Probabilistic Models and Their Impact on the Accuracy of Reconstructed Ancestral Protein Sequences

Tal Pupko[1,*], Adi Doron-Faigenboim[1], David A. Liberles[2], Gina M. Cannarozzi[3]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

[2]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

[3]Institute of Computational Science, ETH Zurich 8092 Zürich, Switzerland.

*Corresponding author
Email: talp@post.tau.ac.il

Abstract

Evolutionary models are at the heart of numerous bioinformatic and molecular evolution challenges such as searching for remote homologous sequences, phylogenetic reconstruction, and detecting positive and purifying selection. In this chapter, we review probabilistic evolutionary models used for reconstructing ancestral protein sequences, and discuss their impact on the accuracy of the reconstructed sequences. We discuss various aspects of current models, such as among site rate variation, variation of the substitution matrix among positions, non-homogeneity and non-stationarity as well as the covarion process. We also present an algorithmic approach that uses external information to increase the accuracy of ancestral reconstruction. Model selection, Bayesian approaches for ancestral reconstruction, the handling of missing characters and gapped positions and the integration of structural information on ancestral sequence reconstruction are also discussed. Finally, computational aspects of joint and marginal ancestral sequence reconstruction are presented.

1. Probabilistic Evolutionary Models

Recent large-scale sequencing efforts are changing the dogma of biological research. To understand and utilize the ever-increasing sequence databases, one must use sequence evolutionary models (reviewed in Whelan et al. 2001). These models are fundamental in various bioinformatics applications, such as protein structure prediction, protein function prediction, sequence motif finding, active site prediction, evolutionary studies, gene prediction, comparative genomics, RNA structure predictions, tree reconstruction and ancestral sequence reconstruction (ASR). When using evolutionary models, we have to be careful in choosing the model assumptions. There are many examples where the use of unrealistic models of sequence evolution leads to erroneous conclusions (Pupko et al. 2002b; Sullivan and Swofford 1997). Novel models of sequence evolution are continuously being developed both in terms of modeling choices and computational tools. Advanced statistical techniques are used to learn parameters of these models and to predict with them. Special effort is directed to better take into account realistic biological phenomena, removing possible sources of error from existing oversimplified models. In this chapter, the effect of the model assumptions on ASR will be discussed. Existing methods for sequence reconstruction are based on the Maximum Parsimony (MP) criterion or on probabilistic models. ASR using probabilistic models is based on either the Maximum Likelihood (ML) or the Bayesian paradigms. This chapter will focus on probabilistic based methods for ASR of protein coding sequences. However, for completeness, in the following section we briefly describe ASR based on the MP criterion, which was widely used before probabilistic ASR methodology was developed.

2. Ancestral Sequence Reconstruction Based On the Maximum Parsimony Criterion

The idea of maximum parsimony (MP) is to identify the ancestral states at each node of a tree that minimize the number of character changes needed to explain the observed differences among the sequences at the leaves. Algorithms for ASR based on this criterion were developed by Fitch (1971), Sankoff (1975), and Sankoff and Rousseau (1975). These algorithms use dynamic programming, ensuring efficient reconstruction. The Fitch algorithm, introduced with nucleotide sequence data, penalizes equally any change among the four character states (*A*, *C*, *G*, and *T*). For the reconstruction of a specific position, the algorithm proceeds by assigning to each node of the tree a set of character states that are compatible with minimum number of changes. The algorithm processes the tree in post-order, i.e., each tree node is visited only after its descendants are visited. Thus, the algorithm starts by assigning character sets to the leaves of the tree. If, for example, a leaf is labeled by the character *A*, a set $\{A\}$ is assigned to that leaf. Next, an internal node for which both descendents have already been visited is evaluated. The set assigned to this internal node is the intersection of the sets at its two descendant nodes if this intersection is not empty, or the union of the two sets if the intersection is empty. If the new set is a union, one change is counted so that the number of changes is the number of the union operations. The next step is to traverse the tree from root to leaves, in pre-order, to determine the ancestral states for internal nodes. Initially the

ancestral state at the root is equal to the character state in its set. If this set includes more than one character different equally parsimonious reconstructions exist. Then, each descendent of the root is evaluated as follows: if the ancestral state at the root is a member of the set of the descendent node the same ancestral state is assigned to the descendent node, otherwise another state from the set in the descendent node is chosen. This procedure is applied for each node in the tree. This procedure will find some of the most parsimonious reconstructions but not all. To guarantee that all most parsimonious reconstructions are found, comparisons involving the outgroups of a node must be performed (Harvey and Pagel 1991; Maddison et al. 1984).

To exemplify the algorithm of ASR using the Fitch's algorithm consider the simple 5-taxa tree in figure 1:

