Estimating the Deviation From a Molecular Clock

Luay Nakhleh[1], Usman Roshan[1], Lisa Vawter[2], and Tandy Warnow[1]

[1] Department of Computer Sciences, University of Texas, Austin, TX 78712;
{nakhleh,usman,tandy}@cs.utexas.edu
[2] Informatics Department, Aventis Pharmaceuticals, Bridgewater, NJ 08807-0800;
lisa.vawter@aventis.com

Abstract. We address the problem of estimating the degree to which the evolutionary history of a set of molecular sequences violates the strong molecular clock hypothesis. We quantify this deviation formally, by defining the "stretch" of a model tree, with respect to the underlying ultrametric tree (indicated by time). We then define the "minimum stretch" of a dataset on a tree, and show how this can be computed optimally in polynomial time. We also present a polynomial time algorithm for computing a lower bound on the stretch of a given dataset on any tree. We then explore the performance of standard techniques in systematics for estimating the deviation of a dataset from a molecular clock. We show that standard methods, whether based upon maximum parsimony or maximum likelihood, can return infeasible values (i.e. values for the stretch which cannot be realized on any model tree), and often under-estimate the true stretch. This suggests that current estimations of the degree to which datasets deviate from a molecular clock may significantly underestimate these deviations. We conclude with some suggestions for further research.

1 Introduction.

A phylogenetic tree is a rooted tree in which the leaves represent the given set of taxa (species, DNA sequences, etc.), and the internal nodes represent the ancestral taxa. The inference of these phylogenetic trees plays a role in many aspects of biological research, including drug design, the understanding of human migrations, the origins of life, etc.

Most phylogenetic methods produce unrooted trees (or produce rooted trees whose roots are unreliable). While in some applications the topology of the unrooted phylogenetic tree is sufficient, for most applications the rooted tree is desirable. For example, the famous *African Eve* study assumed that the location of the root was reliable, and used that location in order to infer that humans evolved out of Africa. Yet rooting a phylogenetic tree is often difficult to do. The most reliable technique seems to be to use an outgroup (a taxon that should attach to the true tree by an edge off the root); yet if the taxon is too closely related it may not be an outgroup, and if it is too distantly related, it may be difficult to reconstruct the location of the attachment to the remainder of the phylogeny,

due to changes of character states over time which result in the distantly related taxon looking essentially random with respect to the remaining taxa.

Other approaches for locating the root assume that the dataset is evolving via a (roughly) molecular clock, which asserts that the expected number of times a random site will change in t time units is (roughly) proportional to t. The assumption that a strong molecular clock underlies the data can be tested through the log-likelihood ratio test [4], but there are no tests for estimating the degree of deviation from a strong molecular clock. Furthermore, the log-likelihood ratio test is computationally intensive if used appropriately, as it should be used with an exact method for finding the maximum likelihood tree (a computationally intractable problem).

The accuracy (or lack thereof) of the molecular clock hypothesis is of significant interest to biologists (see [19, 2, 8, 14, 13, 20, 5, 1] for just a few of the papers that address this question). One of the reasons for this interest is that datasets that conform closely to a molecular clock can be analyzed for times at which speciation (or gene duplication) events occurred, thus enabling a more fine-grained analysis of the molecular processes in the dataset.

In this paper we present a formal definition of the deviation from the molecular clock in a dataset on a tree, which we call the *stretch*. We then present two $O(n^2)$ algorithms: the first computes the optimal stretch of a given tree for a given dataset, and the second computes a tree with the optimal stretch for a given dataset. Furthermore, we describe methods that biologists use for computing the deviation of a dataset from the molecular clock, and provide an empirical evidence which shows that the values obtained by using those methods may be infeasible.

2 Background and Definitions.

We define the terms that are used in the biological literature which pertain to this paper.

Definition 1. *A* **phylogenetic tree** *for a set S of taxa is a rooted tree whose leaves are labeled by the taxa in S, and whose internal nodes represent the (hypothetical) ancestors of the taxa.*

Phylogenetic trees represent the evolutionary history of sets of taxa (genes, species, etc.). If the taxa under consideration have evolved at equal rates from a common ancestor (at the root of the tree), then the number of evolutionary events on every root-to-leaf path in the model tree will tend to be approximately equal. This is the "strong molecular clock" assumption. More formally, it implies the following. If we weight the edges of the model tree by the expected number of

times a random site changes on the edge, then the lengths of all root-to-leaf paths in the model tree will be the same. In other words, the model tree is ultrametric.

