Adaptive Distance Measures for Resolving K2P Quartets:

Metric Separation versus Stochastic Noise

Ilan Gronau Shlomo Moran Irad Yavneh

February 10, 2010

Abstract

Distance based phylogenetic reconstruction methods use the *evolutionary distances* between species in order to reconstruct the tree spanning them. The evolutionary distance between two species, which is computed from their DNA (or protein) sequences, is typically considered as a fixed function of these sequences, predetermined by the assumed model of evolution. This paper continues the line of research which attempts to adjust to each given set of input sequences a distance function which maximizes the expected accuracy of the reconstructed tree. Specifically, we present methods for selecting distance functions that considerably improve the accuracy of quartets constructed by the four point method in Kimura's 2 parameter model, where special emphasis is given to the case of non-homogenous quartets.

1 Introduction.

Distance based methods are among the most common approaches for reconstructing evolutionary trees (and the only ones guaranteed to run in polynomial time) [Saitou and Nei 1987, Gascuel 1997, Bruno et al. 2000]. These methods use inter-species *evolutionary distances*, which are computed from their DNA (or protein) sequences, in order to reconstruct the tree. Inter-species distance estimation usually assumes some stochastic model of evolution. Common models of evolution are characterized by the distinct types of substitutions they assume. For example, Kimura's 2 parameter (K2P) model [Kimura 1980] studied here distinguishes between two types of mutations: transitions and transversions (see Section 2.1). Evolutionary distances in these models are usually defined by the *total substitution rate*, which is the expected number of substitutions per site. Thus, most known reconstruction methods treat the expected accuracy of their input distance estimates as some fixed factor [Erdos et al. 1999, Atteson 1999, Gascuel 1997].

This paper continues the new approach for evolutionary distances introduced in [Gronau et al. 2009]. Rather than treating evolutionary distances as predetermined entities defined by the total substitution rate, this approach adjusts to each given set of input sequences a *substitution rate* (SR) function which maximizes the expected accuracy of the reconstructed tree. Roughly speaking, an SR function (defined in Section 2.1) assigns weights to the different types of substitutions. The total substitution rate, for example, is an SR function which assigns uniform weights to all types of substitutions.

In this paper we focus on the problem of selecting an SR function for the purpose of quartet reconstruction. Specifically, we present strategies for selecting SR functions that considerably improve the performance of the *four point method* (FPM), which is essentially the only distance based algorithm for quartets [Sattath and Tversky 1977, Erdos et al. 1999]. The expected accuracy of FPM in resolving a quartet is affected by the following two factors (see [Atteson 1999, Erdos et al. 1999]):

Metric separation: the ratio between the weight (or length) of the internal edge of the quartet and the weights of its external edges.

Estimation error: the expected difference between the estimated distances and the true substitution rates.

Intuitively, we would like to adapt to each input quartet an SR function which increases its metric separation and decreases the estimation errors of its interspecies distances.

The process of site substitution is often assumed to be *homogenous* [Lio and Goldman 1998], which in the K2P model implies that the ratio between transitions and transversions (ti/tv) remains constant throughout the tree. In homogeneous trees, all SR functions are *metric equivalent* in the sense that the metrics they induce are proportional to each other. Hence, they all imply the same metric separation. However, they may differ (even very much) in their expected estimation error. This fact is demonstrated in the preliminary experimental study on reconstruction of homogeneous K2P quartets presented in [Gronau et al. 2009]. A strategy which chooses an SR function according to the (relative) estimation error is shown to significantly outperform the standard SR functions in this model, despite the fact that all SR functions are metric equivalent.

The situation is more involved when the substitution process is not homogenous, since in this case different SR functions may imply non-proportional metrics. To illustrate this, consider a non-homogenous K2P quartet, in which the ti/tv ratio of the internal edge e is 10, while the ti/tv ratio of all external edges is 1. Then the weight assigned to e by the SR function which counts only transitions is 10 times larger (relative to the weights of the external edges) than its weight according to the SR function which counts only transversions. Thus, even if the estimation error in counting transitions is twice larger, say, than the estimation error in counting transversions, the four point method is still more likely to recover the correct split by using the SR function which counts only transitions than by using the one counting only transversions.

