# What Can and What Cannot be Inferred From Pairwise Sequence Comparisons ?

Ellen Baake

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

**Running head:** Inference from pairwise sequence comparisons

**Keywords:** Phylogenetic reconstruction; Markov processes on trees; additive distances; rate heterogeneity; identifiability.

phone: (+49) 89 5902570,

Fax: (+49) 89 5902474,

email: baake@zi.biologie.uni-muenchen.de

appeared in: Math. Biosci. **154** (1998), 1–21.

# 1 Abstract

We address questions of identifiability in molecular phylogeny, the art of reconstructing the history of a sample of sequences given just the sequences at the leaves of the phylogenetic tree. Here, the 'history' consists of the tree topology, plus the transition probabilites which define the Markov process of sequence evolution along the branches of the tree.

It is assumed that sequences have infinite length, and the pairwise joint distributions of letters at the leaves is taken to be known. We focus on two cases:

1) If the sites of a sequence evolve identically and independently, the topology can be reconstructed, but the one-way edge transition matrices cannot. However, the *return-trip* transition matrices are reconstructible for every edge, up to conjugacy in the case of internal edges.

2) If a rate factor varies from site to site, different topologies may produce identical pairwise joint distributions, even *under the same distribution of rate factors.* Consequently, identifiability of the topology is lost on the basis of pairwise sequence comparisons, even if the distribution of rate factors is known.

The results are discussed in the context of additive measures of phylogenetic distance.

# 2 Introduction

Questions of identifiability abound in the sciences. Only in lucky cases do experiments yield the parameters of interest directly; usually, these must be inferred from observed quantities. In the context of historical considerations, like the reconstruction of evolutionary trees, making full use of the available information is imperative, since we are left with the results of Nature's experiment; no additional experiments can be performed to unravel hidden quantities. Possibilities and limits of inference thus gain a special importance.

The data available to molecular phylogeny are the sequences of contemporary organisms. When these are considered as the leaves of an evolutionary tree, the task is to extract as much information as possible about the underlying tree, and the evolutionary process that took place along it.

First of all, this requires some framework in terms of a general model class of the evolutionary process. Every *node of the tree* is identified with a sequence. A *sequence* consists of an infinite number of *sites*, which are occupied by letters from a given alphabet (e.g. the nucleotide alphabet {A,G,C,T}). Sequences are subject to change due to substitution of letters by other letters at (some of) the sites. According to a standard assumption (for a review of the field, see [1]), this is a Markov process which acts identically and independently at every site of the sequence; in this respect, sites are *indistinguishable.* Also, substitution proceeds independently on the edges of the tree, but the transition matrices may vary from edge to edge. The inference task then consists of the determination of the graph structure or the 'topology' of the tree (a discrete problem), as well as the Markov transition matrices across the edges (a continuous problem).

Maybe a word is in order on the assumption of infinite sequences—an idealization which can never be met with 'real' data. Considering them, however, provides us with information about the *consistency* of reconstruction methods: If one can tell two models (or trees) apart for infinte sequences, then one can tell them apart using sufficiently long (but finite) sequences. But if one cannot distinguish between them on the basis of infinitely long sequences, then one cannot hope to distinguish between them given just finite sequences (of any length).

To a large extent, molecular phylogeny relies on *pairwise* comparisons of sequences and the resulting estimates of the joint probabilities of letters at corresponding sites of pairs of aligned sequences. It is now well known (e.g. [2, 3, 4]) that, under the above assumptions and some mild extra conditions, knowledge of the pairwise joint probabilities allows the determination of the tree topology. On the other hand, it has been shown [2, 5] that joint probabilities are insufficient to reconstruct the Markov transition matrices for the edges.

For biological reasons, we are primarily interested in continuous-time processes. Here, complications arise when an overall rate factor in the continuously-generated process is allowed to vary from site to site; i.e. the substitution process is still independent, but no longer identical at the sites. This is the well-known problem of rate heterogeneity. If sites are mistakenly assumed indistinguishable although rate heterogeneity is present, the topology may not be recovered consistently; see [6] and the references therein. It may even be shown that *every* tree topology may produce the same *full* (as opposed to pairwise!) joint probabilities of letters at the leaves [7], if the distribution of rate factors may be chosen in an arbitrary way for every topology. If the distribution is known, however, one might hope for a more favourable situation. In the special case of homogeneous, stationary, and reversible processes at least, knowledge of the rate factor distribution restores identifiability of both topology and transition matrices from pairwise joint distribution of letters [8, 9]. The situation outside this special model class is, however, largely unexplored.

In this paper, we will complement results on identifiability from pairwise joint distributions in both the case of indistinguishable sites, and in the case with rate heterogeneity. As to the former, we will show that the situation is better than it looks. Although the Markov matrices for *one-way* transitions across the edges are not identifiable, the probabilities for *return-trip* transitions (forward and back again along an edge or path) *may* be determined, up to a similarity transform for internal paths. In the case with rate heterogeneity, we will assume that the distribution of rate factors is known, and show that the situation is as bad as it may get. A family of examples will be constructed for which two different topologies exhibit *identical* pairwise joint distributions, even if *the distribition of rate factors is the same for both topologies.* As a consequence, identifiability of the topology is lost. The results will be discussed in the context of pairwise distance measures. Let us now start by establishing the framework more precisely.

# 3 Markov processes on phylogenetic trees

Let us briefly recall the tools required to build phylogenetic trees; the reader is also referred to [4], [5], and [10].

The terminal nodes (or leaves) of a tree $\mathcal{T}$ (i.e. a connected graph without cycles) are labelled $1, \ldots, N$; the internal nodes are of degree at least 3, and unlabelled. $\mathcal{T}$ is unrooted, and its undirected edges (or branches) are unordered pairs of nodes, $\{r, s\}$. We will, however, be mainly concerned with directed edges $e = (s, r)$, pairs of nodes ordered such that $e$ is oriented from $r$ to $s$. Similarly, $p\{r, s\}$ and $p(s, r)$ will denote the undirected path with terminal nodes $r$ and $s$, and the directed path from $r$ to $s$, respectively, where $r$ and $s$ need not be adjacent. Note that the path between two nodes in a tree is unique, in contrast to the situation in more general graphs. One special way of introducing orientations for *all* edges of a tree is to introduce a root $\rho$, either by distinguishing an existing node (i.e. of degree $\neq 2$), or by subdividing an existing edge by an additional node, which then has degree 2. The resulting rooted tree is denoted as $(\rho, \mathcal{T})$, and its edges $e = (s, r)$ are oriented such that $r$ is on the path from $\rho$ to $s$.

Rooted trees are required to describe the evolutionary process forward in time. Let $(\rho, \mathcal{T})$ be the 'true tree' along which sequence evolution has taken place. In line with standard assumptions of sequence evolution (for review, see [1]), we consider sequences as infinite words in an $m$-letter alphabet. The substitution process acts on them independently and identically at every site of the sequence, and independently, but not necessarily identically, at every edge of the tree[1]. With a suitable indexing of letters, let $\boldsymbol{\xi}^{(w)} := (\xi_1^{(w)}, \ldots, \xi_m^{(w)})^T$ be the vector of marginal probabilities of letters (i.e. the relative frequencies in a sequence of infinite length) at node $w$; these quantities are tacitly assumed to exist. Of course, $\xi_\alpha^{(w)} \geq 0$ for all letters $\alpha$, and $\sum_\alpha \xi_\alpha^{(w)} = 1$. In the case of the nucleotide alphabet, $m$ equals 4, and $\boldsymbol{\xi}^{(w)}$ holds the base composition of the sequence at node $w$. Let now $P^{(e)}$ be the Markov matrix of transition probabilities for edge $e$. Namely, $P_{\beta\alpha}^{(e)} := P_{\beta\leftarrow\alpha}^{(e)} \geq 0$ denotes the probability of $\alpha$ being replaced by $\beta$ across $e$, where $\sum_\beta P_{\beta\alpha}^{(e)} = 1$. Given the *evolutionary tree* $\mathcal{E} := (\rho, \mathcal{T}, P; \boldsymbol{\xi})$, to be read as 'tree $\mathcal{T}$, rooted at $\rho$, with transition matrices $P^{(e)}$ for edges directed away from $\rho$, and letter frequencies $\boldsymbol{\xi}^{(\rho)}$ at the root', the stochastic process, proceeding from the root to the leaves of the tree, is well-defined. That is, the full joint distributions of letters are uniquely determined for all nodes of the tree. In particular, the letter frequencies are defined recursively via

$$\boldsymbol{\xi}^{(s)} = P^{(s,r)} \boldsymbol{\xi}^{(r)} . \tag{1}$$

We are particularly, but not exclusively, interested in transition matrices arising from continuous-time processes, since they provide the adequate description of gradual sequence change. They may be represented as

$$P^{(e)} := \exp(Q^{(e)} \tau^{(e)}) , \tag{2}$$

where $Q^{(e)}$ is the instantaneous matrix of transition *rates* ($Q_{\beta\alpha}^{(e)} \geq 0$ for $\beta \neq \alpha$, $\sum_\beta Q_{\beta\alpha}^{(e)} = 0$ for all $\alpha$), and $\tau^{(e)} = \tau^{(s,r)} = \tau^{(r,s)} > 0$ is the time duration of that edge.

[1]Note that it is sometimes assumed that the *letters* at the sites are independent and identically distributed random variables, e.g. [11]. This is, however, not necessary; what *is* actually required are the above assumptions on the *substitution process*. In fact, sites may be highly correlated to start with (imagine a pseudogene evolving from a gene, for example) and only gradually lose their correlation in the course of the substitution process.