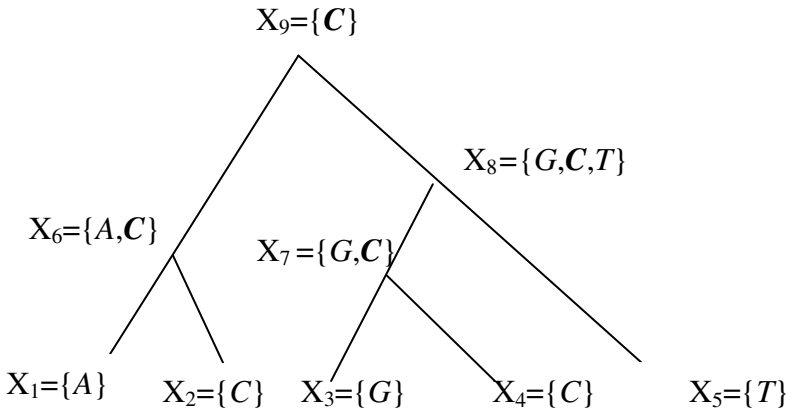

Fig. 1

For the character illustrated, the data observed are $X_1 = A$, $X_2 = C$, $X_3 = G$, $X_4 = C$, and $X_5 = T$. At the leaves the character sets are simply: $X_1 = \{A\}$, $X_2 = \{C\}$, $X_3 = \{G\}$, $X_4 = \{C\}$, and $X_5 = \{T\}$. At node X_6, the intersection of the sets of its two descendants X_1 and X_2 is $\{A\} \cap \{C\} = \phi$. Hence, the union set is assigned: $\{A\} \cup \{C\} = \{A, C\}$. Likewise, the union set of X_3 and X_4 is assigned at node X_7, i.e., $\{G\} \cup \{C\} = \{G, C\}$. Now the set at node X_8 can be determined, since the intersection of sets X_5 and X_7 is again empty, the union of these sets $\{G, C, T\}$ is assigned. Finally the set in the root (X_9) is the intersection of the sets X_8 and X_6: $\{A, C\} \cap \{G, C, T\} = \{C\}$. Three union operations were needed thus a minimum of three changes is needed for this reconstruction. In the next step, the ancestral states are determined (marked in bold type in figure 1) by traversing the tree in pre-order (from the root to the leaves). First the state *C* is determined at the root; the state at X_8 is also set to C since this state is the ancestral state in the parent (X_9) and is a member of the set at that node (X_8). Similarly, the state at X_6 is *C* since this state is the ancestral state in the parent, as well as a member of the set at node X_6. Finally, the state at node X_7 is assigned and is equal to *C*.

The Sankoff (1975) algorithm is a generalization of Fitch's algorithm. Instead of assuming all state changes are equally likely, it allows different costs for different

character changes. Similarly to the Fitch's algorithm, the tree is visited in post-order followed by pre-order steps.

Both algorithms may reconstruct more than one ancestral state for each node. When the results are ambiguous two different methods of assignment: acceleration transformation (ACCTRAN) or delayed transformation (DELTRAN) can be applied (Swofford and Maddison 1987). ACCTRAN assumes that the character changes happen at the earliest possible point and thus prefer reversals over convergences. DELTRAN tries to delay the changes and thus maximizes parallelism.

3. Ancestral Sequence Reconstruction Using Probabilistic Models

In phylogeny, probabilistic models (both maximum likelihood and Bayesian approaches) are considered the state-of-the-art methods for tree reconstruction (Holder and Lewis 2003). Felsenstein's (1981) seminal work showing how to efficiently compute the likelihood of a tree, together with the efficient computer program PHYLIP (Felsenstein 2005, distributed from 1980) boosted interest in probabilistic models for phylogeny. During the next two decades, such probabilistic approach replaced the previously more common MP approach as numerous studies demonstrated the many shortcomings of the later (e.g., Holder and Lewis 2003). For instance, MP is inherently biased toward overestimating the number of ‘‘common to rare’’ changes (Eyre-Walker 1998). Furthermore, this method does not supply statistically robust means for discriminating among equally parsimonious reconstructions (Yang et al. 1995). Unlike MP, the probabilistic approach also allows the statistical testing of various hypotheses, such as testing whether two tree topologies are significantly different or testing for a monophyletic origin of a clade (reviewed in Goldman et al. 2000).

The history of ASR followed that of tree reconstruction, albeit with a delay. Until the concept of probabilistic sequence reconstruction was introduced in the early 90's (Gonnet and Benner 1991; Koshi and Goldstein 1996; Schluter 1995; Yang et al. 1995), MP was the method of choice (e.g., Jermann et al. 1995; Stewart 1995). However, it is clear that the same MP shortcomings that have been mentioned in the context of tree reconstruction are also valid for ASR that is parsimony based. Later Koshi and Goldstein (1996), and Pupko et al. (2000) developed efficient algorithms for both joint and marginal reconstruction using the probabilistic approach (see below for a detailed explanation of these concepts).

A vital advance in the development of evolutionary models was the consideration of heterogeneity of evolutionary rates among sequence sites (Yang 1993). Yang has shown that a model that takes into account such among-site-rate-variation significantly increases the tree likelihood. In the case of ASR, Pupko et al. (2002c) showed via simulations that failure to account for among site rate variation also reduces the accuracy of ASR and results in lower likelihood for the reconstructed sequences. As can be expected, a general pattern emerges: models that better fit data for tree reconstruction are also better for ASR. Areas in which model improvements have been attempted include: accounting for rate variation between different amino-acids (the substitution matrix) and the variation of this substitution matrix among different sites of a protein, and among different branches of the phylogenetic tree. The tree and its associated branches are also considered as part of the probabilistic model. Thus, the impact of taking into account

uncertainties in tree topology and model parameters within a Bayesian approach on the accuracy of ASR was also explored. In the subsequent sections we describe each aspect in details.

4. How Ancestral Sequences are Computed Using Probabilistic Models

To exemplify the concept of probabilistic ASR consider the following simple 4-taxa tree.

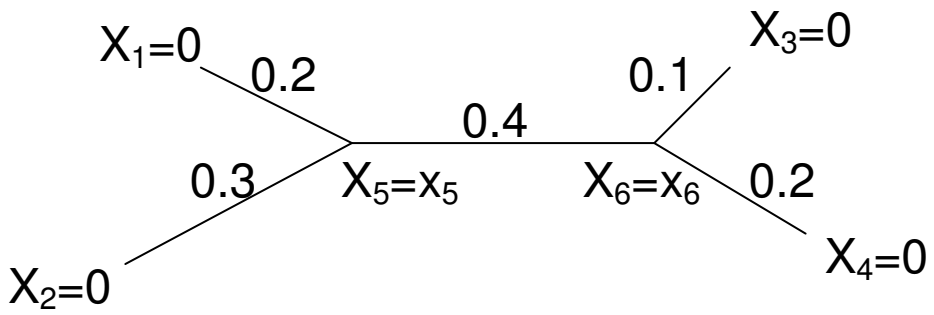

Fig. 2

For simplicity we consider a two-state (0 or 1) alphabet (for example, polar versus non polar amino-acids). The ancestral character assignments, x_5 and x_6 at the internal nodes X_5 and X_6 are unknown. Numbers above branches indicate branch lengths, i.e., average number of substitution per sequence site. In all probability-based models, the probabilities are expressed in terms of summations and multiplications of $P_{ij}(t)$ factors, the probability that character i will be replaced by character j along a branch of length t. The $P_{ij}(t)$ factors are usually expressed in a matrix form $P(t)$, so that $[P(t)]_{ij} = P_{ij}(t)$. The matrix $P(t)$ can be computed by $P(t) = e^{Qt}$, where Q is the instantaneous rate matrix and t the branch length (Felsenstein 2004). The probability model also contains initial probabilities: for each amino acid x, $P(x)$ denotes the probability of observing x at the root of the tree. The likelihood of the data describes the probability of observing the characters at the leaves given the tree topology, the branch lengths and the $P_{ij}(t)$ factors. Thus, the following expression represents the likelihood of the tree in figure 2.