Since SR functions which reduce the estimation error for a given quartet do not necessarily increase the weight of its internal edge (and vice versa), it is plausible that a good strategy for selecting an SR function should take into account both metric separation and estimation error. In this paper we propose several adaptive strategies and test their performance on homogeneous and non-homogeneous K2P quartets. The performance of these strategies is compared to two standard constant strategies: one which always chooses the SR function that counts the total number of substitutions (introduced originally in [Kimura 1980]), and one which always chooses the SR function that counts only transversions (implied by the CFN binary substitution model [Cavender 1978, Farris 1973, Neymann 1971]). Our comparison shows that the adaptive strategies generally outperform the constant ones, reducing the error rate in reconstruction by 50% or more in some cases. The results also indicate that unless the quartet is highly non-homogeneous, considering only estimation error results in reasonable reconstruction accuracy.

2 Background.

2.1 Substitution Rate Functions for the K2P Model.

Kimura's 2 Parameter model (K2P) is a rather simple DNA substitution model, which assumes that all *transition*-type (ti) substitutions (A $\leftrightarrow$ G, C $\leftrightarrow$ T) have the same rate, designated by α, and all *transversion*-type (tv) substitutions ({A,G} $\leftrightarrow$ {C,T}) have the same rate, designated by β. It is typically assumed also that $\alpha > \beta$, since transitions are more prevalent than transversions. The rate parameters α, β associated with a specific evolutionary path indicate the expected number of substitution events of each type along the path. Assuming a uniform distribution on the bases (which is a standard assumption for the K2P model), then along a path with parameters (α, β) there will be on average α ti substitutions and 2β tv substitutions. The ratio $\frac{\alpha}{2\beta}$ is called the ti/tv ratio and is often denoted by R.

A K2P model tree T is an unrooted phylogenetic tree in which each edge (u, v) is assigned rate parameters α_{uv}, β_{uv}. The model tree is said to be *homogeneous* if the ti/tv ratio is constant throughout the tree. The underlying assumption behind homogeneity is that proportion between the various substitution rates is dictated by bio-chemical properties which remain constant throughout time. Although it is a common assumption, there are several clear non-homogeneous cases documented in the literature (see e.g., [Hendy et al. 1994, Herbeck et al. 2005]). The edge rate parameters provided by the model tree imply rates for every path in the tree. Denote by α_{uv}, β_{uv}, the rates corresponding to the path connecting u and v in T. Then rates are additive in the following sense: if w lies on the path connecting u and v in T, then $\alpha_{uv} = \alpha_{uw} + \alpha_{wv}$ and $\beta_{uv} = \beta_{uw} + \beta_{wv}$. This additivity simply follows from linearity of expectation.

Additive metrics play a central role in distance-based reconstruction [Sattath and Tversky 1977]. An *additive metric* on the model tree T is a metric D over the vertices of T which obeys the following property: for every three vertices u, v, w of T s.t. w lies on the path connecting u and v, $d(u, v) = d(u, w) + d(w, v)$. We are specifically interested in additive metrics which are functions of the rate parameters, i.e., $D : V(T) \times V(T) \to \mathbb{R}^+$, s.t. $d(u, v) = f(\alpha_{uv}, \beta_{uv})$. A function f which transforms rates into additive distances is called a *substitution rate function* (SR function, in short - see [Gronau et al. 2009]). Due to the additivity of rates, any positive linear combination of the rates $f(\alpha, \beta) = c_\alpha \alpha + c_\beta \beta$ is a valid SR function[1]. The standard metric considered for the K2P model, which counts the total expected number of substitutions, uses the following SR function: $f(\alpha, \beta) = \alpha + 2\beta$. Two other natural examples of SR functions are the one which counts only transitions, $f(\alpha, \beta) = \alpha$, and the one which counts only transversions, $f(\alpha, \beta) = 2\beta$.

