Distance Measures in Terms of Substitution Processes

Ellen Baake[1] and Arndt von Haeseler[2]

Zoologisches Institut, Universität München, Luisenstr. 14, D-80333 München, Germany.

Running head: Distances and substitution processes

appeared in: Theor. Pop. Biol. **55** (1999), 166–175.

[1]phone: +49-89-5902570,
Fax: +49-89-5902474,
email: baake@zi.biologie.uni-muenchen.de

[2] present address: Max-Planck-Institut für evolutionäre Anthropologie, Inselstr. 22, D-04103 Leipzig, Germany.
Fax: +49-341-52119,
email: haeseler@eva.mpg.de

1 Abstract

Phylogenetic reconstruction from DNA or amino acid sequences relies heavily on suitable distance measures. A number of new distance measures (asynchronous, LogDet, and paralinear distances) have been proposed recently which possess the desired property of tree-additivity under fairly general models of sequence evolution, but are not well understood from the mechanistic point of view. We review them here in a unifying framework, which is the substitution process in continuous time. The emerging interpretation will also clarify the relationship among these distance measures.

We also tackle situations with site-to-site variation of substitution rates, which is well-known to cause non-additive distances and inconsistent branch lengths. For homogeneous, stationary, time-reversible models, this may be repaired provided the distribution of rates is known. In contrast, we will show that, for non-stationary models, different tree topologies may produce identical joint distributions of letters in pairs of sequences, given the same distribution of rates. This precludes the existence of *any* tree-additive pairwise distance measure.

2 Introduction

The task of inferring phylogenetic trees from aligned sequences is, in the majority of applications, accomplished via pairwise distance methods. There are simple and fast algorithms available for this purpose, but the result depends crucially on the notion of distance chosen; for a review of the field, see Swofford *et al.* (1995). It is well-known (see e.g. Waterman *et al.*, 1977) that, given ideal data, the topology of the tree may be recovered correctly if the distance is tree-additive, i.e., if there exists a tree such that the distance between any two sequences equals the sum over the edge lengths on the path joining them. These branch lengths need not have an immediate biological interpretation.

The data available for defining distances is the set of divergence matrices as extracted from site-by-site comparison of pairs of sequences. Among the many possible measures of pairwise distance that one might like to extract from these, a distinguished one is the expected number of substitution events that have occurred per site since the sequences in question diverged from a common ancestor. It is obviously tree-additive and, if rates of evolution are not too different across the tree, it even yields edge lengths that are roughly proportional to time. However, only under very favourable circumstances is it possible to infer the number of substitutions from the divergence matrix, namely if the process of sequence evolution is homogeneous across the tree, stationary, and reversible; this so-called time-reversible distance goes back to Lanave *et al.* (1984). Since the underlying assumptions are rarely satisfied, alternative distance measures have been proposed that may be used under more general models of sequence evolution. The earliest, to our knowledge, was the 'asynchronous distance' (Barry and Hartigan, 1987), which is, however, neither symmetric (and thus not a distance in the true sense), nor tree-additive. These disadvantages were remedied by two related quantities, namely the LogDet distance (Steel, 1994) and the paralinear distance (Lake, 1994). Since the asynchronous, LogDet, and paralinear distances all involve logarithms of determinants of the divergence matrices, we will call them the 'LogDet family'. These quantities have proved useful in both applications and theory (e.g. Zharkikh, 1994; Lockhart *et al.*, 1994; Gu and Li, 1996*a*; Chang, 1996*a*). Unlike the case with the number of substitutions, however, the LogDet distance measures have been lacking a mechanistic interpretation. It is known that, if the base frequencies are all equal, they are proportional to the number of substitutions, and otherwise are close to this (e.g. Lake, 1994); but apart from this, their meaning is not well-understood.

The purpose of the present paper is twofold. First, we will employ the substitution process in continuous time to review distance measures, in particular the LogDet family, in a unifying framework. Being patently adequate to the biological situation, this description will yield a mechanistic interpretation of the corresponding branch lengths, and, at the same time, clarify the relationship among the distance measures in question. We will then turn to situations with rate heterogeneity across sites, which, if not taken into account, is well-known to cause non-additive distances and inconsistent branch lengths. Whereas the time-reversible distance may be corrected for rate heterogeneity given the distribution of rates (cf. Gu and Li, 1996*b*; Waddell and Steel, 1997), no such remedy is known for the LogDet family. We will investigate the nature of this problem.

3 The Problem and the Framework

The sequences collected are understood as the terminal nodes (or leaves) of a tree $\mathcal{T}$ (i.e. a connected graph without cycles), and labelled $1, \ldots, N$ (see Fig. 1). Nodes will be identified with sequences throughout. The tree is rooted at vertex ρ, which induces an orientation of its edges (or branches): Edge $e = (w, v)$ is a pair of nodes ordered such that v is on the path from ρ to w. Every edge has a 'duration'

(A1) $$\tau^{(w,v)} := t^{(w)} - t^{(v)} > 0 \ ,$$

where $t^{(u)}$ is the time of 'speciation' at node u. It is further assumed that the tree is bifurcating, i.e. all internal nodes are of degree 3, except for the root vertex, which may have degree 2 or 3.

We consider sequences as infinite words in an m-letter alphabet. For definiteness, this is implied to be a nucleotide alphabet, i.e., $m = 4$. We will correspondingly speak of the relative frequencies of the letters in a sequence of infinite length (they are tacitly assumed to exist !) as the 'base composition'. However, any other alphabet may be used interchangeably.

Let $\mathcal{T}$ be the 'true tree' along which sequence evolution has taken place. In line with standard assumptions of sequence evolution (for review, see Swofford *et al.*, 1995), we consider a substitution process which acts in continuous time, independently on every edge of the tree, and in the same way at every site of the sequence. The action on the base composition may then be written as a system of linear differential equations that live on the edges of the tree:

$$\dot{\boldsymbol{x}}^{(w,v)}(t) = Q^{(w,v)} \boldsymbol{x}^{(w,v)}(t), \quad \boldsymbol{x}^{(w,v)}(t^{(v)}) = \boldsymbol{\xi}^{(v)}. \tag{1}$$

With a suitable indexing of the letters, $\boldsymbol{x}^{(e)}(t) := (x_1^{(e)}(t), \ldots, x_m^{(e)}(t))^T$ is the vector of relative frequencies at time t on edge e with $\sum_\alpha x_\alpha^{(e)}(t) = 1$. $\boldsymbol{\xi}^{(v)}$ is the base composition at node v. Finally, $Q^{(e)}$ is the matrix of instantaneous substitution rate constants for edge e; namely, for $\beta \neq \alpha$, $Q_{\beta\alpha}^{(e)} := Q_{\beta\leftarrow\alpha}^{(e)} \geq 0$ denotes the rate at which nucleotide α is replaced by nucleotide β, and

$$Q_{\alpha\alpha}^{(e)} := -\sum_{\beta\neq\alpha} Q_{\beta\alpha}^{(e)} \, , \tag{2}$$

which guarantees that Q has vanishing column sums[1].

