TREES AND Ps AND THINGS THAT SNEEZE

MARKOV PROCESS MODELS OF SITE SUBSTITUTION

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science in Mathematics
in the
University of Canterbury
by
Christopher Tuffley

University of Canterbury
1997

Contents

List of Figures

Abstract

An increasingly important tool in phylogenetics, a field which lies somewhere between mathematics and biology and seeks to deduce the evolutionary relationships between present day species, is the comparison of molecular sequences such as DNA and protein sequences. In making meaningful comparisons it is helpful to model the process by which the sequences came to differ. Many such models have at their heart certain Markov-style assumptions, since the "memoryless" feature of the Markov property seems appropriate to the site substitution process. This thesis looks at two problems related to the most basic Markov process model of site substitution, on which most more complicated (and hopefully more realistic) models are based, and takes a first look at a recently suggested model of Fitch and Markowitz's 1970 "covarion" hypothesis, comparing the covarion model with models of the better known rates-across-sites hypothesis.

We show that the LogDet transformation, which under mild conditions allows tree reconstruction under the basic model, is in a sense unique, in that $\phi = \log\det$ is the only continuous homomorphism from $n \times n$ stochastic matrices with positive determinant into the real numbers under addition, up to scalar multiples. This result limits the form of possible alternatives to the LogDet transformation that might weaken the conditions under which it is valid.

We introduce the *reconstruction quotient* and prove two structure theorems for it in the case of the very simple two state fully symmetric model. The reconstruction quotient is obtained from a space of weighted trees by identifying trees that no reconstruction technique will be able to distinguish between. We show that under the two state fully symmetric model, the reconstruction quotient corresponding to a fixed tree is always contractible, and that the quotient obtained from the set of all four leaf binary trees is also contractible.

Finally, we take the first analytic look at a model of the covarion hypothesis, an alternative approach to accounting for differing selective constraints to the competing idea of rates-across-sites. We calculate some of the basic quantities required for tree reconstruction under this model, and compare it with rates-across-sites, seeking to find conditions under which they can and cannot be distinguished.

Chapter 1

Introduction

Phylogenetics is the study of evolutionary relationships and classification. Much of the current thrust of research, including this thesis, is biologically motivated, but many of the techniques are applicable to other classification problems, such as the classification of languages. Although many of the questions to be answered, such as "Are humans more closely related to chimpanzees or gorillas?", are not inherently mathematical, mathematics and statistics have come to play important roles. Many of the structures used to represent phylogenetic information, such as trees, quartets, splits and metrics, are mathematical in nature, and statistical techniques are required to deduce relationships and give us an indication of how confident we can be in them.

One of the central problems of phylogenetics is to reconstruct the tree describing the evolutionary history of some set of objects of interest, from observations of their present state. In the biological case this tree will be some small part of the "Tree of Life": the tree on which the species on earth today, from the ones that sneeze (humans, cats,...) through the ones that make others sneeze (pollen bearing plants, various bacteria,...) and beyond (yeast, sharks, snow algae,...), are each represented by a leaf, with the edges or branches of the tree representing ancestral relationships much as in a family tree. Such trees, together with estimates of associated quantities such as the temporal length of each edge, can help to answer such questions as whereabouts on earth the human race originated and whether the mammals radiated before or after dinosaurs had disappeared. Nor are all such questions purely curiosity driven: the tree of strains of a given virus and knowledge of the populations in which they are found can help to track the virus's spread.

Since the advent of molecular biology the comparison of molecular sequences has become an important tool in tree reconstruction. In order for such comparisons to be meaningful it is necessary to model fairly accurately the substitution process by which ancestral sequences gave rise to the sequences we observe today. Most substitution models that have been proposed are probabilistic in nature, and frequently incorporate Markov-style assumptions. This thesis looks at three problems related to some Markov process based models of site substitution.

After reviewing some basic concepts regarding trees and Markov processes in Chapter 2, we describe the "basic" substitution model, on which the other models we will study are based, in Chapter 3. The next three chapters each look at one of the three problems we examine.

An important issue in reconstructing trees from sequence data is whether the sequences actually contain enough information to do so. The LogDet transformation, due to Chang and Hartigan [8] and independently to Steel [36], shows that they do under the basic model if some mild restrictions are placed on it. This transformation is based mainly on the fact that the map $\phi = \log\det$ is a homomorphism from certain semigroups of matrices into the additive real numbers. The map ϕ also has the property of being continuous, which is desirable since we would like our output to depend continuously on our input. In Chapter 4, **The LogDet transformation: there can be only one**, we show that ϕ is the only map with these two properties, up to scalar multiples. This result limits the form of possible alternatives to the LogDet transformation that might weaken the

conditions under which it is valid.

In Chapter 5 we introduce the **reconstruction quotient**, which is also related to the question of how much information is contained in sequence data. Given a space of edge-weighted trees and a substitution model, the reconstruction quotient is the space obtained by identifying weighted trees that generate the same data under the model. The structure of this space is relevant to phylogenetics, since sampling errors may prevent "real data" from corresponding to "ideal data" from any weighted tree, making it necessary to approximate the observed data by a point in the space of ideal data. We study the reconstruction quotient under the very simple two state fully symmetric model, showing firstly that the LogDet transformation captures all the information present in the data under this model and then proving two structure theorems. Both of these theorems show that certain reconstruction quotients are contractible, and so in a sense "simple".

The final chapter, **More realistic models: the covarion hypothesis and rates-across-sites**, looks in detail at two more realistic substitution models that have been proposed. The basic model studied in previous chapters treats every site as evolving at the same rate, and in practice this does not seem to be the case: some sites seem to be evolving rapidly while some appear to change not at all. This is attributed to differing selective constraints at different sites, and methods of taking this into account have been suggested. One of these, rates-across-sites, assumes that the constraints do not change with time, while a second, the covarion hypothesis, suggests that they change as changes occur elsewhere in the sequence. Although the covarion hypothesis was proposed in 1970 by Fitch and Markowitz [14], it is less well studied than the competing rates-across-sites hypothesis. We take the first analytic look at a recently proposed model of the covarion hypothesis, calculating some of the quantities required for tree reconstruction under this model, and comparing it with rates-across-sites models with the aim of finding conditions under which the two models can be distinguished.

Chapter 2

Basic concepts and definitions

2.1 Trees

2.1.1 Vertex-labelled trees

Phylogenetics uses various types of vertex-labelled trees to express the evolutionary relationships between the species in some set of interest. Most of our trees will be labelled and we will frequently drop explicit reference to the labelling.

The most important vertex-labelled trees are the *leaf-labelled trees.* An **unrooted leaf-labelled tree** is a connected acyclic graph with no vertices of degree two and such that each **leaf** (vertex of degree less than or equal to one) is given a unique label from some label set S. Typically $S = [n] := \{1, \ldots, n\}$, where n is the number of leaves of the tree; we will usually write n for $|S|$. A **rooted leaf-labelled tree** is defined similarly to an unrooted one, but has a distinguished vertex called the **root**, which is allowed to have degree two. Usually the root will not be a leaf, and we will denote it by ρ.

The non-leaf vertices of a tree are called **internal vertices**; similarly an edge that is not adjacent to a leaf is an **internal edge**. Edges adjacent to a leaf will sometimes be called **external** or **pendant edges**. It is sometimes useful to consider trees in which some of the internal vertices are labelled, or have vertices with multiple labels, subject to the restriction that all vertices of degree less than or equal to two are labelled and each label labels only one vertex; we will call such trees **S-labelled trees**, where S is the labelling set. We will later need to consider **S-labelled forests:** this will be a partition α of S, together with an α^i-labelled tree for each part α^i of α. Note that two vertex-labelled trees (or forests) with the same label set S are only considered to be the same if they are isomorphic as graphs and the isomorphism preserves the labelling.

The trees that carry the most phylogenetic information are the **binary** trees. In the unrooted case, this is a leaf-labelled tree in which each internal vertex has degree three; in the rooted case, this is a leaf-labelled tree in which the root has degree two and all other internal vertices have degree three. The leaf-labelled trees with the least phylogenetic information are the **star trees**, which are the trees with vertices $\{0, 1, \ldots, n\}$ and edges $\{\{0, 1\}, \ldots, \{0, n\}\}$ for each $n \geq 3$.

Given a graph G, we will denote its vertex set by $V(G)$, and its edge set by $E(G)$. The edge between vertices u and v will usually be denoted by $\{u, v\}$ unless it is to be considered as directed from u to v, in which case we write (u, v). In rooted trees we will usually assume that all edges are directed away from the root. Some examples of trees are shown in figure 2.1.

2.1.2 Splits

A very useful concept in studying vertex-labelled trees is that of the **split**. Deleting an edge e from a tree T divides it into two connected components and thereby gives rise to a bi-partition $\sigma = \{A, B\}$

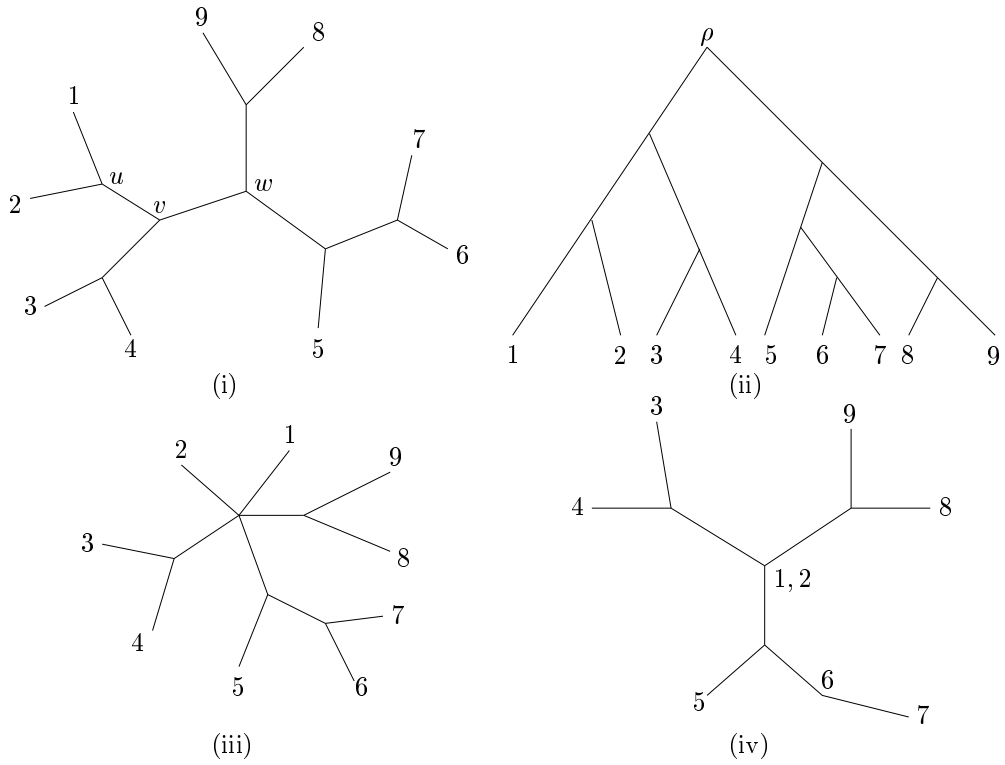


Figure 2.1: Some examples of trees. (i) is an unrooted binary tree. The edges $\{1, u\}$ and $\{2, u\}$ are external edges while $\{u, v\}$ and $\{v, w\}$ are internal edges. (ii) is a rooted binary tree, obtained from (i) by subdividing the edge $\{v, w\}$ and rooting the tree at the newly created vertex. (iii) is a non-binary unrooted leaf-labelled tree, while (iv) is a [9]-labelled tree.

of the labelling set, where A and B are the sets of labels in each component. The bi-partition σ is called the **split** corresponding to the edge e, and we will denote the set of splits of T by $\sigma(T)$. Referring to figure 2.1 (i), the split corresponding to the edge $\{u, v\}$ is $\{\ \{1, 2\}, \{3, 4, 5, 6, 7, 8, 9\}\ \}$ and that corresponding to $\{v, w\}$ is $\{\ \{1, 2, 3, 4\}, \{5, 6, 7, 8, 9\}\ \}$. A set of splits Σ is said to be **compatible** if there is a tree T such that $\Sigma \subseteq \sigma(T)$.

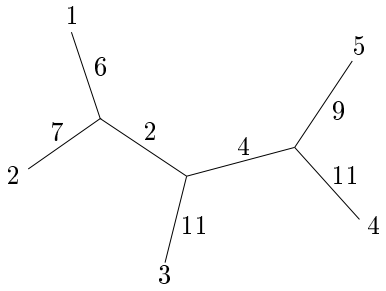
There are a number of important results regarding splits, which we state below.

Theorem 2.1 (Buneman [3])

1. *A tree T is determined by $\sigma(T)$, and can be recovered in polynomial time.*
2. *Two splits $\sigma = \{A, B\}$, $\sigma' = \{C, D\}$ are compatible if and only if at least one of $A \cap C$, $A \cap D$, $B \cap C$ and $B \cap D$ is empty.*
3. *A set of splits Σ is compatible if and only if it is pairwise compatible.*

Splits also give a convenient partial order of vertex-labelled trees: we write $T_1 \leq T_2$ if and only $\sigma(T_1) \subseteq \sigma(T_2)$. This corresponds to the idea that T_1 can be obtained from T_2 by contracting the edges corresponding to $\sigma(T_2) \setminus \sigma(T_1)$. Referring again to figure 2.1, tree (iii) has been obtained from tree (i) by contracting the edges $\{u, v\}$ and $\{v, w\}$, while (iv) has been obtained from (iii) by further contracting the edges adjacent to leaves 1, 2 and 6, so we have (iv)<(iii)<(i).

We will later find it convenient to extend this partial order to vertex-labelled forests; in this context the corresponding idea is that one forest may be obtained from the other via edge deletions and contractions.



d	1	2	3	4	5
1	–	13	19	23	21
2		–	20	24	22
3			–	26	24
4				–	20
5					–

Figure 2.2: An example of a tree metric. The distance between two labels is found by summing the edge weights along the path between the vertices they label.

2.1.3 Tree-additive distances

A very useful tool in reconstructing trees is the concept of a **tree metric**, sometimes called a **tree-like** or **tree-additive distance**, since they are not always metrics. A vertex-labelled tree T whose edges have been given non-negative weights induces a natural distance d_{ij} between elements i and j of the labelling set: simply sum the edge weights along the path from the vertex i labels to the vertex j labels (see figure 2.2). Such a distance function (it is in fact a metric if all weights are positive and there are no vertices with more than one label) is called a tree-additive distance.

Tree-additive distances are completely characterised by the **four-point condition** [3, 35, 45]: given any four points $a, b, c, d \in S$, we have

$$d_{ab} + d_{cd} \leq \max\{d_{ac} + d_{bd}, d_{ad} + d_{bc}\}. \tag{2.1}$$

Furthermore, if all edge weights are positive, T and the edge weighting λ are unique and may be recovered from d. This may be done quickly, which makes tree-additive distances a useful tool in tree reconstruction.

We will later want to allow edges to have "infinite" length. Rather than work with the interval $[0, \infty]$ we will work with $[0, 1]$ and consider our weights to be multiplicative rather than additive; the resulting "distance" ρ we will call a **multiplicative distance**. It is easy to see that taking minus the logarithm of a multiplicative distance gives an additive distance, so the four point condition becomes

$$\rho_{ab}\rho_{cd} \geq \min\{\rho_{ac}\rho_{bd}, \rho_{ad}\rho_{bc}\}, \tag{2.2}$$

and we may reconstruct the tree and edge weights provided they lie in $(0, 1)$.

2.1.4 Characters

Characters are the basic data for tree reconstruction. They are simply functions from the label set S into some set $\mathcal{A}$ of **"states"**. For example, given a set of aligned sequences, the rule "nucleotide at site k" for $k = 1, \ldots, N$ gives a set of characters with state set $\mathcal{A} = \{A, G, C, T\}$. This is the main example we will have in mind, but in the past *morphological* characters (for example, "number of legs" gives a morphological character with state set Z) have been important. We will often use χ to denote a character.

A function $\hat{\chi} : V(T) \mapsto \mathcal{A}$ is a **state function**. If $\ell : S \mapsto V(T)$ is the labelling of T, we will say that $\hat{\chi}$ **extends** χ or that $\hat{\chi}$ is an **extension** of χ on T if $\hat{\chi} \circ \ell = \chi$, and denote this by $\hat{\chi} \upharpoonright \chi$.

2.2 Markov chains and processes

Stated informally, a Markov process is a "memoryless" random process: in order to try to predict what it will do next, only the most recent piece of information known is of any use. Intuitively this

seems a reasonable assumption to make when modelling the evolution of a DNA sequence—what has happened in the past is irrelevant and all that matters is its current state—and all substitution models we consider will be based on a Markov process.

There is an extensive theory of Markov chains and processes and some references are [16, 21, 23]. All our Markov processes will have finite state space, and we will write r for the number of states.

2.2.1 Discrete time Markov chains

A family of random variables $\{X_i | i \in \mathsf{N} \cup \{0\}\}$ taking values in $\mathcal{A} = \{1, \ldots, r\}$ is a **discrete time Markov chain** if it satisfies the **Markov property:** for every n and all states $i_0, i_1, \ldots, i_n$ we have

$$\mathsf{P}[X_n = i_n | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] = \mathsf{P}[X_n = i_n | X_{n-1} = i_{n-1}]. \tag{2.3}$$

The set $\mathcal{A}$ is called the **state space** of the chain.

The behaviour of such a chain is determined by the row vector $\pi^{(0)} = (\pi_i^{(0)})$ of initial probabilities, where

$$\pi_i^{(0)} = \mathsf{P}[X_0 = i],$$

and the **transition matrices** $p^{(n,n-1)}$, which are the conditional probabilities

$$p_{ij}^{(n,n-1)} = \mathsf{P}[X_n = j | X_{n-1} = i].$$

The conditional probability $\mathsf{P}[X_{k+n} = j | X_k = i]$ is given by the ij-entry of the n-step transition matrix $p^{(k+n,k)}$, which may be written in terms of the one step transition matrices $p^{(m,m-1)}$ as

$$p^{(k+n,k)} = p^{(k+1,k)} p^{(k+2,k+1)} \cdots p^{(k+n,k+n-1)}.$$

The vector $\pi^{(n)}$ of probabilities $\mathsf{P}[X_n = i]$ may be written

$$\pi^{(n)} = \pi^{(0)} p^{(1,0)} p^{(2,1)} \cdots p^{(n,n-1)}.$$

The chain is said to be **homogeneous** (sometimes referred to as **stationary**) if $p^{(n,n-1)} = p^{(1,0)}$ for all n. In this case we may simply write $p = p^{(1,0)}$ and obtain $\pi^{(n)} = \pi^{(0)} p^n$.

Note that a transition matrix P is a **stochastic matrix**, that is it satisfies the following two conditions:

1. all its entries are nonnegative, i.e. $P_{ij} \geq 0$ for all i, j, and
2. each row sums to one, which may be written $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the (column) vector of ones.

Similarly $\pi^{(n)}$ satisfies $\pi^{(n)}\mathbf{1} = 1$. The eigenvalues of a stochastic matrix P can be shown to have modulus less than or equal to one, so that $|\det P| \leq 1$ also.

A simple example of a discrete time Markov chain is a random walk on a circle. Consider the following problem:

> There are n people sitting in a ring, one of whom takes a swig from a keg of beer, and then passes it left or right with a 50% probability independently of what has happened before. The process repeats until everyone has had at least one swig, then stops. Show that the probability that the keg stops at a particular (non-starting) person is independent of that person's position.

The situation described in this problem is a homogeneous Markov chain with n states (the n people) and transition matrix

$$P = \begin{pmatrix} 0 & 1/2 & 0 & 0 & \cdots & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & \cdots & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & \cdots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 1/2 & 0 & 0 & 0 & \cdots & 1/2 & 0 \end{pmatrix}.$$

(However this isn't the easiest way to solve the problem).

2.2.2 Continuous time Markov processes

Continuous time Markov processes are defined similarly to discrete time Markov chains, but now the index is continuous rather than discrete. More precisely, a family of random variables $\{X_t | 0 \leq t \in \mathsf{R}\}$ taking values in the state space $\mathcal{A}$ satisfies the **Markov property** if, for all times $t_0 < t_1 < \cdots < t_n$ and any states $i_0, i_1, \ldots, i_n$, we have

$$\mathsf{P}[X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}, \ldots, X_{t_0} = i_0] = \mathsf{P}[X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}]. \tag{2.4}$$

The analogy with equation (2.3) should be clear.

The analogue of the transition matrix is the **transition function**, defined by

$$p_{ij}(s,t) = \mathsf{P}[X_t = j | X_s = i].$$

The process is **homogeneous** if

$$p_{ij}(s,t) = p_{ij}(0, t-s)$$

for all $i, j \in \mathcal{A}$ and $s, t \geq 0$, in which case we write $p_{ij}(t-s)$ for $p_{ij}(s,t)$ and $P(t)$ for the $r \times r$ matrix with entries $p_{ij}(t)$. We consider only homogeneous processes. Note that if $\{X_t\}$ is a homogeneous Markov process then, for any $\tau > 0$, the stochastic process $\{X_{n\tau} | n \in \mathsf{N} \cup \{0\}\}$ is a homogeneous Markov chain, with transition matrix $P = P(\tau)$.

In general, a homogeneous Markov process may be written in terms of a **rate matrix** R as

$$P(t) = \exp(tR),$$

where for a matrix A we have

$$\exp(A) = \sum_{i=0}^{\infty} \frac{1}{n!} A^n.$$

Typically, we will define our Markov processes by specifying R and the (row vector) initial distribution of states π. The rate matrix may be found from $P(t)$ by right-differentiation at 0: that is

$$\lim_{h \to 0^+} \frac{P(h) - I}{h} = R.$$

The off diagonal elements R_{ij} may be interpreted as the "rate" at which state i turns into state j and are all non-negative. The rows of R sum to zero which may be interpreted as "conservation of state".

A Markov process with rate matrix R and initial distribution π is **stationary** if $\pi R = \mathbf{0}$, where $\mathbf{0}$ is the (column) vector of zeros. This is a necessary and sufficient condition that $\pi P(t) = \pi$ for all t, that is that the distribution of states does not vary with time. If $\{X_t | -\infty < t < \infty\}$ is a Markov process with stationary distribution π, define the **reversed process** $\{Y_t | -\infty < t < \infty\}$ by $Y_t = X_{-t}$. The reversed process is also Markov with stationary distribution π, and we say that $\{X_t\}$ is **reversible** if $\{X_t\}$ and $\{Y_t\}$ have the same transition function. A necessary and sufficient condition for reversibility is that $\pi_i R_{ij} = \pi_j R_{ji}$ for all i, j. If we let $\Pi = \mathrm{diag}(\pi)$, this condition is that ΠR should be symmetric.

Stationary and reversible Markov processes are frequently used in modelling DNA substitution, not so much for being more realistic, but because they are more mathematically tractable. Firstly, in a tree setting, reversibility allows us to re-root the tree arbitrarily, and so deal with unrooted rather than rooted trees; secondly, a stationary and time-reversible rate matrix is always diagonalisable so the Markov process always has a spectral representation (see Keilson [23, pp. 32–35]). Since ΠR is symmetric so is $\Pi^{1/2} R \Pi^{-1/2}$ which therefore has real eigenvalues $\{\lambda_j\}$ and orthonormal eigenvectors $\{u_j\}$ (related to the eigenvectors $\{v_j\}$ of R by $v_j = \Pi^{-1/2} u_j$ and $R v_j = \lambda_j v_j$). We then find that

$$J(t) = \Pi P(t) = \sum_{j=1}^{r} e^{\lambda_j t} w_j w_j^{\mathrm{T}}, \tag{2.5}$$

where $w_j = \Pi^{1/2} u_j$ and the superscripted T denotes transposition.

We conclude by looking at two state continuous time Markov processes briefly, as they will be relevant in later work on the covarion model.

A two state continuous time Markov process has rate matrix of the form

$$R = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

where $\alpha, \beta \geq 0$. Left eigenvectors of 0 are scalar multiples of $\pi = (\beta/(\alpha+\beta), \alpha/(\alpha+\beta))$; we then have

$$\Pi R = \frac{\alpha\beta}{\alpha+\beta} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

so that any stationary two state process is also reversible. To calculate $P(t)$ for the stationary process we diagonalise R and find that

$$P(t) = \frac{1}{\alpha+\beta} \begin{pmatrix} \alpha h(t) + \beta & \alpha(1-h(t)) \\ \beta(1-h(t)) & \alpha + \beta h(t) \end{pmatrix} \tag{2.6}$$

where $h(t) = \exp(-t(\alpha+\beta))$.

2.3 A bit of biology

Since phylogenetics is a part of mathematical biology it is inevitable that some terminology of biological origin will creep into usage. We explain (or perhaps rather, translate into mathematics) some of the main terms that will crop up below, and give a little bit of background. For more details see a biology text, or, for a presentation perhaps more appealing to a mathematician, Hofstadter's *Gödel, Escher, Bach: an Eternal Golden Braid* [20].

The **taxa** are the objects to be classified. A **sequence** will be a word in some alphabet, which we will usually call the **state space**, and a **site** is a particular position in a sequence, for example "the eighth letter". A **substitution** is a change in a sequence at a particular site, as the sequence evolves. In the case of DNA the state space is usually $\{A, G, C, T\}$ corresponding to the four nucleotides adenine, guanine, cytosine and thymine. Adenine and guanine are *purines* and cytosine and thymine are *pyrimidines*. A substitution from a purine to a purine or a pyrimidine to a pyrimidine is called a *transition*, while a change to or from a purine from or to a pyrimidine is called a *transversion*. Since DNA sequences are the main application most people have in mind, many models tailored to four states or to DNA in particular have been proposed (for example, models incorporating a "transition bias", since observationally transitions appear to be more common than transversions), but where possible we will work in a greater generality that allows such models as special cases.

DNA sequences code for amino acids according to the genetic code. *Codons* (ordered triples of nucleotides) each code for one of twenty amino acids or certain punctuation marks, and DNA sequences are sometimes analysed by translating them into amino acid sequences and then applying a substitution model at the amino acid level.

An important issue when comparing sequences is that of **alignment**. In comparing two sequences we naturally want to compare corresponding regions; usually, this means comparing the same site of each sequence. A difficulty then arises when comparing sequences from different species as they often have different lengths. When this occurs it is necessary to insert gaps in one or more sequences in such a way that all have the same length and corresponding sites of each sequence line up, a process called *alignment*. Since this thesis is about substitution models rather than alignment, we will always assume that the alignment has been taken care of by someone else and will ignore the issue of how to interpret sites where one or more sequences have a gap.

Chapter 3

Substitution models

We now combine the two main themes of the previous chapter—trees and Markov processes—and consider Markov processes on trees, as they are used to model nucleotide substitution.

Phylogenetically, we interpret each vertex of a tree as representing a species, with edges denoting immediate ancestor-descendent relationships. The labelled vertices represent the species with which they are labelled, and unlabelled vertices represent inferred ancestral species. In rooted trees the root represents a common ancestor from which all other species in the tree are descended. We do not allow vertices of degree two, except possibly at the root or at labelled vertices, since we are primarily interested in speciation events, where the tree "branches". A central problem of phylogenetics is to reconstruct this tree simply from observations of the species in the labelling set.

The main observations used in the modern approach to phylogenetics are the molecular sequences that characterise each species, such as DNA and protein sequences. Increasing numbers of sequences of increasing length are being determined all over the world, resulting in a wealth of phylogenetic data. In order to make use of this data it is important to model as best as we can the process by which the ancestral sequence gave rise to the sequences we see today. Such models are the main subject of this thesis, and in this chapter we describe the "basic" model on which all other substitution models we will study are in some way based.

Any exercise in modelling involves making simplifying assumptions about the process being modelled. Sometimes these assumptions are unstated and lie hidden in our mental picture of what we are modelling—an often unspoken assumption in phylogenetics being that the evolutionary relationships we are trying to determine are best described by a tree and not some other type of graph—so it is perhaps important to make this picture clear. We will imagine that there was some ancestral sequence (species), and that as time passed random errors (substitutions) occurred in this sequence. Sometimes this gave rise to two distinct sequences that both "survived" so that we had two sequences both behaving as the original one did. This continued until finally we had some number of sequences, which are what we are able to observe today, whose ancestral relationships are described by a rooted tree. What we want then is something that will simulate this process and generate "observations", or present day sequences.

The substitution models we will consider do not incorporate a "speciation" process by which the tree "forks". Rather, they treat the tree as a parameter and simulate the substitution process given a particular rooted tree. Nor do they treat the sequences strictly as sequences: to keep things tractable, they treat each site independently. So for us a substitution model will be a "character generator", that generates characters given a particular tree and the values of any parameters in the model. In this framework, the tree reconstruction problem is to determine the tree (and perhaps the values of the parameters) from the frequencies of the characters.

3.1 The "basic" model

The basic substitution model consists of

1. a rooted vertex-labelled tree T;
2. a set of states $\mathcal{A}$;
3. a distribution π of states at the root ρ, and
4. a transition matrix P^e on each edge $e \in E(T)$.

We then imagine that the state at the root "evolves" down the tree, generating a state function $\hat{\chi} : V(T) \mapsto \mathcal{A}$ (in other words, we have a family of $\mathcal{A}$-valued random variables, indexed by the vertex set). We suppose that this takes place such that there is a total order $\leq$ of the vertices, respecting ancestry (so if u is on the path from v to the root then $u \leq v$), such that

$$\mathsf{P}\left[\hat{\chi}(v) = i \,\middle|\, \bigwedge_{w<v} \hat{\chi}(w)\right] = \mathsf{P}[\hat{\chi}(v) = i | \hat{\chi}(w_0)], \tag{3.1}$$

for all $i \in \mathcal{A}$ and $v \in V(T)$, where w_0 is the immediate ancestor of v. Furthermore, if $e = (u, v)$ is an edge of T, then

$$P^e_{ij} = \mathsf{P}[\hat{\chi}(v) = j | \hat{\chi}(u) = i].$$

With these assumptions, the probability of generating a given state function $\hat{\chi}$ may be written

$$\mathsf{P}[\hat{\chi}|T, \pi, \{P^e\}] = \pi_{\hat{\chi}(\rho)} \prod_{(u,v)\in E(T)} P^{(u,v)}_{\hat{\chi}(u)\hat{\chi}(v)}. \tag{3.2}$$

Characters are obtained by restricting the state function to the labelled vertices, so the probability of generating a given character χ is found by summing (3.2) over all state functions extending χ; that is

$$\mathsf{P}[\chi|T, \pi, \{P^e\}] = \sum_{\hat{\chi}\uparrow\chi} \mathsf{P}[\hat{\chi}|T, \pi, \{P^e\}].$$

Figure 3.1 shows how to calculate $\mathsf{P}[\chi|T, \pi, \{P^e\}]$ directly from the definition for a simple tree and character. Note however that the number of extensions of a character can grow exponentially with the number of taxa so computationally it is impractical to calculate $\mathsf{P}[\chi|T, \pi, \{P^e\}]$ in this way. In practice dynamic programming techniques that work from the leaves up are used to calculate $\mathsf{P}[\chi|T, \pi, \{P^e\}]$ efficiently.

A set of aligned molecular sequences of length k is thought of as a set of k characters. Characters are usually assumed to evolve "i.i.d.", that is identically and independently distributed. However, the basic model with the i.i.d assumption does not appear to be very realistic (some sites appear to evolve "faster" than others, and some appear to change very slowly or not at all), and Chapter 6 of this thesis will compare two alternative models that attempt to be more realistic without sacrificing the tractability of an i.i.d. model.

Note that condition (3.1) is different to the usual notion of a Markov process on a graph G, that of a Markov random field. This is a family of $\mathcal{A}$ valued random variables $\{X_v | v \in V(G)\}$, as before, but now we assume that

$$\mathsf{P}\left[X_{v_0} = i \,\middle|\, \bigwedge_{v\in V(G)} (X_v = i_v)\right] = \mathsf{P}\left[X_{v_0} = i \,\middle|\, \bigwedge_{v\in N(v_0)} (X_v = i_v)\right], \tag{3.3}$$

where $N(v_0)$ is the set of neighbours of v_0. Joe Chang, in a private communication to Mike Steel, suggests that these two notions are equivalent on trees, and that one direction (Markov random field implies condition (3.1)) may be proved using the Hammersley-Clifford Theorem. This result states that a random field having the Markov property (3.3) is equivalent to it having a Gibbs distribution (see for example [27]).

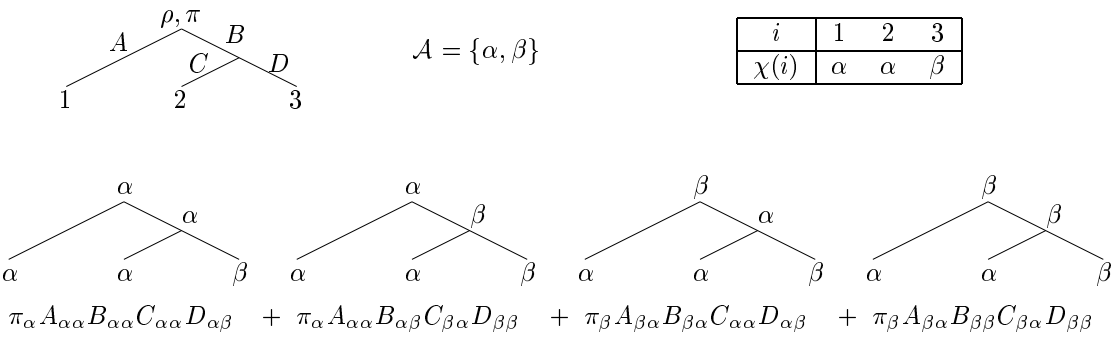


Figure 3.1: Calculating $\mathsf{P}[\chi|T, \pi, \{P^e\}]$ for a simple tree and character. Consider the tree with root distribution π and transition matrices A, B, C and D as shown, with state space $\mathcal{A} = \{\alpha, \beta\}$. The probability of generating the character that assigns state α to leaves 1 and 2 and state β to leaf 3 is given by a sum over all possible assignments of states to the remaining vertices.

3.2 Continuous time substitution models

In a continuous time substitution model we restrict the class of allowed transition matrices, drawing them from a continuous time Markov process. Each edge e is given a non-negative weight τ_e and P^e is taken to be $\exp(\tau_e R)$ for a fixed rate matrix R. The process is frequently chosen to be stationary and reversible; a further assumption that is sometimes made is the **molecular clock** assumption that the distance from the root to any leaf (that is, the sum of the τ_e along the path from the root to the leaf) should be the same for every leaf. We will usually assume stationarity and reversibility but will not explicitly build in the molecular clock.

When $r = 4$ there are certain standard choices for R, corresponding to certain assumptions about which nucleotides may be substituted for more easily. The hierarchy of these choices is shown in [38, page 434]. One of these standard choices is the **K3ST** or Kimura three substitution-type model [26], in which R is assumed to have the form

$$R_K = \begin{pmatrix} -\delta & \alpha & \beta & \gamma \\ \alpha & -\delta & \gamma & \beta \\ \beta & \gamma & -\delta & \alpha \\ \gamma & \beta & \alpha & -\delta \end{pmatrix},$$

where $\delta = \alpha + \beta + \gamma$. If we further require that $\beta = \gamma$ then we obtain the **Kimura two parameter** model (K2P) [25]; setting $\alpha = \beta = \gamma$ gives the **Jukes-Cantor** model (JC) [22]. The K3ST model assumes that each nucleotide occurs with equal frequency and has a different rate for transitions ($A \leftrightarrow G$ or $C \leftrightarrow T$) and two types of transversions ($A \leftrightarrow T$ or $C \leftrightarrow G$, and $A \leftrightarrow C$ or $G \leftrightarrow T$).

A standard choice sometimes made when $r = 2$ is the **Cavender-Farris** model (CF) [5, 10] which has rate matrix

$$R = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

This and the Jukes-Cantor model are particular cases of the **fully-symmetric** model, which for r states has rate matrix a multiple of $\mathbf{1}\mathbf{1}^{\mathrm{T}} - rI_r$ (that is $1 - r$ on the main diagonal and 1 everywhere else). This model appears in Neyman [31], and a slightly modified version in which edges are allowed to have "infinite" length is studied in [41].

3.3 Reconstruction techniques

Our aim in modelling the substitution process is to turn character frequency data into trees. The two main approaches based on substitution models are methods that seek to invert the model (often distance based) and maximum likelihood methods.

Distance methods, of which the LogDet transformation we will meet in the next chapter is an example, aim to turn character frequency data into a tree-additive distance that is realised by the tree that generated the data. The tree can then be recovered, due to the uniqueness theorems for tree metrics, and the edge weightings can usually be interpreted to give an indication of the divergence times. Although this sounds like the ideal way to reconstruct trees, there are some difficulties. Chiefly, since the observed character frequencies only approximate the expected frequencies, the distance obtained usually is not a tree-additive distance, and it is necessary to try to find a tree-additive distance that is in some sense "close" to the observed distance. Finding methods of doing so is consequently an important problem in phylogenetics.

Maximum likelihood methods seek to choose the tree that best explains the data under the assumed model. This usually involves a two step optimisation procedure: first the probability of generating the data on a given tree is optimised over the parameters of the model (for example, under the assumption of a K3ST model, this would involve maximising over α, β, γ and the edge lengths), then this maximum is optimised over all possible trees. Again the edge lengths can usually be interpreted to give divergence times. The main difficulty with this approach is that it is very computationally intensive, especially when the number of taxa is large. The number of binary trees on n leaves is $(2n-5)!! = 1.3.5\ldots(2n-5)$ which grows very rapidly, so it soon becomes infeasible to optimise over all n taxa trees. Usually heuristic approaches are used: a "good guess" at the tree is built up in some way, and then various rearrangements are made to try to find a tree with a higher likelihood value.

Chapter 4

The LogDet transformation: there can be only one

4.1 The LogDet transformation

An important issue in reconstructing trees from sequence data is whether the sequences actually contain enough information about the tree to do so. The LogDet transformation, due to Chang and Hartigan [8] and independently to Steel [36] (see also Lake [28] and Zarkikh [46]), shows that under some very mild assumptions, tree reconstruction is in fact possible under the basic model. To motivate this transformation, consider the tree in figure 4.1.

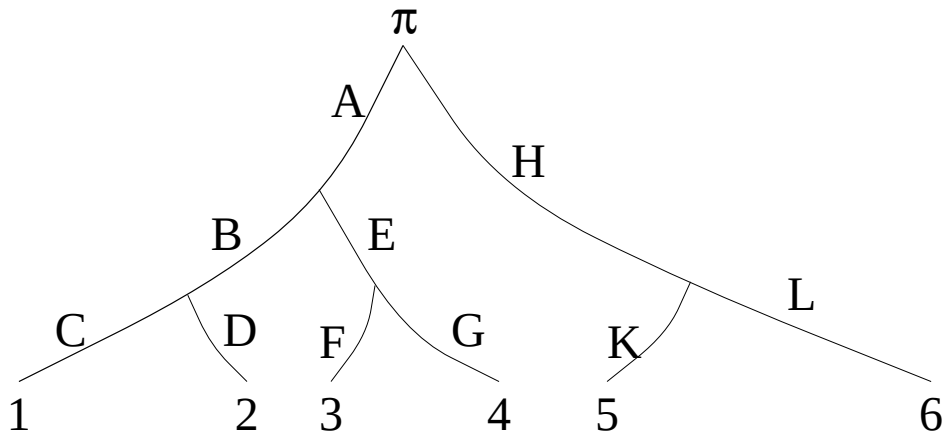


Figure 4.1: An example tree to motivate the LogDet transformation. The numbers are the leaf labels, the letters the transition matrices, and π is the root distribution.

The **joint probability matrix** $J(x, y)$ of taxon x and taxon y is the matrix with ij-entry

$$J_{ij}(x, y) = \mathsf{P}[(\chi(x) = i) \wedge (\chi(y) = j)].$$

Consider $J(1, 6)$. We have

$$\begin{aligned} J_{ij}(1,6) &= \sum_{k,l,m,p} \pi_k A_{kl} B_{lm} C_{mi} H_{kp} L_{pj} \\ &= (C^{\mathrm{T}} B^{\mathrm{T}} A^{\mathrm{T}} \Pi H L)_{ij} \end{aligned}$$

where $\Pi = \mathrm{diag}(\pi)$, so that $J(1,6) = C^{\mathrm{T}} B^{\mathrm{T}} A^{\mathrm{T}} \Pi H L$. Provided the determinant of Π and every transition matrix on the path from leaf 1 to leaf 6 is nonzero, we can take the logarithm of the absolute value of the determinant of $J(1,6)$ to get

$$\log|\det J(1,6)| = \log|\det C| + \log|\det B| + \log|\det A| + \log|\det \Pi| + \log|\det H| + \log|\det L|.$$

The right-hand side has split into a sum with a contribution from each edge, which suggests that

$$d_{xy} = -\log|\det J(x,y)| \tag{4.1}$$

(the negative sign is because the determinant of a transition matrix has modulus less than or equal to one) might be a tree-additive distance, with edge weights roughly equal to $-\log|\det P^e|$, perhaps with some correction due to the root distribution. This turns out to be the case, under the assumption that

$$\text{for every edge } e,\ \det P^e \notin \{0, \pm 1\};\ \pi_i > 0 \text{ for every state } i. \tag{4.2}$$

Writing $\pi_k(v)$ for $\mathsf{P}[\hat{\chi}(v) = k]$, the correct choice of edge weighting (Steel [36]) is

$$\lambda(e) = -\log|\det P^e| - \frac{1}{2}\log \prod_{k \in \mathcal{A}} \pi_k(v)$$

if $e = (v, w)$ and w is a leaf, and

$$\lambda(e) = -\log|\det P^e| - \frac{1}{2}\log \prod_{k \in \mathcal{A}} \pi_k(v) + \frac{1}{2}\log \prod_{k \in \mathcal{A}} \pi_k(w)$$

if $e = (v, w)$ and w is not a leaf.

Note that in reconstructing T from d, there is no information as to the placement of the root; we instead obtain the unrooted tree $T^{-\rho}$ obtained from T by deleting the root ρ and identifying the two incident edges if ρ has degree two, or simply regarding T as unrooted if ρ has degree other than two. Hence we have:

Theorem 4.1 (Steel [36]) *Each leaf-labelled tree, up to the placement of its root, is uniquely defined by the character frequencies it generates under the basic model with assumption (4.2).*

More recently Chang [6] gives conditions under which the transition matrices, in addition to the tree topology, may be reconstructed from distributions of triples.

4.2 There can be only one

What makes the LogDet transformation work? It should be clear that, restricting our attention for the moment to transition matrices with positive determinant, the main reason it works is because the map $\phi = \log\det$ is a homomorphism into the real numbers under addition. A second property of ϕ is that it is continuous, which is important since we would like our output trees to depend continuously on the input. We show here that $\phi = \log\det$ is the only map with these two properties, up to scalar multiples.

We first make some notational definitions. Let $M_n = M_n(\mathsf{R})$ be the set of $n \times n$ matrices with real entries, and let $GL_n = GL_n(\mathsf{R}) \subseteq M_n$ be the group of non-singular $n \times n$ matrices. We denote certain subsets of M_n as follows, where $\mathbf{1}$ is the vector of ones.

$$\begin{aligned} R_n &= \{M \in M_n | M\mathbf{1} = \mathbf{1}\} \\ R_n^+ &= \{M \in R_n | \det M > 0\} \\ T_n &= \{M \in R_n | M_{ij} \geq 0\ \forall\ i, j\} \\ T_n^+ &= T_n \cap R_n^+ \end{aligned}$$

R_n is the set of matrices with row sums one, and R_n^+ consists of those matrices with positive determinant. T_n is the set of transition matrices, and T_n^+ is the set of transition matrices with positive determinant.

Clearly R_n is closed under multiplication, since if $A, B \in R_n$ then $AB\mathbf{1} = A\mathbf{1} = \mathbf{1}$; further, if $A \in R_n$ is non-singular, then $A\mathbf{1} = \mathbf{1}$ so $A^{-1}A\mathbf{1} = A^{-1}\mathbf{1}$, giving $A^{-1} \in R_n$. It follows that R_n^+ is a group. T_n is also closed under multiplication, so T_n and T_n^+ are sub-semigroups of R_n and R_n^+ respectively. We note also that if $M \in T_n$ then $|\det M| \leq 1$.

Theorem 4.2 *If $\psi : T_n^+ \mapsto (\mathsf{R}, +)$ is a continuous homomorphism then ψ is a scalar multiple of $\phi = \log\det$.*

Proof We show that ψ lifts to a continuous homomorphism $\bar{\psi} : R_n^+ \mapsto (\mathsf{R}, +)$ such that $\bar{\psi}|_{T_n^+} = \psi$, and that $\bar{\phi} = \log\det$ is the only continuous homomorphism of R_n^+ into $(\mathsf{R}, +)$ up to scalar multiples. It follows that $\psi = \bar{\psi}|_{T_n^+}$ is a scalar multiple of $\log\det$.

Lemma 4.1 *R_n^+ is generated by T_n^+.*

Proof Since $T_n^+ \subseteq R_n^+$ the group $\langle T_n^+ \rangle$ generated by T_n^+ is contained in R_n^+. We show $\langle T_n^+ \rangle = R_n^+$ by showing that R_n^+ is connected, which implies it is generated by each neighbourhood of the identity. We then show that there is $a \in T_n^+$ and an open neighbourhood U of a in R_n^+ that is contained in T_n^+. Then $a^{-1}U \subseteq \langle T_n^+ \rangle$ is an open neighbourhood of I in R_n^+ and so generates R_n^+, giving $R_n^+ \subseteq \langle T_n^+ \rangle$.

To show that R_n^+ is connected we change basis for R^n to an orthonormal basis $\mathcal{B}$ with first element $\mathbf{1}/||\mathbf{1}||$. This transformation may be written $M \mapsto TMT^{-1}$ for invertible T and so gives a (topological) isomorphism of R_n^+ onto L_n^+, where L_n^+ is the set of matrices of the form

$$\begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & M \end{pmatrix} \quad : \quad c \in \mathsf{R}^{n-1}, \quad M \in GL_{n-1}^+ \, .$$

Subtracting t times the first column from the ith column for t from 0 to c_i, for each column in turn, gives a path from

$$\begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & M \end{pmatrix} \quad \text{to} \quad \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & M \end{pmatrix} ,$$

so we need only show that $\{A \in L_n^+ | A_{1j} = 0, j = 2, \ldots, n\}$ is connected. Since this is homeomorphic to $GL_{n-1}^+(\mathsf{R})$ which is connected [32, page 22], R_n^+ is connected.

To see that there is a and U as claimed, consider

$$a = \exp(\mathbf{1}\mathbf{1}^{\mathrm{T}} - nI_n) = \begin{pmatrix} 1-p & \frac{p}{n-1} & \cdots & \frac{p}{n-1} \\ \frac{p}{n-1} & 1-p & & \vdots \\ \vdots & & \ddots & \frac{p}{n-1} \\ \frac{p}{n-1} & \cdots & \frac{p}{n-1} & 1-p \end{pmatrix},$$

where $p = \frac{n-1}{n}(1 - \exp(-n))$. We have $\det a = e^{-n(n-1)} > 0$ and $a_{ij} > 0$ for each i, j so $a \in T_n^+$. Since det is continuous, we may choose an open ball about a in R_n such that if $b \in U$ then $0 < \det b < 1$ and $b_{ij} > 0$ for all i, j. Hence U is a neighbourhood of a contained in T_n^+. This completes the proof.

•

We now show that a continuous homomorphism $\psi : T_n^+ \mapsto (\mathsf{R}, +)$ lifts to a continuous homomorphism $\bar{\psi} : R_n^+ \mapsto (\mathsf{R}, +)$ such that $\bar{\psi}|_{T_n^+} = \psi$. To do this we require the following algebraic lemma.

Lemma 4.2 *Let G, H be groups and S a generating sub-semigroup of G. If, for each $a, b \in S$ there exists $m, n \in S$ such that $am = bn$, then any homomorphism $\phi : S \mapsto H$ lifts to a homomorphism $\bar{\phi} : G \mapsto H$ such that $\bar{\phi}|_S = \phi$. Further, if G and H are topological groups, ϕ is continuous and there is U open in G contained in S, then $\bar{\phi}$ is continuous.*

Proof Since S generates G, each $g \in G$ may be written as a finite product of the form $g = s_1^{\epsilon_1} \cdots s_k^{\epsilon_k}$ where each s_i is in S and each $\epsilon_i \in \{\pm 1\}$. The map given by

$$\bar{\phi}(s_1^{\epsilon_1} \cdots s_k^{\epsilon_k}) = \phi(s_1)^{\epsilon_1} \cdots \phi(s_k)^{\epsilon_k}$$

will be the required homomorphism provided it is well defined. It suffices to show that if $s_1^{\epsilon_1} \cdots s_k^{\epsilon_k} = 1_G$ then $\phi(s_1)^{\epsilon_1} \cdots \phi(s_k)^{\epsilon_k} = 1_H$.

Using the homomorphism property of ϕ in S, we may assume that plus and minus signs alternate in any expression of the form $s_1^{\epsilon_1} \cdots s_k^{\epsilon_k}$. We now use the fact that for each $a, b \in S$ there is $m, n \in S$ such that $am = bn$ to gather all the inverses at one end, since this allows us to write $a^{-1}b$ as mn^{-1}. Further,

$$\phi(a)\phi(m) = \phi(am) = \phi(bn) = \phi(b)\phi(n)$$

and hence $\phi(a)^{-1}\phi(b) = \phi(m)\phi(n)^{-1}$, so that in re-writing the product in G we do not change the value of the product in H. Thus each expression $s_1^{\epsilon_1} \cdots s_k^{\epsilon_k} = 1_G$ may be re-written as $s_1' s_2'^{-1} = 1_G$ with $\phi(s_1)^{\epsilon_1} \cdots \phi(s_k)^{\epsilon_k} = \phi(s_1')\phi(s_2')^{-1}$. But if $s_1' s_2'^{-1} = 1_G$ then $s_1' = s_2'$, so $\phi(s_1')\phi(s_2')^{-1} = 1_H$, and therefore $\bar{\phi}$ is well defined.

Note that as a consequence of the above argument, every element of G may be written in the form ab^{-1} for $a, b \in S$.

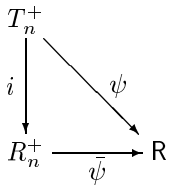
Suppose now that G and H are topological groups, ϕ is continuous and there is U open in G contained in S. To show that $\bar{\phi}$ is continuous we need only show that for every neighbourhood N of 1_H there is a neighbourhood N' of 1_G such that $\bar{\phi}(N') \subseteq N$. Choose $a \in U$. Then $\phi(a)N$ is a neighbourhood of $\phi(a)$; since ϕ is continuous, there is a neighbourhood N'' of a contained in U such that $\phi(N'') \subseteq \phi(a)N$. Then $N' = a^{-1}N''$ is a neighbourhood of 1_G, and

$$\bar{\phi}(N') = \bar{\phi}(a)^{-1}\bar{\phi}(N'') = \phi(a)^{-1}\phi(N'') \subseteq \phi(a)^{-1}\phi(a)N = N,$$

so $\bar{\phi}$ is continuous.

•

Corollary 4.1 *Any continuous homomorphism $\psi : T_n^+ \mapsto (\mathsf{R}, +)$ lifts to a continuous homomorphism $\bar{\psi} : R_n^+ \mapsto (\mathsf{R}, +)$ such that $\bar{\psi}|_{T_n^+} = \psi$. That is, such that the diagram*



commutes, where $i : T_n^+ \mapsto R_n^+$ is the inclusion map.

Proof In order to use Lemma 4.2, we must show that for each $A, B \in T_n^+$ there are $M, N \in T_n^+$ such that $AM = BN$; for the open set U we may use the neighbourhood U of a in the proof of Lemma 4.1.

Let $\pi = (1/n, \ldots, 1/n)$ and let $P(t) = \exp(tR_\pi)$ where $R_\pi = \mathbf{1}\pi - I_n$. Then $P(t) \in T_n^+$ for all $t \geq 0$, and $P(t) \to \mathbf{1}\pi$ as $t \to \infty$. Consider $F(t) = B^{-1}AP(t)$.

As $t \to \infty$,

$$F(t) \to B^{-1}A\mathbf{1}\pi = B^{-1}(A\mathbf{1})\pi = B^{-1}\mathbf{1}\pi = \mathbf{1}\pi.$$

Choose an open neighbourhood V of $\mathbf{1}\pi$ in R_n such that if $C \in V$ then $C_{ij} > 0$ for all i, j and $|\det C| < 1/2$. Since $F(t) \to \mathbf{1}\pi$, there is τ such that $F(\tau) \in V$. Then $B^{-1}AP(\tau)$ has positive entries, and $\det B^{-1}AP(\tau) > 0$, so $B^{-1}AP(\tau) = N \in T_n^+$. Letting $M = P(\tau) \in T_n^+$ we obtain $AM = BN$ and the lemma follows.

•

We now show that any continuous homomorphism $\bar{\psi}$ from R_n^+ into $(\mathsf{R}, +)$ must be a scalar multiple of $\bar{\phi} = \log\det$. Since $(\mathsf{R}, +)$ is abelian, the kernel of $\bar{\psi}$ must contain the commutator subgroup of R_n^+. Calculating this subgroup is the substance of the following lemma.

Lemma 4.3 *The commutator subgroup $R_n^{+\prime}$ of R_n^+ is*

$$R_n^+ \cap SL_n(\mathsf{R}) = \{M \in R_n^+ \,|\, \det M = 1\}.$$

Proof Again we use the isomorphism of R_n^+ to L_n^+ from the proof of Lemma 4.1. This isomorphism preserves det, so we must show that $L_n^{+\prime} = L_n^+ \cap SL_n(\mathsf{R})$. Since $\det ABA^{-1}B^{-1} = 1$, one inclusion is immediate. For the reverse inclusion, we must show that all matrices of the form

$$\begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & C \end{pmatrix} \quad : \quad c \in \mathsf{R}^{n-1} \quad C \in SL_{n-1}^+$$

are in $L_n^{+\prime}$.

We first note that, since $SL_n' = SL_n$ for all n, the commutator subgroup of GL_{n-1}^+ is SL_{n-1}. Hence each $C \in SL_{n-1}$ may be written either as a commutator of matrices in GL_{n-1}^+ or as a product of such commutators. Suppose C is a commutator in GL_{n-1}^+, say $C = ABA^{-1}B^{-1}$. Since $(kA)B(kA)^{-1}B^{-1} = C$, we may assume that A does not have 1 as an eigenvalue. $A^{-1} - I_{n-1}$ is then non-singular, and we may put

$$d^{\mathrm{T}} = c^{\mathrm{T}}B(A^{-1} - I_{n-1})^{-1}.$$

The commutator of $\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & A \end{pmatrix}$ and $\begin{pmatrix} 1 & d^{\mathrm{T}} \\ \mathbf{0} & B \end{pmatrix}$ is then

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & A \end{pmatrix}\begin{pmatrix} 1 & d^{\mathrm{T}} \\ \mathbf{0} & B \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & A^{-1} \end{pmatrix}\begin{pmatrix} 1 & -d^{\mathrm{T}}B^{-1} \\ \mathbf{0} & B^{-1} \end{pmatrix} &= \begin{pmatrix} 1 & d^{\mathrm{T}} \\ \mathbf{0} & AB \end{pmatrix}\begin{pmatrix} 1 & -d^{\mathrm{T}}B^{-1} \\ \mathbf{0} & A^{-1}B^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & -d^{\mathrm{T}}B^{-1} + d^{\mathrm{T}}A^{-1}B^{-1} \\ \mathbf{0} & ABA^{-1}B^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & C \end{pmatrix}, \end{aligned}$$

since

$$-d^{\mathrm{T}}B^{-1} + d^{\mathrm{T}}A^{-1}B^{-1} = d^{\mathrm{T}}(A^{-1} - I_{n-1})B^{-1} = c^{\mathrm{T}}.$$

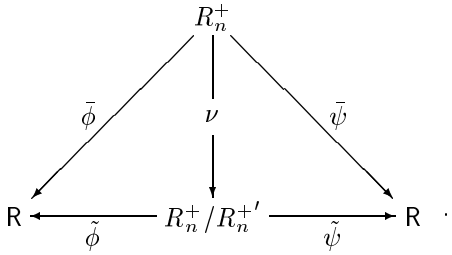
If C is a product of commutators in GL_{n-1}^+ then $C = C'D$ where D is a commutator and C' a product of commutators. From above, $\begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & D \end{pmatrix}$ and $\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & C' \end{pmatrix}$ are in $L_n^{+\prime}$, so

$$\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & C' \end{pmatrix}\begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & D \end{pmatrix} = \begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & C'D \end{pmatrix} = \begin{pmatrix} 1 & c^{\mathrm{T}} \\ \mathbf{0} & C \end{pmatrix}$$

is in $L_n^{+\prime}$. The lemma follows.

•

Let $\bar{\psi}$ be a continuous homomorphism of R_n^+ into $(\mathsf{R}, +)$. Since R_n^+ is connected and $\bar{\psi}$ continuous, $\bar{\psi}(R_n^+)$ is connected. The only connected subgroups of $(\mathsf{R}, +)$ are R itself and $\{0\}$; we need consider only the first case. Since $\bar{\phi} = \log\det$ is a continuous homomorphism of R_n^+ onto R with kernel $R_n^+ \cap SL_n = R_n^{+\prime}$, and $\ker\bar{\psi} \supseteq R_n^{+\prime}$, if $\ker\bar{\psi} \neq R_n^+$ then $\ker\bar{\psi} = \ker\bar{\phi} = R_n^{+\prime}$. We therefore obtain the following diagram, where ν, $\bar{\phi}$ and $\bar{\psi}$ are the natural maps:



The continuity of $\bar{\phi}$ and $\bar{\psi}$ imply that $\tilde{\phi}$ and $\tilde{\psi}$ respectively are continuous; moreover if $\bar{\phi}$ is an open mapping then $\tilde{\phi}^{-1}$ is continuous (see [33, Theorem 11]). Since log is open, we need only show that $\det : R_n^+ \mapsto (\mathsf{R}^+, \cdot)$ is open; this means showing that for every neighbourhood U of I_n in R_n^+, there is a neighbourhood U' of 1 in R such that $\det(U) \supseteq U'$.

Choose $1 > \epsilon > 0$ such that the matrix

$$J_a = I_n - aE_{11} + aE_{12}$$

(that is the matrix with 11-entry $1 - a$, 12-entry a, ones on the rest of the diagonal and zeros elsewhere) is in U whenever $|a| < \epsilon$. Since $\det J_a = 1 - a$, we have $(1 - \epsilon, 1 + \epsilon) \subseteq \det(U)$ and therefore det is open. It follows that $\tilde{\psi} \circ \tilde{\phi}^{-1} : (\mathsf{R}, +) \mapsto (\mathsf{R}, +)$ is a continuous isomorphism of $(\mathsf{R}, +)$.

Now the only continuous isomorphisms of $(\mathsf{R}, +)$ are scalar multiplication. To see this, suppose μ is such an isomorphism and let $r \in \mathsf{Q}$, with $r = p/q$ in lowest terms. Then

$$q\mu\left(\frac{p}{q}\right) = \mu(p) = p\mu(1).$$

Hence $\mu\left(\frac{p}{q}\right) = \frac{p}{q}\mu(1)$, so $\mu(r) = r\mu(1)$ for all $r \in \mathsf{Q}$. Since Q is dense in R and μ continuous, we have $\mu(x) = x\mu(1)$ for all $x \in \mathsf{R}$. Therefore

$$\tilde{\psi} \circ \tilde{\phi}^{-1} = k\,\mathrm{id}$$

for some k, and composing on the right with $\tilde{\phi}$ we get $\tilde{\psi} = k\tilde{\phi}$. The result follows.

•

4.2.1 The structure of R_2

The structure of R_2 is particularly simple and admits a second direct proof of Theorem 4.2, which we give here.

Each matrix on R_2 may be written in the form

$$m_{a,b} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

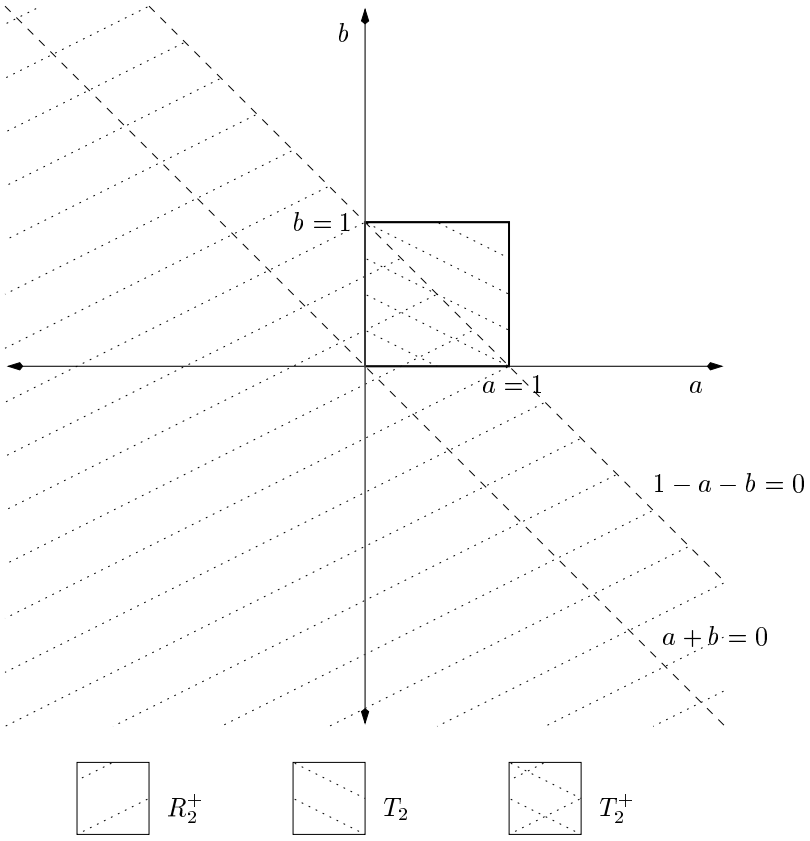


Figure 4.2: The structure of R_2.

for $a, b \in \mathsf{R}$, and conversely every such matrix belongs to R_2. $m_{a,b}$ has determinant $1 - a - b$, so level curves of det are lines with slope -1; in particular the singular matrices lie on $1 - a - b = 0$, and the matrices of determinant one lie on $a + b = 0$. T_2 is the box $[0, 1] \times [0, 1]$, and T_2^+ the triangle $\{(a, b) | 0 \leq a < 1, 0 \leq b < 1 - a\}$ (see figure 4.2). From equation (2.6) we see that the part of the line through the origin with slope β/α that lies below the line $1 - a - b = 0$ may be written as the one-parameter subgroup $\{\exp(tR) | t \in \mathsf{R}\}$, where

$$R = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}. \tag{4.3}$$

In particular every matrix in T_2^+ may be expressed as a transition matrix of a continuous time Markov process.

Let $\psi : T_2^+ \mapsto (\mathsf{R}, +)$. It is easily checked that if $S_R = \{\exp(tR) | t \geq 0\}$, where R is of the form (4.3), then $\psi|_{S_R}$ is a scalar multiple of $\log\det$. Consider in particular the sets $M_a = \{(a, 0) | a \in [0, 1)\} \subseteq T_2^+$ and $M_b = \{(0, b) | b \in [0, 1)\} \subseteq T_2^+$, which correspond to the rate matrices

$$\begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$$

respectively. We may write $\psi|_{M_a} = \lambda_a \log\det$ and $\psi|_{M_b} = \lambda_b \log\det$.

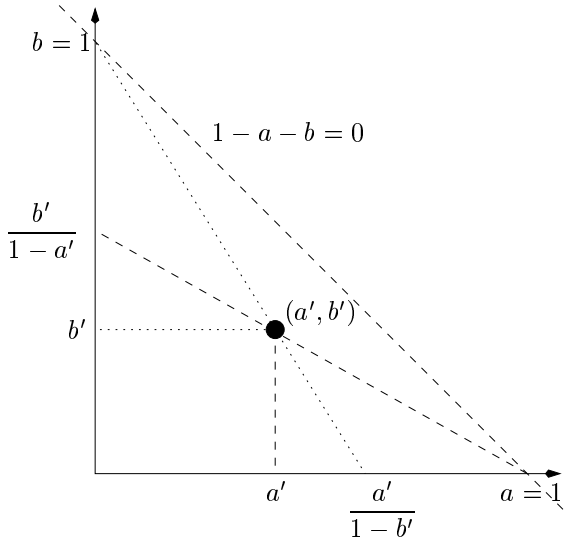


Figure 4.3: Writing T_2^+ as $M_a M_b$ and as $M_b M_a$. The dotted lines give a factorisation of (a', b') as a matrix in M_a times a matrix in M_b, while the dashed lines through (a', b') give a factorisation of (a', b') as a matrix in M_b times a matrix in M_a.

Let $(a', b') \in T_2^+$. We have

$$\begin{pmatrix} 1 - \frac{a'}{1-b'} & \frac{a'}{1-b'} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b' & 1-b' \end{pmatrix} = \begin{pmatrix} 1-a' & a' \\ b' & 1-b' \end{pmatrix} \tag{4.4}$$

and

$$\begin{pmatrix} 1 & 0 \\ \frac{b'}{1-a'} & 1 - \frac{b'}{1-a'} \end{pmatrix} \begin{pmatrix} 1-a' & a' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1-a' & a' \\ b' & 1-b' \end{pmatrix}, \tag{4.5}$$

(see figure 4.3), so $T_2^+ = M_a M_b = M_b M_a$. Applying $\exp \circ \psi$ to both sides of both (4.4) and (4.5), we get

$$\left(1 - \frac{a'}{1-b'}\right)^{\lambda_a} (1-b')^{\lambda_b} = \left(1 - \frac{b'}{1-a'}\right)^{\lambda_b} (1-a')^{\lambda_a}$$

which on rearranging gives

$$\left(\frac{1-a'-b'}{1-a'-b'+a'b'}\right)^{\lambda_a - \lambda_b} = 1.$$

It follows that $\lambda_a = \lambda_b = \lambda$, and hence $\psi = \lambda \log\det$.

Chapter 5

The reconstruction quotient

5.1 Introduction

Since the data for reconstructing trees comes from character frequencies, if two distinct (weighted) trees generate the same character frequencies no reconstruction technique will be able to distinguish between them. The LogDet transformation shows that, under the basic model, generically this is not the case: provided that $\pi_i > 0$ for all i and for every transition matrix M we have $\det M \notin \{0, \pm 1\}$, the underlying tree generating the characters may be reconstructed unambiguously. In this chapter we consider this problem in more detail for the modified version of the two state fully symmetric model in which we allow "infinite" edge lengths.

Theoretical methods of tree reconstruction usually assume that we have "ideal" data, that is, that we can calculate $\mathsf{P}[\chi]$ exactly for each character χ. In practice, since we only ever have finite sequences, this is not the case, and it may be that our estimates of $\mathsf{P}[\chi]$ do not correspond to ideal data from any weighted tree. When this occurs it may be necessary to approximate the observed data by a point in the space of ideal data. A first step towards doing this well is having a good picture of what the approximating space looks like. This question of what the space of ideal data looks like is one that has been raised by Joe Felsenstein (private communication to Mike Steel).

A second issue is that of what trees are to be considered "close". Consider the tree metrics given by the weighted trees shown in figure 5.1. As $\epsilon \to 0$, the metrics given by the binary trees tend to that given by the star tree, so it might be natural to think of these three weighted trees as being "close" for small values of ϵ, even though the underlying tree topologies are different. Since we would ideally like our output trees to depend in a continuous manner on our input data, a good understanding of this point is also important.

With these two related problems in mind we introduce the reconstruction quotient and prove two structure theorems for it in the case of the two state fully symmetric model. By a suitable choice of topology, the set of all transition matrix valued edge-weighted trees may be made into a topological space. Since no reconstruction method can distinguish between two trees that generate the same character frequencies, it is natural to identify weighted trees with the same image under the character frequency map. The resulting quotient space we will call the *reconstruction quotient*, and provided a good choice was made for the topology on the space of weighted trees, the topology on the quotient space should give a good insight into the problems raised above.

Work by Chang [6] giving conditions under which the transition matrices may be reconstructed in addition to the tree topology is also relevant to these problems.

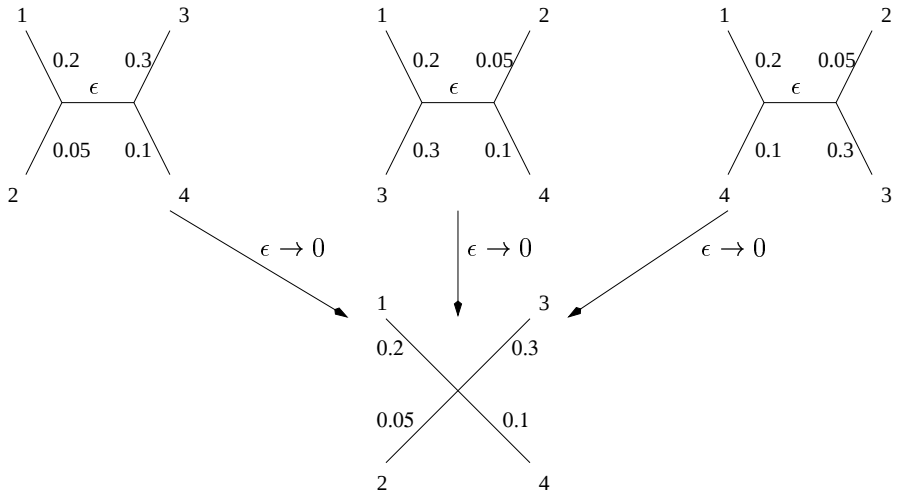


Figure 5.1: Consider the three binary trees with edge weights as shown, where $\epsilon > 0$. As $\epsilon \to 0$ the tree metric given by the edge weightings will tend towards that given by the weighted star tree, so it is natural to think of these weighted trees as "close" for small values of ϵ.

5.2 The two state fully symmetric model

The two state fully symmetric model has root distribution $\pi = (1/2, 1/2)$ and rate matrix

$$R = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

By (2.6) the transition matrices then have the form

$$P^e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix}, \tag{5.1}$$

where $p_e = \frac{1}{2}(1 - \exp(-2\tau_e))$ and is called the **mutation probability** of the edge. The mutation probabilities give a convenient description of the model: letting p be the vector of mutation probabilities we have simply

$$\mathsf{P}[\hat{\chi}|T,p] = \frac{1}{2} \prod_{\substack{\{u,v\}\in E(T):\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_{\{u,v\}}) \prod_{\substack{\{u,v\}\in E(T):\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} p_{\{u,v\}} \quad . \tag{5.2}$$

The mutation probability p_e is a monotonically increasing function of τ_e, satisfying $0 \leq p_e < 1/2$ on $[0,\infty)$. It is often convenient to modify this model to allow the possibility $p_e = 1/2$ and this is the model we will be considering here.

If we write a character $\chi : [n] \mapsto \{\alpha, \beta\}$ as the subset $\sigma = \chi^{-1}(\alpha)$, an alternative description of $\mathsf{P}[\chi|T,p]$ is given by

$$\mathsf{P}[\chi|T,p] = \frac{1}{2^n} \sum_{\substack{X\subseteq[n]:\\ |X|\equiv 0 \bmod 2}} (-1)^{|\sigma\cap X|} \prod_{e\in P(T,X)} (1-2p_e) \tag{5.3}$$

where $P(T,X)$ is the set of edges of T used by an odd number of paths when the vertices of X are matched arbitrarily by paths [18, 19, 39] (note that $P(T,X)$ is independent of the way in which the vertices are matched). We will write z_e for $1-2p_e$, and z for the vector $(z_e)_{e\in E(T)}$. Clearly $z_e \in [0,1]$ for each e.

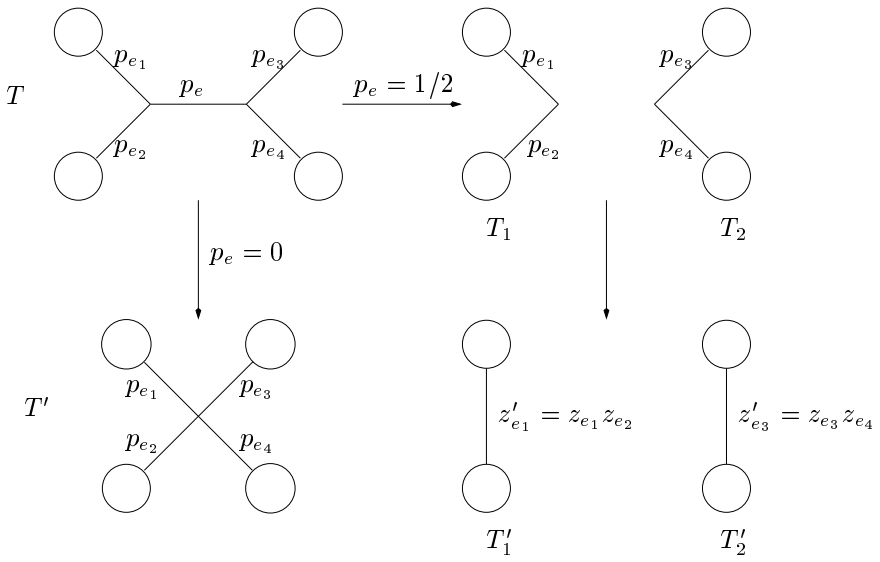


Figure 5.2: The effect of mutation probabilities of 0 and 1/2. Circles denote rooted subtrees. When $p_e = 0$, the effect is to contract e; when $p_e = 1/2$, the effect is to delete e. In the latter case we may suppress unlabelled vertices of degree two provided we weight the resulting edges appropriately.

5.3 Labelled forests

The transition matrix (5.1) has determinant $1 - 2p_e$, so we have $\det P^e \in \{0, \pm 1\}$ when $p_e \in \{0, 1/2\}$. Let T be a tree and suppose $p_e = 0$ for some edge $e = \{u, v\}$. If $\hat{\chi}(u) \neq \hat{\chi}(v)$ then $\mathsf{P}[\hat{\chi}|T, p] = 0$, so in calculating $\mathsf{P}[\chi|T, p]$ we need sum over only those extensions of χ such that $\hat{\chi}(u) = \hat{\chi}(v)$. For such extensions the edge e contributes a factor of 1 to $\mathsf{P}[\hat{\chi}|T, p]$, so that $\mathsf{P}[\chi|T, p] = \mathsf{P}[\chi|T', p']$ where T', p' are the tree and edge weighting obtained from T, p by contracting e (see figure 5.2). So the effect of having $p_e = 0$ is that e may be considered to be contracted, or to have zero length.

Suppose now that $p_e = 1/2$. Let T_1, T_2 be the trees obtained from T by deleting e and $\hat{\chi}_1$, $\hat{\chi}_2$ the restrictions of $\hat{\chi}$ to the vertices of T_1 and T_2 respectively. Let p_1 and p_2 be the restrictions of p to the edges of T_1 and T_2. Edge e contributes a factor of $1/2$ to $\mathsf{P}[\hat{\chi}|T, p]$ regardless of whether $\hat{\chi}(u) = \hat{\chi}(v)$ or not, so that

$$\begin{aligned} \mathsf{P}[\hat{\chi}|T, p] &= \frac{1}{4} \prod_{\substack{\{x,y\} \in E(T) \setminus e: \\ \hat{\chi}(x) = \hat{\chi}(y)}} (1 - p_{\{x,y\}}) \prod_{\substack{\{x,y\} \in E(T) \setminus e: \\ \hat{\chi}(x) \neq \hat{\chi}(y)}} p_{\{x,y\}} \\ &= \mathsf{P}[\hat{\chi}_1|T_1, p_1]\mathsf{P}[\hat{\chi}_2|T_2, p_2]. \end{aligned}$$

It follows that $\mathsf{P}[\chi|T, p] = \mathsf{P}[\chi_1|T_1, p_1]\mathsf{P}[\chi_2|T_2, p_2]$ for any character χ, where χ_1 and χ_2 are the restrictions of χ to the labels of T_1 and T_2 respectively, so the effect of $p_e = 1/2$ is to delete e, dividing T into two parts. The resulting trees T_1 and T_2 may have unlabelled vertices of degree two, but by suppressing such vertices and weighting the resulting edge appropriately as in figure 5.2 we may deal simply with trees without such vertices.

The above discussion motivates our introduction of labelled forests. Given a label set L, a **tree structure** on L or an **L-labelled tree** is a tree T and a labelling $\ell : L \mapsto V(T)$ such that every vertex of degree less than or equal to two is in the image of L. This terminology is due to Bandelt and Dress [1] and Warnow [43] respectively, and is a more natural setting for the study of splits

and tree metrics than the leaf-labelled tree. We obtain an **L-labelled forest** similarly be dropping the requirement that the graph is a tree and allowing it to be a forest. In other words, we have a partition α of L together with an α^i-labelled tree $T(\alpha^i)$ for each part α^i of α. We will usually require that $L = [n]$ and refer to them as n-labelled forests.

Note that $\mathsf{P}[\chi|F, p]$ is given by the product of $\mathsf{P}[(\chi|_{\alpha_F^i})|T(\alpha_F^i), p]$ over the parts α_F^i of α_F.

5.4 The reconstruction quotient

We now introduce the reconstruction quotient, our main object of interest in this chapter. Although we make all our definitions for labelled forests, our primary interest is leaf-labelled trees.

Given an n-labelled forest F, let $\mathcal{W}(F)$ be the set of edge weightings

$$z : E(F) \mapsto [0, 1]$$

with the topology induced by the Euclidean metric

$$||z - z'|| = \left(\sum_{e \in E(F)} (z_e - z'_e)^2 \right)^{\frac{1}{2}}.$$

Clearly $\mathcal{W}(F)$ is homeomorphic to $[0, 1]^{|E(F)|}$; if F is a binary tree this is $[0, 1]^{2n-3}$. Now define

$$\mathsf{P}_F : \mathcal{W}(F) \mapsto [0, 1]^{2^n} : z \mapsto (\mathsf{P}[\chi|F, z])_{\{\chi|\chi:[n]\mapsto\{\alpha,\beta\}\}},$$

that is, P_F takes z to the vector of frequencies of characters generated on F with mutation probabilities $p_e = (1 - z_e)/2$.

We extend these definitions to sets of n-labelled forests in a natural way as follows. If S is such a set, let

$$\mathcal{W}(S) = \coprod_{F \in S} \mathcal{W}(F)$$

with the disjoint union topology (so U is open in $\mathcal{W}(S)$ if and only if $U \cap \mathcal{W}(F)$ is open for all $F \in S$), and define P_S by $\mathsf{P}_S(z) = \mathsf{P}_F(z)$ if $z \in \mathcal{W}(F)$. The **reconstruction quotient** of S, $\mathcal{R}(S)$, is the space obtained from $\mathcal{W}(S)$ by identifying points with the same image under P_S,

$$\mathcal{R}(S) := \mathcal{W}(S)/\ker \mathsf{P}_S.$$

In particular, we are interested in the structure of $\mathcal{R}(T)$, where T is a binary tree, and $\mathcal{R}(\mathcal{T}_n)$, where $\mathcal{T}_n$ is the set of all binary trees on n leaves.

A second space of interest is the **LogDet quotient** $\mathcal{L}D(S)$. This is the quotient of $\mathcal{W}(S)$ obtained by identifying weighted trees that the LogDet transformation is unable to distinguish between. For the model described here, the joint probability matrices have the form

$$J(i,j) = \begin{cases} \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & \text{if } i,j \text{ lie in different parts of } \alpha_F \\ \frac{1}{2}\begin{pmatrix} 1-p_{ij} & p_{ij} \\ p_{ij} & 1-p_{ij} \end{pmatrix} & \text{if } i,j \text{ lie in the same part of } \alpha_F, \end{cases}$$

where

$$1 - 2p_{ij} = \prod_{e \in P_{ij}} (1 - 2p_e),$$

in which P_{ij} is the path from leaf i to leaf j. We then have

$$\det J(i,j) = \frac{1}{4}(1-2p_{ij}).$$

Hence, defining the **path product function** $\pi^F : \mathcal{W}(F) \mapsto [0,1]^{\binom{n}{2}}$ by

$$\pi^F_{ij}(z) = \begin{cases} 0 & \text{if } i,j \text{ lie in different parts of } \alpha_F \\ \prod_{e \in P_{ij}} z_e & \text{if } i,j \text{ lie in the same part of } \alpha_F \end{cases}$$

for each two element subset $\{i,j\} \subseteq [n]$, and extending this to sets of forests as for P_F, we have

$$\mathcal{LD}(S) = \mathcal{W}(S)/\ker \pi^S.$$

Since the entries of the joint probability matrices are sums of character frequencies, the LogDet quotient is always a quotient of the reconstruction quotient. We show here that for the two state fully symmetric model we have in fact $\mathcal{LD}(S) = \mathcal{R}(S)$.

Lemma 5.1 *For any set S of n-labelled forests we have $\mathcal{LD}(S) = \mathcal{R}(S)$ under the two state fully symmetric model.*

Proof We must show that given n-labelled forests F_1 and F_2 and weightings $z_1 \in \mathcal{W}(F_1)$, $z_2 \in \mathcal{W}(F_2)$, if $\pi^{F_1}(z_1) = \pi^{F_2}(z_2)$ then $\mathsf{P}_{F_1}(z_1) = \mathsf{P}_{F_2}(z_2)$. By contracting edges where $z_e = 1$ and deleting edges where $z_e = 0$, we obtain forests F_1', F_2' and weightings $z_1' \in \mathcal{W}(F_1')$, $z_2' \in \mathcal{W}(F_2')$ such that for $i = 1,2$, $\pi^{F_i'}(z_i') = \pi^{F_i}(z_i)$, $\mathsf{P}_{F_i'}(z_i') = \mathsf{P}_{F_i}(z_i)$, and $z'_{ie} \in \{0,1\}$ for no edge $e \in E(F_i')$.

Now $\pi^{F_i'}_{jk}(z_i') > 0$ if and only if j and k lie in the same part of $\alpha_{F_i'}$, so we must have $\alpha_{F_1'} = \alpha_{F_2'} = \alpha$. We consider each part α^m of α separately. Restricting $\pi^{F_i'}$ to the labels in α^m we obtain a multiplicative distance on α^m such that all edge weights lie in $(0,1)$. By the uniqueness theorems for tree metrics we then have $T_1'(\alpha^m)$ equal to $T_2'(\alpha^m)$ and z_1' restricted to $E(T_1'(\alpha^m))$ equal to z_2' restricted to $E(T_2'(\alpha^m))$. It follows that $\mathsf{P}_{F_1}(z_1) = \mathsf{P}_{F_2}(z_2)$ and we are done.

•

Two approaches to studying these quotient spaces readily present themselves. The first is through the images $\mathsf{P}_S(\mathcal{W}(S))$ and $\pi^S(\mathcal{W}(S))$. Since both P_S and π^S are continuous maps with compact domain, they are closed, and as maps from $\mathcal{W}(S)$ to $\mathsf{P}_S(\mathcal{W}(S))$ and $\pi^S(\mathcal{W}(S))$ respectively they are surjective. It follows that

$$\mathcal{R}(S) \cong \mathsf{P}_S(\mathcal{W}(S)) \quad \text{and} \quad \mathcal{LD}(S) \cong \pi^S(\mathcal{W}(S)).$$

This is the approach we will take here and we will study these spaces using elementary methods. However a second perhaps more fruitful approach is through the language of cellular complexes (a reference is [29]). We may introduce a cell structure for $\mathcal{W}(T)$ (and so for $\mathcal{W}(S)$) via the faces of the cube $\mathcal{W}(T)$. If $\mathcal{W}_{(E_c,E_d)}(T)$ is the face

$$\mathcal{W}_{(E_c,E_d)}(T) = \{z \in \mathcal{W}(T) | z_e = 0 \ \forall e \in E_d, z_e = 1 \ \forall e \in E_c\}$$

then $\pi^T(\mathcal{W}_{(E_c,E_d)}(T)) = \pi^F(\mathcal{W}(F))$ where F is the forest obtained from T by contracting the edges in E_c and deleting the edges in E_d. It may then be shown that $\ker \pi^S$ is a cellular equivalence relation, so we obtain a cell structure for $\mathcal{LD}(T)$. A natural object of study using this approach is the *face poset*, which turns out to be a suitable restriction of the set of n-labelled forests, partially ordered by the relation "F_1 may be obtained from F_2 by edge contractions and deletions." Figure 5.3 shows this poset for $n = 3$. The notion of *shellability* [2, section 4.7] may apply and it may be possible to deduce the topological type of these spaces.

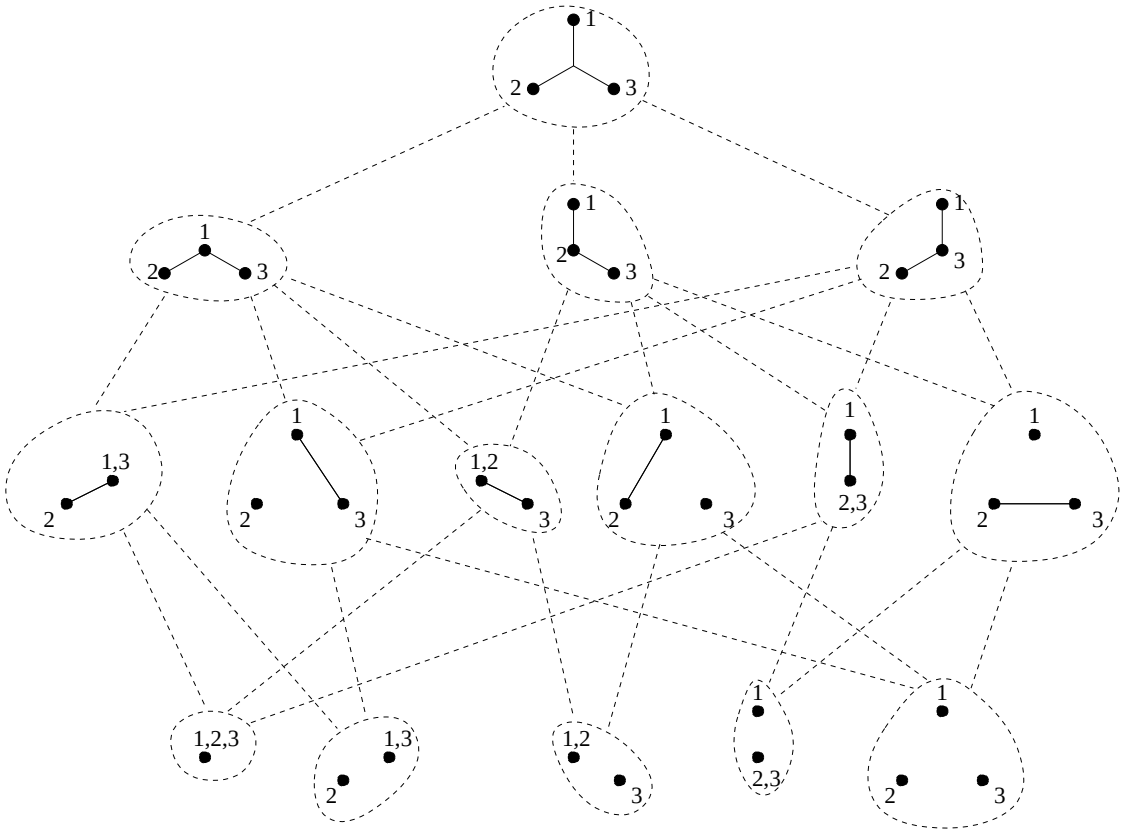


Figure 5.3: Hasse diagram of the face poset when $n = 3$. The number of edges of each forest corresponds to the dimension of the corresponding face and the partial order is given by containment. The forests at the bottom are zero dimensional cells (vertices), the forests in the next level up are one dimensional cells (edges) with faces the vertices lying beneath them, and so on. The resulting complex looks rather like three rhombuses sewn together to form a ball and then stuffed.

5.5 Two structure theorems

We now prove two structure theorems for the reconstruction quotient under the two state fully symmetric model, namely that $\mathcal{R}(T)$ is contractible for a fixed tree T, and that $\mathcal{R}(\mathcal{T}_4)$ is contractible. The map π^F is simpler than the map P_F so we study these spaces via the image under π^F.

We denote the interval $[0,1]$ by I and homotopy equivalence by $\simeq$.

5.5.1 The reconstruction quotient is contractible for a fixed tree

In this section we prove our first structure theorem for the reconstruction quotient, that $\mathcal{R}(T)$ is contractible for a fixed tree T. We show this by constructing homotopies in the space of edge weightings that are sufficiently well behaved on equivalence classes that they may be pushed down to $\pi^T(\mathcal{W}(T))$, and contract $\mathcal{R}(T)$ to a point in steps.

Given a forest F let

$$\mathcal{W}_0(F) = \{z \in \mathcal{W}(F) | z_e = 0 \text{ for some } e \in E(F)\}.$$

If $z \in \mathcal{W}(F) \setminus \mathcal{W}_0(F)$ then $(\pi^F)^{-1}(\pi^F(z)) = \{z\}$ since contracted edges may be "popped" unambiguously using knowledge of the underlying forest. Thus all non-singleton equivalence classes are contained in $\mathcal{W}_0(T)$, which will allow us to restrict our attention to forests that may be obtained from T by edge deletions. Our main tool will be the following lemma, which says that the reconstruction quotient of a forest F is homotopically equivalent to the reconstruction quotient of the set of forests that may be obtained from F by deleting a single edge.

Lemma 5.2 *For any forest F, $\pi^F(\mathcal{W}_0(F))$ is a strong deformation retract of $\pi^F(\mathcal{W}(F))$.*

Proof Let $H_F : \mathcal{W}(F) \times I \mapsto \mathcal{W}(F)$ be the projection in the $(1,1,\ldots,1)$ direction given by

$$H_F(z,t) = z - t\left(\min_{e \in E(F)} z_e\right)(1,1,\ldots,1).$$

Then H_F is a strong deformation retraction of $\mathcal{W}(F)$ onto $\mathcal{W}_0(F)$. We show that there is a map $\tilde{H}_F$ such that the diagram

$$\begin{array}{ccc} \mathcal{W}(F) \times I & \xrightarrow{H_F} & \mathcal{W}(F) \\ \downarrow{\scriptstyle \pi^F \times \mathrm{id}} & & \downarrow{\scriptstyle \pi^F} \\ \pi^F(\mathcal{W}(F)) \times I & \xrightarrow[\tilde{H}_F]{} & \pi^F(\mathcal{W}(F)) \end{array}$$

commutes, and that $\tilde{H}_F$ has the required properties.

Firstly, $\tilde{H}_F = \pi^F \circ H_F \circ (\pi^F \times \mathrm{id})^{-1}$ is well defined, since $\pi^F|_{(\mathcal{W}(F)\setminus\mathcal{W}_0(F))}$ is one to one, and $H_F(\cdot,t)|_{\mathcal{W}_0(F)}$ is the identity. Next, it is continuous, because H_F and π^F are continuous and $\pi^F \times \mathrm{id}$ is a closed map. Finally, the conditions for $\tilde{H}_F$ to be a strong deformation retraction carry over from the corresponding conditions for H_F, and we have the result.

•

Given a tree T, let $\mathcal{D}(T)$ be the set of forests that may be obtained from T by edge deletions. Each forest $F \in \mathcal{D}(T)$ is completely determined by T and the partition α_F of $[n]$. We order $\mathcal{D}(T)$ by the relation "F_1 may be obtained from F_2 by edge deletions," denoted $F_1 \leq F_2$, to obtain a

poset with a unique maximal element (the tree T) and a unique minimal element (the forest F_0 with no edges). Furthermore every maximal chain in $\mathcal{D}(T)$ has the same length $\ell = |\alpha_{F_0}| - 1$. A poset P with unique maximal and minimal elements and such that every chain has the same length is called *graded*, and the *rank* $\lambda(x)$ of the element x is the length of the subposet $\{y \in P | y \leq x\}$. In our case the rank of the forest F depends only on the number of parts of the partition α_F and is given by $\lambda(F) = |\alpha_{F_0}| - |\alpha_F|$. Figure 5.4 illustrates this poset for the four taxa tree that groups taxa 1 and 2 together.

Let $F' \prec F$ denote the relation "F' may be obtained from F by deleting a single edge" and observe that

$$\pi^F(\mathcal{W}_0(F)) = \bigcup_{F' \prec F} \pi^{F'}(\mathcal{W}(F')).$$

By Lemma 5.2, $\bigcup_{F' \prec F} \pi^{F'}(\mathcal{W}(F'))$ is a strong deformation retract of $\pi^F(\mathcal{W}(F))$ via the map $\tilde{H}_F$. Our aim is to glue these maps together for each rank to obtain strong deformation retractions of

$$\bigcup_{\substack{F \in \mathcal{D}(T) \\ \lambda(F) = i}} \pi^F(\mathcal{W}(F)) \quad \text{onto} \quad \bigcup_{\substack{F \in \mathcal{D}(T) \\ \lambda(F) = i-1}} \pi^F(\mathcal{W}(F))$$

for each i. Let F_1 and F_2 have the same rank, $F_1 \neq F_2$, and consider the intersection $\pi^{F_1}(\mathcal{W}(F_1)) \cap \pi^{F_2}(\mathcal{W}(F_2))$. If F_1 may be obtained from T by deleting edges E_1 and F_2 by deleting edges E_2, then

$$\pi^{F_1}(\mathcal{W}(F_1)) \cap \pi^{F_2}(\mathcal{W}(F_2)) = \pi^{F_3}(\mathcal{W}(F_3))$$

where F_3 is the forest obtained by deleting $E_1 \cup E_2$. In particular

$$\pi^{F_1}(\mathcal{W}(F_1)) \cap \pi^{F_2}(\mathcal{W}(F_2)) \subseteq \pi^{F_i}(\mathcal{W}_0(F_i))$$

for $i = 1, 2$. Since the maps $\tilde{H}_{F_i}(\cdot, t)|_{\pi^{F_i}(\mathcal{W}_0(F_i))}$ are the identity, $\tilde{H}_{F_1}$ and $\tilde{H}_{F_2}$ agree on the intersection of their domains, so we may use the map gluing theorem to obtain a strong deformation retract of $\pi^{F_1}(\mathcal{W}(F_1)) \cup \pi^{F_2}(\mathcal{W}(F_2))$ onto $\pi^{F_1}(\mathcal{W}_0(F_1)) \cup \pi^{F_2}(\mathcal{W}_0(F_2))$. It follows that for $1 \leq i \leq \lambda(T)$, we have

$$\bigcup_{\substack{F \in \mathcal{D}(T) \\ \lambda(F) = i}} \pi^F(\mathcal{W}(F)) \quad \simeq \quad \bigcup_{\substack{F \in \mathcal{D}(T) \\ \lambda(F) = i-1}} \pi^F(\mathcal{W}(F))$$

as desired. This gives $\pi^T(\mathcal{W}(T)) \simeq \pi^{F_0}(\mathcal{W}(F_0))$, and since $\pi^{F_0}(\mathcal{W}(F_0))$ is a singleton, it follows that $\mathcal{R}(T)$ is contractible for any tree T.

5.5.2 The four taxa reconstruction quotient is contractible

In this section we prove a second structure theorem for the reconstruction quotient, namely that $\mathcal{R}(\mathcal{T}_4)$ is contractible. The method is similar to that used in section 5.5.1: we construct suitable homotopies in the space of edge weights that may be pushed down into $\pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4))$ using $\pi^{\mathcal{T}_4}$, contracting $\pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4))$ to a point in steps.

Denote the tree that groups taxon i with taxon 1 by T_{1i} and let the edge weightings be as in figure 5.5. Let $\rho = (\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}) \in \pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4))$. From the four-point condition (2.2) we have

$$\begin{aligned}
\rho \in \pi^{\mathcal{T}_4}(\mathcal{W}(T_{12})) \quad &\text{if} \quad \rho_{12}\rho_{34} \leq \rho_{13}\rho_{24} = \rho_{14}\rho_{23} \\
\rho \in \pi^{\mathcal{T}_4}(\mathcal{W}(T_{13})) \quad &\text{if} \quad \rho_{13}\rho_{24} \leq \rho_{12}\rho_{34} = \rho_{14}\rho_{23} \\
\rho \in \pi^{\mathcal{T}_4}(\mathcal{W}(T_{14})) \quad &\text{if} \quad \rho_{14}\rho_{23} \leq \rho_{13}\rho_{24} = \rho_{12}\rho_{34}.
\end{aligned}$$

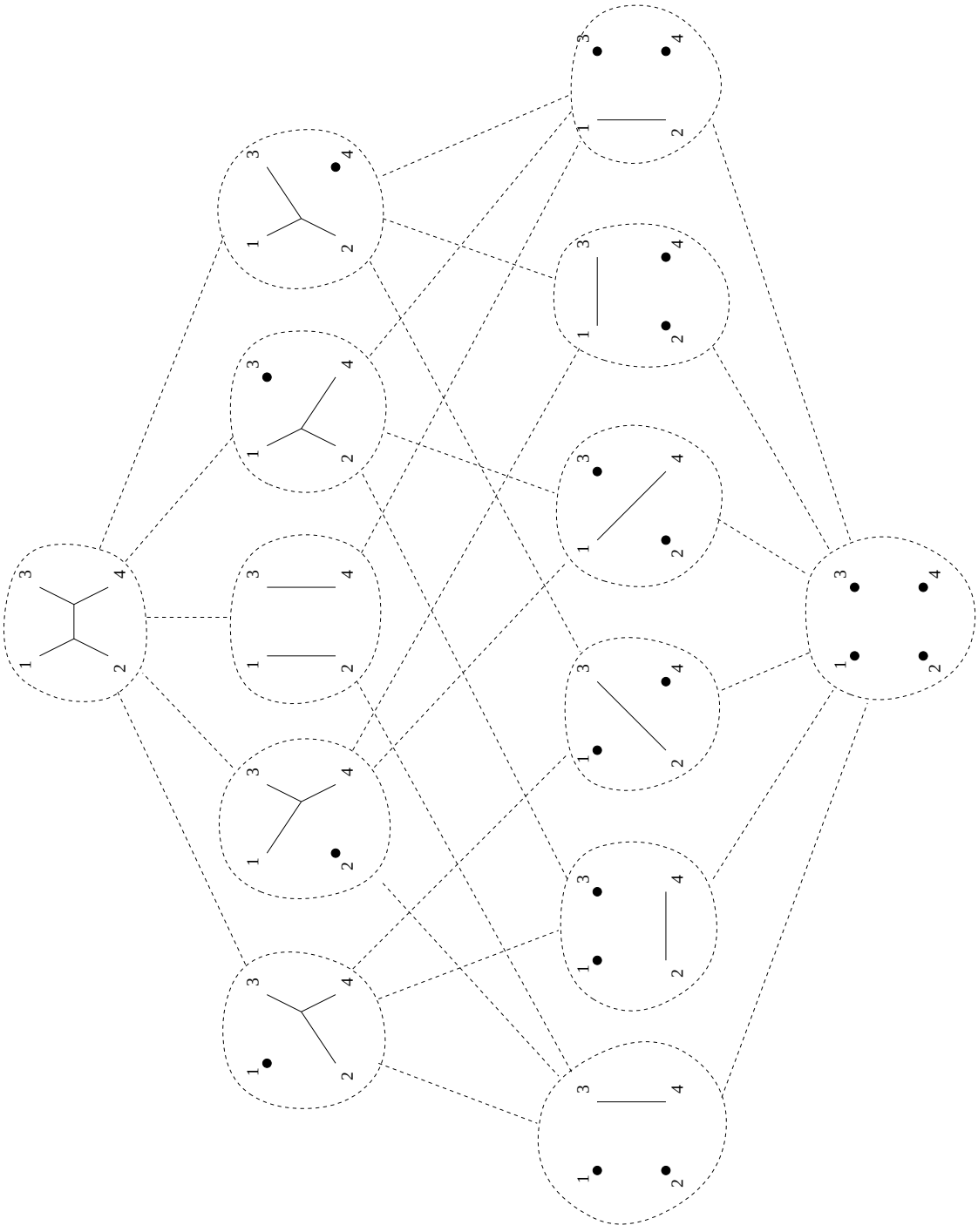


Figure 5.4: Hasse diagram of the poset of forests obtained by edge deletions from the four taxa tree that groups taxon 1 with taxon 2. The tree itself is the unique maximal element and the forest with no edges the unique minimal element. The rank of a forest is its height above the minimal element; this is well defined since every maximal chain has length three. Forests of the same rank have the same number of connected components.

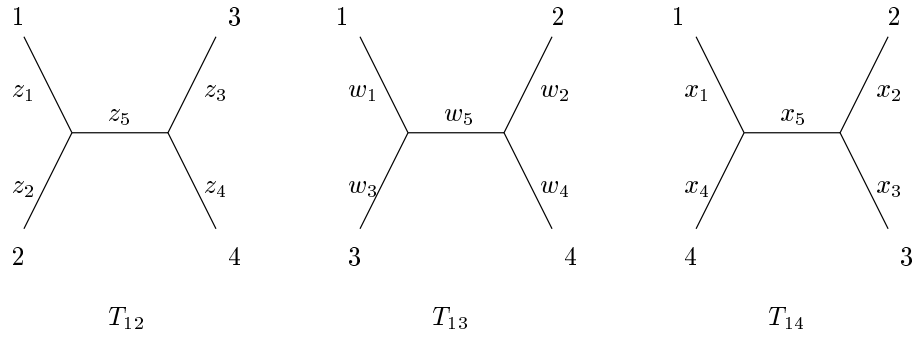


Figure 5.5: The three four taxa binary trees T_{12}, T_{13} and T_{14}, with their edge weightings.

Hence, if $i \neq j$ then

$$\begin{aligned}\pi^{\mathcal{T}_4}(\mathcal{W}(T_{1i})) \cap \pi^{\mathcal{T}_4}(\mathcal{W}(T_{1j})) &= \bigcap_{k=2}^{4} \pi^{\mathcal{T}_4}(\mathcal{W}(T_{1k})) \\ &= \left\{\rho \in \pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4)) | \rho_{12}\rho_{34} = \rho_{13}\rho_{24} = \rho_{14}\rho_{23}\right\}.\end{aligned}$$

If $z \in \mathcal{W}(T_{12})$ we have $\pi^{\mathcal{T}_4}(z) \in \bigcap_{k=2}^{4} \pi^{\mathcal{T}_4}(\mathcal{W}(T_{1k}))$ when $z_1z_2z_3z_4 = z_1z_2z_3z_4z_5$, which implies $z_5 = 1$ or $z_1z_2z_3z_4 = 0$, with similar conditions holding for $w \in \mathcal{W}(T_{13})$ and $x \in \mathcal{W}(T_{14})$. Thus for $y \in \mathcal{W}(T_{1i})$ we have $\left(\pi^{\mathcal{T}_4}\right)^{-1}\left(\pi^{\mathcal{T}_4}(y)\right) = \{y\}$ if and only if $y \notin \mathcal{W}_0(T_{1i}) \cup \{y_5 = 1\}$.

For $z \in [0,1]^5$ let $s(z)$ be the stereographic projection of z onto the set

$$A = [0,1]^5 \cap \left(\left(\bigcup_{i=1}^{5}\{z_i = 0\}\right) \cup \{z_5 = 1\}\right)$$

from the point $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2}, \frac{3}{2}, \frac{1}{2})$, and define $H : [0,1]^5 \times I \mapsto [0,1]^5$ by

$$H(z,t) = (1-t)z + t \cdot s(z).$$

Then H is a strong deformation retract of $[0,1]^5$ onto A. Viewing H as a map from $\mathcal{W}(T_{1i}) \times I$ to $\mathcal{W}(T_{1i})$ and arguing as in Lemma 5.2 we obtain a strong deformation retract H_{1i} of $\pi^{\mathcal{T}_4}(\mathcal{W}(T_{1i}))$ onto $\pi^{\mathcal{T}_4}(\mathcal{W}_0(T_{1i}) \cup \{y_5 = 1\})$. Since H_{1i} and H_{1j}, $2 \leq i, j \leq 4$, agree on the intersection of their domains we may use the map gluing theorem to obtain a strong deformation retract $\tilde{H}$ of $\pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4))$ onto $\pi^{\mathcal{T}_4}\left(\left(\bigcup_{i=2}^{4}\mathcal{W}_0(T_{1i})\right) \cup \mathcal{W}(T_{1234})\right)$, where T_{1234} is the star tree on four taxa.

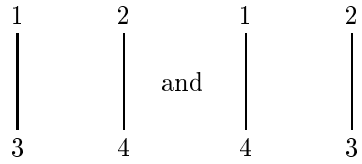
From Lemma 5.2, $\pi^{\mathcal{T}_4}(\mathcal{W}_0(T_{1234})) \subseteq \pi^{\mathcal{T}_4}\left(\bigcup_{i=2}^{4}\mathcal{W}_0(T_{1i})\right)$ is a strong deformation retract of $\pi^{\mathcal{T}_4}(\mathcal{W}(T_{1234}))$, and the map $\tilde{H}_{T_{1234}}$ agrees with the identity map on $\pi^{\mathcal{T}_4}\left(\bigcup_{i=2}^{4}\mathcal{W}_0(T_{1i})\right)$ on the intersection of their domains so that using the map gluing theorem we obtain

$$\pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4)) \simeq \pi^{\mathcal{T}_4}\left(\bigcup_{i=2}^{4}\mathcal{W}_0(T_{1i})\right).$$

Now $\pi^{\mathcal{T}_4}\left(\bigcup_{i=2}^{4}\mathcal{W}_0(T_{1i})\right)$ is simply $\pi^{\mathcal{T}_4}(\mathcal{W}_0(T_{12}))$ with the addition of the sets

$$\pi^{\mathcal{T}_4}\left(\{w \in \mathcal{W}(T_{13}) | w_5 = 0\}\right) \quad \text{and} \quad \pi^{\mathcal{T}_4}\left(\{x \in \mathcal{W}(T_{14}) | x_5 = 0\}\right),$$

which are the images under $\pi^{\mathcal{T}_4}$ of the spaces of edge weights of the forests



(5.4)

respectively. If P is the poset of trees obtained from $\mathcal{D}(T_{12})$ (the poset in figure 5.4) by adding the trees in (5.4) and removing T_{12}, then we may argue as in section 5.5.1 to show that

$$\bigcup_{\substack{F \in P \\ \lambda(F)=i}} \pi^F(\mathcal{W}(F)) \simeq \bigcup_{\substack{F \in P \\ \lambda(F)=i-1}} \pi^F(\mathcal{W}(F))$$

for $i = 1, 2$. We then have $\pi^{\mathcal{T}_4}(\mathcal{W}(\mathcal{T}_4)) \simeq \pi^{F_0^4}(\mathcal{W}(F_0^4))$, where F_0^4 is the four taxa forest with no edges, so that $\mathcal{R}(\mathcal{T}_4)$ is homotopic to a point.

5.6 Discussion

We have proved two structure theorems for the reconstruction quotient, namely that $\mathcal{R}(T)$ is contractible for any tree T and that $\mathcal{R}(\mathcal{T}_4)$ is contractible, results showing that these spaces are in some sense "simple". We leave as open problems further study of these and related spaces. Some questions of interest are determining whether the methods used here can be extended to $\mathcal{R}(\mathcal{T}_n)$ when $n \geq 5$, and the determination of topological type. $\mathcal{R}(\mathcal{T}_3)$ appears to be homeomorphic to a ball, but the structure of $\mathcal{R}(T)$ is less clear when T is a tree on n leaves where $n \geq 4$.

If these questions are successfully answered, further work could look at the geometry of these spaces in more detail or examine quotients corresponding to more complicated substitution models.

Chapter 6

More realistic models: the covarion hypothesis and rates-across-sites

6.1 Introduction

The basic model assumes that each site evolves i.i.d. at the same rate, according to the simple Markov-style assumptions of equation (3.1). However, this single-rate assumption appears to be unrealistic: some sites appear to change very slowly or not at all, while others appear to evolve very rapidly. This is thought to be due to differing functional or structural constraints at each site. For example, the state at a particular site may be so critical to the survival of the organism it codes for that any change there is lethal. A change at such a site would accordingly leave no record. On the other hand, a site that codes for something less important (or even for nothing at all) would be able to change freely. Accordingly, models incorporating some variation of rates across sites have been proposed and studied (see for example [7, 24, 37, 44]) to try and take this into account.

An alternative approach to accounting for differing selective constraints is Fitch and Markowitz's "concomitantly variable codons" or "covarion" hypothesis [14]. This proposes that at any given time, some sites are invariable due to functional or structural constraints, but that as mutations are fixed elsewhere in the sequence these constraints may change, so that sites that were previously invariable may become variable and vice versa. The pool of variable sites is therefore changing with time (see figure 6.1). Since its proposal 27 years ago it has been argued that evidence supports the covarion hypothesis, both on biochemical grounds, and by providing a better description of certain data [12, 13, 30]. However, in contrast to the rates-across-sites models, little is known about the analytic properties of covarion-style models.

In this chapter we present and analyse a simple covarion-style model. Although the motivation for this model clearly says that the i.i.d. assumption is not valid, without it the mathematics becomes much more difficult. We therefore keep this assumption and model only the behaviour of a covarion-style process, with a two state Markov process that acts as a "switch", turning sites "on" (variable) and "off" (invariable). We do not impose any restrictions on the Markov process that operates at the variable sites other than that it is stationary and reversible. Using techniques from the theory of Markov processes such a model may be analysed and compared with rates-across-sites models in terms of the expected frequencies of characters the models should generate. This is the first step in comparing the two models, since if they cannot be distinguished between using infinite sequences there is no prospect of distinguishing between them with finite sequences.

The i.i.d. assumption may be justified as an approximation to the covarion hypothesis by the

following remarks. We are concerned here with the limiting frequencies of characters in sequences as the sequence length becomes large and without reference to the order in which the characters occur along the sequence. If the dependency between sites is spatially localised (perhaps under some reordering of the sites) then the frequencies of the characters will converge towards those generated under an i.i.d. model. This follows from an argument similar to the proof of Bernstein's theorem (see for example Rényi [34, page 379]) which requires only that the correlation between the sites, re-ordered if necessary, falls off sufficiently quickly. In our setting the assumption of local dependency between sites is reasonable. This type of approach is already commonly employed (albeit tacitly) when modelling a distribution of rates-across-sites. In real sequences, high rates are often associated with particular positions in the sequence (such as the third position in a codon, since it can frequently change without changing the amino acid coded for) or proximity to other high rate sites (so-called hypervariable regions), so the sites are clearly neither independent nor identically distributed. Nevertheless, since the dependency is local, it is usual to suppose that the rate at each site is chosen i.i.d. from some distribution, and the resulting i.i.d. model produces indistinguishable character frequencies to the original model as the sequence length tends to infinity.

In section 6.3.1 we find an expression for the joint probability matrix of states for two species separated by an evolutionary distance τ. This allows τ to be determined from the expected proportion of sites where two species differ and so gives a tree-additive distance. We then compare this with the equivalent expression under a rates-across-sites model in section 6.3.4, in order to address the question of whether the covarion model will give different results to rates-across-sites when several sequences are analysed by comparing each pair in turn. We show that a covarion model gives identical results to a suitably chosen rates-across-sites model if only the trace of the joint probability matrix (that is, the probability the two species are in the same state at a given site) is considered, and give a partial answer to the question of when covarion and rates-across-sites models can give identical results if the full joint probability matrix is considered.

In section 6.4 we show that that the two models can, in principle, be distinguished when there are at least four monophyletic groups of species. This result if based on the construction of a distance which is tree-additive under certain versions of the covarion model but which, in general, will not be additive under a rates-across-sites model. The distance constructed does not require knowledge of the parameters of the model and so shows that sequences generated by the covarion model do in fact contain information about the structure of the underlying tree.

A joint paper with Mike Steel [40] based on the work in this chapter was submitted to *Mathematical Biosciences* in October 1996 and has been reviewed. This chapter is based on the revised manuscript which was resubmitted in June 1997.

6.2 The models

6.2.1 A covarion-style model

We model a covarion-style process with two parts: a "switch" process, and an "observable" process, which operates while the switch is "on". Only the state of the observable process, and not that of the switch process, is able to be measured.

The switch is governed by a two state continuous time Markov process with state space $\mathcal{O} = \{\texttt{on}, \texttt{off}\}$ and rate matrix

$$S = \begin{pmatrix} -s_1 & s_1 \\ s_2 & -s_2 \end{pmatrix}$$

where $s_i > 0$ for each i. It is assumed to have the stationary initial distribution $\sigma = (\sigma_1, \sigma_2)$ where

$$\sigma_1 = \frac{s_2}{s_1 + s_2}, \quad \sigma_2 = \frac{s_1}{s_1 + s_2},$$

so that it is stationary and time-reversible.

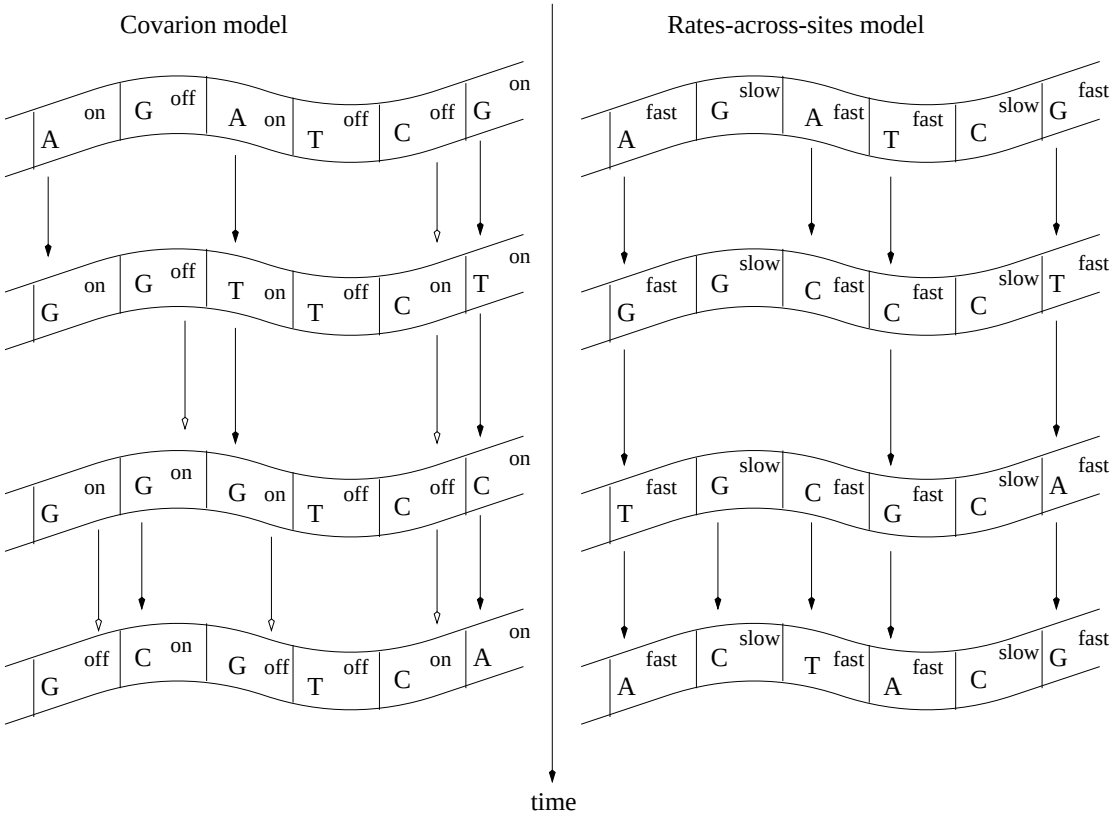


Figure 6.1: Contrasting a covarion style process and rates-across-sites. Under a covarion style process, each site is either "on" or "off". Sites that are off are unable to change state, but may later turn on (due to state changes elsewhere in the sequence) and be able to change. Under rates-across-sites, sites evolve at different rates (shown here as "fast" and "slow"), with faster sites changing more frequently than slower ones. The rate at a given site is assumed constant across the entire tree.

While the switch is in state `off`, the observable process is unable to change state; however, when the switch is in state `on`, the observable process is governed by a second stationary and time reversible Markov process with state space $\mathcal{A} = [r]$, rate matrix R satisfying $R_{ij} > 0$ if $i \neq j$, and initial distribution π. This assumption that $R_{ij} > 0$ if $i \neq j$ allows us to apply the Perron-Frobenius Theorem ([21, page 125] or [16, page 134]) to $I + \frac{1}{k}R$, where $k > \max\{|R_{ii}|\}$, to conclude that (i) 0 is an eigenvalue of R of multiplicity one, (ii) the remaining eigenvalues of R are negative, and (iii) π has positive entries. We write $C = (R, S)$ for the covarion model C with observable process rate matrix R and switch process rate matrix S.

This model may be alternatively formulated in terms of a single time-reversible Markov process with state space $\mathcal{A} \times \mathcal{O}$ (which we identify with $[2r]$ according to $(i, \texttt{on}) \mapsto i$, $(i, \texttt{off}) \mapsto i + r$), initial distribution $\pi' = (\sigma_1\pi_1, \ldots, \sigma_1\pi_r, \sigma_2\pi_1, \ldots, \sigma_2\pi_r)$ and $2r \times 2r$ rate matrix

$$R' = \begin{pmatrix} R - s_1 I_r & s_1 I_r \\ s_2 I_r & -s_2 I_r \end{pmatrix},$$

where I_r denotes the $r \times r$ identity matrix. We assume that we are unable to distinguish between the states $(i, \texttt{on})$ and $(i, \texttt{off})$. The probability of generating a given character is more easily calculated using this formulation than the first. As usual, if each edge e of the tree is given a non-negative

weight τ_e, the transition matrices P^e are given by

$$P^e = \exp(\tau_e R').$$

The probability of generating a particular character is given by a sum over all possible assignments of states in $\mathcal{A} \times \mathcal{O}$ to the remaining vertices of the tree. In practice this can be found quickly using a simple modification of the usual dynamic programming technique.

It is easily checked that R' is stationary and time-reversible whenever R and S are. Further, both formulations lead to the same random process with state space $\mathcal{A}$.

6.2.2 Rates-across-sites

A *rates-across-sites* model $D = (Q, \mathcal{D})$ consists of a stationary and time-reversible continuous time Markov process with rate matrix Q and initial distribution θ, and a distribution $\mathcal{D}$ of rates ν, which may be either discrete or continuous. We denote the cumulative distribution function of $\mathcal{D}$ by $F_{\mathcal{D}}$.

Each site evolves according to rate matrix νQ where ν is chosen i.i.d. according to $\mathcal{D}$. The rate at a given site is assumed constant across the whole tree. This kind of model has been well studied, see for example [7, 24, 37, 44].

6.2.3 Lumpability

The second formulation of the covarion model above shows that we may regard it as a Markov process $\mathcal{F}(t)$ with state space $\mathcal{A} \times \mathcal{O}$ in which we are unable to distinguish between the states $(i, \texttt{on})$ and $(i, \texttt{off})$. Calculations involving this model would be much simpler if it could be shown that the resulting random process on $\mathcal{A}$ was also Markov. A Markov process $X(t)$ with state space $\mathcal{B}$ for which there is a partition $\mathcal{B} = B_1 \cup \ldots \cup B_k$ such that the random process $Y(t) = B_i$ if $X(t) \in B_i$ is Markov is said to be *lumpable* with respect to the partition $\{B_i\}$.

Consider a stationary and reversible Markov chain $X(n)$ with transition matrix P and initial distribution p such that $p_i > 0$ for all i. Burke and Rosenblatt [4, Theorem 1] give the following necessary and sufficient condition for $X(n)$ to be lumpable: $Y(n)$ is Markovian if and only if for any fixed $\beta = 1, \ldots, k$,

$$\sum_{j \in B_\beta} P_{ij} = \mathsf{P}[X(n+1) \in B_\beta | X(n) = i] = C_{B_\alpha, B_\beta}$$

has the same value for all i in any given collapsed set of states B_α, $\alpha = 1, \ldots, k$. Applying this to the induced Markov chain $\mathcal{F}_\tau(n) = \mathcal{F}(n\tau)$ for each $\tau > 0$, lumpability with respect to the partition given by $B_i = \{(i, \texttt{on}), (i, \texttt{off})\} = \{i, i+r\}$ would imply

$$P_{ij}(\tau) + P_{i,j+r}(\tau) = P_{i+r,j}(\tau) + P_{i+r,j+r}(\tau)$$

for $1 \leq i, j \leq r$ and $\tau > 0$. Differentiating at 0 when $i \neq j$ gives

$$R'_{ij} + R'_{i,j+r} = R'_{i+r,j} + R'_{i+r,j+r},$$

and since $R'_{i,j+r} = R'_{i+r,j} = R'_{i+r,j+r} = 0$, we get $R'_{ij} = 0$ also. Thus, $\mathcal{F}(t)$ is not lumpable with respect to this partition, and we cannot analyse the covarion model by simply treating it as a Markov process on $\mathcal{A}$.

6.3 The two taxa tree

Here we calculate the joint probability matrix for the two taxa tree (that is, the matrix whose ij entry is the probability that taxon 1 is in state i and taxon 2 is in state j), and give conditions

under which a suitably chosen rates-across-sites model will agree with a covarion model on all two taxa trees. We also consider the limiting cases of the covarion model as the rate of the switch tends either to zero or to infinity.

6.3.1 Under the covarion model

The joint probability matrix may be calculated using either of the two formulations of the covarion model. We present the calculation via the first formulation. Using the second formulation the ij entry of this matrix is found by summing the probability that taxon 1 is in state $(i, \mathtt{o}_1)$ and taxon 2 is in state $(j, \mathtt{o}_2)$ for $\mathtt{o}_i = \mathtt{on}$, $\mathtt{o}_i = \mathtt{off}$, $i = 1, 2$.

Time reversibility implies we may assume the tree is rooted at either of the leaves. Let the process operate for time τ on the edge between the two taxa and write $J_C(\tau)$ for the joint probability matrix. We regard τ as the "length" of the edge. Put $\Pi = \mathrm{diag}(\pi)$ and let $J(t)$ be the joint probability matrix of the unswitched observable process (that is, the Markov process with rate matrix R and initial distribution π operating in the absence of the switch) for time t. If the occupation time of state $\mathtt{on}$ in time τ is the random variable $X(\tau)$, then, as far as the observable process is concerned, the edge has effective length $X(\tau)$. The joint probability matrix, given the value of $X(\tau)$, is then $J(X(\tau))$. It follows that

$$\begin{aligned} J_C(\tau) &= \mathsf{E}[J(X(\tau))] \\ &= \mathsf{E}\left[\sum_{j=1}^{r} e^{\lambda_j X(\tau)} w_j w_j^{\mathrm{T}}\right] \\ &= \sum_{j=1}^{r} \mathsf{E}[e^{\lambda_j X(\tau)}] w_j w_j^{\mathrm{T}}, \end{aligned}$$

where we have used the spectral representation from equation (2.5). From Darroch and Morris [9] the moment generating function $\mathsf{E}[e^{\lambda X(\tau)}]$ of $X(\tau)$ is given by

$$\mathsf{E}[e^{\lambda X(\tau)}] = \sigma^{\mathrm{T}} e^{\tau(S+\lambda D)} \mathbf{1}, \tag{6.1}$$

where $D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Diagonalising $S + \lambda D$ we obtain the following:

Lemma 6.1 *The joint probability matrix $J_C(\tau)$ is given by*

$$J_C(\tau) = \sum_{j=1}^{r} [c_j^+ e^{\mu_j^+ \tau} + c_j^- e^{\mu_j^- \tau}] w_j w_j^{\mathrm{T}}, \tag{6.2}$$

where μ_j^+ and μ_j^- are the positive and negative roots respectively of

$$\mu^2 + (s_1 + s_2 - \lambda_j)\mu - s_2\lambda_j = 0,$$

and

$$c_j^+ = \frac{-(s_1 + s_2 + \mu_j^+)\mu_j^-}{(s_1 + s_2)(\mu_j^+ - \mu_j^-)} \quad \textit{and} \quad c_j^- = \frac{(s_1 + s_2 + \mu_j^-)\mu_j^+}{(s_1 + s_2)(\mu_j^+ - \mu_j^-)} \quad .$$

We note that, as might be expected, the eigenvalues of R' are $\{\mu_j^{\pm} | j \in [r]\}$.

A common measure of the extent to which two sequences differ is the proportion of sites at which they disagree, known as the *dissimilarity*. The expected proportion of such sites is given by one minus the trace of the joint probability matrix. From equation (6.2) we obtain

$$\mathrm{trace}(J_C(\tau)) = \sum_{j=1}^{r} [c_j^+ e^{\mu_j^+ \tau} + c_j^- e^{\mu_j^- \tau}] \mathrm{trace}\left(w_j w_j^{\mathrm{T}}\right).$$

For the zero eigenvalue $\lambda_1 = 0$ we have $\mu_1^+ = 0$, $\mu_1^- = -(s_1 + s_2)$ and $w_1 = \pi^{\mathrm{T}}$, so that we have

Lemma 6.2 *The probability that two sequences have the same state at a given site is given by*

$$trace(J_C(\tau)) = \pi\pi^{\mathrm{T}} + \sum_{j=2}^{r}[c_j^+ e^{\mu_j^+ \tau} + c_j^- e^{\mu_j^- \tau}]trace\left(w_j w_j^{\mathrm{T}}\right). \tag{6.3}$$

In order to proceed any further with this calculation we need to be able to calculate $\mathrm{trace}(w_j w_j^{\mathrm{T}}) = \mathrm{trace}(\Pi u_j u_j^{\mathrm{T}})$ for the remaining eigenvalues, which requires some knowledge of R. However, in the case of the equi-frequent stationary distribution $\pi = (1/r, \ldots, 1/r)$ we have simply $\mathrm{trace}(w_j w_j^{\mathrm{T}}) = \mathrm{trace}(\frac{1}{r} u_j u_j^{\mathrm{T}}) = 1/r$, so that

$$\mathrm{trace}(J_C(\tau)) = \frac{1}{r} + \frac{1}{r}\sum_{j=2}^{r}[c_j^+ e^{\mu_j^+ \tau} + c_j^- e^{\mu_j^- \tau}].$$

We conclude this section by establishing some properties of the coefficients in expression (6.2) that are helpful in determining the behaviour of the covarion model.

Lemma 6.3

(i) *μ^+ and μ^- are real increasing functions of λ satisfying $\mu^- \leq -(s_1 + s_2) < -s_2 < \mu^+ \leq 0$ on $(-\infty, 0]$.*

(ii) *$c_j^+, c_j^- \geq 0$ (with equality only for $\lambda = 0$, when $c^- = 0$) and $c_j^+ + c_j^- = 1$.*

(iii) *$trace(w_j w_j^{\mathrm{T}}) > 0$ and $\sum_{j=1}^{r} trace(w_j w_j^{\mathrm{T}}) = 1$.*

Proof (i) Suppose $\lambda_1 < \lambda_2$ and consider the functions

$$f_{\lambda_j}(\mu) = \mu^2 + (s_1 + s_2 - \lambda_j)\mu - s_2\lambda_j.$$

If $f_{\lambda_1}(\mu) = f_{\lambda_2}(\mu)$ then we find $\mu = -s_2$, at which point $f_{\lambda_j}(-s_2) = -s_1 s_2$ and $f'_{\lambda_j}(-s_2) = s_1 - s_2 - \lambda_j$. Thus the situation is as illustrated in figure 6.2, and since μ_j^+ and μ_j^- are the roots of $f_{\lambda_j} = 0$ it follows that $\mu_1^- < \mu_2^-$ and $\mu_1^+ < \mu_2^+$. For the inequalities, we have $\mu^- = -(s_1 + s_2), \mu^+ = 0$ when $\lambda = 0$, and $f_{\lambda_j}(-s_2) = -s_1 s_2 < 0$ so $\mu^- < -s_2 < \mu^+$, since the f_{λ_j} are right-way up parabolas.

The inequalities in (ii) follow from (i), and the equality $c_j^+ + c_j^- = 1$ may be verified directly. For (iii) we have $\mathrm{trace}(w_j w_j^{\mathrm{T}}) = w_j^{\mathrm{T}} w_j = |w_j|^2 > 0$ and

$$\sum_{j=1}^{r} \mathrm{trace}(w_j w_j^{\mathrm{T}}) = \mathrm{trace}\sum_{j=1}^{r} w_j w_j^{\mathrm{T}} = \mathrm{trace}(J(0)) = \mathrm{trace}(\Pi) = 1.$$

•

6.3.2 Under rates-across-sites

In the rates-across-sites case, put $\Theta = \mathrm{diag}(\theta)$ and let Q have eigenvalues $\{\alpha_j\}$. Arguing as for the covarion model, if $\Theta^{1/2} Q \Theta^{-1/2}$ has orthonormal eigenvectors $\{y_j\}$, then the joint probability matrix $J_D(\tau)$ of the rates-across-sites model D is given by

$$J_D(\tau) = \sum_{j=1}^{r} \mathsf{E}[e^{\alpha_j \nu \tau}] z_j z_j^{\mathrm{T}},$$

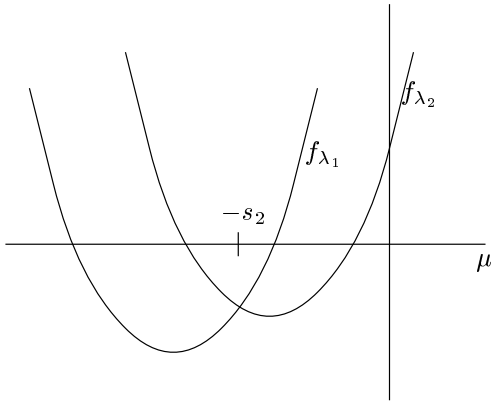


Figure 6.2: f_{λ_1} lies above f_{λ_2} on $(-s_2, \infty)$ and below on $(-\infty, -s_2)$, with $f_{\lambda_1}(-s_2) = f_{\lambda_2}(-s_2) < 0$. Hence $\mu_1^- < \mu_2^- < \mu_1^+ < \mu_2^+$.

where $z_j = \Theta^{1/2} y_j$. We may write this as

$$J_D(\tau) = \sum_{j=1}^{r} M(\alpha_j \tau) z_j z_j^{\mathrm{T}} \tag{6.4}$$

where $M(x) = \mathsf{E}[e^{\nu x}]$ is the moment generating function of $\mathcal{D}$, given by the Lebesgue-Stieltjes integral

$$M(x) = \int_0^\infty e^{\nu x} dF_{\mathcal{D}}(\nu).$$

If $F_{\mathcal{D}}$ has a continuous derivative $f_{\mathcal{D}}$ (its probability density function) this is simply

$$M(x) = \int_0^\infty e^{\nu x} f_{\mathcal{D}}(\nu) d\nu,$$

while if $\mathcal{D}$ has only finitely many rates $\nu_1, \ldots, \nu_k$ and $\mathsf{P}[\nu = \nu_i] = p_i$ we have

$$M(x) = \sum_{i=1}^{k} p_i e^{\nu_i x}.$$

As in the covarion case the probability the two sequences have the same state at a given site is given by

$$\mathrm{trace}(J_{\mathcal{D}}(\tau)) = \theta\theta^{\mathrm{T}} + \sum_{j=2}^{r} M(\alpha_j \tau) \mathrm{trace}(z_j z_j^{\mathrm{T}}); \tag{6.5}$$

in the equi-frequent stationary distribution case $\theta = (1/r, \ldots, 1/r)$ this is

$$\mathrm{trace}(J_{\mathcal{D}}(\tau)) = \frac{1}{r} + \frac{1}{r} \sum_{j=2}^{r} M(\alpha_j \tau).$$

These calculations of are not new and similar or equivalent calculations appear in various papers dealing with rates-across-sites models, such as [17, 24, 42].

6.3.3 Recovering the evolutionary distance under the two models

Equation (6.4) may be written

$$J_D(\tau) = \Theta M(\tau Q), \tag{6.6}$$

where M is the moment generating function of $\mathcal{D}$ applied to matrices. This expression has the advantage of enabling us to calculate the expected number of substitutions K between the two taxa without requiring knowledge of Q, via

$$K = -\text{trace}\left\{\Theta\left[M^{-1}\left(\Theta^{-1}J_D(\tau)\right)\right]\right\} \tag{6.7}$$

[17, 42]. Here M^{-1} is the inverse of the moment generating function, again applied to matrices. This expression gives a tree-additive distance, and since row i of $J_D(\tau)$ sums to θ_i, requires knowledge only of $\mathcal{D}$ to reconstruct the tree from $J_D(\tau)$.

If both Q and $\mathcal{D}$ are known we may express K in terms of just the trace of $J_D(\tau)$ as

$$K = -\text{trace}(\Theta Q) f_D^{-1}(\text{trace}(J_D(\tau)), \tag{6.8}$$

where $f_D(\tau) = \text{trace}(J_D(\tau))$ is given by equation (6.5). Note that f_D^{-1} exists since f_D is monotone decreasing.

The property of (6.4) that allows it to be written in the form (6.6) (namely, M is applied to products of the form $\alpha_j\tau$) does not hold for (6.2), and it appears that a transformation analogous to (6.7) does not exist for the covarion model. However, if R and S are known (or estimated) then, as in (6.8), we may express K in terms of $\text{trace}(J_C(\tau))$ as

$$K = -\text{trace}(\Pi R)\sigma_1 f_C^{-1}(\text{trace}(J_C(\tau)),$$

where $f_C(\tau) = \text{trace}(J_C(\tau))$ is given by equation (6.3). Again f_C is monotone decreasing (by Lemma 6.3) so f_C^{-1} exists.

Note that in applications, the joint probability matrix (J_C or J_D) is estimated from the observed joint frequency matrix $\hat{J}$. Since J_C and J_D are both symmetric, it is usual practice to take the symmetrised matrix $(\hat{J} + \hat{J}^{\text{T}})/2$ as the estimate.

6.3.4 Pairwise comparisons of sequences

Simultaneous pairwise comparisons of several sequences are frequently used as a method of building trees, for example through the construction of tree-additive distances. Here we address the issue of whether the covarion model will give different results to rates-across-sites when making such comparisons. For a fixed $\tau = \tau_1$, if the rates are distributed according to the distribution of $X(\tau_1)/\tau_1$ then we have $J_C(\tau_1) = J_D(\tau_1)$ for $C = (R, S)$ and $D = (R, \mathcal{D})$, so that the covarion model gives identical results to a suitably chosen rates-across-sites model if only one pair of sequences is examined. However, the distribution of $X(\tau)/\tau$ depends on τ which opens the possibility that the models may give different results if more than one pair is considered.

A common measure of the dissimilarity of two sequences is one minus the trace of the joint probability matrix, which is the probability that they disagree at a given site. In applications this is estimated from the proportion of sites at which the aligned sequences from the two taxa differ. We show here (Theorem 6.1) that given any covarion model, there is always a rates-across-sites model that will generate exactly the same data if only the trace is considered. We also characterise the conditions (Theorem 6.2) under which $C = (R, S)$ and $D = (Q, \mathcal{D})$ satisfy $J_C(\tau) = J_D(\tau)$ for all τ. Models satisfying this equality will give identical results under any form of pairwise comparison and on any tree; however, models that do not satisfy this equality may still give identical results on certain trees.

A related question is whether it is possible to distinguish between the covarion model and rates-across-sites on the basis of simultaneous pairwise comparisons of several sequences. On this

question the results of this section are largely negative and suggest that pairwise comparisons are inadequate to distinguish the covarion model from rates-across-sites. Thus, a test of the two models will probably require the simultaneous comparison of three or more sequences. Section 6.4 gives an alternative approach to distinguishing between the covarion and rates-across-sites models based on such a comparison.

We begin with a preliminary result. Stationary and reversible rate matrices with exactly one distinct non-zero eigenvalue will be of relevance to us in what follows, so we give a characterisation of them here.

Lemma 6.4

(i) *Given a distribution π of states there is a stationary and reversible rate matrix R_π having π as its stationary distribution and possessing exactly one distinct non-zero eigenvalue, namely*

$$R_\pi = \mathbf{1}\pi - I_r,$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$.

(ii) *If the stationary and reversible rate matrix R with stationary distribution π and $|R_{ij}| > 0$ for all i, j has exactly one distinct non-zero eigenvalue $-\lambda$, then $R = \lambda R_\pi$.*

Proof (i) We have $(\mathbf{1}\pi - I_r)\mathbf{1} = \mathbf{1}(\pi\mathbf{1}) - \mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}$, so the rows of R_π sum to zero. All the off-diagonal entries are positive and therefore R_π is a rate matrix. $\pi R_\pi = \pi(\mathbf{1}\pi - I_r) = (\pi\mathbf{1})\pi - \pi = \pi - \pi = \mathbf{0}^{\mathrm{T}}$, so R_π has stationary distribution π, and if $i \neq j$ then $(R_\pi)_{ij} = \pi_j$, so $\pi_i(R_\pi)_{ij} = \pi_i\pi_j = \pi_j(R_\pi)_{ji}$. Hence R_π is reversible.

The matrix $\mathbf{1}\pi$ has rank 1 and hence null space of dimension $r - 1$, so 0 is an eigenvalue of multiplicity $r - 1$. The remaining eigenvalue is 1 since $(\mathbf{1}\pi)\mathbf{1} = \mathbf{1}$. Hence R_π has eigenvalues -1 (multiplicity $r - 1$) and 0 (multiplicity 1).

(ii) Consider the matrix $Q = I_r + \frac{1}{\lambda}R$, which has eigenvalues 0 (multiplicity $r - 1$) and 1 (multiplicity 1). By the reversibility assumption R has a full complement of eigenvectors, so Q has null space of dimension $r - 1$ and hence rank 1. Further Q has row sums equal to 1, from which $Q = \mathbf{1}v$ where $\sum_{i=1}^{r} v_i = 1$. In fact we must have $v = \pi$, since the left eigenvector corresponding to 1 is π. Hence $R = \lambda(Q - I_r) = \lambda(\mathbf{1}\pi - I_r) = \lambda R_\pi$.

•

Theorem 6.1 *For any covarion model C there is a rates-across-sites model D such that*

$$\mathit{trace}(J_D(\tau)) = \mathit{trace}(J_C(\tau))$$

for all $\tau \geq 0$.

Proof By (6.3) and Lemma 6.3, $\mathrm{trace}(J_C(\tau))$ has the form

$$\mathrm{trace}(J_C(\tau)) = \pi\pi^{\mathrm{T}} + \sum_{j=2}^{r}[c'^{+}_j e^{\mu^{+}_j \tau} + c'^{-}_j e^{\mu^{-}_j \tau}]$$

where $c'^{+}_j, c'^{-}_j > 0$ and $\sum_{j=2}^{r}[c'^{+}_j + c'^{-}_j] = 1 - \pi\pi^{\mathrm{T}}$. If R has k distinct non-zero eigenvalues we may collect terms in $e^{\mu^{\pm}\tau}$ for each eigenvalue, writing $\mathrm{trace}(J_C(\tau))$ in the form

$$\mathrm{trace}(J_C(\tau)) = a_0 + \sum_{i=1}^{2k} a_i e^{-\nu_i \tau},$$

where $a_i, \nu_i > 0$ for each i and $\sum_{i=0}^{2k} a_i = 1$.

Let $\mathcal{D}$ be the discrete distribution of rates such that

$$\mathsf{P}[\nu = \nu_i] = \frac{a_i}{1 - a_0} \qquad i = 1, \ldots, 2k.$$

Then $\mathcal{D}$ is well-defined, and if $D = (R_\pi, \mathcal{D})$ then by (6.5) and Lemmas 6.3 and 6.4,

$$\begin{aligned}
\text{trace}(J_D(\tau)) &= \pi\pi^{\mathrm{T}} + \sum_{j=2}^{r} M(-\tau)\text{trace}(z_j z_j^{\mathrm{T}}) \\
&= \pi\pi^{\mathrm{T}} + M(-\tau)(1 - \pi\pi^{\mathrm{T}}) \\
&= a_0 + (1 - a_0)\sum_{i=1}^{2k} \frac{a_i}{1 - a_0} e^{-\nu_i \tau} \\
&= a_0 + \sum_{i=1}^{2k} a_i e^{-\nu_i \tau} \\
&= \text{trace}(J_C(\tau)).
\end{aligned}$$

•

Theorem 6.2

(i) *For a given covarion model $C = (R, S)$, there is a rates-across-sites model $D = (Q, \mathcal{D})$ such that*

$$J_C(\tau) = J_D(\tau)$$

for all $\tau \geq 0$ if and only if R has only one distinct non-zero eigenvalue, in which case $\mathcal{D}$ is a discrete two rate distribution and Q is a scalar multiple of R.

(ii) *For a given rates-across-sites model $D = (Q, \mathcal{D})$, there is a covarion model $C = (R, S)$ such that*

$$J_D(\tau) = J_C(\tau)$$

for all $\tau \geq 0$ if and only if Q has only one distinct non-zero eigenvalue and $\mathcal{D}$ is a discrete two rate distribution, with both rates greater than zero.

Proof Suppose $J_C(\tau) = J_D(\tau)$ for all τ. Since they agree for $\tau = 0$, when $J_C(\tau) = \Pi$ and $J_D(\tau) = \Theta$, we must have $\theta = \pi$. Multiply $J_C(\tau) = J_D(\tau)$ on the left and right by $\Pi^{-1/2}$ to get

$$\sum_{j=1}^{r} C_j(\tau) u_j u_j^{\mathrm{T}} = \sum_{j=1}^{r} M(\alpha_j \tau) y_j y_j^{\mathrm{T}},$$

where $C_j(\tau) = c_j^+ e^{\mu_j^+ \tau} + c_j^- e^{\mu_j^- \tau}$. Now $u_j^{\mathrm{T}} u_k = \delta_{jk}$ implying $\Pi^{-1/2} J_C(\tau) \Pi^{-1/2}$ has eigenvalues $\{C_j(\tau)\}$ and corresponding eigenvectors $\{u_j\}$. Similarly $\Pi^{-1/2} J_D(\tau) \Pi^{-1/2}$ has eigenvalues $\{M(\alpha_j \tau)\}$. So there must be some ordering for which $C_j(\tau) = M(\alpha_j \tau)$ for each j. We will suppose that the functions have been ordered in this way.

Write $M_j(\tau)$ for $M(\alpha_j \tau)$. For the zero eigenvalue ($j = 1$) we have $C_1(\tau) = 1 = M_1(\tau)$ so we need only worry about the non-zero eigenvalues ($j \geq 2$). The M_j ($j \geq 2$) have the property that $M_k(\alpha_l \tau / \alpha_k) = M_l(\tau)$, that is we may transform from one to another simply by re-scaling τ. Clearly the C_j must satisfy this also. Suppose $C_k(\gamma\tau) = C_l(\tau)$. Then

$$C_k(\gamma\tau) = c_k^+ e^{\mu_k^+ \gamma\tau} + c_k^- e^{\mu_k^- \gamma\tau} = C_l(\tau) = c_l^+ e^{\mu_l^+ \tau} + c_l^- e^{\mu_l^- \tau},$$

so we must have $\gamma\mu_k^+ = \mu_l^+$, $\gamma\mu_k^- = \mu_l^-$, $c_k^+ = c_l^+$ and $c_k^- = c_l^-$ since exponential functions are independent (note that $\mu_j^- < \mu_j^+$ which precludes the possibility of matching $\gamma\mu_k^+$ with μ_l^-, etc.). Hence, from the definition of c_l^-, we have

$$\begin{aligned} c_l^- &= \frac{(s_1+s_2+\mu_l^-)\mu_l^+}{(s_1+s_2)(\mu_l^+-\mu_l^-)} \\ &= \frac{(s_1+s_2+\gamma\mu_k^-)\gamma\mu_k^+}{(s_1+s_2)(\gamma\mu_k^+-\gamma\mu_k^-)} \\ &= \frac{(s_1+s_2+\gamma\mu_k^-)\mu_k^+}{(s_1+s_2)(\mu_k^+-\mu_k^-)} \end{aligned}$$

which by hypothesis equals

$$c_k^- = \frac{(s_1+s_2+\mu_k^-)\mu_k^+}{(s_1+s_2)(\mu_k^+-\mu_k^-)}.$$

Hence $\gamma = 1$. Lemma 6.3 (i) then implies that $\lambda_k = \lambda_l$ and it follows that R has only one distinct non-zero eigenvalue λ. Now $M_k(\tau) = C_k(\tau) = C_j(\tau) = M_j(\tau)$, $2 \le j,k \le r$, so Q has only one distinct non-zero eigenvalue also, and since R and Q both have stationary distribution π, by Lemma 6.4 they are both scalar multiples of R_π. $\mathcal{D}$ has moment generating function $M(\tau) = c_\lambda^+ e^{\mu_\lambda^+\tau} + c_\lambda^- e^{\mu_\lambda^-\tau}$ and so is two rate with both rates greater than zero.

Conversely, if $R = -\lambda R_\pi$ then

$$J_C(\tau) = \pi^{\mathrm{T}}\pi + [c_\lambda^+ e^{\mu_\lambda^+\tau} + c_\lambda^- e^{\mu_\lambda^-\tau}] \sum_{j=2}^{r} w_j w_j^{\mathrm{T}}.$$

Let $\mathcal{D}$ be the two rate distribution such that

$$\mathsf{P}[\nu = |\mu_\lambda^*|] = c_\lambda^*, \qquad * = +,-.$$

Then $\mathcal{D}$ is well defined and if $D = (R_\pi, \mathcal{D})$ we have

$$\begin{aligned} J_D(\tau) &= \pi^{\mathrm{T}}\pi + \sum_{j=2}^{r} M(-\tau) w_j w_j^{\mathrm{T}} \\ &= \pi^{\mathrm{T}}\pi + [c_\lambda^+ e^{\mu_\lambda^+\tau} + c_\lambda^- e^{\mu_\lambda^-\tau}] \sum_{j=2}^{r} w_j w_j^{\mathrm{T}} \\ &= J_C(\tau). \end{aligned}$$

It remains to show that if $D = (\gamma R_\pi, \mathcal{D})$ where $\mathcal{D}$ is a two rate distribution such that

$$\mathsf{P}[\nu = \nu_i] = \rho_i, \quad i = 1,2$$

then we may choose a covarion model $C = (R,S)$ such that $J_C(\tau) = J_D(\tau)$ for all τ. By scaling ν_1 and ν_2 if necessary we may assume that $\gamma = 1$, and $0 < \nu_1 < \nu_2$. We must then find $\lambda < 0$, and $s_1, s_2 > 0$ such that

$$\mu_\lambda^+ = -\nu_1, \quad \mu_\lambda^- = -\nu_2 \quad \text{and} \quad c_\lambda^- = \rho_2,$$

and then take $R = -\lambda R_\pi$ (note that the third condition implies $c_\lambda^+ = \rho_1$). Using

$$(\mu+\nu_1)(\mu+\nu_2) = (\mu-\mu_\lambda^+)(\mu-\mu_\lambda^-) = \mu^2 + (s_1+s_2-\lambda)\mu - s_2\lambda$$

we obtain the system of equations

$$\begin{aligned} \nu_1\nu_2 &= -s_2\lambda \\ \nu_1+\nu_2 &= s_1+s_2-\lambda \\ \rho_2 &= \frac{(s_1+s_2-\nu_2)\nu_1}{(\nu_1-\nu_2)(s_1+s_2)} \end{aligned}$$

which may be solved (uniquely) to give

$$\begin{aligned} \lambda &= -\frac{\nu_1^2\rho_1+\nu_2^2\rho_2}{\nu_1\rho_1+\nu_2\rho_2} \\ s_1 &= \frac{\rho_1\rho_2\nu_1\nu_2(\nu_1-\nu_2)^2}{(\nu_1\rho_1+\nu_2\rho_2)(\nu_1^2\rho_1+\nu_2^2\rho_2)} \\ s_2 &= \frac{\nu_1\nu_2(\nu_1\rho_1+\nu_2\rho_2)}{\nu_1^2\rho_1+\nu_2^2\rho_2}. \end{aligned}$$

This defines the required covarion model.

•

6.3.5 Limiting cases

We consider the limiting cases of the covarion model when the switch is very slow ($s_1, s_2 \to 0$) and very fast ($s_1, s_2 \to \infty$), keeping s_1/s_2 (the ratio of "off" sites to "on" sites) constant.

For very slow switches we expect few changes between the states `on` and `off` to occur, so that sites in state `on` will tend to remain in state `on`, and sites in state `off` will tend to remain in state `off`. In the limiting case $s_1, s_2 \to 0$ we expect σ_2 of the sites to be invariant and σ_1 of them to be variable. Calculating this limit we find

$$J_C(\tau) \to \sigma_2 J(0) + \sigma_1 J(\tau)$$

as expected.

For fast switches we expect sites to flip back and forth between `on` and `off` very rapidly, and each spend about the same amount of time in state `on`. Differentiating equation (6.1) with respect to λ and setting $\lambda = 0$ we find that the expected time in state `on` is $\sigma_1\tau$, so in the limiting case $s_1, s_2 \to \infty$ with s_1/s_2 constant we expect

$$J_C(\tau) \to J(\sigma_1\tau).$$

Calculating this limit we find this is indeed the case.

6.4 A tree-additive distance on monophyletic groups under the covarion model

One approach to testing the covarion model against rates-across-sites models is to examine the sites that are varied and unvaried in two widely separated groups of closely related species. Under the rates-across-sites model, if a given site is in the same state for each member of a group of closely related taxa, then it is likely that the rate of evolution at that site is slow. Since the rate does not change across the tree, we might expect little change to occur in another group of closely related species that is widely separated from the first. On the other hand, under the covarion model if each species has the same state at a given site it seems likely that the site was off for much of the time. In a distant part of the tree the switch might be on so we no longer expect the unvaried

sites in the two groups to match up. This observation was made by Fitch [11], and examined by Miyamoto and Fitch [30], who compared Cu, Zn superoxide dismutase (SOD) sequences from seven mammals and seven plants with simulated sequences generated under covarion and gamma distribution rates-across-sites models, finding that the covarion hypothesis explained the evolution of the protein better than rates-across-sites.

The following discussion is also motivated by Fitch's observation. For a certain class of events and parameters of a covarion model we obtain a tree-additive distance between monophyletic groups of species that will not in general be tree-additive under rates-across-sites models. This shows firstly that infinite sequences can in fact distinguish between the two models, and secondly that infinite sequences do contain information about the tree without requiring knowledge of the parameters of the model. Standard statistical techniques (such as maximum likelihood for tree reconstruction) may then be used to address these questions given finite sequences.

The class of covarion models for which this is relevant includes those whose underlying observable process is based on the Kimura [26] three-substitution-type model (K3ST) or one of its submodels (the Kimura [25] two parameter (K2P) and Jukes-Cantor [22] (JC) models).

6.4.1 Separable events

We describe a class of events that give rise, under the covarion model, to a tree-additive distance that is not in general tree-additive under a rates-across-sites model.

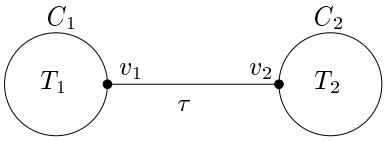


Figure 6.3: The tree joining two monophyletic groups of species C_1 and C_2. The circles denote the rooted subtrees T_1 and T_2, the roots being v_1 and v_2 respectively. The edge $\{v_1, v_2\}$ has length τ.

Suppose E is an event involving an r-state character χ on a set C of species, for example the events

$$E^s = \text{“}\chi(i) \text{ is the same state for all } i \in C\text{”}$$

and

$$E^d = \text{“}\chi(i) \text{ is not the same state for all } i \in C\text{”}.$$

Given two monophyletic groups C_1 and C_2 of species with corresponding rooted trees T_1 and T_2, the tree joining them will be as shown in figure 6.3. Let E_i be the event “E occurs for group C_i” and $\mathtt{0}_i$ the state `on` or `off` of the switch at the vertex v_i for $i = 1, 2$. We say that the event E is *separable* under the covarion model (R, S) if

$$\mathsf{P}[E_1 \wedge E_2 | \mathtt{0}_1 = \mathtt{o}_1, \mathtt{0}_2 = \mathtt{o}_2] = \mathsf{P}[E_1 | \mathtt{0}_1 = \mathtt{o}_1]\mathsf{P}[E_2 | \mathtt{0}_2 = \mathtt{o}_2] \tag{6.9}$$

for all $\mathtt{o}_1, \mathtt{o}_2 \in \{\mathtt{on}, \mathtt{off}\}$. Note that the separability of a given event may depend on R and S. An analogous condition that might be satisfied by a rates-across-sites model $(Q, \mathcal{D})$ is the following *independence condition:*

$$\mathsf{P}[E_1 \wedge E_2 | \nu] = \mathsf{P}[E_1 | \nu]\mathsf{P}[E_2 | \nu]. \tag{6.10}$$

Let

$$\begin{aligned} p_{12} &= \mathsf{P}[E_1 \wedge E_2], \\ p_i &= \mathsf{P}[E_i], \quad i = 1, 2 \end{aligned}$$

and further in the case of the covarion model let

$$\begin{aligned} p_i^{\mathtt{on}} &= \mathsf{P}[E_i|0_i = \mathtt{on}] \\ p_i^{\mathtt{off}} &= \mathsf{P}[E_i|0_i = \mathtt{off}] \\ \delta_i &= p_i^{\mathtt{on}} - p_i^{\mathtt{off}} \end{aligned}$$

for $i = 1, 2$. Then under conditions (6.9) and (6.10) we have the following:

Lemma 6.5

(i) *If E is separable under the covarion model (R, S) then*

$$p_{12} - p_1 p_2 = \sigma_1 \sigma_2 e^{-(s_1+s_2)\tau} \delta_1 \delta_2. \tag{6.11}$$

(ii) *If the independence condition holds for the rates-across-sites model $(Q, \mathcal{D})$ then $p_{12} - p_1 p_2$ does not depend on τ.*

Theorem 6.3 *For a tree with several monophyletic groups $C_1, \ldots, C_n$ ($|C_i| \geq 2$ for each i) at its tips the distance*

$$\rho_{ij} = -\ln |p_{ij} - p_i p_j| \qquad i \neq j$$

is tree-additive under a covarion model for which E is separable, but in general is not under a rates-across-sites model for which the independence condition holds.

Proof of Lemma 6.5 and Theorem 6.3 In the covarion case

$$\begin{aligned} p_{12} &= \sum_{o_1, o_2} \mathsf{P}[E_1 \wedge E_2 | 0_1 = o_1, 0_2 = o_2] \mathsf{P}[0_1 = o_1, 0_2 = o_2] \\ &= \sum_{o_1, o_2} \mathsf{P}[E_1|0_1 = o_1] \mathsf{P}[E_2|0_2 = o_2] \mathsf{P}[0_1 = o_1, 0_2 = o_2] \end{aligned}$$

since E is separable, and

$$p_1 p_2 = \sum_{o_1, o_2} \mathsf{P}[E_1|0_1 = o_1] \mathsf{P}[E_2|0_2 = o_2] \mathsf{P}[0_1 = o_1] \mathsf{P}[0_2 = o_2].$$

Thus

$$p_{12} - p_1 p_2 = \sum_{o_1, o_2} \mathsf{P}[E_1|0_1 = o_1] \mathsf{P}[E_2|0_2 = o_2] \left(\mathsf{P}[0_1 = o_1, 0_2 = o_2] - \mathsf{P}[0_1 = o_1] \mathsf{P}[0_2 = o_2] \right). \tag{6.12}$$

Now from equation (2.6) the joint probability matrix for the switch operating for time τ is

$$(\mathsf{P}[0_1 = o_1, 0_2 = o_2]) = \sigma_1 \sigma_2 \begin{pmatrix} \frac{s_2}{s_1} + e^{-(s_1+s_2)\tau} & 1 - e^{-(s_1+s_2)\tau} \\ 1 - e^{-(s_1+s_2)\tau} & \frac{s_1}{s_2} + e^{-(s_1+s_2)\tau} \end{pmatrix}$$

so the matrix of $\mathsf{P}[0_1 = o_1, 0_2 = o_2] - \mathsf{P}[0_1 = o_1] \mathsf{P}[0_2 = o_2]$ is

$$\sigma_1 \sigma_2 e^{-(s_1+s_2)\tau} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Hence, from (6.12),

$$p_{12} - p_1 p_2 = \sigma_1 \sigma_2 e^{-(s_1+s_2)\tau} (p_1^{\mathtt{on}} - p_1^{\mathtt{off}})(p_2^{\mathtt{on}} - p_2^{\mathtt{off}}) = \sigma_1 \sigma_2 e^{-(s_1+s_2)\tau} \delta_1 \delta_2$$

as claimed.