2.2 Obtaining Pairwise Distance Estimates.

The first step in distance based reconstruction is to obtain estimates of inter-taxon distances induced by some additive metric on the model tree. This estimation process strongly relies on our ability to estimate the rate parameters α_{ij}, β_{ij} for every taxon pair i, j, just by observing the two sequences S_i, S_j. The statistical estimators $\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}$ of these parameters are computed by applying the following formulas [Kimura 1980]:

$$\widehat{\alpha}_{ij} = -\frac{1}{2}\ln(1 - Q_{ij} - 2P_{ij}) + \frac{1}{4}\ln(1 - 2Q_{ij}) \quad ; \quad \widehat{\beta}_{ij} = -\frac{1}{4}\ln(1 - 2Q_{ij}) \ , \tag{1}$$

where P_{ij} denotes the fraction of transition-type differences between S_i and S_j and Q_{ij} denotes the fraction of transversion-type differences[2]. If the terms in the logarithms of Equation 1 are positive, then the above expressions compute the maximum-likelihood estimates of α_{ij}, β_{ij}. On the other hand, if one of the terms is negative or zero, then the path is said to be *saturated*. It is common to consider such cases as a failure to estimate the distance.

We refer to $\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}$ as the *observed rates* for the path connecting i and j in T. Distance estimates are computed from the observed rates simply by applying the appropriate SR function f, i.e.: $\widehat{d}_{ij} = f(\widehat{\alpha}_{ij}, \widehat{\beta}_{ij})$ is an estimate of the "true" additive distance $d(i, j) = f(\alpha_{ij}, \beta_{ij})$. As mentioned earlier, the standard distance estimation formula for K2P suggested in [Kimura 1980] uses the 'total rate' SR function, $\widehat{d}_{ij} = \widehat{\alpha}_{ij} + 2\widehat{\beta}_{ij}$. In this study we show that a conscious choice of SR function based on the observed data (rather than a blind a-priori choice) can lead to a significant improvement in accuracy of reconstruction.

2.3 The Four-Point Method for Quartet Reconstruction.

A quartet tree is a phylogenetic tree over 4 taxa (denoted here by $1, 2, 3, 4$). It is common to assume that this tree is not a star tree, in which case it corresponds to one of the following *splits*:

[1] In [Gronau et al. 2009] we show that *all* SR functions for the K2P model are of this form.

[2] In [Gronau et al. 2009], P and Q are denoted by by $\widehat{p_\alpha}$ and $2\widehat{p_\beta}$, respectively.

(1 2 | 3 4) , (1 3 | 2 4) , (1 4 | 2 3). Unlike reconstructing larger trees, reconstructing quartet trees from pairwise distances is a rather straightforward task. In fact, all known reconstruction algorithms are reduced to the same algorithm when restricted to a single quartet. This algorithm is often referred to as the *four-point method* (FPM) (see, e.g., [Buneman 1971, Sattath and Tversky 1977, Erdos et al. 1999]). FPM is based on the following well known fact [Sattath and Tversky 1977]: if D is an additive metric on the quartet tree, then the quartet split is $(ij|kl)$ (where $\{i,j,k,l\} = \{1,2,3,4\}$) if and only if $d_{ij} + d_{kl} < d_{ik} + d_{jl} = d_{il} + d_{jk}$. Hence, FPM determines the split by computing the three sums of estimated distances, $\widehat{d}_{12} + \widehat{d}_{34}$, $\widehat{d}_{13} + \widehat{d}_{24}$, $\widehat{d}_{14} + \widehat{d}_{23}$, and returning the split $(ij|kl)$ corresponding to the minimal sum $\widehat{d}_{ij} + \widehat{d}_{kl}$ (if there is a tie for the minimum, then FPM returns "fail").

3 Strategies for choosing an SR Function.

In this paper we explore strategies for choosing SR functions, where the primary objective is to increase the accuracy of reconstruction. The practical goal is to design strategies which choose SR functions according to the *observed* taxon sequences (or the observed rates). However, it is useful to consider first the problem of fitting an SR function to the model tree rather than to the observed input. We identify two main factors which should be taken into account when choosing an SR function for a given model tree: the metric induced by the SR function on the model tree, and the expected error in distance estimation. Specifically, we want the stochastic estimation error to be relatively small compared to the weights of internal edges (see [Atteson 1999, Erdos et al. 1999] for a more formal discussion).

An important observation is that proportional SR functions are *equivalent* in the sense that they induce proportional distance estimates, and hence lead to the same reconstructed topology. Therefore, our strategies are designed to be *scale-free*, meaning that they do not differentiate between proportional SR functions. For this reason, when taking metric considerations into account, we consider the *normalized metric* induced by an SR function on the model tree.