The $Q^{(e)}$ are constant for a given edge but may differ from edge to edge. To ensure the continuity of the process, we additionally require that no jumps occur in the base composition at the nodes, i.e.

$$\boldsymbol{\xi}^{(w)} = \boldsymbol{x}^{(w,v)}(t^{(w)})\,. \tag{3}$$

Equations (1) and (3) together constitute a multipoint boundary value problem on a tree; the base composition at the root serves as the initial value for the whole process. Thus, the problem may be solved by proceeding from the root to the leaves. In particular, for every node,

$$\boldsymbol{\xi}^{(w)} = P^{(w,v)}\boldsymbol{\xi}^{(v)}\,. \tag{4}$$

Here,

$$P^{(w,v)} := \exp(Q^{(w,v)}\tau^{(w,v)}) \tag{5}$$

is the Markov matrix of transition probabilities for that edge.

We will assume that

(A2) $\quad \boldsymbol{\xi}^{(\rho)} > 0$

componentwise, that is, all letters are represented in the root sequence in non-vanishing frequencies. Further, it will be assumed throughout that

(A3) $\quad \widetilde{Q}^{(e)}$ is irreducible for all e,

where $\widetilde{Q}^{(e)}$ is obtained from $Q^{(e)}$ by setting the diagonal elements to zero. A matrix A is called irreducible if, for every pair of indices (β,α), there exists an integer n s.t. $(A^n)_{\beta\alpha} > 0$. This implies that all states communicate. Assumptions **(A1)** and **(A3)** guarantee that the transition matrices $P^{(e)}$ have strictly positive entries only, i.e., every state is populated from every other state. Together with **(A2)** and Eq. (4), it follows that $\boldsymbol{x}^{(e)}(t) > 0$ for all e and t.

We have thus formulated a continuous-time analogue of the discrete Markov process on a tree. The latter was set up by, e.g., Barry and Hartigan (1987) and Steel (1994), on whom we have partially leaned for notation. It is, however, essential to point out some differences.

[1]Note that, in some contexts (e.g. Swofford *et al.*, 1995), it is the *transpose* of Q which is termed substitution rate matrix.

Due to irreducibility and Eq. (2), one has $Q_{\alpha\alpha} < 0$ for all α, and thus $-\infty < \mathrm{tr}(Q) < 0$. Together with **(A1)** and Eq. (5), this implies

$$0 < \det(P^{(e)}) < 1\,, \tag{6}$$

where we have used the identity

$$\exp(\mathrm{tr}(A)) = \det(\exp(A))\,, \tag{7}$$

which holds for every square matrix A.

On the other hand, every Markov transition matrix with strictly positive eigenvalues may be represented in terms of a (continuous-time) rate matrix by inverting Eq. (5):

$$Q^{(e)}\tau^{(e)} = \log P^{(e)}\,. \tag{8}$$

For more on the relationship between continuous and discrete Markov processes, see Karlin and Taylor (1981), Chapter 14; for the application to sequence evolution, see also Steel, Hendy, and Penny (1993).

For transition matrices with one or more non-positive eigenvalue, the log is not defined, and there is no corresponding continuous-time rate matrix. Although we thus cannot hope for the most general mathematical results, the latter processes can be safely excluded from biological considerations.

Eq. (8) must also be relied on in situations where the model changes *during* one edge. Such cases certainly exist: After all, whether or not a path is subdivided by a node depends on what species have been observed. If u is an extra node (of degree 2) that subdivides edge (w,v) into (w,u) and (u,v) with distinct substitution rate matrices, we may think of a rate matrix for (w,v) via

$$Q^{(w,v)}\tau^{(w,v)} := \log[\exp(Q^{(w,u)}\tau^{(w,u)})\exp(Q^{(u,v)}\tau^{(u,v)})] \tag{9}$$

provided this is a legitimate rate matrix; this need not be the case if, loosely speaking, $Q^{(w,u)}$ and $Q^{(u,v)}$ fail to commute 'too badly'. Eq. (9) may be evaluated with the help of the Baker-Campbell-Hausdorff formula (e.g. Miller, 1972, Pp. 161–162):

$$\begin{aligned}\log(\exp(A)\exp(B)) &= A + B + \frac{1}{2}[A,B] + \frac{1}{12}[A,[A,B]] - \frac{1}{12}[B,[B,A]] \\ &\quad + (\text{higher-order commutators})\,,\end{aligned} \tag{10}$$

where $[R,S] := RS - SR$ is the commutator of square matrices R and S. Since commutators have vanishing trace, we may note, for later use:

$$\mathrm{tr}(Q^{(w,v)}\tau^{(w,v)}) = \mathrm{tr}(Q^{(w,u)}\tau^{(w,u)}) + \mathrm{tr}(Q^{(u,v)}\tau^{(u,v)})\,. \tag{11}$$

The idea of edge subdivision is easily generalized to the case where the substitution process changes with time in a continuous way, i.e. $Q^{(e)} = Q^{(e)}(t)$, but we will not bother to spell this out explicitly.