Definition 2. *An edge-weighted tree T is called* **ultrametric** *if it can be rooted in such a way that the lengths of all the root-leaf paths in the tree are equal. An ultrametric distance matrix D is the matrix of leaf-to-leaf distances in an ultrametric tree.*

One way of estimating edge weights for a tree is to let $t(e)$ equal the time indicated by the edge e, so that the weight on edge e is given by $\lambda(e) = p \cdot t(e)$, where p is the expected number of events per unit time. Similarly, if $t_{i,j}$ is the time since i and j diverged from a common ancestor (i.e. $2t_{i,j} = \sum_{e \in P_{ij}} t(e)$, where $P_{i,j}$ is the path in T between i and j), then $D_{ij} = 2pt_{i,j}$ is an ultrametric matrix.

Note that given a rooted ultrametric tree (and hence its matrix U), we can assign "heights" to the nodes of the tree as follows. Given a node v, let i and j be two leaves in the tree below v. Set $height(v) = U_{i,j}$. By construction, this is well-defined, and if v and w are nodes in T such that v is on the path from w towards the root (i.e. $v > w$), then $height(v) \geq height(w)$. Hence, an alternative way of defining an ultrametric matrix is as follows:

Definition 3. *A matrix D is an ultrametric matrix if there exists a rooted tree T with function $height(v) \geq 0$ defined for each node $v \in V(T)$, such that whenever $v > w$ (i.e. v is on the path from w to the root of T) we have $height(v) > height(w)$.*

Given a rooted tree T with heights assigned to the nodes, we can then compute the distance $D_{i,j}$ between any two leaves i and j as follows:

$$D_{i,j} = height(lca_T(i,j)),$$

where $lca_T(i,j)$ denotes the most recent common ancestor of i and j. For most stochastic models of evolution it is possible to estimate model distances in a statistically consistent manner (see [10]); this means that the estimates of model distances converge to the model distances as the lengths of the sequences generated on the tree increase.

This has the following consequence for estimating the deviation from a molecular clock: Suppose we are given sequences generated by a stochastic process operating on a model tree, and we apply a statistically consistent estimator for the pairwise distances, thus obtaining a matrix $\{d_{ij}\}$ of distances. If the matrix d is exactly correct (i.e. d is the model distance matrix), then we can set times at the internal nodes, thus defining an ultrametric matrix U, so

that $\lambda_{i,j} = U_{i,j}$. However, when the estimates d_{ij} are not exactly correct, then instead we will seek to minimize the following quantity:

Definition 4. *We define the* **stretch** *of the ultrametric matrix U with respect to the dissimilarity matrix d as follows:*

$$Stretch_d(U) = max\{\frac{d_{ij}}{U_{ij}}, \frac{U_{ij}}{d_{ij}}\}.$$

$Stretch_d(U)$ is thus an estimate of how much the evolutionary rate deviates from the molecular clock on the tree. For example, if the speed-up or slow-down on each edge is bounded between c and $1/c$ (for some positive constant c), then $Stretch_d(U) \leq c$. Furthermore, although our matrix d of estimated pairwise distances will not generally be exactly correct, for long sequences they will be close to the model distances, and so the value computed in this way will be a reasonable estimate of a lower bound of the degree of speed-up or slow down in the model tree.

The relationship between the stretch of the ultrametric matrix U with respect to the corrected distance matrix d and the deviation of the rates of change for the dataset from a strict molecular clock is thus straightforward. If the molecular clock hypothesis applies to the dataset, then as the sequence length increases, it will be possible to label the internal nodes of the model tree so that the value computed by this formula is close to 1. On the other hand, if we cannot label internal nodes so as to obtain an ultrametric matrix U so that $Stretch_d(U)$ is close to 1, then we might suspect that the molecular clock hypothesis does not hold on the dataset (and furthermore, the magnitude of $\min_U Stretch_d(U)$ will allow us to assess the degree to which it fails to hold).

This discussion suggests two computational problems:

- *Problem 1: Min Stretch Tree.* The input is a dissimilarity matrix d, and the objective is to find an ultrametric matrix U with a minimum $Stretch_d(U)$ among all possible ultrametric matrices. We call this $Stretch(d)$:

$$Stretch(d) = \min_U\{Stretch_d(U)\}$$

- *Problem 2: Min Stretch Fixed Tree T.* The input is a dissimilarity matrix d and a rooted tree T. Our goal is to find an ultrametric assignment of heights to the nodes of the tree T, thus defining an ultrametric matrix U, so that U minimizes $Stretch_d(U)$. We call this $Stretch_d(T)$.

The first problem is of interest because the minimum stretch obtained for any ultrametric tree is by necessity a lower bound on the stretch of the model tree on the matrix of estimated pairwise distances. The second problem arises

when we use techniques such as maximum parsimony, maximum likelihood (see [7] for details), and neighbor joining [16] to infer trees from biomolecular sequence datasets.

In this paper, we show that both these problems can be solved exactly in polynomial time, using techniques from [3]. We solve the first problem through the use of the general algorithm given in [3], as we show in Section 3. We solve the second problem by a modification to another algorithm in [3], as we show in Section 4. Both algorithms run in $O(n^2)$ time, i.e., linear in the input size.