Under rates-across-sites with the independence condition holding,

$$\begin{aligned}\mathsf{P}[E_1 \wedge E_2] &= \int_0^\infty \mathsf{P}[E_1 \wedge E_2|\nu] dF_{\mathcal{D}}(\nu) \\ &= \int_0^\infty \mathsf{P}[E_1|\nu]\mathsf{P}[E_2|\nu] dF_{\mathcal{D}}(\nu)\end{aligned}$$

which does not depend on τ, and similarly

$$\mathsf{P}[E_i] = \int_0^\infty \mathsf{P}[E_i|\nu] dF_{\mathcal{D}}(\nu)$$

does not depend on τ, so that $p_{12} - p_1 p_2 = \mathsf{P}[E_1 \wedge E_2] - \mathsf{P}[E_1]\mathsf{P}[E_2]$ does not depend on τ either.

Since ρ_{ij} does not depend on the length of the edge between T_i and T_j in the rates-across-sites case, we may rearrange the tree on the groups without changing the value of ρ_{ij}, so the tree on the groups is not uniquely determined by ρ. In the covarion case, if the edge between T_x and T_y has total length τ_{xy} then

$$\begin{aligned}\rho_{xy} &= -\ln|p_{xy} - p_x p_y| \\ &= -\ln\left(\sigma_1\sigma_2 e^{-(s_1+s_2)\tau_{xy}}|\delta_x||\delta_y|\right) \\ &= -\ln(\sigma_1\sigma_2) + (s_1+s_2)\tau_{xy} - \ln|\delta_x| - \ln|\delta_y|.\end{aligned}$$

Referring to figure 6.4 we have $\tau_{ij} = \tau_i + \tau_j$, $\tau_{ik} = \tau_i + \tau_m + \tau_k$ etc., and Theorem 6.3 follows.

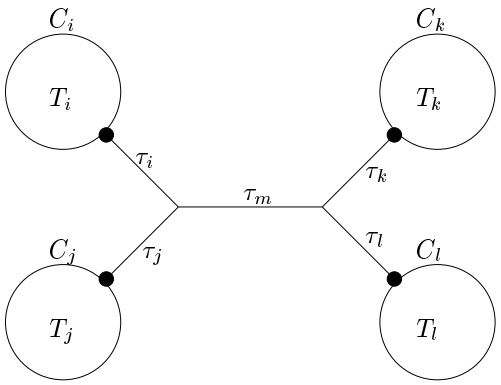


Figure 6.4: The tree on the four monophyletic groups of species C_i, C_j, C_k and C_l. The τ_x are the edge lengths.

•

Note that the set of equations

$$\rho_{ij} = -\ln(\sigma_1\sigma_2) + (s_1+s_2)\tau_{ij} - \ln|\delta_i| - \ln|\delta_j|$$

where $1 \le i < j \le 4$ is a system of six linear equations in the six unknowns $\ln(\sigma_1\sigma_2)$, $(s_1+s_2)\tau_k - \ln|\delta_k|$, $1 \le k \le 4$ and $(s_1+s_2)\tau_5$. However this system is singular (since $\rho_{13} + \rho_{24} = \rho_{14} + \rho_{23}$) and only $(s_1+s_2)\tau_5$ may be solved for uniquely if only the ρ_{ij} are known.

Further, although ρ_{ij} is not in general tree-additive under a rates-across-sites model, the four-point condition may still hold, albeit in the form

$$\rho_{ij} + \rho_{kl} = \rho_{ik} + \rho_{jl} = \rho_{il} + \rho_{jk}. \tag{6.13}$$

For example, if trees T_i, T_j, T_k and T_l are exactly the same, this will certainly be the case. Less restrictively, if the T_x are all two taxa trees with their leaves separated by a distance τ_x, and we assume the fully symmetric model ($Q_{ij} = \alpha$ if $i \neq j$ and $Q_{ii} = (1-r)\alpha$) then the ρ_{xy} may be calculated relatively easily and it appears that (6.13) holds for any choice of τ_i, τ_j, τ_k and τ_l if and only if $\mathcal{D}$ is a discrete one or two rate distribution.

6.4.2 Examples of separable events

We begin by giving a sufficient condition for separability, which will allow us to show that under a model that regards the states somewhat interchangeably, any event that respects that interchangeability will be separable. We will then be able to find some examples of separable events.

Let A_i be the state of the observable process at vertex v_i. Then

Lemma 6.6

(i) *Under the covarion model (R, S), if E_i is independent of A_i for $0_i = \texttt{on}$ and $0_i = \texttt{off}$ then E is separable.*

(ii) *Under the rates-across-sites model $(Q, \mathcal{D})$, if E_i is conditionally independent of A_i given ν then the independence condition holds.*

Proof The proofs of parts (i) and (ii) are entirely similar so we prove only (i). For any reversible Markov tree model we have

$$\mathsf{P}[E_1 \wedge E_2 | 0_1 \wedge 0_2, A_1 \wedge A_2] = \mathsf{P}[E_1 | 0_1 \wedge A_1]\mathsf{P}[E_2 | 0_2 \wedge A_2]. \tag{6.14}$$

Let $p_i(a_i) = \mathsf{P}[E_i | 0_i \wedge (A_i = a_i)]$. Then from (6.14),

$$\mathsf{P}[E_1 \wedge E_2 | 0_1 \wedge 0_2] = \sum_{a_1, a_2 \in \mathcal{A}} p_1(a_1) p_2(a_2) \mathsf{P}[(A_1 = a_1) \wedge (A_2 = A_2) | 0_1 \wedge 0_2].$$

Also,

$$\mathsf{P}[E_i | 0_i] = \sum_{a_i \in \mathcal{A}} p_i(a_i) \mathsf{P}[A_i = a_i | 0_i],$$

so that

$$\begin{aligned} \Delta &= \mathsf{P}[E_1 \wedge E_2 | 0_1 \wedge 0_2] - \mathsf{P}[E_1 | 0_1]\mathsf{P}[E_2 | 0_2] \\ &= \sum_{a_1, a_2 \in \mathcal{A}} p_1(a_1) p_2(a_2) \left(\mathsf{P}[(A_1 = a_1) \wedge (A_2 = a_2) | 0_1 \wedge 0_2] - \mathsf{P}[A_1 = a_1 | 0_1]\mathsf{P}[A_2 = a_2 | 0_2] \right). \end{aligned}$$

Now, if E_i is independent of A_i for $0_i = \texttt{on}$ and for $0_i = \texttt{off}$ then we may write $p_i(a_i) = p_i$ and so

$$\begin{aligned} \Delta &= p_1 p_2 \left(\sum_{a_1, a_2 \in \mathcal{A}} \mathsf{P}[(A_1 = a_1) \wedge (A_2 = a_2) | 0_1 \wedge 0_2] - \sum_{a_1, a_2 \in \mathcal{A}} \mathsf{P}[A_1 = a_1 | 0_1]\mathsf{P}[A_2 = a_2 | 0_2] \right) \\ &= p_1 p_2 (1 - 1) = 0. \end{aligned}$$

Hence E is separable.

•

Given a permutation $\sigma \in S_r$ (the symmetric group on r objects), the permutation matrix P_σ corresponding to σ is the matrix whose $i\sigma(i)$-entry is 1 with all other entries being zero. If R is an $r \times r$ matrix then $P_\sigma R$ is the matrix that results if the rows of R are swapped according to σ, while RP_σ^{T} is the matrix that results if the columns of R are swapped according to σ. Consequently $P_\sigma R P_\sigma^{\mathrm{T}}$ is the result of swapping both the rows and columns.

The map $R \mapsto \sigma R = P_\sigma R P_\sigma^{\mathrm{T}}$ defines a group action on the set of all $r \times r$ rate matrices. If $R = P_\sigma R P_\sigma^{\mathrm{T}}$ we say that R is *invariant* under σ; further, if G is a subgroup of S_r and R is invariant under σ for all $\sigma \in G$ we say that R is invariant under the action of G. We note that since $P_\sigma^{\mathrm{T}} = P_{\sigma^{-1}} = P_\sigma^{-1}$, the set of matrices invariant under the action of G is closed under multiplication.

As an example, consider the matrices

$$R_K = \begin{pmatrix} -\delta & \alpha & \beta & \gamma \\ \alpha & -\delta & \gamma & \beta \\ \beta & \gamma & -\delta & \alpha \\ \gamma & \beta & \alpha & -\delta \end{pmatrix} \quad \text{and} \quad R_C = \begin{pmatrix} -\delta & \alpha & \beta & \gamma \\ \gamma & -\delta & \alpha & \beta \\ \beta & \gamma & -\delta & \alpha \\ \alpha & \beta & \gamma & -\delta \end{pmatrix} \tag{6.15}$$

where $\delta = \alpha + \beta + \gamma$. It is easily checked that R_K (which is the matrix used in the K3ST model) is invariant under the action of

$$K_4 = \{\mathrm{id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\},$$

which is isomorphic to the Klein 4-group, while R_C is invariant under the action of

$$C_4 = \{\mathrm{id}, (1\ 2\ 3\ 4), (1\ 3)(2\ 4), (1\ 4\ 3\ 2)\}$$

which is isomorphic to the cyclic group Z_4.

In a similar way we may define a group action on state characters by $\chi \mapsto \sigma\chi$ where $\sigma\chi(i) = \sigma(\chi(i))$. Since events involving states of the taxa are sets of characters, this extends to an action on such events by

$$\sigma E = \{\sigma\chi | \chi \in E\}.$$

Again, if $\sigma E = E$ for all $\sigma \in G$ we say that E is invariant under the action of G. As an example, if C is a set of species then the event E^s that a given site takes the same state at each species is invariant under the action of any subgroup of S_r. For a less trivial example, consider the events on two species with four states (1,2,3 and 4) given by

$$E^2 = \text{“the states differ by 2” (e.g. 1 and 3 or 4 and 2)}$$

and

$$E^{1,3} = \text{“the states differ by 1 or 3” (e.g. 1 and 2 or 1 and 4).}$$

Again it is easily checked that E^2 and $E^{1,3}$ are invariant under the action of both K_4 and C_4. Note that if the states are the nucleotides A, C, G and T in that order then E^2 is the event “the states differ by a transition” while $E^{1,3}$ is the event “the states differ by a transversion”.

The usefulness of these concepts in the present context is given by the following theorem. Invariance of the rate matrix under the action of a group G breaks the state space up into classes of states that “look the same”. If there is just one class of states that “look the same” (that is, if there is just one orbit under the action of G; such an action is called transitive), then any event invariant under G will be separable.

Theorem 6.4 *Let R be a stationary and time-reversible $r \times r$ rate matrix and let E be an event involving r-state characters on monophyletic sets of species. If both R and E are invariant under the action of some $G \subseteq S_r$ that acts transitively on $[r]$, then*

(i) *E is separable under the covarion model (R, S) for any switch matrix S;*

(ii) *the independence condition holds for E under the rates-across-sites model $(R, \mathcal{D})$ for any distribution $\mathcal{D}$.*

Proof The result follows from a simple symmetry argument. For part (ii), the transition matrices $P^e = \exp(\tau_e R)$ inherit invariance under the action of G from R, so that

$$P^e_{\sigma(\alpha)\sigma(\beta)} = P^e_{\alpha\beta}$$

for all $\alpha, \beta \in [r]$ and all edges e of T_i. It follows that on renaming all states according to $\sigma \in G$ we have $\mathsf{P}[E_i|A_i = a_i] = \mathsf{P}[E_i|A_i = \sigma(a_i)]$. Since G acts transitively on $[r]$, (ii) now follows from Lemma 6.6 (ii).

For part (i) we use the second formulation of the covarion model and argue similarly to part (ii). We again have

$$P^e_{(\sigma(\alpha),\mathtt{o})(\sigma(\beta),\mathtt{o}')} = P^e_{(\alpha,\mathtt{o})(\beta,\mathtt{o}')}$$

for all $\alpha, \beta \in [r]$, $\mathtt{o}, \mathtt{o}' \in \{\mathtt{on}, \mathtt{off}\}$ and all edges e of T_i, so on renaming all states according to σ we get

$$\mathsf{P}[E_i|(A_i, \mathtt{0}_i) = (a_i, \mathtt{o}_i)] = \mathsf{P}[E_i|(A_i, \mathtt{0}_i) = (\sigma(a_i), \mathtt{o}_i)].$$

Since G acts transitively on $[r]$, E_i is independent of A_i for $\mathtt{0}_i = \mathtt{on}$ and $\mathtt{0}_i = \mathtt{off}$, so E is separable by Lemma 6.6 (i).

•

Theorem 6.4 will allow the construction of many examples of separable events, and we state some examples as a corollary below. Part (ii) is of interest in its own right. Fu and Li [15], in constructing certain quadratic invariants, showed that the events E^s, E^2 and $E^{1,3}$ defined above satisfy the independence condition on four taxa trees if all transition matrices have the form R_K, without placing any conditions on the location of the root or the distribution of states there. The proof of part (ii) above requires only that all transition matrices are invariant under the action of G (they need not be generated by a single continuous time Markov process); no requirements are placed on the distribution of states at the root, the number of taxa or the number of states. This extends Fu and Li's result, fitting it into a much broader framework.

We will say that R is *permutable* if it is invariant under the action of some $G \subseteq S_r$ that acts transitively on $[r]$. Then:

Corollary 6.1 (Some examples of separable events)

1. *The events E^s and E^d above are separable under the covarion model (R, S), and satisfy the independence condition under the rates-across-sites model $(R, \mathcal{D})$, whenever R is permutable. In particular, R is permutable if it has one of the following forms:*

   (i) *$R = R_K$, where R_K is as given in (6.15) and is the form of the matrix used in the K3ST model. This includes as special cases the Kimura two parameter model ($\beta = \gamma$) and the Jukes-Cantor model ($\alpha = \beta = \gamma$).*

   (ii) *R is the $r \times r$ matrix given by $R_{ij} = \alpha$ if $i \neq j$ and $R_{ii} = (1-r)\alpha$, for any r. This gives the fully symmetric model, and includes as particular cases the Cavender-Farris model ($r = 2$) and the Jukes-Cantor model ($r = 4$).*

2. *The events on pairs of species E^2 (differ by a transition) and $E^{1,3}$ (differ by a transversion) are separable under the covarion model (R, S), and satisfy the independence condition under the rates-across-sites model $(R, \mathcal{D})$, whenever R is of the form R_K.*

(Note that the matrix R_C is not time-reversible unless $\alpha = \gamma$, in which case it is of the form R_K).

6.5 Discussion

We have presented and analysed a simple covarion-style model, comparing it to the better known rates-across-sites models.

We have shown that, even for infinitely long sequences, a covarion model will give identical results to a suitably chosen rates-across-sites model when making simultaneous comparisons of pairwise dissimilarities between a collection of sequences. Consequently, if one wishes to test between these two models using real (finite length) sequences it is necessary to consider further properties of the data than just pairwise dissimilarities. We also showed how the expected pairwise dissimilarities could be transformed so as to estimate the evolutionary distance between the two sequences, however this required knowledge of the underlying rate matrices R and S.

In section 6.4, following an observation of Fitch [11], we showed how certain versions of the covarion model could be used to construct a tree-additive distance on monophyletic groups of species, again for infinitely long sequences but this time without using knowledge of the underlying rate matrices R and S. The significance of this result for real sequences is two-fold. Firstly, it shows that tree-like information can be recovered from sufficiently long sequences under the covarion-style model, given knowledge of monophyletic groupings. The particular tree-additive distance described could be used directly on real sequences, provided they are reasonably long, in much the same way as similar logarithmic transformations are routinely used in phylogenetics. Alternatively, more powerful (but also more computationally intensive) statistical techniques such as maximum likelihood could be employed—our result simply shows that tree-like information is there in the sequences to be recovered.

Secondly, because the tree-additive distance is not, in general, tree-additive under the rates-across-sites model, this shows that the two models can indeed be distinguished given sufficiently long sequences. A useful project for future work would be the development of such tests. A test that did not depend on restrictions to the model such as the separability condition of equation (6.9) would be particularly desirable.

Acknowledgements

First of all I would like to thank Mike Steel, who has been more than just a supervisor. Carrying an early draft of what is now Chapter 4 up Zurbriggens Ridge to the high peak of Cook has to count as beyond the call of duty. Maybe soon I'll have a chance to go up to Rapaki Rock and tackle what was to have been my second practical requirement, climbing *Body and Soul.*

The New Zealand Marsden Fund has helped fund these studies, through contract UOC 516.

Chapter 6 on the covarion model is joint work with Mike Steel, and we thank Dr David Penny, Dr Walter Fitch and two anonymous reviewers for their helpful comments. The work in Chapters 4 and 5 is my own, unless stated otherwise. Thanks and a *beeeeeeep!* go to Vincent Moulton for comments and advice.

Thanks also go to Chris Stephens, for keeping on working on that maximum likelihood problem I gave him in spite of my efforts to forget it, and also for providing entertaining distractions; and to Michael Burns, for helping connect some matrices (keep up the good work Mike, even matrices need to reach out and connect sometimes).

Christopher Tuffley
June 1997

References

[1] H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7:309–343, 1986.

[2] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. Ziegler. *Oriented Matroids*. Number 46 in Encyclopedia of mathematics and its applications. Cambridge University Press, 1993.

[3] P. Buneman. The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Mathematics in the archaeological and historical sciences*, pages 387–395. Edinburgh University Press, 1971.

[4] C. J. Burke and M. Rosenblatt. A Markovian function of a Markov chain. *Ann. Math. Stat.*, 29:1112–1122, 1958.

[5] J. A. Cavender. Taxonomy with confidence. *Mathematical Biosciences*, 40:270–280, 1978.

[6] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.

[7] J. T. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134:189–215, 1996.

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

[9] J. N. Darroch and K. W. Morris. Passage time generating functions for continuous-time finite Markov chains. *Journal of Applied Probability*, 5:414–426, 1968.

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

[11] W. M. Fitch. The nonidentity of invariable positions in the cytochrome c of different species. *Biochemical Genetics*, 5:231–241, 1971.

[12] W. M. Fitch. Rate of change of concomitantly variable codons. *J. Mol. Evol.*, 1:84–96, 1971.

[13] W. M. Fitch and F. J. Ayala. The superoxide dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. USA*, 91:6802–6807, 1994.

[14] W. M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4:579–593, 1970.

[15] Y.-X. Fu and W.-H. Li. Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree. *Mathematical Biosciences*, 105:229–238, 1991.

[16] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Clarendon Press, Oxford, 1982.

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

[18] M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38(4):310–321, 1989.

[19] M. D. Hendy and D. Penny. A framework for the qualitative studies of evolutionary trees. *Systematic Zoology*, 38(4):297–309, 1989.

[20] D. R. Hofstadter. *Gödel, Escher, Bach: an Eternal Golden Braid.* Penguin Books, 1979.

[21] D. L. Isaacson and R. W. Madsen. *Markov Chains: Theory and Applications.* Wiley series in probability and mathematical statistics. John Wiley and Sons, 1976.

[22] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

[23] J. Keilson. *Markov Chain Models—Rarity and Exponentiality*, volume 28 of *Applied Mathematical Sciences.* Springer-Verlag, 1979.

[24] C. Kelly and J. Rice. Modelling nucleotide evolution: A heterogeneous rate analysis. *Mathematical Biosciences*, 133:85–109, 1996.

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

[26] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. USA*, 78:454–458, 1981.

[27] R. Kindermann and J. L. Snell. *Markov Random Fields and their Applications.* Contemporary Mathematics. American Mathematical Society, 1980.

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

[29] A. T. Lundell and S. Weingram. *The topology of CW complexes.* University series in higher mathematics. Von Nostrand Reinhold, 1969.

[30] M. M. Miyamoto and W. M. Fitch. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.*, 12(3):503–513, 1995.

[31] J. Neyman. Molecular studies of evolution: a source of novel statistical problems. In S. S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, 1971.

[32] A. L. Onishchik and E. B. Vinberg. Foundations of Lie theory. In A. L. Onishchik, editor, *Lie Groups and Lie Algebras I*, volume 20 of *Encyclopedia of Mathematical Sciences.* Springer-Verlag, 1988.

[33] L. S. Pontryagin. *Topological Groups.* Gordon and Breach, second edition, 1966.

[34] A. Rényi. *Probability Theory*, volume 10 of *North-Holland Series in Applied Mathematics and Mechanics*. North-Holland Publishing Company, Amsterdam, 1970. English translation by L. Vekerdi.

[35] Y. A. Smolensky. A method for linear recording of graphs. *USSR Comput. Math. Phys.*, 2:396–397, 1969.

[36] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.*, 7(2):19–23, 1994.

[37] M. A. Steel, L. A. Székely, and M. D. Hendy. Reconstructing trees when sequence sites evolve at variable rates. *Journal of Computational Biology*, 1(2):153–163, 1994.

[38] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Marble, editors, *Molecular Systematics*, chapter 11, pages 407–514. Sinauer Associates, 2nd edition, 1996.

[39] L. A. Székely, P. L. Erdös, M. A. Steel, and D. Penny. A Fourier inversion formula for evolutionary trees. *Appl. Math. Lett.*, 6(2):13–16, 1993.

[40] C. Tuffley and M. Steel. Modelling the covarion hypothesis of nucleotide substitution. Submitted to Mathematical Biosciences, 1996.

[41] C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59(3):581–607, 1997.

[42] P. J. Waddell and M. A. Steel. General time reversible distances with unequal rates across sites. Research report 143, Dept. of Mathematics and Statistics, University of Canterbury, New Zealand, 1996.

[43] T. J. Warnow. Tree compatibility and inferring evolutionary history. *Journal of Algorithms*, 16:388–407, 1993.

[44] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ across sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.

[45] K. A. Zaretsky. Reconstruction of a tree from the distances between its pendant vertices. *Uspekhi Math. Nauk (Russian Mathematical Surveys)*, 20:90–92, 1965.

[46] A. Zarkikh. Estimation of evolutionary distance between nucleotide sequences. *J. Mol. Evol.*, 39:315–329, 1994.