Let us now turn to the divergence matrices. Following Barry and Hartigan (1987) or Steel (1994), let i and j be terminal sequences and $a = a(i,j)$ their most recent common ancestor, i.e. the last vertex common to the paths $p(j,\rho)$ and $p(i,\rho)$ from the root to the terminal sequences in question (see Fig. 1). Note that $p(\ell,k)$, the (directed) path from node k to node ℓ in a tree, is unique, in contrast to the situation in more general graphs. The $m \times m$ matrix $H^{(j,i)}$ of joint probabilities of letters in these sequences (i.e. $H^{(j,i)}_{\beta\alpha}$ is the probability that a site has letter β in sequence j and letter α in sequence i) is

$$H^{(j,i)} = P^{(j,a)}\Xi^{(a)}(P^{(i,a)})^T = (H^{(i,j)})^T \quad \text{for } i \neq j\,. \tag{12}$$

Here, $\Xi^{(u)} := \mathrm{diag}\{\boldsymbol{\xi}^{(u)}\}$ (the 'base composition matrix' at node u), $\mathrm{diag}\{\boldsymbol{v}\} := \mathrm{diag}\{v_1, \ldots, v_m\}$ for an m-vector v, and

$$P^{(k,a)} = \overleftarrow{\prod_{e\in p(k,a)}} P^{(e)}\,, \quad k = i,j\,. \tag{13}$$

The product ranges over all directed edges on the path $p(k,a)$ from the ancestor a to the sequence under consideration. Since the transition matrices need not commute, attention must be paid to the order of multiplication (from right to left, as one ascends from the root to the leaves of the tree).

Let us now turn to the inference problem and assume we have collected a set of divergence matrices $H^{(j,i)}$ from pairwise comparison of (ideal, infinite) sequences. The most modest inference problem is then to identify the topology of the underlying *unrooted* tree as obtained from $\mathcal{T}$ by fusing the edges incident to ρ in case ρ is of degree 2, and removing the orientation of the edges. This inference task requires that we have extracted *tree-additive* distances from the divergence matrices, i.e., distances that may, for every pair of sequences, be written as

$$d^{(j,i)} = \sum_{e\in p(j,i)} \varepsilon^{(e)} \tag{14}$$

with positive and symmetric edge lengths, $\varepsilon^{(w,v)} = \varepsilon^{(v,w)} > 0$. If this is the case, the underlying tree is unique and may be reconstructed in polynomial time (for the theory of additive evolutionary trees, see Waterman *et al.* (1977) or Bandelt and Dress (1986)).

In dealing with distance measures, the definition of the distance usually comes first. It is then decomposed into edge lengths, which must be proved to be additive, symmetric, and positive (see Barry and Hartigan, 1987; Lake, 1994; Steel, 1994). The latter may be a technical and non-intuitive step. We shall, therefore, start out with the process on the edges and then use these to gain an understanding of distance measures.

4 Number of substitutions and time-reversible distance

To set the scene, we will briefly recall the simplest additive distance measure within our own framework. To this end, let us consider the substitution process on one selected edge, starting at $t = 0$ for convenience, and suppress the upper index, i.e.

$$\dot{\boldsymbol{x}} = Q\boldsymbol{x}, \ \boldsymbol{x}(0) = \boldsymbol{\xi}\,. \tag{15}$$

From the irreducibility assumption **(A3)** and the Perron-Frobenius theorem (see e.g. Karlin and Taylor, 1975, Pp. 475–484), one concludes that there is a unique stationary state $\boldsymbol{\pi} > 0$, which attracts all solutions no matter where they start:

$$\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{\pi}\,, \quad \text{independent of } \boldsymbol{\xi}\,. \tag{16}$$

The continuous approximation of the number of substitutions per site that have taken place until t, $z(t)$, is given by

$$\dot{z} = \sum_{\alpha} \sum_{\beta\neq\alpha} Q_{\beta\alpha} x_\alpha = -\sum_{\alpha} Q_{\alpha\alpha} x_\alpha\,, \quad z(0) = 0\,; \tag{17}$$

note that z is strictly increasing since $Q_{\alpha\alpha} < 0$ and $x_\alpha > 0$. If the process is stationary, i.e. $\boldsymbol{x}(t) \equiv \boldsymbol{x}(0) = \boldsymbol{\pi}$ (a standard assumption in tree reconstruction), z increases at a constant rate $\dot{z} = -\mathrm{tr}(Q\Pi)$, where $\Pi := \mathrm{diag}\{\boldsymbol{\pi}\}$. Thus, z is proportional to time,

$$z(t) = -t\,\mathrm{tr}(Q\Pi)\,, \tag{18}$$

which is the usual definition of the number of substitutions (cf. Gu and Li, 1996*b*). If the process is not stationary, the substitution rate need not be constant across an edge, but will converge towards $-\mathrm{tr}(Q\Pi)$ when $\boldsymbol{x}(t)$ converges to $\boldsymbol{\pi}$, either monotonically or in damped oscillations. Note, however, that for rate matrices with $Q_{\alpha\alpha}$ independent of α, like the Jukes-Cantor and Kimura-2-parameter models, the substitution rate is independent of the base composition and thus independent of the initial conditions, as has been remarked by Rzhetsky and Nei (1995).

For a specified edge e, let $\varepsilon_S^{(e)}$ be its length in terms of the number of substitutions that have occurred along it, i.e.

$$\varepsilon_S^{(e)} := z(\tau^{(e)})\,. \tag{19}$$

For pairwise distances of terminal nodes, the situation is greatly simplified under the assumptions that the model is

(R1) homogeneous, i.e. identical on all edges, $Q^{(e)} \equiv Q$,

(R2) stationary, $\boldsymbol{x}^{(e)}(t) \equiv \boldsymbol{\pi}$,

(R3) and reversible, $Q\Pi = \Pi Q^T$.

It is then well known (Lanave *et al.*, 1984; Gu and Li, 1996*b*; Waddell and Steel, 1997) that the number of substitutions that have occurred on the path $p(j,i)$, via $a(j,i)$, may be inferred from the divergence matrix as the so-called 'time-reversible distance', $d_{TR}^{(j,i)}$. For, in this case,

$$\begin{aligned} H^{(j,i)} &= P^{(j,a)}\Pi(P^{(i,a)})^T = P^{(j,a)}P^{(i,a)}\Pi \qquad (20)\\ &= \exp(Q\tau^{(j,i)})\Pi = H^{(i,j)}\,, \qquad (21)\end{aligned}$$

where stationarity and reversibility have been used, and

$$\tau^{(j,i)} := \sum_{e\in p(j,i)} \tau^{(e)} \tag{22}$$

is the duration of the path from i to j (twice the divergence time of i and j in the present case). $d_{TR}^{(j,i)}$ now follows from Eq. (18) as

$$\begin{aligned} d_{TR}^{(j,i)} &= z(\tau^{(j,i)}) = -\mathrm{tr}(\Pi\, Q\tau^{(j,i)}) \qquad (23)\\ &= -\mathrm{tr}(\Pi\,\log(H^{(j,i)}\Pi^{-1})) = d_{TR}^{(i,j)}\,. \qquad (24)\end{aligned}$$

Using Eqs. (19), (22), and the linearity of the trace, one sees that d_{TR} is tree-additive:

$$d_{TR}^{(j,i)} = \sum_{e \in p(j,i)} \varepsilon_S^{(e)} . \tag{25}$$

Note that, with assumption **(R1)**, we have restricted ourselves to a situation with a molecular clock (cf. Swofford *et al.*, 1995), but non-clocklike situations are readily incorporated by redefining the $\tau^{(e)}$.