3 Finding the stretch when the topology is not fixed.

In [3], Farach *et al.* described an $O(n^2)$ algorithm for finding optimal ultrametric trees with respect to an input distance matrix. We use this algorithm in order to solve the optimal stretch problem for the case where the tree is not given. We will describe the general problem they address, and show how our first issue is a special case of their general problem. Consequently, their $O(n^2)$ algorithm solves this problem.

Definition 5. THE GENERAL ULTRAMETRIC OPTIMIZATION PROBLEM:

- **Input:** *A distance matrix M and two functions $f(x, \varepsilon)$ and $g(x, \varepsilon)$ which take two real arguments such that both f and g are monotone non-decreasing on their first argument, f is monotone non-increasing on ε and g is monotone non-decreasing on ε.*
- **Output:** *The smallest ε such that there exists an ultrametric matrix U in which for all (i, j), $f(M[i, j], \varepsilon) \leq U_{ij} \leq g(M[i, j], \varepsilon)$.*

We show how our problem can be stated in these terms by defining the functions $f(x, \varepsilon)$ and $g(x, \varepsilon)$ appropriately.

Because our goal is to find an ultrametric matrix U with the minimum stretch, we want to minimize the value of the following:

$$max\{\tfrac{M[i,j]}{U_{ij}}, \tfrac{U_{ij}}{M[i,j]}\}.$$

In other words, we want to find the minimum value of ε, such that there exists an ultrametric matrix U satisfying

$$\max\{\tfrac{M[i,j]}{U_{ij}}, \tfrac{U_{ij}}{M[i,j]}\} \leq \varepsilon.$$

We solve for U_{ij} and obtain the following:

$$\tfrac{M[i,j]}{U_{ij}} \leq \varepsilon \text{ and } \tfrac{U_{ij}}{M[i,j]} \leq \varepsilon$$

which is equivalent to

$$\frac{M[i,j]}{\varepsilon} \leq U_{ij} \leq \varepsilon M[i,j].$$

The problem is reduced now to the following:
Given a distance matrix M, find the smallest ε such that there exists an ultrametric matrix U so that for all i, j, $\frac{M[i,j]}{\varepsilon} \leq U_{ij} \leq \varepsilon M[i,j]$. In other words, we wish to solve the General Ultrametric Optimization problem, with $f(x,\varepsilon) = \frac{x}{\varepsilon}$ and $g(x,\varepsilon) = \varepsilon x$.

Hence, we have:

Theorem 1. *We can solve the Min Stretch Tree problem in $O(n^2)$ time, using the algorithm in [3].*

The algorithm in [3] is therefore useful directly in solving our first problem. As we will show, the techniques in that algorithm are also useful for solving our second problem.

4 Finding the stretch when the topology is fixed.

In this section we describe a polynomial time algorithm for solving the problem of finding the minimum stretch of a fixed tree. More formally, given a tree topology T and a distance matrix M, defined on the leaves of T, the algorithm finds an ultrametric assignments of heights to the nodes of T, thus defining an ultrametric matrix U, so that U minimizes $Stretch_M(U)$.

Lemma 1. *For a given tree T, and upper and lower matrices, M_l and M_h, there exists an ultrametric assignment of heights to the internal nodes of T if the following condition holds for every internal node $v \in T$:*

$$min\ M_h[i,j]\ \geq\ max\ M_l[p,q],$$

where i, j, p, and q are leaves of T such that $lca(i,j) = v$, and $v \geq lca(p,q)$.

Proof. Given the tree T, M_l and M_h, we let the height of node $v \in T$ be

$$height(v)\ =\ max\ M_l[p,q],$$

where p and q are leaves of T such that $v \geq lca(p,q)$.

To complete the proof, we show the following two properties of the height function as we defined it:

- *If $v > w$ then $height(v) \geq height(w)$.* Since $v > w$, it follows that $S = \{(i,j) : i,j \text{ are leaves below } w\} \subseteq \{(i,j) : i,j \text{ are leaves below } v\} = S'$. Therefore, $height(v)\ =\ max\ M_l[p,q]\ \geq\ max\ M_l[i,j]\ =\ height(w)$, where $(p,q) \in S$ and $(i,j) \in S'$.

– $M_l[i,j] \leq U_{ij} \leq M_h[i,j]$ *for all* i *and* j. This follows from the definition of the height function, the fact that $U_{ij} = height(v)$, where $v = lca(i,j)$, and the given assumption.

Theorem 2. *Given a tree T, and a distance matrix M, we can find their deviation from ultrametricity in polynomial time.*

Proof. For each internal node $v \in T$, we define $p(v) = \{max\ M[i,j] :\ i,j$ are below $v\}$, and $q(v) = \{min\ M[i,j] :\ lca(i,j) = v\}$.

