CHAPTER 9

Maximum parsimony and the phylogenetic information in multistate characters

Mike Steel and David Penny

9.1 Introduction

In this chapter we investigate some of the statistical issues that surround the maximum parsimony (MP) method. Such issues have long been of interest, since the pioneering work of Farris (1973) and Felsenstein (1978). The latter was particularly interested in the question of statistical consistency: would MP select a correct tree under a simple finite-state Markov model, as the number of characters became large? Although much more is now known about the (necessary and sufficient) conditions for this to occur there is still a lot that isn't. More recently, there has also been interest in other types of statistical questions. For example, when will MP and maximum likelihood (ML) select the same tree on any given data, and under what sort of model(s) is MP an ML method?

This chapter considers this last question, and describes some new sufficient conditions for such an equivalence. We are particularly interested here in settings that involve a large state space. Traditionally most of the biological studies involving MP have involved a state space that is small (typically 2 or 4 or 20) and fixed (independent of the number of taxa). Indeed much standard software for parsimony (including PAUP*) appears to have problems dealing with a state space that has size of more than (say) 64. However increasingly there is interest in genomic characters such as gene order where the underlying state space may be very large (Rokas and Holland 2000; Moret *et al.* 2001, 2002; Gallut and Barriel 2002). For example, the order of k genes in a signed circular genome can take any of $2^k(k-1)!$ values. In these models whenever there is a change of state—for example a re-shuffling of genes by a random inversion (of a consecutive subsequence of genes)—it is likely that the resulting state (gene arrangement) is a unique evolutionary event, arising for the first time in the evolution of the genes under study. At this point the reader may object that the observed number of states in such a situation can never exceed the number n of extant species and so this is the only bound that matters. However when we come to investigate the stochastic properties of MP under simple models of state transition, it is the potential rather than the observed number of states that is important. Having a large state space allows for a low level of predicted homoplasy, leading to one of the links we report below between MP and ML.

A related central question we consider in this chapter is how many characters are needed to unambiguously recover a phylogenetic tree? We consider this both for random models of state transition, and in the deterministic setting. We also consider the question of when, on a fixed tree, we can expect the most-parsimonious reconstruction of a character to correspond exactly with its actual evolution.

This chapter is organized so the first three sections are largely 'model-free' (beyond the assumption of evolution on a tree), and the remaining three sections are based on simple Markov models of character evolution. We begin by recalling some background and definitions that

are required to state our results, and by reviewing some basic combinatorial properties of MP.

9.2 Preliminaries

Throughout this chapter, X will denote a set of n extant species or individuals. A *character (on X, over a set R of character states)* is any function χ from X into some finite set R. Throughout this chapter, we let r denote the size of R. Suppose we have a tree $T=(V,E)$. We say that T is a *tree on* X if X is a subset of V, and all vertices of T of degree 1 or 2 are contained in X. If, in addition, X is precisely the set of leaves of T we say that T is a *phylogenetic X-tree*, and if, furthermore, every vertex of T has degree 3 we say that T is *fully resolved*. Two phylogenetic X-trees are regarded as equivalent if the identity mapping from X to X induces a graph isomorphism between the two trees. Further background and mathematical details concerning phylogenetic trees can be found in Semple and Steel (2003).

The MP method for reconstructing a tree on X from a collection of characters on X can be described as follows.

Suppose we have a tree $T=(V,E)$ on X, and a character $\chi: X \to R$. A function $\bar{\chi}: V \to R$ is said to be an *extension* of χ since it describes an assignment of states to *all* the vertices of T that agrees with the states that χ stipulates at the leaves.

Let $\mathrm{ch}(\bar{\chi}, T) := |\{e=\{u,v\} \in E : \bar{\chi}(u) \neq \bar{\chi}(v)\}|$. Given a character $\chi: X \to R$, the *parsimony* score of χ on T, is defined by

$$l(\chi, T) := \min_{\bar{\chi}: V \to R,\, \bar{\chi}|X = \chi} \{\mathrm{ch}(\bar{\chi}, T)\}$$

where $\bar{\chi}|X$ denotes the restriction of $\bar{\chi}$ to X. A map $\bar{\chi}$ that extends χ and which minimizes $\mathrm{ch}(\bar{\chi}, T)$ is called a *minimal extension* (or most-parsimonious extension) of χ on T. Let

$$h(\chi, T) = l(\chi, T) - |\chi(X)| + 1$$

be the *homoplasy* of χ on T. By necessity, $h(\chi, T) \geq 0$ and when $h(\chi, T) = 0$ we say that χ is *homoplasy-free* on T. This condition is exactly equivalent to a statement that, informally, says the following: regardless of where T is rooted, one can evolve states down the tree (from the root to the leaves) in such a way that (1) the leaf states are specified by χ and (2) there is no convergent or reverse evolution (for a more formal rendition of this equivalence, see Semple and Steel 2002).

Suppose we are given a sequence $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters on X. The *parsimony* score of $\mathcal{C}$ on T, denoted $l(\mathcal{C}, T)$, is defined by

$$l(\mathcal{C}, T) := \sum_{i=1}^{k} l(\chi_i, T)$$

Any tree T on X that minimizes $l(\mathcal{C}, T)$ is said to be a *maximum parsimony* (MP) tree for $\mathcal{C}$, and the corresponding l-value is the *parsimony* or MP score of $\mathcal{C}$, denoted $l(\mathcal{C})$. Similarly, we may define

$$h(\mathcal{C}, T) := \sum_{i=1}^{k} h(\chi_i, T)$$

the total homoplasy of $\mathcal{C}$ on T, and the tree(s) on X that minimize h are precisely the MP trees (since $h(\mathcal{C}, T) = l(\mathcal{C}, T) + \text{constant}$, where the constant depends on $\mathcal{C}$ and not T). This minimal value of h we write as $h(\mathcal{C})$.

As is well known, the problem of finding an MP tree for $\mathcal{C}$ is computationally intractable (NP-hard), as shown by Foulds and Graham (1982). One might therefore ask for a more reasonable goal. For example, is it possible to determine splits that are shared by all (or some) MP trees? One sufficient condition that allows for the identification of such splits was described recently by David Bryant, and can be stated as follows (from Bryant 2003, Lemma B6). Recall that two binary characters are *compatible* if there exists a tree T on which they are both homoplasy-free (this is equivalent to the condition that at most three (of the four possible) pairs of states are assigned by these two characters).

Proposition 9.2.1. Suppose $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ is any sequence of binary characters on X. Let χ be any nontrivial binary character that is compatible with all the characters in $\mathcal{C}$. Then there exists an MP tree T for $\mathcal{C}$ that contains the X-split defined by χ. Furthermore, if χ is one of the characters in $\mathcal{C}$ then *every* MP tree for $\mathcal{C}$ contains the X-split defined by χ.

9.2.1 Bounds on the MP score of data

For a single character $\chi : X \to R$ it is easily shown that

$$\min_T \{l(\chi, T)\} = |\chi(X)| - 1 \tag{1}$$

and that

$$\max_T \{l(\chi, T)\} = |(X)| - m \tag{2}$$

where m is the largest number of species in X that are assigned the same state (formally $m = \max\{|\chi^{-1}(\alpha)| : \alpha \in R\}$). In (1) and (2) T ranges over all phylogenetic X-trees (or, equivalently, over all fully resolved phylogenetic X-trees).

For a collection $\mathcal{C}$ of characters it is also useful to determine lower bounds on $l(\mathcal{C})$. We first recall an easily computed lower bound. Form a graph by taking X as the set of vertices, and placing an edge between each pair of vertices (this produces the 'complete graph on X'). Weight each edge $\{x, y\}$ by the number of characters f in $\mathcal{C}$ for which $f(x) \neq f(y)$, then construct a minimum-length-spanning tree for this graph. This last task can be accomplished using one of the well-known polynomial-time techniques, such as Kruskal's algorithm or Prim's algorithm. Let $L(\mathcal{C})$ denote the sum of the weights of the edges in this tree. Then, $l(\mathcal{C}) \geq \frac{1}{2} L(\mathcal{C})$. Furthermore, the factor of $\frac{1}{2}$ is (asymptotically) optimal for a lower bound based on this approach due to Foulds (1984); however by adopting a more complex polynomial-time approach a slightly better approximation to $l(\mathcal{C})$ is possible (see Prömel and Steger 2000). Here we describe a quite different type of lower bound, which has the advantage of coinciding with $l(\mathcal{C})$ when the homoplasy $h(\mathcal{C})$ is zero (in contrast to the minimum-length-spanning tree bound, which does not have this property in general).