In any case, we will avoid degenerate cases and restrict ourselves to situations where

**(A1)** $\boldsymbol{\xi}^{(\rho)} > 0$ componentwise,

**(A2)** $P^{(e)}$ has strictly positive entries for all $e$, and

**(A3)** $\det P^{(e)} \neq 0, \pm 1$.

Let us now change our point of view to that of the present-day observer who has collected $N$ sequences, to be identified with the leaves of $\mathcal{T}$. He seeks to reconstruct $\mathcal{T}$, but will, in general, remain ignorant of the location of $\rho$. The edges will, therefore, lack a distinguished orientation. We are thus led to consider the backward process (i.e. relative to the orientation induced by $\rho$) on an edge or path. The reverse process is obtained from the forward one by replacing probability fluxes from $\alpha$ to $\beta$ by those from $\beta$ to $\alpha$ (see e.g. [12]). Due to (A1) and (A2),

$$\boldsymbol{\xi}^{(w)} > 0 \tag{3}$$

for all nodes $w$, and the backward transition matrices are well-defined:

$$P^{(r,s)} = \Xi^{(r)} \big(P^{(s,r)}\big)^T \big(\Xi^{(s)}\big)^{-1} . \tag{4}$$

Here, $P^{(r,s)}$ denotes the *reversed* transition matrix with respect to $P^{(s,r)}$, $T$ denotes transpose, and $\Xi^{(w)} := \operatorname{diag}\{\boldsymbol{\xi}^{(w)}\}$. We will loosely speak of $\Xi^{(w)}$ as the frequency matrix at node $w$.

With transition matrices in both directions now available, we are ready to write down the transition matrices over arbitrary directed paths in the tree:

$$P^{(s,r)} = \overset{\leftarrow}{\prod_{e \in p(s,r)}} P^{(e)} \tag{5}$$

with $P^{(r,r)} = I$ (the identity matrix) for the empty path $p(r,r)$. Since the transition matrices need not commute, attention must be paid to the order of multiplication.

Our interest now centres on the pairwise joint probabilities of letters in pairs of aligned sequences. They are conveniently collected into matrices $H^{(s,r)}$, where $H^{(s,r)}_{\beta\alpha}$ is the probability that a site has letter $\alpha$ in sequence $r$, and letter $\beta$ in sequence $s$. Due to Eqns. (4) and (5), they may be expressed as

$$H^{(s,r)} = P^{(s,w)} \Xi^{(w)} (P^{(r,w)})^T = (H^{(r,s)})^T \,, \tag{6}$$

where $w$ may be any node in $p(s,r)$. Note that $\boldsymbol{\xi}^{(s)}$ and $\boldsymbol{\xi}^{(r)}$ hold the row and column sums, respectively, of $H^{(s,r)}$, and $\sum_{\beta,\alpha} H^{(s,r)}_{\beta\alpha} = 1$.

If the process has a continuous-time representation, the situation is greatly simplified under the assumptions that the model is

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

**(S)** stationary, $\boldsymbol{\xi}^{(e)} \equiv \boldsymbol{\pi}$,

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

Here, $\boldsymbol{\pi}$ is the Perron-Frobenius eigenvector of $Q$, and $\Pi := \mathrm{diag}\{\boldsymbol{\pi}\}$. For more details on continuous-time substitution processes, see, e.g., [13]. For HSR processes (i.e. processes for which (H),(S), and (R) hold), the forward and reverse transition matrices are identical,

$$P^{(s,r)} = P^{(r,s)} \,, \tag{7}$$

and the divergence matrices take the simple form

$$H^{(s,r)} = \exp(Q\tau^{(s,r)})\Pi = H^{(r,s)} \,, \tag{8}$$

where

$$\tau^{(s,r)} := \sum_{e \in p(s,r)} \tau^{(e)} \tag{9}$$

is the duration of the path from $r$ to $s$ (for consistency, $\tau^{(r,r)} = 0$ for the empty path).

Whereas pairwise joint distributions of letters may be written down for arbitrary nodes $r, s$ in the tree if $\mathcal{E}$ is known, our present-day observer can only access them for the leaves. To be more precise, comparisons of finite sequences will give him estimates of the so-called *divergence matrices* $H^{(j,i)}$, $i, j = 1, \ldots, N$. We will now embark on what information he can retrieve from these when the estimates converge to their true values, i.e. in the limit of infinite sequence length.

# 4 Indistinguishable sites: Identifiability of return-trip transition probabilities

Returning to the general process obeying (A1)–(A3) and using the divergence matrices in the representation

$$H^{(j,i)} = P^{(j,i)}\Xi^{(i)} \tag{10}$$

(cf. Eq. (6)), it is obvious that the transition matrix for the path from leaf $i$ to leaf $j$ is identifiable as

$$P^{(j,i)} = H^{(j,i)}(\Xi^{(i)})^{-1}\,; \tag{11}$$

see [10]. Note that $\Xi^{(i)}$ is a function of $H^{(j,i)}$ via the column sums. Although these *'leaf-to-leaf transition matrices'* do not allow the reconstruction of the stochastic process, they give access to the tree topology. For it is now well known that the so-called *paralinear distance*

$$d^{(i,j)} := -\frac{1}{2}(\log|\det(P^{(i,j)})| + \log|\det(P^{(j,i)})|) \tag{12}$$

is a *tree-additive distance measure*, i.e. for any pair of leaves it may be written as

$$d^{(j,i)} = \sum_{e\in p(j,i)} \varepsilon^{(e)} \tag{13}$$

with positive and symmetric edge lengths $\varepsilon^{(s,r)} = \varepsilon^{(r,s)} > 0$; cf. [3], see also [5]. A tree-additive distance measure is sufficient to recover the topology of the unrooted tree, as has been reviewed in [14] and [15].

On the grounds of the topology thus identified, one would wish to infer the collection of $P^{(e)}$ and $\boldsymbol{\xi}^{(r)}$, but Chang and Hartigan [2, 5] have demonstrated that this is impossible on the basis of pairwise joint distributions. For later use, let us recall their argument here in a modified form that is adapted to our needs. Consider an unrooted tree of leaves $1, 2, 3$, with one interior node $m$. Let $S$ be an invertible matrix with unit column sums (single elements of which may, however, be negative, i.e. $S$ need not be a stochastic matrix), and consider the following transformation of the frequency and transition matrices:

$$\widetilde{P}^{(i,m)} := P^{(i,m)}S^{-1}\,, \quad i = 1, 2, 3, \tag{14}$$

$$\widetilde{\Xi}^{(m)} := S\,\Xi^{(m)}S^{T}\,. \tag{15}$$

This induces the transformation of the reverse transition matrices, cf. Eq. (4):

$$\widetilde{P}^{(m,i)} = S P^{(m,i)} . \tag{16}$$

If $\widetilde{P}^{(m,i)}$ and $\widetilde{P}^{(i,m)}$ have non-negative entries for $i = 1, 2, 3$, and $S\,\Xi^{(m)} S^T$ is a diagonal matrix with non-negative entries (explicit examples have been provided in [5], but a general characterization of admissible matrices $S$ seems difficult), then the transformed quantities qualify as stochastic matrices and frequency matrix, respectively. Also, it is easily verified that $\widetilde{\Xi}^{(m)} = \mathrm{diag}\{\tilde{\boldsymbol{\xi}}^{(m)}\}$, where $\tilde{\boldsymbol{\xi}}^{(m)} = S\,\boldsymbol{\xi}^{(m)}$.

Obviously, the divergence matrices are invariant under this transformation,

$$H^{(j,i)} = P^{(j,m)} \Xi^{(m)} \big(P^{(i,m)}\big)^T = \widetilde{P}^{(j,m)} \widetilde{\Xi}^{(m)} \big(\widetilde{P}^{(i,m)}\big)^T , \quad i \neq j, \tag{17}$$

and so is the base composition of the leaves: $\boldsymbol{\xi}^{(i)} = P^{(i,m)} \boldsymbol{\xi}^{(m)} = \widetilde{P}^{(i,m)} \tilde{\boldsymbol{\xi}}^{(m)}$. This is exactly Chang and Hartigan's example of non-identifiability of the edge transition matrices from pairwise joint distributions of letters at the leaves. Let us now proceed by observing that the transition matrix for the return trip from any leaf $i$ to $m$ and back again is also invariant under the transformation (14)–(16):

$$P^{(i,m)} P^{(m,i)} = P^{(i,m)} \Xi^{(m)} \big(P^{(i,m)}\big)^T \big(\Xi^{(m)}\big)^{-1} = \widetilde{P}^{(i,m)} \widetilde{P}^{(m,i)} . \tag{18}$$

We will denote products $P^{(s,r)} P^{(r,s)}$ as *round-trip transition matrices*, as opposed to the *one-way* edge (or path) transition matrices $P^{(s,r)}$. Applying Eq. (4) to $P^{(i,m)}$ and $P^{(m,i)}$, we note that the round-trip transitions are reversible, even if the one-way transitions are not:

$$\begin{aligned} P^{(i,m)} P^{(m,i)} \Xi^{(i)} &= P^{(i,m)} \Xi^{(m)} \big(P^{(i,m)}\big)^T && (19) \\ &= \Xi^{(i)} \big(P^{(i,m)} P^{(m,i)}\big)^T . && (20) \end{aligned}$$

It is now easy to see that it is just these return-trip transition matrices that are identifiable from the divergence matrices. Indeed, we have

**Proposition 1** *Consider an evolutionary tree $\mathcal{E} = (\rho, \mathcal{T}, P; \boldsymbol{\xi})$, and an undirected path $p \in \mathcal{T}$ whose endpoints are not of degree 2. Then the following matrices are identifiable from pairwise joint distributions of letters at the leaves:*

*a) the leaf-to-leaf transition matrices* $P^{(i,j)}$ *and* $P^{(j,i)}$*, if* $p = p\{i, j\}$ *with* $i$ *and* $j$ *leaves.*

*b) the return-trip transition matrix* $P^{(i,m)}P^{(m,i)}$*, if* $p = p\{i, m\}$*,* $i$ *a leaf and* $m$ *an internal node.*

*c) the return-trip transition matrices up to a similarity transform, namely* $P^{(i,m)}P^{(m,n)}P^{(n,m)}(P^{(i,m)})^{-1}$*, if* $p = p\{m, n\}$*,* $m, n$ *are internal nodes, and* $i$ *is any leaf for which edge* $\{m, n\} \notin p\{i, m\}$*.*

PROOF: a) is a recapitulation of Eq. (11).

b) We use the fact that every interior node $m$ of degree $d$ has an associated partition $\{A_1, \ldots, A_d\}$ of nonempty sets of terminal nodes (see Fig. 1). Without loss of generality, we assume that $A_1$ contains node $i$. Choose $j \in A_2$, and $k \in A_3$. As is illustrated in Fig. 2, we may then infer $P^{(i,m)}P^{(m,i)}$ from the leaf-to-leaf transition matrices via

$$\begin{aligned} P^{(i,m)}P^{(m,i)} &= P^{(i,m)}P^{(m,j)}(P^{(k,m)}P^{(m,j)})^{-1}P^{(k,m)}P^{(m,i)} \\ &= P^{(i,j)}(P^{(k,j)})^{-1}P^{(k,i)}, \end{aligned} \tag{21}$$

where we have used the Markov property (2), and the invertibility of the $P^{(e)}$ as guaranteed by (A3), and Eqns. (3) and (4).

c) Edge $\{m, n\}$ induces the partitioning of leaves

$$\{\{A_1, \ldots, A_d\}, \{B_1, \ldots, B_e\}\},$$

where $d$ and $e$ are the degrees of $m$ and $n$, respectively (see Fig. 1). Let us again assume that $A_1$ contains node $i$, and choose $j \in A_2$, $k \in B_1$, and $\ell \in B_2$. As illustrated in Fig. 2, one then has

$$P^{(i,m)}P^{(m,n)}P^{(n,m)}(P^{(i,m)})^{-1} = P^{(i,\ell)}(P^{(k,\ell)})^{-1}P^{(k,j)}(P^{(i,j)})^{-1}. \tag{22}$$

Since $i, j, k$ and $\ell$ are leaves, the assertions b) and c) follow from part a). $\square$

REMARK: Identifiability of the return-trip transition matrices up to conjugacy in part c) implies that the characteristic polynomial may be recovered, and thus all information contained in it, like the eigenvalues.

The fact that the one-way edge transition matrices are not identifiable is clearly related to the fact that the letter frequencies at internal nodes cannot be uniquely inferred (see Chang's transformation). In the indistinguishable sites situation tackled in this section, one is left with identifiability of the return-trip matrices. In the case with rate heterogeneity to be treated next, the inaccessibility of certain letter frequencies will even destroy the identifiability of the topology.

# 5 Rate heterogeneity: Non-identifiability of the topology

Beginning with the early work of Benzer [16], it was gradually realized that the sites in a sequence may evolve at very different rates, which reflect the various degrees of functional constraint, or unequal biochemical predisposition to mutate, or both. Such rate heterogeneity is now considered a common phenomenon, which interferes with inference, see e.g. [17], and [1, 19] for reviews.

Rate heterogeneity has been mainly tackled under HSR models (e.g. [6, 8, 9]). We will formalize the general situation here. The basis for the description is the continuous-time substitution process. It is assumed that, on every edge $e$, all sites evolve according to the same 'structural' model $Q^{(e)}$ times one out of $n+1$ rate factors $\mu_k$, with probability $\nu_k$, $0 \leq k \leq n$, $\sum_k \nu_k = 1$. This model may be interpreted in two ways (cf. [7]): (i) Every site of the sequence has its individual well-defined, but *unknown* rate factor. (ii) For every site, a rate factor is randomly and independently selected from a distribution $\vartheta$ which is *constant* over the sites. The distinction between (i) and (ii) bears no consequences for the questions considered here, since we do not attempt to determine the rate parameters of the sites individually (as was done in [18]); only the distribution matters. Note that, in any case, rate factors are firmly associated with sites throughout the tree. Thus, the sites remain independent, but no longer indistinguishable. For ease of presentation, we will restrict ourselves to discrete probability distributions for the rate factors, but note that continuous ones like the popular gamma distribution (cf. [1, 19]) may be tackled in much the same way. Running out of positions to place indices, we will now use the lower one to distinguish between rate factors.

The Markov transition matrix (2) must be replaced by the transition matrices for the sites evolving at rate factor $\mu_k$:

$$P_k^{(e)} = \exp(Q^{(e)}\tau^{(e)}\mu_k)\,, \quad k = 0, \ldots, n\,. \tag{23}$$

The recursion for the relative letter frequencies $\boldsymbol{\xi}_k$ for the sites with rate factor $\mu_k$ reads

$$\boldsymbol{\xi}_k^{(s)} = P_k^{(s,r)} \boldsymbol{\xi}_k^{(r)} \tag{24}$$

in generalization of Eq. (1). In line with (A1), it is assumed that $\boldsymbol{\xi}_k^{(\rho)} > 0$ for all $k$. The matrix of pairwise joint distributions for a pair of sequences is then obtained by averaging the corresponding quantities over all rate factors:

$$H^{(s,r)} = \sum_{k=0}^{n} P_k^{(s,w)} \Xi_k^{(w)} (P_k^{(r,w)})^T \, \nu_k \,, \tag{25}$$

where $w \in p(s,r)$, and transition matrices over directed paths have been used in obvious generalization of Eqs. (4) and (5).

It is important to pay attention to the dependence of the letter frequencies on the rate factor. Even if they are all identical at some node, they tend to diverge across the tree due to Eq. (24), unless the process is stationary.

The evolutionary tree may now be specified as $\mathcal{E} = (\rho, \mathcal{T}, \mathcal{P})$, where $\mathcal{P} := (Q, \vartheta; \boldsymbol{\xi})$ determines the stochastic process. Here, $Q$ is the collection of $Q^{(e)}$ for those edges directed away from $\rho$, $\vartheta$ denotes the distribution of rate factors, and $\boldsymbol{\xi}$ is now to be understood as $(\boldsymbol{\xi}_k^{(\rho)})_{k=0,\dots,n}$.

From the point of view of inference, several cases have been considered with various constraints on the model class. First of all, one has to distinguish whether or not the process is HSR for all rates (which implies $\boldsymbol{\xi}_k^{(w)} \equiv \boldsymbol{\pi}$ for all $w$ and $k$). It must then be decided how much *a priori* information about the distribution of rate factors is available: $\vartheta$ may be considered as completely specified, specified up to one or several free parameters, or an arbitrary probability distribution (recall, however, that $\vartheta$ remains the same for all sites). The HSR case with $\vartheta$ completely specified is comparably well-behaved. It has been shown to be formally identical with the situation with constant rates, if the matrix exponential in Eq. (2) is replaced by the moment-generating function of $\vartheta$. Consequently, *pairwise* joint distributions are sufficient to determine the tree topology, the branch durations, and the substitution rate matrix, see [8, 9, 20, 21]. If, still with an HSR substitution process, $\vartheta$ is specified up to a free parameter, then the $\tau^{(e)}$ and the parameters of both $Q$ and $\vartheta$ may be simultaneously optimized on the basis of the *full* joint disbributions (with the help of maximum-likelihood procedures, e.g. [9, 22, 23]), although identifiability does not seem to have been proved formally. If $\vartheta$ is arbitrary, however, even the HSR situation is hopeless from the point of view of inference. It has, in fact, been shown that, for a special substitution rate matrix (the Cavender-Farris model [24, 25]), $\vartheta$ and the $\tau^{(e)}$ may be arranged for *any* tree to produce the same *full* joint distribution of letters at the leaves [7]. It follows that, in general, $\mathcal{T}$ and $\vartheta$ may not be determined independently of each other; one will always have to rely on external information, e.g. in the form of a parametrized version of $\vartheta$.

Results dwindle as soon as the HSR assumption is abandoned. It is our aim now to explore identifiability in this realm. To this end, we will assume the even favourable case that $\vartheta$ is completely specified, but will allow $Q^{(e)}$ to vary from edge to edge. What can be inferred from pairwise joint distributions in this case? The answer is disappointing: Not even the topology may be reconstructed. To show this, we will now construct a family of examples where two different topologies give rise to identical pairwise joint distributions, given the same $\vartheta$. (A numerical example, without proof, was given in a previous paper [13]). We will restrict ourselves to the simplest such situation, namely, a 4-species tree, a two-letter alphabet, and a fraction of the sites invariable, whereas the others evolve at a constant rate. The examples will exploit the fact that the base composition of variable and invariable sites may differ, maybe due to changes in the letter frequencies of the variable sites in deeper parts of the phylogeny. The construction procedure will involve three steps. First, we will clarify the notation and reformulate the problem. Second, we will see that for every *triple* of terminal nodes, there exists a whole *family* of HSR processes (for the variable sites) which give rise to identical pairwise joint distributions of letters at the leaves. We will finally attach to such an HSR triple a fourth branch with a *non*-stationary process. The attachment will be possible at different positions in the triple, which results in different quartet topologies without changing the pairwise joint distributions.

**Notation and preliminary considerations**. Consider a two-letter alphabet with letters + and −, say. Half the sites are invariable, whereas the others have rate factor 1, i.e. $\mu_0 = 0$, $\mu_1 = 1$, $\nu_0 = \nu_1 = 1/2$. Let $\varphi$ be the fraction of invariable sites with letter '+', $\boldsymbol{\varphi} := (\varphi, 1-\varphi)^T$ the vector of letter frequencies for the invariable sites, and $\Phi := \mathrm{diag}\{\boldsymbol{\varphi}\}$ the corresponding frequency matrix. Likewise, $\xi^{(w)}$ is the fraction of variable sites at node $w$ which read '+', $\boldsymbol{\xi}^{(w)} = (\xi^{(w)}, 1-\xi^{(w)})^T$, and $\Xi^{(w)} := \mathrm{diag}\{\boldsymbol{\xi}^{(w)}\}$. Further, the $Q^{(e)}$ and $P^{(e)}$ are the rate and transition matrices for the variable sites. We have dropped the rate index, so that the lower index is free to enumerate letters again.

With $\vartheta$ fixed for the remainder of the paper, we let $\mathcal{P} := (Q; (\boldsymbol{\xi}, \boldsymbol{\varphi}))$ specify the stochastic process in the evolutionary tree $\mathcal{E} = (\rho, \mathcal{T}, \mathcal{P})$. We now seek situations where evolutionary trees $\mathcal{E}$ and $\bar{\mathcal{E}} = (\bar{\rho}, \bar{\mathcal{T}}, \bar{\mathcal{P}})$, $\bar{\mathcal{P}} := (\bar{Q}; (\bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\varphi}}))$, produce identical pairwise joint distributions of letters at the leaves, i.e. identical $H^{(j,i)}$, $i, j = 1, \dots, N$. If this is the case, we will say that $\mathcal{E}$ and $\bar{\mathcal{E}}$ are *pair-equivalent*.

With $\vartheta$ as specified, the divergence matrices read

$$H^{(j,i)} = \frac{1}{2}(V^{(j,i)} + \Phi), \quad i, j = 1, \dots, N, \tag{26}$$

where $V^{(s,r)}$ denotes the matrix of pairwise joint distributions of the variable sites in sequences $r$ and $s$ of $\mathcal{E}$,

$$V^{(s,r)} = P^{(s,w)} \Xi^{(w)} (P^{(r,w)})^T, \quad w \in p(s,r)\,. \tag{27}$$

The characterization of divergence matrices is greatly simplified by taking into account their normalization and symmetries. Let us consider

$$\eta^{(s,r)} := H^{(s,r)}_{+-} + H^{(s,r)}_{-+} = \eta^{(r,s)} \tag{28}$$

$$\zeta^{(s,r)} := H^{(s,r)}_{+-} - H^{(s,r)}_{-+} = \frac{1}{2}(\xi^{(s)} - \xi^{(r)}) = -\zeta^{(r,s)}\,, \quad \text{and} \tag{29}$$

$$\mu^{(r)} := H^{(s,r)}_{++} + H^{(s,r)}_{-+} = \frac{1}{2}(\xi^{(r)} + \varphi)\,, \tag{30}$$

where we have used Eq. (26) and $\xi^{(r)} = V^{(s,r)}_{++} + V^{(s,r)}_{-+}$. The symmetry of $\eta$ and antisymmetry of $\zeta$ are due to $H^{(s,r)} = (H^{(r,s)})^T$ (cf. Eq. (6)), and $\eta^{(r,r)} = \zeta^{(r,r)} = 0$ since $H^{(r,r)}$ is diagonal. Note that $\eta^{(s,r)}$ is the average Hamming distance per site of sequences $r$ and $s$, i.e. the relative proportion of sites where $r$ and $s$ differ. In contrast, $\zeta^{(s,r)}$ is the change in frequency of '+' sites across the path from $r$ to $s$, and

$$\zeta^{(s,r)} = \zeta^{(s,w)} + \zeta^{(w,r)} \quad \text{for } w \in p(s,r). \tag{31}$$

Finally, $\mu^{(r)}$ is the frequency of '+' sites in sequence $r$.

Clearly, the off-diagonal elements of our divergence matrices are uniquely determined by $\eta^{(j,i)}$ and $\zeta^{(j,i)}$, $j > i$, $i = 1, \dots, N$. To determine the diagonal elements, the frequency of '+' sites at one reference node (say leaf 1) is additionally required. Using Eqns. (28)–(31), one finds

$$H^{(j,i)}_{++} = \mu^{(i)} - \frac{1}{2}(\eta^{(j,i)} - \zeta^{(j,i)}) = \mu^{(1)} + \frac{1}{2}(\zeta^{(j,1)} - \eta^{(j,i)})\,. \tag{32}$$

Finally, the $H^{(j,i)}_{--}$ are determined by the normalization of $H^{(j,i)}$, its elements summing up to 1. We have thus proved

**Lemma 1** *The family of divergence matrices $H^{(j,i)}$, $i = 1, \ldots, N$, is uniquely determined by $\mu^{(1)}$ and $\eta^{(j,i)}$, $\zeta^{(j,i)}$, $j > i$, $i, j = 1, \ldots, N$.*

Next, we have to consider the divergence matrices in terms of the parameters of $\mathcal{P}$. To this end, we introduce the parametrization

$$Q^{(e)} = \begin{pmatrix} -u^{(e)} & v^{(e)} \\ u^{(e)} & -v^{(e)} \end{pmatrix}, \; u^{(e)}, v^{(e)} > 0\,, \tag{33}$$

where we assume $u^{(e)} + v^{(e)} = 1$ throughout. Evaluation of the matrix exponential yields

$$P^{(e)} = \exp(Q^{(e)}\tau^{(e)}) = \begin{pmatrix} 1 - p^{(e)} & q^{(e)} \\ p^{(e)} & 1 - q^{(e)} \end{pmatrix}, \tag{34}$$

where

$$p^{(e)} = u^{(e)}\,(1 - e^{-\tau^{(e)}}), \quad q^{(e)} = v^{(e)}\,(1 - e^{-\tau^{(e)}})\,. \tag{35}$$

We will first turn to homogeneous processes for the variable sites, i.e. $u^{(e)} \equiv u$, $v^{(e)} \equiv v = 1 - u$, which are also stationary, $\boldsymbol{\xi}^{(r)} \equiv \boldsymbol{\pi} = (v, u)^T$. For a Markov model with two states only, like the one considered here, it is easily verified that homogeneity and stationarity jointly imply reversibility, i.e. the process is HSR for the variable sites. Its instantaneous (overall) substitution rate (i.e. $\sum_{\beta\neq\alpha}\sum_{\alpha} Q_{\beta\alpha}\pi_\alpha$) equals $2\omega$, where $\omega := uv$. Explicitly, the divergence matrices read (cf. Eq. (26),(27)):

$$H^{(j,i)} = \frac{1}{2}\begin{pmatrix} v(v + e^{-\tau^{(j,i)}}) + \varphi & uv(1 - e^{-\tau^{(j,i)}}) \\ uv(1 - e^{-\tau^{(j,i)}}) & u(u + e^{-\tau^{(j,i)}}) + 1 - \varphi \end{pmatrix}. \tag{36}$$

With the given constraints on the parameters, and recalling that the position of the root is arbitrary, we may parametrize the evolutionary tree as $\mathcal{E} = (\mathcal{T}, \mathcal{P})$, where $\mathcal{P} := (u, \boldsymbol{\tau}; \varphi)$ denotes the stochastic process and $\boldsymbol{\tau}$ is the collection of the $\tau^{(e)}$. This will now enable us to establish

**Lemma 2** (Pair-equivalence of reversible triples). *Given a triple $\mathcal{T}_3$ of terminal nodes $1, 2, 3$, adjacent to one internal node $m$. Let the substitution process on $\mathcal{T}_3$ be HSR for the variable sites, i.e. $\mathcal{P} = (u, \boldsymbol{\tau}; \varphi)$. Then there is a neighbourhood $U$ of $u$ such that for all $\bar{u} \in U$, there exist $\bar{\boldsymbol{\tau}}$ and $\bar{\varphi}$ so that the evolutionary trees $\mathcal{E} = (\mathcal{T}_3, \mathcal{P})$ and $\bar{\mathcal{E}} = (\mathcal{T}_3, \bar{\mathcal{P}})$ are pair-equivalent.*

Proof: Let us abbreviate edge indices $(i, m)$ by $(i)$, so that the branch durations in $\mathcal{E}$ are $\boldsymbol{\tau} = (\tau^{(1)}, \tau^{(2)}, \tau^{(3)})$, as shown in Fig. 3. We will, however, retain the indices $(j, i)$ for paths from leaf to leaf, and recall from Eq. (9) that branch durations are additive, which means for the triple

$$\tau^{(i,j)} = \tau^{(i)} + \tau^{(j)}, \quad i \neq j,\ i, j = 1, \ldots, N\,. \tag{37}$$

Due to the stationarity of the process, the $\zeta^{(i,j)}$ all vanish. In view of Lemma 1, we must then establish that the quantities $\eta^{(1,2)}, \eta^{(1,3)}, \eta^{(2,3)}$, and $\mu^{(1)}$ remain unchanged when $\mathcal{P}$ is replaced by $\bar{\mathcal{P}}$. For our special model class, they read, in terms of the parameters $u, \boldsymbol{\tau}$, and $\varphi$ (cf. Eq. (36)):

$$\eta^{(i,j)} = \omega(1 - e^{-\tau^{(i)} - \tau^{(j)}})\,, \qquad (\omega = u(1-u)) \tag{38}$$

$$\mu^{(1)} = \frac{1}{2}(1 - u + \varphi)\,. \tag{39}$$

Existence of $U$ and proper ranges for $\bar{\boldsymbol{\tau}}$ and $\bar{\varphi}$ may then be established by a direct application of the Implicit Function Theorem, as long as $\bar{u}$ is close enough to $u$. For later use, we will also invert Eqs. (38) and (39) explicitly. Replacing $u$ by $\bar{u}$ and $\omega$ by $\bar{\omega} = \bar{u}(1 - \bar{u})$, one finds

$$\bar{\varphi} = \varphi + \bar{u} - u\,, \tag{40}$$

$$\bar{\tau}^{(i)} = \frac{1}{2}(\bar{\tau}^{(i,j)} + \bar{\tau}^{(i,k)} - \bar{\tau}^{(j,k)}) \tag{41}$$

for pairwise different $i, j, k$. Here, $\bar{\tau}^{(j,i)} = f(\tau^{(j,i)}; \omega, \bar{\omega})$ with

$$f(\tau; \omega, \bar{\omega}) := -\log(1 - \frac{\omega}{\bar{\omega}}(1 - e^{-\tau})) = \log\omega - \log(\omega e^{-\tau} + \alpha)\,, \tag{42}$$

and $\alpha := \bar{\omega} - \omega$. Note that the decomposition of branch durations into edge durations via inversion of Eq. (37) may be understood as 'tree reconstruction', see [14]. $\square$

We are now ready to attach an additional branch to our reversible triple. The substitution process on the new branch will not be HSR; we must therefore return to *rooted* trees. Let us consider the rooted quartets with leaves

$\{1,2,3,4\}$ of Fig. 4. They are $\mathcal{T}$, the 'star phylogeny' with $\rho$ as the only internal node, as well as two non-degenerate quartets $\hat{\mathcal{T}}$ and $\check{\mathcal{T}}$, both with internal nodes $\rho$ and $\sigma$. In line with previous notation, let index $(i)$ denote the terminal branch leading to leaf $(i)$, i.e. replace $(i,\rho)$ or $(i,\sigma)$; see Fig. 4. Index $(0)$ stands for the interior branch index $(\sigma,\rho)$. The substitution process is taken to be HSR for the variable sites on the subtree $\mathcal{T}_3$ of $\mathcal{T}$ which consists of the leaves $\{1,2,3\}$ and the root $\rho$, as indicated by the solid lines in Fig. 4. That is, $u^{(i)} \equiv u$, $v^{(i)} \equiv v$ for $i = 0, \dots, 3$, and $\boldsymbol{\xi}^{(\rho)} = (v,u)^T$. In contrast, letter frequencies are assumed to change on branch 4, i.e. $u^{(4)} = u + \varepsilon$, $v^{(4)} = v - \varepsilon$, $\varepsilon > 0$ for definiteness. With these constraints on the model class, the process may be parametrized as $\mathcal{P} = (u, \varepsilon, \boldsymbol{\tau}; \varphi)$, where, this time, $\boldsymbol{\tau} = (\tau^{(0)}, \tau^{(1)}, \dots, \tau^{(4)})$. As an exception from the general agreement, we set $\tau^{(0)}$ to 0 for the star phylogeny. This way, there is a continuous transition between the evolutionary trees belonging to $(\rho, \hat{\mathcal{T}})$ and $(\rho, \check{\mathcal{T}})$, respectively, via the star tree.

**Proposition 2** (Pair-equivalence of quartets.) *Given the rooted quartets $(\rho,\mathcal{T})$, $(\rho,\hat{\mathcal{T}})$ and $(\rho,\check{\mathcal{T}})$ of Fig. 4, and the substitution process $\mathcal{P} = (u, \varepsilon, \boldsymbol{\tau}; \varphi)$ on $(\rho,\mathcal{T})$, with $0 < u < \frac{1}{2}$, $0 < \varphi < 2u$, and $0 = \tau^{(0)} < \tau^{(1)} < \tau^{(4)} < \tau^{(2)} < \tau^{(3)}$. Then there is a neighbourhood $U$ of $0$ such that for every positive $\varepsilon \in U$, there exist $\hat{\mathcal{P}} = (\hat{u}, \hat{\varepsilon}, \hat{\boldsymbol{\tau}}; \hat{\varphi})$ and $\check{\mathcal{P}} = (\check{u}, \check{\varepsilon}, \check{\boldsymbol{\tau}}; \check{\varphi})$, with $\hat{\varepsilon}$ and $\check{\varepsilon}$ also positive, such that $\mathcal{E} = (\rho,\mathcal{T},\mathcal{P})$, $\hat{\mathcal{E}} = (\rho,\hat{\mathcal{T}},\hat{\mathcal{P}})$, and $\check{\mathcal{E}} = (\rho,\check{\mathcal{T}},\check{\mathcal{P}})$ are pair-equivalent.*

Proof. Let us first observe that, for the model class considered, the stationarity of the process on all branches except (4) implies that $\zeta^{(i)} = 0$ for $i = 0, \dots, 3$. For the same reason, $\zeta^{(4,i)} \equiv \zeta^{(4)}$ (recall from Eq. (31) that $\zeta$ is additive over branches). Following Lemma 2, it is therefore sufficient to establish the identity of $\mu^{(1)}$, $\zeta^{(4)}$, and $\eta^{(j,i)}$, $j > i$, $i,j = 1, \dots, 4$, for $\mathcal{E}, \hat{\mathcal{E}}$, and $\check{\mathcal{E}}$. Explicit evaluation according to Eqns. (27)and (36) gives

$$\mu^{(1)} = \frac{1}{2}(v+\varphi) \tag{43}$$

$$\zeta^{(4)} = \frac{1}{2}\varepsilon(1-e^{-\tau^{(4)}}) \tag{44}$$

$$\eta^{(i,j)} = \omega(1-e^{-\tau^{(i,j)}}) + (v-u)\zeta^{(4)}(\delta_{i,4}+\delta_{j,4}). \tag{45}$$

We will use $\mathcal{E}$ as the reference situation and compare $\hat{\mathcal{E}}$ and $\check{\mathcal{E}}$ with it in turn. Let us start with $\hat{\mathcal{E}}$.

We replace $u$ by $\hat{u}$ and $\omega$ by $\hat{\omega} = \hat{u}(1-\hat{u})$ in Eq. (45) and introduce the abbreviations $\alpha := \hat{\omega} - \omega = -(\hat{u}-u)(\hat{u}-v)$ and $\beta := -2\zeta^{(4)}(v-u-\hat{v}+\hat{u}) = -2\zeta^{(4)}(\hat{u}-u)$. In order to determine $\hat{u}$ and $\hat{\boldsymbol{\tau}}$ from the $\eta^{(j,i)}$, one has to solve the following system of nonlinear equations:

$$
\begin{array}{rrcl}
(Q12): & \hat{\omega}e^{-\hat{\tau}^{(1)}-\hat{\tau}^{(2)}-\hat{\tau}^{(0)}} & = & \alpha + e^{-\tau^{(1)}-\tau^{(2)}} \\
(Q13): & \hat{\omega}e^{-\hat{\tau}^{(1)}-\hat{\tau}^{(3)}} & = & \alpha + \omega e^{-\tau^{(1)}-\tau^{(3)}} \\
(Q14): & \hat{\omega}e^{-\hat{\tau}^{(1)}-\hat{\tau}^{(4)}-\hat{\tau}^{(0)}} & = & \alpha + \beta + \omega e^{-\tau^{(1)}-\tau^{(4)}} \\
(Q23): & \hat{\omega}e^{-\hat{\tau}^{(2)}-\hat{\tau}^{(3)}-\hat{\tau}^{(0)}} & = & \alpha + \omega e^{-\tau^{(2)}-\tau^{(3)}} \\
(Q24): & \hat{\omega}e^{-\hat{\tau}^{(2)}-\hat{\tau}^{(4)}} & = & \alpha + \beta + \omega e^{-\tau^{(2)}-\tau^{(4)}} \\
(Q34): & \hat{\omega}e^{-\hat{\tau}^{(3)}-\hat{\tau}^{(4)}-\hat{\tau}^{(0)}} & = & \alpha + \beta + \omega e^{-\tau^{(3)}-\tau^{(4)}} .
\end{array}
$$

Multiplying (Q14) with (Q23) and (Q12) with (Q43), one observes that this gives equal left hand sides, so from the right hand sides one obtains

$$
\hat{u} = v + \frac{\varepsilon\, e^{-\tau^{(2)}}(1-e^{-\tau^{(4)}})}{e^{-\tau^{(4)}} - e^{-\tau^{(2)}}} \quad \text{and} \quad \hat{v} = u - \frac{\varepsilon\, e^{-\tau^{(2)}}(1-e^{-\tau^{(4)}})}{e^{-\tau^{(4)}} - e^{-\tau^{(2)}}} . \tag{46}
$$

Inserting this back into $\alpha$ and $\beta$, one finds

$$
\frac{\alpha+\beta}{\alpha} = e^{\tau^{(2)}-\tau^{(4)}} . \tag{47}
$$

Next, $e^{-2\hat{\tau}^{(0)}}$ is obtained as the quotient of the products of the right-hand sides of (Q12) and (Q34), and those of (Q13) and (Q24), i.e.

$$
\hat{\tau}^{(0)} = \frac{1}{2}\log\frac{(\omega e^{-\tau^{(1)}-\tau^{(3)}}+\alpha)(\omega e^{-\tau^{(2)}-\tau^{(4)}}+\alpha+\beta)}{(\omega e^{-\tau^{(1)}-\tau^{(2)}}+\alpha)(\omega e^{-\tau^{(3)}-\tau^{(4)}}+\alpha+\beta)} \tag{48}
$$

(note that this is still a formal solution at this stage).

We now exploit the fact that (Q12), (Q13), (Q23), and Eq. (43) jointly describe the pair-equivalence of reversible triples $(\mathcal{T}_3, u, \tau^{(1)}, \tau^{(2)}, \tau^{(3)}; \varphi)$ and $(\mathcal{T}_3, \hat{u}, \hat{\tau}^{(1)}, \hat{\tau}^{(2)} + \hat{\tau}^{(0)}, \hat{\tau}^{(3)}; \hat{\varphi})$, as indicated by the solid lines in Fig. 4, and established in Lemma 2. Consequently,

$$
\hat{\varphi} = \varphi + \hat{u} - u , \tag{49}
$$

and

$$
\hat{\tau}^{(1)} = \bar{\tau}^{(1)}, \quad \hat{\tau}^{(2)} = \bar{\tau}^{(2)} - \hat{\tau}^{(0)}, \quad \hat{\tau}^{(3)} = \bar{\tau}^{(3)} \tag{50}
$$

where the $\bar{\tau}^{(i)}$ are from Eqns. (41) and (42) with $\bar{\omega}$ set to $\hat{\omega} = \hat{u}\hat{v}$. $\tau^{(4)}$ is then obtained by dividing the difference between (Q41) and (Q43) by that between (Q21) and (Q23) to yield

$$\hat{\tau}^{(4)} = \tau^{(4)} + \hat{\tau}^{(2)} - \tau^{(2)} . \tag{51}$$

It is easily checked that Eqs. (46)–(51) constitute a solution of (Q12)–(Q34). We must finally consider $\zeta^{(4)}$ from Eq. (44), which, when solved for $\hat{\varepsilon}$ and $\hat{\varphi}$, results in

$$\hat{\varepsilon} = \varepsilon \, \frac{1 - e^{-\tau^{(4)}}}{1 - e^{-\hat{\tau}^{(4)}}} \tag{52}$$

to complete the (formal) solution.

Last but not least, it must be checked that the quantities in $\hat{\mathcal{P}}$ are all well-defined, i.e. in their proper ranges. To this end, let us first consider the case $\varepsilon = 0$, which yields $\hat{\varepsilon} = 0$, $\hat{\tau}^{(0)} = 0$, $\hat{u} = v$, $\hat{v} = u$, and $\hat{\varphi} = \varphi + v - u$. These constitute a decent substitution process on the *star tree* $(\rho, \mathcal{T})$, since $\hat{u}\hat{v}, \hat{\varphi} \in (0, 1)$ due to the assumed constraints on the parameters of $\mathcal{P}$. Now let $\varepsilon > 0$. The parameters of $\hat{\mathcal{P}}$ being continuous in $\varepsilon$, we are done except for $\hat{\tau}^{(0)}$ which must be checked for positivity. Some algebraic manipulation with Eqns. (47) and (48) shows that $\hat{\tau}^{(0)} > 0$ if and only if

$$\alpha(e^{-\tau^{(2)}} + e^{-\tau^{(1)}+\tau^{(2)}-\tau^{(3)}} - e^{-\tau^{(1)}} - e^{-\tau^{(3)}}) > 0 . \tag{53}$$

$\alpha$ is negative for small $\varepsilon$, since $u - v < 0$ and $e^{-\tau^{(4)}} - e^{-\tau^{(2)}} > 0$ (cf. Eq. (46)). To see that the term in brackets is negative as well, one applies the convexity of the exponential function in the form

$$e^{-z_1} + e^{-z_2} < e^{-x_1} + e^{-x_2} \quad \text{if} \;\; z_1 + z_2 = x_1 + x_2, \; x_1 < z_1, z_2 < x_2 \tag{54}$$

to the arguments $x_1 = \tau^{(1)}, x_2 = \tau^{(3)}, z_1 = \tau^{(1)} + \tau^{(3)} - \tau^{(2)}, z_2 = \tau^{(2)}$. So $\hat{\tau}^{(0)}$ is indeed positive for small $\varepsilon$.

We are finally ready to consider $\check{\mathcal{T}}$. The same procedure is applied but with indices 1 and 2 permuted. In particular, we now have $\frac{\alpha+\beta}{\alpha} = e^{\tau^{(1)}-\tau^{(4)}}$ instead of Eq. (47), and $\check{\tau}^{(0)}$ turns out to be positive if and only if

$$\alpha(e^{-\tau^{(1)}} + e^{\tau^{(1)}-\tau^{(2)}-\tau^{(3)}} - e^{-\tau^{(2)}} - e^{-\tau^{(3)}}) > 0 . \tag{55}$$

This time, $\alpha > 0$ (since $e^{-\tau^{(4)}} > e^{-\tau^{(2)}}$), and convexity (Eq. (54)) as applied to $x_1 = \tau^{(1)}, x_2 = \tau^{(2)} + \tau^{(3)} - \tau^{(1)}, z_1 = \tau^{(2)}, z_2 = \tau^{(3)}$) renders the term in brackets positive, too, which completes the proof. $\square$

We have thus constructed evolutionary trees with different topologies which are indistinguishable in terms of their pairwise joint distributions of letters at the leaves. Table 1 shows a numerical example of such a situation. Let us remark that, for the model class considered, the third quartet topology (with leaves 3 and 4 grouping together) is ruled out in the sense that it leads to a *negative* duration of the interior branch.

Altogether, the construction procedure may be interpreted as follows. The situation with $\varepsilon = 0$ describes a substitution process which is HSR on *all* branches of the star tree. This is invariant under interchange of $u$ and $v$, which leads to a pair-equivalent process on the same star tree. Choosing $\varepsilon > 0$ breaks this symmetry and makes the star tree bifurcate into two.

# 6 Discussion

Pairwise sequence comparisons, and the resulting (estimates of the) divergence matrices, have been mainly seen from the point of view of tree-additive distance measures, i.e. functions $d^{(i,j)} = h(H^{(i,j)})$ which may be decomposed into additive branch lengths according to Eq. (13). In a previous paper [13], the so-called LogDet family of additive distance measures was considered in terms of the substitution process. It was shown that the branch lengths belonging to the paralinear distance (Eq. (12)), namely $\varepsilon^{\{r,s\}} = -\frac{1}{2}\log(\det(P^{(r,s)}P^{(s,r)}))$, possess a mechanistic interpretation: They equal the arithmetic mean of the *relative* substitution count for the forward and the reverse process across edge $\{r,s\}$, where the averaging reflects our ignorance of the position of the root, i.e. the direction of the edges. The present identifiability result for the indistinguishable sites case says much more: Not only the branch lengths may be reconstructed, but also details about the substitution process itself. The return-trip transition matrices $P^{(r,s)}P^{(s,r)}$ are identifiable from the pairwise joint distribution of letters at the leaves if $r$ is a terminal node, and identifiable up to conjugacy otherwise. As a consequence, all information contained in the characteristic polynomial of the return-trip matrices is available, in particular the eigenvalues, $\lambda_i^{\{r,s\}}$, $i = 1, \ldots, m$. Paralinear branch lengths use exactly this information: $\varepsilon^{\{r,s\}} = -\frac{1}{2}\sum_{i=1}^{m}\log\lambda_i^{\{r,s\}}$. The failure to reconstruct the *one-way* transition probabilities is due to ignorance of the location and base composition of the root. Put differently, identification of the return-trip transition matrices may be viewed as a matrix version of branch length reconstruction from additive distances, without the prior collapsing of the divergence matrices to scalar quantities.

With the additional complication of rate heterogeneity, it is now well-established that, in the HSR case, the number of substitutions may be retrieved from the divergence matrices *provided* the distribution of rate factors is known [8, 9]; an additive distance measure, and with it the identifiability of the topology, is thus recovered. Our example of a non-HSR process, however, demonstrates that different topologies may be compatible with identical pairwise joint distributions of letters at the leaves. As a consequence, identifiability of the topology is lost; in particular, *no additive distance measure can exist*, even if the distribution of rate factors is known. The example exploits the simplest situation, namely one with invariable sites whose base composition differs from that of the variable sites. Such a situation is by no means academic, for as the base composition becomes more heterogeneous over taxa, sites with different substitution rates also change base composition with respect to each other (see the review in [1]). This problem has, for example, emerged as relevant in hominoid mitochondrial sequence data [9].

Pairwise sequence comparisons are at the heart of distance-based tree reconstruction methods, like neighbour joining [26]. In contrast, maximum-likelihood procedures make use of the full joint distributions of letters at the leaves, see e.g. [11, 27]. The methods have complementary ranges of application. With maximum likelihood, the Markov process is parametrized explicitly, and these parameters are determined, often with high accuracy. The number of parameters must be kept reasonable, however, i.e. the model class must be restricted severely. For instance, all current implementations are restricted to HSR models. Pairwise distance methods, however, do not employ explicit parametrizations, nor do they aim at determining individual parameters; only certain invariants (like the branch lengths) are recovered. As a consequence, they can cope with non-HSR processes, as long as adequate distance measures, like the members of the LogDet family, are available. They do, however, ignore some information, which increases the variance of branch length estimates and may make tree reconstruction less reliable in practice. Neither method seems to be suited to non-HSR processes with rate heterogeneity: There is no additive distance measure, and the number of parameters is a nightmare for maximum-likelihood procedures (and present-day computers).

It has recently been shown that joint distributions of letters over *triples* of leaves allow full reconstruction of the *one-way* edge transition matrices, provided these belong to the class of matrices which are reconstructible from their columns (in the notation of the present paper); see [5] and the application in [28]. Put differently, the one-way transition matrices are identifiable up to permutations of their columns. This ambiguity is also inherent in Markov processes which are continuously generated, since the corresponding transition matrices are *not* reconstructible from columns. In fact, *different* sets of one-way edge transition and frequency matrices (i.e. different $\mathcal{P}$'s) may give rise to *the same* distributions of letters over *triples* of leaves; the example rate matrix on p. 66 of [5] may be used to demonstrate this (Joe Chang, personal communication). Nevertheless, it should be remarked that, in our rate heterogeneity example, the joint distributions of letters at triples of leaves do *not* coincide for the different topologies, although the pairwise joint distributions do. It thus remains an open question whether, for a given distribution of rates, *topologies* may be distinguished on the basis of joint distributions of letters over triples of leaves.


## Acknowledgements

It is my pleasure to thank Joe Chang, Arndt von Haeseler, and Mike Steel for fruitful discussions, and for critically reading the manuscript. I am grateful to Michael Baake for his persistent but constructive criticism. This work was supported by the German Science Foundation (DFG).


# References


[1] D. L. Swofford, G. J. Olsen, P. J. Waddell, D. M. Hillis, Phylogenetic inference, In *Molecular Systematics*, D. M. Hillis, C. Moritz, and B. K. Mable, (Eds.), Sinauer, Sunderland, 1996, pp. 407–514.

[2] J.T. Chang, J.A. Hartigan, Reconstruction of evolutionary trees from pairwise distributions on current species, In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, E.M. Keramids, (Ed.), Interface Foundation, Fairfax Station, VA, 1991, pp. 254–257.

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

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

[5] J. T. Chang, Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency, *Math. Biosci.* 137 (1996) 51–73.

[6] J. T. Chang, Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters, *Math. Biosci.* 134 (1996) 189–215.

[7] M.A. Steel, L.A. Székely, M.D. Hendy, Reconstructing trees when sequence sites evolve at variable rates, *J. Comp. Biol.* 1 (1994) 153–163.

[8] X. Gu, W.-H. Li, A general additive distance with time-reversibility and rate variation among nucleotide sites, *Proc. Natl. Acad. Sci. USA* 93 (1996) 4671–4676.

[9] P. J. Waddell, M. A. Steel, 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.

[10] D. Barry and J. A. Hartigan, Asynchronous distance between homologous DNA sequences, *Biometrics* 43 (1987) 261–276.

[11] J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach, *J. Mol. Evol.* 17 (1981) 368–76.

[12] J. G. Kemeny, J. L. Snell, *Finite Markov chains*, Springer, New York, 1981, p. 105.

[13] E. Baake, A. von Haeseler, Distance measures in terms of substitution processes, *Theor. Pop. Biol.*, in press.

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

[15] H.-J. Bandelt, A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. Appl. Math.* 7 (1986) 309–343.

[16] S. Benzer, On the topography of the genetic fine structure, *Proc. Natl. Acad. Sci. USA* 47 (1961) 403–415.

[17] J. Wakeley, Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA, *J. Mol. Evol.* 37 (1993) 613–623.

[18] R. Nielsen, Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA, *Syst. Biol.* 46 (1997), 346–353.

[19] Z. Yang, Among-site rate variation and its impact on phylogenetic analyses, *TREE* 11 (1996) 367–372.

[20] F. Rodriguez, J.L. Oliver, A. Marin, J.R. Medina, The general stochastic model of nucleotide substitution, *J. Theor. Biol.* 142 (1990) 485–501.

[21] Z. Yang, S. Kumar, Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites, *Mol. Biol. Evol.* 13 (1996) 650–659.

[22] Z. Yang, Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Mol. Biol. Evol.* 10 (1993) 1396–1401.

[23] X. Gu, Y.-X. Fu, W.-H. Li, Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites, *Mol. Biol. Evol.* 12 (1995) 546–557.

[24] J.S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.* 22 (1973) 250–256.

[25] J.A. Cavender, Taxonomy with confidence, *Math. Biosci.* 40 (1978) 271–280.

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

[27] K. Strimmer, A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, *Mol. Biol. Evol.* 13 (1996) 964–969.

[28] J.A. Lake, Phylogenetic inference: How much evolutionary history is knowable ? *Mol. Biol. Evol.* 14 (1997) 213–219.

## Figure captions

Fig.1: Left: Partitioning of leaves, $\{A_1, \ldots, A_4\}$, as induced by an internal node $m$ of degree 4. Right: Partitioning of leaves, $\{\{A_1, A2, A_3\}, \{B_1, B_2\}\}$, as induced by an internal edge $\{m, n\}$.

Fig.2: Reconstruction of return-trip transition matrices from leaf-to-leaf transition matrices according to Propositon 1. The bullets denote the starting points. Solid lines symbolize leaf-to-leaf transition matrices, dashed lines their inverses. Paths traversed through a solid line and a dashed line in immediate succession but in opposite direction cancel out. Left: Reconstruction of $P^{(i,m)}P^{(m,i)}$ for leaf $i$ and internal node $m$. Right: Reconstruction of $P^{(i,m)}P^{(m,n)}P^{(n,m)}(P^{(i,m)})^{-1}$ for internal nodes $m, n$ and leaf $i$.

Fig.3: Pair-equivalence of reversible triples. Evolutionary trees $\mathcal{E} = (\mathcal{T}_3, u, \boldsymbol{\tau}; \varphi)$ and $\bar{\mathcal{E}} = (\mathcal{T}_3, \bar{u}, \bar{\boldsymbol{\tau}}; \bar{\varphi})$ produce the same pairwise joint distributions of letters at the leaves (see Lemma 2).

Fig.4: Pair-equivalence of quartets. Evolutionary trees $\mathcal{E} = (\rho, \mathcal{T}, \mathcal{P})$, $\hat{\mathcal{E}} = (\rho, \hat{\mathcal{T}}, \hat{\mathcal{P}})$ and $\check{\mathcal{E}} = (\rho, \check{\mathcal{T}}, \check{\mathcal{P}})$ produce the same pairwise joint distributions of letters at the leaves (see Proposition 2). Solid lines indicate the subtree $\mathcal{T}_3$ with an HSR process for the variable sites; dashed lines stand for the non-stationary process on the branch which leads to leaf 4.

Table 1: Example of pair-equivalent evolutionary trees $\mathcal{E}, \hat{\mathcal{E}}$, and $\check{\mathcal{E}}$ as in Fig. 4. Parameters: $u = 0.40, \varepsilon = 0.05, \boldsymbol{\tau} = (0.00, 0.30, 0.90, 1.20, 0.60), \varphi = 0.50$; $\hat{u} = 0.66, \hat{\varepsilon} = 0.04, \hat{\boldsymbol{\tau}} = (0.10, 0.16, 1.14, 1.65, 0.84), \hat{\varphi} = 0.76$; $\check{u} = 0.51, \check{\varepsilon} = 0.05, \check{\boldsymbol{\tau}} = (0.03, 0.28, 0.80, 1.05, 0.58), \check{\varphi} = 0.61$. The table lists letter frequencies, transition matrices and divergence matrices. Quantities with overstrike characters belong to $\hat{\mathcal{E}}$ and $\check{\mathcal{E}}$, respectively. Note that all figures are rounded to two decimal places. For details, see Proposition 2.

| $\boldsymbol{\varphi}$ | $\boldsymbol{\xi}^{(\rho)}$ | $P^{(0)}$ | $P^{(1)}$ | $P^{(2)}$ | $P^{(3)}$ | $P^{(4)}$ |
|---|---|---|---|---|---|---|
| $\begin{pmatrix}.50\\.50\end{pmatrix}$ | $\begin{pmatrix}.60\\.40\end{pmatrix}$ | $\begin{pmatrix}1.0 & 0.0\\0.0 & 1.0\end{pmatrix}$ | $\begin{pmatrix}90 & .16\\.10 & .84\end{pmatrix}$ | $\begin{pmatrix}.76 & .35\\.23 & .64\end{pmatrix}$ | $\begin{pmatrix}.72 & .42\\.28 & .58\end{pmatrix}$ | $\begin{pmatrix}.80 & .25\\.20 & .75\end{pmatrix}$ |
| $\hat{\boldsymbol{\varphi}}$ | $\hat{\boldsymbol{\xi}}^{(\rho)}$ | $\hat{P}^{(0)}$ | $\hat{P}^{(1)}$ | $\hat{P}^{(2)}$ | $\hat{P}^{(3)}$ | $\hat{P}^{(4)}$ |
| $\begin{pmatrix}.76\\.24\end{pmatrix}$ | $\begin{pmatrix}.34\\.66\end{pmatrix}$ | $\begin{pmatrix}.94 & .03\\.06 & .97\end{pmatrix}$ | $\begin{pmatrix}.90 & .05\\.10 & .95\end{pmatrix}$ | $\begin{pmatrix}.55 & .23\\.45 & .77\end{pmatrix}$ | $\begin{pmatrix}.46 & .27\\.54 & .73\end{pmatrix}$ | $\begin{pmatrix}.60 & .17\\.40 & .83\end{pmatrix}$ |
| $\check{\boldsymbol{\varphi}}$ | $\check{\boldsymbol{\xi}}^{(\rho)}$ | $\check{P}^{(0)}$ | $\check{P}^{(1)}$ | $\check{P}^{(2)}$ | $\check{P}^{(3)}$ | $\check{P}^{(4)}$ |
| $\begin{pmatrix}.61\\.39\end{pmatrix}$ | $\begin{pmatrix}.49\\.51\end{pmatrix}$ | $\begin{pmatrix}.98 & .02\\.02 & .98\end{pmatrix}$ | $\begin{pmatrix}.87 & .12\\.13 & .88\end{pmatrix}$ | $\begin{pmatrix}.72 & .27\\.28 & .73\end{pmatrix}$ | $\begin{pmatrix}.67 & .32\\.33 & .68\end{pmatrix}$ | $\begin{pmatrix}.75 & .19\\.25 & .81\end{pmatrix}$ |

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

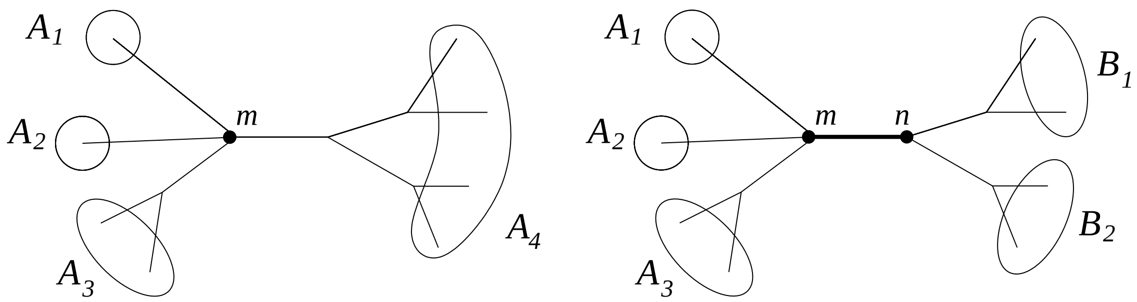
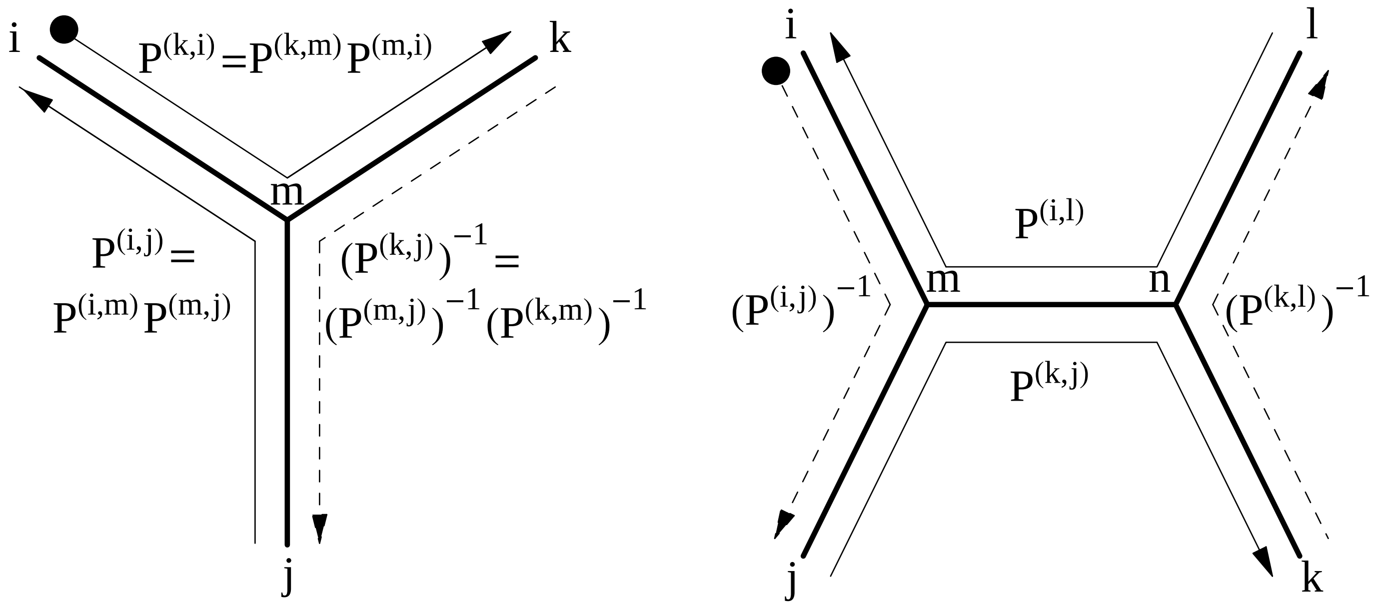



Figure 1:



i
$P^{(k,i)} = P^{(k,m)} P^{(m,i)}$
k
m
$P^{(i,j)} =$
$P^{(i,m)} P^{(m,j)}$
$(P^{(k,j)})^{-1} =$
$(P^{(m,j)})^{-1} (P^{(k,m)})^{-1}$
j
i
l
$P^{(i,l)}$
m
n
$(P^{(i,j)})^{-1}$
$(P^{(k,l)})^{-1}$
$P^{(k,j)}$
j
k


Figure 2:

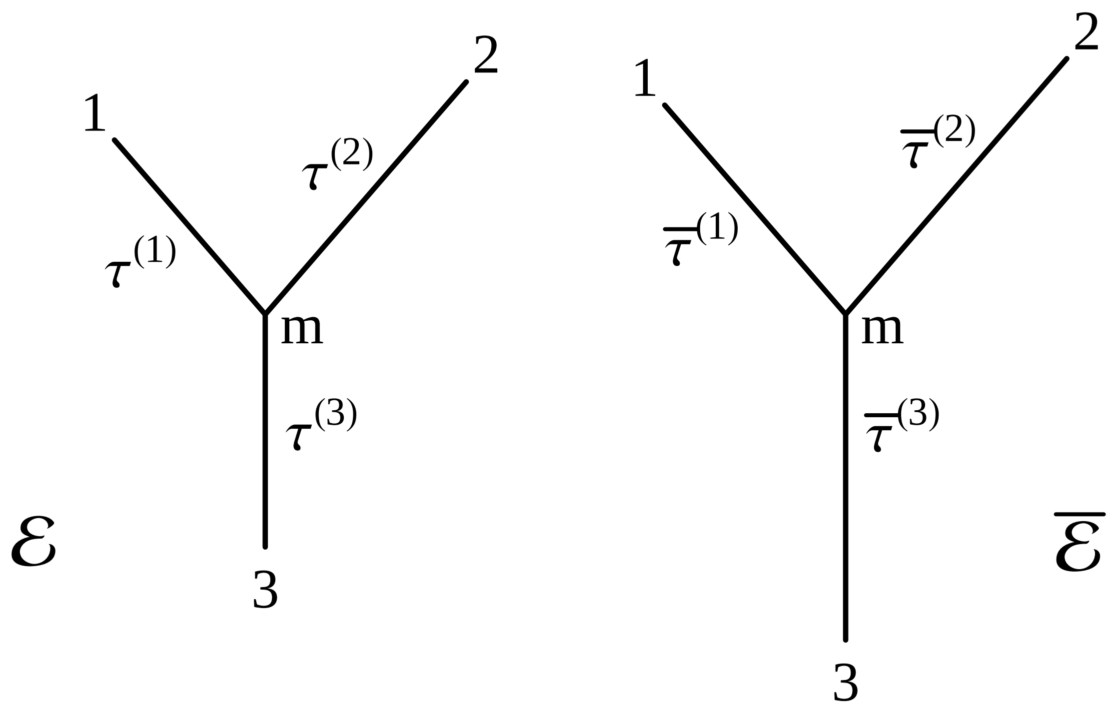
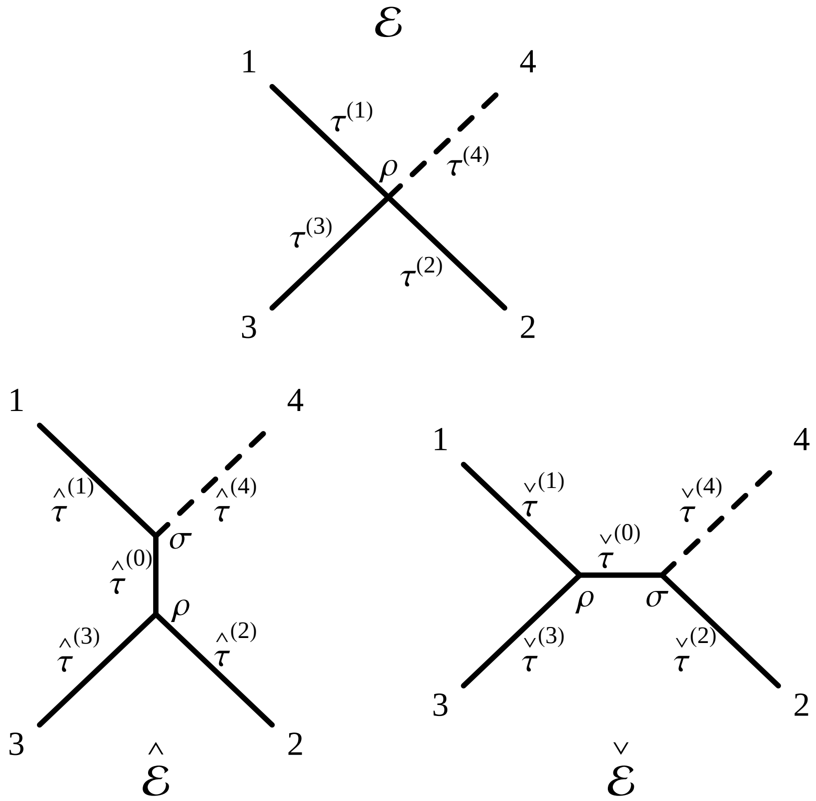

1
2
$\tau^{(1)}$
$\tau^{(2)}$
m
$\tau^{(3)}$
$\mathcal{E}$
3
1
2
$\overline{\tau}^{(1)}$
$\overline{\tau}^{(2)}$
m
$\overline{\tau}^{(3)}$
$\overline{\mathcal{E}}$
3

Figure 3:



$\mathcal{E}$
1
4
$\tau^{(1)}$
$\rho$
$\tau^{(4)}$
$\tau^{(3)}$
$\tau^{(2)}$
3
2
1
4
$\hat{\tau}^{(1)}$
$\hat{\tau}^{(4)}$
$\sigma$
$\hat{\tau}^{(0)}$
$\rho$
$\hat{\tau}^{(3)}$
$\hat{\tau}^{(2)}$
3
2
$\hat{\mathcal{E}}$
1
4
$\check{\tau}^{(1)}$
$\check{\tau}^{(4)}$
$\check{\tau}^{(0)}$
$\rho$
$\sigma$
$\check{\tau}^{(3)}$
$\check{\tau}^{(2)}$
3
2
$\check{\mathcal{E}}$


Figure 4: