# Distance Corrections on Recombinant Sequences

David Bryant[1], Daniel Huson[2], Tobias Kloepper[2], and Kay Nieselt-Struwe[2]

[1] McGill Centre for Bioinformatics
3775 University
Montréal, Québec, H3A 2B4
Canada
bryant@mcb.mcgill.ca,
[2] University of Tuebingen,
Center for Bioinformatics Tuebingen,
Sand 14
D-72076 Tuebingen, Germany
huson,kloepper,nieselt@informatik.uni-tuebingen.de

**Abstract.** Sequences that have evolved under recombination have a 'mosaic' structure, with different portions of the alignment having evolved on different trees. In this paper we study the effect of mosaic sequence structure on pairwise distance estimates. If we apply standard distance corrections to sequences that evolved on more than one tree then we are, in effect, correcting according to an incorrect model. We derive tight bounds on the error introduced by this model mis-specification and discuss the ramifications for phylogenetic analysis in the presence of recombination.

## 1 Introduction

Generally, phylogenetic analysis works under the assumption that the homologous sequences evolved along a single, bifurcating tree. Recombination, gene conversion and hybridisation can all lead to violations of this basic assumption and give rise to 'mosaic' sequences, different parts of which evolved along different trees [12,21].

Simulation experiments have established that a mosaic sequence structure can have a marked effect on phylogenetic reconstruction and evolutionary parameter estimation [13,17]. Our goal in this article is to characterise this effect theoretically. Standard distance corrections assume that the sequences evolved on a single evolutionary tree, so if we correct distance estimates using these methods we are essentially correcting according to an incorrect model. We show that the effect of this model mis-specification is relatively small and derive explicit bounds for the bias introduced by this failure to account for mosaic sequence structure.

The result has important applications in conventional phylogenetic analysis. As we shall discuss in Section 5.1, our characterisation of distance corrections on mosaic sequences provides theoretical explanations for the various forms of bias

G. Benson and R. Page (Eds.): WABI 2003, LNBI 2812, pp. 271–286, 2003.

observed experimentally in [17]. We can also apply the result to discuss the effect of rate heterogeneity on phylogenetic reconstruction. Our observations complement the inconsistency results of [5] by limiting the zone for which distance based methods are inconsistent.

However our principal motivation for this investigation was to better understand the behavior of distance based phylogenetic network algorithms like split decomposition [2] and NeighborNet [3]. We show that recombinant phylogenetic information is indeed retained in corrected distance matrices, and justify the family of network approaches that decompose distance matrices into weighted collections of split metrics. NeighborNet and split decomposition are two members of this family, though there is potential, and perhaps need, for several more.

Perhaps most importantly, we can finally provide a theoretical interpretation of the form and branch lengths of the splits graph. A splits graph is not a reconstruction of evolutionary history: the internal nodes in a splits graph should not be identified with ancestral sequences [18]. These networks had long been justified only in the weak sense that they 'represent' some kind of structure in the data. As we will discuss, we can consistently view a splits graph as an estimation of the splits appearing in the input trees (or, under a Bayesian intepretation, the splits in trees with a high posterior probability).

To illustrate, suppose that we have a collection of sequences that evolved on two different trees. Even so we compute and correct distances over all the sites, giving a 'wrongly corrected' distance matrix $d$. Suppose that one third of the sites evolved on $T_1$ and two thirds on $T_2$. If $d_{T_1}$ is the matrix of path length distances for $T_1$ and $d_{T_2}$ is the distance matrix for $T_2$ then, as we will show, the 'wrongly corrected' distance $d$ will closely approximate the weighted sum $1/3d_{T_1} + 2/3d_{T_2}$, as the sequence length increases. Since split decomposition is consistent on distance matrices formed from the sum of two distance matrices, the splits graph produced will exactly represent the splits in $T_1$ and $T_2$. Furthermore, the weights of the splits in this graphs will be a weighted sum of the corresponding branch lengths in $T_1$ and $T_2$ (where a split has length 0 in a tree that doesn't contain it).

This interpretation of splits graphs ignores some fundamental limitations of the various network methods: existing methods are consistent on particular collections of distance matrices and, as with tree based analysis, are affected by sampling error and model mis-specification. However having the correct theoretical interpretation should enable researchers to better design network methods that overcome these difficulties.

## 2 Background

### 2.1 Markov Evolutionary Models

We briefly outline the aspects of Markov processes we need for the paper. For further details, refer to [15,19] or any text book on molecular evolution.

Sequence evolution along a branch is typically modelled using a Markov process. The process is determined by an $n \times n$ *rate matrix* $Q$, where $Q_{ij} > 0$ for all $i \neq j$ and $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. Nucleotide models have $n = 4$, while amino acid models have $n = 20$. We assume that the process is *time reversible*, which means that there exist positive $\pi_1, \ldots, \pi_n$ such that $\sum_{i=1}^{n} \pi_i = 1$ and $\pi_i Q_{ij} = \pi_j Q_{ji}$ for all $i, j$. The values $\pi_i$ correspond to the equilibrium frequency for the process. We assume that the process starts in equilibrium. Let $\Pi$ denote the diagonal matrix with $\pi_1, \pi_2, \ldots, \pi_n$ on the diagonal.

Suppose that we run the process for time $t$. The probability of observing state $j$ at time $t$ conditional on being at state $i$ at time 0 equals $P_{ij}(t)$, where $P(t)$ is the *evolutionary matrix*

$$P(t) = e^{Qt} = \sum_{m=0}^{\infty} \frac{Q^m t^m}{m!}.$$

If we assume that the states are in equilibrium at the start then the probability of having $i$ at time 0 and $j$ at time $t$ equals $X_{ij}(t) = \pi_i P_{ij}(t)$. The matrix $X(t) = \Pi P(t)$ is called the *divergence matrix.*

The *mutation rate* $r_Q$ is the expected number of mutations per unit time, and can be shown to equal the sum of the off-diagonal elements of $\frac{d}{dt} X(t)|_{t=0} = \Pi Q$. Hence

$$r_Q = \sum_{i=1}^{n} \sum_{j \neq i} \pi_i Q_{ij} = -\sum_{i=1}^{n} \pi_i Q_{ii} = -\mathrm{tr}(\Pi Q)$$

where $\mathrm{tr}(A)$ denotes the trace of a matrix $A$. The expected number of changes between time 0 and time $t$ therefore equals $r_Q t$. This is the standard unit for measuring evolutionary divergence.

This general description includes many specific Markov models. The simplest for nucleotide sequences is the Jukes-Cantor model [9]. The rate matrix for this model is

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

which has only one parameter $\alpha > 0$. Substituting into the above formulae we see that the evolutionary matrix for this model is specified by

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & \text{when } i = j; \\ \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \text{when } i \neq j, \end{cases}$$

while the mutation rate is $r_Q = 3\alpha$. Thus letting $\alpha = 1/3$ gives a model with expected mutation rate of 1 per unit time.

We assume that evolution of different sites to be independent. We use $s[i]$ to denote the state at site $i$ in sequence $s$. The probability of observing sequence $s_2$ after time $t$ given $s_1$ at time 0 is then given by

$$P(s_2 | s_1, t) = \prod_i P_{s_1[i] s_2[i]}(t).$$

### 2.2 Distance Corrections

Rodríguez et al. [15] describe a general method for estimating the evolutionary distance $r_Q t$ between two sequences $s_1$ and $s_2$. Let $F_{ij}$ denote the proportion of sites for which $s_1$ has an $i$ and $s_2$ has a $j$ (Actually, we can obtain better results using $\frac{1}{2}(F_{ij} + F_{ji})$ since $F$ should be symmetric). The general time reversible (GTR) correction is given by

$$\hat{d} = r_Q \hat{t} = -\mathrm{tr}(\Pi \log(\Pi^{-1} F))$$

where log is the matrix logarithm defined by

$$\log(I + A) = A - \tfrac{1}{2}A^2 + \tfrac{1}{3}A^3 - \tfrac{1}{4}A^4 + \cdots .$$

The correction formula is consistent: if $F = X(t)$ then

$$\begin{aligned} -\mathrm{tr}(\Pi \log(\Pi^{-1} F)) &= -\mathrm{tr}(\Pi \log(e^{Qt})) \\ &= -\mathrm{tr}(\Pi Q)t \\ &= r_Q t. \end{aligned}$$

Most of the standard corrections can be derived from this general formula. For example, under the Jukes-Cantor model, suppose that we have observed that the proportion of changed sites equals $p$. Our estimate for $X(t)$ would then be the matrix with $p/12$ on the off-diagonal (there are 12 such entries) and $(1-p)/4$ on the diagonal. Substituting into the general formula, we obtain the standard Jukes-Cantor correction $r_Q t = -\frac{3}{4} \log(1 - \frac{4}{3}p)$.

### 2.3 Trees, Splits, Splits Graphs, and Distance Matrices

A *phylogenetic tree* is a tree with no vertices of degree two and leaves identified with the set of taxa $X$. A *split* $A|B$ is a partition of the taxa set into two non-empty parts. Removing an edge from a phylogenetic tree $T$ induces a split of the taxa set. The set of splits that can be obtained in this way from $T$ is called the *splits of* $T$ and denoted $\Sigma(T)$. A given set of splits is *compatible* if it is contained within the set of splits of some tree.

A *splits graph* is a bipartite connected graph $G$ with a partition $E(G) = E_1 \cup E_2 \cup \cdots \cup E_k$ of $E(G)$ into disjoint sets such that no shortest path contains more than two edges from the same block and, for each $i$, $G - E_i$, consists of exactly two components. (This definition is equivalent to the definition of [6]). Some of the vertices are labelled by elements of $X$ so that each edge cut $E_i$ induces a split of $X$. The set of these splits is denoted $\Sigma(G)$. Note that every set of splits can be represented by a splits graph, but this graph is not necessarily unique. Every phylogenetic tree is a splits graph with every edge in a different block.

Suppose that we assign lengths to the edges of $T$. The *additive distance* $d_T(x, y)$ between two taxa $x, y$ equals the sum of the edge lengths along the path separating them. We can also assign length to the edges in a splits graph

$G$, where all edges in the same block are assigned the same length. The distance $d_G(x, y)$ between two taxa in $G$ is then the length of the shortest path connecting them.