Using Lemma 1, we want to find the "tightest" M_h and M_l such that for every internal node $v \in T$, the following holds:

$$min\ M_h[i,j] \geq max\ M_l[a,b],$$

where $i, j, a,$ and b are leaves of T, such that $lca(i,j) = v$, and $lca(a,b) \leq v$.

Using the same monotone functions $f(x,\varepsilon)$ and $g(x,\varepsilon)$, that we defined before, for each node $v \in T$, we find the smallest $\varepsilon(v)$ such that

$$\varepsilon(v) \cdot q(v) \geq \frac{p(v)}{\varepsilon(v)}$$

Solving for $\varepsilon(v)$, we obtain

$$\varepsilon(v) \geq \sqrt{\frac{p(v)}{q(v)}}$$

To obtain the minimum value of $\varepsilon(v)$, we choose $\varepsilon(v) = \sqrt{\frac{p(v)}{q(v)}}$. The deviation of the distance matrix M, given the tree topology T, is the maximum value of $\varepsilon(v)$, where v is any internal node in the tree T.

The algorithm we have described runs in $O(n^2)$ time. We can compute $lca(i,j)$ for all pairs of leaves in $O(n^2)$ time by first doing a linear-time preprocessing of the tree followed by constant time calculations for each pair of leaves (See [6]). Therefore, by processing the distance matrix once, we can compute $q(v)$ for each v. At the same time, for each node v, we initialize the value of $p(v)$ to $max\ M[i,j]$, where i and j are below v. Then, in a bottom-up fashion, we update $p(v)$ to $max\{p(v), p(u), p(w)\}$, where u and w are the two children of v. Once $p(v)$ and $q(v)$ are computed, finding $\varepsilon(v)$ takes constant time, and finding the maximum among them takes linear time. Therefore, the algorithm takes $O(n^2)$ time.

5 Simulation Study.

5.1 Introduction

The general design of our simulation study is as follows. We implemented several techniques for estimating the stretch of a given dataset: our own technique

for obtaining a lower bound on the stretch of the model tree on a dataset when the topology is not given (described in Section 3), the technique for the fixed-tree case given in Section 4, as well as the two techniques biologist use (and which we describe below, in Section 5.3). We applied these techniques to a number of datasets obtained by simulating DNA sequence evolution down model trees, under the K2P+Gamma model.

5.2 Model Trees

We used K2P+ gamma [9] model trees. We used the *r8s* software [17] to produce a number of random birth-death trees with a strong molecular clock. Hence as the sequence length increases, the stretch on these datasets on the true tree will tend to 1. To obtain trees that are deviated from the molecular clock, we multiplied each edge in the tree by e^x, where x is a random number drawn uniformly from the range $[-\ln c, \ln c]$. We used six different values of for c: 1, 3, 5, 8, 15, and 25. The expected value of the scaling factor on an edge is $\frac{c-1/c}{2\ln c}$, so the expected deviations are moderate even for the values of c that we examined. The height of the trees generated by *r8s* is 1. To obtain trees with additional heights, we scaled those trees by factors of 0.25, 1, and 4. We set $\alpha = 1$ and ts/tv ratio equals 2 for the K2P+Gamma evolution model. We looked at trees with 20, 40 and 80 taxa, and used sequences of lengths 100, 200, 400 and 800.

5.3 Biological methods for estimating the stretch

Our experimental study examines the accuracy of two methods used by systematists for estimating the stretch of a dataset. The two methods have the same basic structure. First, they obtain an estimate of the phylogenetic tree, using either Maximum Parsimony or Maximum Likelihood. Such methods not only produce tree topologies but also edge lengths. For example, with MP, the edge lengths are the Hamming Distances on each edge, in an optimal labelling of the internal nodes of the tree so as to minimize the sum of the Hamming Distances. In Maximum Likelihood, the edge lengths represent the expected number of times a random site changes on the edge. Given either way of defining edge lengths, we can then define distances between nodes v in the tree and leaves i, and we indicate such a distance by the notation $d(v, i)$. These trees are then rooted using some technique (for example, the "mid-point" technique, whereby the midpoint of the longest path in the tree is identified with the root of the tree). Then, the stretch of the rooted edge-weighted tree T ($w(e)$ denotes the weight of edge e in T) is computed as follows:

$Dev(T, w)) = \max\{\sqrt{\frac{d(v,i)}{d(v,j)}} : v \text{ is a node in } T \text{ and } i \text{ and } j \text{ are leaves below } v\}$

where $d(v, i) = \sum_e w(e)$, where e is an edge on the path from node v to leaf i.

We denote by $Deviation_{\Phi}(T)$ the stretch of tree T as computed by the previous formula, when the branch lengths are estimated using criterion Φ. In this paper we consider $Deviation_{MP}(T)$ and $Deviation_{ML}(T)$, and we used PAUP* 4.0 [18] to assign branch lengths. In addition, we also consider $Deviation_{Model}(T)$, where the model branch lengths are used.

Note that this way of estimating the stretch of a tree with respect to an input does not verify that the internal nodes can be assigned heights so that the resultant values are feasible solutions to the stretch problem. Therefore, one of our objectives in our study was to determine whether these calculations did produce feasible solutions, or not. For the same reason, we do not call the resultant value the *stretch* of the tree with respect to the estimated distances, but rather the *deviation* of the tree with respect to the estimated distances.

There are several places where this technique can err: in particular, in obtaining a good estimate of the rooted tree, and then in assigning edge lengths. We have simplified the problem by studying the accuracy of these methods assuming that the rooted model tree is given to the methods; hence, we only use the methods to infer branch lengths and not to also find the best tree. We then compare the estimated values for the stretch obtained by those methods against the lower bound for the rooted model tree.

5.4 Simulations

We used the program `Seq-Gen` [15] to randomly generate a DNA sequence for the root and evolve it through the tree under the K2P + Gamma model. We calculated K2P+Gamma distances appropriately for the model (see [10]). We then applied the algorithm in Section 3 to compute the tree with the optimal stretch (and hence the optimal stretch). We also applied our algorithm (Section 4) to the dataset on the model topology, as well as the other techniques where the MP and ML branch length estimates of the model topology were computed.

In order to obtain statistically robust results, we used a number of *runs*, each composed of a number of *trials* (a trial is a single comparison), computed the mean for each run, and studied the mean over the runs of these events. This method enables us to obtain estimates of the mean that are closely grouped around the true value. This method was recommended by McGeoch [11] and Moret [12].

6 Results and analysis.

In this section we report on the results of the experimental studies that we carried out according to the description in Section 5. We examine the performance of the following methods for estimating the stretch:

1. The minimum stretch of the dataset: $Stretch(d)$, where d is the distance matrix of the dataset on the model tree.
2. The stretch of the ultrametric model tree (i.e. the model tree before we deviate the branch lengths away from ultrametricity) with respect to the model branch lengths obtained after deviation from ultrametricity. Thus, this is $Stretch_D(U)$ where U is the ultrametric matrix underlying the model tree, and D is the additive matrix of the model tree).
3. The minimum stretch of the dataset on the rooted model tree topology: $Stretch_d(T)$, where T is the model topology.
4. The deviation of the rooted model tree topology with respect to MP branch lengths: $Deviation_{MP}(T)$.
5. The deviation of the rooted model tree topology with respect to model branch lengths: $Deviation_{Model}(T)$.
6. The deviation of the rooted model tree topology with respect to ML branch lengths: $Deviation_{ML}(T)$.

(Recall that the deviation of the rooted model tree is calculated using the technique used by systematists, and hence does not produce a number that is guaranteed to be a feasible solution to the stretch problem.)

The values plotted in the figures are the mean of 30 runs for each experimental setting.

The values of $Deviation_{ML}(T)$ were too large (sometimes in the thousands); see Figure 3. Therefore, we did not plot those values in the graphs, since they almost always either give infeasible solutions or stretch values that are too large compared to the actual stretch.

Figures 2(a), 1(c), 1(d) and 1(a) show clearly that the values of the $Stretch(d)$ and $Deviation_{Model}(T)$ are equal when the the model tree is ultrametric. However, as the deviation from the molecular clock increases, we see that the gap between those two values widens. Therefore, even if we had a method that could estimate the branch lengths of a tree with very high accuracy, the method that biologists use computes values that are far from the values of the true stretch.

By definition, $Stretch(d) \leq Stretch_d(T)$ for all trees T, and Figures 1 and 2 demonstrate this observation empirically. Figures 2(a) and 2(b) demonstrate that our methods "converge" to the true stretch value on ultrametric datasets, as the sequence length increases. In these two figures, we notice that the two

curves corresponding to the values of $Stretch(d)$ and $Stretch_D(U)$ go to 1 as the sequence length increases.

Figures 1(a), 1(b), 1(c) and 2(c) show that $Deviation_{MP}(T)$ are sometimes inconsistent with the definition of the stretch, since we see the values computed by this method are lower than $Stretch(d)$. Figure 2(d) shows the value computed by the same method is greater than the stretch of the dataset, but is lower than the stretch of the dataset on the same topology ($Stretch_D(U)$), which is an inconsistent result. This means that the values computed by $Deviation_{MP}(T)$ are sometimes infeasible.

7 Conclusions.

In this paper we defined the concept of stretch, which is the amount of deviation of a dataset from ultrametricity. We presented two theoretical results in this paper: the first is an $O(n^2)$ algorithm for computing the optimal stretch of any ultrametric matrix for a given dataset, and the second is an $O(n^2)$ algorithm for computing the optimal stretch of a fixed tree with respect to a given dataset.

The experimental study is surprising, and shows that the two standard methods (MP and ML) used by systematists to estimate the degree to which a dataset deviates from the strong molecular clock hypothesis are quite faulty. Both can produce estimates that are not even feasible (i.e. no way of assigning heights to the nodes of any tree would produce such values), and the ML method in particular can produce enormous values, clearly much larger than is needed. More generally, our study suggests that accurate estimates of the deviation from the molecular clock may be beyond what can be inferred using existing stochastic models of evolution.

References

[1] F. Bossuyt and M.C. Milinkovitch. Amphibians as indicators of early tertiary "Out-of-India" dispersal of vertebrates. *Science*, 292:93–95, 2001.

[2] V. Dvornyk, O. Vinogradova, and E. Nevo. Long-term microclimatic stress causes rapid adaptive raditiona of kaiABC clock gene family in a cyanobacterium Nostoc linckia, from 'evolution canyons' I and II, Israel. *Proc. National Academy of Sciences (USA)*, 99(4):2082–2087, 2002.

[3] M. Farach, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13(1):155–179, 1995.

[4] N. Goldman. Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36:182–198, 1993.

[5] X. Gu and W-H. Li. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Natl. Acad. Sci. (USA)*, 95:5899–5905, 1998.

[6] D. Gusfield. *Algorithms on strings, trees, and sequences*. Cambridge University Press, 1997.

[7] D.M. Hillis, C.Moritz, and B.K. Mable. *Molecular Systematics*. Sinauer Associates, Sunderland, MA, 1996.

[8] S. Y. Kawashita, G. F. Sanson, O. Fernandez, B. Zingales, and M. R. Briones. Maximum-likelihood divergence date estimates based on rRNA gene sequences suggest two scenarios of trypanosoma crazi intrapsecific evolution. *Mol. Biol. Evol.*, 18(12):2250–2259, 2001.

[9] M. Kimura. A simple method for estimating evolutional rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[10] W.-H. Li. *Molecular Evolution*. Sinauer Assoc., 1997.

[11] C.C. McGeoch. Analyzing algorithms by simulation: variance reduction techniques and simulation speedups. *ACM Comp. Surveys*, 24:195–212, 1992.

[12] B.M.E. Moret. Towards a discipline of experimental algorithmics. In *Proc. 5th DIMACS Challenge*. available at www.cs.unm.edu/moret/dimacs.ps.

[13] M. Nei, P. Xu, and G. Glazko. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl. Acad. Sci. (USA)*, 98(5):2497–2502.

[14] M. Nikaido, K. Kawai, Y. Cao, M. Harada, S. Tomita, N. Okada, and M. Hasegawa. Maximum likelihood analysis of the complete mitochondirial genomes ofeutherians and a reevaluation of the phylogeny of bats and insectivores. *J. Mol. Evol.*, 53(4-5):508–516, 2001.

[15] A. Rambaut and N.C. Grassly. Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comp. Applic. Biosci.*, 13:235–238, 1997.