$$\sum_{x_5}\sum_{x_6} P(x_5)P_{x_5,x_1}(0.2)P_{x_5,x_2}(0.3)P_{x_5,x_6}(0.4)P_{x_6,x_3}(0.1)P_{x_6,x_4}(0.2). \qquad (1)$$

This likelihood is a sum of 4 different terms, each corresponding to a specific ancestral sequence assignment (x_5=x_6=0, x_5=x_6=1, x_5=0 and x_6=1, x_5=1 and x_6=0). In this case, internal node X5 was arbitrarily chosen as the root of the tree. Most evolutionary models used are time reversible. In mathematical terms, a model is time reversible if $P(i)P_{ij}(t)=P(j)P_{ji}(t)$ for all pairs of characters, i and j. Felsenstein (1981) showed that for time reversible models, the position of the root of the tree does not effect the likelihood score. The joint character assignment (x_5, x_6) which contributes the most to the above likelihood is called the joint ancestral sequence reconstruction and is explicitly given by the expression:

$$\text{argmax}_{x_5,x_6}\left(P(x_5)P_{x_5,x_1}(0.2)P_{x_5,x_2}(0.3)P_{x_5,x_6}(0.4)P_{x_6,x_3}(0.1)P_{x_6,x_4}(0.2)\right). \quad (2)$$

The maximum of the above expression is the likelihood of the joint reconstruction.

In essence, x_5 and x_6 are two non-independent random variables over the [0, 1] set. The term of equation 1 above, in which x_5=0, and x_6=0, corresponds to $P(x_5=0,x_6=0,data)$, whereas "data" refers to x_1=0, x_2=0, x_3=1, and x_4=0. The probability of x_5=0 and x_6=0 given the data is simply

$$P(x_5=0,x_6=0\mid data)=\frac{P(x_5=0,x_6=0,data)}{P(data)} \quad (3)$$

$P(data)$ is exactly the expression given in equation (1) above. The four possible values $P(x_5, x_6 \mid data)$ can be expressed in a tabular form:

	$x_6=0$	$x_6=1$	sum
$x_5=0$	$P(x_5=0,x_6=0\mid data)$	$P(x_5=0,x_6=1\mid data)$	$P(x_5=0\mid data)$
$x_5=1$	$P(x_5=1,x_6=0\mid data)$	$P(x_5=1,x_6=1\mid data)$	$P(x_5=1\mid data)$
sum	$P(x_6=0\mid data)$	$P(x_6=1\mid data)$	1

From these joint probabilities, one can easily compute the marginal probabilities. For example, the probability that character 0 was the ancestral state at node X_6, given the available current sequence is $P(x_6=0\mid data)=P(x_5=0,x_6=0\mid data)+P(x_5=1,x_6=0\mid data)$. Thus, if we are interested in the best character assignment to node X_6, we should compare $P(x_6=0\mid data)$ and $P(x_6=1\mid data)$ — if the former is higher, 0 is the most likely reconstruction at this node, otherwise it is 1. As was shown by Pupko et al. (2000), joint and marginal reconstructions are not always the same. If for example $P(x_5=0,x_6=0\mid data)$ = 0.4, $P(x_5=0,x_6=1\mid data)$=0.3, $P(x_5=1,x_6=0\mid data)$=0.05, $P(x_5=1,x_6=1\mid data)$=0.25, the highest would be the first, indicating character 0 at both internal nodes. However, the marginal probabilities of assigning 1 to node 6 would be 0.55, indicating that this is the most likely marginal reconstruction. A similar explanation of such ancestral reconstruction probabilities can be found in Koshi and Goldstein (1996) and Yang et al. (1995).

5. Ancestral Sequence Reconstruction Taking Into Account Among Site Rate Variation

While heterogeneity of the number of amino-acid replacements in different sites, can result from the stochastic nature of the process, it was understood that the observed pattern of variability in the number of replacement significantly deviates from the expected pattern under a model which assumes homogenous rate at all sites (Uzzell and Corbin 1971). This rate heterogeneity stems from the fact that not all sites in a protein are subject to the same evolutionary constraints. Sites that are important to maintaining the structure of function of a protein, such as the active site residues, are usually highly

conserved, while other sites evolve at higher rates. Among site rare variation (ASRV) models aim to mathematically express this heterogeneity of evolutionary rates across sites.

In ASRV models it is assumed that each site has a fixed rate *r*, which indicates how fast this position evolves relative to the average rate over all positions (Yang 1993). Thus, a site with a rate of 2 evolves twice as fast as the average, i.e., the expected number of substitutions in this site is twice that of the average. More formally, when a position evolves at a rate *r*, we assume that the rate matrix *Q* underlying the site's Markovian's process is multiplied by *r*. We note that since $P(t) = e^{Qt}$, the same replacement probabilities will be obtained by either multiplying the rate matrix by *r* or by multiplying the branch length *t* by the same factor *r*. This equivalency shows that the likelihood of a position that evolves at a rate *r* can be computed by first multiplying all the branch lengths of the tree by *r*, and then computing the likelihood of that position with *Q*. This allows computing the likelihood of all sites with the same *Q*, rather with a different *Q* matrix for each rate, thus significantly reducing computation times.