Both $d_T$ and $d_G$ have an equivalent formulation in terms of splits. The *split metric* $\delta_{A|B}$ for a split $A|B$ is the (pseudo-)metric

$$\delta_{A|B}(x, y) = \begin{cases} 1 & \text{if } x, y \text{ are on different sides of } A|B; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\lambda_{A|B}$ denote the length of the edge (or in the case of splits graphs, edges) corresponding to $A|B$. Both $d_T(x, y)$ and $d_G(x, y)$ equal the sum of the edge lengths for all of the splits that separate $x$ and $y$. Hence

$$d_T(x, y) = \sum_{A|B \in \Sigma(T)} \lambda_{A|B}\delta_{A|B}(x, y) \quad \text{and} \quad d_G(x, y) = \sum_{A|B \in \Sigma(G)} \lambda_{A|B}\delta_{A|B}(x, y).$$

Split decomposition [2] and NeighborNet [3] both take a distance metric $d$ and compute a decomposition

$$d = \epsilon + \sum_{A|B} \lambda_{A|B}\delta_{A|B}$$

of $d$ into a positive combination of split metrics and an error term $\epsilon$. Furthermore, both methods are *consistent* over a large class of metrics. If a set of splits $\mathcal{S}$ is *weakly compatible* and

$$\bar{d}(x, y) = \sum_{A|B \in S} \bar{b}(A|B)\delta_{A|B}$$

then split decomposition will recover the splits $A|B$ as well as the coefficients $\bar{b}$ [2]. NeighborNet will recover this decomposition when the set of splits $\mathcal{S}$ is circular [4].

## 3 Correcting Distances Estimated from Mosaic Sequences

In a mosaic alignment, different sites evolved along different trees. Correction formulae, such as those described above, make the assumption that the sequences evolved on the same tree. When we apply these corrections to mosaic sequences we are correcting according to an incorrect model.

We show here that the distance correction formulae work just how we would hope, at least up to a small error term. Correcting a heterogeneous collection of sequences using a homogeneous model does introduce error, but the error is quite small compared to the distances themselves.

Suppose that the sequences evolved under the same model on $k$ different trees $T_1, T_2, \ldots, T_k$. Furthermore, for each $i$, suppose that the proportion of sites coming from $T_i$ is $q_i$. Let $s_1$ and $s_2$ be the sequences for two taxa, and let

$d_1, d_2, \ldots, d_k$ be the expected number of mutations on the path between these taxa on trees $T_1, T_2, \ldots, T_k$. We use

$$\mathbb{E}[d] = \sum_{i=1}^{k} q_i d_i$$

and

$$\mathrm{var}[d] = \sum_{i=1}^{k} q_i (d_i - \mathbb{E}[d])^2$$

to denote the mean and variance of the $d_i$'s.

Let $\hat{F}_{ab}$ denote the proportion of sites with an $a$ in sequence $s_1$ and a $b$ in sequence $s_2$. Then $\hat{F}$ will approach

$$F = \sum_{i=1}^{k} q_i e^{Qt_i}$$

as the sequences become sufficiently long. We obtain upper and lower bounds on the distance estimate computed from $F$.

First, however, we need to prove a small result in matrix analysis.

**Lemma 1.** *Suppose that all eigenvalues of an $n \times n$ matrix $X$ are real and non-negative, and that there is a diagonal matrix $D > 0$ such that $DX$ is symmetric. Then* $\mathrm{tr}(DX) \geq 0$.

*Proof*

Since $D > 0$ the inverse $D^{-1}$ and square root of the inverse $D^{-\frac{1}{2}}$ both exist. Define the matrix $Y$ by

$$Y = D^{-\frac{1}{2}}(DX)D^{-\frac{1}{2}} = D^{\frac{1}{2}} X D^{-\frac{1}{2}}.$$

Then $Y$ is symmetric and has the same non-negative eigenvalues as $X$. It follows that $Y$ is positive semi-definite with non-negative diagonal entries. As $D^{-\frac{1}{2}} > 0$, the matrix $DX$ has non-negative diagonal entries and $\mathrm{tr}(DX) \geq 0$. □

**Theorem 1** *Let $F = \sum_{i=1}^{k} q_i e^{Qt_i}$ and let $\rho_Q$ denote the constant $\frac{\mathrm{tr}(\Pi Q^2)}{(r_Q)^2}$. Then*

$$\mathbb{E}[d] - \tfrac{1}{2}\rho_Q \mathrm{var}[d] \leq -\mathrm{tr}(\Pi \log(\Pi^{-1}F)) \leq \mathbb{E}[d].$$

*Proof*

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $Q$. One of these is zero and all others are negative. Let $v_1, v_2, \ldots, v_n$ be a linearly independent set of eigenvectors, where $Qv_j = \lambda_j v_j$ for all $j$. Define $t_i = \frac{d_i}{r_Q}$ for all $i$, and $\bar{t} = \frac{\mathbb{E}[d]}{r_Q}$. Then

$$\begin{aligned}\mathbb{E}[d] - (-\mathrm{tr}(\Pi \log(\Pi^{-1}F))) &= \mathrm{tr}(\Pi(\log(\Pi^{-1}F) - Q\bar{t})) \\ &= \mathrm{tr}(\Pi A)\end{aligned}$$

where $A = (\log(\mathit{\Pi}^{-1}F) - Q\bar{t})$. The matrix $A$ has the same eigenvectors as $Q$. Let $\alpha_j$ denote the eigenvalue of $A$ corresponding to the eigenvector $v_j$. We derive lower and upper bounds on $\alpha_j$.

For the lower bound we have

$$\begin{aligned}\alpha_j &= \log(\sum_{i=1}^{k} q_i e^{\lambda_j t_i}) - \lambda_j \bar{t} \\ &\geq \sum_{i=1}^{k} q_i \log(e^{\lambda_j t_i}) - \lambda_j \bar{t} \\ &= 0\end{aligned}$$

with the inequality following from the concavity of the logarithm (or Jensen's inequality). It follows that $A$ has only non-negative eigenvalues so, by Lemma 1, $\mathrm{tr}(\mathit{\Pi} A) \geq 0$. Thus

$$0 \leq \mathrm{tr}(\mathit{\Pi} A) = \mathrm{tr}(\mathit{\Pi} \log(\mathit{\Pi}^{-1}F)) - \mathrm{tr}(\mathit{\Pi} Q)\bar{t} = \mathrm{tr}(\mathit{\Pi} \log(\mathit{\Pi}^{-1}F)) + \mathbb{E}[d]$$

For the upper bound, let $x = \sum_{i=1}^{k} q_i(e^{\lambda_j t_i} - 1)$. Then $-1 < x \leq 0$ so $\log(1+x) \leq x - \frac{1}{2}x^2$ and

$$\begin{aligned}\alpha_j &= \log\left(1 + \sum_{i=1}^{k} q_i(e^{\lambda_j t_i} - 1)\right) - \lambda_j \bar{t} \\ &\leq \left(\sum_{i=1}^{k} q_i(e^{\lambda_j t_i} - 1)\right) - \tfrac{1}{2}\left(\sum_{i=1}^{k} q_i(e^{\lambda_j t_i} - 1)\right)^2 - \lambda_j \bar{t}\end{aligned}$$

If we set $y = \lambda_j t_i$ then $y \leq 0$ and $y \leq e^y - 1 \leq y + \frac{1}{2}y^2$. Thus

$$\begin{aligned}\alpha_j &\leq \sum_{i=1}^{k} q_i(\lambda_j t_i + \tfrac{1}{2}\lambda_j^2 t_i^2) - \tfrac{1}{2}\left(\sum_{i=1}^{k} q_i \lambda_j t_i\right)^2 - \lambda_j \bar{t} \\ &= \tfrac{1}{2}\lambda_j^2 \left[\sum_{i=1}^{k} q_i t_i^2 - \left(\sum_{i=1}^{k} q_i t_i\right)^2\right] \\ &= \tfrac{1}{2}\tfrac{\lambda_j^2}{(r_Q)^2}\mathrm{var}[d]\end{aligned}$$

Thus $(\frac{\mathrm{var}[d]}{2}Q^2/r_Q^2 - A)$ has non-negative eigenvalues, and

$$\begin{aligned}0 &\leq \mathrm{tr}(\mathit{\Pi}(\tfrac{\mathrm{var}[d]}{2}Q^2/r_Q^2 - A)) \\ &= \tfrac{\mathrm{var}[d]}{2}\mathrm{tr}(\mathit{\Pi} Q^2)/r_Q^2 - \mathrm{tr}(\mathit{\Pi} \log(\mathit{\Pi}^{-1}F)) + \mathrm{tr}(\mathit{\Pi} Q)\bar{t} \\ &= \tfrac{1}{2}\rho_Q \mathrm{var}[d] - \mathrm{tr}(\mathit{\Pi} \log(\mathit{\Pi}^{-1}F)) - \mathbb{E}[d].\end{aligned}$$

□

This general result implies error bounds for all the standard distance corrections. For example, under Jukes-Cantor, we have $\rho_Q = \frac{4}{3}$ so if the proportion of observed changes approaches $p$ then

$$\mathbb{E}[d] - \frac{2}{3}\mathrm{var}[d_i] \leq -\frac{3}{4}\log(1-\frac{4}{3}p) \leq \mathbb{E}[d]$$

For K2P with parameter $\kappa = 2\frac{\text{expected num. transitions}}{\text{expected num. transversions}}$ we obtain the bound

$$\mathbb{E}[d] - \frac{\kappa^2+2\kappa+3}{(\kappa+2)^2}\mathrm{var}[d] \leq -\mathrm{tr}(\Pi\log(\Pi^{-1}F)) \leq \mathbb{E}[d].$$

The divergences between well aligned molecular sequences are typically small. In this case, the error bound $\frac{1}{2}\rho_Q\mathrm{var}[d]$ comes close to zero. Thus, when distances are small, the corrected distances approximate the convex combination of the distances from the different blocks of the mosaic sequences. We therefore have

**Corollary 1.** *Let $\hat{\mathbf{d}}$ be the corrected distances estimated from mosaic sequences evolved on trees $T_1,\ldots,T_k$, where for each $i$, the proportion of sites evolved on $T_i$ is $q_i$. Let $\mathbf{d}_i$ be the distance matrix estimated from only those sites evolving on $T_i$. Then*

$$\hat{\mathbf{d}} \approx \sum_{i=1}^{k} q_i\mathbf{d}_i.$$

Even if the bias *is* sufficient to have a significant effect on distance estimates, this bias is well characterised. The lower bound $\mathbb{E}[d] - \frac{1}{2}\mathrm{var}[d]$ is very tight. The distance correction differs from $\mathbb{E}[d] - \frac{1}{2}\mathrm{var}[d]$ only by a term of order $O(d^3)$, as can be easily demonstrated from a Taylor series expansion.

