, which is the number of clades that belong to one tree but not to the other. When different topologies are found, one should prefer the one with best likelihood value, or best AIC (or similar criterion) value, when evolutionary models used for tree inference involve different numbers of parameters. However, the difference may be slight and nonsignificant, so one cannot reject the topology with a lower fit to data. We thus counted the number of alignments where M and LG topologies differ, and where M is significantly better than LG, using a KH test on AIC values with $P < 0.01$.

Finally, we checked that the observed topological differences comprised some clades with significant support. Indeed, the topological impact would be low if all differences corresponded to poorly supported clades. To this end, we performed bootstrap analyses and compared the topologies and clade supports of EX_EHO_CONF/MIX (our best model), LG, and JTT. We counted the number of clades with notable bootstrap support (BP1 $\geq$ 50%) in one tree, which were not recovered in the other tree, or had a much lower support in this tree (BP2 + 50% $\leq$ BP1). For example, one clade with BP1 = 40% was not counted, even when it was not recovered in the other tree; on the opposite, one clade with BP1 = 80% in one tree was counted when it was found in the other tree with BP2 = 20%. This measure thus summarizes the topological and clade support differences. We used only 50 bootstrap replicates for computing time reasons, but the differences in clade supports corresponding to our gap of 50% between BP1 and BP2 are highly significant (P value $\sim$0.0 using a z-test for 2 proportions). Moreover, 50% of bootstrap support was shown to be optimal in terms of topological error (Berry and Gascuel 1996; see also Holder et al. 2008).

Results are displayed in Table 1. We see that all models frequently infer topologies that differ from LG topologies. Moreover, these topologies are significantly better than LG topologies in most cases. For example, EX_EHO_CONF/MIX finds 230 topologies that differ from LG ones, and among these 225 correspond to significant AIC gains. In practice, one should thus retain EX_EHO_CONF/MIX topology and discard LG one with 225 alignments among 300. The difference with JTT is even larger, as EX_EHO_CONF/MIX topology should be retained with 248 alignments. The topological distance between these trees is also appreciable. Comparing LG and EX_EHO_CONF/MIX, the average distance equals 0.136, meaning that on average, the LG and EX_EHO_CONF/MIX topologies differ with $\sim$11 clades (among $\sim$106 on average). With JTT the average topological distance with EX_EHO_CONF/MIX trees is even larger and corresponds to $\sim$14 clades on average.

TABLE 1. Topological impact

	MIX	PART	CONF/LG	CONF/MIX
EX				
≠LG top	210 (158)	219 (171)	172 (162)	207 (191)
R&F LG top	0.127	0.140	0.061	0.117
EHO				
≠LG top	177 (122)	200 (108)	157 (110)	186 (147)
R&F LG top	0.098	0.117	0.053	0.103
EX_EHO				
≠LG top	214 (184)	232 (208)	195 (190)	230 (225)
R&F LG top	0.132	0.157	0.075	0.136

Notes: These results are obtained with the 300 HSSP test alignments. " ≠LG top" is the number of alignments (among 300) where given model and LG topologies differ; the bracketed number is the number of cases where given model topology has significantly better AIC value than LG topology (KH test on AIC values with $P < 0.01$; each topology is analyzed with its own model). "R&F LG top" is the average Robinson and Foulds (1979) topological distance between given model and LG topologies for all 300 alignments; this distance is normalized between 0 (both trees are identical) and 1 (they do not share any clade in common). MIX = mixture model; PART = partition model; CONF/LG = confidence-based model using LG for nonreliable sites; CONF/MIX = confidence-based model using mixtures for nonreliable sites; EX = solvent accessibility–based site partition; EHO = secondary structure–based site partition; EX_EHO = combination of EX and EHO site partitions.