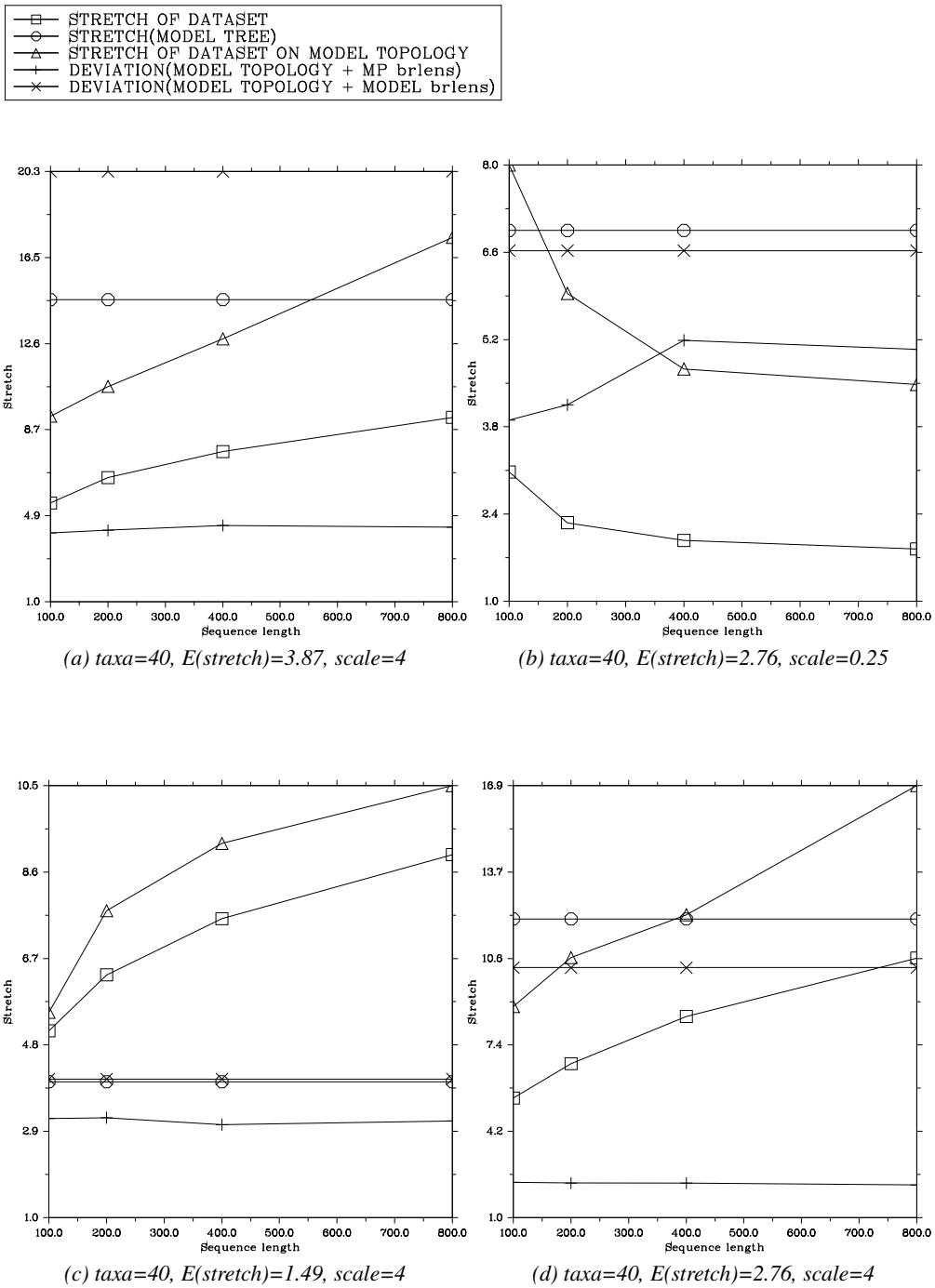
[16] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[17] Michael Sanderson. available from http://loco.ucdavis.edu/r8s/r8s.html.

[18] D. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0. 1996.

[19] L. Vawter and W.M. Brown. Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science*, 234(4773):194–196, 1986.

[20] Z. Yang, I.J. Lauder, and H.J. Lin. Molecular evolution of the hepatitis B virus genome. *J. Mol. Evol.*, 41(5):587–596, 1995.



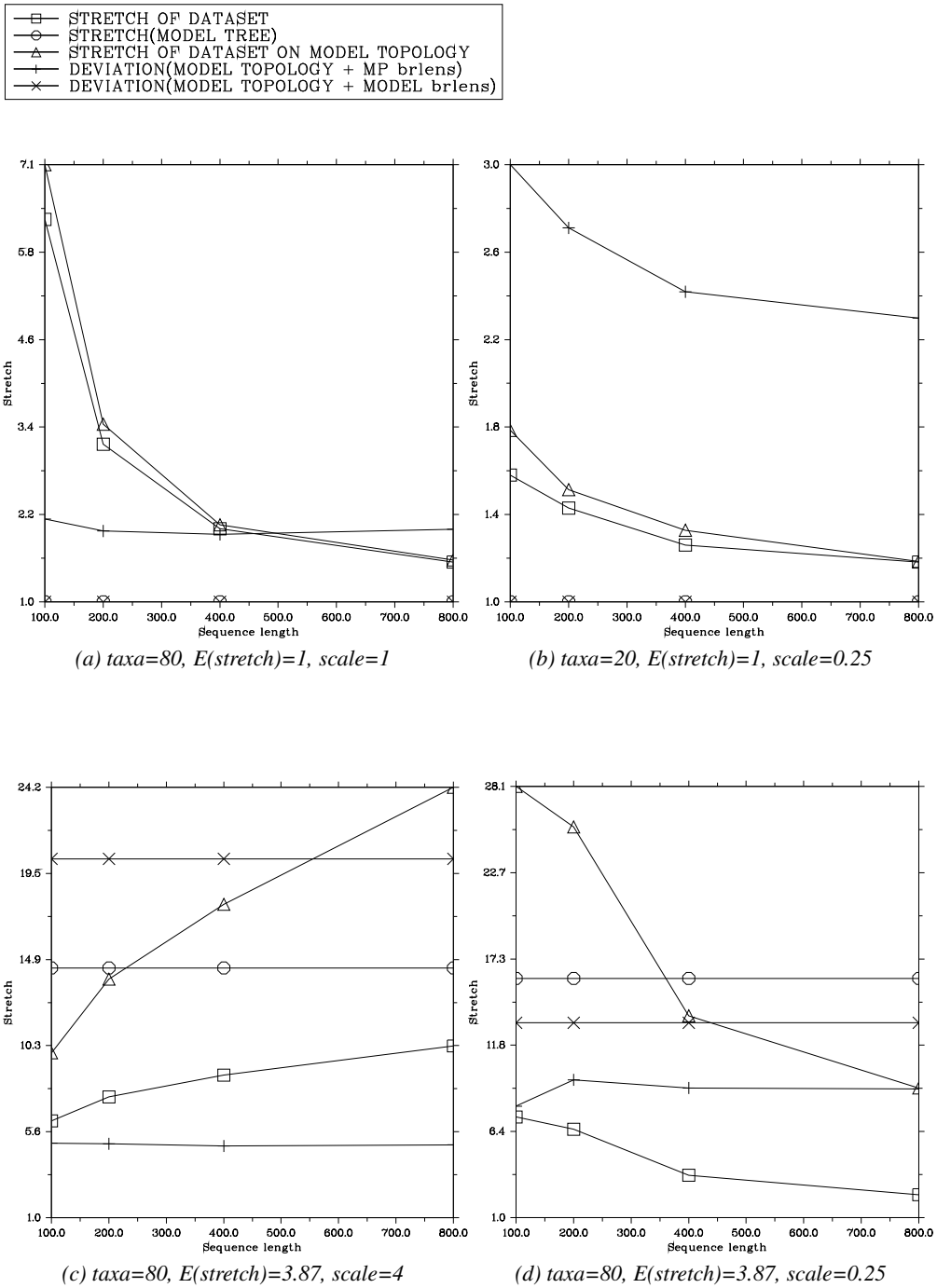
(a) taxa=40, E(stretch)=3.87, scale=4

(b) taxa=40, E(stretch)=2.76, scale=0.25

(c) taxa=40, E(stretch)=1.49, scale=4

(d) taxa=40, E(stretch)=2.76, scale=4

Fig. 1.



(a) taxa=80, E(stretch)=1, scale=1

(b) taxa=20, E(stretch)=1, scale=0.25

(c) taxa=80, E(stretch)=3.87, scale=4

(d) taxa=80, E(stretch)=3.87, scale=0.25

Fig. 2.

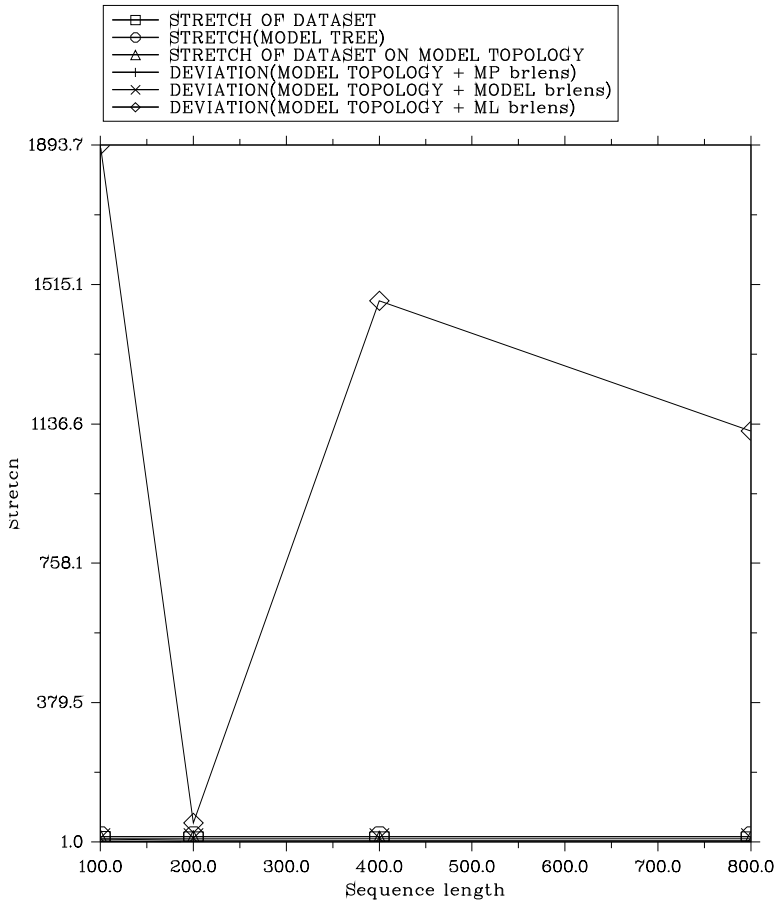


Fig. 3. taxa=20, E(stretch)=3.87, scale=0.25