Note also that the error bound $\frac{1}{2}\rho_Q\mathrm{var}[d]$ depends only on the variance of the block distances, and not on the number of different blocks. By taking $k$ to infinity, we see that Theorem 1 extends directly to a continuous distribution on input distances.

## 4 Experimental Results

We performed two separate experiments to assess the tightness of the approximation established in Theorem 1 for the distance between two sequences. The parameters for each run were

- The number of contiguous blocks $k$, set to $k = 2$ or 5.
- The height $h$ (in expected mutations) of the root in the coalescent.

The $k$ contiguous segments were determined by randomly selecting $k-1$ breakpoints without repetition. This gave the proportions $q_1, q_2, \ldots, q_k$ of sites from each tree. The distances $d_i$ for each contiguous segment were sampled by constructing a coalescent with 30 leaves and height $h$, using the protocol outlined

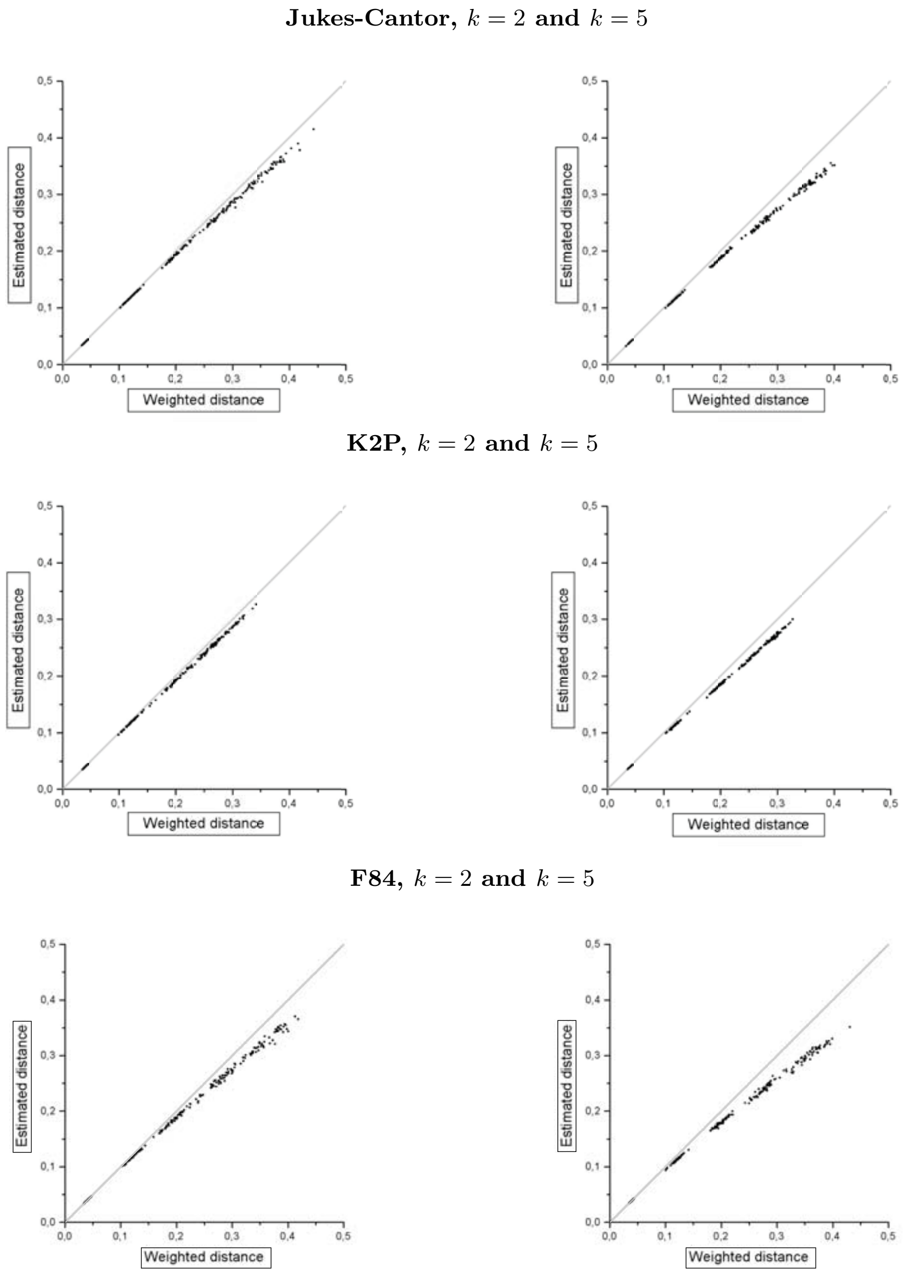

**Fig. 1.** Results of the first experiment. The estimated distance is computed using a distance correction applied to the whole sequence. The weighted distance is the mean $\mathbb{E}[d]$ of the distances for each contiguous block. Results are presented for Jukes-Cantor, Kimura 2-parameter (K2P) and the Felsenstein 84 model (F84).

in [10], then taking the distance between two fixed leaves. (In practice this sampling can be performed without constructing the tree by simply determining the time taken for the two lineages to coalesce.) A new coalescent tree is sampled for each segment.

For the first experiment we selected the Jukes-Cantor (JC), Kimura 2-parameter (K2P), and Felsenstein 84 (F84) models [19], scaled so that they had rates of $-\mathrm{tr}(\Pi Q) = 1$. We used $\kappa = 2$ for K2P and F84, and equilibrium frequencies $\pi_A = 0.37, \pi_C = 0.4, \pi_G = 0.05, \pi_T = 0.18$. for F84. These equal the emperical frequencies observed by [20] for Human mtDNA. For each we computed $F = \Pi(\sum_{i=1}^{k} q_i e^{Qt_i})$ analytically, then computed the *estimated distance* $-\mathrm{tr}(\Pi \log(\Pi^{-1}F))$ and the *weighted distance* $E[d]$.

The results for this first experiment are presented in figure 1. We plot estimated distance $-\mathrm{tr}(\Pi \log(\Pi^{-1}F))$ versus the weighted distance $\mathbb{E}[d]$. The fit is quite close, even when the distances become quite large. The actual error (weighted distance - estimated distance) is extremely close to the error bound from Theorem 1, differing only in the 5th or 6th decimal place (data not shown).

For the second experiment we wanted to compare the estimated distances to weighted distances for randomly generated sequences. We selected the $t_i$'s and $q_i$'s as for the first experiment. We used SEQGEN [14] to randomly evolve two mosaic sequences of length 1200, where the sequences were evolved separately for each contiguous segment. Even with no recombination, the presence of sampling error means that the corrected distance computed from the sequences will differ from the distance used to generate the sequences. As we wanted to distinguish sampling error from the error introduced by recombination, we re-estimated the distances for each contiguous segment. That is, for each $i = 1, 2, \ldots, k$ we computed $a_i$, the corrected distance computed from the sequences in the $i$th contiguous segment.

Because some of the segments were short they were often saturated. We resampled all cases when saturation occurred.

The results for this study are presented in figure 2, just for the Jukes-Cantor case. Once again, the estimated distances closely approximate the weighted distances. The first two plots present the results for $k = 2$ and $k = 5$ blocks. The lower plots compare the sampling error, measured as the absolute difference between the estimated distance and the average of the distance used to generate the sequences, versus the error due to recombination. The plots indicate that the two values are of roughly the same magnitude, with the sampling error being somewhat larger.

## 5 Applications

### 5.1 The Consequences of Recombination on Traditional Phylogenetic Analysis

Schierup and Hein [17] conducted an extensive simulation experiment to assess the effect of mosaic sequence structure on features of reconstructed phylogenies.

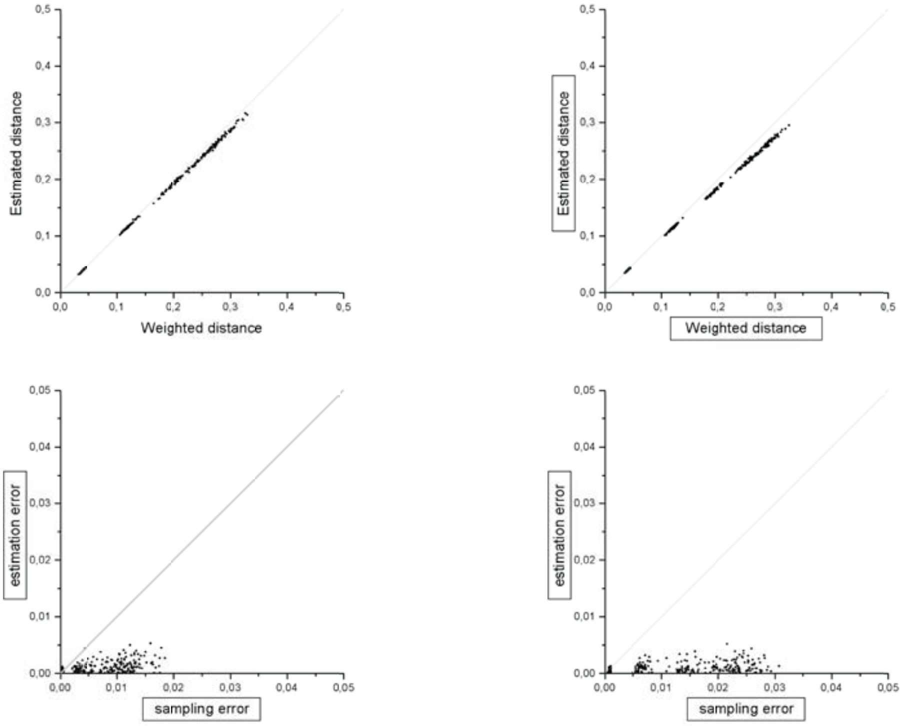

**Fig. 2.** Results of the second experiment. The top two plots give estimated distance versus weighted distance for JC ($k = 2, 5$). The two lower graphs plot the sampling error versus the error from the approximation .

They used the recombinant coalescent algorithm of [8] to generate genealogies with varying rates of recombination, then evolved simulated DNA sequences along these genealogies. Distances were corrected according to the Jukes-Cantor model, and trees were constructed using a least squares heuristic. As the amount of recombination increased, Schierup and Hein observed

1. a tiny decrease in the average distance between sequences.
2. a decrease in the time to the most recent common ancestor of all taxa, and also in the average time back to the common ancestors of pairs of taxa
3. an increase in the total length (sum of branch lengths) of the topology

All of these observations can be predicted from Theorem 1. We showed that the estimated distances will under-estimate the average distances from the various input trees. This explains the decrease in average pairwise distances. The fact that this decrease is very small indicates that the estimated distances are close to the convex combinations of the input distances.

The decrease in tree height and average time to least common ancestors is therefore due more to the presence of conflicting signal than the negative bias predicted by Theorem 1. The time to the most recent common ancestor in any clock based phylogeny equals half the maximum divergence between taxa. A given pair of taxa will most likely not be maximally diverged in all of the input phylogenies, so the maximum divergence between sequences will decrease when we take a convex combination from different trees.

The increase in total tree length follows from the proof of consistency for minimum evolution [16], at least for ordinary least squares (see [7]). If we estimate branch lengths from a distance matrix from an incorrect tree, the tree length will be longer than for the correct tree. Thus the incompatibilities introduced by increased recombination will increase tree length.

### 5.2 Inconsistency of Phylogenetic Analysis under Variable Rate Models

Chang [5] showed that distance based and maximum likelihood methods can be inconsistent when evolutionary rates vary across sites. Using Theorem 1, we can limit the zone of inconsistency for these methods. Variation in evolutionary rates means that each site evolved on the same tree but the branch lengths can differ. We can rescale so that the expected distance between two taxa equals the distance between them in $T$. The inconsistency is due to the variance in the distance caused by the varying rates. Bounding this rate variance makes Neighbor-joining a consistent method.

**Theorem 2** *Suppose that sequences are evolved along a phylogeny $T$ under a stochastic model with rate matrix $Q$ and variable evolutionary rates. Let $\epsilon$ be the expected number of mutations along the shortest branch of $T$. If*

$$\mathrm{var}[d] < \frac{\epsilon}{\rho_Q}$$

*then Neighbor-Joining (and most other distance based methods) applied to corrected distances will return $T$ with sufficiently long sequences.*

*Proof*
We prove the result for a finite number $k$ of possible evolutionary rate histories, though the result extends immediately to a continuous rate distribution. Each rate history corresponds to an assignment of branch lengths to $T$. For each $i$ we let $T_i$ denote $T$ with branch lengths modified according to the $i$th possible rate history. Let $d_i$ denote the additive distance for $T_i$. Since each of $T_1, \ldots, T_k$ has the same topology, the distance matrix $\mathbb{E}[d]$ formed from their weighted averages is also additive on $T$.

From Theorem 1 the corrected distance $-\mathrm{tr}(\Pi \log(\Pi^{-1}F))$ differs from $\mathbb{E}[d]$ by at most $\frac{1}{2}\rho_Q \mathrm{var}[d] = \frac{1}{2}\epsilon$. Neighbor-joining therefore returns the correct tree $T$ [1]. □

### 5.3 Split Decomposition and NeighborNet

Models for generating sequences under recombination combine three parts: the sampling of the recombination history, the sampling of breakpoints, and the evolution of the sites. After the recombination history and breakpoints are determined, each site has an associated phylogenetic tree. While the trees are correlated, we usually assume independence of each site given the trees (e.g. [11,17]). The problem of reconstructing the complete recombination history therefore requires a reconstruction of the contributing phylogenies.

Suppose that we have a mosaic alignment with a proportion of $q_1$ of the sites evolved on $T_1$, $q_2$ of the sites evolved on $T_2$,...,$q_k$ sites evolved on $T_k$. For each $i$ let $d_i$ denote the distance matrix for $T_i$. Since the trees $T_i$ are highly correlated there will be less variation in the individual distances than if the trees had been sampled independently. If the sequences are sufficiently long, the distance estimates computed for the whole sequence will approximate the weighted average

$$\bar{d}(x,y) = \sum_{i=1}^{k} q_i d_i(x,y).$$

Each tree $T_i$ can be decomposed into a non-negative linear combination of split metrics, as we observed in Section 2.3. For each tree $T_i$, let $b_i(A|B)$ denote the length of the branch corresponding to $A|B$, with $b_i(A|B) = 0$ if $A|B$ is not a split of $T_i$. We set

$$\bar{b}(A|B) = \sum_{i=1}^{k} q_i b_i(A|B) \geq 0.$$

Let $\mathcal{S}$ equal the union $\Sigma(T_1) \cup \cdots \cup \Sigma(T_k)$ of the splits of $T_1, \ldots, T_k$. Then

$$d_i(x,y) = \sum_{A|B \in \mathcal{S}} b_i(A|B)\delta_{A|B}(x,y)$$

and

$$\begin{aligned}\bar{d}(x,y) &= \sum_{i=1}^{k} \sum_{A|B \in \mathcal{S}} q_i b_i(A|B)\delta_{A|B}(x,y) \\ &= \sum_{A|B \in \mathcal{S}} \bar{b}(A|B)\delta_{A|B}(x,y).\end{aligned}$$

We therefore have

1. The distance matrix $\bar{d}$ is in the positive cone generated by the split metrics for splits in the trees $T_1, \ldots, T_k$
2. The coefficient of $\delta_{A|B}$ in this sum equals the sum of the branch lengths corresponding to $A|B$ in the input tree, where the branch lengths in $T_i$ are weighted by a factor $q_i$.

Split decomposition and NeighborNet both take a distance metric $d$ and compute a decomposition

$$d = \epsilon + \sum_{A|B} \lambda_{A|B} \delta_{A|B}$$

of $d$ into a positive combination of split metrics and an error term $\epsilon$. Furthermore, both methods are *consistent* over a large class of metrics. If the set of splits $\mathcal{S}$ from the trees $T_1, \ldots, T_k$ is *weakly compatible* and

$$\bar{d}(x,y) = \sum_{A|B \in S} \bar{b}(A|B) \delta_{A|B}$$

then split decomposition will recover the splits $A|B$ as well as the coefficients $\bar{b}$ [2]. In particular, if $k = 2$ then $\mathcal{S}$ is weakly compatible and split decomposition will recover the splits. If $\mathcal{S}$ is *circular* then NeighborNet will also recover both the splits and coefficients.

The splits in a splits graph therefore represent an estimate of the splits in the trees generating the mosaic sequences. The lengths of the edges represent an estimate of the corresponding branch lengths, weighted by the frequencies.

## 6 Discussion – Error and the Phylogenetic Analysis of Recombination

We have established a general result for distance corrections on mosaic sequences, and studied applications of this result to phylogenetic tree and network construction. The result characterises the signal present in distance matrices derived from recombinant sequences, a fundamental step towards the design of new methods for recovering this conflicting phylogenetic signal.

Building on the observations made in Theorem 1 we can identify (at least) four sources of error that could cause the splits graph representation to be incorrect, even under the assumption that we have the correct evolutionary model. The first is the (negatively biased) error term introduced by the approximation in Theorem 1. This factor is tiny, however, and is likely to only affect splits with small branch lengths or splits that only contribute to a small fraction of the sites.

A second source of error is sampling. Here the combined network approach seems to have an advantage over sliding window approaches. A consequence of the Theorem is that the variance of the estimate $\mathbb{E}[d]$ is not significantly different than the standard variance estimate for an alignment without recombination. On the other hand, if we knew which sites evolved from which trees and estimated these distances separately, the variances would be far higher.

The third source of error is the systematic error introduced by split decomposition and NeighborNet. For certain classes of splits, both methods are consistent. However if the splits in the trees contributing to the mosaic are not weakly compatible (or in the case of NeighborNet, not circular) then the resulting splits graph could be misleading. There is a need to characterise how both methods respond to these model violations, and scope for developing further methods for decomposing distance matrices.

When the splits in $\mathcal{S}$ are not weakly compatible a fourth complication can arise. A distance matrix can have two or more distinct decompositions into split metrics. Consider the three clock-like trees $T_1, T_2, T_3$ in figure 3. The average of the distance matrices generated by these trees is exactly equal to the distance matrix for $T^*$. Even if the the proportions were not exact, it would always be possible to decompose the estimated distance matrix into the non-negative sum of six (rather than seven) split metrics [2]. This problem of non-recoverability will become worse the more complicated the recombinations are, and poses a severe challenge for the development, and experimental analysis, of recombination reconstruction algorithms.

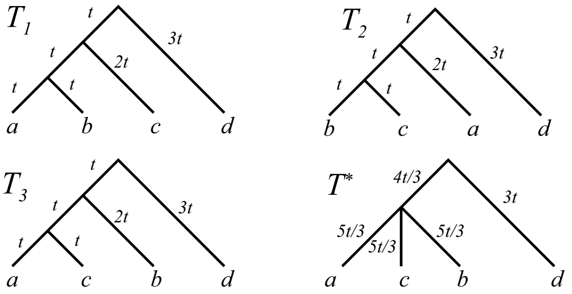

**Fig. 3.** An example of non-recoverability. The three trees $T_1, T_2, T_3$ generate the same distance matrix as the single tree $T^*$.

## References

1. K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.
2. H.-J. Bandelt and A.W.M. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92:47–105, 1992.
3. D. Bryant and V. Moulton. Neighbornet: An agglomerative algorithm for the construction of planar phylogenetic networks. In R. Guigo and D. Gusfield, editors, *Workshop in Algorithms for Bioinformatics (WABI)*, number 2452 in LNCS, pages 375–391. Springer-Verlag, 2002.
4. D. Bryant and V. Moulton. Consistency of the neighbornet algorithm for constructing phylogenetic networks. Technical report, School of Computer Science, McGill University, 2003.
5. J. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134:189–215, 1996.

6. A.W.M. Dress and D. Huson. Computing phylogenetic networks from split systems (manuscript).
7. O. Gascuel, D. Bryant, and F. Denis. Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50:621–627, 2001.
8. R.R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.
9. T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York, 1969.
10. M.K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11:459–468, 1994.
11. M.K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–1401, 2000.
12. J. Maynard-Smith. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34:126–129, 1992.
13. D. Posada and K. Crandall. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54:396–402, 2002.
14. A. Rambaut and N.C. Grassly. Seq-gen: An application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235–238, 1997.
15. F. Rodriguez, J. Oliver, A. Marin, and R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142:485–501, 1990.
16. A. Rzhetsky and M. Nei. Theoretical foundation of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10:1073–1095, 1993.
17. M. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879–891, 2000.
18. K. Strimmer, C. Wiuf, and V. Moulton. Recombination analysis using directed graphical models. *Molecular Biology and Evolution*, 18:97–99, 2001.
19. D. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B.K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer, 2nd edition, 1996.
20. K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526, 1993.
21. C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, 18:1929–1939, 2001.