Bootstrap analyses result in 171 alignments where LG and EX_EHO_CONF/MIX trees differ by at least one clade with notable support (BP1 $\geq$ BP2 + 50%, see above); the total number of such clades equals 338 with all 300 HSSP test alignments. Comparing JTT and EX_EHO_CONF/MIX, we found 213 alignments where both trees have notable differences, corresponding to 509 clades in total. These numbers of clades (<2 per alignment on average) are low compared with the topological differences obtained with the standard Robinson and Foulds topological distance. This effect was expected, as it is commonly observed in phylogenetics that clades with strong support tend to be recovered with all models and inference methods. To scale the differences observed between JTT, LG, and EX_EHO_CONF/MIX, we thus used the same protocol to compare LG + Γ4 (as in former experiments) and LG without gamma distribution of rates across sites; we found 209 alignments where trees show notable differences, corresponding to 510 clades in total. These results are very close to those of JTT versus EX_EHO_CONF/MIX. This indicates that the topological impact of our models versus standard models (JTT, WAG, and LG) is in the same range as that induced by the use of a gamma distribution of rates, a finding which is consistent with the results above on likelihood gains.

Table 1 (without regard on branch supports) seems to indicate that the topological impact of CONF/LG models is lower than that of CONF/MIX ones. Most notably, the topological distance with LG trees is about twice lower for CONF/LG than for CONF/MIX, thus showing some (expected) closeness between CONF/LG and LG. However, when only counting clades with notable support difference (BP1 $\geq$ BP2 + 50%, see above), CONF/LG has similar impact as CONF/MIX. Comparing EX_EHO_CONF/LG and JTT (LG), we found 225 (168) alignments where both inferred trees have notable differences, corresponding to 530 (323) clades in total. With JTT, these results are even a bit higher than those obtained by EX_EHO_CONF/MIX (209 alignments corresponding to 510 clades). Moreover, as expected, there is a few notable differences between EX_EHO_CONF/LG and EX_EHO_CONF/MIX (61 clades in total). This further supports the use of CONF/LG models, which obtain similar likelihood values and topologies as CONF/MIX models, but require much lower computing times.

All together, these results demonstrate that our models have a notable topological impact. Although it is not clear whether the resulting topologies are closer to the true phylogenies, they are, in most cases, different from the topologies inferred using standard models (JTT, WAG, LG, etc.), with significantly higher likelihood values and new well-supported clades. These alternative topologies should thus be of great interest for phylogeneticists. Just as was observed with other substitution model improvements (e.g., transition/transversion for DNA in Kimura 1980 or gamma-distributed site rates in ML methods by Yang 1993), it is most likely that the high likelihood gains obtained here will also result in more accurate phylogeny reconstructions.

CONCLUSIONS AND PERSPECTIVES

We have described simple phylogenetic models to benefit from the structural annotations available for a large number of proteins (~50,000 entries in PDB and ~35% of Uniprot covered by HSSP). These models continue previous research, mainly by Koshi and Goldstein (1995), Thorne et al. (1996), and Goldman et al. (1998). Our results confirm their observations that evolutionary processes are quite heterogeneous depending on site structural configurations and that highly significant likelihood gains are obtained when accounting for this heterogeneity. The main difference between these previous studies and our paper is that we explicitly use available structural annotations, whereas their approaches were based on mixtures or HMMs, where all structural categories were envisaged for each site, a solution imposed by the low number of 3D structures available in the 1990s. Moreover, our models cope with the uncertainty inherent in structural annotations and include a gamma distribution of rates across sites, which is an essential component in substitution modeling. All together, our models provide likelihood gains over standard models (JTT, WAG, and LG) that are of the same magnitude as the gain obtained by adding a gamma distribution of site rates to standard models. The topological impact of these new models is also notable and (again) in the same range as that induced by the use of a gamma distribution with standard models. We thus believe that accounting for structural annotations in protein phylogenetics should become routine in near future using our models or their foreseeable refinements.

Several research directions deserve to be explored. First, some tools should be implemented to connect the phylogeny domain to structural bioinformatics. For example, extracting and synthesizing the structural annotations associated with multiple alignments is still a difficult task, despite the progress of some recent Web servers (e.g., Pollastri et al. 2007). Moreover, our models could be refined in several ways, for example, using different structural categories (e.g., 3 exposure classes, other groupings of the 8 DSSP secondary structure states), or studying solutions with variable confidence coefficients (χ) among site categories or along the sequences. Finally, we have described general models for globular proteins, but it is well known that substitution processes vary among life domains (e.g., apicomplexa or viruses) and protein groups (e.g., mitochondrial or membrane proteins). Specific substitution rate matrices and models should be estimated and implemented to analyze these data and nonglobular proteins.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by the Mitosys project of ANR BioSys.

Acknowledgments

Sincere thanks to Nick Goldman, Stéphane Guindon, Mark Holder, Clemens Lakner, Nicolas Lartillot, Juliette Martin, David Posada, Aylwyn Scally, and Hidetoshi Shimodaira for discussions, comments, and help.

References

Akaike H. 1974. A new look at statistical model identification. IEEE Trans. Automat. Control. 19:716–722.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L.L. 2002. The Pfam protein families database. Nucleic Acids Res. 30:276–280. Available from: http://pfam.cgb.ki.se/.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. 2000. The protein data bank. Nucleic Acids Res. 28:235–242. Available from: http://www.pdb.org.

Berry V., Gascuel O. 1996. Interpretation of bootstrap trees: threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999–1011.

Branden C., Tooze J. 1999. Introduction to protein structure. 2nd ed. New York: Garland Publishing.

Bryant D., Galtier N., Poursat M.A. 2005. Likelihood calculations in phylogenetics. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. p. 33–62.

Chothia C., Lesk A.M. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826.

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

Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple method of sequence data. Mol. Biol. Evol. 14: 685–695.

Gascuel O., Guindon S. 2007. Modelling the variability of evolutionary processes. In: Gascuel O., Steels M., editors. Reconstructing evolution: new mathematical and computational advances. Oxford: Oxford University Press. p. 65–99.

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

Guindon S., Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59(3):307–321.

Holder M.T., Sukumaran J., Lewis P.O. 2008. A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. Syst. Biol. 57:814–821.

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

Hordijk W., Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics. 21(24):4338–4347.

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

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

Kabsch W., Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Klosterman P.S., Uzilov A.V., Bendaña Y.R., Bradley R.K., Chao S., Kosiol C., Goldman N., Holmes I. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics. 7(1):428.

Koshi J.M., Goldstein R.A. 1995. Context-dependent optimal substitution matrices. Protein Eng. 8:641–645.

Lanave C., Preparata G., Saccone C., Serio G. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

Le S.Q., Gascuel O. 2008. An improved general amino-acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Le S.Q., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.

Le S.Q., Lartillot N., Gascuel. 2008. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. B. 363:3965–3976.

Lio P., Goldman N., Thorne J.L., Jones D.T. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics. 14:726–733.

Pagel M., Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. p. 121–142.

Pollastri G., Martin A.J.M., Mooney C., Vullo C. 2007. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics. 8:201.

Raman P., Cherezov V., Caffrey M. 2006. The membrane protein data bank. Cell. Mol. Life Sci. 63:36–51.

Rannala B., Yang Z.Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Robinson D., Foulds L. 1979. Comparison of weighted labeled trees. Berlin (Germany): Springer-Verlag (Lecture notes in mathematics; vol. 748), p. 119–126.

Schneider R., de Daruvar A., Sander C. 1997. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 25: 226–230.

Shimodaira H. 1997. Assessing the error probability of the model selection test. Ann. Inst. Stat. Math. 49:395–410.

Shrake A., Rupley J.A. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79:351–372.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Providence (RI): American Mathematical Society (Lectures on mathematics in the life sciences; vol. 17). p. 57–86.
Thorne J.L., Goldman N., Jones D.T. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. 13:666–673.
Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.
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. 2006. Computational molecular evolution. Oxford: Oxford University Press.