The rate at each site is in general unknown. Whereas one can try to estimate the most likely rate at each site (e.g., Nielsen 1997; Pupko et al. 2002a), when the goal is to reconstruct the tree topology or the ancestral characters, it is preferable to include the various possible rates in the computation in a probabilistic manner (Felsenstein 2001). Thus, it is usually assumed that several rates are allowed at each site, each rate with a specific probability. Given such a rate distribution, the likelihood of the data is computed by summing the likelihood over all possible rates, taking into account their probabilities. A discrete approximation of the gamma distribution (Yang et al. 1994) is by far the most widely used rate distribution.

In the paper of Yang (1995) on probabilistic ancestral sequence reconstruction, ASRV was not taken into account. Yang suggested that "the relative contributions to the likelihood by different reconstructions at a site is unlikely to change significantly when the branch lengths are multiplied by a constant…" Although this statement may be true regarding one most likely character at a specific node at a specific site, it is not true for the probability vector: the probabilities of each character at each node and site (`xxxChapterGina`). The most likely character is just the one that maximizes the probability vector. This probability vector is often taken to represent the confidence interval of the reconstruction, and moreover, it is used in various applications such as detecting co-evolving substitutions and substitution mapping (Bollback 2006; Dimmic et al. 2005; Dutheil et al. 2005) and detecting radical replacements (Pupko et al. 2003). It is intuitively expected that the most likely reconstruction will be less sensitive to model assumptions. Yet, this is clearly not the case for probability vectors and for applications which use such vectors as part of their computations. Furthermore, it was later shown that taking into account ASRV significantly increases the accuracy of the most likely reconstruction, especially for sequences that are highly diverged (Pupko et al. 2002c). Thus, ASRV should be especially critical when highly diverged sequences are reconstructed, which is often the case (e.g., Chang et al. 2002; Thornton 2001).

In some cases, computing ancestral sequences using ASRV model is not trivial. When marginal reconstruction is needed, computational time is linear with the number of sequences whether or not ASRV is assumed. However, computing joint reconstruction assuming ASRV is exponential with the number of sequences. To this end, Pupko et al.

(2002c) developed an efficient branch-and-bound algorithm that, although exponential in the worst case, can handle dozens of sequences in most practical cases. It should be noted that this algorithm *guarantees* finding the joint ML reconstruction, i.e., it is not a heuristic approach.

One limitation of all ASRV models discussed above is that they assume a constant rate throughout evolution, which is not always the case. This issue is discussed in section 7.

6. Model Selection and Ancestral Sequence Reconstruction

Which rate distribution is best for ancestral sequence reconstruction? As in phylogenetic reconstruction, the "best" distribution cannot be determined a-priori and should be determined based on the available data. The most widely used distribution for modeling ASRV is the discrete gamma distribution. Is the discrete gamma distribution the best? In terms of likelihood it seems that allowing a proportion of the sites to be invariable and having the rest of the rates sampled from a gamma distribution results in an improved likelihood (Gu et al. 1995). However, this "Gamma + Invariant" model is also not ideal, as the proportion of invariant sites seems to be highly sensitive to the amount of sequences used in the analysis. Recently, Mayrose et al. (2005) suggested using a mixture of gamma distributions to better account for the complicated pattern of ASRV. This model significantly increases the likelihood and it is expected that it will also increase the accuracy of ancestral sequence reconstructions and methods that rely on them.

Complex evolutionary models are usually more realistic from the biological point of view. Yet, they often require the estimation of additional parameters from the same amount of data, and hence, the standard error of each estimated parameter is increased. This tradeoff between rich models with many parameters and simple models with a few parameters is the topic of extensive research in statistics, which is known as model selection (Burnham and Anderson 2003). In general, criteria such as the Akaike Information Criterion (AIC) and the likelihood ratio test (LRT) (Posada and Buckley 2004) are often used to compare different models. Hence, it is advisable to search for the best-fitting model for the given data using programs such as ProTest (Abascal et al. 2005), and only then to perform ASR using this best model. Theoretically, it is best to estimate both the ancestral sequences and the model parameters simultaneously. However, a two stage approach, consisting of first finding the parameters that maximize the likelihood of the data (average over all reconstructions), and then fixing these parameters for the ASR is more efficient and should provide essentially identical results as compared to the simultaneous approach.

7. The Instantaneous Rate Matrix and its Impact on Ancestral Sequence Reconstruction

As stated above, current likelihood models are based on two components: the rate matrix Q which determines the substitution probabilities and the ASRV parameters. The determination of Q is critical to any evolutionary model. As it turns out, different approaches for the determination of Q were developed for DNA (coding or non-coding),

RNA, amino-acid, and codons. As the focus of this book is on ancestral protein sequences, we will describe in detail only amino-acid and codon based matrices.

Most amino-acid matrices are "empirical". Empirical matrices are derived from large datasets and hence accurate estimates of amino-acid replacement probabilities can be obtained. A few empirical amino-acid replacement matrices have been previously proposed (Adachi and Hasegawa 1996; Dayhoff et al. 1978; Gonnet et al. 1992; Jones et al. 1992; Whelan and Goldman 2001). These matrices are extensively used in amino-acid based applications such as programs for multiple sequence alignment, detecting distant homologs, phylogenetic tree, and ancestral sequence reconstruction (e.g., Altschul et al. 1997; Penny and Hasegawa 1997; e.g., Thompson et al. 1994). The strength of empirical matrices stems from the fact that they are derived from averaging over many genes from various organisms, and thus, the rate matrix entries usually correspond to accurate estimation for the *average* replacement probabilities. However, in these matrices, no information about replacement probabilities can be learned from the specific data analyzed, i.e., Q has no free parameters. This concern can be a major difficulty when the protein analyzed evolves in a manner distinct from the "average" protein.

Should the same amino-acid matrix be used to model the evolution of all positions of a given protein? It is well known that not all regions within a protein evolve under the same evolutionary constraints. For example, transmembrane regions of proteins are known to evolve under different evolutionary constraints than non-transmembrane regions. Jones, Taylor, and Thornton (1994) have computed specific amino-acid replacement matrices for transmembrane and non-transmembrane domains. Other context-dependent matrices were developed for secondary structures (alpha helices, beta sheets and loops) or for buried versus exposed structural elements (Koshi and Goldstein 1995). Unfortunately, these matrices are not commonly used for ASR, although it is expected that using an alpha helix based matrix when reconstructing the ancestral characters of an alpha helix region will result in more accurate reconstruction compared to a general matrix such as the widely used JTT matrix of Jones et al. (1992). In addition, specific amino-acid matrices were developed for the mitochondria (Adachi and Hasegawa 1996) and for chloroplasts genomes (Adachi et al. 2000). These matrices reflect both the different mutation pattern in these genomes (correlated with the different genetic codes used in these genomes, the different replication machinery, etc.), and the different selection pressures. With the advent of more sophisticated algorithms for constructing amino-acid replacement models (e.g., Muller et al. 2002; Muller and Vingron 2000), it is expected that more such context-dependent matrices will be developed.

An effort was made to create mechanistic models for amino-acid replacement probabilities - models which include parameters that are fitted to each data analyzed. For example, in some models Q_{ij} depends on the difference in fitness between the pair of amino-acid, which is based on some physical-chemical properties, such as alpha helical propensity or hydrophobicity (Koshi and Goldstein 1998; Koshi et al. 1997). Another approach is to construct amino-acid based matrices from codon models (Yang et al. 1998).These models provide an intriguing alternative to the commonly used empirical matrices. More research is needed to test their applicability for phylogenetic tree inference and for ASR.

All the models suggested above assumed that the same *Q* matrix models all sites. These models, thus, do not allow heterogeneity of substitution pattern among sites. Models that take into account among-site substitution-variation are similar in concept to discrete ASRV models, in the sense that in the former, each site can evolve according to some predefined rate category, and in the latter, each site can evolve according to some predefined *Q* matrix. Such an approach was used by Dimmic et al. (2000), in which several *Q* matrices were used to model mitochondrial proteins - each such matrix constructed so that it can model different selection forces underlying the evolution of the various sites of these proteins.

Goldman and Yang (1994) and Muse and Gaut (1994) were among the first to suggest mechanistic codon-based evolutionary models. The more sophisticated Goldman and Yang (1994) model takes into account the transition-transversion bias, the codon frequencies, and the different replacement probabilities between amino acids based on the Grantham (1974) physico-chemical distance matrix. However, these models did not account for the heterogeneity of the evolutionary selection pressure among protein sites. Nielsen and Yang (1998) and Yang et al. (2000) further developed mechanistic Bayesian models that accounts for such selection heterogeneity. In their model, a prior distribution of the ratio of the nonsynonymous substitutions rate (*Ka)* to the synonymous substitutions rate (*Ks*) is assumed. Sites showing *Ka/Ks* values significantly lower than 1 are regarded as undergoing purifying selection and therefore may have a functionally or structurally important role. Sites showing *Ka/Ks* values significantly higher than 1 are indicative of positive Darwinian selection, suggesting adaptive evolution. However, unlike the model of Goldman and Yang (1994), these models ignore the fact that distinct amino acids differ in their replacement rates. Recently, an empirical codon substitution matrix was developed by Schneider et al. (2005). This model was used to estimate synonymous distance between coding sequences (Schneider et al. 2006). Another direction is the derivation of codon matrices from empirical amino-acid matrices (Doron- Faigenboim and Pupko. *in preparation*).

Current efforts in codon models focus on their applicability for detecting positive selection. However, it is clear that for coding sequences, these models can be very helpful for phylogenetic reconstruction and ASR. In this respect, the different substitution rates between amino-acids must be taken into account as in the original paper of Goldman and Yang (1994).

8. Covarion Models and Their Impact on Ancestral Sequence Reconstruction

Several new approaches show promise in generating better models of protein evolution. The realization that different sites in a protein evolve at different rates has led to the use of the abovementioned ASRV models. However, these models assume that over the course of evolution, the rate of substitution remains unchanged at a given protein site. Recent studies have shown this may not always be the case (e.g., Lopez et al. 2002). Sites that are highly conserved in one part of the tree may be variable in the rest of the tree and vice versa. Although this notion, termed covarion, was described decades ago by Fitch and Markowitz (1970), an evolutionary probabilistic model for the covarion process was only recently developed (Galtier 2001)+(`xxxChapterGina`). Galtier's covarion model

assumes that the rate itself is not fixed for each site, but rather follows a continuous time Markov process along the tree, with a specific rate-change parameter. The higher this rate switching parameter, the more common the rate jumps. In his pioneering work Galtier applied such a covarion model to infer the ancestral GC content of the most recent common ancestor (MRCA) of extant life forms. The moderate GC content found in this study brings into question the previously accepted hypothesis of a thermophilic origin of the MRCA. Gaucher et al. (2003) also used ASR to study whether the ancient bacteria were thermophiles. In this work, ancestral elongation factors sequences were inferred, synthesized in the lab and finally, their activity was empirically measured as a function of the temperature. Interestingly, the optimal temperature of the ancestral protein found strengthens the "classical" view that ancient bacteria were thermophiles.

It is difficult to draw many conclusions about the effect of covarion on ASR, since in Galtier's study the covarion model was used to study ancestral GC content rather than to infer ancestral sequences. Thus, the effect of the covarion assumption on the accuracy of the reconstructed sequences was not tested. Furthermore, the model used was a non-homogenous one (see section 8), and thus no analysis was performed to determine the effect of accounting for the covarion in standard homogenous models. However, especially when highly diverged sequences are analyzed, ASR is expected to be much more accurate when covarion models are introduced. This is especially true as it was shown that refraining to take into account the covarion process underestimates the number of multiple substitutions (Galtier 2001).

9. Deviations from Homogenous Stationary Reversible Models

Standard evolutionary models assume the following three assumptions: (1) the process is homogenous, so that the same Q matrix models the evolution in all branches. (2) The process is stationary, so that on average, the same character frequencies hold for all branches. (3) The process is reversible, so $P(i)Q_{ij} = P(j)Q_{ji}$. These concepts are explained in e.g., Galtier and Gouy (1998). All these assumptions are known to be violated in many biological examples (e.g., Chang and Campbell 2000; Galtier and Gouy 1995; Lockhart et al. 1994). The main reason for these assumptions is to avoid the introduction of too many parameters, thus risking over fitting the data (see section 5). In addition, models that do not make these assumptions are computationally intensive, for some tasks, such as finding the ML tree. Thus, these models may not be applicable even for moderately sized datasets.

When the tree topology is known, as is often the case in many ASR studies, this computational limitation is not a real hurdle. This enables the use of such complex models, and the subsequent gains of insights from these models regarding the function of the ancestral sequences.

One such example is the nonhomogenous model developed by Galtier and Gouy (1998). In their model, the G+C content is allow to change over time, so that each branch along the tree can have a distinct G+C content. They have shown via simulation studies that ML inference with this model can accurately infer the ancestral G+C content. They then used a variant of this model (allowing also ASRV) to reconstruct the G+C content in the MRCA (section 7).

Yap and Speed (2005) compared three models of nucleotide substitutions: the standard reversible one (REV), a stationary non reversible one (STAT) and a non stationary non reversible one (NONSTAT). They found that the NONSTAT model significantly improved the likelihood and could be used to root simple trees. In their NONSTAT model, the nucleotide frequencies at the root can be different from the stationary frequencies. Although the authors did not explore the possible effect of the NONSTAT model on ASR, their model suggests a way to test if the character frequencies at extant sequences reliably reflect those of the ancestral sequence. The effects and impact of the different assumptions, i.e., non-reversibility and non-stationarity on ASR await further investigation.

10. Using Side Information

Accurate ancestral sequence reconstruction depends on how well the model fits the data. Each model contains parameters such as tree topology, branch lengths, and the transition/transversion ratio that are estimated from the data. This estimation depends on the amount of information the data contain: in general, the more sequences and positions available for analysis, the higher the accuracy of each estimated parameter. However, when "data" are concerned, one can separate between two types of data. The first is the protein alignment whose ancestral sequences are to be inferred. We term this the "ASR data". The second type of data is any information that is not directly related to the protein sequence analyzed, yet can contribute to the accuracy of the ASR. We term this second type "side information".

Assume our goal is to reconstruct the ancestral *cytochrome b* of human, chimpanzee, gorilla and baboon. A naïve approach would consider adding an outgroup (e.g., mouse), and using these ASR data to search for the maximum likelihood tree topology and branch lengths and then to compute the ancestral sequence at the ancestor of these primates. A first improvement to this approach is not to search the tree topology based on cytochrome *b* sequences alone, but rather estimate the species tree based on a large set of orthologous sequences available for all these organisms. In this case, the species tree is assumed to reflect the topology of the gene trees, and this tree should be used for the ASR, rather than the ML tree based on the ASR data alone.

Such an approach was used for example in Krishnan et al. (2004) when reconstructing ancestral primate mitochondrial DNA. Even when a Bayesian approach, which takes into account many alternative trees, is considered, one should compute the trees' posterior probabilities not from the ASR data alone, but rather also from all other available side information. When using information external to the sequences to build a gene tree, one should be careful to evaluate the fit of the gene tree to the imposed species tree. Techniques for such evaluation have been previously developed using the parsimony (Berglund et al. 2006) or the Bayesian (Arvestad et al. 2003) frameworks.

Assuming the tree topology is known, should we infer branch lengths from the ASR data or can we use side information for more accurate branch lengths estimation? In other words, for the above example, assume that in addition to the cytochrome *b* sequences, we also have the sequences of cytochrome *c* oxidase subunit 1 from these four primates and mouse – how can this information be used to more accurately estimate the branch lengths of the cytochrome *b* tree? One approach is to concatenate these two

datasets and infer the branch lengths from the concatenated data. However, as has been previously shown, concatenation is by far the least correct method for combining different datasets (Pupko et al. 2002b; Yang 1996). In essence, concatenation assumes that all genes evolve at the same rate, an assumption that is known to be wrong: some genes are highly conserved (slow evolving), while others are highly variable (fast evolving). An alternative approach (the separate model) is to assume that branch lengths of each dataset are independent of each other. This assumes that no information is shared regarding evolutionary rates of different species. It is known that mouse evolves faster than human for all genes, so that the branch leading to human should be, on average, shorter than the branch leading to mouse. When assuming the separate model, this information is ignored, and a possible increase in branch length accuracy is overlooked. Furthermore, the separate model assumes a free parameter for each branch for each dataset: a very large number of parameters resulting in decreased accuracy for each branch length.

A better approach, first suggest by Yang (1996), would be to assume a proportional model, in which a base tree topology and branch lengths are assumed to be shared among all members of the dataset(see also Bevan et al. 2005; Pupko et al. 2002b). The tree for each gene in the dataset has the same topology as this base tree, but its branches are multiplied by a fixed rate factor, the evolutionary rate of this gene. To exemplify this point consider a hypothetical case in which the analysis of many genes from human and chimpanzee resulted in an estimate that the chimpanzee evolves 1.05 times faster than human. It can then be assumed that this same ratio exists between the branch lengths leading to human and chimpanzee in the analyzed ASR data. Knowing such relative evolutionary rates between organisms can significantly reduce the number of free parameters. We note, that for some genes, the proportionality assumption can be rejected, if for example, a rapid evolution of the gene in a specific lineage has occurred. Therefore, we suggest to first test which model best fits the entire data analyzed (concatenation, proportional or separate analysis) and then use the best model for branch length estimation.

Finally, should we use other vertebrate cytochrome *b* sequences if our goal is only to reconstruct the human-chimpanzee-gorilla-baboon ancestral sequence? In general, the answer is yes. All model parameters (tree topology and branch lengths, the alpha parameter of the gamma distribution, the transition/transversion ratio in case of DNA based model, amino acid frequencies, etc.) are more accurately inferred in cases when more data are available. One should always be cautious to specific cases in which this general more data - more accuracy assumption fails. The identification of such cases is strongly linked with methods aimed at identifying functional shifts in proteins, and is an active area of current research (e.g., Gaucher et al. 2002; Gu 2003; Pupko and Galtier 2002). See also section 7.

11. Taking into Account Uncertainties in the Tree Topology, Branch Lengths and Model Parameters

In the empirical Bayesian approach of ASR, one computes the most likely tree topology, branch lengths and model parameters in the first computational step. Then, the ancestral sequences are reconstructed using these maximum likelihood estimates (MLE) as fixed

values. This approach fails to consider the uncertainty in the MLE and may result in overestimated confidence.

Bayesian methods for ASR were suggested to overcome this uncertainty (Huelsenbeck and Bollback 2001; Schultz and Churchill 1999). Schultz and Churchill (1999) studied the effect of different prior distributions on the posterior probabilities of ancestral states in the simple case of a two-character-states model. Their approach was fully Bayesian: the parameters of the prior distribution were fixed and were not influenced by the data.

Huelsenbeck and Bollback (2001) reconstructed ancestral DNA sequences using the HKY85 substitution model (Hasegawa et al. 1985) with ASRV. Uncertainties in the tree topology and branch lengths, ASRV and substitution model parameters were considered. Prior distributions on model parameters are assumed. In the case of ASRV, the Bayesian method is hierarchical: the rate at each site is assumed to follow a gamma distribution with parameter α, and this α is assumed to be derived from a uniform distribution between zero and ten. Computing exact posterior probabilities is computationally infeasible, so the Markov chain Monte Carlo (MCMC) technique was used to efficiently approximate these probabilities (see Mau et al. 1999 for an example of applying MCMC for tree reconstruction).

Consider the goal of reconstructing the ancestral sequence at a specific node of a given tree. One problem arises: when integrating over the space of all possible trees, how trees in which this node does not exist should be considered in the MCMC computation? Huelsenbeck and Bollback (2001) considered only trees in which this node exists in their computation. However, Pagel et al. (2004) have shown that such a method of disregarding irrelevant trees introduces a bias in the estimation of the posterior probability. In essence, they suggest that the uncertainty in the existence of the node must also be taken into account in the ASR.

12. Gaps and Unknown Characters

All the models described thus far do not explicitly consider gapped positions and unknown characters. A common technique in phylogenetic tree reconstruction is to exclude from the analysis all positions in which at least one sequence contains a gap. However, for ASR, the goal is to infer the most likely ancestral sequence and thus, it is essential to determine whether a character or gap is the ancestral state.

One approximation to escape this problem is to consider a gap as a missing character (e.g., Pupko et al. 2000; Yang 1997). In this approach, all possible character states are considered at the gapped position, in the ASR computation. One problem with this approach is that the ancestral sequence is always longer or equal in length compared to the longest sequence – clearly an unrealistic result.

An alternative approach is to represent a gap by adding an additional character to the model (thus creating an alphabet of size 21 for amino acid, or 5 for DNA/RNA). Since gaps are considered as all other characters, the probability of a gap being replaced by any other character and vice versa must be determined. There are two main difficulties with this approach. First, the probabilities of such hypothetical transitions from each amino acid to a gap and vice versa are unknown. More importantly, this approach assumes independency among sites. Thus, an insertion of two residues will be considered

as two independent “character to gap transitions”, rather than the more parsimonious explanation of a single insertion of two amino-acids.

Clearly, it is more realistic to consider insertions and deletions of more than one residue as components of the evolutionary model. To this end, the tree-based HMM (T-HMM) scheme was developed (Mitchison and Durbin 1995; Mitchison 1999). In this approach three hidden states are allowed in each position: match, insertion, and deletion. Each “match state” emits a character. A Markovian substitution scheme between these hidden states is assumed in two dimensions: the spatial dimension across the sequence and the temporal dimension along evolutionary times. Qian and Goldstein (2003) have used this approach for finding remote homologs. To this end, they applied ASR methodology to build sequence profiles that were then used in the homology search. However, the evaluation of the impact of these models on ASR awaits further studies.

A different approach to deal with gaps was suggested by Edwards and Shields (2004). This algorithm first approximates the probabilities of gaps at each position and internal node, using a two-states-character model (0 for a gapped position, 1 for any other character). Once the ancestral state (0/1) for each node was determined, the non-gapped sites are estimated in an informal likelihood approach using probabilities derived from empirical substitution matrices. Although they show that their method is not as accurate as ML, their novelty is in dividing the ASR algorithm to two separate tasks: first reconstruct gapped versus non gapped positions and than use this reconstruction for ASR of un-gapped position.

13. Using Structural and Physicochemical Based Information When Reconstructing Ancestral Proteins

The different purifying selection resulting from different constraints on the structure, stability, foldability, and function of a protein should be considered as part of the evolutionary model. There are a few interesting directions towards this goal. For example it is possible to include information of secondary structures or surface accessibility in the model. This is done by considering amino acid substitution matrices for specific secondary structures or for buried versus exposed residues (these models are discussed in section 6).

Ideally, reconstructed proteins should be stable, foldable and functional. With current computational techniques, it is difficult to predict if this is the case for a reconstructed sequence. Towards this goal, biophysical model of protein evolution were developed that can be used to study the relationship between reconstructed ancestral proteins and their stability (DePristo et al. 2005; Rastogi and Liberles 2005; Taverna and Goldstein 2000; Taverna and Goldstein 2002). Recently, such a model was used to study the accuracy of various ASR methods. In a simulation study, the thermodynamic properties of the reconstructed sequence were compared with these properties in the “true” ancestral sequences (Williams et al. 2006). Such methods are likely to have impact on the detection of co-evolving substitutions, and as such on ASR, since non independent evolution of residues in a protein is a direct result from purifying selection forces acting to maintain interactions between amino-acid sites, and thus maintain protein stability and proper folding.

The ongoing effort to improve existing models for protein evolution, the endeavor to develop more realistic models, and the integration of these models with efficient ASR methods should increase our ability to accurately infer ancestral sequences and genomes, a vital element in evolutionary research.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-5

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459-68

Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348-58

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-402

Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19 Suppl 1:i7-15

Berglund AC, Steffansson P, Betts MJ, Liberles DA (2006) Optimal Gene Trees from Sequences and Species Trees Using a Soft Interpretation of Parsimony. J Mol Evol *in press*

Bevan RB, Lang BF, Bryant D (2005) Calculating the evolutionary rates of different genes: A fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Systematic Biology 54:900-915

Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88

Burnham KP, Anderson DR (2003) Model Selection and Multi-Model Inference. Springer, 2 edition

Chang BS, Campbell DL (2000) Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. Mol Biol Evol 17:1220-31

Chang BS, Jonsson K, Kazmi MA, Donoghue MJ, Sakmar TP (2002) Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol 19:1483-9

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. National Biomedical Research Foundation, Washington

DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6:678-87

Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21 Suppl 1:i126-35

Dimmic MW, Mindell DP, Goldstein RA (2000) Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput:18-29

Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22:1919-28

Edwards RJ, Shields DC (2004) GASP: Gapped Ancestral Sequence Prediction for proteins. BMC Bioinformatics 5:123

Eyre-Walker A (1998) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686-90

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-76

Felsenstein J (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. J Mol Evol 53:447-55
Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Inc. Sundeland, MA
Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
Fitch WM (1971) Toward Defining Course Of Evolution - Minimum Change For A Specific Tree Topology. Systematic Zoology 20:406-&
Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-93
Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-73
Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-21
Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871-9
Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315-21
Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285-8
Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. Syst Biol 49:652-70
Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-36
Gonnet GH, Benner SA (1991) Computational Biochemistry Research at ETH. ETH, Zurich, p 1-18
Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256:1443-5
Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862-4
Gu X (2003) Functional divergence in protein (family) sequence evolution. Genetica 118:133-41
Gu X, Fu YX, Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546-57
Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford
Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-74
Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275-84
Huelsenbeck JP, Bollback JP (2001) Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol 50:351-66

Jermann TM, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57-9
Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82
Jones DT, Taylor WR, Thornton JM (1994) A mutation data matrix for transmembrane proteins. FEBS Lett 339:269-75
Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Eng 8:641-5
Koshi JM, Goldstein RA (1996) Probabilistic reconstruction of ancestral protein sequences. J Mol Evol 42:313-20
Koshi JM, Goldstein RA (1998) Models of natural mutations including site heterogeneity. Proteins 32:289-95
Koshi JM, Mindell DP, Goldstein RA (1997) Beyond mutation matrices: physical-chemistry based evolutionary models. In: Miyano S, Takagi T (eds) Genome informatics. Universal Academy Press, Tokyo, p 80-89
Krishnan NM, Seligmann H, Stewart CB, De Koning AP, Pollock DD (2004) Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. Mol Biol Evol 21:1871-83
Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering Evolutionary Trees under a More Realistic Model of Sequence. Mol Biol Evol 11:605-612
Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1-7
Maddison WP, Donoghue MJ, Maddison DR (1984) Outgroup Analysis And Parsimony. Systematic Zoology 33:83-103
Mau B, Newton MA, Larget B (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1-12
Mayrose I, Friedman N, Pupko T (2005) A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21 Suppl 2:ii151-ii158
Mitchison G, Durbin R (1995) Tree-based maximal likelihood substitution matrices and hidden Markov models. J Mol Evol 41:1139-1151
Mitchison GJ (1999) A probabilistic treatment of phylogeny and sequence alignment. J Mol Evol 49:11-22
Muller T, Spang R, Vingron M (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 19:8-13
Muller T, Vingron M (2000) Modeling amino acid replacement. J Comput Biol 7:761-76
Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715-24
Nielsen R (1997) Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. Syst Biol 46:346-53
Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-36
Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673-84

Penny D, Hasegawa M (1997) Molecular systematics. The platypus put in its place. Nature 387:549-50
Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793-808
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002a) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 Suppl 1:S71-7
Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci 269:1313-6
Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M (2002b) Combining multiple data sets in a likelihood analysis: which models are the best? Mol Biol Evol 19:2294-307
Pupko T, Pe'er I, Hasegawa M, Graur D, Friedman N (2002c) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. Bioinformatics 18:1116-23
Pupko T, Pe'er I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17:890-6
Pupko T, Sharan R, Hasegawa M, Shamir R, Graur D (2003) Detecting excess radical replacements in phylogenetic trees. Gene 319:127-35
Qian B, Goldstein RA (2003) Detecting distant homologs using phylogenetic tree-based HMMs. Proteins 52:446-53
Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28
Sankoff D (1975) Minimal Mutation Trees Of Sequences. Siam Journal On Applied Mathematics 28:35-42
Sankoff D, Rousseau P (1975) Locating Vertices Of A Steiner Tree In An Arbitrary Metric Space. Mathematical Programming 9:240-246
Schluter D (1995) Uncertainty in ancient phylogenies. Nature 377:108-10
Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6:134
Schneider A, Gonnet GH, Cannarozzi GM (2006) Synonymous codon substitution matrices. ICCS 2006. LNCS 3992: 630-637
Schultz TR, Churchill GA (1999) The role of subjectivity in reconstructing ancestral character states: A Bayesian approach to unknown rates, states, and transformation asymmetries. Systematic Biology 48:651-664
Stewart CB (1995) Molecular evolution. Active ancestral molecules. Nature 374:12-3
Sullivan J, Swofford DL (1997) Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. J Mammal Evol 4:77-86
Swofford DL, Maddison WP (1987) Reconstructing Ancestral Character States Under Wagner Parsimony. Mathematical Biosciences 87:199-229
Taverna DM, Goldstein RA (2000) The distribution of structures in evolving protein populations. Biopolymers 53:1-8

Taverna DM, Goldstein RA (2002) Why are proteins so robust to site mutations? J Mol Biol 315:479-84
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-80
Thornton JW (2001) Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proc Natl Acad Sci U S A 98:5671-6
Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089-96
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 18:691-9
Whelan S, Lio P, Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 17:262-72
Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. Plos Computational Biology 2:598-605
Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396-401
Yang Z (1996) Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. J Mol Evol 42:587-96
Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-6
Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316-24
Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-50
Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-49
Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-11
Yap VB, Speed T (2005) Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol Biol 5:2