Definition 3.1 (Normalized Metric)**.** *The* normalized metric *induced by an SR function f on a model tree T is the metric proportional to the additive metric induced by f on T, in which the diameter (the maximal distance) is* 1.

Metric considerations essentially differentiate between SR functions which have different normalized metrics. As mentioned in the introduction, when the model tree is homogeneous, all SR functions are *metric-equivalent*, i.e. their normalized metrics are identical. This is because the ti/tv ratio R is constant throughout the tree, implying that the α-metric induced on the model tree is $2R$-times larger than the β-metric. Therefore, for homogeneous model trees, metric considerations are moot. However, when the model tree is non-homogeneous, these considerations may be quite important. When considering quartet trees, for instance, we intuitively prefer SR functions which give a high ratio between the weight of the internal edge and the weights of external edges. Experimental results presented in Section 4 demonstrate that even in rather simple non-homogeneous quartet trees, metric considerations can be crucial.

Distance estimation error is rooted in the fact that the observed rates $\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}$, are noisy estimates of the rate parameters in the actual model tree. This estimation error and its propagation to distance estimation error are studied in [Gronau et al. 2009]. The estimation error implied by an SR function on a given path in the model tree is measured by the *mean square error* (MSE), $\mathsf{E}\left[(\widehat{d} - d)^2\right]$, where $\widehat{d}$ is the estimated distance (a random variable) and d is the actual additive distance associated with that path. In [Gronau et al. 2009] we developed a closed-form approximation formula for $\mathsf{MSE}[\widehat{d}]$, employing the delta method [Oehlert 1992]. This approximation formula expresses $\mathsf{MSE}[\widehat{d}]$ as a function of the sequence length k, the actual rate parameters of the path α, β, and the SR

coefficients c_α, c_β used in obtaining $\widehat{d}$:

$$\mathsf{MSE}[\widehat{d}] \;=\; \mathsf{E}\left[(\widehat{d}-d)^2\right] \;\approx \tag{2}$$
$$\tfrac{1}{16k}\left(c_\alpha^2\left((e^{4\beta}-1)^2+2(e^{4\alpha}-1)(e^{4\beta}+1)\right) \;-\; 2c_\alpha c_\beta(e^{4\beta}-1)^2 \;+\; c_\beta^2(e^{8\beta}-1)\right) .$$

Typically, a normalized version of this noise estimate is used – one which depends only on the ratio $\frac{c_\alpha}{c_\beta}$, rather than on the actual scale of c_α and c_β (see [Gronau et al. 2009] and Section 3.1 below). The approximation of Equation 2 is shown in [Gronau et al. 2009] to be very tight as long as the rates α, β are not too large. Moreover, it is shown that substituting the observed rates $\widehat{\alpha}, \widehat{\beta}$, for α, β, also results in an accurate estimate of the error (on average). This allows us to easily employ noise considerations with respect to the observed data, which is not always straightforward when it comes to metric considerations.

3.1 Strategies for Quartet Reconstruction.

In order to explore the effect of metric factors and noise factors, we focus on the problem of quartet reconstruction. A strategy for choosing an SR function in such a case maps the twelve observed rates $\{\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}\}_{\{i,j\}\subset\{1,2,3,4\}}$ onto a pair of SR coefficients c_α, c_β. Since, as observed earlier, proportional SR functions are equivalent, our strategies consider only SR coefficients $c_\alpha, c_\beta \geq 0$ s.t. $c_\alpha + c_\beta = 1$. We test three strategies: one based on metric considerations, one based on noise considerations and one based on combined metric and noise considerations. Each of the three strategies scores the various SR functions according to one of the following terms, which depends on the observed rates as well as the SR function itself:

Relative separation: assume a labelling i, j, k, l of the 4 taxa s.t. $\widehat{d}_{ij} + \widehat{d}_{kl} \;\leq\; \widehat{d}_{ik} + \widehat{d}_{jl} \;\leq\; \widehat{d}_{il} + \widehat{d}_{jk}$, then the relative separation is defined as –

$$\frac{\frac{1}{2}\left((\widehat{d}_{ik} + \widehat{d}_{jl}) \;-\; (\widehat{d}_{ij} + \widehat{d}_{kl})\right)}{(\widehat{d}_{ij} + \widehat{d}_{kl})} .$$

Average Relative Error:

$$\frac{1}{6}\sum_{i,j}\frac{\mathsf{MSE}[\widehat{d}_{ij}]}{\widehat{d}_{ij}^2} .$$

Distinguishability: the separation squared divided by the average MSE:

$$\frac{\left((\widehat{d}_{ik} + \widehat{d}_{jl}) \;-\; (\widehat{d}_{ij} + \widehat{d}_{kl})\right)^2}{\frac{1}{6}\sum_{i,j}\mathsf{MSE}[\widehat{d}_{ij}]} .$$

$\mathsf{MSE}[\widehat{d}_{ij}]$ is approximated by substituting $\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}$ in Equation 2.

Our *metric strategy* chooses the SR function which maximizes the relative separation, completely disregarding the expected error in distance estimation. Note that, assuming that the distance estimates are accurate, the relative separation equals the ratio between the weight of the internal edge and the weight-sum of all external edges. Our *noise strategy* chooses the SR function which minimizes the average relative error, disregarding metric separation altogether. The third is a *combined strategy* which chooses the SR function which maximizes the distinguishability. Distinguishability takes both metric and noise considerations into account. Note that all three strategies are scale-free in the sense that they give proportional SR functions the same score.

4 Experimental Results.

4.1 The Experimental Setting.

We conducted experiments on simulated data in order to evaluate the three proposed strategies. Each experiment estimates the reconstruction error rates of each strategy on a specific K2P model quartet. Each experiment consists, therefore, of a predetermined number of simulations (5000 in the results shown here) of site substitutions along a DNA sequence of predetermined length (500 in the results shown here). Sequence evolution in each simulation propagates along the model quartet according to its preset rate parameters. The simulation results in four taxon sequences, from which the observed rates $\{\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}\}_{\{i,j\}\subset\{1,2,3,4\}}$ are extracted according to Equation 1[3].

After obtaining the observed rates, the various strategies are applied to choose an SR function. Recall that each of the three strategies chooses an SR function which maximizes or minimizes some measure. Finding this SR function is done by scanning values for c_α in the interval $[0, 1]$ in steps of 0.01 (where $c_\beta = 1 - c_\alpha$). After the SR function is chosen and distances are computed according to it, the split is reconstructed using FPM and compared to the true split of the model quartet. Once all the simulations are completed, the error rates of each strategy are recorded. These error rates correspond to the number of times FPM failed to return the correct split divided by the total number of simulations. In addition to the error rates of the three *adaptive* strategies mentioned in Section 3.1, the error rates for two standard *constant* strategies are also recorded: the strategy which always chooses the total substitution rate $f_{\text{total\_rate}}(\alpha, \beta) = \alpha + 2\beta$, and the strategy which always chooses the transversion count $f_{tv\_only}(\alpha, \beta) = 2\beta$.

In order to obtain some lower-bound estimate on the reconstruction error rate of each model quartet, we compute the SR function which *a-posteriori* leads to the lowest error rate on that quartet across the simulations done in the experiment. This is done by scanning values for c_α in the interval $[0, 1]$ in steps of 0.01, and choosing the coefficient which leads to the minimal number of reconstruction failures across the simulations of that experiment. As we do not know how to compute this SR function directly from the parameters α, β (without running all the simulations on the actual model quartet), we refer to it as the *oracle's best* SR function. The oracle's best SR function does not necessarily imply what makes a good SR function, but its error rate provides an estimate on the best performance of any strategy.

The experiments focus on model quartets in which all the external edges share the same ti/tv ratio (R_{ext}), but the ti/tv ratio of the internal edge (R_{int}) is allowed to be different. This simple pattern of model quartets allows us to carefully investigate the effect of introducing non-homogeneity into the model. Note that in such quartets, the best metric separation is obtained by one of the extreme SR functions corresponding either to $c_\alpha = 0$ or to $c_\beta = 0$. For instance, assume that $R_{int} < R_{ext}$, and let $\alpha_{int}, \beta_{int}$, denote the rate parameters of the internal edge and let $\alpha_{ext}, \beta_{ext}$, denote the rate parameters of some external edge. Now, since $R_{int} < R_{ext}$, then $\frac{\alpha_{int}}{\alpha_{out}} < \frac{\beta_{int}}{\beta_{out}}$. In other words, the ratio between the weight of the internal edge and the weight of any external edge is smaller under the α-metric than it is under the β-metric. Furthermore, any non-trivial convex combination of these metrics also implies a smaller ratio. So $c_\alpha = 0$ is the SR coefficient leading to optimal quartet metric. Similarly, if $R_{int} > R_{ext}$, then $c_\beta = 0$ is the SR coefficient leading to optimal quartet metric. Obviously, if $R_{int} = R_{ext}$, then the quartet is homogeneous, in which case all SR functions imply the same normalized metric.

4.2 Results.

The results presented here correspond to symmetric model quartets in which all external edges are identical in rate parameters. This is done for simplicity of presentation; similar results were obtained

[3]If one of the paths is saturated (i.e., one of the expressions in the logarithms of Equation 1 is negative), this simulation is discarded and replaced with a valid one (in which distances can be estimated). Saturation happens only when α is very large.

for non-symmetric quartets. A symmetric model quartet is defined by 4 parameters: L_{int}, R_{int}, which denote the total rate ($\alpha + 2\beta$) and the ti/tv ratio ($\frac{\alpha}{2\beta}$) of the internal edge, and L_{ext}, R_{ext}, which denote the total rate and the ti/tv ratio of all external edges. The results of four batches of experiments are presented. In each batch one parameter of the model quartet is varied in some range while the other parameters remain constant. The X-axis of each graph indicates the value of the variable parameter, and the Y-axis corresponds to the error rates. Recall that all experiments consist of 5000 simulations of sequence evolution on sequences of length 500.

The first batch of experiments (Fig. 1) considers *homogeneous* model quartets in which $R_{ext} = R_{int} = 2$ and $L_{ext} = 5L_{int}$, where the scale of the quartet is changed by varying L_{ext} in the interval $[0.05, 1]$. In this scenario the noise strategy emerges as the best one, as its error rate is clearly lower than that of the other strategies and is actually very close to that of the oracle's best SR function. The combined strategy has very good performance as well, beating the two constant strategies in nearly all cases. The metric strategy, on the other hand, has quite high error rates, specifically when the quartet becomes longer. This is due to two factors. First, since the quartet is homogeneous, metric considerations do no really apply (although preferring high separation in the observed distances still make sense). Second, as the quartet becomes longer, the estimation error of $\widehat{\alpha}_{ij}, \widehat{\beta}_{ij}$ increases, and error becomes a crucial factor. Fig. 1 also supports an observation in [Gronau et al. 2009], that the 'total rate' SR function estimates rather accurately the distances along short paths (where α, β are both short), whereas the 'tv-only' SR function has high accuracy along long paths (assuming $\beta < \alpha$).

The second batch of experiments (Fig. 2) explores what happens when heterogeneity is introduced into the model. A quartet of intermediate scale is taken, $L_{ext} = 0.5, L_{int} = 0.1$, where $R_{ext} = 2$ remains constant and R_{int} varies in the interval $[0.5, 11]$. This batch demonstrates that even when the quartet is non-homogeneous, the noise strategy leads to very good performance. However, in the extreme case where $R_{int} \gg R_{ext}$ metric considerations become important: both the metric and the combined strategies outperform the noise strategy, and the total rate SR function becomes near-optimal (because, when $R_{int} > R_{ext}$, this function implies high metric separation on the model quartet). Another interesting observation about this graph is that the 'total rate' SR function has near constant error rates throughout the various experiments. This is because the metric it induces on the model quartet does not change throughout the experiments (since L_{int}, L_{ext} are constant), and the changes made to the internal edge do not have a high influence on the estimation errors of the observed rates.

The first two batches clearly indicate the importance of noise considerations when choosing an SR function. However, metric considerations are definitely crucial in some cases. The third batch of experiments (Fig. 3) explores a series of small scale quartets in which $L_{ext} = 0.2, L_{int} = 0.04$. $R_{ext} = 10$ is constant in this figure, and is larger or equal to R_{int} which grows from 0.5 to 10. In this case, 'tv-only' SR function induces the best metric separation on the model quartet (since $R_{int} < R_{ext}$). However, only in the extreme case where $R_{int} \ll R_{ext}$ does this SR function outperform the adaptive metric and combined strategies. Because this batch of experiments considers short model quartets (with low substitution rates), the error in estimation of the observed rates is small and does not play an important part in choosing a good SR function.

The fourth and last batch of experiments demonstrates that metric considerations can be crucial also when $R_{int} > R_{ext}$. This batch explores model quartets where $L_{ext} = 0.6, L_{int} = 0.3$ and $R_{int} \leq R_{ext}$: $R_{ext} = 2$ is constant and R_{int} varies in the interval $[2, 10]$ (Fig. 4). The increased ratio between L_{int} and L_{ext} (compared to previous experiment batches) amplifies the influence of the metric considerations. In this case, noise still plays an important role (since the quartet is not short), which is why the noise strategy has reasonable performance when R_{int} is not too large. However, as the quartet becomes less homogeneous (by increasing R_{int}), the metric strategy outperforms the noise strategy, and eventually also the combined strategy.

5 Concluding Remarks.

The experimental results presented in Section 4 may be summarized by the following general observations:

1. All three adaptive strategies proposed are in general superior to the two constant ones; they result in as much as 50% lower error rates in many cases.
2. In many scenarios, the 'noise' strategy is superior to the 'metric' strategy – even for rather non-homogeneous quartets.
3. The combined strategy seems to be the only one of the three which never gives much higher error rates than the oracle's best lower bound.

This line of research opens up several directions for further research. The combined strategy suggested here performs very well, but it is often outperformed by one of the other two. This implies the need to find a better combination of metric vs. noise considerations. Another problem is to develop strategies which choose an SR function using a closed-form formula, rather than searching for it using some optimization criterion. Although this search takes very little time, a closed-form formula has obvious advantages. A practical task related to this work is to use the approaches suggested here in reconstructing larger trees (spanning many taxa). A first step in that direction is to use quartet based algorithms (see, e.g., [Strimmer and von Haeseler 1996]), whose input is quartets resolved using the strategies suggested here.

References

[Atteson 1999] Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25, 251–278.

[Bruno et al. 2000] Bruno, W., Socci, N., Halpern, A., 2000. Weighted Neighbor Joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17 (1), 189–197.

[Buneman 1971] Buneman, P., 1971. The recovery of trees from measures of dissimilarity. Mathematics in the Archeological and Historical Sciences, 387–395.

[Cavender 1978] Cavender, J., 1978. Taxonomy with confidence. Math Biosci 40, 271–280.

[Erdos et al. 1999] Erdos, P., Steel, M., Szekely, L., Warnow, T., 1999. A few logs suffice to build (almost) all trees (I). Random Structures and Algorithms 14, 153–184.

[Farris 1973] Farris, J., 1973. A probability model for inferring evolutionary trees. Systematic Zoology 22, 250–256.

[Gascuel 1997] Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14 (7), 685–695.

[Gronau et al. 2009] Gronau, I., Moran, S., Yavneh, I., 2009. Towards optimal distance functions for stochastic substitution models. J Theor Biol 260 (2), 294–307.

[Hendy et al. 1994] Hendy, U., Penny, D., Steel, M., april 1994. A discrete Fourier analysis for evolutionary trees. Proc Natl Acad Sci 91 (8), 3339–3343.

[Herbeck et al. 2005] Herbeck, J. T., Degnan, P. H., Wernegreen, J. J., 2005. Nonhomogeneous Model of Sequence Evolution Indicates Independent Origins of Primary Endosymbionts Within the Enterobacteriales (gamma-Proteobacteria). Mol Biol Evol 22 (3), 520–532.

[Kimura 1980] Kimura, M., Dec. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16 (2), 111–120.

[Lio and Goldman 1998] Lio, P., Goldman, N., dec 1998. Models of molecular evolution and phylogeny. Genome Research 8 (12), 1233–1244.

[Neymann 1971] Neymann, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S., Jackel, Y. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, pp. 1–27.

[Oehlert 1992] Oehlert, G., 1992. A note on the delta method. The American Statistician 46 (1), 27–29.

[Saitou and Nei 1987] Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406–425.

[Sattath and Tversky 1977] Sattath, S., Tversky, A., 1977. Additive similarity trees. Psychometrica 42 (3), 319–345.

[Strimmer and von Haeseler 1996] Strimmer, K., von Haeseler, A., 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Molecular Biology and Evolution 13, 964–969.