Let $\mathcal{F}$ be a family of subsets of $\{1, \ldots, k\}$ with the property that each number $1, 2, \ldots, k$ appears in the same number of sets from $\mathcal{F}$. In this case we say that $\mathcal{F}$ is *uniformly covering*. Let $v(\mathcal{F})$, or more briefly v, denote this number of sets from $\mathcal{F}$ that each number appears in (formally, $v(\mathcal{F}) = |\{S \in \mathcal{F} : j \in S\}|$ for each $j \in \{1, \ldots, k\}$). One natural example of such a family is the collection $\mathcal{F}^p$ of all subsets of $\{1, \ldots, k\}$ of fixed size p (i.e. $\mathcal{F}^p := \{S \subseteq \{1, \ldots, k\} : |S| = p\}$), for which $v(\mathcal{F}^p) = \binom{k-1}{p-1}$. A second class of examples is where $\mathcal{F}$ is a partition of $\{1, \ldots, k\}$ into nonoverlapping subsets in which case $v(\mathcal{F}) = 1$.

Given a sequence $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters on X, and a set $S \subseteq \{1, \cdots, k\}$, let $\mathcal{C}_S = (\chi_j : j \in S)$ and let

$$h^{\mathcal{F}} := \sum_{S \in \mathcal{F}} h(\mathcal{C}_S)$$

The following result extends the 'partition theorem' of Hendy *et al.* (1980).

Proposition 9.2.2. Let $\mathcal{F}$ be a uniformly covering family $\mathcal{F}$ of subsets of $\{1, \ldots, k\}$, let $\mathcal{C}$ be a sequence of characters. Then,

$$h(C) \geq \frac{1}{v(\mathcal{F})} h^{\mathcal{F}}$$

Proof. Let T_0 denote an MP tree for $\mathcal{C}$, and let $h'(j) := h(\chi_j, T_0)$. For $S \subseteq \{1, \ldots, k\}$, let $h'(\mathcal{C}_s) := \sum_{j \in S} h'(j)$. Thus, $h'(\mathcal{C}_s) \geq h(\mathcal{C}_s)$ and so

$$\begin{aligned} h(\mathcal{C}) = \sum_{j=1}^{k} h'(j) &= \frac{1}{v(\mathcal{F})} \sum_{S \in \mathcal{F}} h'(\mathcal{C}_s) \geq \frac{1}{v(\mathcal{F})} \sum_{S \in \mathcal{F}} h(\mathcal{C}s) \\ &= \frac{1}{v(\mathcal{F})} h^{\mathcal{F}} \end{aligned}$$

where the second equality is justified by the identity:

$$\begin{aligned} \sum_{S \in \mathcal{F}} h'(\mathcal{C}_s) = \sum_{S \in \mathcal{F}} \sum_{j \in s} h'(j) &= \sum_{j=1}^{k} \sum_{S : j \in S} h'(j) \\ &= \sum_{j=1}^{k} v(\mathcal{F}) h'(j) \end{aligned}$$

For applications one would construct a family $\mathcal{F}$ of (small) subsets of X that cover each element of X the same number of times, and compute $h(\mathcal{C}_S)$ for each small subset. As a special case, if we take $\mathcal{F} = \mathcal{F}^{(2)}$ (so that $v = k - 1$) and note that $h(\mathcal{C}_S) \geq 1$ whenever $\mathcal{C}_S$ is incompatible, then we obtain the following bound for any collection $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters:

$$l(C) \geq \sum_{i=1}^{k} (|\chi_i(X)| - 1) + \frac{In(C)}{k-1}$$

where $In(\mathcal{C})$ is the number of pairs of characters in $\mathcal{C}$ that are incompatible. The 'partition theorem' from Hendy *et al.* (1980), which states that if $\mathcal{F}$ is a partition of $\{1,\ldots,k\}$ then $l(\mathcal{C}) \geq \sum_{S\in\mathcal{F}} l(\mathcal{C}_S)$, also follows directly from Proposition 9.2.2.

Note that the requirement of Proposition 9.2.2 that $\mathcal{F}$ covers each element of X the same number of times can be weakened by adopting a linear programming approach. That is, if we let h be the minimal value of $\sum_{i=1}^{k} x_i$ subject to the linear inequality constraints, $x_i \geq 0$ for all $i=1,\ldots,k$, and $\sum_{j\in S} x_j \geq h(\mathcal{C}_S)$ for all $S\in\mathcal{F}$, then clearly $l(\mathcal{C}) \geq \sum_{i=1}^{k} (|\chi_i(X)| - 1) + h$.

9.3 How phylogenetically informative is a single *r*-state character?

In this section we consider the question of to how to quantify the phylogenetic information a single r-state character carries (*a priori*, without regard to other characters, or to the character's fit on an existing tree). Let $\chi : X \to R$ be a character. One measure of the phylogenetic information content of χ, based on compatibility, is the following:

$$I(\chi) = -\log(p(\chi)) \tag{3}$$

where $p(\chi)$ is the proportion of fully resolved phylogenetic X-trees for which χ is homoplasy-free. For example, if χ assigns the same state to all species in X or, at the other extreme, a separate state to each species in X then $I(\chi) = 0$, as we should expect, since every such character is homoplasy-free on all trees.

A measure of phylogenetic content is only useful if it can be readily computed. For the measure I described in (3) it might seem tempting to approximate this quantity by simulation: simply generate fully resolved trees at random and count what proportion of them allow χ to be homoplasy-free. However this turns out to be generally impractical once X becomes large, for the obvious reason: even if you simulate a huge number of large trees at random, it is likely that few if any of them will provide a homoplasy-free fit for χ. Fortunately it turns out that I can be easily computed by a simple exact formula, and without recourse to simulations. That such a formula exists is truly remarkable, and is due to a little-known but nontrivial result from Carter *et al.* (1980). Using straightforward algebra one can easily derive the following result from Theorem 2 of that paper.

Proposition 9.3.1. Suppose that a character χ partitions a set of n species into classes of size $a_1, a_2, \ldots, a_r$. Then

$$I(\chi) = \sum_{j=3}^{n-r+1} (1 - b_j)\log(2j-3)$$

where $b_j = |\{i : a_i \geq j\}|$.

For example, consider a character χ that partitions 20 species into classes of size 6, 4, 4, 3, 2, and 1. Then

$$I(\chi) = -3\log(3) - 2\log(5) + \log(11) + \log(13) + \cdots + \log(27)$$

In this example, $b_3 = 4$ (giving rise to the $-3(=1-b_3)$ multiplier for $\log(3)$), $b_4 = 3$, $b_5 = b_6 = 1$, and $b_j = 0$ for $j > 6$.

Proposition 9.3.1 may be useful for deciding how to construct and select between possible character codings, for example for genomic data. Ideally we would like $I(\chi)$ to be as large as possible, and achieving this may assist in tuning certain coding procedures. Further aspects of this information measure have also been explored recently using simulations by Dezulian and Steel (2004). At this point we will simply note an interesting consequence of Proposition 9.3.1. Firstly, if we fix r, the number of classes that X is partitioned into, then $I(\chi)$ is largest when all of the classes have (approximately) the same size. Let $I_{\max}(n,r)$ be this largest value of $I(\chi)$ over all characters χ that partition a set of size n into r non-empty sets. We may ask how this quantity varies as a function of r. Clearly if $r=1$ or $r=n$ then $I_{\max}(n,r) = 0$. Consequently, there is some intermediate value, between $r=1$ and $r=n$, where $I_{\max}(n,r)$ is largest. A plot of $I_{\max}(120,r)$ is shown in Fig. 9.1. Under the I measure, the most informative character for

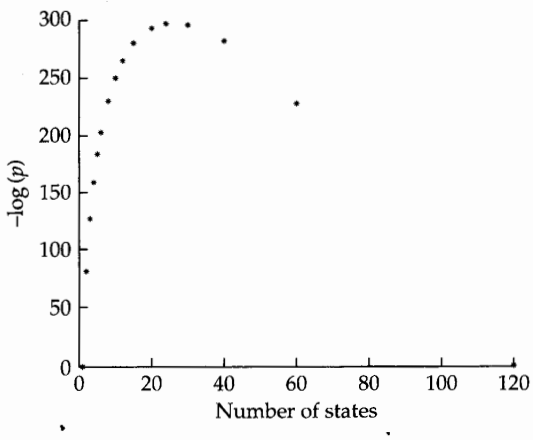

Figure 9.1 Distribution of −log of the number of fully resolved phylogenetic trees on 120 species for a homoplasy-free character that partitions the species into r equally sized sets.

$n = 120$ is one that partitions the taxa into 24 groups, each of size 5.

Other measures of the informativeness of a character are possible and have been proposed; for example, following Farris (1989), one can consider the difference

$$\delta(\chi) = \max_T\{l(\chi, T)\} - \min_T\{l(\chi, T)\}$$

where the terms on the right-hand side of this last equation are given by (1) and (2). Note that if, as above, we fix r, the number of classes that X is partitioned into by χ, then $\delta(\chi)$ is largest when all of the classes have (approximately) the same size. Let $\delta_{\max}(n, r)$ be this largest value of $\delta(\chi)$ over all characters χ that partition a set of size n into r non-empty sets. We may ask how this quantity varies as a function of r. Clearly if $r = 1$ or $r = n$ then, as with the measure based on I, we have $\delta_{\max}(n, r) = 0$. Note that if r divides n, then applying (1) and (2) gives

$$\delta_{\max}(n, r) = n(1 - \frac{1}{r}) - r + 1$$

Maximizing this expression for r we find the maximal value of this expression as r varies (over the real numbers) occurs precisely when $r = \sqrt{n}$. For the example discussed above with $n = 120$ the character that maximizes δ partitions the taxa into fewer groups (namely 10 or 12) than the 24-fold partition that maximizes I.

9.3.1 Coding gene order as multistate character data

It is instructive to consider the types of genomic data for which we may expect, simultaneously, both low homoplasy due to a large state space, and yet phylogenetically informative characters.

For gene-order data, one approach (that has been called "maximum parsimony on multistate encodings") was proposed by Bryant (2000) and tested by Wang *et al.* (2002). Suppose one has n genomes. We will take these to be circular, and consider the genes as signed (oriented) and we will suppose that the genomes have been edited so that each of them contains the set of N genes, which we can label 1, 2, ..., N. A circular gene ordering then can be regarded as a signed circular permutation, for example (1, −4, −3, −2) (which is equivalent to (−4, −3, −2, 1) or to (4, −1, 2, 3), etc.). The coding procedure considered by Bryant (2000) and Wang *et al.* (2002) is based on the observation that each gene order induces a sequence of length $2N$ by considering the gene that immediately follows each given gene in either direction. Given a collection of genomes, this allows one to define a sequence of characters $\chi_1, \ldots, \chi_{2N}$ (on a state space of size $2N$) as follows. For each i between 1 and N, set $\chi_i(j) = \pm k$ if $\pm k$ immediately follows gene i in genome j; and for each i between $N+1$ and $2N$, $\chi_i(j) = \pm k$ if $\pm k$ immediately follows gene — in genome j. For example, for $j = (1, -4, -3, -2)$ the sequence $(\chi_i(j) : i = 1, 2, \ldots, 8)$ is $(-4, 3, 4, -1, 2, 1, -2, -3)$.

The method of Gallut and Barriel (2002) has a similar flavor. In their approach each gene is associated with the (unordered) pair of genes that appear on either side of it. Thus if there are n genomes, each consisting of N genes, then this coding method produces N characters that have a state space of size $\binom{N}{2}$.

Other methods of coding are also possible, and these are currently being investigated (Dezulian and Steel, unpublished work).

9.4 The smallest number of multistate characters required for tree reconstruction

In this section we consider two related questions: given a fully resolved phylogenetic tree T with

n leaves, what is the smallest possible number of characters for which (1) T is the unique MP phylogenetic tree for these characters, and (2) T is the unique phylogenetic tree for which the characters have no homoplasy? If we call these two numbers, respectively, $n_1(T)$ and $n_2(T)$ it is clear that $n_1(T) \leq n_2(T)$. It might be expected that both these quantities would grow with the size of the tree, yet it has recently been shown that this is not so, provided no bound is placed on the size of the state space. More precisely, we have the following result, from Huber *et al.* (2002).

Theorem 9.4.1. For any fully resolved phylogenetic tree T, on any number of species, the quantities $n_1(T)$ and $n_2(T)$ are at most 4.

When a bound is placed on the size of the state space, then an elementary counting argument shows that both $n_1(T)$ and $n_2(T)$ cannot be bounded by any fixed number that is independent of the number n of leaves of T. This begs the question: how fast must $n_1(T)$ and $n_2(T)$ grow with n? In the case of *binary* characters it is well known that

$$n_2(T) = n - 3$$

since every one of the $n-3$ interior edges of the fully resolved tree T must be distinguished by at least one of the binary characters. Furthermore, for r-state characters, it was shown by Semple and Steel (2002) that

$$n_2(T) \geq \frac{n-3}{r-1}$$

and it seems that this bound is fairly close to the true value. The behavior of $n_1(T)$ has received less investigation, and consequently little is known about how large $n_1(T)$ might be. However the following result shows that $n_1(T)$ must grow at least logarithmically with n (at least for some trees).

Proposition 9.4.2. For any given state space size r, there is a positive constant c such that for each n there exists a fully resolved phylogenetic tree T with n leaves, for which $n_1(T) \geq c \cdot \log(n)$.

Proof. Suppose that to each fully resolved phylogenetic X-tree T we can associate a sequences $\mathcal{C}_T$ of k characters on X for which T is the unique MP phylogenetic tree. Then the number $B(n)$ of fully resolved phylogenetic trees on a set of size n must be less or equal to the number of sequences of k characters on a set of n species. This latter number is r^{nk} where $n = |X|$, which we may rewrite as $e^{nk\log(r)}$. Now $B(n) = \prod_{i=3}^{n}(2i-5)$ and it can be shown (using Stirling's approximation for $n!$) that for a constant $\beta > 0$ we have $B(n) > e^{\beta n \log(n)}$. Thus $B(n) \leq r^{nk}$ implies that $k \geq c \log n$ where $c = \beta/\log(r)$. This completes the proof.

It seems plausible that this lower bound on $n_1(T)$ is not too far from the true value, even for binary characters, and so we offer the following.

Conjecture 9.4.3. There exists a constant $c > 0$ such that, for any fully resolved phylogenetic tree T, there exists a sequence of at most $\lfloor c \cdot \log(n) \rfloor$ binary characters on X for which T is the unique MP phylogenetic tree, where n denotes as usual the number of leaves of T.

Proposition 9.2.1 places interesting constraints on the sorts of sequences of characters that this last conjecture requires. Namely, any split that is not in T must be incompatible with at least one of the (at most) $c \cdot \log(n)$ characters in the collection promised by the conjecture. Can such a small set of binary characters be incompatible with virtually all other binary characters? We end this section by describing a result that shows that this is indeed possible. The proof is given in Appendix 9.1.

Proposition 9.4.4. There exists a set $\mathcal{C}$ of $\log_2(n)$ binary characters on a set X of size $n(=2^k)$ with the following property: any binary character on X that is compatible with every character in $\mathcal{C}$ is a trivial character.

A further interesting feature of the type of data sets that would be required to verify Conjecture 9.4.3 is that many of the characters would need to have large homoplasy values on the tree T. The effectiveness of such data sets in recovering trees is in line with recent observations by Källersjö *et al.* (1999).

9.4.1 Reconstructing ancestral states

In the previous section we considered the question of defining a tree using parsimony. Now we will consider the analogous question for the

'small parsimony' (i.e. fixed-tree) problem. Given a phylogenetic tree X-tree, T, and a character $\chi : X \to R$ that has evolved on T, when are the states that were present at the ancestral vertices of the tree identical to the most-parsimonious reconstruction? We will present a sufficient condition (on the evolution of the character) that guarantees the historical accuracy of the ancestral-state reconstructions. Essentially this sufficient condition is that substitutions that occur are 'well-separated' in the tree (that is, they do not occur too close to each other in the tree). Apart from its intrinsic interest, this result will also be useful later in providing a limiting Poisson distribution for the parsimony score of a tree, under low substitution rates.

Theorem 9.4.5. Suppose that T is a phylogenetic X-tree, and consider the assignment of states $\overline{\chi} : V(T) \to R$ corresponding to the evolution of some character on T. Let $\chi = \overline{\chi}|X$ be the observed states on the extant set of species (leaves of T). Suppose furthermore that the evolution of the character is such that any two edges of T on which a net transition occurs are separated by at least three other edges of T. Then $\overline{\chi}$ is a minimal extension of χ on T; moreover it is the only minimal extension of χ on T.

Proof. Suppose that T is a phylogenetic X-tree, and $\overline{\chi} : V(T) \to R$. Suppose furthermore that for any two edges $\{u, v\}$ and $\{u', v'\}$ for which $\overline{\chi}(u) \neq \overline{\chi}(v)$ and $\overline{\chi}(u') \neq \overline{\chi}(v')$ there are at least three other edges separating $\{u, v\}$ and $\{u', v'\}$. Let $\chi = \overline{\chi}|X$. Then we claim that $\overline{\chi}$ is the unique minimal extension of χ on T. To establish this claim, let $\overline{\chi}'$ be a minimal extension of χ on T; we will show that for each vertex v of T we have $\overline{\chi}'(v) = \overline{\chi}(v)$.

Let us root tree T on vertex v and direct all the edges of T away from v. For any vertex u in this rooted tree, let $S(u)$ denote the set of states assigned to u by applying the first pass of the Fitch–Hartigan algorithm (Fitch 1971; Hartigan 1973) to the pair (T, χ). We will establish the following. Claim: suppose that u is an internal vertex of T and that $v_1, v_2, \ldots v_k$ are the vertices of T that are immediate descendents of u. Then

$$S(u) = \begin{cases} \{\overline{\chi}(v_1), \overline{\chi}(v_2)\}, & \text{if } k = 2 \text{ and } \overline{\chi}(v_1) \neq \overline{\chi}(v_2) \\ \{\overline{\chi}(u)\}, & \text{otherwise} \end{cases}$$

The proof of this claim is by induction on the height h of u (i.e. h is the number of edges separating u from a most distant descendant leaf). When $h = 1$ the claim holds, since the assumption on $\overline{\chi}$ implies that all but at most one (of the two or more) descendant leaves of u has the same state under χ. Suppose the claim holds for all internal vertices of height h and that u has height $h+1$. By the assumption on $\overline{\chi}$ one of the following two cases applies: (i) $\overline{\chi}(v_i) = \overline{\chi}(u)$ for all $i \in \{1, \ldots, k\}$; (ii) $\overline{\chi}(v_i) = \overline{\chi}(u)$ for all but at most one i.

In case (i), we may apply the induction hypothesis to the vertices $v_1, \ldots v_k$ which each have height at most h. It follows that $\overline{\chi}(u) \in S(v_i)$ for all i. Furthermore there is at most one vertex v_i for which $S(v_i) \neq \{\overline{\chi}(u)\}$ since if there were two such vertices, then we would obtain two edges on which $\overline{\chi}$ changes state, yet which are separated by only two edges in T. Consequently, by the Fitch–Hartigan recursion we deduce that $S(u) = \{\overline{\chi}(u)\}$.

Consider now case (ii). We may suppose that $\overline{\chi}(v_1) \neq \overline{\chi}(u)$. Consider first the case where $k > 2$. Applying the induction hypothesis to $v_1, \ldots, v_k$ and invoking the assumption on $\overline{\chi}$ we have that $S(v_1) = \{\overline{\chi}(v_1)\}$, and for all $i > 1$ we have $S(v_i) = \{\overline{\chi}(u)\}$. It now follows by the Fitch–Hartigan recursion (remembering that $k > 2$) that $S(u) = \{\overline{\chi}(u)\}$. Thus we have established the second part of the claim. It remains to consider the other possibility for case (ii), namely $k = 2$. Again we apply the induction hypothesis on v_1, v_2 and invoke the assumption on $\overline{\chi}$ to deduce that $S(v_1) = \{\overline{\chi}(v_1)\}$ and $S(v_2) = \{\overline{\chi}(v_2)\}$; hence $S(u) = \{\overline{\chi}(v_1), \overline{\chi}(v_2)\}$, as required to justify the claim.

Now let us take $u = v$, the vertex we have selected as our putative root for T. Since T is a phylogenetic tree, v has degree at least three, so by the claim we have $S(v) = \{\overline{\chi}(v)\}$. However, since v is the root of the tree for the recursion, $S(v)$ is precisely the set of states that can occur at v across all possible minimal extensions of χ on T (Hartigan 1973). Thus we have shown that all such minimal extensions (in particular $\overline{\chi}'$) assign vertex v the state as that specified by $\overline{\chi}$. Since we can repeat this argument for any vertex v in T the theorem now follows.

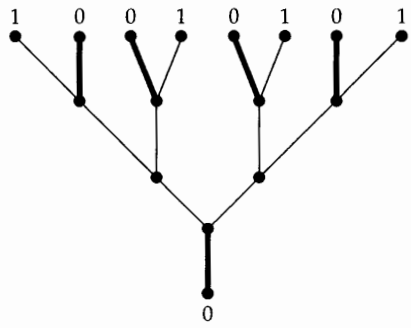

Figure 9.2 Example showing that two-edge separation does not suffice for Theorem 9.4.5.

Note that Theorem 9.4.5 is no longer true if we weaken the edge-separation requirement from three edges to two. For example, consider the tree and character χ shown in Fig. 9.2. Then the extension $\overline{\chi}$ of χ defined by making substitutions precisely on the five (bold) edges incident with leaves in state 0 as indicated in Fig. 9.2 is not a minimal extension of χ, even though each pair of bold edges is separated by at least two other edges. For this example, the minimal extension is provided by assigning state 0 to all the interior vertices of the tree. Note also that an ancestral-state reconstruction satisfying the requirements of Theorem 9.4.5 is not necessarily the 'true' reconstruction, it is merely the unique most-parsimonious reconstruction. Nevertheless, as we will see in the next section (Proposition 9.5.1), certain stochastic models of character evolution imply that this unique most-parsimonious reconstruction is also likely to be historically accurate, provided the substitution probabilities are uniformly small.

9.5 The Poisson model

In this section and the next we consider the simplest tree-based model for the evolution of characters with state space R, which we will refer to here simply as the *Poisson model on R* (with parameters (T, p)). In this model, we have a tree T on X, select any element $x_0 \in X$ as a reference vertex, and direct all edges of T away from x_0. We will regard the value from R assigned to vertex x_0 as being given (it would make little difference to the arguments below if we allowed the state at x_0 to be random). The model assigns states from R recursively to the remaining vertices of the tree according to the following scheme: if $e = \{u, v\}$ is an edge of T directed from u to v and u has been assigned state α, then, with probability $1 - p(e)$ we assign v state α, otherwise, with probability $p(e)$ we select uniformly at random one of the other $r - 1$ states (different to α) and assign this state to v. The assignments are made independently across edges, and the value $p(e)$ is called the *substitution probability* associated with edge e. It is natural to constrain $p(e)$ to lie in the interval $\left[0, \frac{r-1}{r}\right]$; the reason for the upper bound is that, if we realise this model by a continuous-time Markov process, the probability of a net substitution over any period of time is always less than $\frac{r-1}{r}$. We will say that the mapping $e \to p(e)$ is *admissible* if the $p(e)$ values all lie within this allowed interval.

When $r = 4$, this model is essentially the same as what is often referred to as the Jukes–Cantor model. For general values of r, this model was investigated in 1970 by Jerzy Neyman (1971), and has more recently been studied by Paul Lewis (2001) as a starting framework for likelihood analysis for certain morphological characters. This model has been christened in the bioinformatics literature under a variety of titles, including the Neyman r-state model and the r-state Jukes–Cantor model.

Given the pair (T, p) where $T = (V, E)$ is a tree on X, and p is an admissible assignment of transition probabilities, and given a map $\overline{\chi} : V \to R$, let $\Pr(\overline{\chi}|T, p)$ denote the probability that the vertices in T take values specified by $\overline{\chi}$ under the Poisson model on R with parameters (T, p). More formally, $\Pr(\overline{\chi}|T, p) = \Pr(\cap_{v \in V - \{x_0\}} \{\eta(v) = \overline{\chi}(v)\})$, where $\eta(v)$ is the random variable state assigned to v under the model. By the assumptions of the model, we have

$$\Pr(\overline{\chi}|T, p) = \prod_{\{u,v\} \in E: \overline{\chi}(u) \neq \overline{\chi}(v)} \frac{p(e)}{r-1} \prod_{\{u,v\} \in E: \overline{\chi}(u) = \overline{\chi}(v)} (1 - p(e)) \tag{4}$$

For any character $\chi : X \to R$, let

$$\Pr(\chi|T, p) = \sum_{\overline{\chi} \in c(\chi)} \Pr(\overline{\chi}|T, p)$$

where $c(\chi) = \{\overline{\chi} : V \to R : \overline{\chi}|X = \chi\}$.

9.5.1 Distribution of the parsimony score

Theorem 9.4.5 has the following consequence for the (limiting) distribution of the parsimony score of a character under the Poisson model.

Proposition 9.5.1. Consider a process on a fully resolved phylogenetic tree T with n leaves, and let

$$h = \max\{p(e) : e \in E\}\sqrt{n}$$

and

$$\mu = \sum_{e \in E(T)} p(e)$$

Generate a character χ by this process on T and let $\overline{\chi}$ denote the states at all the vertices of T. Then, for small values of h, the most-parsimonious reconstruction of χ is likely to be both unique and historically accurate, and the parsimony score $\mathcal{L} = l(\chi, T)$ of a character χ generated by this process on T is closely approximated by a Poisson distribution with mean μ. More precisely, for any value of h we have (i) $\Pr[\overline{\chi}$ is the unique MP reconstruction of χ on $T] \geq 1 - 28h^2$, (ii) $|\Pr(\mathcal{L} = k) - e^{-\mu}\frac{\mu^k}{k!}| < 32h^2$, and (iii) $\Sigma_{k=0}^{\infty}|\Pr(\mathcal{L} = k) - e^{-\mu}\frac{\mu^k}{k!}| < 60h^2$.

To illustrate this result, suppose that a fully resolved phylogenetic X-tree has $n = 10\,000$ leaves, and the substitution probability $p(e)$ on each edge is (say) 2×10^{-4}. In this case we can take $h = 2 \times 10^{-2}$, and so we may approximate $\mathcal{L}$ closely by a Poisson distribution with mean 4.

Notice that, in Proposition 9.5.1, a small value of h does not necessarily imply a small value for μ if the number of leaves in the tree T is large.

Proof of Proposition 9.5.1

Let A be the event that substitutions occur on some pair of edges that are separated by two or fewer edges. The number of ordered pairs of edges that are separated by two or fewer edges is at most $(2n-3) \cdot (4+8+16)$ since $(2n-3)$ is the number of edges of T and since $(4+8+16)$ bounds the number of edges of T that are separated by 0, 1 or 2 other edges from any given edge of T. Thus the number of unordered pairs of edges that are separated by two or fewer edges is at most $\frac{1}{2} \cdot (2n-3) \cdot (4+8+16) < 28n$, and so, by the Bonferroni inequality,

$$\Pr(A) < 28n \cdot \left(\frac{h}{\sqrt{n}}\right)^2 \leq 28h^2 \tag{5}$$

which, together with Theorem 9.4.5 establishes part (i).

Let $\mathcal{L}^*$ denote the random number of edges of T on which there is a substitution. Thus $\mathcal{L} \leq \mathcal{L}^*$, and $\mathcal{L}^*$ has a limiting Poisson distribution since it is the sum of an increasing (with n) number of independent 0/1 random variables, where the probability that each variable takes the value 1 converges to 0 (with n). Moreover, Le Cam's inequality (Le Cam 1960) gives

$$\sum_{k=0}^{\infty} |\Pr(\mathcal{L}^* = k) - e^{-\mu}\frac{\mu^k}{k!}| < 2\sum_e p(e)^2 \tag{6}$$

By the law of total probability,

$$\begin{aligned}\Pr(\mathcal{L} = k) =& \Pr(\mathcal{L} = k|A^c)\Pr(A^c) \\ &+ \Pr(\mathcal{L} = k|A)\Pr(A)\end{aligned} \tag{7}$$

and

$$\begin{aligned}\Pr(\mathcal{L}^* = k) =& \Pr(\mathcal{L}^* = k|A^c)\Pr(A^c) \\ &+ \Pr(\mathcal{L}^* = k|A)\Pr(A)\end{aligned} \tag{8}$$

where A^c is the complementary event of A.

Now, conditional on the event A^c, Theorem 9.4.5 guarantees that $\mathcal{L} = \mathcal{L}^*$ (with probability 1); that is, $\Pr(\mathcal{L} = k|A^c) = \Pr(\mathcal{L}^* = k|A^c)$. Applying this identity to (7) and (8) gives

$$\begin{aligned}|\Pr(\mathcal{L} = k) - \Pr(\mathcal{L}^* = k)| &= |\Pr(\mathcal{L} = k|A) \\ &- \Pr(\mathcal{L}^* = k|A)|\Pr(A) \leq \Pr(A) < 28h^2\end{aligned}$$

where the last inequality is from (5). Furthermore, (6) implies that

$$|\Pr(\mathcal{L}^* = k) - e^{-\mu}\frac{\mu^k}{k!}| < 4h^2$$

Combining these last two inequalities gives

$$|\Pr(\mathcal{L} = k) - e^{-\mu}\frac{\mu^k}{k!}| < 32h^2$$

which establishes part (ii). Similarly,

$$\sum_{k=0}^{\infty} |\Pr(\mathcal{L} = k) - \Pr(\mathcal{L}^* = k)|$$
$$\leq \sum_{k=0}^{\infty} |\Pr(\mathcal{L} = k|A) - \Pr(\mathcal{L}^* = k|A)|\Pr(A)$$
$$\leq 2\Pr(A) < 56h^2$$

from which part (iii) now follows.

9.6 Links between MP and ML

Given a sequence $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters on X, we put

$$\Pr(\mathcal{C}|T, p) = \prod_{i=1}^{k} \Pr(\chi_i|T, p),$$
$$L(T|\mathcal{C}) = \sup_p (\Pr(\mathcal{C}|T, p)),$$

$$\Pr(\mathcal{C}|T, p)_{\text{mp}} = \prod_{i=1}^{k} \max(\Pr(\overline{\chi_i}|T, p)|\overline{\chi_i} \in c(i))$$
$$L_{\text{mp}}(T|\mathcal{C}) = \sup_p (\Pr(\mathcal{C}|T, p)_{\text{mp}})$$

where the supremum is taken over all admissible choices of p and $c(i) = c(\chi_i)$ is the set of extensions of χ_i to V. Note that $\Pr(\mathcal{C}|T, p)$ is the probability of generating the k characters by independent and identical evolution under a Poisson model with parameters (T, p).

Similarly one has analogous definitions for the 'no common mechanism' Poisson model, in which each character evolves independently under a Poisson model on R but where p in the parameter pair (T, p) for this model takes admissible values that are permitted to vary freely between the characters. Specifically, let

$$\Pr(\mathcal{C}|T, (p_1, \ldots, p_k)) = \prod_{i=1}^{k} \Pr(\chi_i|T, p_i)$$

and

$$L_{\text{ncm}}(T|\mathcal{C}) = \sup_{(p_1, \ldots, p_k)} (\Pr(\mathcal{C}|T, (p_1, \ldots, p_k)))$$

where the supremum is taken over all k-tuples $(p_1, \ldots, p_k)$ where each p_i is admissible.

Recall that $L(T|\mathcal{C})$ and L_{ncm} are referred to as the *maximum (average) likelihood* or ML score, and $L_{\text{mp}}(T|\mathcal{C})$ as the *most-parsimonious likelihood* or MPL score, of T given $\mathcal{C}$ (cf. Barry and Hartigan 1987; Steel and Penny 2000).

The distinction between these two forms of likelihood is as follows: the ML score of T is the largest probability (over all admissible choices of substitition probabilities p) of generating the observed sequence of characters at the leaves of T but without specifying or conditioning on any particular assignment of sequences of characters at the interior vertices of the tree (these are effectively 'averaged over'). In contrast the MPL score of T is the largest probability (over all admissible choices of substitution probabilities p) of generating any particular assignment of sequence of characters to all the vertices of the tree, so that the sequences assigned to the tips are the observed sequences.

A tree T on X is said to be an ML tree or an MPL tree for $\mathcal{C}$ if $L(T|\mathcal{C}) \geq L(T'|\mathcal{C})$ or $L_{\text{mp}}(T|\mathcal{C}) \geq L_{\text{mp}}(T'|\mathcal{C})$, respectively, holds for all other trees T' on X. The problem of finding an MPL tree given only $\mathcal{C}$ was recently shown to be NP-hard by Addario-Berry *et al.* (2004) (where the method is referred to as "ancestral maximum likelihood",). Finding an MP tree from $\mathcal{C}$ is also NP-hard (Foulds and Graham 1982); most likely so too is the problem of finding an ML tree for $\mathcal{C}$.

We say that an MP, ML, or MPL tree for $\mathcal{C}$ is *irreducible* if we cannot collapse any edge of T to obtain another such tree for $\mathcal{C}$.

We now describe three links between two tree reconstruction methods, one of which (ML) is based explicitly on an underlying Markov model for the evolution of characters on a tree (the Poisson model), while the other method—MP—is based solely on a minimality principle.

9.6.1 Link 1: no common mechanism and an extension

MP is an ML estimator for phylogenetic trees under the 'no common mechanism' model described above. In particular, a tree T maximizes $L_{\text{ncm}}(T|\mathcal{C})$ precisely if T is an MP tree for $\mathcal{C}$. This result, established in Tuffley and Steel (1997), extended the result

for $r=2$ that was described by Penny *et al.* (1994). Here we describe a further slight extension of this result where we allow the size of the state space of the Poisson model to vary from character to character. In this case it can be shown that a weighted form of MP is an ML estimator for a phylogenetic tree under the 'no common mechanism' model.

First recall that character-weighted parsimony is directly analogous to standard MP; given a sequence $(\chi_1, \ldots, \chi_k)$ of characters and a weighting function w: $\{1, \ldots, k\} \rightarrow \mathbb{R}^{\geq 0}$ we simply replace $l(\mathcal{C}, T)$ by its weighted version $l_w(\mathcal{C}, T) = \Sigma_{i=1}^{k} w(i) l(\chi_i, T)$. We then have the following result.

Theorem 9.6.1. Suppose $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ are characters on X. Consider the model in which all characters evolve independently on a phylogenetic tree T and that each character χ_i evolves according to some Poisson model on a state space of size r_i according to admissible edge parameters that are free to vary from character to character. Then the (average) ML method ranks phylogenetic trees on X in exactly the same order as the weighted MP method provided that each character χ_i is assigned weight $\log(r_i)$.

Proof. The proof relies on a key result from Tuffley and Steel (1997): for any character $\chi: X \rightarrow R$, and any phylogenetic X-tree T' we have

$$\sup_{p'} \Pr(\chi | T', p') = r^{-l(\chi, T')} \tag{9}$$

where the supremum is over all admissible p'. Consequently,

$$\begin{aligned} L_{\text{ncm}}(T' | \mathcal{C}) &= \prod_{i=1}^{k} r_i^{-l(\chi_i, T')} \\ &= \exp\left(-\sum_{i=1}^{k} \log(r_i) l(\chi_i, T')\right) \\ &= \exp\left(-l_w(\mathcal{C}, T')\right) \end{aligned}$$

where w is the character weight function defined by $w(i) = \log(r_i)$. Consequently the tree(s) T' that maximize $L_{\text{ncm}}(T'|\mathcal{C})$ are precisely the tree(s) that minimize $l_w(\mathcal{C}, T)$, as claimed.

Note that if the size (r_i) of the state space for character χ_i is unknown for some or all values of i, then in an ML framework we might optimize these variables (r_i) subject to the obvious constraint that $r_i \geq |\chi_i(X)|$. In that case Theorem 9.6.1 holds if we replace the character weight $\log(r_i)$ by $\log(|\chi_i(X)|)$.

9.6.2 Link 2: large state space

In this section, we describe a quite different link between MP and ML. In contrast to the aforementioned link we consider here the 'common mechanism' setting for which the two methods are in general quite different, since they may select different trees (Felsenstein 1973). However when the number of states is sufficiently large, then once again ML trees are always MP trees. As we will see this may be relevant to the use of certain genomic data (such as gene order) for inferring phylogenies, as in this case the underlying state space may be very large. The proof of the following result—which also relies on the identity (9)—can be found in Steel and Penny (2004).

Theorem 9.6.2. Suppose $\mathcal{C} = (\chi_1, \chi_2, \ldots, \chi_k)$ is a sequence of k characters on X over a state space R of size $r \geq 4^{nk}$. Under the model in which the characters evolve independently according to the same Poisson model on R, any ML tree for $\mathcal{C}$ is an MP tree for $\mathcal{C}$.

9.6.3 Link 3: dense sampling of sequences

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ be a collection of aligned sequences of length k on $r \geq 2$ states. Equivalently, we may view $\mathcal{S}$ as a sequence $\mathcal{C}_{\mathcal{S}} = (\chi_1, \ldots, \chi_k)$ where χ_i is an r-state character on X. If we write S_i as $S_i(1), \ldots S_i(k)$, then $S_i(l) = \chi_l(i)$ for all $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, k\}$. Let d_H denote the Hamming metric on $\mathcal{S}$, defined by setting $d_H(S_i, S_j) = |\{l: S_i(l) \neq S_j(l)\}|$. We will suppose that the sequences in $\mathcal{S}$ are distinct: that is, $d_H(S_i, S_j) > 0$ for all $i \neq j$. Let $G_{\mathcal{S}}$ be the graph with vertex set $\mathcal{S}$ and with an edge connecting any two sequences that differ in exactly one coordinate. Equivalently, $G_{\mathcal{S}} = (\mathcal{S}, E)$ where

$$E = \{(S_i, S_j) : d_H(S_i, S_j) = 1\}$$

In the context of molecular genetics, $G_{\mathcal{S}}$ is the 'haplotype graph' described, for example, in

Excoffier and Smouse (1994). We say that $\mathcal{S}$ is *ample* if $G_{\mathcal{S}}$ is connected. It is easily shown that if $\mathcal{S}$ is an ample collection of sequences then the set of spanning trees of $G_{\mathcal{S}}$ (i.e. the trees in $G_{\mathcal{S}}$ on vertex set $\mathcal{S}$) is precisely the set of irreducible MP trees for $\mathcal{C}_{\mathcal{S}}$. Consequently, $\mathcal{C}_{\mathcal{S}}$ has MP score $n-1$.

Theorem 9.6.3 below implies that when $\mathcal{S}$ is ample, then any spanning tree for $\mathcal{C}_{\mathcal{S}}$ is also an MPL tree for $\mathcal{C}_{\mathcal{S}}$ under this model. That is, we cannot improve the MPL score by introducing additional 'Steiner points' (hypothetical ancestral sequences). As an aside, this result provides another case where a particular instance of an NP-hard problem (namely that described by Addario-Berry *et al.* 2004) has a simple, polynomial-time solution. We note also that the Buneman complex (Buneman 1971) or, equivalently, the median network Bandelt *et al.* (1995) of a collection of X-splits provides natural examples of ample sets of sequences. The proof of the following result can be found in Steel and Penny (2004).

Theorem 9.6.3. Suppose that $\mathcal{S}$ is ample. Then, under the model in which the characters evolve independently under the same Poisson model on R, the MP trees and the MPL trees for $\mathcal{C}_{\mathcal{S}}$ coincide. Furthermore, the MPL value is given by

$$L_{\mathrm{mp}}(T|\mathcal{C}_{\mathcal{S}}) = \left[\frac{1}{k(r-1)}(1-\frac{1}{k})^{k-1}\right]^{n-1}$$

where k is the length of the sequences, and r is the size of the state space.

9.7 More general models; the probability of homoplasy-free evolution

In this section we investigate a more general class of Markov processes than the simple Poisson model. For these models we ask the question of how likely it is that a character has evolved without homoplasy. This question has been investigated for the two-state Poisson model (and pairs of taxa) by Chang and Kim (1996). Here we consider more general processes on a larger state space, and for many taxa. Consequently we obtain bounds rather than the exact expressions that are possible in the simpler setting of Chang and Kim (1996).

To introduce the more general class of Markov processes, we note that many processes involving simple reversible models of change can be modeled by a random walk on a regular graph. To explain this connection, suppose there are certain 'elementary moves' that can transform each state into some 'neighboring' states. In this way we can construct a graph from the state space, by placing an edge between state α and state β precisely if it is possible to go from either state to the other in one elementary move. The graph so obtained is said to be *regular*, or more specifically *d-regular* if each state is adjacent to the same number d of neighboring states.

For example, aligned sequences of length N under the r-state Poisson model can be regarded as a random walk on the set of all sequences of length N over R; here an elementary move involves changing the state at any one position to some other state (chosen uniformly at random from the remaining $r-1$ states). Thus the associated graph has r^N vertices and it is $N(r-1)$-regular.

As another example, consider a simple model of (unsigned) genome rearrangement where the state space consists of all permutations of length N (corresponding to the order of genes $1, \ldots, N$) and an elementary move consists of an inversion of the order of the elements of the permutation between positions i and j, where this pair is chosen uniformly at random from all such pairs between $\{1, \ldots, N\}$. In this case the state space has size $N!$ and the graph is d-regular for $d = \binom{N}{2}$.

Both of the graphs we have just described have more structure than mere d-regularity. To describe this we recall the concept of a Cayley graph. Suppose we have a (non-abelian or abelian) group $\mathcal{G}$ together with a subset $\mathcal{S}$ of elements of $\mathcal{G}$, with the properties that $1_{\mathcal{G}} \notin \mathcal{S}$ and $s \in \mathcal{S} \Rightarrow s^{-1} \in \mathcal{S}$. Then the *Cayley graph* associated with the pair $(\mathcal{G}, \mathcal{S})$ has vertex set $\mathcal{G}$ and an edge connecting g and g' whenever there exists some element $s \in \mathcal{S}$ for which $g = g' \cdot s$. To recover the above graph on aligned sequences of length N over an r-letter alphabet, we may take $\mathcal{G}$ as the (abelian) group $(\mathbb{Z}_r)^N$ and the set $\mathcal{S}$ of all N-tuples that are the identity element of $\mathbb{Z}_r$ except on one coordinate. To recover the graph described above for unsigned genome rearrangements we may take $\mathcal{G}$ to be the

(non-abelian) symmetric group on N letters and $\mathcal{S}$ to be the elements corresponding to inversions.

The demonstration that such graphs are Cayley graphs has an important consequence: it implies that they also have the following property. A graph $\mathcal{G}$ is said to be *vertex-transitive* if, for any two vertices u and v there is an automorphism of $\mathcal{G}$ that maps u to v. Informally, a graph is vertex-transitive if it "looks the same, regardless of which vertex one is standing at." Clearly a (finite) vertex-transitive graph must be d-regular for some d, and it is an easy and standard exercise to show that every Cayley graph is vertex-transitive (however not every vertex-transitive graph is a Cayley graph, and not every regular graph is vertex-transitive). Thus, there are three properly nested classes of graphs:

$$\text{Cayley graphs} \subset \text{vertex-transitive graphs} \subset \text{regular graphs}$$

Given a connected graph $\mathcal{G}$ a (simple) *random walk on a graph* is a walk on the vertices of $\mathcal{G}$ that, from any given position, selects as its next state one of the neighboring vertices (selected uniformly at random). This random process forms a reversible Markov chain. The proof of the following result is given in Appendix 9.1.

Lemma 9.7.1. Suppose $W_0, W_1, \ldots$ is a random walk on a d-regular graph G. Then, for any two distinct vertices u, v, and any $n \geq 0$,

$$\Pr(W_n = v | W_0 = u) \leq \frac{1}{d} \tag{10}$$

Furthermore, if G is vertex-transitive then

$$\Pr(W_n = u | W_0 = u) = \Pr(W_n = v | W_0 = v) \tag{11}$$

Consider now a continuous-time Markov process $(X_t; t \geq 0)$ on a finite state space R, and with rate matrix Q. Thus, for any two distinct states α, β, $Q_{\alpha\beta}$ is the instantaneous rate at which state α changes to state β. Suppose that for some fixed positive integer d and some fixed positive real number q we have the following property: for each state $\alpha \in R$ there is some neighborhood $N(\alpha) \subseteq R - \{\alpha\}$ of size d for which, for all $\beta \neq \alpha$ we have

$$Q_{\alpha\beta} = \begin{cases} q, & \text{if } \beta \in N(\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Associated with any such process there is a corresponding graph with vertex set R and where the edge set E is defined by $E = \{\{\alpha,\beta\}: Q_{\alpha\beta} \neq 0, \alpha \neq \beta\}$. Note that this graph is d-regular, and substitution events under a model satisfying (12) corresponds to a random walk on the associated graph. Accordingly we will call any continuous-time Markov process that satisfies (12) a *d-regular walk process*. The equilibrium distribution of any such process is uniform.

Lemma 9.7.2. Let $(X_t; t \geq 0)$ be a d-regular walk process. Then, for any two distinct states α, β, and any values s, $t \geq 0$,

$$\Pr(X_{t+s} = \beta | X_t = \alpha) \leq \frac{1}{d}$$

Proof. For this Markov process, consider the associated graph (R, E). Let M denote the random number of transitions between states, during the interval between time t and $t+s$. Then $\Pr(X_{t+s} = \beta \mid X_t = \alpha)$ can be written as

$$\sum_{m \geq 0} Pr(X_{t+s} = \beta | M = m, X_t = \alpha) \times \Pr(M = m | X_t = \alpha). \tag{13}$$

Now, $\Pr(X_{t+s} = \beta \mid M = m, X_t = \alpha)$ is precisely the probability that for a random walk W_n on the graph (R, E) we have $W_m = \beta$ conditional on $W_0 = \alpha$, and by Lemma 9.7.1 this is at most $\frac{1}{d}$. Applying this to the expression for $\Pr(X_{t+s} = \beta \mid X_t = \alpha)$ given by (13) completes the proof.

The following result shows that for such a Markov process if d is much larger than $2n^2$ (the number of species) then any character generated on a tree with n species will almost certainly be homoplasy-free on that tree.

Proposition 9.7.3. Suppose characters evolve on a phylogenetic tree T according to a d-regular walk process. Let $p(T)$ denote the probability that the resulting randomly generated character χ is homoplasy-free on T. Then

$$p(T) \geq 1 - \frac{(2n-3)(n-1)}{d}$$

where $n = |X|$.

Proof. Consider a general Markov process on T with state space R. Suppose that for each arc (u, v) of T and each pair α, β of distinct states in R, the conditional probability that state β occurs at v given that α occurs at u is at most p. Then, from Proposition 7.1 of Semple and Steel (2003) we have $p(T) \geq 1 - (2n-3)(n-1)p$. By Lemma 9.7.2 we may take $p = \frac{1}{d}$. The result now follows.

As an example to illustrate Propostion 9.7.3 consider the simple model for random inversions of (unsigned) gene orders mentioned above. If we have L genes then $d = \binom{L}{2}$ and so if we have (say) $n = 10$ genomes each consisting of the same set of $L = 100$ (unsigned) genes that have evolved on a phylogenetic tree, the probability that this character is homoplasy-free on that tree is at least 0.97.

9.8 Results for infinite and large state spaces

Finally, we turn to the question of how many characters we need to reconstruct a large tree if the characters evolve under a Markov model on a large state space.

Markov models for genome rearrangement such as the (generalized) Nadeau–Taylor model (Nadeau and Taylor 1984; Moret *et al.* 2002) confer a high probability that any given character generated is homoplasy-free on the underlying tree, provided the number of genes is sufficiently large relative to $|X|$ (Semple and Steel 2002). In this setting the appropriate limiting model is to assume that every time a substitution occurs a completely new and unique state arises: such a model may be viewed as the phylogenetic analogue of what is known in population genetics as the 'infinite alleles model' of Kimura and Crow (1964).

Mossel and Steel (2004a) recently investigated such a 'random cluster' model on a phylogenetic tree T, which operates as follows. For each edge e let us independently either cut this edge—with probability $p(e)$—or leave it intact. The resulting disconnected graph (forest) G partitions the vertex set $V(T)$ of T into non-empty sets according to the equivalence relation that $u \sim v$ if u and v are in the same component of G. This model thus generates random partitions of $V(T)$, and thereby of X by connectivity, and we will refer to these partitions as $\overline{\chi}$ and χ, respectively. For an element $x \in X$ we will let $\chi(x)$ denote the equivalence class containing x. We call the resulting probability distribution on partitions of X the *random cluster model* with parameters (T, p) where p is the map $e \mapsto p(e)$. A central result from Mossel and Steel (2004d) was that the number of characters required to correctly reconstruct a fully resolved phylogenetic tree with n leaves grows (with n) at the rate $\log(n)$ provided upper and lower limits to p are specified (and the upper limit is less than 0.5). More precisely, let us suppose for the rest of this section that each value $p(e)$ lies between a value $p_{\min}$ and value $p_{\max}$ where $0 < p_{\min} \leq p_{\max} < 0.5$.

For this model Mossel and Steel (2004d) established the following result: if one independently generates at least

$$\frac{2}{\beta}\log\left(\frac{n}{\sqrt{\epsilon}}\right) \tag{14}$$

characters under this model, where

$$\beta = p_{\min}\left(\frac{1-2p_{\max}}{1-p_{\max}}\right)^4 \tag{15}$$

then with probability at least $1-\epsilon$, T is the only phylogenetic tree on which the characters are homoplasy-free; furthermore T can be reconstructed from the characters in polynomial time (simulations conducted by Dezulian and Steel (2004) show that even fewer characters may suffice for accurate tree reconstruction than (14) requires, although a logarithmic dependence on n is still provably necessary).

We now provide a similar result for certain regular walk processes on a finite state space. We will show that for a subclass of d-regular walk processes, and provided d grows at least as fast as $n^2\log(n)$ (where n is the number of leaves of T), then we can generate enough homoplasy-free characters to reconstruct T correctly.

First we describe a subclass of regular walk processes. Suppose that R is a group, and for some subset S (closed under inverses and not containing the identity element of R) we have $Q_{\alpha\beta} = q$ if and only if there exists some element $s \in S$ for which $\beta = \alpha \cdot s$, otherwise for any distinct pair α, β we have $Q_{\alpha\beta} = 0$. Such a process we will call a *group*

walk process (on the generating set S). Clearly a group walk process is a regular process, and the graph (R, E) associated with the regular walk process is the Cayley graph for the pair (R, S). Random walk processes have a further useful property on trees: for each arc $e=(u,v)$ of $T=(V,E)$ consider the event $\Delta(e)$ that the state that occurs at v is different from the state that occurs at u (i.e. there has been a net transition across the edge). By Lemma 9.7.1 (and the fact that the Cayley graph for (R,S) is vertex transitive), it follows that the events $(\Delta(e), e\in E)$ are independent. Let $p'_{min}=\min\{\Pr(\Delta(e)): e\in E\}$, $p'_{\max}=\max\{\Pr(\Delta(e)): e\in E\}$, and for any $\epsilon>0$ let

$$c_\epsilon = \frac{1+\log(\frac{1}{\sqrt{\epsilon}})}{\beta'\epsilon} \tag{16}$$

where $\beta' = p'_{\min}\left(\frac{1-2p'_{\max}}{1-p'_{\max}}\right)^4$.

We are now ready to state a result for certain Markov processes on large (but finite!) state spaces, which brings together several ideas presented above. Informally, Theorem 9.8.1 states that, for a group walk process, a growth of around $n^2\log(n)$ in the size of the generating set is sufficient (with all else held constant) for producing a sequence of homoplasy-free characters that define T.

Theorem 9.8.1. Suppose characters evolve independently on a fully resolved phylogenetic tree T according to a group walk process on a generating set of size d, where

$$d \geq c_\epsilon \cdot n^2 \log(n)$$

with c_ϵ given by (16) and with $p_{\max} < \frac{1}{2}$. Then with probability at least $1-2\epsilon$ we can correctly reconstruct the topology of T by generating $\lceil \frac{2}{\beta'}\log(\frac{n}{\sqrt{\epsilon}})\rceil$ characters and applying a method such as MP or maximum compatibility.

As an example, consider the group walk process for (unsigned) gene-order reversal mentioned earlier. In this case, for L genes, we have $d=\binom{L}{2}$. Theorem 9.8.1 shows that provided L grows at the rate (with n) at least some constant times $n\sqrt{\log(n)}$ then one can hope to recover fully resolved phylogenetic trees with n leaves from a (logarithmic with n) number of such independent gene-order characters.

Outline of the proof of Theorem 9.8.1. A detailed proof of Theorem 9.8.1 can be found in Mossel and Steel (2004b). Here we simply outline the argument and indicate how it depends on earlier results.

Generate $k=\left\lceil \frac{2}{\beta'}\log(\frac{n}{\sqrt{\epsilon}})\right\rceil$ characters under a group walk process satisfying condition (12) on a rooted phylogenetic tree. Consider the event H that all of these characters are homoplasy-free on T. Since a group walk process is a regular walk process, satisfying (12), using Proposition 9.7.3 it can be shown that $\mathcal{P}[H] \geq 1-\epsilon$. Furthermore the probability that T will be correctly reconstructed (using MP or maximum compatibility) from k characters produced by a coupled random cluster model (with $\beta=\beta'$) is at least $1-\epsilon$ by (14) (recalling that $p_{\max} < \frac{1}{2}$). Now, the original k characters induce the same partitions as the coupled random cluster characters whenever event H holds, and $\mathcal{P}[H] \geq 1-\epsilon$. Consequently, by the Bonferroni inequality, the joint probability that event H holds and that the k characters produced by the coupled process recover T is at least $1-2\epsilon$. Thus the probability that the original k characters recover T is at least this joint probability, and so at least $1-2\epsilon$, as claimed.

We end this section by noting that a related result—namely the statistical consistency of MP for certain Markov processes on a sufficiently large state space—was established in Steel and Penny (2000). The main difference between that result and Theorem 9.8.1 is that statistical consistency is a limiting statement; it says that as the number of characters becomes large, the probability of recovering the correct tree converges to 1. Theorem 9.8.1 meanwhile provides an explicit bound on the probability of correctly reconstructing the correct tree from a certain given number of characters.

9.9 Concluding comments

MP has continued to provide mathematicians with a rich variety of problems for study. Often these problems have led to elegant and surprising solutions, including the bichromatic binary tree theorem (Carter *et al.* 1990; Erdős and Székely 1993; Steel 1993), the min-max theorem of Erdős

and Székely (1992), and the guaranteed embedding of MP trees in median networks due to Bandelt *et al.* (1995). In this chapter we have considered further problems, particularly those concerning the statistical aspects of applying MP to character data on a large state space, and for which some solutions have been proposed. However the reader would be wrong to conclude that MP for even two-state character data is completely understood. Indeed the following problem is still open: under the two-state Poisson process is there a value $p > 0$ so that MP is statistically consistent for all fully resolved trees (having any number of leaves) under the constraint $p(e) = p$ for all edges of the tree? The fact that such a basic question is still open suggests there still await challenges for investigators in future.

9.10 Acknowledgments

We thank the New Zealand Marsden Fund and the New Zealand Institute for Mathematics and its Applications (NZIMA) for supporting this research. We also thank Andrew Hugall for posing a question that led to Theorem 9.6.1, and Joseph Felsenstein, Michael Sanderson, and Cécile Ané for helpful comments on an earlier version of this chapter.

Appendix 9.1 Proof of Proposition 9.4.4, and Lemma 9.7.1

Proof of Proposition 9.4.4. Let $X = \{0,1\}^k$ and let $\mathcal{C} = \{A_i|B_i, i = 1, \ldots k\}$ where $A_i := \{x \in X : x_i = 1\}$ and $B_i = X - A_i$. We claim that $\mathcal{C}$ has the property described. To this end, suppose that $A \mid B$ is an X-split that is compatible with every character in $\mathcal{C}$. Let $\mathbf{1} = (1, 1, \ldots, 1) \in X$. Without loss of generality (by interchanging A and B, as well as A_i and B_i if necessary) we may suppose that $\mathbf{1} \in A$ and, for each i, $\mathbf{1} \in A_i$. Note that, by definition, $|A_i| = 2^{k-1}$; also we have $A_i \cap A \neq \emptyset$ for all i. Thus the compatibility of $A \mid B$ with $A_i \mid B_i$ ensures that

$$\text{for each } i \text{ either} A_i \subseteq A \text{ or } A \subseteq A_i \text{ or } B_i \subseteq A \qquad (17)$$

We then consider two cases:

(*i*) $|A| < 2^{k-1}$

(*ii*) $|A| \geq 2^{k-1}$

In case (i) condition (17) and the equality $|A_i| = 2^{k-1}$ ensures $A \subseteq A_i$ for all i. But this means that $A = \{(1,1,\ldots,1)\}$ and so $A \mid B$ is a trivial character. In case (ii) condition (17) and the equality $|A_i| = 2^{k-1}$ ensures that for each i either $A_i \subseteq A$ or $B_i \subseteq A$; in the first case we will let $y_i = 0$ and in the second case we will let $y_i = 1$. Let $y = (y_1, \ldots, y_k)$. Then $A = X - \{y\}$ and so again $A \mid B$ is a trivial character.

Proof of Lemma 9.7.1. We prove the first claim by induction on n. The result trivially holds for $n = 0$, and for $n = 1$ we have $\Pr(W_1 = v | W_0 = u) \in \{0, \frac{1}{d}\}$ since the graph is d-regular, and so (10) holds. Suppose (10) holds for $n = k$. Then by the elementary theory of Markov chains,

$$\Pr(W_{k+1} = v | W_0 = u) = \sum_w \Pr(W_1 = w | W_0 = u) \Pr(W_k = v | W_0 = w) \qquad (18)$$

Letting $N(u)$ denote the set of vertices that neighbor u the right-hand term in (18) is

$$\frac{1}{d} \sum_{w \in N(u)} \Pr(W_k = v | W_0 = w) = \frac{1}{d} \sum_{w \in N(u)} \Pr(W_k = w | W_0 = v) \qquad (19)$$

where the equality in (19) arises since the chain-transition matrix is symmetric and so $\Pr(W_k = v \mid W_0 = w) = \Pr(W_k = w \mid W_0 = v)$. Combining (18) and (19) we have

$$\Pr(W_{k+1} = v | W_0 = u) = \frac{1}{d} \sum_{w \in N(u)} \Pr(W_k = w | W_0 = u) \leq \frac{1}{d}$$

so that (10) holds for $n = k+1$, establishing the induction step and thereby the lemma.

The proof of (11) in Lemma 9.7.1 is similar but easier.