5 The LogDet family

The pleasant property that makes homogeneous, stationary, reversible processes so amenable to study is that probability fluxes are the same in the forward and the reverse direction. As a consequence, the substitution rate matrix (times the duration of the path) may be inferred from the divergence matrix (cf. Eq. (21)). At the same time, any convenient placement of the root will yield the same number of substitutions; this is used in most, if not all, tree-reconstruction programs (Felsenstein, 1981). If, however, one of the assumptions **(R1)**–**(R3)** is violated, then the (unknown) position of the root (and the induced orientation of the edges and probability fluxes along them) is crucial to inference. In the extreme case where the root is *assumed* to be positioned at node i, and thus all edges on the path from i to j are directed away from i, we would have

$$H^{(j,i)} = P^{(j,i)} \Xi^{(i)}, \tag{26}$$

where $\Xi^{(i)}$ may be read off $H^{(j,i)}$ in terms of its column sums; it is worth emphasizing that the base composition need not be stationary. Following Barry and Hartigan (1987), one would thus infer

$$P^{(j,i)} = H^{(j,i)} (\Xi^{(i)})^{-1}; \tag{27}$$

note that the inverse of $\Xi^{(i)}$ exists, since $\boldsymbol{x}^{(e)}(t)$ is positive for all e and t. At the other extreme, where evolution is assumed to have proceeded from j to i (i.e. all fluxes are reversed), one has, of course, $P^{(i,j)} = H^{(i,j)}(\Xi^{(j)})^{-1}$. The transition matrices $P^{(j,i)}$ and $P^{(i,j)}$ are related via

$$\begin{aligned} H^{(i,j)} &= (H^{(j,i)})^T \\ \Leftrightarrow P^{(i,j)} \Xi^{(j)} &= \Xi^{(i)} (P^{(j,i)})^T \\ \Leftrightarrow \quad P^{(i,j)} &= \Xi^{(i)} (P^{(j,i)})^T (\Xi^{(j)})^{-1} . \end{aligned} \tag{28}$$

Here, $P^{(i,j)}$ is the *reversed* transition matrix w.r.t. $P^{(j,i)}$, which results from replacing probability fluxes from i to j by those from j to i (see e.g. Kemeny and Snell (1981), Pp. 24–26).

We can thus infer the overall transition matrices (recall that these are products along the path according to Eq. (13)) given the extreme placements of the root. If the base composition is changing, however, the number of substitutions that would have occurred cannot be inferred from these transition matrices. In search for other quantities to be counted, let us, therefore, return to the 'true' evolutionary process on the rooted tree $\mathcal{T}$, and consider the *relative* change in frequencies along an edge, as opposed to the *absolute* rates in Eq. (1):

$$(X^{(w,v)})^{-1}\dot{X}^{(w,v)} = \frac{d}{dt}\log X^{(w,v)}, \tag{29}$$

where $X^{(w,v)} := \operatorname{diag}\{\boldsymbol{x}^{(w,v)}\}$, and hence Eq. (29) is also valid elementwise and gives the proper analogue of Eq. (1). This is an idea borrowed from the geometry of population genetics (Akin, 1979, Pp. 37–43) where it has provided insight into concepts of genetic distance between populations as based on allele frequencies. Let us define a corresponding *relative* substitution counter:

$$\begin{aligned} \dot{y}^{(w,v)} &= \sum_\alpha \frac{1}{x_\alpha^{(w,v)}} \sum_{\beta\neq\alpha} Q_{\beta\alpha}^{(w,v)} x_\alpha^{(w,v)} = -\operatorname{tr}(Q^{(w,v)})\,, && (30) \\ y^{(w,v)}(t^{(v)}) &= 0. && (31) \end{aligned}$$

Unlike $\dot{z}$ from Eq. (17), $\dot{y}$ is independent of the base composition; $-\operatorname{tr}(Q^{(w,v)})$ may be interpreted as an 'overall substitution rate constant'. A suitable edge length is thus

$$\varepsilon_{AS}^{(w,v)} := y^{(w,v)}(t^{(w)}) = -\operatorname{tr}(Q^{(w,v)}\,\tau^{(w,v)}) = -\log(\det(P^{(w,v)}))\,, \tag{32}$$

where Eqs. (5) and (7) have again been used. If $Q^{(w,v)}$ is the rate matrix for a *subdivided* edge as in Eq. (9), Eq. (32) is replaced by

$$\varepsilon_{AS}^{(w,v)} = -(\operatorname{tr}(Q^{(w,u)}\tau^{(w,u)}) + \operatorname{tr}(Q^{(u,v)}\tau^{(u,v)}))\,, \tag{33}$$

where Eq. (10) has been employed.

Recall that (w,v) is the *oriented* edge in the *rooted* tree. Thus, $Q^{(w,v)}$, $\dot{y}^{(w,v)}$ and $\varepsilon_{AS}^{(w,v)}$ belong to the 'true' direction of evolution, which we will identify with the 'forward' one. But to the observer of terminal nodes, who is ignorant of the direction, the process could as well have run backwards. Let us therefore consider the relative substitution counter for the reversed process, which results from the forward one by replacing fluxes $Q^{(w,v)}_{\beta\alpha} x^{(w,v)}_{\alpha}$ by $Q^{(w,v)}_{\alpha\beta} x^{(w,v)}_{\beta}$ for $\alpha \neq \beta$, and interchanging the roles of $t^{(v)}$ and $t^{(w)}$:

$$\begin{aligned} \dot{y}^{(v,w)} &= \sum_{\alpha} \sum_{\beta \neq \alpha} \frac{1}{x^{(w,v)}_{\alpha}} Q^{(w,v)}_{\alpha\beta} x^{(w,v)}_{\beta} = -\mathrm{tr}(Q^{(v,w)}) \,, &(34)\\ y^{(v,w)}(t^{(w)}) &= 0 \,, &(35)\end{aligned}$$

where $Q^{(v,w)}$ is the rate matrix of the reversed process:

$$Q^{(v,w)}_{\beta\alpha} := \begin{cases} (X^{(w,v)}(Q^{(w,v)})^T (X^{(w,v)})^{-1})_{\beta\alpha} \,, & \beta \neq \alpha \\ -\sum_{\gamma\neq\alpha} (X^{(w,v)}(Q^{(w,v)})^T (X^{(w,v)})^{-1})_{\gamma\alpha} \,, & \beta = \alpha \,; \end{cases} \qquad (36)$$

compare with the reversed *discrete* transition matrix in Eq. (28). Note that the reversed rate matrix depends on the current base composition and thus on time; as to this problem, see also Kemeny and Snell (1981), Pp. 24–26.

Using Eqs. (1), (29), the fundamental theorem of integration, (10), (7), and (28), one obtains the chain of equations for the corresponding edge lengths:

$$\begin{aligned} \varepsilon^{(v,w)}_{AS} &:= y^{(v,w)}(t^{(v)}) = -\int_{t^{(w)}}^{t^{(v)}} \sum_{\alpha} \sum_{\beta\neq\alpha} (x^{(w,v)}_{\beta})^{-1} Q^{(w,v)}_{\beta\alpha} x^{(w,v)}_{\alpha} dt &(37)\\ &= -\int_{t^{(w)}}^{t^{(v)}} \mathrm{tr}[(X^{(w,v)})^{-1}\dot{X}^{(w,v)} + Q^{(w,v)}]dt &(38)\\ &= -\int_{t^{(w)}}^{t^{(v)}} \mathrm{tr}[\frac{d}{dt}\log X^{(w,v)} + Q^{(w,v)}]dt &(39)\\ &= -\mathrm{tr}[\log(\Xi^{(v)}) - \log(\Xi^{(w)}) + Q^{(w,v)}(t^{(v)} - t^{(w)})] &(40)\\ &= -\log[\det(\Xi^{(v)} \exp(Q^{(w,v)}(t^{(v)} - t^{(w)}))(\Xi^{(w)})^{-1})] &(41)\\ &= -\log[\det(P^{(v,w)})] \,. &(42)\end{aligned}$$

We chose the subscript AS since Barry and Hartigan's (1987) 'asynchronous distance' between terminal nodes i and j,

$$d^{(j,i)}_{AS} := -\log(\det(H^{(j,i)}(\Xi^{(i)})^{-1})) \qquad (43)$$

turns out to coincide with the length of the path directed from i to j. For, from Eqs. (32) and (42), one has

$$\begin{aligned} d_{AS}^{(j,i)} &= -\log(\det(P^{(j,i)})) \\ &= -\log(\det(\overleftarrow{\prod_{e \in p(j,i)}} P^{(e)})) \\ &= \sum_{e \in p(j,i)} \varepsilon_{AS}^{(e)}\,; \end{aligned} \tag{44}$$

see also Barry and Hartigan (1987) and Zharkikh (1994). Note that, for the sake of compatibility within the LogDet family, our definition (43) differs by a factor of 1/4 from Barry and Hartigan's original one.

Obviously, $d_{AS}^{(j,i)} \neq d_{AS}^{(i,j)}$ unless $\det(\Xi^{(j)}) = \det(\Xi^{(i)})$ – thus the 'asynchronous distance' is not a true distance. Also, although d_{AS} is a sum over branch lengths in the sense of Eq. (44), it should *not* be mistaken to be *tree-additive* – note that, in general, $\varepsilon_{AS}^{(w,v)} \neq \varepsilon_{AS}^{(v,w)}$!

To get rid of this ambiguity, let us, instead, take the arithmetic mean of the forward and backward counts:

$$\varepsilon_{PL}^{(w,v)} := \frac{1}{2}(\varepsilon_{AS}^{(w,v)} + \varepsilon_{AS}^{(v,w)}) = \varepsilon_{PL}^{(v,w)}\,. \tag{45}$$

This quantity is symmetric, and positive (starting at 0, it increases with time at a rate which is always positive; cf. the positive right-hand sides of Eqs. (30) and (34)). In labelling ε_{PL}, we have anticipated that these edge lengths belong to the paralinear distance. The latter was defined by Lake (1994) as

$$d_{PL}^{(j,i)} := -\log[\det((\Xi^{(j)})^{-\frac{1}{2}} H^{(j,i)} (\Xi^{(i)})^{-\frac{1}{2}})]\,, \tag{46}$$

which is just the arithmetic mean of the forward and reversed 'asynchronous distances' (cf. Eq. (43)):

$$d_{PL}^{(j,i)} = \frac{1}{2}(d_{AS}^{(j,i)} + d_{AS}^{(i,j)})\,. \tag{47}$$

Taking together Eqs. (45) and (47), one has, for the branch lengths, too:

$$d_{PL}^{(j,i)} = \sum_{e \in p(j,i)} \varepsilon_{PL}^{(e)}\,, \tag{48}$$

which is now *tree-additive* due to the symmetry of ε_{PL}.

Let us finally turn to LogDet distances. They were defined as

$$d_{LD}^{(j,i)} := -\log(\det(H^{(j,i)})) \tag{49}$$

by Steel (1994). Comparison with Eq. (46) shows that

$$d_{LD}^{(j,i)} = d_{PL}^{(j,i)} - \frac{1}{2}\log(\det(\Xi^{(j)})) - \frac{1}{2}\log(\det(\Xi^{(i)}))\,, \tag{50}$$

and thus

$$\begin{aligned} \varepsilon_{LD}^{(w,v)} &= \begin{cases} \varepsilon_{PL}^{(w,v)} & \text{if } (w,v) \text{ is an internal edge} \\ \varepsilon_{PL}^{(w,v)} - \frac{1}{2}\log(\det(\Xi^{(u)})) & \text{if } u \in \{w,v\} \text{ is a terminal node} \end{cases} \\ &= \varepsilon_{LD}^{(v,w)}\,. \end{aligned} \tag{51}$$

Obviously, the only difference lies in an extension of the terminal branches for LogDet distances relative to paralinear distances. Thus, these terminal branch lengths lack the mechanistic interpretation that is available for the internal ones. From this point of view, LogDet distances appear to be conceptually less homogeneous–but, of course, this property does not interfere with additivity.

6 Rate heterogeneity

We are finally ready to take into account rate heterogeneity (or 'site-to-site variation'). This has bothered sequence analysts since the early work of Benzer (1961) and Uzzel and Corbin (1971). For stationary, homogeneous, time-reversible models, good and bad news have since been reported. It is now well-known that, if rate heterogeneity is present but not taken into account in the distance estimate, the number of substitutions is underestimated, regardless of whether the substitution rate matrix is assumed to be known (Golding, 1983; Chang, 1996*b*), or estimated from the divergence matrix (Kelly and Rice, 1996). In either case, underestimation may cause inconsistency in the tree reconstruction procedure, as analyzed in detail by Chang (1996*b*). If, however, the distribution of rates across sites is known or reliably estimated, the problem may be corrected for, and the true number of substitutions recovered to yield additive distances again (cf. Gu and Li, 1996*b*; Waddell and Steel, 1997). In contrast, the situation is far from being understood in the non-stationary case. For this, we shall now present an example which eliminates all hope that *any* pairwise distance measure might distinguish between tree topologies, *even if the distribution of rates is known.*

To this end, let us recall the general situation with rate heterogeneity. It is assumed that, on every edge, all sites evolve according to the same 'structural' model $Q^{(e)}$, but at one out of $n+1$ rates μ_k, with probability ν_k, $0 \leq k \leq n$; these rates are firmly associated with sites throughout the tree. For ease of presentation, we will restrict ourselves to discrete probability distributions but note that continuous ones like the gamma distribution may be tackled in much the same way. Since we are running out of positions to place indices, we will now use the lower one to distinguish between rates. Thus, for every k:

$$\dot{\boldsymbol{x}}_k^{(e)} = \mu_k \, Q^{(e)} \boldsymbol{x}_k^{(e)} \,. \tag{52}$$

Now, $\boldsymbol{x}_k^{(e)}$ is to be interpreted as the base composition of sites of rate μ_k on edge e.

The observed divergence matrix for a pair of sequences is then an average over all rates:

$$H^{(j,i)} = \sum_{k=0}^{n} P_k^{(j,a)} \Xi_k^{(a)} (P_k^{(i,a)})^T \nu_k \,, \tag{53}$$

where

$$P_k^{(\ell,a)} := \overleftarrow{\prod_{e \in p(\ell,a)}} \exp(Q^{(e)} \tau^{(e)} \mu_k) \tag{54}$$

in generalization of definitions (5) and (13).

What makes things difficult is the fact that the process is no longer Markovian (Chang, 1996*b*). Put differently, changes in the base composition (be they caused by non-homogeneity, or non-stationarity) will cause the base composition to depend on the substitution rate at that site, in contrast to the stationary situation.

We shall now demonstrate that no additive distance measure can exist in the nonstationary case (as a function of the divergence matrix), even if the distribution of rates is known. Here, we will provide a numerical example where two different tree topologies produce identical sets of divergence matrices, given the same distribution of rates. A formal proof, together with the construction procedure for a family of examples, is given elsewhere (Baake, 1998), in a more general context of identifiability.

To make things precise, we restrict ourselves to the simplest such situation, namely, a two-letter alphabet, a 4-species tree, and invariable sites. Let us assume throughout that one-half of the sites is invariable, whereas the other half evolves at rate 1, i.e. $\mu_0 = 0$, $\mu_1 = 1$, and $\nu_0 = \nu_1 = 1/2$. Let us then consider topology A of Fig. 2 with terminal nodes $1, \dots, 4$, and internal nodes ρ and σ. Let $\Phi := \mathrm{diag}\{\boldsymbol{\xi}_0^{(\rho)}\}$ be the base composition matrix of the invariable sites, and $\Xi := \mathrm{diag}\{\boldsymbol{\xi}_1^{(\rho)}\}$ the base composition matrix of the variable sites at the root. Further, the $P^{(e)} := P_1^{(e)}$ are the Markov transition matrices for the variable sites, where we have dropped the rate index for ease of notation. According to Eq. (53), the divergence matrices read

$$H^{(j,i)} = \frac{1}{2}(\Phi + P^{(j,a)}\Xi^{(a)}(P^{(i,a)})^T) \tag{55}$$

for all distinct i and j, where $\Xi^{(a)}$ is the base composition matrix at the common ancestor of i and j. Let us now turn to topology B in Fig. 2 with base composition and transition matrices $\widehat{\Phi}$, $\widehat{\Xi}$, and $\widehat{P}^{(e)}$, respectively, which give rise to

$$\widehat{H}^{(j,i)} = \frac{1}{2}(\widehat{\Phi} + \widehat{P}^{(j,a)}\widehat{\Xi}^{(a)}(\widehat{P}^{(i,a)})^T)\,. \tag{56}$$

We now claim that Φ, Ξ, and the $P^{(e)}$, on the one hand, and $\widehat{\Phi}$, $\widehat{\Xi}$, and the $\widehat{P}^{(e)}$, on the other hand, may be arranged such that $H^{(j,i)} = \widehat{H}^{(j,i)}$ for all $i \neq j$. That this is indeed the case is demonstrated in the example in Tab. 1. We may conclude that no function of the divergence matrix will be able to distinguish between topologies A and B, and thus no additive pairwise distance measure may exist.

7 Discussion

Pairwise distance measures are invaluable to phylogenetic inference for at least two reasons. Firstly, they directly give rise to tree reconstruction methods like neighbour joining (Saitou and Nei, 1987), which have a comparatively simple structure and are computationally efficient. Secondly, they are indispensable for the theoretical underpinning of statistically more elaborate methods like maximum likelihood (Felsenstein, 1981), which make better use of the information in the data but are computationally by far more demanding. For both classes of methods, pairwise joint distributions set the starting point for the investigation of identifiability and consistency. Even more so, this is true for the joint distribution of *triples* of sequences, whose properties have been analyzed in detail (Chang, 1996*a*). It has, in fact, been shown that these allow the identification of the full set of Markov transition matrices, provided these belong to a class of matrices that may be reconstructed from the set of their columns. But the transition matrices arising from the continuous-time process (via Eq. (5)) do not belong to this class, and so this identifiability result is, unfortunately, not relevant to our class of problem.

At the same time, Chang (1996*a*) demonstrated that the joint distribution of *pairs* is not sufficient to determine the $P^{(e)}$ completely, not even for the models in his special class; there remains some ambiguity related to the base composition at internal nodes. Identifiability of the *topology* remains, however, via the LogDet and paralinear distance measures, which are additive even if the substitution model changes from edge to edge (and even if it changes continuously, see the remark in Section 3). We have shown in this paper that, apart from the identification of the topology, these measures also yield interpretable branch lengths: The length of a branch is the arithmetic mean of the integral of the (negative of the) trace of the substitution rate matrix of the forward and the reversed process on the branch, where the averaging reflects our ignorance of the direction. They may be interpreted in terms of a relative counting process which counts substitution events weighted with the inverse of the frequency of the letter that is undergoing substitution. Let us remark that Barry and Hartigan (1987) themselves noted this weighted counting process but felt uncomfortable with it. We claim, however, that it gives perfectly reasonable branch lengths, in line with the relative change in frequencies as well-known from the Antonelli-Shahshahani metric and related distance concepts in population genetics (Antonelli and Strobeck, 1977; Shahshahani, 1979; Akin, 1979, Pp. 37–42).

Whereas the case without rate heterogeneity is, by and large, well-behaved, the situation changes when substitution rates vary across sites. If the rate distribution is unknown and arbitrary, Steel, Székely, and Hendy (1994) have shown that, under certain substitution models, it is possible for *every* tree to produce the *same* joint distribution of letters at the leaves. If the rate distribution is known, however, recovery of the number of substitutions from the divergence matrix is possible for the homogeneous, stationary, reversible case (Gu and Li, 1996*b*; Waddell and Steel, 1997). But no such remedy is known if the base composition is changing; see the discussions by Swofford *et al.* (1995) and Waddell and Steel (1997). Our example demonstrates that things are even worse: Even under the same distribution of rates, different topologies may produce identical divergence matrices. As a consequence, no additive pairwise distance measure may exist.

In this particular example, chosen for its simplicity, the parameter to be varied was the base composition of the invariable sites. It might be argued that, under favourable circumstances, this quantity might be accessible from the sequence data. But it is clear that, in situations with more rate classes (or even a continuous distribution of rates), the same ambiguity will exist – but here the estimation problem certainly becomes forbidding.

It might also be argued that, even if we cannot distinguish between topologies on the basis of pairwise distances alone, this might well be possible if joint distributions of *triples* are considered. We do, indeed, not know the answer at this stage. It is obvious, however, that even a maximum-likelihood procedure would require the estimation (or specification) of all of the $\boldsymbol{x}_k$ at at least one point in the tree, which seems a daunting task.

Acknowledgements

It is our pleasure to thank Joe Chang and Mike Steel for illuminating discussions, and for critically reading the manuscript. Financial support from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

References

AKIN, E. 1979. The Geometry of Population Genetics, *Lect. Notes Biomath.*, vol. 31. Springer, New York.

ANTONELLI, P. L. and C. STROBECK. 1977. The geometry of random drift I. Stochastic distance and diffusion. Adv. Appl. Prob. **9**, 238–249.

BAAKE, E. 1998. What can and what cannot be inferred from pairwise sequence comparisons? Math. Biosci. **154** (1998), 1–21.

BANDELT, H.-J. and A. DRESS. 1986. Reconstructing the shape of a tree from observed dissimilarity data. Adv. Appl. Math. **7**, 309–343.

BARRY, D. and J.A. HARTIGAN. 1987. Asynchronous distance between homologous DNA sequences. Biometrics **43**, 261–276.

BENZER, S. 1961. On the topography of the genetic fine structure. Proc. Natl. Acad. Sci. USA **47**, 403–415.

CHANG, J. T. 1996*a*. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Math. Biosci. **137**, 51–73.

CHANG, J. T. 1996*b*. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. **134**, 189–215.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. **17**, 368–376.

GOLDING, B. 1983. Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol. Biol. Evol. **1**, 125–142.

GU, X. and W.-H. LI. 1996*a*. Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol. **13**, 1375–1383.

GU, X. and W.-H. LI. 1996*b*. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc. Natl. Acad. Sci. USA **93**, 4671–4676.

KARLIN, K. S. and H. M. TAYLOR. 1975. A first course in stochastic processes, 2nd edition. Academic Press, London.

KARLIN, K. S. and H. M. TAYLOR. 1981. A second course in stochastic processes. Academic Press, London.

KELLY, C. and J. RICE. 1996. Modelling nucleotide evolution: A heterogenous rate analysis. Math. Biosci. **133**, 85–109.

KEMENY, J. G. and J. L. SNELL. 1981. Finite Markov chains. Springer, New York.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA **91**, 1455–1459.

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

Lockhart, P., M. Steel, M. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**, 605–612.

Miller, W. 1972. Symmetry groups and their applications. Academic Press, London.

Rzhetsky, A. and M. Nei. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. **12**, 131–151.

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

Shahshahani, S. 1979. A new mathematical framework for the study of linkage and selection. *Memoirs Amer. Math. Soc.* **211**.

Steel, M., M. Hendy, and D. Penny. 1993. Invertible models of sequence evolution. Technical Report 93/102, School of Mathematical and Information Scienes, Massey University, Palmerston North, New Zealand.

Steel, M., L. Székely, and M. Hendy. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comp. Biol. **1**, 153–163.

Steel, M. A. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. **7**, 19–23.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1995. Phylogenetic inference. Pp. 407–514 *in*: D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular Systematics. Sinauer, Sunderland.

Uzzel, T. and K. W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**, 1089–1096.

Waddell, P. J. and M. A. Steel. 1997. General time reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites. Mol. Phyl. Evol. **8** (1997), 398–414.

Waterman, M. S., T. F. Smith, M. Singh, and W. A. Beyer. 1977. Additive evolutionary trees. J. Theor. Biol. **64**, 199–213.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39**, 315–329.

8 Figure Legends

Fig. 1: A rooted (phylogenetic) tree with $N = 7$ terminal nodes. Open circles: terminal nodes (or leaves), representing the observed sequences. Bullets: internal nodes. $a^{(j,i)}$ is the most recent common ancestor of leaves i and j. The bold line indicates $p(j, i)$, the path from i to j. Substitution rate matrices $Q^{(w,v)}$ act on edges (w, v).

Fig. 2: Topologies A and B which produce identical sets of divergence matrices. Trees refer to the variable sites (with base composition matrices Ξ and $\widehat{\Xi}$ at the root ρ), whereas the 'immobile' dots correspond to the invariable sites (with base composition matrices Φ and $\widehat{\Phi}$). Markov transition matrices are $P^{(w,v)}$ and $\widehat{P}^{(w,v)}$ for topologies A and B, respectively. Numerical values are given in Table 1.

Table 1: Example of base composition and transition matrices for topologies A and B of Fig. 2 which produce identical divergence matrices

Φ[a]	Ξ	$P^{(1,\rho)}$	$P^{(2,\sigma)}$	$P^{(3,\sigma)}$	$P^{(4,\rho)}$	$P^{(\sigma,\rho)}$
$\begin{pmatrix}.76 & .00\\ .00 & .24\end{pmatrix}$	$\begin{pmatrix}.33 & .00\\ .00 & .67\end{pmatrix}$	$\begin{pmatrix}.91 & .05\\ .09 & .94\end{pmatrix}$	$\begin{pmatrix}.60 & .17\\ .40 & .83\end{pmatrix}$	$\begin{pmatrix}.55 & .23\\ .45 & .77\end{pmatrix}$	$\begin{pmatrix}.46 & .27\\ .54 & .73\end{pmatrix}$	$\begin{pmatrix}.94 & .03\\ .06 & .97\end{pmatrix}$
$\widehat{\Phi}$	$\widehat{\Xi}$	$\widehat{P}^{(1,\rho)}$	$\widehat{P}^{(2,\rho)}$	$\widehat{P}^{(3,\sigma)}$	$\widehat{P}^{(4,\sigma)}$	$\widehat{P}^{(\sigma,\rho)}$
$\begin{pmatrix}.61 & .00\\ .00 & .39\end{pmatrix}$	$\begin{pmatrix}.48 & .00\\ .00 & .52\end{pmatrix}$	$\begin{pmatrix}.88 & .12\\ .12 & .88\end{pmatrix}$	$\begin{pmatrix}.75 & .20\\ .25 & .80\end{pmatrix}$	$\begin{pmatrix}.72 & .27\\ .28 & .73\end{pmatrix}$	$\begin{pmatrix}.67 & .32\\ .33 & .68\end{pmatrix}$	$\begin{pmatrix}.98 & .02\\ .02 & .98\end{pmatrix}$

$H^{(2,1)}$[b]	$H^{(3,1)}$	$H^{(4,1)}$	$H^{(3,2)}$	$H^{(4,2)}$	$H^{(4,3)}$
$\begin{pmatrix}.47 & .07\\ .08 & .38\end{pmatrix}$	$\begin{pmatrix}.47 & .08\\ .08 & .37\end{pmatrix}$	$\begin{pmatrix}.46 & .09\\ .09 & .36\end{pmatrix}$	$\begin{pmatrix}.45 & .10\\ .10 & .35\end{pmatrix}$	$\begin{pmatrix}.44 & .11\\ .10 & .35\end{pmatrix}$	$\begin{pmatrix}.44 & .11\\ .11 & .34\end{pmatrix}$

[a] Figures are rounded to two significant digits throughout. Any discrepancy that might occur between $H^{(j,i)}$ and $\widehat{H}^{(j,i)}$ is due to rounding errors.

[b] $H^{(j,i)} = \widehat{H}^{(j,i)}$ for $i \neq j$. $H^{(j,i)}$ and $\widehat{H}^{(j,i)}$ are from Eqs. (55) and (56), respectively.

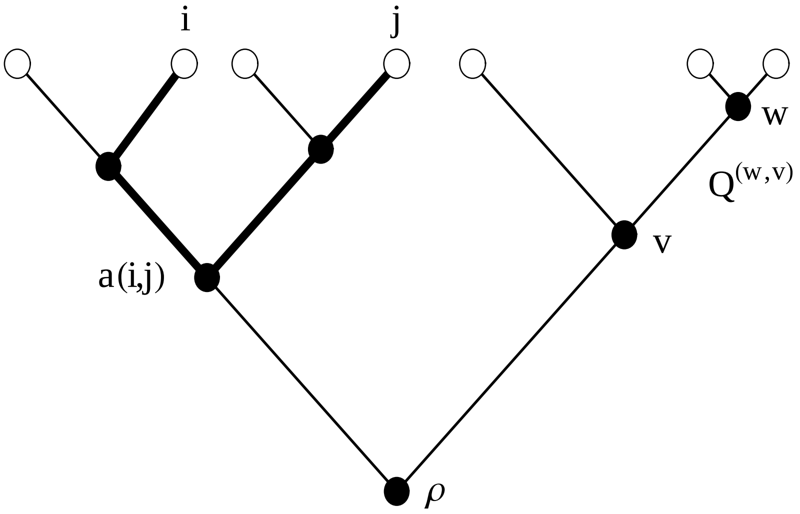
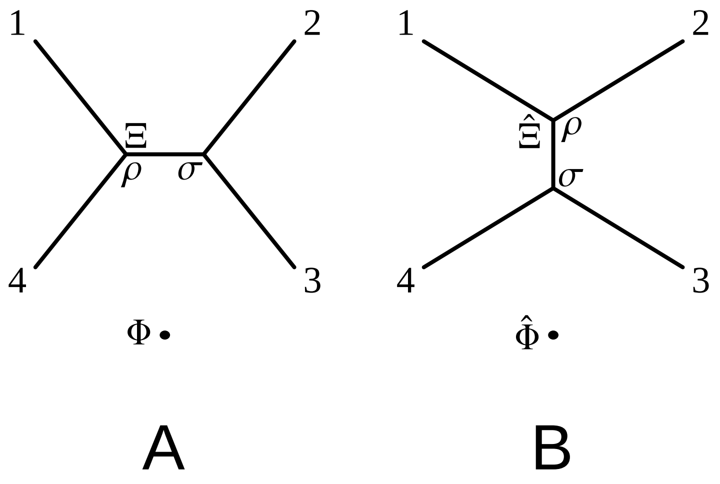


Figure 1:



$\Uparrow$

Figure 2: