

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**DISTRIBUTIONS ON BICOLOURED
EVOLUTIONARY TREES**

A thesis presented in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
in Mathematics at
Massey University

Michael Anthony Steel
February, 1989

Massey University Library

Thesis Copyright Form

Title of thesis: DISTRIBUTIONS ON BICOLOURED
EVOLUTIONARY TREES

- (1) (a) I give permission for my thesis to be made available to readers in the Massey University Library under conditions determined by the Librarian.
- (b) ~~I do not wish my thesis to be made available to readers without my written consent for _____ months.~~
- (2) (a) I agree that my thesis, or a copy, may be sent to another institution under conditions determined by the Librarian.
- (b) ~~I do not wish my thesis, or a copy, to be sent to another institution without my written consent for _____ months.~~
- (3) (a) I agree that my thesis may be copied for Library use.
- (b) ~~I do not wish my thesis to be copied for Library use for _____ months.~~

Signed

M. A. Steel

Date

1/3/89

The copyright of this thesis belongs to the author. Readers must sign their name in the space below to show that they recognise this. They are asked to add their permanent address.

NAME AND ADDRESS

M. A. Steel

c/- P&L Read No 1 Le Rd
Palmerston North NZ

DATE

1/3/89

ABSTRACT

A central and challenging problem in contemporary biology is how to accurately reconstruct evolutionary trees from DNA sequence data. This thesis addresses three themes from this endeavour -- comparison, consistency and confidence intervals -- by analysing distributions arising from phylogenetic trees.

Toward the first theme, the distribution of the symmetric difference metric on pairs of binary and phylogenetic trees is studied, and a number of new results obtained. These theorems, as well as a result on another tree metric answer previous conjectures in this area. Also under the theme of comparison, we analyse distributions on bicoloured trees arising from the principle of parsimony. A streamlined proof is given of an elegant theorem which allows an efficient comparison of how much better a maximum parsimony tree fits given data than a randomly-chosen tree. A dual distribution, where the tree is fixed and the data varies is also analysed, answering a recent unsolved problem.

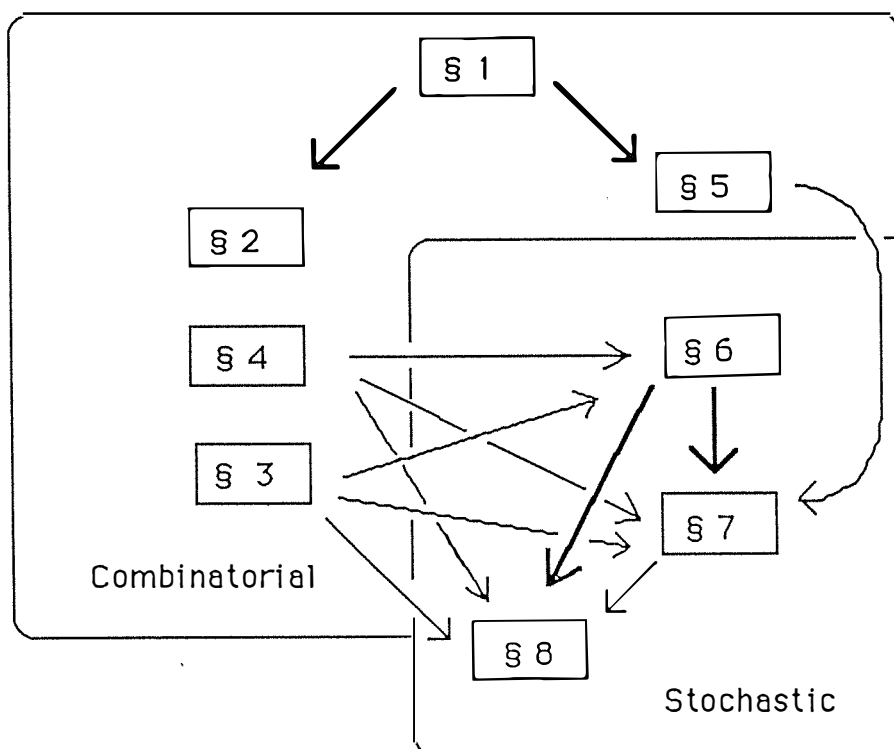
We then consider the theoretical accuracy of tree-building methods, concentrating on the statistical property of consistency. Under a simple stochastic model on bicoloured trees, conditions for the consistency of frequently-used methods based on parsimony and compatibility are examined. It is shown that even in "best possible" conditions both methods can be inconsistent, though a strong sufficient condition for compatibility is given. The analysis is extended for a molecular clock.

Finally, procedures are described for placing confidence intervals around phylogenies, and limitations on the sort of confidence intervals possible are given. Ways to efficiently implement these procedures are then considered -- in particular, approximate methods, applications to sets of taxa of size four, and simplifications under a molecular clock.

The rate that sequence data must grow as a function of the number of taxa for confidence intervals to converge to a single tree is also considered.

The arguments in this thesis are primarily combinatorial and stochastic.

In the hope that their implications will also interest biologists, some space has been given to motivating and explaining the biological relevance of the results presented.



**Chart illustrating the flow of results
from one section to another**

Contents:	Page:
Introduction	1
Notation/Table of Symbols	3
 Section One: Counting Trees	
--trees	5
--resolved subtrees	11
--forests	15
--bicoloured trees	18
 Section Two:	
Distribution of the Symmetric Difference Metric	
--distribution generating functions	29
--distribution on pairs of trees	34
--absolute (non-asymptotic) inequalities	38
--asymptotic range of the distribution	43
--monotonicity	46
--description of the metric from below	48
--distribution on $PT(n)$	51
--comparison with other metrics	54
 Section Three: Subtree constraints	
--induced subtrees and minimal similarity	56
--spanning sets	62
--consensus trees	66
--efficiency	68
 Section Four: Sequence and Dissimilarity Data	
--constraints with two colours	73
--lower bounds on information loss	76
 Section Five: Combinatorics of Parsimony	
--vector spaces of edge sets	82
--(weakly-) connecting trees and forests	84
--distributions arising from parsimony	90
--path/edge duality	100

Section Six: Analysis of Cavender's model

--stochastic preliminaries	108
--the model	109
--central observations	111
--invariants	113
--partition probabilities	118

Section Seven: Consistency

--selection procedures	127
--convergence and consistency	127
--consistent recovery of trees from dissimilarities	132
--consistency of parsimony and compatibility	135
--sufficient conditions	147
--consistency under a molecular clock	160

Section Eight: Confidence Intervals

--confidence intervals	168
--approximate methods	172
--exact solutions	176
--efficiency (I)	178
--efficiency (II)	180
--a χ^2 test (molecular clock)	186

Appendix	193
-----------------	-----

References	196
-------------------	-----

Introduction

"There is one unique, true hierarchical nesting pattern in nature, waiting to be discovered. All that we have to do is develop methods for discovering it," Richard Dawkins.

The last two decades have witnessed a flourishing of methods aimed at constructing phylogenetic trees from genetic data. Yet despite the extensive use of these procedures relatively little attention has been paid to determining their accuracy. Consequently, disputes on how best to reconstruct evolutionary trees have arisen (see for example Diamond [1988]) and different data or different methods often give differing trees, as the introduction of Felsenstein [1987], concerning the *Pan-Homo-Gorilla* clade, makes clear. It is thus desirable to determine how these methods perform under simple models which describe how variations in the underlying genetic data arise on the evolutionary tree linking the taxa in question.

This thesis considers these problems from a combinatorial and stochastic viewpoint. Section one enumerates various classes and properties of phylogenetic trees, and presents a solution of an open problem by Day [1986]. This is followed by a streamlined proof of an important new theorem from Carter *et al.* [1988], which is applied in section five.

Section two examines the question of how similar one would expect "randomly-chosen" trees to be, so as to better understand the significance of disparities in tree reconstruction, mentioned above. This continues work in Hendy, Little and Penny [1984], and Day [1983] and settles two conjectures raised in the former paper, as well as providing a deeper understanding of the metric used in that paper.

We then examine from a combinatorial viewpoint three issues related to tree building: section three considers how the structure of subtrees constrains the structure of the parent tree(s); section four quantifies the loss of information in working with dissimilarity data rather than sequence data, and section five examines combinatorial aspects of parsimony, which is the principle behind the most widely-used tree building method. In this section we give an efficient method for measuring how much better the maximum parsimony tree fits data than a "randomly-chosen" tree. A useful invariance result is also derived and answers a dual question of how many edge changes are required to fit "random" data to a tree.

These and other results are applied in sections six, seven and eight to address the question of the theoretical accuracy of tree-building methods under a simple model proposed in 1978 by James Cavender. This model and its immediate consequences are outlined in section six. In section seven, the statistical consistency of various classes of tree-building methods is examined. Particular attention is paid to parsimony and a closely related method, compatibility, and new necessary and sufficient conditions for consistency are obtained. In section eight the question of how to find confidence intervals around trees is investigated. This question has received relatively little attention as suggested in the recent and comprehensive review of tree-building methods by Felsenstein [1988]. Finally the rate of convergence (to a single tree) of these confidence intervals is examined and a positive result obtained.

For the first half of this thesis, the arguments are mostly combinatorial, relying on generating functions and tree decompositions, combined with two technical results for which the reader is referred to Goulden and Jackson [1983]--the principle of inclusion and exclusion (for generating functions), and the Lagrange inversion formulae. Where exact solutions appear intractable or difficult, asymptotic methods have been used. From section six, probability-based arguments are exploited, particularly properties of the multinomial distribution.

Acknowledgement:

I would like to thank Mike Hendy and David Penny for their valuable suggestions and guidance in the preparation of this thesis. I also wish to thank the following for their helpful comments:

Michael Carter, Joe Felsenstein, Ron Graham, Charles H. Little, Terry Moore and Ingrid Rinsma.

Notation:

$[x^n]f(x)$ denotes the coefficient of x^n in $f(x)$. Similarly for

$[x_1^{n_1} \dots x_k^{n_k}]f(x_1, \dots, x_k)$.

$f'(x)$ denotes the (formal) derivative of $f(x)$.

$f(n) = O(g(n))$ means $f(n)/g(n)$ is bounded as $n \rightarrow \infty$.

$f(n) \sim g(n)$ means $\lim_{n \rightarrow \infty} f(n)/g(n) = 1$.

$\approx$ means approximately equal to.

For a set X , $[X]_k$ is the set of subsets of X of size k .

$[n]_k = [\{1, \dots, n\}]_k$, $[n]^+$ is the set of even subsets of $\{1, \dots, n\}$ (including $\emptyset$).

nC_k denotes the binomial coefficient $n!/(k!(n-k)!)$

$x, X, \dots$ denotes vectors; $j = [1, 1, \dots, 1]^t$.

J denotes a square matrix with all entries $+1$.

$x \cdot y$ is the inner product of x and y

$x > y$ (resp. $x \geq y$) means $x_i > y_i$ (resp. $x_i \geq y_i$) for all i .

$x^y = \prod_i x_i^{y_i}$, $\|x - y\|$ is the Euclidean distance between x and y .

$\mathbb{R}^k$, $(\mathbb{R}^+)^k$, $\mathbb{N}^k$ denotes respectively k -tuples of reals, positive reals and nonnegative integers (when $k=1$ the superscript is suppressed).

$P(A)$ denotes the probability of event A .

$E[X]$, $\text{Var}[X]$ denote the expectation and variance of random variable X .

$(\alpha_1 \alpha_2)(\alpha_3, \dots, \alpha_{k-2})(\alpha_{k-1}, \alpha_k)$ denotes a binary caterpillar tree, J_k

endpoint labelled as shown in fig. 0.1.

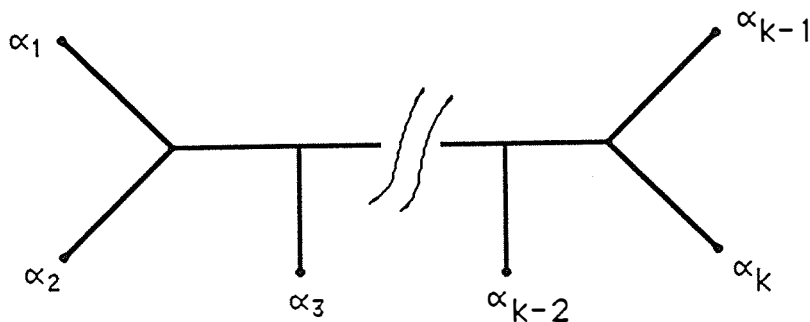


Figure 0.1

Table of Symbols	Page		Page
<u>Trees</u>		<u>Stochastic</u>	
$IE(T), IV(T), E(T)$	5	$s(T,p), s(\sigma; T,p)$	110
$BPT(L), PT(L)$	6	$s(T \mid q), s(\sigma; T \mid p)$	110
$BPT(n), PT(n), PT(n,f)$	6	$r(T,p)$	116
$b(n), p(n)$	6	$s^{\wedge}(T,p)$	133
$bpt(n), pt(n)$	8	$S(k), D(k)$	139
$RPT(n), RBT(n)$	9	$P(k), Q(k)$	142
$q(s,T), q(s,n), Q(T,x), q_k(T), P(T,x)$	29	BCI, CCI	169
T/e	30		
$BS(n,a)$	43		
$T[S]$	56		
$w(\sigma,T), w(D,T)$	91	<u>Combinatorial</u>	
$[E, E_0]$	100	$\langle Q \rangle$	62
r_X	100	$\psi(T)$	62
$\pi(T,X)$	100	δ	73
$WBT(n)$	109	X_{σ}, X	73
$\Omega(T), \Omega I(T), \Omega P(T)$	135	$\sigma, \sigma_0, \Omega(n)$	73
$T(k)$	136	$\chi, C(n)$	73
$RWT(n)$	138	$\Delta_f(n,c)$	79
$T_1(k,p), T_1(k), T_2(k)$	143	$p(C)$	82
$B(\lambda,L), B(\lambda), B(n,\lambda,L)$	148	$X[S], p[S], q[S]$	112
$BE(\lambda), BH(\lambda)$	160	φ	127
$RT_1(k,\lambda)$	161	A	172
		$\langle Q \rangle_{res}, \delta(Q)$	178

§1: COUNTING TREES

"We have for twenty species more than a gram molecular weight of evolutionary trees," Walter Fitch.

1.1 Introduction

A central problem involved in constructing phylogenies is the rate of growth in the number of potential candidates as a function of the number of taxa. We begin therefore by defining and enumerating the types of trees suitable for phylogenetic analysis. The asymptotic distribution of certain properties of these classes is then derived, solving an open problem posed by Day [1986] and providing results which are used in section two. The enumeration and properties of forests of phylogenetic trees is also considered and compared with classical results. We then consider the enumeration of binary trees according to the number of edge changes required to fit a colouring to the endpoints. The results in section one rely on tree decompositions, and in some cases, the Lagrange inversion formula.

Trees

1.2 Definitions

A *tree* is an acyclic ^{connected} simple graph.

Let L be a set of $n \geq 2$ labels. A *phylogenetic tree on L* , is a tree with n vertices of degree one, called *pendant vertices*, each labelled with a distinct element from L , and with the remaining (internal) vertices of degree at least three, and unlabelled. For such a tree, the n edges incident with a pendant vertex are called *pendant edges*, and the remaining edges are *internal*. We let $E(T)$ (resp. $IE(T)$, $IV(T)$) denote the set of edges (resp. internal edges, internal vertices) of T .

Two phylogenetic trees on L are considered equivalent if there is a graph isomorphism between them which preserves the labelling on the pendant vertices. More generally, if two phylogenetic trees are graph isomorphic with their labellings suppressed, we say they are *topologically equivalent*.

Let $PT(L, f)$ denote the set of phylogenetic trees with f internal edges on label set L . Let $PT(L) = \cup \{PT(L, f); 0 \leq f \leq n-3\}$, and $BPT(L) = PT(L, n-3)$, the set of *binary phylogenetic trees*, for which each internal vertex has degree three. For $L = \{1, \dots, n\}$ denote these sets $PT(n, f)$, $PT(n)$ and $BPT(n)$ respectively. The following result is from Felsenstein [1978 (1), p.29].

1.3 Lemma

The size of $PT(n, f)$ is determined recursively as follows:

$$\begin{aligned} |PT(n, f)| &= (n+f-2) |PT(n-1, f-1)| + (f+1) |PT(n-1, f)|; \quad n \geq 4, \\ |PT(3, 0)| &= 1, \quad |PT(3, f)| = 0, \quad f > 0. \end{aligned}$$

For $n \geq 3$, lemma 1.3 gives

$$|BPT(n)| = (2n-5)!! = 1.3.5 \dots (2n-5) = (2n-4)! / (n-2)! 2^{n-2}$$

For convenience, we let $b(n) = |BPT(n)|$ and $p(n) = |PT(n)|$.

1.4 Corollary

Let $PT^+(n)$ (resp. $PT^-(n)$) denote the set of trees in $PT(n)$ with an even (resp. odd) number of internal edges. Then

$$|PT^+(n)| - |PT^-(n)| = (-1)^{n+1} (n-2)! \quad n \geq 3.$$

Proof: Let $p(n,x) = \sum_{f \geq 0} |PT(n,f)| x^f$. By lemma 1.3,

$$p(n,x) = ((n-1)x+1)p(n-1,x) + (x^2+x)^d / d_x p(n-1,x). \text{ Putting}$$

$x = -1$ gives $p(n,-1) = (2-n)p(n-1,-1)$, which together with $p(3,-1) = 1$,

$$\text{gives } |PT^+(n)| - |PT^-(n)| = p(n,-1) = (-1)^{n+1}(n-2)! \quad \S$$

A central theorem in the enumeration of labelled trees is the following result:

1.5 Theorem Moon [1970]

The number of trees on n labelled vertices of degrees $d(1), \dots, d(n)$ is $n^{-2} C_{d(1)-1, \dots, d(n)-1}$, if $\sum_i d(i) = 2n-2$, and 0 otherwise. $\S$

This theorem gives an alternative description of $|PT(n,f)|$ as a sum of $f+1$ terms, each expressed in terms of Stirling numbers of the second kind. (Aigner [1979] gives a table of the Stirling numbers $S(m,r)$ for $m, r \leq 8$).

1.6 Theorem

$$|PT(n,f)| = \sum_{0 \leq s \leq f} \binom{n+f-1}{s} C_s (-1)^s S(n+f-1-s, f+1-s).$$

(For example, $|PT(6,2)| = {}^7C_0 S(7,3) - {}^7C_1 S(6,2) + {}^7C_2 S(5,1) = 105$).

Proof:

Any $T \in PT(n,f)$ has $n+f+1$ vertices, n of which are labelled and of degree 1 and $f+1$ of which are unlabelled and of degree $d_1, \dots, d_{f+1} \geq 3$.

Now only the trivial automorphism leaves every endpoint of a tree fixed (Harary and Mowshowitz [1975]).

Thus if we regard this second set of vertices as labelled, apply theorem 1.5 and then "unlabel" these internal vertices we obtain:

$$\begin{aligned} |PT(n,f)| &= [(f+1)!]^{-1} \sum_{\{d \geq 3j, d.j=2f+n\}} (n+f-1)! / (d_1-1)! \dots (d_{f+1}-1)! \\ &= [(f+1)!]^{-1} \sum_{\{x=(x(1), \dots, x(f+1)) \geq j, x.j=n+f-1\}} (n+f-1)! / x(1)! \dots x(f+1)! \end{aligned}$$

by the substitutions $x(i) = d_i - 1$ for $i=1, \dots, f+1$.

Thus, by the principle of inclusion and exclusion we have:

$$\begin{aligned} |PT(n,f)| &= [(f+1)!]^{-1} \times \sum_S (-1)^S \cdot {}^{f+1}C_S \times \\ &\quad \sum_{\{x=(x(1), \dots, x(f+1-s)) \geq j, x.j=n+f-1-s\}} (n+f-1)! / x(1)! \dots x(f+1-s)! \quad (*) \end{aligned}$$

Now for positive integers m, k the summation

$\sum_{\{x=(x(1), \dots, x(k)) \geq j, x.j=m\}} m! / x(1)! \dots x(k)!$ is the number of ways of placing m labelled objects into k labelled sets, so that each set contains at least one element.

But $S(n, k)$ is the number of ways of partitioning n objects into k non-empty sets so that

$k!S(n, k) = \sum_{\{x=(x(1), \dots, x(k)) \geq j, x.j=m\}} m! / x(1)! \dots x(k)!$ as in Anderson [1974, p.58].

Thus the summation term in (*) is

$$(n+f-1)!(f+1-s)!S(n+f-1-s, f+1-s) / (n+f-1-s)!$$

Substituting this into (*) gives the result. §

Frequently in following constructions, the labelling of trees is unimportant. Thus it is useful to let $pt(n)$, $pt(n, f)$ and $bpt(n)$ denote a set of distinct representatives for the topological classes of $PT(n)$, $PT(n, f)$ and $BPT(n)$ respectively. To enumerate $bpt(n)$ we use the following result.

1.7 Lemma

Let $K(n)$ be the set of unlabelled trees with n vertices of degree ≤ 3 .

Then for $n \geq 4$, there is a bijection from $bpt(n)$ to $K(n-2)$.

Proof:

For $T \in bpt(n)$, let T' be the unlabelled tree obtained from T by deleting all its pendant vertices and pendant edges (T' is the *derived* tree of T in Hendy, Little and Penny [1984]). Since T has $n-2$ internal vertices, we have $T' \in K(n-2)$, and the process is clearly invertible. §

1.8 Corollary

$|bpt(n)| \sim b\theta^n/n^{5/2}$, where $b \approx 0.03301$, $\theta \approx 2.4832535$.

Proof:

The asymptotic value for $|K(n)|$ is derived by Otter [1948, p.597-598] as $(\beta^3 \theta^{-9/2} / 4\sqrt{\pi}) \theta^n / n^{5/2}$, with values for θ , (written as α^{-1} by Otter), and β given. Harding [1971] corrects an error in Otter's calculations to show $\beta \approx 4.4220432$. The result now follows by the previous lemma. §

1.9 Definition

For $L' = \{1, \dots, n\} \cup \{\infty\}$, let $RPT(n)$ be the set of *rooted phylogenetic trees* on n pendant vertices, defined as follows. By convention $RPT(1)$ is the tree consisting of a distinguished vertex (called a *root*) connected to a labelled pendant vertex. For $n \geq 2$, $RPT(n)$ is the set of trees obtained by distinguishing an internal vertex of $T \in PT(n)$ or adding a distinguished vertex to the midpoint of an edge of T . Define the set of *rooted binary trees* on n pendant vertices, denoted $RBT(n)$, analogously.

It is easily shown that $|RPT(n)| = p(n+1)$. Thus to count $PT(n)$ it suffices to count $RPT(n)$, for which the appropriate tool is the exponential generating function. Let x mark the number of non-root pendant vertices, and $R(x)$ denote the exponential generating function for $RPT(n)$. A standard tree decomposition gives the following result.

1.10 Proposition Foulds and Robinson [1984]

$$2R(x) = \exp(R(x)) - 1 + x.$$

Asymptotically we have the following result, where $p = 2\ln(2)-1$.

1.11 Lemma

$$(a): p(n)/n! \sim p^{1-n} n^{-5/2} \sqrt{p/4\pi}$$

$$(b): p(n)^{-1}(\sum_{f \geq 0} f |PT(n, f)|) - np^{-1}(1-\ln(2)) = O(1)$$

$$(c): p(n)^{-1}(\sum_{f \geq 0} f(f-1) |PT(n, f)|) - n^2 p^{-2}(1-\ln(2))^2 = O(n).$$

Proof:

These results follow from closely related results in Foulds and Robinson [1984], the only modification required being that the first and second moments corresponding to (b) and (c) are evaluated in that paper with respect to the total number of vertices, v , rather than the number of internal edges, f , requiring the substitutions: $f = v - (n+1)$ and $f(f-1) = v(v-1) - 2(n+1)v + n^2 + 3n + 2$. §

Resolved Subtrees

We now answer a question raised by Day [1986]. Given $T \in PT(n)$ and a subset $A = \{i, j, k, l\}$ of distinct elements of $\{1, \dots, n\}$, we say A is *resolved* by T if the vertices can be paired such that the path between the first pair is vertex disjoint from the path between the other pair.

Establishing the proportion $\tau(n)$ of unresolved quartets for a randomly-chosen tree in $PT(n)$ is of interest to taxonomists because of other results in Day's paper. This paper reduces the evaluation of the averages of various measures of similarity between randomly-chosen phylogenetic trees to the evaluation of $\tau(n)$. For example, proposition 1 (1) of Day's paper states that the expected proportion of resolved quartets on which two phylogenetic trees do not conflict is $2(1-\tau(n))^2/3$, while for binary trees the expected proportion is simply $2/3$. The next theorem shows that these two values are asymptotically equal, refuting a suggestion in Day [1986] that $1-\tau(n)$ might converge "to a value close to, but distinct from, one."

1.12 Lemma

For power series $f(x)$,

(1): $[x^k]f'(x)f''(x) = 0.5(k+1)[x^{k+1}](f'(x))^2$, where ' denotes differentiation with respect to x .

(2): if $[x^n]f(x) \sim f_0 p^{-n} n^{-0.5}$, then $[x^n]f(x)^2 \sim f_0^2 \pi p^{-n}$.

Proof:

(1): Let $g(x) = \sum_i a_i x^i = (f'(x))^2$. Then since $g'(x) = (f'(x)^2)' = 2f'(x)f''(x)$

we have $[x^k]f'(x)f''(x) = 0.5[x^k]g'(x)$. But $g'(x) = \sum_i i a_i x^{i-1}$ so that

$[x^k]g'(x) = (k+1)[x^{k+1}]g(x)$, as required.

(2): $[x^n]f(x)^2 = \sum_{0 \leq i \leq n} [x^i]f(x)[x^{n-i}]f(x)$, and a careful but straightforward analytical argument shows that asymptotically this expression is:

$$f_0^2 \rho^{-n} \sum_{1 \leq i < n} i^{-0.5} (n-i)^{-0.5} = f_0^2 \rho^{-n} n^{-1} \sum_{1 \leq i < n} 1 / \sqrt{((i/n)(1-i/n))} \sim$$

$$f_0^2 \rho^{-n} \int_{1/n \leq x \leq 1-1/n} dx / \sqrt{(x(1-x))} \sim f_0^2 \rho^{-n} \pi. \S$$

1.13 Theorem

Let $\tau(n)$ denote the expected proportion of unresolved 4-trees in a tree randomly chosen from $PT(n)$.

(Thus $\tau(5) = 25/5 \times 26 = 0.1923$, $\tau(6) = 570/15 \times 236 = 0.1610$, as in table 4 of Day [1986]).

Then with $R(x)$ and ρ as above:

$$(a): \tau(n) = (2\rho(n))^{-1} (n-3)! [x^{n-3}] (R'(x))^2$$

$$(b): \tau(n) \sim \sqrt{(\rho^\pi / 4n)}.$$

Proof:

Let $[n]_4$ be the set of all subsets of $\{1, \dots, n\}$ of size 4, called *quartets*, and let $[[n]_4]_k$ be the set of all subsets of $[n]_4$ of size k . For $Q \in [[n]_4]_k$ let $U(Q)$ denote the number of $T \in PT(n)$, such that T is unresolved on all the quartets of Q (and possibly others).

Letting U_k be the sum of $U(Q)$ over all $Q \in [[n]_4]_k$, and let $U(x) = \sum_k U_k x^k$.

By the principle of inclusion and exclusion, $E(x) = U(x-1)$ is the ordinary generating function for the number of trees having precisely a given number of quartets unresolved.

In particular, $\tau(n) = E'(1)/n_{C_4} \rho(n)$, where $'$ denotes differentiation with respect to x . Thus $\tau(n) = U'(0)/n_{C_4} \rho(n) = U_1/n_{C_4} \rho(n)$.

Now $U(Q)$ is clearly the same for any $Q \in [[n]_4]_1$, so choosing $Q_0 = \{1, 2, 3, 4\}$ we have $U_1 = {}^nC_4 U(Q_0)$. Hence $\tau(n) = U(Q_0) / p(n)$. (1)

Now suppose $T \in \text{PT}(n)$ is unresolved on Q_0 . Then there exists a unique internal vertex $v(T)$ of T linking the 6 paths joining pairs of pendant vertices with labels in Q_0 .

Furthermore, T can be represented uniquely by one of the two trees in fig. 1.1 (a),(b), where $T_i \in \text{RPT}(n_i)$, $i=1, \dots, 4$, $n_i \geq 1$, is a pendant subtree containing the vertex labelled i , and rooted at a vertex v_i adjacent to $v(T)$, and T^* is one of the following:

Case 1: $T^* = \emptyset$ (so that in case (b), v^* is a pendant vertex).

Case 2: $T^* \in \text{RPT}(k)$, $k \geq 2$, and the root is $v(T)$ for T in fig.1.1(a) and v^* in fig. 1.1(b).

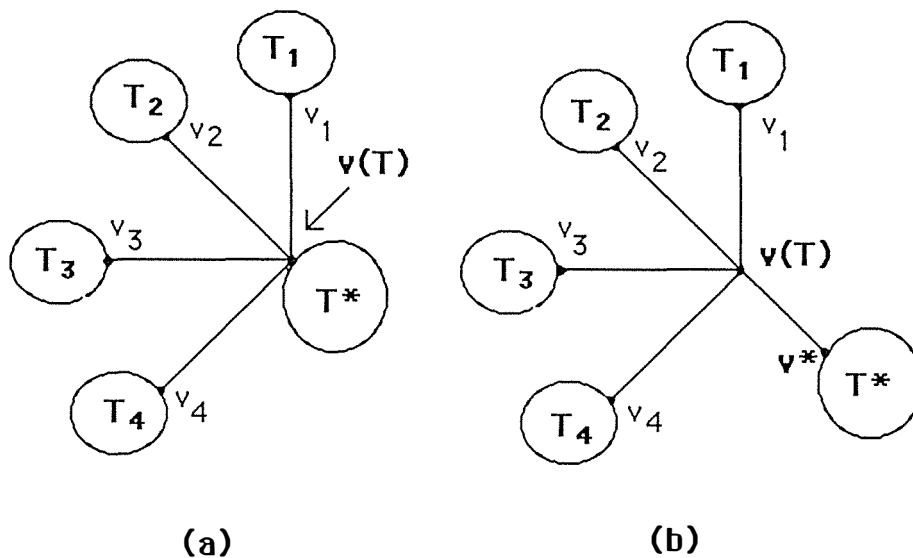


Figure 1.1

Furthermore, any $T \in \text{PT}(n)$ which is described by fig. 1.1 is clearly unresolved on Q_0 .

Using this decomposition, and letting $r(k) = |RPT(k)|$ we have,

$$U(Q_0) = \sum_{\{n \in \mathbb{N}^4: n.j=(n-4)\}}^{(n-4)} c_{n_1, n_2, n_3, n_4} \prod_i r(n_i+1)$$

(case 1, tree representation (a))

$$+ \sum_{\{n \in \mathbb{N}^4: n.j=(n-5)\}}^{(n-4)} c_{n_1, n_2, n_3, n_4, 1} \prod_i r(n_i+1)$$

(case 1, tree representation (b))

$$+ 2 \sum_{\{k, n \in \mathbb{N}^4: k \geq 2, n.j+k=(n-4)\}}^{(n-4)} c_{n_1, n_2, n_3, n_4, k} \prod_i r(n_i+1) r(k)$$

(case 2).

(In all cases i ranges from 1 to 4, $n.j = \sum_i n_i$ and $\mathbb{N} = \{0, 1, 2, \dots\}$).

Noting that $r(1) = 1$, rewrite the last two of the above three terms as

$$2 \sum_{\{k \in \mathbb{N}, n \in \mathbb{N}^4: k \geq 1, n.j+k=(n-4)\}}^{(n-4)} c_{n_1, n_2, n_3, n_4, k} \prod_i r(n_i+1) r(k) \\ - \sum_{\{n \in \mathbb{N}^4: n.j=(n-5)\}}^{(n-4)} c_{n_1, n_2, n_3, n_4, 1} \prod_i r(n_i+1) r(1)$$

$$\text{Thus } U(Q_0) = (n-4)! [x^{n-4}] (R'(x))^4 (1+2R(x)-x). \quad (2)$$

$$\text{Now by proposition 1.10, } 1+2R(x)-x = \exp(R(x)). \quad (3)$$

Differentiating this equation with respect to x gives

$$2R'(x)-1 = R'(x)\exp(R(x)), \text{ so that } R'(x) = (2-\exp(R(x)))^{-1}. \quad (4)$$

Differentiating again gives,

$$R''(x) = R'(x)\exp(R(x))/(2-\exp(R(x)))^2 = \exp(R(x))/(2-\exp(R(x)))^3 \text{ (by (4))}.$$

This together with (3) and (4) gives,

$$U(Q_0) = (n-4)! [x^{n-4}] R'(x) R''(x) = 0.5(n-3)! [x^{n-3}] (R'(x))^2 \text{ by lemma 1.12 (a)}$$

Combining this with (1) gives the required result.

(b): This now follows from part (a), lemma 1.11 and lemma 1.12 (b). §

Forests

Let $F(n,k)$, (resp. $N(n,k)$) denote the set of forests consisting of exactly k rooted phylogenetic (resp. binary) trees having exactly n pendant vertices in total. Let $F(n) = \cup_k F(n,k)$ and $N(n) = \cup_k N(n,k)$.

1.14 Theorem

- (1): $|F(n)| = 2p(n+1)$
- (2): If $f(n) = |N(n)|$, then $f(n) = (2n-3)f(n-1) + f(n-2)$.
- (3): $f(n) \sim b(n+1)e$.
- (4): $|N(n,k)| = (2n-k-1)! / (n-k)!(k-1)! 2^{n-k}$

Proof:

(1): $\exp(R(x))$ is the exponential generating function for $F(n)$, by the "logarithmic connection" (Goulden and Jackson, [p.187, 1983]). But $\exp(R(x)) = 2R(x) - 1 + x$, by proposition 1.10, which gives (1).

(2): If $B(x)$ denotes the exponential generating function for rooted binary trees having n non-root pendant vertices, then by a standard tree decomposition, $B(x) = 1/2 B^2(x) + x$, giving $B(x) = 1 - \sqrt{1-2x}$, as in Carter *et al.* [1988]. Again, $\exp(B(x))$ is the exponential generating function for $N(n)$.

Now it is easily checked that $1 + B'(x) = (1-2x)(B''(x) + B'(x)^2)$, which can be rearranged as:

$$(B'(x)^2 + B''(x)) = (4B'(x) + 2xB'(x)^2 + 2xB''(x)) - 3B'(x) + 1.$$

Multiplying by $\exp(B(x))$ and integrating twice gives:

$\exp(B(x)) = 2x\exp(B(x)) - 3 \int \exp(B(x)) + \int \int \exp(B(x))$, which translates into the recursion in (2).

(3): By the Lagrange inversion formula,

$$[t^n] \exp(B(x)) = n^{-1} [\lambda^{n-1}] \exp(\lambda) (1-\lambda/2)^{-n}$$

$$\text{Thus } f(n) = n! [t^n] \exp(B(x)) = \sum_{i \geq 0} \frac{2^{i+1-n} (2n-2-i)!}{i!(n-i-1)!}$$

$$= \frac{(2n-2)!}{(n-1)! 2^{n-1}} \sum_{i \geq 0} \frac{2^i \cdot (2n-2-i)!(n-1)!}{i!(2n-2)!(n-i-1)!}$$

Thus $\lim_{n \rightarrow \infty} f(n)/b(n+1) = \sum_{i \geq 0} (i!)^{-1} = e$, as required.

(4) is proved by Carter *et al.* [1988]. §

1.15 Theorem

Let $\mu_F(n)$ (resp. $\mu_N(n)$) denote the average number of components in $F(n)$ (resp. $N(n)$).

$$(1): \mu_F(n) = 0.5 + (4 |F(n,2)| - np(n))/2p(n+1) \sim 1 + \ln(2)$$

$$(2): \mu_N(n) \sim 3.$$

Proof:

$$(1): \text{Let } F(x,y) = \sum_{n,k} |F(n,k)| x^n y^k / n! = \sum_{k \geq 1} y^k R(x)^k / k! = \exp(yR(x)) - 1.$$

$$\text{Then } \mu_F(n) = |F(n)|^{-1} n! \partial / \partial y |_{y=1} F(x,y) = |F(n)|^{-1} n! [x^n] R(x) \exp(R(x)).$$

Now $\exp(R(x)) = 2R(x) + 1 - x$, and $|F(n)| = 2p(n+1)$, thus

$$\mu_F(n) = n! [x^n] (2R^2(x) + R(x) - xR(x)) / 2p(n+1).$$

But $|F(n,2)| = n! [x^n] R^2(x) / 2!$ which gives the first part of (1).

Now by lemma 2 of Meir, Moon and Mycielski [1983], together with lemma 1.11(a),

$[x^n]R^2(x) \sim 2R(\rho)\rho^{(n+1)}/n!$, where $R(\rho) = \ln(2)$, as shown by Foulds and Robinson [1984]. Thus $\mu_F(n) \sim 2\ln(2) - 0.5(\lim_{n \rightarrow \infty} np(n)/\rho^{(n+1)}) + 0.5$, and the result follows by lemma 1.11(a).

(2): By a similar argument, $\mu_N(n) = f(n)^{-1}n![x^n]B(x)\exp(B(x))$. By the Lagrange inversion formula,

$$\begin{aligned} n![x^n]B(x)\exp(B(x)) &= (n-1)![\lambda^{n-1}](1+\lambda)e^{\lambda}(1-\lambda/2)^{-n} \\ &= (n-1)![\lambda^{n-1}]e^{\lambda}(1-\lambda/2)^{-n} + (n-1)![\lambda^{n-2}]e^{\lambda}(1-\lambda/2)^{-n}. \end{aligned}$$

$$\text{Thus } n![x^n]B(x)\exp(B(x)) = f(n) + \sum_{0 \leq i \leq n-2} \frac{2^{i+2-n} \cdot (2n-3-i)!}{i!(n-2-i)!}$$

$$= f(n) + 2b(n+1) \sum_{0 \leq i \leq n-2} \frac{2^i \cdot (2n-3-i)! (n-1)!}{i!(2n-2)!(n-i-2)!}$$

$\sim f(n) + 2b(n+1)e$, and the result follows since $f(n) \sim b(n+1)e$. §

1.16 Remark

The average number of labelled trees in all forests of p points approaches $3/2$ (Moon [1970]). For unlabelled trees the limit is 2.191837 (approx.) for rooted trees and 1.755510 (approx.) for unrooted, (Palmer and Scwenk [1979]).

Bicoloured Trees

1.17 Definition

For $a+b=n$, let $f_m(a,b)$ denote the number of trees $T \in \text{BPT}(n)$ in which the pendant vertices labelled $\{1, \dots, a\}$ are coloured A, and the rest are coloured B, so that a minimum of exactly m edges must have differently-coloured endpoints in order to extend the colouring to all the vertices of the tree.

In Carter *et al.* [1988] it is shown that

$$f_m(a,b) = (m-1)!(2n-3m) |N(a,m)| |N(b,m)| b(n) / b(n-m+2); \quad (1)$$

where $|N(k,m)|$ is the number of forests consisting of m rooted labelled binary trees on a total of k endpoints, as given in theorem 1.14 (4).

A proof of this result, which is used in section five, is now given which avoids both the messy calculations and the use of a computer package to manipulate expressions in Carter *et al.* [1988].

1.18 Notation

${}^m C_k$ denotes the usual binomial coefficient, except for ${}^{-m} C_k$ ($m > 0$)

which denotes $[x^k](1-x)^{-m} = {}^{m+k-1} C_k$ (and not $(-1)^k \cdot {}^{m+k-1} C_k$).

1.19 Lemma

$$a+x C_a \cdot {}^{-k} C_{a+x} = a+k-1 C_{k-1} \cdot {}^{-a-k} C_x$$

Proof: Immediate.

1.20 Lemma

Let $F(n,m) = [x^{m-1}](1-x)^{2m-2n}(1-x/2)^{-m}$ $n \geq 2m$.

Then $F(n,m) = b(n+1)/(m-1)!b(n-m+2)$.

Proof:

Let $\Phi(\lambda) = (1-\lambda)^2/(1-\lambda/2)$. The solution of $w(t) = t\Phi(w(t))$ with $w(0)=0$

is $w = 1 - (1+2t)^{-0.5}$. Let $f(\lambda) = (2n-1)^{-1}(1-\lambda)^{-2n+1}$. By the Lagrange inversion formula, $[t^m]f(w) = m^{-1}[x^{m-1}]f'(x)\Phi^m(x) = m^{-1}F(n,m)$.

Thus $F(n,m) = m(2n-1)^{-1}[t^m](1+2t)^{(2n-1)/2}$, and the result follows. §

(This lemma resembles lemma 2 of Carter *et al.* [1988], but does not follow from it).

1.21 Theorem

$$f_m(a,b) = (m-1)!(2n-3m) \frac{N(a,m)N(b,m)b(n)}{b(n-m+2)}$$

Proof:

Define generating functions T_1, T_2, T_3 for rooted binary trees (exponential in x and y which mark pendant vertices coloured A and B respectively, and ordinary in z which marks the number of edge changes), as in Carter *et al.* Thus T_1 (resp. T_2) enumerates those rooted binary trees in which every minimal colouring (of the internal vertices of the tree) assigns the root vertex the colour A (resp. B), while T_3 enumerates all remaining trees.

Removing the root vertex from a tree with more than one vertex to give two smaller rooted trees, we have, by Fitch's algorithm (definition 5.18),

$$T_1 = \frac{1}{2}T_1^2 + T_1T_3 + x$$

$$T_2 = \frac{1}{2}T_2^2 + T_2T_3 + y$$

$$T_3 = \frac{1}{2}T_3^2 + zT_1T_2$$

as given by Carter *et al.*

$$\text{Now } (2n-3)f_m(a,b) = a!b![x^a y^b z^m](T_1+T_2+T_3), \quad (1)$$

since both sides count the number of rooted trees requiring m changes to fit an a/b colouring, where the possible states of the root are left unspecified. (This is a minor but useful departure from Carter *et al.*). We now use the multivariate Lagrange formula for monomials, rather than the full multivariate Lagrange inversion formula as employed in by Carter *et al.*

Applying this formula, (refer Goulden and Jackson, p.25):

$$[x^{k_1} y^{k_2} z^{k_3}](T_1+T_2+T_3) =$$

$$(k_1 k_2 k_3)^{-1} \sum_{\alpha=1,2,3} \sum_{\mu(\alpha)} \Delta(\alpha) \prod_{i=1,2,3} [w_i \mu_{i1}(\alpha) \dots w_3 \mu_{i3}(\alpha)] \Phi_i^{k_i} \quad (2)$$

where $\Delta(\alpha)$ is the determinant of $[\delta_{ij} k_i - \mu_{ij}(\alpha)]_{ij}$,

$$\Phi_1 = (1 - w_1/2 - w_3)^{-1}, \quad \Phi_2 = (1 - w_2/2 - w_3)^{-1}, \quad \Phi_3 = w_1 w_2 (1 - w_3/2)^{-1}, \text{ and}$$

the second summation is over all matrices of the form

$$\mu(\alpha) = [\mu_{ij}(\alpha)] = \begin{bmatrix} k_1 - k_3 - \delta_{1\alpha} & 0 & x_1 \\ 0 & k_2 - k_3 - \delta_{2\alpha} & x_2 \\ k_3 & k_3 & x_3 \end{bmatrix}$$

with $x_1, x_2, x_3 \geq 0$ and $x_1 + x_2 + x_3 = k_3 - \delta_{3\alpha}$ (the condition that $\mu_{31} = \mu_{32} = k_3$ is

imposed by the vanishing of the product term in (2) corresponding to $i=3$ for all other values of μ_{31} and μ_{32}).

$\Delta(\alpha)$ is then the determinant

$$\begin{vmatrix} k_3 + \delta_{1\alpha} & 0 & x_1 \\ 0 & k_3 + \delta_{2\alpha} & x_2 \\ k_3 & k_3 & k_3 - x_3 \end{vmatrix}$$

which takes the values

$k_3 x_1$ ($\alpha=1$), $k_3 x_2$ ($\alpha=2$) and k_3^2 ($\alpha=3$).

For the product term corresponding to $i=1$ in (2) we have

$$\begin{aligned} [w_1^{k_1-k_3-\delta_{1\alpha}} w_3^{x_1}] \Phi_1^{-k_1} &= \\ 2^{-(k_1-k_3-\delta_{1\alpha})} \cdot k_1-k_3-\delta_{1\alpha}+x_1 C_{k_1-k_3-\delta_{1\alpha}}^{-k_1} C_{k_1-k_3-\delta_{1\alpha}+x_1} & \\ = 2^{-(k_1-k_3-\delta_{1\alpha})} \cdot 2k_1-k_3-\delta_{1\alpha}-1 C_{k_1-1}^{-2k_1+k_3+\delta_{1\alpha}} C_{x_1} &\text{ by lemma 1.19.} \end{aligned}$$

By symmetry, the term corresponding to $i=2$ is obtained from this expression by replacing k_1 with k_2 and x_1 with x_2 .

For $i=3$ we have

$$[w_1^{k_3} w_2^{k_3} w_3^{x_3}] \Phi_3^{-k_3} = [w_3^{x_3}] (1-w_3/2)^{-k_3} = 2^{-x_3} \cdot k_3 C_{x_3}$$

Thus the summation term in (2) for $\alpha=1$ is

$$\beta_1 \sum_{\mathbf{x}} x_1^{-2k_1+k_3+1} C_{x_1}^{-2k_2+k_3} C_{x_1}^{-k_3} C_{x_3} 2^{-x_3} \quad (3)$$

where $\beta_1 = k_3 \cdot 2k_1-k_3-2 C_{k_1-1}^{2k_2-k_3-1} C_{k_2-1}^{-2n+2k_3+1}$, $n = k_1+k_2$, and the summation is over all non-negative $\mathbf{x} = (x_1, x_2, x_3)$ such that $x_1+x_2+x_3=k_3$.

Let $P(x) = (1-x)^{-1}$. Then since $P(x)^r = \sum_{j \geq 0} {}^{-r}C_j x^j$, (with ${}^{-r}C_j$ as in 1.18)

$x^d / dx P^r(x) = \sum_{j \geq 0} j \cdot {}^{-r}C_j x^j$ so that the sum in (3) is just

$$[x^{k_3}] (x^d / dx P(x)^{2k_1-k_3-1}) P(x)^{2k_2-k_3} P(x/2)^{k_3}.$$

But $(d/dx P(x)^{2k_1-k_3-1}) = (2k_1-k_3-1)P(x)^{2k_1-k_3}$.

Hence (3) is just $\beta_1(2k_1-k_3-1)[x^{k_3-1}]P(x)^{2n-2k_3}P(x/2)^{k_3}$. (4)

For $\alpha=2$ the term is again given symmetrically, with appropriate β_2 .

For $\alpha=3$, the term inside the summation sign of (2) is

$$\beta_3 \sum_{\mathbf{x}} x_1^{-2k_1+k_3} x_2^{-2k_2+k_3} x_3^{-k_3} 2^{-x_3} \quad (5)$$

where

$$\beta_3 = k_3^2 2^{-(2n-2k_3)} 2^{2k_1-k_3-1} C_{k_1-1}^{2k_2-k_3-1} C_{k_2-1} \text{ and } x_1+x_2+x_3=k_3-1.$$

Hence (5) is just $\beta_3[x^{k_3-1}]P(x)^{2n-2k_3}P(x/2)^{k_3}$. (6)

Combining (1) to (6), (with $k_1=a$, $k_2=b$, $k_3=m$) gives

$$f_m(a,b) = a!b!(2n-3)^{-1}(abm)^{-1}(\beta_1(2a-m-1)+\beta_2(2b-m-1)+\beta_3)F(n,m), \text{ with}$$

$F(n,m)$ as in lemma 1.20.

$$\text{But } \beta_1(2a-m-1)+\beta_2(2b-m-1)+\beta_3 = (2n-3m)\beta_3/m.$$

$$\text{Hence } f_m(a,b) = F(n,m)(2n-3)^{-1}(2n-3m) 2^{-(2n-2m)} \times \\ (2a-m-1)!(2b-m-1)!/(a-m)!(b-m)!$$

and the result follows immediately from lemma 1.20, and theorem 1.14 (4). §

In view of the combinatorial nature of the factors in $f_m(a,b)$, and the difficulties of extending the above approach to r -colourings, Carter *et al.* ask for a structural derivation of $f_m(a,b)$. Such a decomposition is now given for the special case: $a=b=m$.

1.22 Lemma

Let $H(n,k) = \sum_{\{x=(x_1,\dots,x_n) \geq j: \sum_i x_i = k\}} \prod_{1 \leq i \leq n} x_i$.

Then $H(n,k) = {}^{-2n}C_{k-n}$ (${}^{-n}C_k$ as in 1.18).

Proof: For $n \geq 2$, $H(n,k) = \sum_{j \geq 0} j \cdot H(n-1, k-j)$.

Thus if we let $H(n,x) = \sum_{k \geq 0} H(n,k)x^k$ we have for $n \geq 2$,

$$H(n,x) = H(n-1,x)(x+2x^2+3x^3+\dots) = H(n-1,x) \cdot x/(1-x)^2.$$

Hence $H(n,x) = (x/(1-x)^2)^{(n-1)}H(1,x)$, and since

$$H(1,x) = x+2x^2+3x^3+\dots = x/(1-2x)^2, \text{ we have } H(n,x) = x^n(1-x)^{-2n}.$$

Thus $H(n,k) = [x^k]H(n,x) = [x^{k-n}](1-x)^{-2n} = {}^{-2n}C_{k-n}$ as required. §

1.23 Lemma

(1): For $T \in \text{BPT}(2m)$, a colouring of the endpoints of T requires m edge changes, if and only if there are m disjoint paths in T , each with differently coloured endpoints.

(2): For such a colouring, the set of disjoint paths is unique.

Proof:

Part (1) follows immediately from Menger's theorem (see Harary [1969]).

For part (2) suppose there exist two path sets Π_1, Π_2 with $\Pi_1 \neq \Pi_2$. For $i=1,2$, Π_i defines a permutation f_i on S , defined by $f_i(x) = y$, if x and y are joined by a path in Π_i . Note that $f_i(f_i(x)) = x$ for all $x \in S$.

Since $\Pi_1 \neq \Pi_2$ we have $f_2(f_1(x)) \neq x$ for some $x \in S$. However regarding the composition $f_2 \cdot f_1$ as an element of the group of permutations on S , we have $(f_2 \cdot f_1)^N(x) = x$ for some integer $N > 1$. Since $f_2 \cdot f_1(x) \neq x$, we can represent T as in fig.1.2 where $\pi_1 \in \Pi_1$ is the path joining x and $f_1(x)$ and $\pi_2 \in \Pi_2$ is the path joining $f_1(x)$ and $f_2 \cdot f_1(x)$, (with $f_2 \cdot f_1(x)$ a label of T_r), and the shaded edges indicate that $T_1, \dots, T_{r-1}$ and $T_{r+1}, \dots, T_s$ may or may not exist, depending on the position along π_1 of T_r .

Now vertex v , shown in fig.1.2, lies on a path from both Π_1 and Π_2 , so that no further paths from these sets can pass through v (thus $(f_2 \cdot f_1)^k(x)$ is "trapped" in T_r for $k > 1$). However this contradicts the equality $(f_2 \cdot f_1)^N(x) = x$ for some $N > 1$, thereby refuting our initial assumption that $\Pi_1 \neq \Pi_2$. §

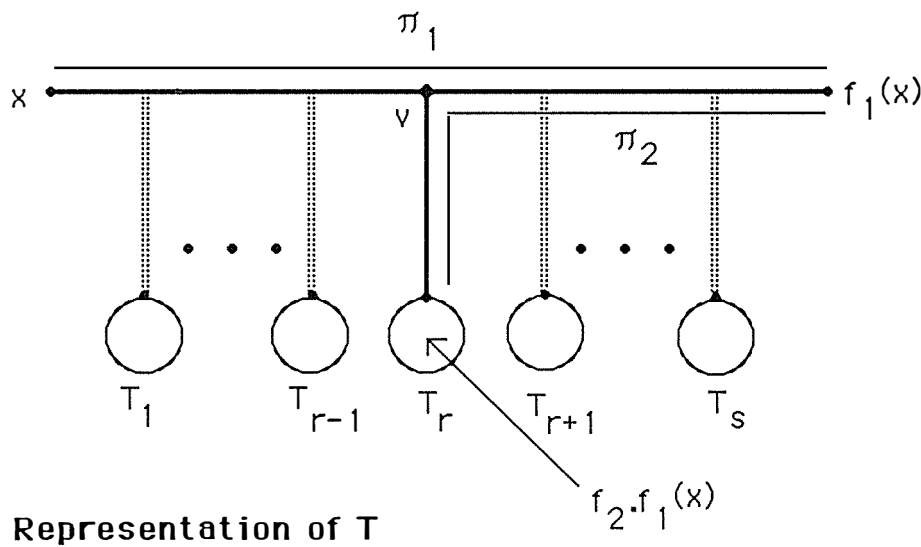


Figure 1.2

1.24 Lemma

$$[x^{k-2}](1-x/2)^{-(k-1)}(1-x)^{-2k} = [t^{k-2}](1-2t)^{-(2k+1)}/2.$$

Proof:

This is just a special case of lemma 2 in Carter *et al.* [1988] with $p=0$, $i=k-2$, $j=2k$. §

1.25 Theorem

$$f_m(m,m) = m!b(2m)/b(m+2).$$

Proof 1:

Let $F(m)$ be the set of trees in $BPT(2m)$ requiring m edge changes to fit a colouring, so that $f_m(m,m) = |F(m)|$. In view of lemma 1.23, $F(m)$ can be constructed as follows: First join each vertex from $\{1,...,2m\}$ coloured A to a vertex coloured B --there are $m!$ ways of doing this. Let E be the set of edges so created. Replace each $e \in E$ with a new vertex $v(e)$ and join these vertices to each other and to $k \geq 0$ new labelled internal vertices of degree 3 so as to form a (not necessarily binary) tree T' whose endpoints form a subset of $\{v(e): e \in E\}$.

Next, letting $d(e)$ be the degree of $v(e)$, construct $\prod_e d(e)!$ trees in $F(m)$ from T' by replacing each $v(e)$ by the $d(e)!$ possible ways of attaching along e the edges that were incident with $v(e)$. Finally unlabel the k new internal vertices created above and take the union over all values of k . An example of this process (with $k=2$) is illustrated in fig. 1.3.

The reason for this somewhat circuitous construction is that it allows us to use theorem 1.5 which gives the number of trees having m vertices of degree $d_1, ..., d_m$ and k vertices of degree 3 as $(m+k-2)! / \prod (d_i-1)! 2^k$.

$$\text{Hence } f_m(m,m) = |F(m)| = m! \sum_k \sum_{d \geq j} \prod_i d_i! \times (m+k-2)! / \prod_i (d_i-1)! 2^k k!$$

where $k + \sum_i d_i = 2m-2$ (the number of internal vertices of any $T \in BPT(2m)$).

Thus $|F(m)| = m! \sum_k \frac{(m+k-2)!}{2^k k!} \times \sum_{\{d \geq j: d+j = 2m-k-2\}} \prod d_i$

$= m! \sum_k \frac{(m+k-2)!}{2^k k!} \times {}^{-2m}C_{m-k-2}$ by lemma 1.22.

$= m!(m-2)! \sum_{\{i,j: i+j=m-2\}} {}^{m+i-2}C_{m-2} 2^{-i} \times {}^{-2m}C_j$

$= m!(m-2)! [x^{m-2}(1-x/2)^{-(m-1)}(1-x)^{-2m}]$

and the result follows by lemma 1.24. §

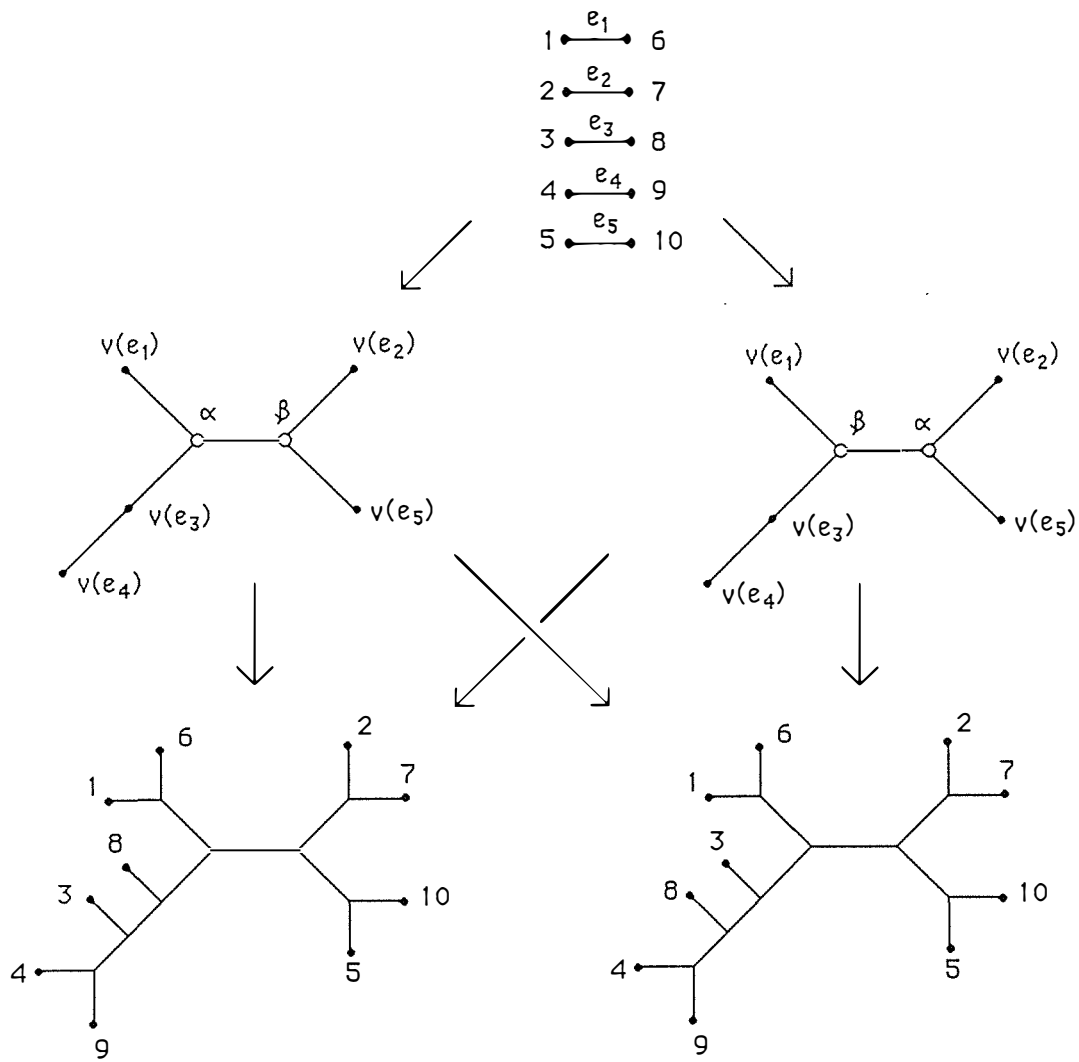


Figure 1.3

Proof 2:

We now give a more direct proof (though one which is likely to be of less use for extension to r -colourings when $r > 2$) of theorem 1.25.

For a partition P of $\{1, \dots, 2m\}$ into m sets of size two, let $G(P)$, denote those trees in $BPT(2m)$ which have a set of m disjoint paths whose endpoints comprise the sets of P . Then $G(P) \cap G(P') = \emptyset$ if $P \neq P'$ by the argument used to establish lemma 1.23 (2), while

$\cup_P G(P) = BPT(2m)$, by a straightforward inductive argument.

Now $|G(P)|$ is clearly (by symmetry) dependent only on m . Thus, since there are $(2m)!/m!2^m$ partitions of X into m sets of size two (Anderson [1974, p.22]), we have

$$|G(P)| = b(2m)/b(m+2) \text{ (since } b(m+2) = (2m)!/m!2^m \text{)}.$$

By lemma 1.23, $f_m(m, m) = \sum_{P \in P_0} |G(P)|$ where P_0 is the collection

of all partitions P of $\{1, \dots, 2m\}$ of type $(1, x_1), (2, x_2), \dots, (m, x_m)$ where

$\{x_1, \dots, x_m\} = \{m+1, \dots, 2m\}$. Clearly $|P_0| = m!$ so that $f_m(m, m) =$

$m!b(2m)/b(m+2)$, as required. §

1.26 Summary

The proofs of the major results in this section (theorems 1.13 and 1.21) demonstrate the usefulness of using generating functions to solve enumerative tree problems. Furthermore theorems 1.5 and 1.25 illustrate that in proving results about our primary object of interest--binary phylogenetic trees--it is sometimes necessary to work with more general classes of trees, a theme that reappears in later sections. Clearly there is further work to be done in enumerating binary trees by the weight of r -colourings, for $r > 2$. While the appropriate set of $(2^r - 1)$ simultaneous quadratic equations can be written down for the generating functions, (as in theorem 1.21 for $r=2$) it is not clear how they could be solved, or indeed whether there is a convenient expression for their solution. It is possible that structural approaches, such as those used in the proof of theorem 1.25 may be more useful.

§2: DISTRIBUTION OF THE SYMMETRIC DIFFERENCE METRIC

"It seems feasible and would be desirable, to estimate by Monte Carlo simulation the means and variances of [the symmetric difference metric] for large values of n ," W.H.E. Day, 1983.

2.1 Introduction

The symmetric difference metric defined on phylogenetic trees is a special case of symmetric difference metrics on sets studied by Restle [1959], and Marczewski and Steinhaus [1958]. The tree metric has been useful in testing evolutionary hypotheses and in examining the methods used to build evolutionary trees as discussed by Penny and Hendy [1985]. An optimally efficient algorithm has been developed by Day [1985] to compute the metric, and its distribution amongst pairs of small trees is described by Day [1983] and Hendy, Little and Penny [1984].

This section extends results from these last two publications to obtain bounds on the distribution of pairs of arbitrarily-large binary trees a given distance apart. As a result, the asymptotic distribution is shown to be Poisson, with $e^{-1/8} \approx 88\%$ of all pairs of binary trees maximally distant, which answers a conjecture by Hendy, Little and Penny [1984]. Asymptotic bounds on the distribution and a monotonicity result are derived, and the distribution is described "from below". The distribution of the metric on the full class of phylogenetic trees is also examined. In particular, the asymptotic mean and variance of this metric is derived, and this confirms a second conjecture by Hendy, Little and Penny [1984].

This section expands on and extends results by Steel [1988]. Some results from that paper have been substantially improved with more elegant proofs and several new theorems.

Properties of the symmetric difference metric on binary trees make it useful for hypothesis testing involving trees derived from homologous DNA sequences, as in Penny, Foulds and Hendy [1982]. The resulting trees may be expected to be similar and it is useful to have a metric for which most trees are far apart. This section concludes with a brief discussion of other metrics.

Distribution Generating Functions

2.2 Definitions

The symmetric difference metric d , which Bourque [1978] and Robinson and Foulds [1981] applied to phylogenetic trees, is defined on $PT(n)$, and so on $BPT(n)$, as follows. For $T \in PT(n)$, deletion of an internal edge of T , $e \in E(T)$ induces a two-set partition $\pi(T, e)$ of $\{1, \dots, n\}$ corresponding to the labels on the two connected components of T with e deleted. For $T_1 \in PT(n_1, f_1)$, $T_2 \in PT(n_2, f_2)$ and $\pi(T_1, e_1) = \pi(T_2, e_2)$ we call e_1, e_2 an *equivalent pair of edges*. If T_1, T_2 have exactly m equivalent pairs of edges then $d(T_1, T_2) = f_1 + f_2 - 2m$. In particular for $T_1, T_2 \in BPT(n)$, $d(T_1, T_2) = 2(n - 3 - m)$.

For $T \in PT(n)$ we recall from Hendy, Little and Penny [1984] the generating functions:

$$P(T, x) = \sum_{k \geq 0} p_k(T) x^k$$

$$Q(T, x) = \sum_{k \geq 0} q_k(T) x^k$$

where $p_k(T)$ (resp. $q_k(T)$) is the number of trees in $PT(n)$ (resp. $BPT(n)$) at distance k from T . Thus for $T \in PT(n, f)$, $Q(T, x)$ has degree $n + f - 3$ and is an even or odd polynomial of parity equal to the numerical parity of its degree.

For $T \in PT(n)$, $s \geq 0$ let $q(s, T)$ denote the number of binary trees having s equivalent edge pairs with T , and let $q(s, n)$ be the average value of $b(n)^{-1} q(s, T)$ over $BPT(n)$. Thus $q(s, n) = b(n)^{-2} \sum_{T \in BPT(n)} q(s, T)$ is the probability that two trees randomly-chosen from $BPT(n)$ have exactly s equivalent pairs of edges. We show that $q(s) = \lim_{n \rightarrow \infty} q(s, n)$ has a Poisson distribution in s with mean $1/8$.

We begin by noting that for $T \in PT(n)$, $P(T, x)$ and $Q(T, x)$ do not depend on the labelling of T , only on its topology. We shall frequently write these and other tree-valued functions which are invariant under topological equivalence without specifying the labelling of the tree. With this in mind we now state from Hendy, Little and Penny [1984]:

2.3 Theorem

Let e be an internal edge of $T \in PT(n)$. Let T/e be the tree formed by contracting e , and let T_1, T_2 be the maximal subtrees of T with e as a pendant edge. Then:

$$P(T, x) = xP(T/e, x) + (1-x^2)P(T_1, x)P(T_2, x)$$

$$Q(T, x) = xQ(T/e, x) + (1-x^2)Q(T_1, x)Q(T_2, x). \quad \S$$

We now give a constructive description of $Q(T, x)$. Let $T \in PT(n, f)$ and let E be a set of internal edges of T . For each edge $e \in E$ cut e in half and place new pendant vertices on each of the two "ends" of e . In this way E defines a collection of trees, T_i , having n_i pendant vertices, for

$i=1, \dots, |E|+1$ (with $T_1=T$, if $E=\emptyset$). Clearly, $\sum_{\{i: 1 \leq i \leq |E|+1\}} n_i = n + 2|E|$.

Let $\Phi(E)$ be the sequence $(n_1, \dots, n_{|E|+1})$, taken in some order, and let

$\langle \Phi(E) \rangle = \prod_{1 \leq i \leq |E|+1} b(n_i)$. Define $r(s, T)$ to be the sum of $\langle \Phi(E) \rangle$ over

all sets of internal edges, E , with $|E| = s$. Finally, let

$$R(T) = R(T, x) = \sum_{s \geq 0} r(s, T) x^s$$

$$q(T) = q(T, x) = \sum_{s \geq 0} q(s, T) x^s$$

(so that $q(T, x) = x^{(n-3+f)/2} Q(T, x^{-1/2})$).

2.4 Lemma

For $T \in PT(n)$, $T_n \in PT(n,0)$,

(a): $q(T,x) = R(T,x-1)$

(b): In the notation of theorem 2.3

$$(i) R(T) = R(T/e) + xR(T_1)R(T_2)$$

$$(ii) q(T) = q(T/e) + (x-1)q(T_1)q(T_2)$$

(c): $R(T,0) = R(T_n,x) = b(n)$.

Proof:

(a): Let $T \in BPT(n)$ and let E be a set of s internal edges of T .

Under the above construction, for each edge $e \in E$, new pendant vertices v_1, v_2 are attached to the ends of a bisection of e .

For $i=1,2$, label v_i with the set of labels of those pendant vertices of T which are no longer joined by a path to v_i when e is cut. Each tree, T_i , ($i=1,...,s+1$) defined by E , thus has a natural label set $L(i)$ for its pendant vertices, so that $T_i \in BPT(L(i))$. This process is illustrated for $s=2$ in fig. 2.1 by the tree J_6 with two distinguished edges.

Now, let $B(T,E)$ be the set of trees in $BPT(n)$ having internal edges equivalent to edges in E . We construct a bijection, F , from $B(T,E)$ to $\prod_i BPT(L(i))$. Given $T' \in B(T,E)$, performing the above edge splitting and labelling procedure on T' produces the label sets $L(1),...,L(s+1)$ and hence an element of $F(T') \in \prod_i BPT(L_i)$. The inverse of F , takes

$(T_1,...,T_{s+1}) \in \prod_i BPT(L(i))$ and identifies all pairs of pendant vertices v_1, v_2 labelled with sets A_1, A_2 such that $A_1 \cup A_2 = L$, the identified vertices then being suppressed to give a tree in $B(T,E)$.

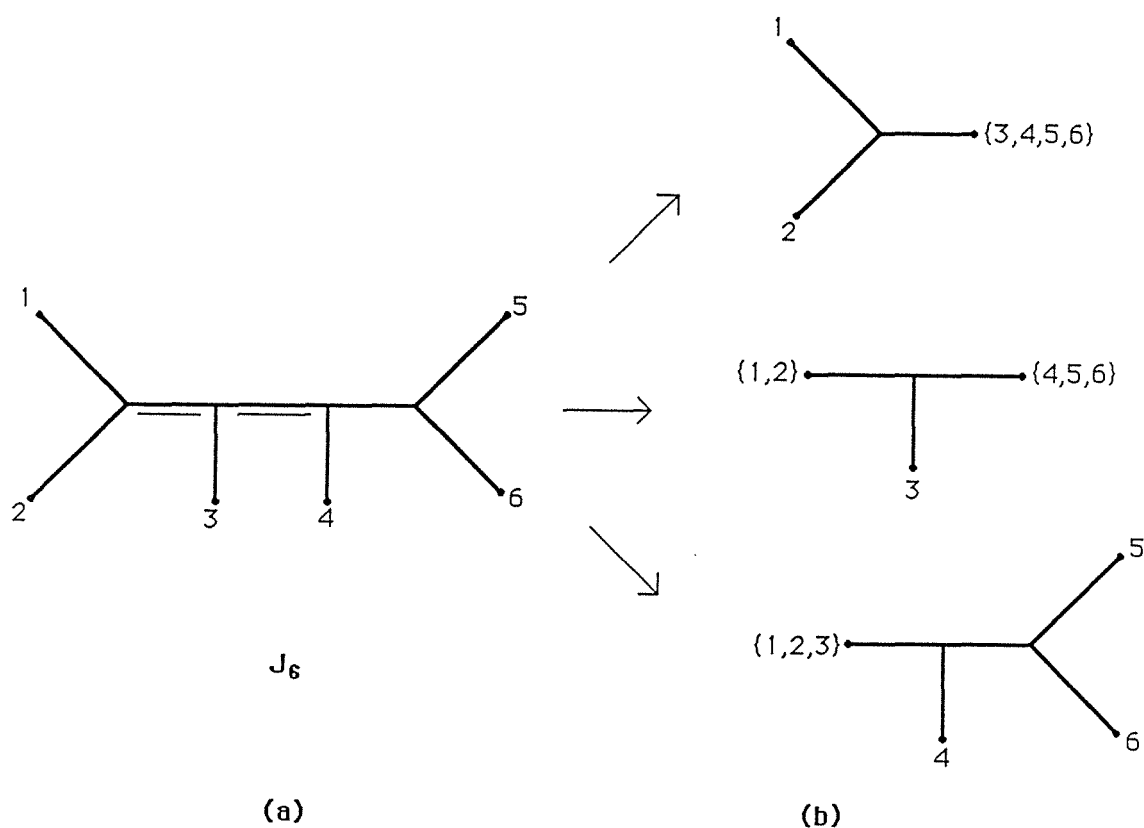


Figure 2.1

Now if $|L(i)| = n_i$ then $|BPT(L_i)| = b(n_i)$, so the bijection gives

$|B(T, E)| = \langle \Phi(E) \rangle$. By the principle of inclusion and exclusion, $r(T, x-1)$ is then the (ordinary) generating function for the number of binary trees equivalent to T on an exact number of internal edges, establishing (a).

Part (b)(i) can be proved directly, or from theorem 2.3 by noting $Q(T, x) = x^{n-3+f} R(T, x^{-2}-1)$. Part (b)(ii) follows from (a), while part (c) follows from the definition of $R(T, x)$. §

2.5 Example

For the caterpillar tree $J_6 \in \text{BPT}(6)$;

$$\begin{aligned} R(J_6, x) &= b(6) + (2b(3)b(5) + b(4)^2)x + 3b(3)^2b(4)x^2 + b(3)^2x^3 \\ &= 105 + 39x + 9x^2 + x^3. \end{aligned}$$

$$q(J_6, x) = R(T, x-1) = 74 + 24x + 6x^2 + x^3$$

$$Q(J_6, x) = 1 + 6x^2 + 24x^4 + 74x^6. \quad \S$$

The following result is easily proved by induction on N for each k using $b(n_1)b(n_2) \leq b(n_1-1)b(n_2+1)$ for $3 \leq n_1 \leq n_2$.

2.6 Lemma

For positive integers $t \geq 3, N, k$;

$$\max\{\prod_{1 \leq i \leq k} b(x_i) : \sum_i x_i = N, x_i \geq t\} = b(t)^{k-1}b(N-(k-1)t).$$

2.7 Theorem

For $T \in \text{PT}(n, f)$, $n \geq 3, s \geq 0$,

(1): $q(s, T) = 0$, for $s > f$,

(2): $q(f, T) = \prod_i b(\partial_i)$, where $(\partial_1, \dots, \partial_{f+1})$ is the degree sequence of the internal vertices of T , and

(3): $q(s, T) \leq {}^f C_s b(n-s)$.

Proof:

We have $q(s, T) \leq \sum_{i \geq 0} {}^{(s+i)} C_s \cdot q(s+i, T) = r(s, T)$.

Thus $q(s,T) \leq r(s,T) = 0$ for $s > f$, giving (1), while for (2) we have from lemma 2.4(a), $q(f,T) = \sum_i (-1)^i \cdot {}^{(f+i)}C_s r(f+i,T) = r(f,T) = \prod_i b(\partial_i)$. For (3) we may assume from (1) that $s \leq f \leq n-3$, so that for a set, E , of s internal edges of T , if $\Phi(E) = (n_1, \dots, n_{s+1})$, then $\sum_i n_i = n+2s \geq 3(s+1) \geq \min\{n_i\} \cdot s$.

Applying lemma 2.6 with $t=3$, $N=n+2s$, $k=s+1$, gives:

$$\langle \Phi(E) \rangle \leq b(3)^s b(n-s) = b(n-s).$$

Since there are fC_s possible choices for E , $r(s,T) \leq {}^fC_s \cdot b(n-s)$, which together with $q(s,T) \leq r(s,T)$ gives the result. §

Distribution on pairs of trees

We now consider the distribution of d on pairs of trees.

2.8 Lemma

For $e_1 \neq e_2 \in E(T)$, if $\pi(T, e_1) = (V, W)$ and $\pi(T, e_2) = (V', W')$ with $|V| = |V'|$, then $V \cap V' = \emptyset$.

Proof:

We can represent T as in fig. 2.2 (refer to following page) where $r \geq 1$, $T_1, \dots, T_r, T_\alpha, T_\beta$, are pendant subtrees of T , each with at least one pendant vertex, and without loss of generality $V = \text{Ver}(T_\alpha)$, the set of vertices of T_α .

Now $V' = \text{Ver}(T_\beta)$ or $V' = \text{Ver}(T_\alpha) \cup \bigcup_{1 \leq i \leq r} \text{Ver}(T_i)$, and since

$|\bigcup_{1 \leq i \leq r} \text{Ver}(T_i)| \geq r \geq 1$, and $|V| = |V'|$, we have $V' = \text{Ver}(T_\beta)$, so that

$V \cap V' = \emptyset$. §

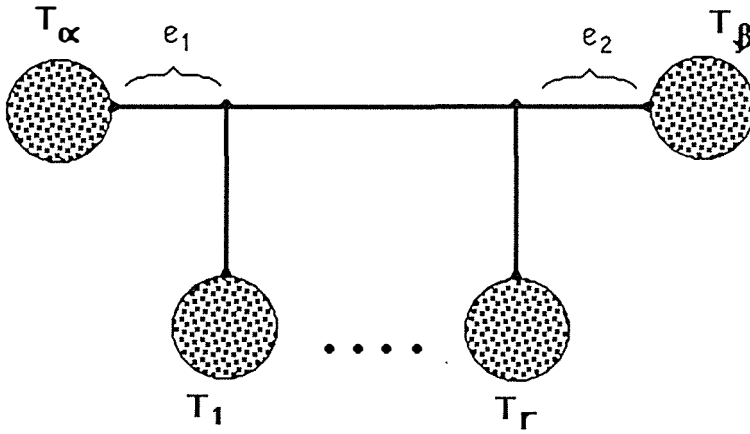


Figure 2.2

We now present the analogue of lemma 2.4 for pairs of trees.

2.9 Definition

For $k \leq n/2$, $\Omega = \{S_1, \dots, S_r\}$ with $S_i \in [n]_k$ for $i=1, \dots, r$, let $B(\Omega)$ denote the set of pairs of trees $T_1, T_2 \in \text{BPT}(n)$ having among their pairs of equivalent edges a set of size r for which Ω is the collection of the sets of labels on the smaller of the induced maximal subtrees.

For $k \leq n/2$, let $\beta(n, k, r)$ denote the sum of $|B(\Omega)|$ over all sets $\Omega \in [n]_k^r$.

2.10 Theorem

$$(1): \beta(n, k, r) = (b^2(k+1)/k!)^r b^{2(n-(k-1)r)} n! / (n-kr)! r!$$

In particular,

$$(2): \beta(n, k+2, 1) / b^{2(n)} = O(n^{-k})$$

$$(3): \lim_{n \rightarrow \infty} \beta(n, 2, r) = b^{2(n-r)} n! / (n-2r)! 2^r.$$

Proof: (1): Suppose that $B(\Omega) \neq \emptyset$. By lemma 2.8, we have $S_i \cap S_j = \emptyset$ for $i \neq j$.

For $i \in \{1, \dots, r\}$, let $L(i) = S_i \cup \{1, \dots, n\} - S_i$ and $L(0) = \{S_1, \dots, S_r\} \cup \{1, \dots, n\} - \cup_i S_i$. Then

there is a bijection from $\prod_{0 \leq i \leq r} \text{BPT}(L(i)) \times \text{BPT}(L(i))$ to $B(\Omega)$ defined as

follows. For $((T_0, T_1, \dots, T_r), (T_0', T_1', \dots, T_r')) \in \prod_{0 \leq i \leq r} \text{BPT}(L(i)) \times \text{BPT}(L(i))$, and

for $i=1, \dots, r$, identify the pendant vertex of T_i and T_0 labelled with sets which

form a two-set partition of $\{1, \dots, n\}$, and suppress the identified vertex to

obtain a tree $T_1 \in \text{BPT}(n)$. Repeating this procedure for T_i' and T_0' gives a

second tree $T_2 \in \text{BPT}(n)$. By construction $(T_1, T_2) \in B(\Omega)$ and the process is

clearly invertible as required. Now since $|L(i)| = k+1$ and $|L(0)| = r+n-kr$,

(since $S_i \cap S_j = \emptyset$ for $i \neq j$) we have:

$|B(\Omega)| = b(k+1)^{2r} b(n-(k-1)r)^2$, if $S_i \cap S_j = \emptyset$ for $i \neq j$ ($B(\Omega) = \emptyset$ otherwise). Now

there are $(kr)! / (k!)^r r!$ ways to partition kr elements into r sets, each of size

k , (Anderson [1974, theorem 3.1, p.22]) and ${}^n C_{kr}$ ways to choose kr elements

from a set of size n , giving $n! / (k!)^r r! (n-kr)!$ choices for Ω in which $B(\Omega) \neq \emptyset$.

The result follows, and gives (2) and (3) immediately. §

2.11 Theorem

$$q(s) = \lim_{n \rightarrow \infty} q(s, n) = e^{-1/8} / 8^s s!$$

Proof:

The proportion of pairs of trees which have at least one pair of equivalent edges partitioning $\{1, \dots, n\}$ into two sets, both of which have at least four elements is bounded above (by Bonferroni's inequality,

refer Bender [p.491, 1974]) by $\sum_{k \geq 3} \beta(n, k, 1) = O(n^{-1})$.

Thus, asymptotically, we need only consider the contribution to $q(s)$ by pairs of equivalent edges which partition the two trees into sets of size $\{2, n-2\}$.

If $N_2(x) = \sum_r \beta(n, 2, r) x^r$, by the principle of inclusion and exclusion the expected proportion of pairs of binary trees which have exactly r equivalent edges of this type is asymptotically $b(n)^{-2} [x^r] N_2(x-1)$.

Now $b^{-2}(n) \beta(n, 2, r) = b^{2(n-r)} n! / 2^r (n-2r)! r! b^{2(n)} which converges to $1/8 r!$ uniformly for $0 \leq r \leq \sqrt{n}$, so that by corollary 4.2 (page 491) of$

Bender [1974], (taking $\lambda(n) = 1/8$, $f(n) = 1$ and $l(n) = \sqrt{n}$) we have

$b(n)^{-2} [x^r] N_2(x-1)$ converges to $e^{-1/8} / 8 r!$. §

2.12 Corollary

If $v(n)$ is the expected distance between two trees in $BPT(n)$, and $\sigma^2(n)$ the variance, then

(a): $(2n-6)-v(n) \sim 0.25$.

(b): $\sigma^2(n) \sim 0.5$.

Proof:

(a): $v(n) = \sum_s q(s, n) (2n-6-2s)$,

thus $(2n-6)-v(n) = (2n-6)(1-\sum_s q(s, n)) + 2\sum_s s q(s, n)$

$$= 2\sum_s s q(s, n) \quad \text{since } \sum_s q(s, n) = 1,$$

and $\lim_{n \rightarrow \infty} \sum_s s q(s, n) = \sum_s s q(s) = 1/8$, by theorem 2.11.

(b): $\sigma^2(n) = \sum_s q(s, n) ((2n-6-2s)-v(n))^2$

$$= ((2n-6)-v(n))^2 \sum_s q(s, n) - 4(2n-6-v(n)) \sum_s s q(s, n) + 4 \sum_s s^2 q(s, n).$$

Letting $n \rightarrow \infty$, using (a) and noting that $\sum_s s^2 q(s) = 0.125 + (0.125)^2$ (by theorem 2.11) gives the result. §

2.13 Remark

Corollary 2.12 (a) confirms the observation in Hendy, Little and Penny [1984] that $v(n)/(2n-6) \sim 1$. More generally, for $T \in \text{BPT}(n)$, let $v(T)$ be the average distance from T to all other trees in $\text{BPT}(n)$. Clearly $v(T)$ depends only on the topological class of T . We now establish further properties of $v(T)$.

Absolute (non-asymptotic) inequalities

2.14 Theorem

(1): $\min_{T \in \text{BPT}(n)} v(T)/(2n-6) \geq 1 - 1/(2n-5) \sim 1$.

(2): For all integers $k \geq 1$, there exists a positive integer n and a set $S \subset \text{bpt}(n)$ of size k , on which $v(T)$ is constant.

Proof:

(1): By theorem 2.3, differentiating $Q(T, x)$ and setting $x=1$, gives

$$v(T) = v(T/e) + 1 - 2b(n_1)b(n_2)/b(n), \text{ where } T_1 \in \text{BPT}(n_1), T_2 \in \text{BPT}(n_2).$$

By induction, $v(T) = v(T_n) + (n-3) - 2r(1, T)/b(n)$, for $T_n \in \text{PT}(n, 0)$.

Now $Q(T_n, x) = b(n)x^{n-3}$ so that $v(T_n) = n-3$.

$$\text{Thus } v(T)/(2n-6) = 1 - 2r(1, T)/b(n)(2n-6).$$

As in theorem 2.7, $r(1, T) \leq (n-3)b(n-1)$ giving $v(T)/(2n-6) \geq 1 - 1/(2n-5)$, as required.

(2): For n even, and $T \in \text{BPT}(n)$, write

$r(1, T) = \sum_{i \geq 2, 2i \leq n} a_T(i) b(i+1) b(n+1-i)$, where $a_T(i)$ is the number of internal edges of T partitioning its pendant vertices into sets of size i and $n-i$.

In particular, $r(1, T)$ is determined by $\{a_T(i) : 2 \leq i \leq n/2\}$. Clearly $ia_T(i) \leq n$, so that $a_T(i) \leq n/i$, and thus if $R(n)$ is the number of possible values $r(1, T)$ can take as T ranges over $\text{BPT}(n)$,

$$R(n) \leq \prod_{2 \leq i \leq n/2} a_T(i) \leq (n/2)(n/3) \dots (n/(n/2)) = ((n/2)!)^{-1} \cdot n^{n/2-1}$$

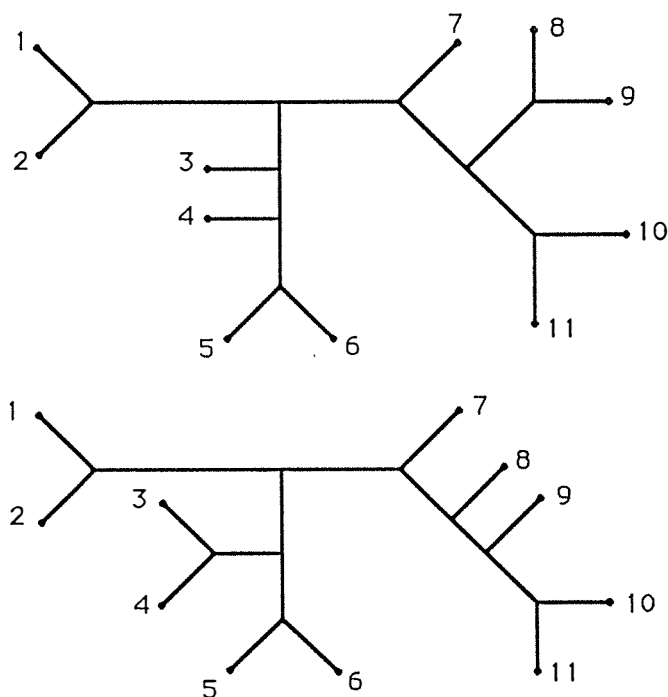
Applying Stirling's approximation gives an asymptotic upper bound on $R(n)$ of $(\sqrt{(2e)})^n / \sqrt{(2\pi n^3)}$. By corollary 1.8, $|\text{bpt}(n)|$ is asymptotically proportional to $n^{-5/2} \theta^n$, where $\theta > \sqrt{(2e)}$, so that $R(n)/|\text{bpt}(n)| \sim 0$, which gives result (2), since $r(1, T)$ determines $v(T)$, as in part (1). §

2.15 Remark

For $k=2$, the non-uniqueness of part (2) of this theorem is realized for $n=11$, by the two trees given in fig. 2.3 (refer to following page). For each tree we have

$$r(1, T) = 4b(3)b(10) + b(4)b(9) + 2b(5)b(8) + b(6)b(7), \text{ giving } v(T) \cong 15.482$$

Although $v(T)$ does not characterize the topology of T , it is not known whether $Q(T, x)$ (equivalently $R(T, x)$ by lemma 2.4) does. §



Two trees in BPT(11) having the same average distance (15.482, approx.) to all other trees in BPT(11).

Figure 2.3

We now establish some interesting properties of $q(0, T)$ and $q(1, T)$ which are required later in this section.

2.16 Theorem

For any $T \in PT(n)$,

(1): $q(0, T) \leq q(0, T/e)$.

(2): $q(0, T) \geq 2q(1, T)$.

(3): $q(1, T) \geq q(1, T/e)$.

Proof:

(1) follows immediately from lemma 2.4 (b)(ii).

For (2), suppose $T_\alpha \in BPT(n)$ and T have exactly one pair of equivalent edges (e_1, e_1') . Represent T_α as $T_\alpha(e_1) = (T_1 T_2)(T_3 T_4)$, as in fig. 2.4(a).

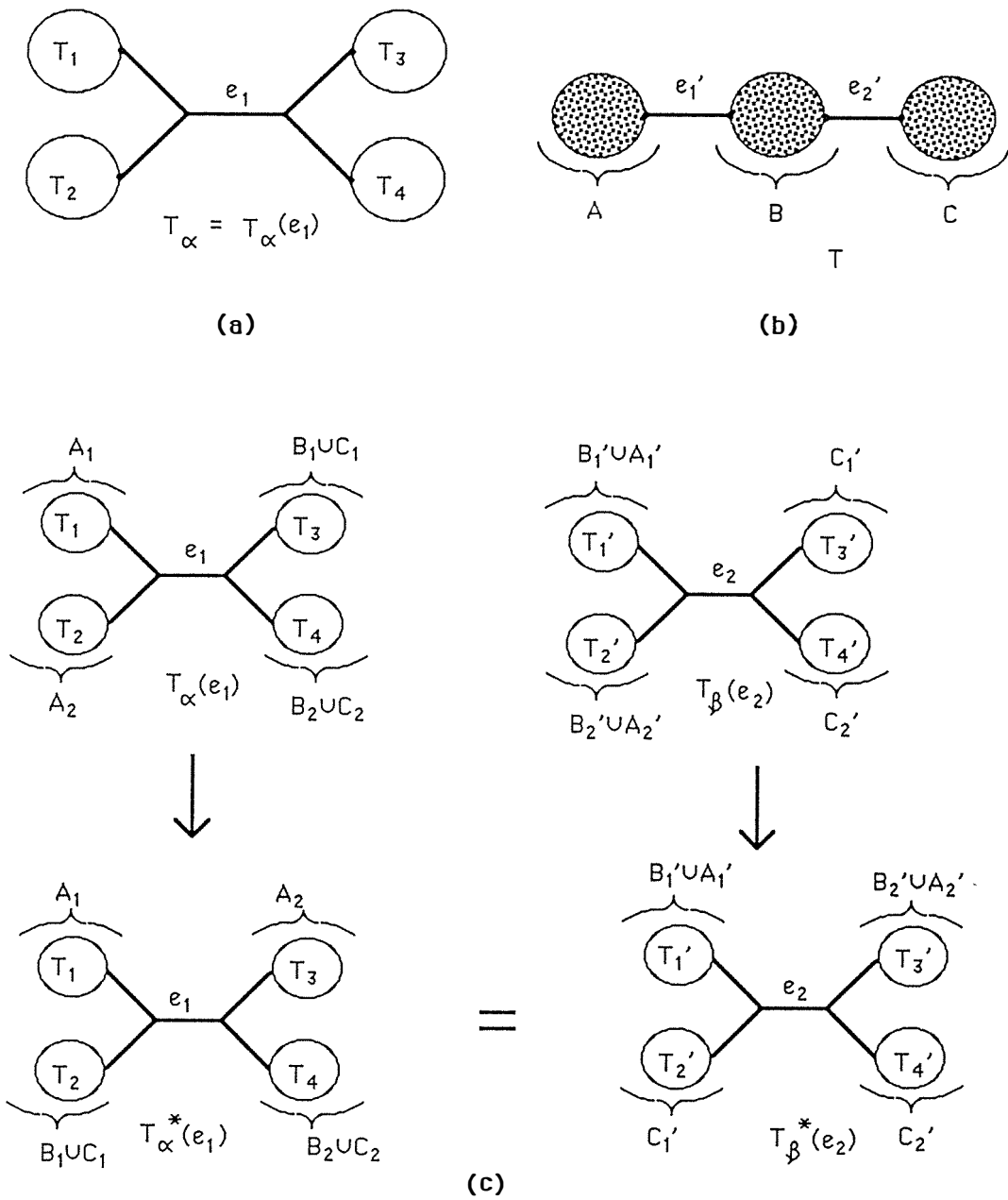


Figure 2.4

Let $T_\alpha^*(e_1) = (T_1 T_3)(T_2 T_4)$ and $T_\alpha^{**}(e_1) = (T_1 T_4)(T_2 T_3)$, which are two trees maximally distant from T . To prove (2) we need to establish that these pairs are distinct for different choices of T_α .

That is, we must check that the pair $T_\alpha^*(e_1), T_\alpha^{**}(e_1)$ do not arise from a different tree $T_\beta \neq T_\alpha$ having exactly one equivalent pair of edges (e_2, e_2') with T .

Case one: if $e_1' = e_2'$ then T_α and T_β have equivalent edge pair (e_1, e_2) so that if $\{T_\alpha^*(e_1), T_\alpha^{**}(e_1)\} \cap \{T_\beta^*(e_2), T_\beta^{**}(e_2)\} \neq \emptyset$ then $T_\alpha = T_\beta$.

Case two: if $e_1' \neq e_2'$ and $\{T_\alpha^*(e_1), T_\alpha^{**}(e_1)\} \cap \{T_\beta^*(e_2), T_\beta^{**}(e_2)\} \neq \emptyset$, we may suppose, without loss of generality, that $T_\alpha^*(e_1) = T_\beta^*(e_2)$.

Represent T as in fig. 2.4(b), where A, B, C are the pendant labels on the subtrees induced by e_1', e_2' . Represent T_α as $T_\alpha(e_1)$, T_β as $T_\beta(e_2)$ and $T_\alpha^*(e_1)$ and $T_\beta^*(e_2)$ as in fig. 2.4(c), where $A_1 \cup A_2 = A_1' \cup A_2' = A$, $B_1 \cup B_2 = B_1' \cup B_2' = B$ and $C_1 \cup C_2 = C_1' \cup C_2' = C$. By construction we have $A_1, A_2, C_1', C_2' \neq \emptyset$. Suppose $T_\alpha^*(e_1) = T_\beta^*(e_2)$. Then since $A_1 \neq \emptyset$, and since T_2', T_4' do not contain elements of A we have $T_1 = T_1'$ (so that $T_2 = T_2'$) or $T_1 = T_3'$ (so that $T_2 = T_4'$). Similarly, $T_3 = T_3'$ or $T_3 = T_1'$.

Thus there are two cases:

- (a) $T_i = T_i'$ for $i=1, \dots, 4$.
- (b) $T_1 = T_3', T_2 = T_4', T_3 = T_1', T_4 = T_2'$.

In both cases it is easily checked from fig. 2.4(c) that $B_1 = B_2 = \emptyset$, so that $B = \emptyset$, a contradiction.

Thus if $\{T_\alpha^*(e_1), T_\alpha^{**}(e_1)\} \cap \{T_\beta^*(e_2), T_\beta^{**}(e_2)\} \neq \emptyset$ we have $e_1 = e_2$ and $T_\alpha = T_\beta$ as required.

For (3) we have from lemma 2.1(b)(ii),

$q(1,T) = q(1,T/e) + q(0,T_1) q(0,T_2) - q(1,T_1)q(0,T_2) - q(0,T_1)q(1,T_2)$. Applying part (2) we have

$$q(1,T) \geq q(1,T/e) - q(0,T_1) q(0,T_2)(1 - 1/2 - 1/2) \geq q(1,T/e), \text{ as required. } \S$$

Asymptotic range of the distribution

Having found the asymptotic average value over $BPT(n)$ of $b(n)^{-1}q(s,T)$, we now calculate its asymptotic range.

2.17 Definition

For $T \in PT(n)$, a *binary vertex* is an internal vertex which is adjacent to exactly two pendant edges. Let $a(T)$ denote the number of binary vertices of T and let $T/2$ be the tree obtained by collapsing all internal edges of T not incident with a binary vertex. Finally let $BS(n,a) \in pt(n,a)$ be the tree obtained by attaching pairs of pendant vertices to a pendant vertices of a star tree in $pt(n-a,0)$. Thus $T/2$ is topologically equivalent to $BS(n,a(T))$.

2.18 Theorem

$$(1): q(s,BS(n,a)) = a C_s \sum_{i \geq 0} (-1)^i \cdot (a-s) C_i \cdot b(n-s-i).$$

$$(2): \text{Let } \eta_n(a) = |\{T \in BPT(n): a(T)=a\}|, \text{ and let}$$

$$\psi(s,n) = \sum_a q(s,BS(n,a)) \eta_n(a). \text{ Then } \psi(s,n) = b(n)^{-2} [x^s] N_2(x-1), \text{ with } N_2(x)$$

as in the proof of theorem 2.11.

$$(3): \text{For } T \in PT(n), s \geq 0, n > 4,$$

$$q(s,T)/b(n) = q(s,BS(n,a))/b(n) + \delta(s,T), \text{ with}$$

$$a = a(T), |\delta(s,T)| < 3^{(s+1)}/2(2n-7).$$

Proof:

(1) follows from lemma 2.4, while (2) follows from the principle of inclusion and exclusion, and (3) follows from theorem 2.3. More detailed proofs can be found in Steel [1988]. §

2.19 Remarks

(1): This theorem gives an alternative proof of the asymptotic Poisson distribution for the symmetric difference metric on BPT(n).

(2): Although we do not require it, there is a closed-form expression for $\eta_n(a)$ derived by Hendy and Penny [1982] who show that:

$$\eta_n(a) = \begin{cases} n!(n-4)!/(n-2a)!a!(a-2)!2^{2a-2}, & \text{for } 2 \leq a \leq \lfloor n/2 \rfloor. \\ 0, & \text{otherwise.} \end{cases}$$

The proof relies on a recurrence for $\eta_n(a)$ which can be written

$$T_n(x) = (n-4 + nx-x)T_{n-1}(x) + 2(x-x^2)^d/dx T_{n-1}(x), \text{ where}$$

$$T_n(x) = \sum_{a \geq 0} \eta_n(a)x^a. \quad \S$$

The next theorem completes our aim of describing the asymptotic range of $b(n)^{-1}q(s,T)$ over BPT(n).

2.20 Theorem

$$\begin{aligned} \sup_{m \geq n, T \in \text{BPT}(m)} \{b(m)^{-1}q(s,T)\} &\sim \begin{cases} 1; & s=0 \\ e^{-1/4}/4^s s!; & s>0 \end{cases} \\ \inf_{m \geq n, T \in \text{BPT}(m)} \{b(m)^{-1}q(s,T)\} &\sim \begin{cases} e^{-1/4}; & s=0 \\ 0; & s>0. \end{cases} \end{aligned}$$

Proof:

Consider a sequence $T(n) \in \text{BPT}(n)$. Since $a(T(n))/n$ is bounded, it has a convergent subsequence $a(T(n_k))/n_k$ whose limit we denote as γ .

Let $a(k) = a(T(n_k))$.

By theorem 2.18 (1) we have:

$$q(s, BS(n_k, a(k))) / b(n_k) = \frac{(s!)^{-1} \sum_{i \geq 0} \frac{(-1)^i \cdot a(a-1) \dots (a-s-i+1)}{i! (2n-5) \dots (2n-2i-2s-3)}}{1}$$

where $a = a(k)$ and $n = n_k$. Writing the second quotient term in this sum as

$$\frac{(a/n)(a-1)/n \dots (a-(s+i-1))/n \cdot 2^{-s-i}}{(1-5/2n) \dots (1-(2i+2s+3)/2n)},$$

we see that for fixed s and i , the i -th term in this sum tends to

$$(s!i!)^{-1} (-1/2)^i \gamma^{s+i} \text{ as } k \rightarrow \infty, \text{ and so } q(s, BS(n_k, a(k))) / b(n_k) \sim \gamma^s e^{-\gamma/2} / s!$$

as $k \rightarrow \infty$. By theorem 2.18 (3) we have

$$q(s, T(n_k)) / b(n_k) \sim q(s, BS(n_k, a(k))) / b(n_k), \text{ as } k \rightarrow \infty.$$

$$\text{Thus } \lim_{k \rightarrow \infty} q(s, T(n_k)) / b(n_k) = \gamma^s e^{-\gamma/2} / s!.$$

Now for any tree $T \in \text{BPT}(n)$, we have $0 \leq a(T) \leq n/2$, so that $0 \leq \gamma \leq 1/2$.

Furthermore for the caterpillar trees J_n which have $a(T) = 2$, we realize

this lower bound of 0, while the upper bound of $1/2$ is realized by the

family of binary trees, $K_{2n} \in \text{bpt}(2n)$, obtained by attaching pairs of

pendant edges to every pendant vertex of a caterpillar tree $T \in J_n$. Then

we have $a(K_{2n})/2n = 1/2$, and the result follows. §

Monotonicity

This subsection and the next further extends results in Steel [1988].

We first show that $q(0,n)$ is monotone increasing in n .

For $s \geq 0$ let $G(s,n) = \{(T,T') : T, T' \in \text{BPT}(n), d(T,T') = 2n-6-2s\}$,

so that $q(s,n) = |G(s,n)| / b(n)^2$.

Given $(T,T') \in \text{BPT}(n) \times \text{BPT}(n)$ there are $(2n-3)^2$ ways of attaching a new pendant vertex v_{n+1} (labelled $n+1$) to the edges of T and T' by a new pendant edge. In this way (T,T') defines a subset $\phi(T,T') \subseteq \text{BPT}(n+1) \times \text{BPT}(n+1)$ of size $(2n-3)^2$.

2.21 Lemma

$$\text{For } (T,T') \in G(s,n), \quad |\phi(T,T') \cap G(j,n+1)| = \begin{cases} n+s, & \text{for } j=s+1, \\ 0, & \text{for } j>s+1. \end{cases}$$

Proof: The only way to increase the number of equivalent edges between T and T' is to bisect an existing one (by adjoining v_{n+1} to one of the s equivalent edge pairs) or create a new one (by adjoining v_{n+1} to pendant edges corresponding to pendant vertices with the same label). In both cases one new equivalent edge pair is produced, and the result follows. $\S$

2.22 Theorem

$q(0,n)$ is monotone increasing in n .

Proof:

We first show that if $(T, T') \in G(1, n)$, then $|\phi(T, T') \cap G(0, n+1)| \geq 8(n-3)$.
 For $(T, T') \in G(1, n)$, let (e, e') be the equivalent pair of edges and let T_1, T_2 (resp. T'_1, T'_2) be the two maximal subtrees of T (resp. T') with e (resp. e') as pendant edge, and with the labels on T_i and T'_i corresponding for $i=1, 2$. Since T_i and T'_i are binary they have the same number of edges (not including e), n_i . Now the equivalent edge pair (e, e') is destroyed and no others produced precisely if v_{n+1} is adjoined to $T_1 - e$ and $T_2' - e'$ or to $T_2 - e$ and $T_1' - e'$, which occurs in $2n_1n_2$ cases. Now $n_1 + n_2 = 2n - 4$ and $n_1, n_2 \geq 2$, so that $2n_1n_2 \geq 8(n-3)$ as required.

Combined with lemma 2.21 (with $s=0$), this gives

$$G(0, n+1) \geq ((2n-3)^2 - n)G(0, n) + 8(n-3)G(1, n); \text{ thus}$$

$$q(0, n+1) \geq q(0, n) + [8(n-3)q(1, n) - nq(0, n)] / (2n-3)^2.$$

Finally we show that the term in square brackets is positive, completing the proof. By theorem 2.5 (1), $q(0, T) \leq q(0, BS(n, a))$, and so $q(0, n) \leq \Psi(0, n)$, where $\Psi(s, n)$ is given by theorem 2.18. By theorem 2.16 (3), $q(1, T) \geq q(1, BS(n, a))$ and hence $q(1, n) \geq \Psi(1, n)$.

Comparison of the first few terms of $\Psi(0, n)$ and $\Psi(1, n)$, shows that $\Psi(1, n) / \Psi(0, n) \geq n / 8(n-3)$ for $n \geq 16$ and since $q(0, n)$ is monotone increasing for $n \leq 16$, by table 4 of Hendy, Little and Penny [1984], the theorem follows. §

2.23 Conjecture

For each $s > 0$, $q(s, n)$ is monotone decreasing in n .

(This conjecture holds for $n < 17$, by table 4 of Hendy, Little and Penny [1984]).

Description of the metric from below

In this section we consider the distribution of binary trees a fixed distance away from a given binary tree. Formally we consider the coefficients $q_{2k}(T)$ of $Q(T,x)$ (see definition 2.2), which are related to the coefficients $q(s,T)$ of $q(T,x)$ by $q_{2k}(T) = q(n-3-k,T)$.

It is easily shown that $q_2(T) = 2n-6$. We now derive an exact expression for $q_4(T)$ and its average value $q_4(n)$ over $BPT(n)$.

2.24 Lemma

For $T \in BPT(n)$, $n \geq 4$, let $\alpha(T)$ be the number of pairs of adjacent internal edges of T . Then $\alpha(T) = n - a(T) - 6$, where $a(T)$ is the number of binary vertices of T (definition 2.17).

Proof:

For $i \in \{0,1,2\}$ let N_i denote the number of internal vertices of T which are adjacent to exactly i pendant vertices. Then $N_1 + N_2 + N_0$ is the number of internal vertices ($n-2$), $N_1 + 2N_2$ is the number of pendant vertices (n) and $N_2 = a(T)$. Now the number of pairs of adjacent internal edges is just $N_1 + 3N_0$. The result follows immediately. $\square$

2.25 Lemma [Steel, 1988]

For $\eta_n(a)$ defined as in theorem 2.18 (2), $T_n(x) = \sum_a \eta_n(a) x^a$,

and $t_n(s) = d^s / dx^s (T_n(x)) \big|_{x=1}$, we have $t_n(s) = 2^{-s} b(n-s) n! / (n-2s)!$

2.26 Theorem

$$q_4(T) = 4^{n-3}C_2 + 6(n-6+a(T))$$

$$q_4(n) = 4^{n-3}C_2 + 6(n-6) + 3n(n-1)/(2n-5).$$

Proof:

$$q_4(T) = r(n-5, T) - (n-4)C_{n-5}r(n-4, T) + (n-3)C_{n-5}r(n-3, T), \text{ by lemma 2.4}$$

and the identity $q_4(T) = q(n-5, T)$.

Now $r(n-5, T) = \alpha(T)b(5) + \beta(T)b(4)^2$ where $\alpha(T)$ is the number of adjacent pairs of internal edges, and $\beta(T) = (n-3)C_2 - \alpha(T)$ is the number of nonadjacent pairs. Since $r(n-4, T) = (n-3)b(4)$, and $r(n-3, T) = 1$, the expression for $q_4(T)$ follows by lemma 2.24. The expression for $q_4(n)$ follows from lemma 2.25 with $s=1$. §

Notice that as $n \rightarrow \infty$, $q_2(T) \sim 2^1 \cdot n^{-3}C_1$ and $q_4(T) \sim 2^2 \cdot n^{-3}C_2$. The next result establishes this pattern for $q_{2k}(T)$.

2.27 Theorem

For any sequence $T(n) \in \text{BPT}(n)$, as $n \rightarrow \infty$,

$$q_{2k}(T(n)) \sim 2^k \cdot n^{-3}C_k \sim (2n)^k / k!$$

Proof:

By definition, $q_{2k}(T) = q(n-3-k, T)$, so that by lemma 2.4(a),

$$q_{2k}(T) = \sum_{0 \leq i \leq k} (-1)^i \cdot (n-3-k+i)C_{(n-3-k)}r(n-3-k+i, T).$$

We first show that $r(n-3-j, T(n)) \sim (n-3)C_j 3^j$ as $n \rightarrow \infty$.

For $T \in \text{BPT}(n)$, let $A(i(1), i(2), \dots; T)$ be the family of all sets E of $n-3-j$ internal edges of T for which $\Phi(E)$ has $i(r)$ copies of $r+3$ for $r = 1, 2, \dots$. Thus

$$r(n-3-j, T) = \sum_i |A(i; T)| \prod_{j \geq 4} b(j)^{i(j)},$$

and $|A(i, T)| = 0$ unless $i(1) + 2i(2) + 3i(3) + \dots = j$, since the $n-3$ internal edges of T are partitioned by each element $E \in A(i; T)$ into two sets: (a) the $n-3-j$ internal edges in A , and (b) the union (over r) of the sets of r internal edges of the $i(r)$ subtrees of T of size $r+3$, induced by A .

For subtrees of size four induced by a set $E \in A(j, 0 \dots 0; T)$, selecting the middle edge of these subtrees gives a bijection between $A(j, 0 \dots 0; T)$ and the set of ways of choosing j internal edges of T with no two edges adjacent (this is generally called the number of j -matchings of T). By Bonferroni's inequality (Bender [p.491, 1974]) we have that

$|A(j, 0 \dots 0; T)|$ lies between $\binom{n-3}{j}$ and $\binom{n-3}{j} - \alpha(T) \cdot \binom{n-5}{j-2}$. By lemma 2.24, $\alpha(T) = O(n)$ so that $|A(j, 0 \dots 0, T(n))| \sim \binom{n-3}{j}$, as $n \rightarrow \infty$.

Now $|A(i, T)| \leq \prod_{r \geq 2} \binom{n-3}{i(r)} \cdot ((n-3) - 2i(2) - 3i(3) - \dots) \binom{n-3-j}{n-3-j}$.

Furthermore, $\binom{n-3}{i(r)} \leq (n-3)^{i(r)}$, and

$$(n-3) - 2i(2) - 3i(3) - \dots \binom{n-3-j}{n-3-j} \leq (n-3)^{j-2i(2)-3i(3)-\dots}$$

Thus $|A(i, T)| \leq (n-3)^{j-i(t)}$, so that if $i(t) \neq 0$ for some $t > 1$, then

$\lim_{n \rightarrow \infty} |A(i, T(n))| / \binom{n-3}{j} = 0$. Thus $r(n-3-j, T(n)) \sim \binom{n-3}{j}^j$ as

$n \rightarrow \infty$, since for fixed j the above summation is finite (clearly $i(r) = 0$

for $r > j$). Thus $q_{2k}(T(n)) \sim \sum_{0 \leq i \leq k} (-1)^i \cdot \binom{n-3-k+i}{n-3-k} \binom{n-3}{k-i} 3^{k-i}$,

as $n \rightarrow \infty$, (since the range of this sum is again independent of n),

and this sum is $\binom{n-3}{k} 3^k \cdot \sum_{0 \leq i \leq k} \binom{k}{i} (-1/3)^i = \binom{n-3}{k} 2^k$, as required. §

Distribution on $PT(n)$

We now consider the distribution of the symmetric difference metric on $PT(n)$. The normalized distance between two trees $T, T' \in PT(n)$ is $d(T, T')$ divided by the maximum possible distance, $2n-6$. Theorem 2.28 (below) shows that, asymptotically, the normalized distance becomes increasingly peaked about its mean $\mu(n)$, which is shown to be less than 1, confirming a conjecture by Hendy, Little and Penny [1984].

2.28 Theorem

Let $\mu(n)$ and $\sigma^2(n)$ denote respectively the mean and variance of the normalized distance between two trees in $PT(n)$.

$$(1): \mu(n) \sim (1 - \ln(2))/p \approx .7943$$

$$(2): \sigma^2(n) = O(n^{-1}).$$

Proof:

(1): A straightforward argument using lemma 1.11 (a) gives a constant C_1 such that:

$$\text{if } n_1, n_2 \geq 3, n_1 + n_2 = n + 2, \text{ then } p(n_1)p(n_2)/p(n) < C_1/n \quad (2)$$

Let $T \in PT(n, f)$ and let $\mu(T) = p(n)^{-1} d /_{dx} P(T, x) |_{x=1}$ be the expected distance between T and trees in $PT(n)$.

By theorem 2.3,

$$d /_{dx} P(T, x) |_{x=1} = P(T/e, 1) + d /_{dx} P(T/e, x) |_{x=1} - 2P(T_1, 1)P(T_2, 1).$$

Hence, $\mu(T) = 1 + \mu(T/e) - 2p(n_1)p(n_2)/p(n)$, where $T_1 \in PT(n_1)$, $T_2 \in PT(n_2)$.

By contracting T to $T_n \in PT(n, 0)$, (2) gives:

$$\mu(T) = f + \mu(T_n) - \epsilon(T), \text{ where } 0 < \epsilon(T) < 2C_1 f / n. \quad (3).$$

Averaging $\mu(T)$ over $PT(n)$, and dividing by $(2n-6)$ to obtain $\mu(n)$, gives

$$\mu(n) = ((p(n)^{-1} \sum_f f |PT(n, f)|) + \mu(T_n) - \epsilon(n)) / (2n-6) \quad (4)$$

with $0 < \epsilon(n) < 2C_1$.

Now the trees at distance f from any $T_n \in PT(n, 0)$ are precisely those with f internal edges. Thus

$$\mu(T_n) = p(n)^{-1} \sum_{f \geq 0} f |PT(n, f)|, \text{ and}$$

$$\mu(n) = (2\mu(T_n) - \epsilon(n)) / (2n-6). \quad (5)$$

Result (a) now follows by lemma 1.11 (a),(b).

(2): The variance $\sigma^2(n)$ of the normalized distance is the average value of $(d(T, T') / (2n-6) - \mu(n))^2$ over all pairs $T, T' \in PT(n)$. Thus letting $D = d/dx$,

$$\sigma^2(n) = (\sum_{T \in PT(n)} \sum_k k^2 P_k(T)) / p(n)(2n-6)^2 - \mu^2(n) =$$

$$(\sum_{T \in PT(n)} D^2 P(T, x) \big|_{x=1} + \sum_{T \in PT(n)} D P(T, x) \big|_{x=1}) / p(n)(2n-6) - \mu^2(n) \quad (6)$$

From theorem 2.3,

$$D^2 P(T, x) \big|_{x=1} = D^2 P(T/e, x) \big|_{x=1} + 2 D P(T/e, x) \big|_{x=1}$$

$$- 4 D((P(T_1, x) P(T_2, x)) \big|_{x=1} - 2 P(T_1, 1) P(T_2, 1)) \quad (7)$$

Now $p(n)^{-1} D((P(T_1, x) P(T_2, x)) \big|_{x=1} \leq 2(2n-8)p(n_1)p(n_2) / p(n)$, (where

$T_1 \in PT(n_1), T_2 \in PT(n_2)$) since for $T' \in PT(m, f)$, $d/dx P(T', x) \big|_{x=1}$ is clearly bounded above by $(m+f-3)p(m) \leq (2m-6)p(m)$.

Let $T \in PT(n, f)$. Dividing (7) by $p(n)$ gives,

$$p(n)^{-1} D^2 P(T, x) \big|_{x=1} = p(n)^{-1} D^2 P(T/e, x) \big|_{x=1} + 2\mu(T/e) - \epsilon_1(T), \text{ with}$$

$0 < \epsilon_1(T) < \epsilon(n) = O(1)$ by (2).

Reducing T to $T_n \in PT(n, 0)$, and using (3),

$$p(n)^{-1} D^2 P(T, x) \big|_{x=1} =$$

$$p(n)^{-1} D^2 P(T_n, x) \big|_{x=1} + 2(\sum_{0 \leq f' \leq f-1} (f' + \mu(T_n))) - \epsilon_2(T) \quad (8)$$

with $0 < \epsilon_2(T) < \epsilon_1(n) = O(n)$.

Now $D^2 P(T_n, x) \big|_{x=1} = \sum_f f(f-1) |PT(n, f)|$, while the second term in (8) is $f(f-1) + 2f\mu(T_n)$. Averaging (8) over $PT(n)$ gives,

$$2p(n)^{-1} \sum_f f(f-1) |PT(n, f)| + 2\mu^2(T_n) - \epsilon_2(n), \text{ with } \epsilon_2(n) = O(n).$$

Substituting this into (6), and noting that

$$\sum_{T \in PT(n)} d / dx P(T, x) \big|_{x=1} / p(n)(2n-6)^2 = \mu(n) / (2n-6), \text{ and}$$

$$2\mu^2(T_n) / (2n-6)^2 - 0.5\mu(n) = O(n^{-1}) \text{ (from (5))}, \text{ we have:}$$

$$\sigma^2(n) = 2(\sum_f f(f-1) |PT(n, f)|) / (2n-6)^2 - 0.5\mu^2(n) + \epsilon_3(n)$$

with $\epsilon_3(n) = O(n^{-1})$. The result now follows from part (a) and lemma 1.11. §

2.29 Remark

Applying Chebyshev's inequality to theorem 2.28 shows that for any number k , the probability that two trees in $PT(n)$ have k or **fewer** equivalent edges tends to zero as n becomes large. This is in contrast to the binary case for which most trees are a maximal distance apart.

Finally, as for $BPT(n)$, we consider briefly the distribution of the d on $PT(n)$ "from below". We evaluate only $P_1(T)$, though $P_2(T)$, $P_3(T)$, ..., could in principle be evaluated by similar though more lengthy arguments.

2.30 Theorem

For $T \in PT(n)$, $P_1(T) = 0.5 \sum_{v \in IV(T)} 2^{\partial(v)} - (n+2f+1)$, where $IV(T)$ is the set of the internal vertices of T and $\partial(v)$ is the degree of v .

Proof:

If $d(T_1, T) = 1$, either T_1 is obtained from T by collapsing an edge of T (in f ways) or T is obtained from T_1 by collapsing an edge, in which case T_1 is obtained from T by replacing an internal vertex v of degree $\partial(v)$ by a tree in $PT(\partial(v), 1)$. From theorem 1.6 we have $|PT(k, 1)| = 2^{k-1} - k - 1$. The result now follows from the identities: $\sum_{v \in IV(T)} \partial(v) = n+2f$, and $|IV(T)| = f+1$. $\S$

Comparison with other metrics:

At least two other metrics have been proposed for comparing phylogenetic trees. One, often referred to as the nearest neighbourhood interchange (NNI), or "crossover" metric, d^* was defined and studied independently by Robinson [1971] and by Moore, Goodman and Barnabas [1973]. Waterman and Smith [1978], (see also Smith and Waterman [1980]) have attempted to analyse this metric further. For $T_1, T_2 \in BPT(n)$, $d^*(T_1, T_2)$ is the minimum length of any chain $\{T_i\}$ of binary trees joining T_1 and T_2 with $d(T_i, T_{i+1}) = 1$ for all i . A useful comparison of d and d^* is given by Day [1983] and Penny and Hendy [1985].

Another metric, d' , based on quartets, and described by Estabrook *et al.* [1985], has also been applied to phylogenetic analysis. One disadvantage of these metrics over d is that the diameter of $BPT(n)$ under both d^* and d' is not known in general, (though the diameter of d^* is reported to be $O(n \log n)$, (Day [1983]), and is given, for $n \leq 9$, in the same paper; while the diameter of d' for $n \leq 10$ is given by Bandelt and Dress [1986]).

Furthermore there is no recursive or constructive description of the distribution of d^* or d' analogous to theorem 2.3 or lemma 2.4, and in the case of d^* , there is no efficient method known for even calculating $d^*(T_1, T_2)$. However as pointed out by Bandelt and Dress [1986], d' has a possible advantage over d of being relatively insensitive to the position of a single taxon. Also since d' and d^* have larger diameters than d , they may provide a finer measure for comparing trees than d . The precise relationship between these three metrics is unknown, apart from $d(T_1, T_2) \leq 2d^*(T_1, T_2)$, implied by the triangle inequality for d . One might ask, for example, whether $d(T_1, T_2) \geq d(T_1, T_3)$ implies a corresponding inequality for d^* or d' .

It is worth noting that by theorem 3 of Robinson and Foulds [1981], d on $PT(n)$ satisfies the property: for all $T_1, T_2 \in PT(n)$, there exists a $T' \in PT(n)$ with $d(T_1, T_2) = d(T_1, T') + d(T', T_2)$. By definition d^* satisfies the same property on $BPT(n)$, so that $(PT(n), d)$ and $(BPT(n), d^*)$ are isomorphic to the distance space of graphs (refer Harary [1975, p.24]).

Finally, all of these metrics are defined on unweighted trees, where edge lengths are irrelevant in any comparisons. For weighted trees, the symmetric difference metric has been adapted to provide a meaningful metric by Robinson and Foulds [1979].

2.31 Summary

This section has described a number of features of the distribution of the symmetric difference metric on phylogenetic trees. In contrast to other metrics, this metric is amenable to analysis, particularly at the asymptotic level. Essentially there are two properties which allow this. First, there are structural descriptions of the metric's distribution (as given in lemma 2.4(a) and theorem 2.10) which manifest in recursive descriptions such as theorem 2.3 and lemma 2.4(b)(ii). Second, the distribution on $BPT(n)$ quickly becomes dominated (as n increases) by the number of equivalent pairs of binary edges between trees (edges partitioning the label set into sets of size 2 and $n-2$). This is ultimately due to the rate at which the number of binary trees, $b(n)$, grows. The second property allows an asymptotic description of the metric (theorems 2.10, and 2.18) leading to Poisson distributions (theorems 2.11 and 2.20). Finally, while much is now known regarding this metric, a number of important questions remain concerning the distribution and diameter of other tree metrics, and their relationship to each other.

§3: Subtree constraints

"The analysis of large data sets could proceed by division into overlapping subsets which are classified separately and then recombined to provide a single classification," A. D. Gordon 1986.

3.1 Introduction

This section examines the restriction on sets of trees in $BPT(L)$ which arises from fixing the structure of some of the subtrees induced by subsets of L . The motivation for this problem is as follows: most tree building methods choose the tree (or trees) which optimizes some criterion. However, for several methods this has been shown to be an NP-complete problem. An alternative approach is to determine the structure of the subtrees of a fixed size of T , which then define a set of trees which are compatible with this induced structure. When all the subtrees of a fixed size are known this presents no problem. However in the reconstruction of phylogenies using statistical methods it is likely that only certain subtrees will be resolved at a statistically-significant level. This raises important questions of how to efficiently describe the set of trees which are consistent with these subtrees, or at least calculate properties of this set, such as its size or its consensus tree. Some of these questions have been considered briefly by a number of authors, in particular Meacham [1983], Constantinescu and Sankoff [1986], Colonius and Schulze [1981], Bandelt and Dress [1986] and Gordon [1986]. We begin this section by considering the extent to which two binary trees can lack common subtrees of a fixed size.

Induced subtrees and minimal similarity

3.2 Definition

Given $T \in PT(L)$, $S \subseteq L$, let $T(S)$ be the minimal subtree of T containing S , and let $T[S]$ be $T(S)$ with all vertices of degree two suppressed. Thus $T[S] \in PT(S)$, the *phylogenetic subtree of T induced by S* .

3.3 Lemma

For $k > 2$, let $f_1(k)$ be the smallest value of n such that for all $T \in \text{BPT}(m)$, $m \geq n$, there exists a path in T of at least k vertices. Then

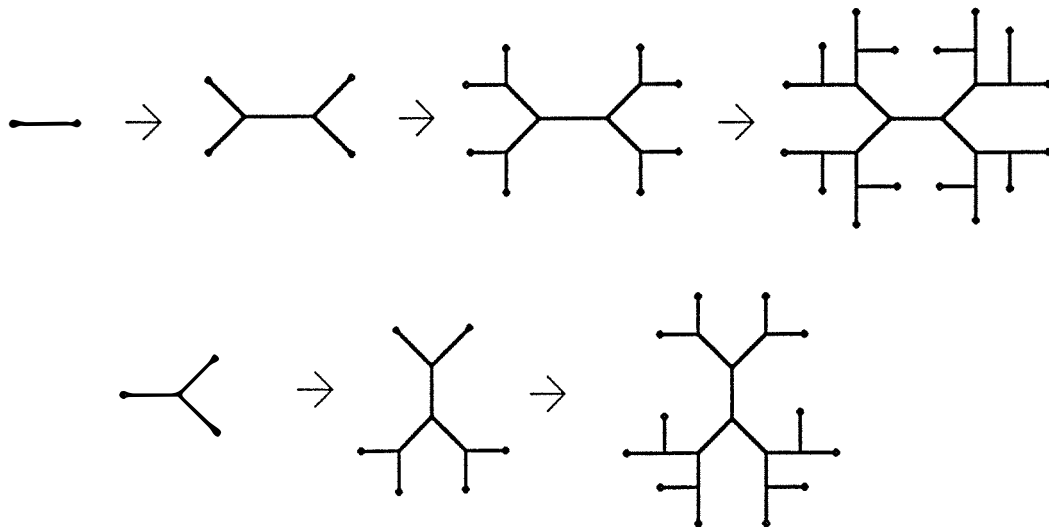
$$f_1(k) = \begin{cases} 3 \cdot 2^{k/2-2} + 1; & \text{if } k \text{ is even,} \\ 2^{(k-1)/2} + 1; & \text{if } k \text{ is odd.} \end{cases}$$

Proof:

Let $g(k)$ be the maximum value of n for which there exists a tree $T \in \text{BPT}(n)$ having longest path at most $k-1$, and let $G(k)$ be the set of such trees.

Then $f_1(k) = g(k) + 1$.

For $T \in G(k)$, let $T^2 \in \text{BPT}(2k)$ be the tree obtained from T by attaching pairs of new pendant vertices and edges to each pendant vertex of T . Define a sequence $QT(k)$ of trees by $QT(k+2) = QT(k)^2$, with $QT(3) = \text{BPT}(2)$, $QT(4) = \text{BPT}(3)$, as shown in fig. 3.1.



The family $QT(k)$ for $2 < k < 10$

Figure 3.1

We first show by induction that $G(k) = \{QT(k)\}$. For $T \in G(k+2)$, consider the binary tree $T^\#$ obtained by removing all pairs of adjacent pendant edges and their pendant vertices (this is less pruning than the derivation discussed in the proof of lemma 1.7). Now for any binary tree, a path of maximal vertex length has as its endpoints, pendant vertices, each of whose associated pendant edge is adjacent to another pendant edge. Thus the paths of $T^\#$ have maximal vertex length $k-1$. Furthermore $T^\# \in G(k)$, for otherwise if $T^\#$ has n_0 pendant vertices there exists $T_1 \in BPT(n_1)$, with $n_1 > n_0$ and $T_1 \in G(k)$. But then $T_1^2 \in BPT(2n_1)$ and T_1^2 has a path of maximal length $k+1$, so that $g(k+2) \geq 2n_1$. Now T has $\leq 2n_0$ pendant vertices and $T \in G(k+2)$, giving $g(k+2) \leq 2n_0$, and so $n_0 \geq n_1$, a contradiction. Applying the hypothesis $G(k) = \{QT(k)\}$, to $T^\# \in G(k)$ we deduce that T consists of $QT(k+2)$ with perhaps some of its pendant vertices and edges deleted. But clearly (all of) $QT(k+2)$ has maximal path length $k+1$, so that $G(k+2) = \{QT(k+2)\}$, as required. Now $QT(k+2)$ has twice as many pendant vertices as $QT(k)$, giving $g(k) = 2g(k-2)$, which together with $g(3) = 2$, $g(4) = 3$, and $f_1(k) = g(k)+1$, establishes the lemma. $\S$

The next theorem says that a common induced subtree of size $k = 4, 5, \dots$ can always be found amongst any two sufficiently large (depending on k) binary trees.

3.4 Theorem

$\forall k \geq 1, \exists n = n(k): \forall T_1, T_2 \in BPT(m), m \geq n, \exists S \in [m]_k$ with $T_1[S] = T_2[S]$.

Proof:

There exist functions f_1, f_2 such that:

- (1): every $T \in \text{BPT}(m)$, $m \geq f_1(n)$ has a path on n vertices,
 (2): any two sequences of length $m \geq f_2(n)$ have a subsequence of length n of the same or reverse order.

Lemma 3.3 gives $f_1(n)$, while the existence of f_2 follows from Ramsey's theorem, though Erdos and Szekeres [1935] show we can take $f_2(n) = n^2 - 2n + 2$.

Let $n(k) = f_1(f_1(f_2(k)))$, and suppose $T_1, T_2 \in \text{BPT}(m)$, $m \geq n(k)$. Then T_1 has a path on a set of $f_1(f_2(k))$ vertices V . For each vertex $v \in V$ select exactly one label in $\{1, \dots, n(k)\}$ as follows: if v is a pendant vertex select its label, otherwise choose the label of any pendant vertex of T_1 whose path back to v is not incident with any other vertex of V . Let S be the set of labels so generated. Then $T_2[S]$ has a path on $f_2(k)$ vertices. Using this path, select a subset S' of S in the same way as before. Then $T_1[S']$ and $T_2[S']$ are both caterpillar trees in $\text{BPT}(S')$, with $|S'| = f_2(k)$. Then there is a subset S'' of S' of size k such that $T_1[S''] = T_2[S'']$, completing the proof. §

3.5 Corollary

$\forall k, r \geq 1, \exists n = n(k, r): \forall P \subseteq \text{BPT}(m), m \geq n, |P| = r, \exists S \in [m]_k$ with $T[S]$ the same for all $T \in P$.

Proof:

Straightforward from the previous theorem. §

3.6 Proposition (Constaninescu and Sankoff [1986])

For $S \subseteq L$, and $T_1 \in \text{BPT}(S)$, we have $|\{T \in \text{BPT}(L): T[S] = T_1\}| = b(|L|) / b(|S|)$.

3.7 Definition

For $T_1, T_2 \in \text{BPT}(n)$, let $A(T_1, T_2, k) = |\{S \in [n]_k \text{ and } T_1[S] = T_2[S]\}|$, and let $\rho(n, k) = \min\{A(T_1, T_2, k) : T_1, T_2 \in \text{BPT}(n)\}$.

3.8 Theorem

- (1): The average value of $A(T_1, T_2, k)$ over all pairs of trees in $\text{BPT}(n)$ is $n C_k / b(k)$.
- (2): $\rho(n, k) / n C_k$ is monotone increasing in n , bounded above by $1 / b(k)$, and has a non-zero limit, $\rho(k)$ as $n \rightarrow \infty$.
- (3): $\rho(k+1) < \rho(k)$.

Proof:

(1) has been established for $k=4$ by W.H.E. Day (reported by Bandelt and Dress [1986]). For the general case consider the set $V(n, k, r) = [[n]_k]_r$, that is, the collection of all sets $\{S_1, \dots, S_r\}$, where $S_1, \dots, S_r$ are distinct subsets of $\{1, \dots, n\}$ of size k . For $\Gamma \in V(n, k, r)$, $\Gamma = \{S_1, \dots, S_r\}$, let $G(\Gamma)$ denote the set of all pairs of trees $T_1, T_2 \in \text{BPT}(n)$, with $T_1[S_i] = T_2[S_i]$ for $i=1, \dots, r$.

Let $n_r = \sum_{\Gamma \in V(n, k, r)} G(\Gamma)$, $N(x) = \sum_{r \geq 0} n_r x^r$ and $E(x) = N(x-1)$. By the principle of inclusion and exclusion, $[x^r]E(x)$ is the number of pairs of trees T_1, T_2 in $\text{BPT}(n)$ for which $A(T_1, T_2, k) = r$. Thus the expected value of $A(T_1, T_2, k)$ over all pairs of trees in $\text{BPT}(n)$ is $b(n)^{-2} [x^r] d / d x E(x) |_{x=1} = n_1$.

$$\begin{aligned}
\text{Now } n_1 &= \sum_{S \in [n]_k} |\{T_1, T_2 \in \text{BPT}(n): T_1[S] = T_2[S]\}| \\
&= \sum_{S \in [n]_k} \sum_{T_1 \in \text{BPT}(n)} |\{T_2 \in \text{BPT}(n): T_1[S] = T_2[S]\}| \\
&= n C_k b(n). (b(n)/b(k)), \text{ by proposition 3.6, and the result follows.}
\end{aligned}$$

The first part of (2) follows by essentially the same argument that Bandelt and Dress [1986] use to establish this result in the case $k=4$. Specifically, choose $T_1, T_2 \in \text{BPT}(n)$ with $A(T_1, T_2, k) = \rho(n, k)$, and suppose $n > m \geq k$. Consider the set Ω of pairs $\{S, \omega\}$ where $\omega \subset S \subset \{1, \dots, n\}$, $|S| = m$, $|\omega| = k$ and $T_1[\omega] = T_2[\omega]$. Construct Ω by first choosing ω (in $A(T_1, T_2, k) = \rho(n, k)$ ways) and then S (in ${}^{n-k}C_{m-k}$ ways) to give $|\Omega| = \rho(n, k) \cdot {}^{n-m}C_{m-k}$. On the other hand we can undercount Ω by first choosing S (in ${}^n C_m$ ways) and then choosing $\omega \subset S$ such that $T_1[S][\omega] = T_2[S][\omega]$ (in $A(T_1[S], T_2[S], k)$ ways). Since $A(T_1[S], T_2[S], k) \geq \rho(m, k)$ we have $\rho(n, k) \cdot {}^{n-m}C_{m-k} \geq {}^n C_m \rho(m, k)$, which gives the required result.

The upper bound of $1/b(k)$ in (2) follows immediately from (1), while the last part of (2) follows from theorem 3.4.

For (3) consider for $T_1, T_2 \in \text{BPT}(n)$, the set Ω_1 of pairs $\{S, x\}$ where $S \in [n]_{k-1}$ and $T_1[S] = T_2[S]$, and $x \in \{1, \dots, n\} - S$. Then $|\Omega_1| = A(T_1, T_2, k-1) {}^{n-k+1}C_1$, but we can undercount Ω_1 by pulling out common subtrees of T_1, T_2 of size k and selecting one label (in ${}^k C_1$ ways) as x , and the remaining set as S to give a pair $\{S, x\}$. In this way we have: $A(T_1, T_2, k) {}^k C_1 \leq A(T_1, T_2, k-1) {}^{n-k+1}C_1$.

Choose T_1, T_2 so that $A(T_1, T_2, k-1) = \rho(n, k-1)$. Then

$$\rho(n, k-1) / {}^n C_{k-1} \geq A(T_1, T_2, k) {}^k C_1 / (n-k+1) {}^n C_{k-1} \geq \rho(n, k) / {}^n C_k$$

since $A(T_1, T_2, k) \geq \rho(n, k)$, as required. §

3.9 Conjecture

$$\lim_{n \rightarrow \infty} \rho(n,k)/n C_k = 1/b(k).$$

(The first nontrivial case, $k = 4$ is conjectured by Bandelt and Dress [1986]).

Spanning Sets

3.10 Definitions

Given $L_1, \dots, L_k \subseteq L$, and $T_i \in \text{BPT}(L_i)$ for $1 \leq i \leq k$. For $Q = \{T_1, \dots, T_k\}$, define the *span* of Q , denoted $\langle Q \rangle$, as $\{T \in \text{BPT}(L) : T_i = T[L_i]\}$. Q is *realizable* if $\langle Q \rangle \neq \emptyset$ and Q *defines* $T \in \text{BPT}(L)$ if $\langle Q \rangle = \{T\}$. Q is a *minimal defining set* for T if Q defines T , but no proper subset of Q defines T .

Henceforth we consider subtrees of size four.

A *quartet spectrum*, Q , is any set of binary 4-trees with labels chosen from $\{1, \dots, n\}$. If every subset of $\{1, \dots, n\}$ of size four labels exactly one tree, the quartet spectrum is *full*. Let $\mathcal{Q}$ denote the set of all quartet spectra, and for $T \in \text{PT}(L)$, let $\psi(T) = \{T[S] : S \in [L]_4\} \cap \mathcal{Q}$.

3.11 Proposition (Bandelt and Dress [1986], Prop. 2(a))

For any $T \in \text{BPT}(n)$, $\psi(T)$ defines T .

3.12 Theorem

(1): If $Q \in \mathcal{Q}$ defines $T \in \text{BPT}(L)$, $|L| \geq 4$, then $|Q| \geq |L| - 3$.

(2): For $T \in \text{BPT}(L)$, there exists a quartet spectrum Q of size $|L| - 3$ which defines T .

Proof:

(1): The result holds for $|L| = 4$; suppose it holds for $|L| = k$, and that Q defines $T \in \text{BPT}(L_1)$, $|L_1| = k+1$. Choose an adjacent pair of pendant edges whose incident pendant vertices are labelled α, β , let $L_2 = L_1 - \{\beta\}$ and let Q^* be the set of quartets obtained by replacing $(xy)(z\beta)$ with $(xy)(z\alpha)$ for all $x, y, z \in L$, $z \neq \alpha$, and deleting $(xy)(\alpha\beta)$ for all $x, y \in L$.

Let $T' \in \text{BPT}(L_2)$ be the tree obtained from T by removing the pendant vertex labelled β and its incident pendant edge. Clearly $T' \in \langle Q^* \rangle$. Suppose $T' \neq T''$ are both in $\langle Q^* \rangle$. Then attaching a new pendant vertex labelled β to the pendant edge of T'' incident with the pendant vertex labelled α gives a tree $T_1 \in \langle Q \rangle$, $T_1 \neq T$, a contradiction. Thus Q^* defines T' , so by hypothesis $|Q^*| \geq k - 3$. Now by the construction, $|Q| \geq |Q^*|$, with equality if and only if $(\alpha\beta)(xy)$ is not in Q , for any $x, y \in L$. But if $(\alpha\beta)(xy)$ is not in Q for any $x, y \in L$, then representing T' as in fig. 3.2 (a) (refer to following page), we see that the trees $(\alpha T_1)(\beta T_2)$, $(\alpha T_2)(\beta T_1)$ and $T = (\alpha\beta)(T_1 T_2)$ in fig. 3.2 (b), (c), (d) are also in $\langle Q \rangle$, a contradiction. Thus $|Q| \geq (k+1) - 3$, as required.

(2): Again we use induction on $|L| = k$. For $k=4$ the result holds, so suppose it holds for $|L| \leq k$, and $T \in \text{BPT}(L)$, $|L| = k+1$. From T choose a pair of adjacent pendant edges whose incident pendant vertices are labelled α, β (so that T is represented by fig. 3.2 (d)) and define T' as before.

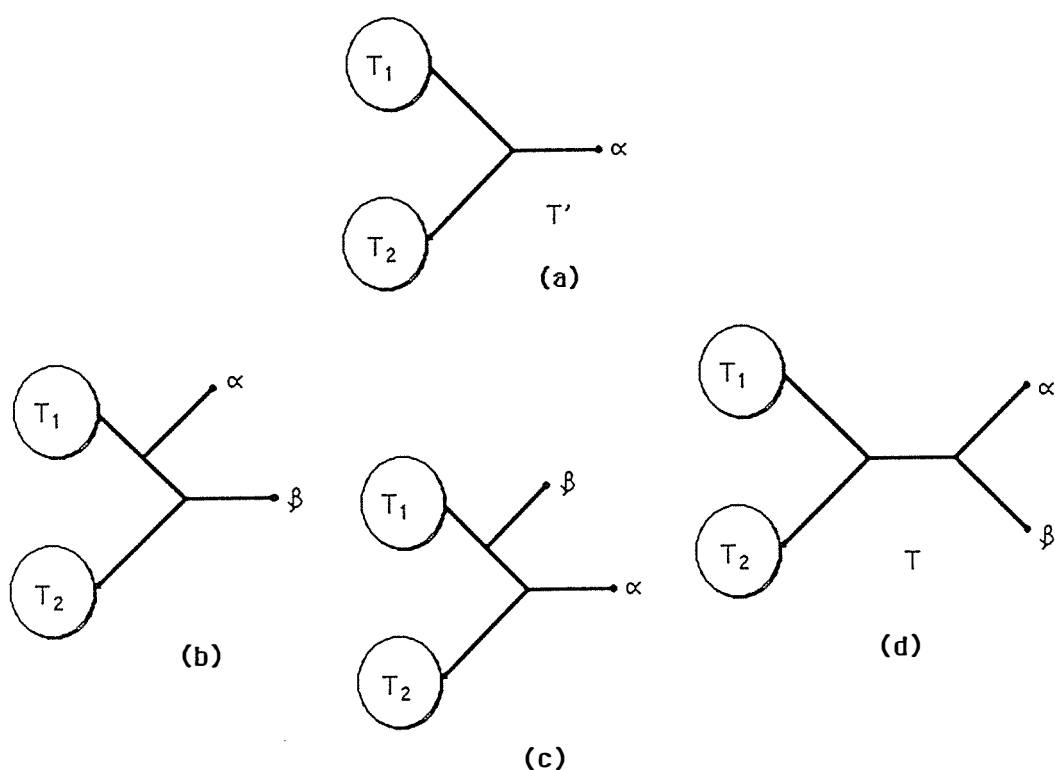


Figure 3.2

Then since $T' \in \text{BPT}(L')$, $|L'| = k$, there is a set Q^* of size $k-3$ which defines T' . Represent T' as in fig. 3.2 (a), and let x be the label of a pendant vertex in T_1 , and y a label of a pendant vertex in T_2 . Let $Q = Q^* \cup \{(\alpha\beta)(xy)\}$. Then $T \in \langle Q \rangle$, and if $T_3 \in \langle Q \rangle$, then since $Q^* \subset Q$, we have $T_3 \in \langle Q^* \rangle$.

If T_3' is obtained from T_3 by deleting the pendant vertex labelled β , and its associated pendant edge, then since Q^* defines T' we have $T_3' = T'$. The position of the new pendant vertex labelled β is now fixed by $(\alpha\beta)(xy) \in Q$, so that $T_3 = T$, and hence $\langle Q \rangle = \{T\}$, as required. $\S$

The proof of part (b) of this theorem gives a constructive (and efficient) method for finding a minimal defining set for T . A natural question is whether amongst quartet spectra, all minimal defining sets for a tree have the same cardinality, a suggestion supported by considering trees with less than 7 pendant vertices. The next example shows that the answer to this question is no.

3.13 Example

Let $Q = \{(35)(67), (13)(45), (16)(23), (57)(12), (72)(34)\}$. Then $\langle Q \rangle = \{T\}$, where T is the caterpillar tree $(16)(273)(45)$. However, deleting the i -th element of Q as they are ordered above for $i=1, \dots, 5$ gives sets Q_i which do not define T . Specifically, $(16)(342)(57) \in Q_1$, $(16)(274)(35) \in Q_2$, $(26)(173)(45) \in Q_3$, $(45)(321)(67) \in Q_4$, $(23)(145)(67) \in Q_5$. Thus Q is a minimal defining set of T of cardinality five, which is one more than the cardinality of the set constructed by the previous theorem. §

3.14 Remark

On the basis of considering small examples one might also conjecture that if $|L_i \cap L_j| \leq 1$ for all $i \neq j$, then Q is realizable. This is not so however, for consider a projective plane of order 3, which has 13 points and 13 lines each with 4 points (Hall [1986]). Taking the lines to be our $L_1, \dots, L_{13}$, we have $|L_i \cap L_j| = 1$, $i \neq j$, and there are 3^{13} possible choices of $Q = \{T_1, \dots, T_{13}\}$ since there are three possible trees on each label set L_i . Now clearly if $Q \neq Q'$ then $\langle Q \rangle \cap \langle Q' \rangle = \emptyset$, so that if each of the sets Q were realizable, $b(13) \geq 3^{13}$. But it can be easily checked that in fact $b(13) < 3^{13}$.

Consensus Trees

3.15 Definition

For subsets L_1, L_2 of L , if $\{L_1, L_2\} = \{\emptyset, L\}$ or $\{L_1, L_2\} = \pi(T, e)$ for some internal edge of T , then L_1 is said to be a *cluster* of T and L_1, L_2 are *complementary clusters*. §

By Proposition 1(a),(b) of Bandelt and Dress [1986], the clusters of T define T , and the conditions on a family of subsets of L to be a cluster of a tree are such that the clusters common to all of a set of trees gives a treelike cluster, thereby allowing the following definition due to Sokal and Rohlf [1981].

3.16 Definition

Given a set of trees $X = \{T_1, \dots, T_k\} \subseteq PT(L)$, the *strict consensus* of X , CX is the tree in $PT(L)$ whose clusters are precisely those clusters belonging to all trees in X . If $X = \emptyset$, define CX to be the star tree.

We now give a description of $\langle Q \rangle$ in terms of $C\langle Q \rangle$.

Consider the set $IV(C\langle Q \rangle)$ of internal vertices of the consensus tree of the span of a quartet spectrum Q . If $v \in IV(C\langle Q \rangle)$ has degree $\partial v > 3$, let $L(v) = \{L_1, L_2, L_3, L_4, \dots, L_{\partial v}\}$ be the collection of sets of labels on the pendant vertices of the pendant subtrees of T having v as an endpoint. Replacing v in $C\langle Q \rangle$ by a binary tree $T(v) \in BPT(L(v))$ gives a new tree $[T(v), C\langle Q \rangle] \in PT(L)$. Let $N(v) = \{T(v) \in BPT(L(v)) : Q \subseteq \Psi([T(v), C\langle Q \rangle])\}$.

Now for $T \in \langle Q \rangle$, and $v \in IV(C\langle Q \rangle)$, $\partial v \geq 3$, T defines a specific tree $T(v) \in BPT(L(v))$, since T is binary. If we denote this tree by (T, Q, v) , consider the function

$H: \langle Q \rangle \rightarrow \prod_{v \in IV(C\langle Q \rangle)} N(v)$, defined by $H(T)_v = (T, Q, v)$.

3.17 Theorem

H is a bijection. In particular $|\langle Q \rangle| = \prod_{v \in IV(C\langle Q \rangle)} b(n_v)$, where $n_v \leq \partial(v)$.

Proof:

H has the following inverse, K . Given $T(v) \in N(v)$ replace v by $T(v)$ for each $v \in IV(C\langle Q \rangle)$ to obtain a tree $T_0 \in BPT(L)$. Suppose T_0 is not in $\langle Q \rangle$, so that Q is not $\subseteq \psi(T_0)$. Let $(ij)(kl) \in Q$, $(ik)(jl) \in \psi(T_0)$. Consider the tree $T_1 = C\langle Q \rangle[S] \in PT(S)$, where $S = \{ijkl\}$. There are two cases to consider (a): T_1 is a binary tree, (b): T_1 is a star tree.

In the first case, if we replace $v \in IV(C\langle Q \rangle)$ by $T(v)$ for any $T(v) \in N(v)$ then, by definition, $Q \subseteq \psi[T(v), \langle C(Q) \rangle]$ so that $(ij)(kl) \in \psi[T(v), \langle C(Q) \rangle]$.

Replacing the other internal vertices to obtain T_0 , we clearly still have $(ij)(kl) \in \psi(T_0)$ a contradiction.

In case (b), let v_0 be the unique vertex of $C\langle Q \rangle$ where the paths $\pi(i, j)$ and $\pi(k, l)$ meet. Replacing v_0 by $T(v_0) \in N(v_0)$ we have $Q \subseteq \psi[T(v_0), \langle C(Q) \rangle]$, and again we find that $(ij)(kl) \in \psi(T_0)$, a contradiction. Thus $T_0 \in C\langle Q \rangle$.

Clearly then HK and KH are identity maps, as required. $\S$

Efficiency

An interesting question is whether there is an efficient way of deciding whether $Q \in \mathcal{Q}$ is realizable--that is whether there is an algorithm for deciding this question whose running time is bounded above by a polynomial function of n . One approach might be to observe that if Q is realizable then there must be a pair $i, j \in L$, with $\{(ix)(jy) : x, y \in L\} \cap Q = \emptyset$. We might then replace the pair (i, j) by new label $\alpha(i, j)$ and replace all occurrences of i or j in Q by $\alpha(i, j)$ (removing any occurrence of $(ij)(xy)$) to obtain a quartet spectrum Q_1 on a set L_1 of $|L| - 1$ labels. Repeating the procedure, eventually either all the labels will get used up and a tree will be constructed, or at the r -th stage no pair with $\{(ix)(jy) : x, y \in L_r\} \cap Q_r = \emptyset$ will exist, or $\{(xy)(zw), (xz)(yw)\} \subseteq Q_r$ for some $x, y, z, w \in L_r$, either of which imply that Q_r is not realizable. All this can be efficiently implemented, but this method depends on making the right choices for which pairs (i, j) to select at each stage. For if the wrong choices are made, at some later stage we may find $\{(xy)(zw), (xz)(yw)\} \subseteq Q_r$, even though Q is realizable. Another approach is suggested by the characterization of when a full quartet spectrum is realizable to which we now turn.

3.18 Definition

Given a quartet spectrum Q , define the *closure* of Q , $cl(Q)$, to be the minimal set containing Q with the property:

for all $(ij)(kl), (ik)(lm) \in cl(Q)$, we have $(ij)(km), (jk)(lm) \in cl(Q)$.

Note that $cl(Q)$ can be efficiently constructed from Q , $cl(cl(Q)) = cl(Q)$ and $\langle Q \rangle = \langle cl(Q) \rangle$, so that Q is realizable if and only if $cl(Q)$ is. Also observe that $cl(\psi(T)) = \psi(T)$.

3.19 Lemma

A full quartet spectrum is realizable if and only if its closure does not contain $(xy)(zw)$ and $(xz)(yw)$.

Proof:

This follows directly from similar results due to Colonius and Schulze [1981], and Bandelt and Dress [1986]. §

For non-full quartet spectra, the example in remark 3.14 shows that even with $cl(Q)=Q$, Q may still be unrealizable. Indeed using Q from example 3.1 and adjoining $(24)(56)$ gives a set Q' with $\langle Q' \rangle = \Phi$ but with $cl(Q') = Q'$, since Q' consists of five of the blocks of the biplane on seven points given by Hall [p. 321, 1986].

We now show that the question of whether or not Q is realizable is equivalent from an efficiency viewpoint to two related questions.

3.20 Lemma

If $L = L_1 \cup L_2$, $L_1 \cap L_2 = \Phi$, and for each $i, j \in L_1$, and $k, l \in L_2$ $\langle Q \cup \{(ik)(jl)\} \rangle = \Phi$ and $\langle Q \cup \{(il)(jk)\} \rangle = \Phi$, then for any $T \in \langle Q \rangle$, L_1 is a cluster of T .

Proof:

If Q satisfies the above condition, then for $T \in \langle Q \rangle$, $T[\{i, j, k, l\}] = (ij)(kl)$, so that the path joining i and j has no common vertex with the path joining k and l . Since this holds for all $i, j \in L_1$, $k, l \in L_2$ it follows (refer Bandelt and Dress [p.318, 1986]) that L_1 is a cluster of T . §

3. 21 Theorem

If one of the following has an efficient algorithm, so do the others:

- (1): Deciding whether Q is realizable.
- (2): Deciding whether Q is realizable, and if so constructing $T \in \langle Q \rangle$.
- (3): Constructing $C \langle Q \rangle$.

Proof:

Clearly efficiency for (2) implies that for (1). Conversely, given Q , and an efficient algorithm for (1), if Q is realizable, Q can be efficiently extended to a full quartet spectrum by choosing at each stage a quartet $\{i,j,k,l\}$ not already in the realizable spectrum so far constructed Q' , and testing the realizability of $Q' \cup \{(ij)(kl)\}$, $Q' \cup \{(ik)(jl)\}$, $Q' \cup \{(il)(kj)\}$. Since Q' is realizable, at least one of these extensions is realizable and the process can be repeated until a full and realizable quartet spectrum Q_0 is obtained. Now given Q_0 , there is an efficient inductive method for building the tree T with $\psi(T) = Q_0$ by using Q_0 to decide where each additional pendant vertex should be attached to the tree so far constructed. Thus efficiency for (1) implies efficiency for (2), as required.

To obtain an efficient algorithm for (3), efficiently construct (by (2)) a tree $T \in \langle Q \rangle$. By definition, the pairs of complementary clusters $\{L_1, L_2\}$ of $C \langle Q \rangle$ are a subset of the complementary clusters of T .

For each internal edge e of T , let $\pi(T, e) = \{L_1, L_2\}$, and for each pair $i, j \in L_1$, $k, l \in L_2$, use the efficient algorithm from (1) to check whether or not $Q \cup \{(ik)(jl)\}$ or $Q \cup \{(il)(kj)\}$ is realizable. If either is realizable (for any i, j, k, l) then clearly L_1 is not a cluster of $C \langle Q \rangle$.

On the other hand if none of these spectra are realizable, then by lemma 3.20, L_1 is a cluster of $C\langle Q \rangle$. and thus the clusters of $C\langle Q \rangle$ can be efficiently generated. The clusters of $C\langle Q \rangle$ can then be used to efficiently construct $C\langle Q \rangle$, by a method such as "TREE POPPING" due to C.A. Meacham, and described in Dress and Bandelt [p.317, 1986].

Finally, an efficient algorithm for (3), gives an efficient algorithm for (1) as follows. Given quartet spectrum Q on label set L , let $L' = L \cup \{\alpha, \beta\}$, with α, β new labels not in L . Let $Q' = Q \cup \{(\alpha\beta)(xy) : x, y \in L\}$. If $\langle Q \rangle = \emptyset$ then $\langle Q' \rangle = \emptyset$, and so by definition $C\langle Q' \rangle$ is a star tree. If $\langle Q \rangle \neq \emptyset$, choose $T \in \langle Q \rangle$ and join one end of a new edge e_0 to any edge of T , and attach two pendant vertices labelled α, β by two new edges to the other endpoint of e_0 , giving a tree $T' \in \text{BPT}(L')$. Thus $\langle Q' \rangle \neq \emptyset$, and by lemma 3.20, $\{\alpha, \beta\}$ is a cluster of $C\langle Q' \rangle$, so that $C\langle Q' \rangle$ is not a star tree. Thus $\langle Q \rangle = \emptyset$ precisely if $C\langle Q \rangle$ is not a star tree, giving an efficient algorithm for (1). §

3.22 Summary

The problem considered in the first part of this section is the extent to which two trees can lack common subtrees of a given size, k . Even for $k=4$ this appears to be a difficult problem (it is equivalent to calculating the diameter of the metric d' discussed in 2.31). After establishing that sufficiently large trees always share a tree of size k (theorem 3.1) we generalized (from $k=4$) some known results aimed at providing lower bounds for the number of shared subtrees of size k (theorem 3.8).

We then addressed the question of how many subtrees of a parent tree are necessary to define it, obtaining both positive and negative results (theorem 3.12 and example 3.13, respectively).

A description of the span of a set of trees of size four is then given in terms of its consensus tree (theorem 3.17). This raises the question of whether one can efficiently construct this consensus tree, a question which is shown (in theorem 3.21) to be equivalent to efficiently deciding the realizability of sets of four-trees. It would be useful to find an efficient algorithm for this question, or else show that it is NP-hard.

§4: Sequence and Dissimilarity Data.

4.1 Introduction

This section defines sequence data and dissimilarities, and derives basic properties of the transformation from the former to the later, required in later sections. We also examine, from an enumerative viewpoint, how much information is lost by this transformation, improving on results by Penny [1982].

4.2 Definitions

A *premetric* $f: \{1, \dots, n\} \times \{1, \dots, n\} \rightarrow \mathbb{R}$, is any function satisfying $f(i, i) = 0$, $f(i, j) \geq 0$, for all $i, j \in \{1, \dots, n\}$.

Given a set $\mathcal{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_r\}$ of r colours, a *sequence space of length c on r colours and n taxa* is a set $S = \{x_1, \dots, x_n\} \subset \mathcal{X}^c$, (thus each x_i is an ordered sequence of colours, which for $r = 4$, might represent a DNA sequence, or for $r=2$ the same sequence in terms of *purine* and pyrimidine bases).

Given a sequence space, $(S, \mathcal{X}, c, r)$, let δ_{ij} be the number of components, called *sites*, of $\mathcal{X}^c$ at which x_i and x_j differ. It is easily checked that $\{\delta_{ij}\}$ is a pseudometric. §

Constraints with two colours

4.3 Definitions

Let $C(n)$ denote the set of all 2^n two-colourings of $\{1, \dots, n\}$, and let $\Omega(n)$ denote the set of the 2^{n-1} ⁽¹⁾partitions--subsets of $\{1, \dots, n\}$ containing 1. A colouring $\chi \in C(n)$, induces a partition corresponding to the elements of $\{1, \dots, n\}$ coloured the same colour as 1. Conversely, for a partition σ , there are precisely two $\chi \in C(n)$ which induce σ . For a sequence space, and $\sigma \in \Omega(n)$, let X_σ be the number of occurrences of σ . Denote $\{1, \dots, n\} \in \Omega(n)$

by σ_0 . § ⁽¹⁾(By an abuse of terminology we call $\sigma \in \Omega(n)$ a partition because σ and its complement partition $\{1, \dots, n\}$).

We have $\delta_{ij} = \sum_{\sigma \in \Omega(n)} \alpha(i, j, \sigma) X_\sigma$ where $\alpha(i, j, \sigma) = 0$ if i and j are either both in σ or neither in σ , and $\alpha(i, j, \sigma) = 1$ otherwise.

Indexing the unordered distinct pairs $\{i, j\}$ and ordering $\Omega(n)$ we have for the associated vectors δ and X , $\delta = MX$, with M a ${}^nC_2 \times 2^{n-1}$, 0-1 matrix.

4.4 Remark

It is easily shown (though we do not require it) that $MM^t = 2^{n-3}(I+J)$,

which is invertible so that for any $\delta \in R^{n(n-1)/2}$ letting

$X = M^t(MM^t)^{-1}\delta$ we have $MX = \delta$. The restrictions on dissimilarities δ derived from sequence data arise from the requirement that $\delta = MX$ for $X \geq 0$.

4.5 Example ($n=4$, $r=2$).

Let $\delta = [\delta_{12}, \delta_{13}, \delta_{14}, \delta_{23}, \delta_{24}, \delta_{34}]^t$, and let $X = [X_1, X_{12}, X_{13}, X_{123}, X_{14}, X_{124}, X_{134}]^t$.

Then $\delta = MX$, where

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Let c^* denotes the number of sites where all taxa are not the same colour and let δ' be the vector obtained from δ by adjoining c^* as a seventh component. Let M_1 be the 7×7 matrix whose first six rows are M , and whose last row is the vector $[1, 1, \dots, 1]$. Then $\delta' = M_1 X$, and M_1^{-1} exists.

Indeed,

$$M_1^{-1} = 0.5 \begin{bmatrix} 0 & 0 & 0 & -1 & -1 & -1 & 2 \\ 0 & 1 & 1 & 1 & 1 & 0 & -2 \\ 1 & 0 & 1 & 1 & 0 & 1 & -2 \\ -1 & -1 & 0 & -1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 & 1 & 1 & -2 \\ -1 & 0 & -1 & 0 & -1 & 0 & 2 \\ 0 & -1 & -1 & 0 & 0 & -1 & 2 \end{bmatrix}$$

Now δ' is induced by a sequence space precisely if for $X = M^{-1}\delta'$, we have $X_\sigma \in \mathbb{Z}^+$ for all $\sigma \in \Omega(4)$. In view of the expression of M^{-1} this condition is equivalent to two conditions on the vector δ' .

- (i): $c^* \geq (\delta_{ij} + \delta_{jk} + \delta_{ik})/2 \in \mathbb{Z}^+$,
- (ii): (a) $c^* \leq (\delta_{13} + \delta_{14} + \delta_{23} + \delta_{24})/2 \in \mathbb{Z}^+$,
- (b) $c^* \leq (\delta_{12} + \delta_{14} + \delta_{23} + \delta_{34})/2 \in \mathbb{Z}^+$,
- (c) $c^* \leq (\delta_{12} + \delta_{13} + \delta_{24} + \delta_{34})/2 \in \mathbb{Z}^+$.

Notice that if δ is induced by a sequence space on two colours, δ does not determine c^* , though the conditions shown give bounds for c^* , such as $l(\delta) \leq c^* \leq 2.l(\delta)$, where $l(\delta) = \max_{i,j} \{\delta_{ij}\}$, a result required in section 7.

4.6 Theorem

(1): If $r=2$, δ satisfies the following property:

P: for any triple i,j,k , $\delta_{ij} + \delta_{jk} - \delta_{ik}$ is an even, non-negative integer.

(2): Any premetric on three or four points satisfying property P in (1) is induced by a sequence space on two colours.

(3): If δ is induced by a sequence space $(S, \mathcal{X}, c, r)$, $|S| = n$, then

$\sum_{i,j} \delta_{ij} \leq cn^2(1-r^{-1})$, and this bound may be realized for all values of c and r .

Proof:

(1): For $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2\}$, $t=1, \dots, c$, let $x_i(t)$ denote the t -th component of $x_i \in \mathcal{X}^c$.

For $i=1, \dots, k$, let $S(i) = \{t: x_i(t) = \mathcal{X}_1\}$. Then $\delta_{ij} = |S(i) \nabla S(j)|$, where ∇ denotes symmetric difference. The result now follows since for arbitrary sets, A and B , we have $|A \nabla B| = |A| + |B| - 2|A \cap B|$.

(2): The case for $n=3$ is straightforward. For $n=4$, we show that property P implies that δ , (together with a suitable value for c^*) satisfies the conditions in example 4.5, which are sufficient for δ to be induced by a sequence space on two colours.

If δ satisfies property P, then $(\delta_{ij} + \delta_{jk} + \delta_{ik})/2 = (\delta_{ij} + \delta_{jk} - \delta_{ik})/2 + \delta_{ik} \in \mathbb{Z}^+$

for all i,j , and condition (i) can be satisfied by taking $c^* =$

$\max_{i,j,k} \{(\delta_{ij} + \delta_{jk} + \delta_{ik})/2\}$. Furthermore for condition (ii)(a),

$$\delta_{13} + \delta_{14} + \delta_{23} + \delta_{24} = (\delta_{13} + \delta_{14} - \delta_{34}) + (\delta_{23} + \delta_{24} - \delta_{34}) + 2\delta_{34}$$

so if δ satisfies property P, $(\delta_{13} + \delta_{14} + \delta_{23} + \delta_{24})/2 \in \mathbb{Z}^+$, and similarly for

(ii)(b) and (ii)(c).

Finally since δ satisfies P, and thereby the triangle inequality, the expressions on the right hand side of (ii)(a),(b),(c) are all less than $\max_{i,j} \{(\delta_{ij} + \delta_{jk} + \delta_{ik})/2\} = c^*$, satisfying that the inequality conditions in (ii).

(3): Define $\epsilon(i,j,t) = 0$ if $x_i(t) = x_j(t)$, otherwise $\epsilon(i,j,t) = 1$.

Then $\delta_{ij} = \sum_t \epsilon(i,j,t)$, and so $\sum_{i,j} \delta_{ij} = \sum_{i,j} \sum_t \epsilon(i,j,t) = \sum_t \sum_{i,j} \epsilon(i,j,t)$

For $s=1, \dots, r$, let $N_s(t) = |\{i: x_i(t) = \gamma_s\}|$. Then

$\sum_{i,j} \epsilon(i,j,t) = \sum_{s,s': s \neq s'} N_s(t) N_{s'}(t) = n^2 - \sum_s N_s(t)^2$, since $\sum_s N_s(t) = n$.

Using this constraint, an elementary application of the method of Lagrange multipliers shows that $\sum_{i,j} \epsilon(i,j,t)$ is maximized when $N_s(t) = n/r$ for $s=1, \dots, r$, in which case $\sum_{i,j} \epsilon(i,j,t) = n^2(1-r^{-1})$ and the result follows. $\square$

Lower bounds on information loss

We now consider how much information is lost in converting sequence data to dissimilarities.

4.7 Theorem

(1): Let $h(n,c)$ denote the number of metrics d on $\{1, \dots, n\}$ of diameter $\leq c$. Then $h(3,c) = c(c^2 + 3c - 2)/2$.

(2): Let $h^*(n,c)$ denote the number of pseudometrics on $\{1, \dots, n\}$ of diameter $\leq c$.

Then $h^*(3,c) = h(3,c+1) - 3c(c+1)/2 = h(3,c)(c+1)/c$.

(3): Let $f(n,c)$ denote the number of metrics on $\{1, \dots, n\}$ of diameter $\leq c$ induced by some sequence space on two colours.

$$\text{Then } f(3,c) = \begin{cases} h(3,c)/2, & \text{if } c \text{ is even} \\ (c-1)(2c^2+11c+1)/8, & \text{if } c \text{ is odd.} \end{cases}$$

(4): For $n \equiv 1, 3 \pmod{6}$,

$$h(n,c) \leq h(3,c)^{n(n-1)/6}, \quad h^*(n,c) \leq h^*(3,c)^{n(n-1)/6},$$

$$f(n,c) \leq f(3,c)^{n(n-1)/6}.$$

(For other values of n similar results apply).

Proof:

(1): We distinguish three cases depending on d_{12}, d_{23}, d_{13} .

(a): all three distances are the same,

(b): two of the distances are the same and strictly larger than the other,

(c): the largest distance, denoted d_1 , is strictly larger than the other two, denoted d_2, d_3 .

Clearly (a) arises in c ways and (b) arises in ${}^3C_2 \cdot {}^cC_2 = 3c(c-1)/2$ ways (in these cases the triangle inequality is not a constraint). For case (c), let $d_2 + d_3 = k + d_1$. By the triangle inequality $k \geq 0$, while since $d_2, d_3 < d_1$, we have $k \leq d_1 - 2$. Now for fixed d_1 , and k in this range this equation has $d_1 - 1 - k$ solutions in d_2, d_3 with $d_2, d_3 < d_1$. Thus over the range $0 \leq k \leq d_1 - 2$ there are $d_1(d_1 - 1)/2$ solutions. Letting d_1 range from 2 to c and noting that

$\sum_{2 \leq j \leq c} jC_2 = {}^{c+1}C_3$, and that there are three possible choices (from d_{12}, d_{23}, d_{13}) for d_1 , we have $3{}^{c+1}C_3$ ways case (c) can arise. The result follows immediately.

(2): Given a metric of diameter $\leq c+1$, we obtain a pseudometric of diameter $\leq c$ by replacing $d(i,j)$ by $d(i,j)-1$, except when

$d(i,j) = d(i,k) + d(k,j)$, for some $i, j, k \in \{1, 2, 3\}$, which occurs in $3c(c+1)/2$

ways.

Conversely, if $d(i,j)$ is a pseudometric of diameter $\leq c$, $d(i,j)+1$ is a metric of diameter $\leq c+1$. The result follows immediately.

(3): Theorem 4.6 (parts (1) and (2)), give the additional constraint on d . The result now follows by the same decomposition as in part (1) of theorem 4.7, and a more lengthy, albeit elementary argument.

(4): First we recall the existence of a $2-(n,3,1)$ design when $n \equiv 1$ or $3 \pmod{6}$, noting that such a design has $n(n-1)/6$ blocks, (Hall [1986]). Thus $\{1, \dots, n\}$ can be expressed as the union of sets of $N = n(n-1)/6$ sets $A_1, \dots, A_N$, each of size 3, with $|A_i \cap A_j| = 1$, for $i \neq j$. Now a metric on $\{1, \dots, n\}$ is a metric on all the sets A_i (though clearly the converse does not always hold). Thus the number of metrics on $\{1, \dots, n\}$, $h(n, c)$ is bounded above by $h(3, c)^{n(n-1)/6}$, and similarly for h^* and f . §

4.8 Remark

This theorem shows that (for $n \equiv 1, 3 \pmod{6}$), the proportion of premetrics on n taxa of maximum diameter c , which are in fact metrics is bounded above by

$$\frac{(c(c^2+3c-2)/2)^{n(n-1)/6}}{c^{n(n-1)/2}}$$

which is asymptotically (in c), $2^{-n(n-1)/6}$. Similarly the proportion induced by a sequence space on two colours is (asymptotically in c) bounded above by $2^{-n(n-1)/3}$. Similar results hold for pseudometrics. We now apply theorem 4.7 to obtain a lower bound on a crude measure of how much information is lost in reducing sequence data to dissimilarities.

We adopt as our measure, the ratio of the number of "essentially distinct" sequences spaces to the number of distinct dissimilarity matrices $[\delta_{ij}]$ these sequences generate--this gives the average compression of the transformation.

We must first define when two sequence spaces are to be regarded as distinct, (the following definition is motivated by the fact that most tree-building methods do not take account of the order of the sites and treat the four nucleotide bases equally).

4.9 Definition

Call two sequence spaces Ω_1, Ω_2 of length c on the same set of colours $\mathcal{X}$, *equivalent* if Ω_1 is obtained from Ω_2 by permuting $\mathcal{X}$ and the order of the components in $\mathcal{X}^c$. Let $\Delta_r(n, c)$ be the set of equivalence classes of sequence spaces of length c on r -colours and n taxa.

4.10 Lemma

$$|\Delta_r(n, c)| = I(n, r) + c - 1 C_c, \text{ where } I(n, r) = \sum_t (t^n / t!) (d_{r-t} / (r-t)!),$$

with d_i the number of derangements on i objects.

(Note that $d_i = i! \sum_{0 \leq j \leq i} (-1)^j / j!$, as in Hall [1986]. Thus $I(n, 2) = 2^{n-1}$ and

$I(n, 4) = (2^{n-1} + 1)(2^{n-2} + 1)/3$, providing convenient expressions for the

sums of the first two columns, and first four columns respectively of the table given by Penny [1982, p.134].)

Proof:

The number of distinct ordered n -tuples $x \in \mathcal{X}^n$ under permutations of $\mathcal{X}$, is simply the number of orbits of $\mathcal{X}^n$ under the action of the full symmetric group G on $\mathcal{X}$. By Burnside's lemma this is $(r!)^{-1} \sum_{g \in G} F(g)$, where $F(g)$ is the number of elements of $\mathcal{X}^n$ fixed by g . Now G has exactly $r C_t d_{r-t}$ elements which fix precisely t elements of $\mathcal{X}$, and hence t^n elements of $\mathcal{X}^n$. Thus $I(n, r) = \sum_t (t^n / t!) (d_{r-t} / (r-t)!)$. §

4. 11 Example

Let $\Phi: \Delta_4(n, c) \rightarrow R^{n(n-1)/2}$ be the map transforming sequences into dissimilarities, and consider $|\Delta_4(n, c)| / |\Phi(\Delta_4(n, c))|$. By theorem 4.7, this ratio is at least $|\Delta_4(n, c)| / h^*(n, c)$, which we denote by $\tau(n, c)$. Table one lists values for $\log_{10} \tau(9, c)$. Thus for sequences of length 100 on 9 taxa there are, on average, at least 10^{178} essentially distinct sequence spaces to every dissimilarity matrix. A similar table has been published by Steel *et al.* [1988].

c		$\log_{10} \tau(9, c)$
15	...	8.5
20	...	18.5
25	...	28.6
50	...	79.9
100	...	178.0
250	...	436.8
500	...	798.9
1000	...	1390.5

Table One: Logarithms of lower bounds on the average information loss involved in converting sequence data of length c on nine taxa to dissimilarities.

4.12 Summary

This section characterizes dissimilarities arising from two-coloured sequences on four taxa (example 4.5 and theorem 4.6 (1,2)) and provides constraints in the more general cases of n taxa and r -colours (theorem 4.6 (1,3)). We also examine the redundancy involved in transforming sequences to dissimilarities by exploiting some of the constraints on induced dissimilarities. While we have neglected some constraints (such as theorem 4.6 (3)) it is likely that a more exact enumerative approach to this problem would be difficult and of little value since the lower bounds given in example 4.11 are already huge. A more valuable exercise would be to determine the nature of this redundancy, particularly as it relates to phylogeny reconstruction. Some initial work in this direction appears in Penny [1982].

§5: Combinatorics of Parsimony.

"Parsimony is another name for economic meanness," Richard Dawkins.

5.1 Introduction

Methods based on principles of parsimony are the most frequently-used method for constructing phylogenetic trees from aligned DNA sequence data (Felsenstein, [1988]). While the problem of finding the tree (or trees) of maximum parsimony is NP-hard (Foulds and Graham [1982]), a branch and bound algorithm has been developed by Hendy and Penny [1982], along with various heuristic methods. In this section, by restricting attention to two character states, we give conditions for when a set of edge changes is minimal for the colouring it induces. We then examine from a combinatorial viewpoint two distributions which arise, in the first case from fitting trees to data, and in the second case from fitting data to trees, in both cases so as to minimize the number of edge changes ("steps") required. Using results from section one, and another tree decomposition, we obtain useful descriptions of these distributions. Finally we set out a combinatorial result which is pivotal to much of the stochastic theory in section six and seven.

Vector spaces of edge sets

5.2 Definitions

For $T \in \text{BPT}(n)$, and $C \subseteq E(T)$, let $p(C) \in \Omega(n)$, denote the labels (including "1") of those pendant vertices of T whose unique path to the the pendant vertex labelled 1 passes through an even number of edges of C . Define an equivalence relation, $\sim$, on $2^{E(T)}$ by $C \sim C'$ precisely if $p(C) = p(C')$.

5.3 Lemma

If $C \sim C'$, $D \sim D'$ then $C \nabla D \sim C' \nabla D'$ where ∇ denotes symmetric difference.

Proof:

For pendant vertex of T labelled x and $C \subseteq E(T)$, let $f(x, C)$ be the number of edges of C on the unique path from 1 to x . Then

$$f(x, C \nabla D) = f(x, C) + f(x, D) - 2f(x, C \cap D) \equiv f(x, C) + f(x, D) \pmod{2}.$$

Now if $C \sim C'$, $D \sim D'$, then $f(x, C) \equiv f(x, C') \pmod{2}$ and $f(x, D) \equiv f(x, D') \pmod{2}$ so that $f(x, C \nabla D) \equiv f(x, C' \nabla D') \pmod{2}$, for all x , giving $C \nabla D \sim C' \nabla D'$. $\S$

5.4 Corollary

$C' \sim C$ precisely if $C' = C \nabla C_0$ where $C_0 \sim \Phi$.

Proof:

If $C' \sim C$ let $C_0 = C \nabla C'$. Then $C' = C \nabla C_0$ and by lemma 5.3,

$C_0 = C \nabla C' \sim C \nabla C = \Phi$. Conversely if $C_0 \sim \Phi$, by lemma 5.3, $C \nabla C_0 \sim C \nabla \Phi = C$. $\S$

Now $\Pi(T) = 2^{E(T)}$, under ∇ , forms a vector space over Z_2 . Let $\Pi_0(T) = \{C \in 2^{E(T)} : C \sim \Phi\}$. By lemma 5.3, $\Pi_0(T)$ forms a subspace of $\Pi(T)$. For $v \in IV(T)$, let $E(v)$ be the set of the three edges of T incident with v .

5.5 Lemma

$E(V) = \{E(v) : v \in IV(T)\}$ forms a basis for $\Pi_0(T)$.

Proof:

Regard a set $C \subseteq E(T)$ as inducing a two-colouring of all the vertices of T . Then for any fixed $\sigma \in \Omega(n)$, colouring the $n-2$ internal vertices of T in all 2^{n-2} possible ways gives 2^{n-2} edge sets E with $p(E) = \sigma$. Taking $\sigma = \sigma_0$, gives $|\Pi_0(T)| = 2^{n-2}$, and so $\Pi_0(T)$ has dimension $n-2$ over Z_2 . Since $E(V)$ has $n-2$ elements we need only check it is linearly independent.

If not, we have $\{E_1, \dots, E_k\} \subseteq E(V)$, $k \geq 1$, with $E_1 \nabla \dots \nabla E_k = \Phi$.

But for any $\{E_1, \dots, E_k\} \subseteq E(V)$, a simple inductive argument on k gives that

$$|E_1 \nabla \dots \nabla E_k| \geq k+2 \geq 3, \text{ as required. } \S$$

(Weakly-) Connecting trees and forests

5.6 Definitions

For $F \subseteq E(T)$, $T \in \text{BPT}(n)$, let

$w(F) = \min\{|F'| : F' \subseteq E(T) \text{ and } F' \sim F\}$, and

$L(F) = \{F' \subseteq E(T) : F' \sim F \text{ and } |F'| = w(F)\}$.

F is *minimal* if $F \in L(F)$, and *strictly minimal* if $L(F) = \{F\}$.

Given $F \subseteq E(T)$, and a (not necessarily binary) subtree T^* of T , let $\xi(F, T^*)$ denote the number of edges of F incident with a vertex of T^* . For a forest K of vertex disjoint subtrees of T define $\xi(F, K) = \sum_{T^* \in K} \xi(F, T^*)$.

5.7 Definitions

A *connecting tree for F* is a subtree T^* of T , not necessarily binary with $|V(T^*)| < 2\xi(F, T^*) - 2$.

A *weakly-connecting forest K* is a forest of k subtrees of T with

(1): no two trees both incident with any edge of T .

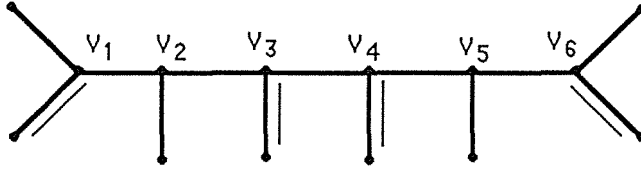
(2): $|V(K)| \leq 2\xi(F, K) - 2k$.

Let $\tau(F)$ be the set of weakly-connecting forests of F , including the empty forest Φ .

A *weakly-connecting tree* is a weakly-connecting forest, consisting of a single tree.

5.8 Example

Fig. 5.1 shows an example of a set $F \in \mathcal{CE}(T)$, $T \in \text{BPT}(8)$, with five weakly-connecting forests (in fact trees), given by $\emptyset$, (v_3, v_4) , (v_1, v_2, v_3, v_4) , (v_3, v_4, v_5, v_6) , $(v_1, v_2, v_3, v_4, v_5, v_6)$.



A tree with a minimal set F of four distinguished edges, having five weakly-connecting trees.

Figure 5.1

5.9 Theorem

For $F \in \mathcal{CE}(T)$, $T \in \text{BPT}(n)$,

- (1): F is minimal precisely if F has no connecting trees.
- (2): F is strictly minimal precisely if F has no weakly-connecting trees.
- (3): If F is minimal, there is a bijection between $L(F)$ and $\tau(F)$.

Proof:

(1): Suppose F has a connecting tree T^* on vertices $v_1, v_2, \dots, v_k$.

Let $C = E(v_1) \nabla \dots \nabla E(v_k)$. Then $|C \nabla F| = |C| + |F| - 2|C \cap F|$. Now $|C| = k+2$, while $2|C \cap F| = 2\xi(T^*, F) > k+2$, so that $|C \nabla F| < |F|$. Furthermore since $C \in \Pi_0(T)$, $C \sim \emptyset$, so that $C \nabla F \sim F$. Thus F is not minimal.

Conversely, suppose F is not minimal. Then for some $C \in \Pi_0(T)$, $|C \nabla F| < |F|$, so that $|C| < 2|C \cap F|$. By lemma 5.5 there is a unique subset $E(v_1), \dots, E(v_k)$ of $E(V)$ such that $C = E(v_1) \nabla \dots \nabla E(v_k)$.

Consider $V' = \{v_1, \dots, v_k\}$ and join v_i by an edge to v_j if v_i and v_j are adjacent in T . In this way V' defines a forest of subtrees $T_1, \dots, T_r$ of T . Let $n(i) = |V(T_i)|$, $\epsilon(T_i) = E(v_1') \nabla \dots \nabla E(v_{n(i)}')$, where $V(T_i) = \{v_1', \dots, v_{n(i)}'\}$. Then we have $|\epsilon(T_i)| = n(i) + 2$. Now $|C| = \sum_i |\epsilon(T_i)|$ and $|C \cap F| = \sum_i \xi(F, T_i)$ since for $i \neq j$, $\epsilon(T_i) \cap \epsilon(T_j) = \emptyset$. Thus, since $|C| < 2|C \cap F|$, $\sum_i (n(i) + 2 - 2\xi(F, T_i)) > 0$ so for at least one $i \in \{1, \dots, r\}$, $n(i) + 2 - 2\xi(F, T_i) > 0$, which, by definition, means that T_i is a connecting tree.

(2): This follows by essentially the same argument as for (1), using weakly-connecting trees rather than connecting trees.

(3): For $L \in L(F)$, we have $L \nabla F \sim \emptyset$ so that $L = F \nabla C$ where

$C = \nabla_{r=1, \dots, k} \nabla_{j=1, \dots, \alpha(r)} E(v_{r,j})$ where $\{v_{r,j} : j=1, \dots, \alpha(r)\}$ forms the

component trees of the forest K generated by $\cup_{r,j} \{v_{r,j}\}$ under the rule that v and v' are adjacent in the forest precisely if they are adjacent in T . Now since $L \in L(F)$, $|L| = |F| + |C| - 2|L \cap F| = |F|$, so that $|C| = 2|L \cap F|$. Also $|V(K)| = \sum_r \alpha(r)$, so that $|C| = \sum_r (\alpha(r) + 2) = |V(K)| + 2k$ and so

$2\xi(F, K) = |L \cap F| = |V(K)| + 2k$, so that $K \in \tau(F)$. This process is clearly invertible, establishing (c). §

5.10 Remark If F is minimal and $K_1, K_2 \in \tau(F)$, K_1 and K_2 are not necessarily vertex disjoint, as example 5.8 shows.

5.11 Theorem

- (a): If F is minimal and $A \subseteq F$ then A is minimal.
 (b): If F is strictly minimal and $A \subseteq F$ then A is strictly minimal.
 (c): If F is minimal and $A, B \subseteq F$, with $A \sim B$, then $A = B$.

Proof:

(a): Suppose A is not minimal. Then A has a connecting tree T^* by theorem 5.9. Now since $A \subseteq F$, $\xi(F, T^*) \geq \xi(A, T^*)$, so that T^* is a connecting tree for F , and so applying theorem 5.9 again, F is not minimal.

(b): Essentially the same argument as for (a), except using weakly-connecting trees.

(c): Suppose $A \sim B$, $A \neq B$. Then $A \nabla B \sim A \nabla A = \emptyset$, by lemma 5.3, and $A \nabla B \neq \emptyset$. Let $E = F \nabla (A \nabla B)$. Since $A \nabla B \sim \emptyset$, $E \sim F$, by corollary 5.4. But $A \nabla B \subseteq A \cup B \subseteq F$, so that $E = F - (A \nabla B)$ which has less edges than F since $|A \nabla B| > 0$. $\S$

We now give a sufficient condition for $FCE(T)$ to be strictly minimal by requiring the edges of F to be separated by at least three other edges.

5.12 Definition

For $\{x, y\} \subseteq V(T)$, let $\Delta(x, y)$ denote the number of edges of T on the path between x and y .

5.13 Lemma

Suppose a tree has n vertices, including a set S of $k \geq 2$ distinguished vertices with $\Delta(x,y) \geq 3$ for all $x,y \in S$. Then $n \geq 2k$.

Proof:

We use induction on k . For $k=2$ the tree has a path with at least 4 vertices so the result holds. Suppose the result holds for all $k \leq k_0 \geq 2$. For a tree T with a set S of k_0+1 distinguished vertices, select a vertex $v_0 \in S$, and an adjacent vertex v_1 not in S . Represent T as in fig. 5.2, where the dotted edges (and incident trees $T_1, \dots, T_r$, and $T_{r+1}, \dots, T_s$) may or may not exist depending on the degree of v_0 and v_1 .

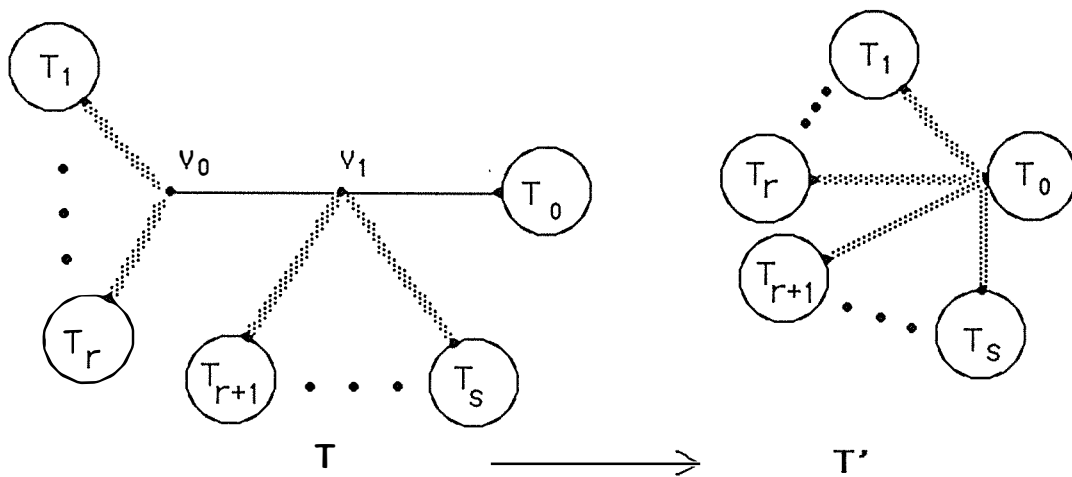


Figure 5.2

Replace T by the tree T' shown in fig. 5.2. Then T' has k_0-1 of the original distinguished vertices from T , and furthermore T' satisfies the conditions of the theorem, so that by induction T' has at least $2k_0$ vertices. But T has exactly two more vertices than T' , completing the induction step. §

5.14 Remark The bound $n = 2k$ is realized, for example, by a hair-comb-shaped tree with distinguished vertices of degree one.

5.15 Lemma

If F has a weakly-connecting tree T^* then F has a weakly-connecting tree T' with $E(T') \cap F = \emptyset$.

Proof:

(1) Suppose T^* weakly connects F , and $E(T^*) \cap F \neq \emptyset$. Delete $E(T^*) \cap F$ from $E(T^*)$ and for each deleted edge remove one of the two incident vertices from $V(T^*)$ (together with the other edge incident with that vertex) to obtain a forest of trees $T_1, \dots, T_r$, with $r \geq 2$.

Thus $|V(T^*)| = r + \sum_i |V(T_i)|$, and $\xi(F, T^*) = \sum_i \xi(F, T_i)$, and since

$|V(T^*)| \leq 2\xi(E, T^*) - 2$ we have

$\sum_i |V(T_i)| \leq \sum_i 2\xi(F, T_i) - 2 - r < \sum_i (2\xi(F, T_i) - 1)$. Thus for at least one i ,

$|V(T_i)| \leq 2\xi(F, T_i) - 2$, so that T_i is a weakly-connecting tree with

$E(T_i) \cap F = \emptyset$, as required. $\S$

5.16 Theorem

Let $F \subseteq E(T)$.

For $x, y \in F$, let $\Delta(x, y)$ denote the number of edges of T between x and y .

If $\Delta(x, y) > 2$ for all $x, y \in F$, F is strictly minimal.

Proof:

Suppose F is not strictly minimal. Then F has a weakly-connecting tree T^* , by theorem 5.9. By lemma 5.15 we may suppose $E(T^*) \cap F = \emptyset$.

Distinguish the set S of vertices of T^* that are incident with an edge of F . Then, by the Δ condition on F , $\Delta(x,y) \geq 3$ for all $x,y \in S$. By lemma 5.13, $|V(T^*)| \geq 2\xi(F, T^*)$, which contradicts the assumption that T^* is weakly-connecting. $\S$

5.17 Remark

It is immediate from theorem 5.9 that if F is minimal (resp. strictly minimal) then $\Delta(x,y) \geq 1$ (resp. ≥ 2). These are clearly not sufficient--indeed the stronger second condition is not sufficient to imply minimality, as illustrated by the tree T in fig. 5.2, with a set F of five edges for which $w(F) = 4$.

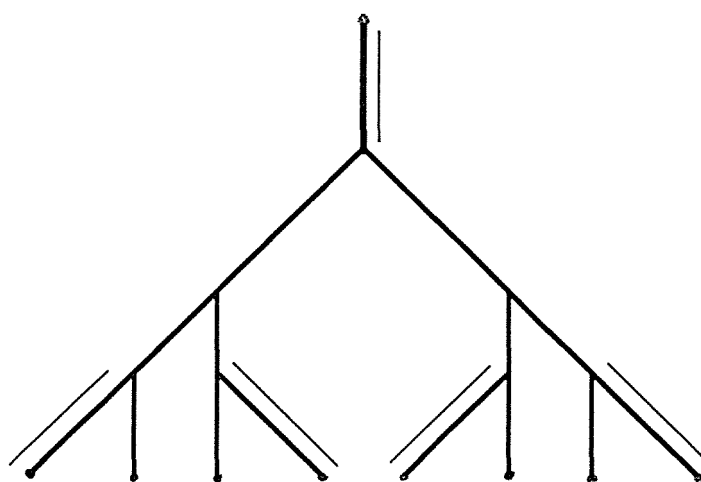


Figure 5.3

Distributions arising from parsimony

We now examine the distribution of binary phylogenetic trees according to the number of edge changes ("steps") required to fit data to the tree. An efficient method is given for calculating the mean of this distribution. This gives a simple measure of how much better a tree of maximum parsimony fits data than a randomly chosen tree. A useful invariant property of a related dual distribution is derived, allowing a complete description of this distribution and the calculation of its mean and variance. Either distribution gives a precise expression for the expected number of steps required to fit random data to any tree.

5.18 Definitions

(1): Given $\sigma \in \Omega(n)$, let $w(\sigma, T)$ be the minimum number of edge changes required to fit σ to T . Thus $w(\sigma, T) = \min\{|C| : C \subseteq E(T), p(C) = \sigma\}$. For a colouring $\chi \in C(n)$, $w(\chi, T)$ is defined similarly.

In this section *data* refers to a sequence space on two colours (definition 4.3), that is, an ordered sequence of two-colourings

$D = (D_1, \dots, D_c)$, $D_i \in C(n)$, and we let $|D|$ denote the length of the sequence (in this case $|D| = c$). Thus D might be a collection of aligned DNA sequences (where the four bases have been paired in some way). For convenience we do not impose the equivalence relation described in definition 4.9.

For data D , the minimum number of steps required to fit D to a fixed tree $T \in \text{BPT}(n)$ is $\sum_{1 \leq i \leq |D|} w(D_i, T)$, which we denote by $w(D, T)$.

(2): Fitch's algorithm.

For a colouring $\chi \in C(n)$ and $T \in \text{BPT}(n)$, Fitch [1971] (see also Hartigan [1973] for a formal justification) gives an efficient method for calculating $w(\chi, T)$, and for finding a minimal colouring of the internal vertices. We describe part of this procedure now. Root T on the midpoint of any edge and direct all edges away from the root, denoted v_0 , to give a rooted tree T^* . Assign to each internal vertex v the set $\{A\}, \{B\}$ or $\{A, B\}$ recursively as follows: pendant vertices are assigned the set containing their colour under χ , and for each vertex directed towards two vertices v_1, v_2 whose sets S_1, S_2 have already been chosen assign vertex v the set $S(v) = \begin{cases} S_1 \cap S_2; & \text{if this set is nonempty,} \\ S_1 \cup S_2; & \text{if } S_1 \cap S_2 = \emptyset. \end{cases}$

Then

(1): $w(\chi, T)$ is the number of internal vertices v (including the root v_0) for which $S(v)$ is defined by the second option ($S_1 \cap S_2 = \emptyset$).

(2): $S(v_0)$ is the set of colours v_0 can take over all colourings of minimal weight of the internal vertices of T^* . §

We examine the following two dual distributions:

(M*): For $T \in \text{BPT}(n)$, let $M(T, c, k) = |\{D: |D| = c \text{ and } w(D, T) = k\}|$.

(N*): For data D , let $N(D, k) = |\{T \in \text{BPT}(n): w(D, T) = k\}|$.

We consider first the distribution N^* .

For $|D| = 1$, $N(D, k) = f_k(m(D), n - m(D))$, where $f_k(a, b)$ is described in section one, and given by theorem 1.21. For $|D| > 1$, we now calculate the mean value of $N(D, k)$.

5.19 Definition

For a colouring χ , let $m(\chi) = \min\{|C_1|, |C_2|\}$, where C_1 (resp. C_2) is the set of elements from $\{1, \dots, n\}$ coloured A (resp. B) under χ .

For data D , let $X(D, a) = |\{i: m(D_i) = a\}|$ and let

$$\psi(n, a) = \sum_{k \geq 0} k \cdot f_k(a, n-a) / b(n).$$

Note that by theorems 1.21 and 1.14 (4), for any n , $\psi(n, 1) = 1$, while for fixed a , $\lim_{n \rightarrow \infty} \psi(n, a) = a$.

5.20 Theorem

Let $\mu(D)$ denote the mean value of $w(D,T)$ averaged over $BPT(n)$.

Then $\mu(D) = \sum_{a \geq 0} X(D,a) \psi(n,a)$.

Proof:

$$\mu(D) = b(n)^{-1} \sum_{\{T \in BPT(n)\}} w(D,T)$$

$$= b(n)^{-1} \sum_{\{T \in BPT(n)\}} \sum_{i \geq 0} w(D_i, T)$$

$$= b(n)^{-1} \sum_{a \geq 0} \sum_{k \geq 0} \sum_{\{T \in BPT(n)\}} \sum_{\{i: m(D_i)=a, w(D_i,T)=k\}} k$$

$$= b(n)^{-1} \sum_{a \geq 0} \sum_{k \geq 0} \sum_{\{i: m(D_i)=a\}} k \cdot |\{T \in BPT(n): w(D_i,T)=k\}|$$

$$= \sum_{a \geq 0} \sum_{k \geq 0} (k f_k(a, n-a) / b(n)) \cdot |\{i: m(D_i)=a\}|$$

$$= \sum_{a \geq 0} X(D,a) \psi(n,a). \quad \S$$

5.21 Remark

Theorem 5.20, together with theorem 1.21 and values for $\log(f_k(a, n-a))$

derived from $\log(t!) = \sum_{1 \leq i \leq t} \log(i)$, and $\log(b(t)) = \sum_{3 \leq i \leq t} \log(2i-5)$, give

values for $\psi(n,a)$ which can be efficiently calculated for large values of n , allowing $\mu(D)$ to be readily calculated in practice.

Table two (refer Appendix) lists the values of $\psi(n,a)$ for $n \leq 20$. $\S$

We now consider the dual distribution M^* .

5.22 Theorem

(1): Let $H_n(y) = \sum_{k \geq 0} n^{-k} c_k y^k$ and $G_n(y) = H_n(2y) + H_{n-1}(2y)$.

Then for any $T \in \text{BPT}(n)$, $M(T, c, k) = [y^k](G_n(y))^c$.

In particular, $M(T, 1, k) = (n^{-k} c_k + n^{-k-1} c_k) 2^k$.

(2): For $T \in \text{BPT}(n)$, the mean $\underline{\mu}(T, c)$, and variance $\underline{\sigma}^2(T, c)$ of the normalized distribution of $M(T, c, k) / \sum_{k \geq 0} M(T, c, k)$ are given by:

(a): $\underline{\mu}(T, c) = c \mu(T) / 2^n$, where $\mu(T) = \sum_{\chi \in C(n)} w(\chi, T)$,

(b): $\mu(T) = 2((3n-2)2^{n-1} + (-1)^n) / 9 \sim n \frac{2^n}{3}$ (as $n \rightarrow \infty$),

(c): $\underline{\sigma}^2(T, c) = c(6n+2 - (6n+1)(-2)^{1-n} - 2^{2-2n}) / 81 \sim 2cn / 27$ (as $n \rightarrow \infty$).

Proof:

Consider a fixed tree $T \in \text{BPT}(n)$ and a colouring χ .

Represent T as in fig. 5.4 (a) (refer to following page), where the endpoints of T_1, T_2 have the colouring induced by χ , and y, z denote the respective colouring by χ of the endpoints shown. Consider the two cases $y=z$ and $y \neq z$. In the first case if we replace T by the tree

$T' \in \text{BPT}(n-1)$ in fig. 5.4 (b), with the colouring χ^* as indicated, then $w(\chi, T) = w(\chi^*, T')$. Conversely, such a χ^* defines uniquely a colouring of T . In case $y \neq z$, replacing T by T'' with the colouring χ^{**} indicated in fig. 5.4 (c), so that $w(\chi, T) = w(\chi^{**}, T'') + 1$. In this case precisely two colourings of T induce the same χ^{**} corresponding to the cases $(y, z) = (a, b)$ and (b, a) .

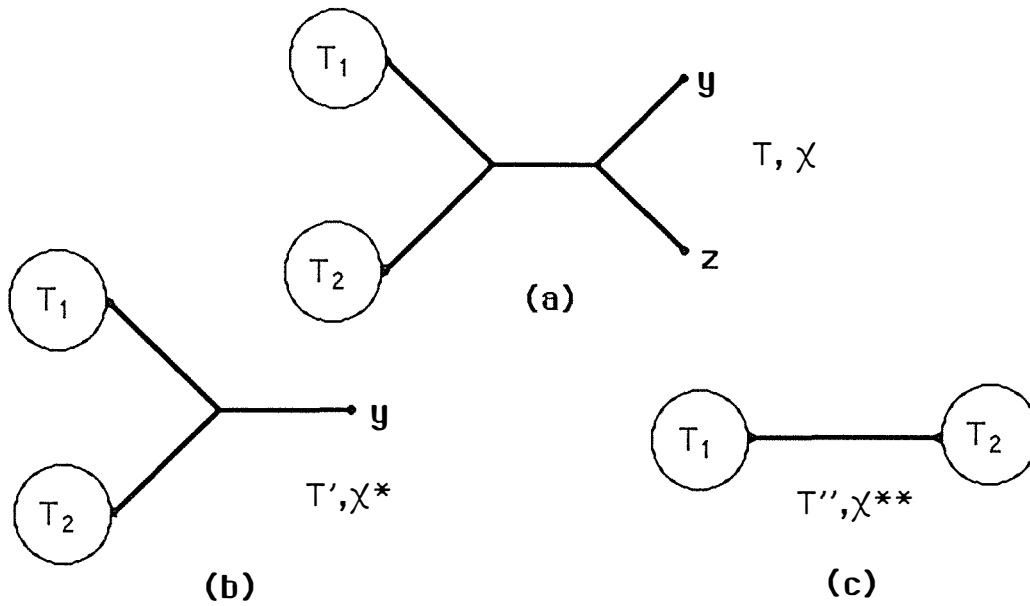


Figure 5.4

In this way, $M(T, 1, k) = M(T', 1, k) + 2M(T'', 1, k-1)$. Since T' and T'' have respectively $n-1$ and $n-2$ endpoints, induction on n shows that $M(T, 1, k)$ is the same for all $T \in \text{BPT}(n)$.

Thus we can let $M(n, 1, k)$ denote $M(T, 1, k)$ for any $T \in \text{BPT}(n)$, and let

$$G_n(y) = \sum_k M(n, 1, k) y^k \text{ giving } G_n(y) = G_{n-1}(y) + 2yG_{n-2}(y).$$

A simple inductive argument with $G_2(y) = 2+2y$, $G_3(y) = 2+6y$, shows that $G_n(y) = H_n(2y) + H_{n-1}(2y)$, where $H_n(x)$ is defined as in the statement of the theorem.

$$\text{Now } M(n, c, k) = \sum_{(k_1, \dots, k_c): k_1 + \dots + k_c = k} \prod_i M(n, 1, k_i),$$

so that $M(n, c, k)$ is the coefficient of y^k in $G_n(y)^c$, as required.

(2)(a): We first note that $\sum_k M(n, c, k) = G_n(1)^c = 2^{nc}$.

$$\text{Now } \underline{\mu}(T, c) = 2^{-nc} \partial_y G_n(y)^c \big|_{y=1} = 2^{-nc} c G_n(1)^{c-1} G_n'(1) = 2^{-n} c G_n'(1).$$

But, $\mu(T) = \sum_{\chi \in C(n)} w(\chi, T) = G_n'(1)$, establishing (a).

(b): From the above recurrence for $G_n(y)$,

$G_n'(1) = G_{n-1}'(1) + 2G_{n-2}'(1) + 2G_{n-2}(1)$. Now $G_n(1) = 2^n$, and the

solution to this recurrence can be shown by induction to be

$G_n'(1) = 2((3n-2)2^{n-1} + (-1)^n)/9$, which gives $\mu(T)$ as required.

(c): Similarly, $G_n''(1) = G_{n-1}''(1) + 2G_{n-2}''(1) + 4G_{n-2}'(1)$, which for

$G_2''(1) = G_3''(1) = 0$, and $G_n'(1)$ given as above, has solution

$G_n''(1) = 8(n-1)((3n-8)2^{n-3} + (-1)^n)/27$. The result for $\underline{\sigma}^2(T, c)$ follows by

noting that $\underline{\sigma}^2(T, c) = (2^{-cn} \partial^2 / \partial^2 y G_n(y)^c |_{y=1}) + \mu(T, c) - \mu^2(T, c)$, and

$$\partial^2 / \partial^2 y G_n(y)^c |_{y=1} = c(c-1)G_n(1)^{c-2}(G_n'(1))^2 + cG_n(1)^{c-1}G_n''(1). \quad \S$$

5.23 Remark

This independence of the underlying tree T in this theorem does not generalize to r -colourings of $BPT(n)$ or two-colourings on $PT(n)$. Indeed even $\mu(T, c)$ depends on T if $BPT(n) = PT(n, n-3)$ is replaced by $PT(n, n-4)$, as the counterexample $T_1, T_2 \in PT(6, 2)$ in fig. 5.5 shows.

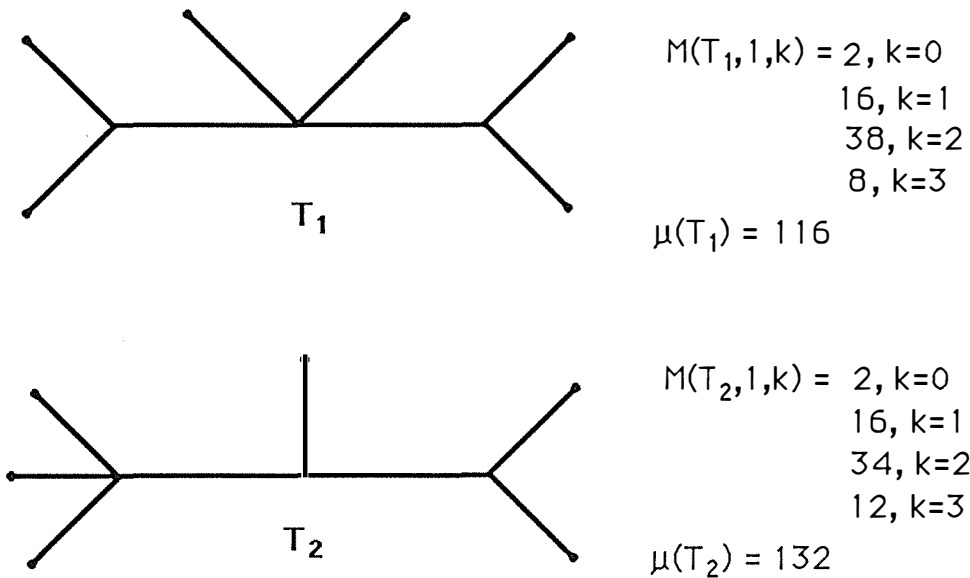


Figure 5.5

The next result gives an asymptotic expression for the expected weight of random data, answering a question posed by Joe Felsenstein and Jim Archie (private communication, Asilomar, June 1988).

5.24 Corollary

Let $w(D) = \min\{w(D, T) : T \in \text{BPT}(n)\}$, and let $w(c, n)$ be the average value of $w(D)$ over all sets of data of length c on n taxa.

Then $\lim_{c \rightarrow \infty} w(c, n)/c = ((3n-2) - (-2)^{1-n})/9$. Thus asymptotically

(in c) the average weight of random data on its best fit binary tree is the same as the weight of random data on any fixed binary tree (or indeed on a randomly-chosen binary tree).

Proof:

Since $w(D) \leq w(D, T)$ for any fixed $T \in \text{BPT}(n)$, part 2 of the previous theorem gives

$$w(c, n)/c \leq \underline{w}(T, c)/c = ((3n-2) - (-2)^{1-n})/9, \quad (1)$$

Let $c = k2^n$, and let $X(D, \chi)$ be the number of occurrences of colouring χ in D . For $\epsilon > 0$, let $G(n, c, \epsilon) = \{D : |D| = c, |X(D, \chi) - k| < c\epsilon \text{ for all } \chi \in C(n)\}$.

$$\text{Then } \lim_{c \rightarrow \infty} 2^{-nc} |G(n, c, \epsilon)| = 1. \quad (2)$$

This essentially follows from the weak law of large numbers, though we give a combinatorial proof as follows.

For any $\chi \in C(n)$ we have $|\{D : |D| = c, X(D, \chi) = r\}| = {}^c C_r (2^n - 1)^{c-r}$, so that

$2^{-nc} \sum_{\{D : |D| = c\}} (X(D, \chi) - k)^2 / c^2 = 2^{-nc} \sum_r (r - k)^2 / c^2 {}^c C_r (2^n - 1)^{c-r}$, which is asymptotically (in c) equal to 0.

Now

$2^{-nc} \sum_{\{D: |D|=c\}} (X(D, \chi) - k) / c^2 \geq \epsilon^2 2^{-nc} |\{D: |D|=c, |X(D, \chi) - k| \geq c\epsilon\}|$, so

that $\lim_{c \rightarrow \infty} 2^{-nc} |\{D: |D|=c, |X(D, \chi) - k| \geq c\epsilon\}| = 0$, and since

$G(n, c, \epsilon) = \{D: |D|=c\} - \cup_{\chi \in C(n)} \{D: |D|=c, |X(D, \chi) - k| \geq c\epsilon\}$, we have

$1 \geq 2^{-nc} |G(n, c, \epsilon)| \geq 1 - \sum_{\chi \in C(n)} 2^{-nc} |\{D: |D|=c, |X(D, \chi) - k| \geq c\epsilon\}| \sim 1$, as

$c \rightarrow \infty$, as required.

For each data set D choose $T(D) \in \text{BPT}(n)$ such that $w(D) = w(D, T(D))$.

Thus $T(D)$ is a maximal parsimony tree for D . Then

$w(c, n) = 2^{-nc} \sum_{\{D: |D|=c\}} w(D, T(D)) \geq 2^{-nc} \sum_{\{D \in G(n, c, \epsilon)\}} w(D, T(D)) \geq$

$2^{-nc} |G(n, c, \epsilon)| (k \sum_{\chi \in C(n)} w(\chi, T(D)) - c\epsilon(n/2) |C(n)|)$,

since $0 \leq w(\chi, T(D)) \leq n/2$ for any $\chi \in C(n)$.

Thus, by (2), $\lim_{c \rightarrow \infty} w(c, n) / c \geq k \sum_{\chi} w(\chi, T(D)) / c - \epsilon(n/2) 2^n$

$= ((3n-2) - (-2)^{1-n}) / 9 - \epsilon(n/2) 2^n$.

Since this holds for any $\epsilon > 0$ we have

$$w(c, n) / c \geq ((3n-2) - (-2)^{1-n}) / 9 \quad (3).$$

The result now follows from the inequalities (1) and (3). §

5.25 Remark

Regarding $\underline{\mu}(T, c)$ as the expected number of steps required to fit random data to a binary tree, we can compare $\underline{\mu}(T, c)$ with the expected number

$\underline{\mu}'(T_n, c)$ of steps required to fit random data of length c to a star tree

$T_n \in \text{PT}(n, 0)$ having no internal edges.

As before $\underline{\mu}'(T_n, c) = c^{\underline{\mu}'(T_n)} / 2^n$, where $\underline{\mu}'(T_n) = \sum_{\chi \in C(n)} w(\chi, T_n)$ and clearly, $\underline{\mu}'(T_n) = \sum_{0 \leq k \leq n} \min\{k, n-k\} n C_k$.

It can be shown (Michael Carter--personal communication) that this sum is $n(2^n - 2\delta_n) / 2$ where

$$\delta_n = \begin{cases} 2^k C_k; & \text{if } n = 2k+1, \\ (2^{-1/(k+1)})^{2k} C_k; & \text{if } n = 2k+2. \end{cases}$$

In either case, $\underline{\mu}'(T_n, c) \sim c^n / 2$, compared with $\underline{\mu}(T(n), c) \sim c^n / 3$ for a binary tree $T(n)$ in $BPT(n)$.

Table three lists $\underline{\mu}(T(n))$, $\underline{\mu}'(T_n)$ and their ratio.

n	$\underline{\mu}(T(n))$	$\underline{\mu}'(T_n)$	$\underline{\mu}(T(n)) / \underline{\mu}'(T_n)$
---	---	---	---
4	18	20	.9
5	46	50	.92
6	114	132	.8636
7	270	308	.8766
8	626	744	.8414
9	1422	1674	.8495
10	3186	3860	.8254
11	7054	8492	.8307
12	15474	19032	.8131
.			.
.			.
.			.
∞			$2/3$

Values for $\underline{\mu}(T(n))$ and $\underline{\mu}'(T_n)$, with their actual and asymptotic ratio.

Table Three

5.26 Remaining questions

It would be desirable to extend the above results in two directions. Firstly, the biologically-interesting case of four (or more generally $r > 2$) colours rather than just two might be considered. Secondly, an efficient means of calculating from data D the variance of the distribution of trees according to $w(D, T)$, even for two-colours, would be desirable. One approach would be to look for a suitable "closed form" analogue of theorem 1.6 for pairs of colours (with $m(a)$ and n replaced by four parameters one of which is required to measure the "overlap" of the two colourings).

Path/Edge Duality

5.27 Definitions

(a): Let $[n]^+$ denote the set of all even subsets of $\{1, \dots, n\}$.

For $X \in [n]^+$ and $T \in \text{BPT}(n)$, let $\pi(T, X)$ denote the set of edges $\bigvee_{x \in X} \pi(1, x)$, where $\pi(1, x)$ is the set of edges of T on the path joining vertices labelled 1 and x (with $\pi(1, 1) = \emptyset$), and $\bigvee$ denotes symmetric difference.

Since T is binary, the sets consisting of $\emptyset$ and the disjoint paths linking endpoints of T form a group under $\bigvee$, so that $\pi(T, X)$ is a set of disjoint paths, $\pi(x_1, x_2), \pi(x_3, x_4), \dots, \pi(x_{2k-1}, x_{2k})$, where $\cup_i \{x_i\} = \{1, \dots, n\}$.

(b): For $T \in \text{BPT}(n)$ with edge set $E_0 = E(T)$, assign an indeterminate x_e to each edge $e \in E_0$.

For $E \subseteq E_0$, let $[E, E_0] = \prod_{e \in E} x_e \prod_{e \in E_0 - E} (1 - x_e)$.

For $X \in [n]^+$, let $r_X = \prod_{e \in \pi(T, X)} (1 - 2x_e)$.

The next theorem, which is exploited in section six, was stated in a probabilistic setting by Hendy [1988]. The following proof is essentially a combinatorial restatement of the original proof, with minor variations.

5.28 Theorem

$$(1): \sum_{\{X \in [n]^+\}} (-1)^{|X \cap \sigma|} (-1)^{|X \cap \sigma'|} = \begin{cases} 0 & \text{for } \sigma \neq \sigma' \\ |[n]^+| = 2^{n-1} & \text{for } \sigma = \sigma'. \end{cases}$$

(Thus $H = [(-1)^{|\sigma \cap X|}]_{\sigma, X}$ is Hadamard).

(2): For any $\sigma \in \Omega(n)$ there is a formal power series identity:

$$\sum_{\{E \subseteq E_0: p(E) = \sigma\}} [E, E_0] \equiv 2^{1-n} \sum_{\{X \in [n]^+\}} (-1)^{|X \cap \sigma|} r_X.$$

Proof:

$$(1): \text{Clearly } |[n]^+| = \sum_{k \geq 0} {}^n C_{2k} = 2^{n-1}.$$

Now $[n]^+$ forms an abelian group under symmetric difference ∇ . For $\sigma_1, \sigma_2 \in \Omega(n)$, consider the homomorphism

$$\begin{aligned} \varphi: ([n]^+, \nabla) &\rightarrow (\{1, -1\}, \cdot) \\ \varphi(X) &= (-1)^{|\sigma_1 \cap X| + |\sigma_2 \cap X|} \end{aligned}$$

If $\sigma \neq \sigma'$, then φ is onto, since we can choose $x \in \sigma_1 - \sigma_2$ or $x \in \sigma_2 - \sigma_1$, and letting $X = \{1, x\}$ gives $\varphi(X) = -1$. Thus $|[n]^+| = 2 |\ker \varphi|$ and the result follows.

(2): We first show that for $E \subseteq E_0$,

$$|E \cap \pi(T, X)| \equiv |p(E) \cap X| \pmod{2} \quad (i)$$

$$\begin{aligned}
\text{We have } |E \cap \pi(T, X)| &= |\nabla_{x \in X} \pi(1, x) \cap E| = |\nabla_{x \in X} (\pi(1, x) \cap E)| \\
&= \sum_{x \in X} |\pi(1, x) \cap E| \pmod{2} = |X| - \sum_{x \in \{1, \dots, n\} - X} |\pi(1, x) \cap E| \\
&= \sum_{x \in \{1, \dots, n\} - X} |\pi(1, x) \cap E| \pmod{2} \text{ since } |X| \equiv 0 \pmod{2}.
\end{aligned}$$

Result (i) now follows since $x \in p(E)$ precisely if $|\pi(1, x) \cap E| \equiv 1 \pmod{2}$.

For $F \subseteq E_0$, and $x \in \{0, 1\}$, let

$$\lambda_F(x) = \sum_{\{E \subseteq E_0 : |E \cap F| \equiv x \pmod{2}\}} [E, E_0]. \quad (\text{ii})$$

Then

$$\lambda_F(0) + \lambda_F(1) = \sum_{\{E \subseteq E_0\}} \prod_{e \in E} x_e \prod_{e \in E_0 - E} (1 - x_e) = \prod_{e \in E_0} (x_e + (1 - x_e)) = 1. \quad (\text{iii})$$

Also, if $e \in E_0 - F$,

$$\begin{aligned}
\lambda_{F \cup \{e\}}(1) &= \sum_{\{E \subseteq E_0 : |E \cap (F \cup \{e\})| \equiv 1 \pmod{2}\}} [E, E_0] \\
&= \sum_{\{E \subseteq E_0 : e \in E, |E \cap F| \equiv 0 \pmod{2}\}} [E, E_0] \\
&\quad + \sum_{\{E \subseteq E_0 : e \in E_0 - E, |E \cap F| \equiv 1 \pmod{2}\}} [E, E_0] \\
&= x_e \sum_{\{E \subseteq E_0 - \{e\} : |E \cap F| \equiv 0 \pmod{2}\}} [E, E_0 - \{e\}] \\
&\quad + (1 - x_e) \sum_{\{E \subseteq E_0 - \{e\} : |E \cap F| \equiv 1 \pmod{2}\}} [E, E_0 - \{e\}]
\end{aligned}$$

Multiplying both terms by $(x_e + (1 - x_e))$ and expanding gives

$$\lambda_{F \cup \{e\}}(1) = x_e \lambda_F(0) + (1 - x_e) \lambda_F(1) \quad (\text{iv})$$

By (iii) and (iv) it follows by induction on $|F|$ that

$$\lambda_F(x) = \frac{1}{2} (1 + (-1)^x \cdot \prod_{e \in F} (1 - 2x_e)). \quad (\text{v})$$

Now we can re-write $\lambda_F(x)$ from (ii) as

$$\lambda_F(x) = \sum_{\sigma \in \Omega(n)} \sum_{\{E \subseteq E_0 : p(E) = \sigma \text{ and } |E \cap F| \equiv x \pmod{2}\}} [E, E_0] \quad (\text{vi})$$

Furthermore, in view of (iii),

$$\lambda_F(1) = \frac{1}{2} (1 - \lambda_F(0) + \lambda_F(1)). \quad (\text{vii})$$

Suppose $F = \pi(T, X)$. Then substituting (vi) into (vii) and using part (i),

$$\lambda_F(1) = 1/2(1 - \sum_{\sigma \in \Omega(n)} (-1)^{|\sigma \cap X|} \sum_{\{E \subseteq E_0: p(E)=\sigma\}} [E, E_0])$$

Identifying this expression for $\lambda_F(1)$ with that given by (v) gives

$$\sum_{\sigma \in \Omega(n)} (-1)^{|\sigma \cap X|} \sum_{\{E \subseteq E_0: p(E)=\sigma\}} [E, E_0] = r_X.$$

The result now follows by part (1). §

5.29 Corollary

Order the edges of T , and $\Omega(n)$ and let x, x' denote vectors of indeterminates indexed over the edges of T . Let $s(x)$ be the vector of values of $\sum_{\{E \subseteq E_0: p(E)=\sigma\}} [E, E_0]$, and similarly for $s(x')$.

We may suppose $\Omega(n)$ is ordered so that for $N = 2^{n-1}$, $s_N(x)$ corresponds to the uniform colouring induced by $E=\Phi$. Then under the usual inner product, $s(x).s(x') = s_N(z)$ with $z_e = x_e + x'_e - 2x_e x'_e$.

Proof:

Using both parts of theorem 5.28,

$$\begin{aligned} s(x).s(x') &= 2^{1-n} \sum_{\{X \in [n]^+\}} \prod_{e \in \pi(T, X)} (1-2x_e) \prod_{e \in \pi(T, X)} (1-2x'_e) \\ &= 2^{1-n} \sum_{\{X \in [n]^+\}} \prod_{e \in \pi(T, X)} (1-2x_e)(1-2x'_e) \\ &= 2^{1-n} \sum_{\{X \in [n]^+\}} \prod_{e \in \pi(T, X)} (1-2z_e) \\ &= s_N(z), \text{ by the second part of the theorem. } \S \end{aligned}$$

5.30 Summary

In this section we have presented a number of attractive and useful properties of parsimony on bicoloured binary trees. In particular we have seen that the distribution of data on such trees is independent of the tree topology, and have derived an exact expression for this distribution (theorem 5.22). This leads to an asymptotic expression for the weight of "random" data on its tree of maximal parsimony (corollary 5.24). We have shown the usefulness of the elegant formula from theorem 1.21, by giving an effective method for comparing the maximum parsimony tree with a randomly-chosen tree (theorem 5.20).

A number of structural results have also been derived by exploiting the vector space structure on edges and paths. A useful exercise would be the extension of these results (and the others discussed above) to r -coloured binary trees, for $r > 2$.

§6: Analysis of Cavender's model.

"The matter unfortunately comes down to questions of the philosophical foundations of statistics, which biologists are unlikely to resolve on their own," J. Felsenstein 1988.

"Every probability is in reality a conditional probability, " Alfred Renyi.

6.1 Introduction

A statistical approach to the problem of reconstructing phylogenies, presupposes a stochastic model describing how variations in sequence data arose between different taxa. This model M , will have as a discrete parameter the underlying tree T linking the taxa in question, and other (generally continuous) parameters $\Gamma \subseteq \Gamma_0$. For any given tree T and variable sequence data, X , the model $M = M(T, \Gamma)$ will assign a unique probability to the event of observing that X takes a particular value x , denoted $P(X=x; M(T, \Gamma))$. Clearly if we wish to be able to always infer the unknown underlying tree we require the following uniqueness condition:

$P(X=x; M(T, \Gamma)) = P(X=x; M(T', \Gamma'))$ for all x implies $T=T'$;

(for otherwise we could not stochastically distinguish T from T' , for appropriate Γ, Γ'). From here there are (at least) three directions which can be taken.

(1): Maximum Likelihood:

Given x and T let $f(T, x) = \max\{P(X=x; M(T, \Gamma)): \Gamma \subseteq \Gamma_0\}$ and let

$f(x) = \max\{f(T, x): T \in \text{BPT}(n)\}$. Maximum likelihood selects those trees T with $f(T, x) = f(x)$.

The concept of maximum likelihood in general has been championed by A.W.F. Edwards [1972] and frequently advocated for phylogenetic analysis by J. Felsenstein (see, for example, Felsenstein [1983]; for a history of this approach dating back to pioneering work by A.W.F Edwards and L.L.Cavalli-Sforza in 1964, see Felsenstein [1988]).

Maximum likelihood has desirable statistical properties, such as consistency and statistical efficiency. However the use of maximum likelihood for phylogenetic analysis raises fundamental difficulties (discussed briefly in section eight) in constructing confidence intervals. Determining exactly (rather than heuristically) the maximum likelihood tree, for a large number of taxa, is also computationally intensive.

(2): Prior distribution D on $M^* = \{M(T, \Gamma) : T \in \text{BPT}(n), \Gamma \subseteq \Gamma_0\}$:

Regarding $P(X=x; M(T, \Gamma))$ as a conditional probability $P(X=x | M(T, \Gamma))$, a distribution (with density function $f(T, \Gamma)$) on M^* allows the use of Bayes rule to calculate a modified (conditional) distribution.

We can then calculate the probability that any presented tree T_1 is the underlying tree, T given the data, as

$$P(T_1 = T | x) = \frac{\int_{\Gamma_0} f(T, \Gamma) P(X=x; M(T, \Gamma)) d\Gamma}{\sum_{T' \in \text{BPT}(n)} \int_{\Gamma_0} P(X=x; M(T', \Gamma)) f(T', \Gamma) d\Gamma}.$$

This approach, can be incorporated into a "Bayesian" approach to phylogenetic analysis, and can also be used classically by postulating a biologically-meaningful prior distribution D , based on branching processes.

This type of approach has been discussed by Felsenstein [1983, p. 248] and prior distributions on $BPT(n)$ motivated by biological considerations have been derived by Harding [1971]. One greatly appealing feature of this approach is that one can talk about, and in principle calculate $P(T_1 = T \mid \mathbf{x})$. A major disadvantage is that this approach involves two models, M and D , and so places more assumptions on nature than approaches (1) or (3). The calculation of $P(T=T_0 \mid \mathbf{x})$ would also be difficult in general. This approach and its difficulties have been referred to by Farris [1973] and Felsenstein [1983], and we shall not consider it further.

(3): Confidence intervals:

Given the data, $\mathbf{X}$, and given a pre-set $\alpha \in (0,1)$, confidence intervals of trees are constructed from $\mathbf{X}$ and α , so that no matter what values T and Γ take in $M(T, \Gamma)$, (1): α is a lower bound on the probability that the data stochastically generated by $M(T, \Gamma)$ will be such that the method will construct a confidence interval containing T , and (2): these confidence intervals eventually contain just T , with probability tending to one, as the length of the sequences tends to infinity.

For such confidence intervals to exist, one requires M to satisfy certain properties, for which sufficient conditions are certain independence assumptions discussed below. This approach has been successfully applied in special cases by Cavender [1978] and Felsenstein [1985], and is developed in section eight.

Stochastic preliminaries

6.2 Lemma

(1): For any family of events $\{E_1, \dots, E_n\}$ in a probability space,

$$P(\cap_i E_i) \geq \sum_i P(E_i) - n + 1.$$

Proof: $P(\cap_i E_i) = 1 - P((\cap_i E_i)') = 1 - P(\cup_i E_i') \geq 1 - \sum_i P(E_i') = 1 - \sum_i (1 - P(E_i)).$

(2): Markov's inequality: For a non-negative random variable X ,

$$P(X > \alpha) \leq E[X^k] / \alpha^k.$$

(3): If $X = (X_1, \dots, X_N)$ has a multinomial distribution with parameters λ and c , then:

(a): $\text{Var}[X_i] = c\lambda_i(1-\lambda_i)$, $\text{Cov}[X_i, X_j] = -c\lambda_i\lambda_j$, $i \neq j$.

(b): if g is a continuous real-valued function on X , then

$$\lim_{c \rightarrow \infty} P(|g(X/c) - g(\lambda)| < \epsilon) = 1 \text{ for all } \epsilon > 0.$$

(c): The conditional distribution of $(X_1, \dots, X_{N-1})$ given that $X_N = k$ is multinomial with parameters $\lambda/(1-\lambda_N)$ and $c-k$.

(d): let $\chi^2 = \sum_i (X_i - cs_i)^2 / cs_i$. Then with N small, c large and $\min_i \{cs_i\}$ not too small, (eg. $N = 4$, $c=1000$, $cs_i \geq 5$), χ^2 is approximated by a chi-square distribution with $N-1$ degrees of freedom.

Proof: For (a), (c) and (d) refer Johnson and Kotz [1969], for (b) refer Bishop, Findberg and Holland [1975, p. 465-472].

The Model

We assume the model proposed in Cavender's seminal paper [1978], which relates to earlier work by Farris [1973]. The model is based on simple if strong assumptions involving independence, which make the model at least partially tractable to analysis, as we show. The model also has a number of desirable properties, for example it satisfies (by lemma 6.13 (4)) the uniqueness condition described in the introduction.

Cavender's model assumes an underlying evolutionary tree T_0 linking the set of species under consideration, with internal vertices representing hypothetical ancestors. For each site z on the aligned DNA sequences, and at each internal vertex of T_0 it is assumed that the corresponding binary state, which belongs to the set $\{A,B\}$, is a random variable. On edge $e = [v_1, v_2]$ let $E(e, z)$ be the event that the state at site z and v_1 differs from the state at site z at v_2 . The original model has the following additional assumptions:

- (1): $\{E(e, z): 1 \leq z \leq c, e \in E(T_0)\}$ are independent events.
- (2): $\{E(e, z): 1 \leq z \leq c\}$ have the same probability, denoted $p(e)$.
- (3): $0 \leq p(e) \leq 0.5$.

6.3 Remark

If assumption (1) fails in general it may hold if the sites are chosen randomly and reasonably far apart. Assumption (3) expresses the belief that a change from one state to the other in a given unit of time is as likely as the reverse change--a belief partially supported by the relatively equal proportion of the four DNA bases.

6.4 Definitions

(1) A *weighted binary tree* is a pair (T, w) where $T \in \text{BPT}(n)$, and w is function from $E(T)$ to $\mathbb{R}^+ \cup \{0\}$. Let $\text{WBT}(n)$ be the set of weighted binary trees with n pendant vertices.

In Cavender's model we call $p(e)$ the *edge weight* and indexing $p(e)$ over the edges of any $T \in \text{BPT}(n)$ gives a vector p and a tree $(T, p) \in \text{WBT}(n)$.

(2): For $(T, p) \in \text{WBT}(n)$, and one site on the aligned DNA sequences, denote the probability of observing a given partition $\sigma \in \Omega(n)$ by $s(\sigma; T, p)$ (or just $s(\sigma)$ if the underlying tree is clear), and let $s(T, p)$, (or just $s(p)$) be the associated vector.

(3): Suppose the state along edge e of the underlying tree T_0 changes according to a Poisson process with rate μ_e . Then if t_e is the temporal length of edge e , $\mu_e t_e$ is the expected number of changes on e , and the endpoints of the e have a different state precisely if the number of changes is odd. It is easily shown that $\mu_e t_e = -0.5 \ln(1 - 2p(e))$, for $p(e) < 0.5$.

(4): Motivated by the Poisson model we call $q(e) = \mu_e t_e$ the *expected number of changes on edge e* , or more simply the *edge length of e* and let q be the associated vector of edge lengths.

(5): For convenience let $s(\sigma; T | q)$ denote the partition frequencies on T having edge weights for which q is the associated vector of edge lengths. Thus $s(\sigma; T | q) = s(\sigma; T, p)$ where $p_e = 0.5(1 - e^{-2q_e})$. Let $s(T | q)$ be the associated vector for σ indexed over $\Omega(n)$.

(6): Because temporal time is additive over the tree, the assumption that μ_e is the same on each edge of T ($\mu_e = \mu$) imposes linear restrictions on the components of q , (and thus polynomial restrictions on the components of p). Such a weighted tree is said to be subject to a *molecular clock* discussed by Zuckerkandl and Pauling [1962].

Central Observations:

6.5 Proposition

(1): For $(T, p) \in \text{WBT}(n)$, and aligned sequences of length c stochastically generated by (T, p) , let $X(c)$ be the random vector obtained by indexing the partition frequencies (the X_σ 's) over $\Omega(n)$. By assumption (2) of Cavender's model and the assumption of the independence of changes at different sites, (implied by assumption (1)), the probability of observing a given value x for X , $P(X=x; T, p)$, has a multinomial distribution, with parameters c and $s(T, p)$. Thus an event (i.e. a set of possible values for X), E , has probability

$$P(E; T, p) = \sum_{x \in E} P(X=x; T, p) = \sum_{x \in E} \binom{c!}{x!} s^x$$

(2): We have $s(\sigma; T, p) = \sum_{\{C \subseteq E(T): p(C)=\sigma\}} \prod_{e \in C} p_e \prod_{e \in E(T)-C} (1-p_e)$

by the assumption of independence of changes on different edges.

Theorem 5.28 gives a useful alternative description of $s(\sigma; T, p)$, and the corollary to that theorem, expressed in terms of edge lengths, becomes $s(T | q).s(T | q') = s(\sigma_0; T | q+q')$. In particular $\|s(T | q)\|^2 = s(\sigma_0; T | 2q)$.

(3): (Induced weighted subtrees and aggregate data)

Frequently we wish to build trees by deciding the structure of subtrees.

Thus given a tree $T \in \text{BPT}(n)$ with edge weights p and partition probabilities s we must relate these weights and probabilities to the partition

probabilities on the subtree $T[S]$ (defined in 3.2) induced by a subset $S \subseteq [n]_k$

of the taxa.

Let $A(e)$ be the path in T consisting of the set of edges which collapse to $e \in E(T[S])$ under the map from T to $T[S]$. Then by a straightforward argument (see for example Hendy [1988]), the probability that the endpoints of the path $A(e)$ are differently coloured is

$0.5(1 - \prod_{e' \in A(e)} (1 - 2p_{e'}))$. Thus the induced edge weights, denoted $p[S]$, and induced edge lengths $q[S]$ are given by

$$p[S]_e = 0.5(1 - \prod_{e' \in A(e)} (1 - 2p_{e'})), \quad q[S]_e = \sum_{e' \in A(e)} q_{e'}.$$

Given $S \in [n]_k$, to discuss the induced partition probabilities it is

convenient to relabel the taxa so that $S = \{1, \dots, k\}$.

The independence assumption in Cavender's model allows us to calculate $s(\sigma; T[S], p[S])$ from the $s(\sigma; T, p)$ by summing out all the possible states of pendant vertices not labelled from S . A straightforward argument gives the following result.

6.6 Proposition

For $\sigma_1 \in \Omega(k)$, $S = \{1, \dots, k\}$

$$s(\sigma_1; T[S], p[S]) = \sum_{\{\sigma \in \Omega(n): \sigma_1 \subseteq \sigma\}} s(\sigma; T, p).$$

6.7 Definition

Finally given partition frequencies X and $S = \{1, \dots, k\}$ we obtain induced partition frequencies $X[S]$, on S defined by $X[S]_\sigma = \sum_{\{\sigma' \in \Omega(n): \sigma \subseteq \sigma'\}} X_{\sigma'}$, the *aggregate of the data relative to S* . Aggregating data allows us to relate the probability of events generated by the weighted induced tree $T[S], p[S]$ to events generated by the weighted parent tree T, p .

Invariants

This section generalizes to WBT(n) results of Hendy [1988] and Cavender and Felsenstein [1987] which describe invariants of Cavender's model on WBT(4).

6.8 Definitions

(1): An *admissible set* S , is a collection of disjoint sets $X_1, \dots, X_k$ each belonging to $[n]^+$ (the set of even subsets of $\{1, \dots, n\}$).

(2): Given an admissible set $S = \{X_1, \dots, X_k\}$, $T \in \text{BPT}(L)$ and $e \in E(T)$ define $v(S, T, e)$ to be $|\{j: e \in \pi(T, X_j)\}|$.

(3): Define relations $\sim_T$ (resp. $<_T$) as follows:

$S \sim_T S'$ (resp. $S <_T S'$) precisely if for each $e \in E(T)$, $v(S, T, e) = v(S', T, e)$, (resp. $v(S, T, e) \leq v(S', T, e)$ and $v(S, T, e) < v(S', T, e)$ for at least one $e \in E(T)$).

(4): Given admissible sets S, S' let $x(\sigma)$ be an indeterminate for each $\sigma \in \Omega(n)$, and let $p(S, S')$ be the polynomial with integer coefficients defined by:

$$p(S, S')(x) = \prod_{X \in S} \sum_{\sigma} (-1)^{|\sigma \cap X|} |^{\sigma \cap X}|_{x(\sigma)} - \prod_{X \in S'} \sum_{\sigma} (-1)^{|\sigma \cap X|} |^{\sigma \cap X}|_{x(\sigma)}.$$

Note that no tree structure is used in defining this polynomial.

6.9 Theorem

For admissible sets (S, S') we have:

(1): If $S \neq S'$ then $p(S, S')$ is not the zero polynomial.

(2): $S \sim_T S'$ precisely if $p(S, S')(x) = 0$ for all $x = s(T, p)$, $p \in [0, 0.5]^{2n-3}$.

(3): $S <_T S'$, precisely if $p(S, S')(x) > 0$ for all $x = s(T, p)$, $p \in (0, 0.5)^{2n-3}$.

(4): The collection of all pairs of admissible sets (S, S') for which $S \sim_T S'$, characterizes T .

Proof:

(1): For $X \subseteq \{1, \dots, n\}$, let $q(X) = \sum_{\sigma \in \Omega(n)} (-1)^{|\sigma \cap X|} x(\sigma)$

and suppose $q(X_1) = \lambda q(X_2)$, for $\lambda = \pm 1$. Identifying the coefficient of $x(\sigma)$ for $\sigma = \{1\}, \{1, 2\}, \{1, 3\}, \dots, \{1, n\}$ in $q(X_1)$ and $q(X_2)$ gives:

$$X_1 = X_2, \text{ if } \lambda = 1; \quad X_1 = X_2 \cup \{1\} \text{ or } X_2 = X_1 \cup \{1\}, \text{ if } \lambda = -1. \quad (*)$$

Order $\Omega(n)$ and consider the ring of polynomials $Z[x(\sigma_1), \dots, x(\sigma_N)]$, $N = 2^{n-1}$.

Now for $S = \{X_1, \dots, X_k\}$, $S' = \{Y_1, \dots, Y_k\}$ we have

$P(S, S')(x) = \prod_{1 \leq i \leq k} q(X_i) - \prod_{1 \leq i \leq k} q(Y_i)$. Thus if $P(S, S')(x)$ is the zero

polynomial we have $\prod_{1 \leq i \leq k} q(X_i) = \prod_{1 \leq i \leq k} q(Y_i)$ in $Z[x(\sigma_1), \dots, x(\sigma_N)]$.

But this ring is a unique factorization domain (Herstein [1975, p.166]) and

for any $X \subseteq \{1, \dots, n\}$, $q(X)$ is a prime element of this ring, so that if

$p(S, S')(x) \equiv 0$ then for some permutation of τ of $\{1, \dots, n\}$, $q(X_i) = \pm q(Y_{\tau(i)})$,

for $i = 1, \dots, n$. Finally since X_i and $Y_{\tau(i)}$ are sets of even cardinality we

must have $X_i = Y_{\tau(i)}$ by $(*)$, as required.

(2):($\Rightarrow$) If $X \in [n]^+$, $\sigma \in \Omega(n)$, then by theorem 5.28,

$$\sum_{\sigma} (-1)^{|\sigma \cap X|} s(\sigma; T, p) = \prod_{e \in \pi(T, X)} (1 - 2p_e), \text{ so that}$$

$$\prod_{X \in S} \sum_{\sigma} (-1)^{|\sigma \cap X|} s(\sigma; T, p) = \prod_{e \in E(T)} (1 - 2p_e)^{v(S, T, e)}. \quad (*)$$

Since the right-hand side of this equation depends only on $v(S, T, e)$, the result follows.

($\Leftarrow$) For an edge $e_0 \in E(T)$ set $p_e = 0$, for all $e \neq e_0$, and set $p_{e_0} = (1 - \alpha)/2$,

giving $\prod_e (1 - 2p_e)^{v(S, T, e)} = \alpha^{v(S, T, e_0)}$. Thus if $p(S, S')(s) = 0$, the equation

(*) above gives $\alpha^{v(S, T, e_0)} = \alpha^{v(S', T, e_0)}$ so that for $\alpha \neq 0, 1$,
 $v(S, T, e_0) = v(S', T, e_0)$.

Since this holds for all $e_0 \in E(T)$, we have $S \sim_T S'$.

(3):($\Rightarrow$): The proof mirrors (2), noting that $(1 - 2p_e)^{v(S, T, e)} > (1 - 2p_e)^{v(S', T, e)}$ if $v(S, T, e) < v(S', T, e)$, for $p_e \in (0, 0.5)^{2n-3}$.

($\Leftarrow$): A similar proof as in (2), except that since $p \in (0, 0.5)^{2n-3}$, we must take $p_e = \epsilon > 0$ for all $e \neq e_0$ and $p_{e_0} = (1 - \alpha)/2$, $\alpha \neq 0, 1$, and let ϵ tend to zero.

(4): Suppose $\{(S_1, S_2): S_1 \sim_T S_2\} = \{(S_1, S_2): S_1 \sim_{T'} S_2\}$, where (S_1, S_2) are admissible sets. If $T \neq T'$ there exists by proposition 3.11 a set $S \in [L]_4$ with $T[S] \neq T'[S]$. Suppose $T[S] = (xy)(zw)$, while $T'[S] = (xz)(yw)$. Let $S_1 = \{\{x, y, w, z\}\}$ and $S_2 = \{\{x, y\}, \{w, z\}\}$. Then S_1, S_2 are admissible sets and $S_1 \sim_T S_2$ but $S_1 \not\sim_{T'} S_2$, which gives the required contradiction. $\S$

6.10 Example

For $n=4$, the admissible sets are Φ , $\{\{i,j\}\}$, $(i,j=1,\dots,4)$, $S_1 = \{\{1,2\},\{3,4\}\}$, $S_2 = \{\{1,3\},\{2,4\}\}$, $S_3 = \{\{1,4\},\{2,3\}\}$ and $S_4 = \{\{1,2,3,4\}\}$.

For $T = (12)(34)$, we have $S_1 \sim_T S_4 <_T S_2 \sim_T S_3$.

Order $\Omega(4)$ as $\{1\}$, $\{1,2\}$, $\{1,3\}$, $\{1,2,3\}$, $\{1,4\}$, $\{1,2,4\}$, $\{1,3,4\}$, $\{1,2,3,4\}$ and order $[4]^+$ as Φ , $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{1,4\}$, $\{2,4\}$, $\{3,4\}$, $\{1,2,3,4\}$.

Let $r = r(T,p)$ be the vector of values of $\prod_{e \in \pi(T,X)} (1-2p_e)$ with X indexed over $[4]^+$ ordered as above, so that the above relations become:

$r_2 r_7 = r_8$, $r_4 r_5 = r_3 r_6$, $r_8 \geq r_4 r_5$, as described by Hendy [1988]. Furthermore we have $r_1=1$, and $r_i \geq 0$ for all i . We now show that with two more conditions these are sufficient for a vector $r \in \mathbb{R}^8$ to be induced by edge change probabilities. This then characterizes the image of $\{s(T,p); p \in [0, 0.5]^5\}$, since s and r are related by an invertible linear transformation given in theorem 5.28.

6.11 Lemma

For $T = (12)(34)$, $\{r(T,p) : p \in [0, 0.5]^5\}$ is precisely the set of real 8-tuples satisfying:

(1): $r_1 = 1$, $r_i > 0$ for $i > 1$.

(2): (a) $r_2 \leq \min\{r_4/r_3, r_3/r_4\}$

(b) $r_7 \leq \min\{r_5/r_3, r_3/r_5\}$

(c) $r_8 \geq r_4 r_5$

(3): $r_2 r_7 = r_8$, $r_4 r_5 = r_3 r_6$.

Proof:

Necessity is easily checked, so suppose $r_1, \dots, r_8$ satisfies conditions

(1) to (3). Let $p_i = \ln(r_i)$, and let K be the matrix:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Now let D be the 5 by 5 submatrix of K on rows 2,3,4,5,8.

$$\text{Thus } D = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \quad \text{and } D^{-1} = 0.5 \begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}$$

Using (1) and (3) to eliminate p_1, p_6 and p_7 , and applying D^{-1} to the resulting subvector p' of p it follows from the conditions in (2) that

$D^{-1}p' \leq 0$ so taking $p_i = 0.5(1 - \exp(-D^{-1}p'_i))$ we have $0 \leq p_i < 0.5$ so that

$p = [p_1, p_2, p_3, p_4, p_5]$ can be regarded as edge change probabilities. But

$p = -2Kq$, where $q_i = -0.5 \ln(1 - 2p_i)$, so that if the edges of $T = (12)(34)$ are

ordered so that pendant edges have the same label as their incident

pendant vertex and the central internal edge is the fifth edge then p

induces $r(T, p)$, as required. $\S$

6.12 Remarks

(a): Conditions (1) to (3) imply $r_i \leq 1$ for $i = 1, \dots, 8$.

(b): For the other two trees in $BPT(4)$ one gets corresponding image sets in the natural way.

Partition probabilities

The following results are required in sections seven and eight.

6.13 Lemma

- (1): $s(\sigma_0; T, p)$ is a strictly decreasing function of each of the components of p .
- (2): For $\sigma \neq \sigma_0$, $s(\sigma_0; T, p) \geq s(\sigma; T, p)$, with equality precisely if $p = 0.5j$.
- (3): For fixed T , the function $p \rightarrow s(p)$ is one-to-one on $[0, 0.5]^{2n-3}$.
- (4): For $T \neq T'$, $p \in (0, 0.5)^{2n-3}$, $p' \in [0, 0.5]^{2n-3}$, we have $s(T, p) \neq s(T', p')$.
- (5): For any event $E \neq \emptyset$, if $p_e > 0$ on each pendant edge of T , then $P(E; T, p) > 0$.
- (6): $P(\sigma; T, p)$ is the same for all $T \in \text{BPT}(n)$ and all $\sigma \in \Omega(n)$, (in which case $P(\sigma; T, p) = 2^{-n+1}$), if and only if at least one edge on every path linking each pair of pendant vertices has weight 0.5.

Proof:

(1): By theorem 5.28, $s(\sigma_0) = 2^{1-n} \sum_{X \in [n]^+} r_X$, where

$r_X = \prod_{e \in \pi(T, X)} (1 - 2p_e)$, which immediately implies (1).

(2): By theorem 5.28, $s(\sigma) = 2^{1-n} \sum_{X \in [n]^+} (-1)^{|\sigma \cap X|} \cdot r_X$, and for $\sigma \neq \sigma_0$ there is a set $X \in [n]^+$ for which $|\sigma \cap X|$ is odd (indeed, by theorem 5.28 (1), $\sum_{X \in [n]^+} (-1)^{|\sigma \cap X|} = 0$). Comparing this with the expression in the proof of (1) for σ_0 gives (2).

(3): For $T \in \text{BPT}(n)$, let $K = [K(X, e)]_{X \in [n]^+, e \in E(T)}$ be the 2^{n-1} by $(2n-3)$ matrix defined by $K(X, e) = 1$ if $e \in \pi(T, X)$, 0 otherwise. Then $K^t K = 2^{n-3}(I + J)$, (Hendy [1988]) which is full rank. Hence K is one-to-one.

Now by theorem 5.28, $s = 1/8Hr$ with $H = [(-1)^{|\sigma \cap X|}]_{\sigma, X}$ (as in theorem 5.28 (1)), $r = \exp(-2Kq)$, where q , as in definition 6.4 (4), is given by $q_e = -0.5 \ln(1-2p_e)$ (it is here that $p_e < 0.5$ is required).

Thus s is a composition of one-to-one functions, and so is one-to-one.

(4): Suppose $T \neq T'$. Then there exists a set $S \subseteq L$, $|S| = 4$ with $T[S] \neq T'[S]$, by proposition 3.11. By proposition 6.5 (3), it suffices to prove the result for WBT(4). But this follows immediately from the invariants described in example 6.10.

(5): Since $E \neq \emptyset$, we can choose $x \in E$. Then

$P(E; T, p) \geq P(X=x; T, p) = \binom{m}{|x|} s^x$. Now for any $\sigma \in \Omega(n)$, let $F(\sigma)$ be the set of pendant edges incident with the pendant vertices having labels in σ .

Then $s(\sigma) \geq \prod_{e \in F(\sigma)} p_e \prod_{e \in E(T) - F(\sigma)} (1-p_e) > 0$. Thus $s(T, p) > 0$ so that $P(E; T, p) > 0$, as required.

(6): By theorem 5.28, if $s(\sigma_0) = 2^{-n+1}$, then

$1 = \sum \{X \in [n]^+\} r_X = 1 + \sum \{X \in [n]^+ : X \neq \emptyset\} r_X$, so that $(1-2p_e) = 0$ for at least one $e \in \pi(T, X)$ and all $X \in [n]^+$.

Conversely, suppose this condition holds. Then for any $\sigma \in \Omega(n)$, by theorem 5.28, $s(\sigma) = 2^{-n+1} \sum \{X \in [n]^+\} (-1)^{|\sigma \cap X|} r_X = 2^{-n+1}$. §

Finally we strengthen part (4) of the previous lemma to obtain a result which is required in section seven. Define $\hat{s}(T, p)$ to be the vector $s(T, p)$ with the component $s(\sigma_0; T, p)$ deleted. We show $\hat{s}(T, p)$ defines T .

6.14 Theorem

If $T \neq T'$, $p, p' \in (0, 0.5)^{2n-3}$, $\beta \in \mathbb{R}$, then $s^\wedge(T, p) \neq \beta s^\wedge(T', p')$.

Proof:

As in part (4) of the previous theorem it suffices to prove this result for $n=4$. Let $T = (12)(34)$, $T' = (13)(24)$, and order the paths on T and T' as $\pi(T, X)$ (resp. $\pi(T', X)$) where the sets X are ordered as in example 6.10.

Let $s_1 = s^\wedge(T, p)$, $s_1' = s^\wedge(T', p')$, and suppose $s_1 = \beta s_1'$.

Then $s(T, p) = [s_1, s_0]^t$, $s(T', p') = [\beta s_1', s_0']^t$, (where $[x, a]^t$ is the vector obtained from x by adjoining a).

Since the components of both vectors sum to 1 we have

$$\beta = (1 - s_0') / (1 - s_0).$$

Using the notation of example 6.10, we have $r(p) = H^t[s_1, s_0]^t$. If H_1 is the submatrix of H on the first 7 rows, we have $r(T, p) = H_1^t s_1 + s_0 j$ (where $j = [1, \dots, 1]^t$), while $r(T', p') = \beta H_1^t s_1' + s_0' j$.

Thus $r_i' = \beta r_i + s_0' - \beta s_0 = \beta r_i + \lambda$, where $\lambda = (s_0' - s_0) / (1 - s_0)$.

We show $\beta > 1$ and $\beta < 1$, a contradiction. First since $r_8' = r_3' r_6'$ we have

$\beta^2 r_3 r_6 + \beta \lambda (r_3 + r_6) - \beta r_8 + \lambda^2 - \lambda = 0$. Now $\beta + \lambda = 1$, so this equation becomes:

$\beta(\beta(r_3 r_6 - r_3 - r_6 + 1) + (r_3 + r_6 - r_8 - 1)) = 0$. Since $\beta \neq 0$, we have

$$\beta = (1 + r_8 - r_3 - r_6) / (1 + r_3 r_6 - r_3 - r_6) = (1 + r_8 - r_3 - r_6) / (1 - r_3)(1 - r_6).$$

Now $r_8 > r_3 r_6$ so that $(1 + r_8 - r_3 - r_6) > (1 + r_3 r_6 - r_3 - r_6) = (1 - r_3)(1 - r_6)$. Thus the numerator of β exceeds the denominator, which is positive since $r_3, r_6 < 1$, giving $\beta > 1$. Now this holds for any p and p' and if $s^\wedge(T, p) = \beta s^\wedge(T', p')$ we have $s^\wedge(T', p) = \beta^{-1} s^\wedge(T, p)$. Thus repeating the argument (replacing T and T') we would deduce that $\beta^{-1} > 1$, which contradicts $\beta > 1$, as required. $\S$

6.15 Theorem

Any linear combination of the coordinates of $\mathbf{s}(T, \mathbf{p})$ is nonzero except on an algebraic subset of $\mathbf{p}$ values in $[0, 0.5]^{2n-3}$.

(Recall that $U \subset \mathbb{R}^k$ is algebraic precisely if $U = \{\mathbf{x} : p(\mathbf{x}) = 0\}$ for some non-zero polynomial $p \in \mathbb{R}[x_1, \dots, x_k]$).

Proof:

Suppose $\sum_{\sigma} \lambda_{\sigma} s(\sigma; T, \mathbf{p}) = 0$ for all $\mathbf{p} \in [0, 0.5]^{2n-3}$. Applying theorem 5.28,

$\sum_{\{X \in [n]^+\}} d_X r_X = 0$ for all $\mathbf{p}$, where $d_X = \sum_{\sigma} (-1)^{|\sigma \cap X|} \lambda_{\sigma}$. We show $d_X = 0$

by induction on $|X|$. For $|X| = 0$, let $p_e = 0.5$ for all $e \in E(T)$, giving $d_{\emptyset} = 0$.

For $|X| = 2k$, set $p_e = 0$ for $e \in \pi(T, X)$, and $p_e = 0.5$ for $e \in E(T) - \pi(T, X)$. Then

$d_X + \sum_{X' \subset X, X' \neq X} d_{X'} = 0$. Since the sets being summed over in this

equation have cardinality less than $2k$, we can apply the induction

hypothesis to deduce that $d_X = 0$. Thus for all $X \in [n]^+$, $d_X = 0$, and since λ_{σ}

$= 2^{1-n} \sum_{\{X \in [n]^+\}} (-1)^{|\sigma \cap X|} d_X$, we have $\lambda_{\sigma} = 0$ for all σ , a contradiction.

Thus $f(\mathbf{p}) = \sum_{\sigma} \lambda_{\sigma} s(\sigma; T, \mathbf{p}) \in \mathbb{R}[p_1, \dots, p_{2n-3}]$, with $f(\mathbf{p}) \not\equiv 0$, so that $\mathbf{p}$

lies in an algebraic set. $\square$

The previous theorem implies that, except for exceptional values for $\mathbf{p}$,

$\sum_{\sigma} c_{\sigma} s(\sigma; T, \mathbf{p}) = x$, implies $c_{\sigma} = x$. Thus $\sum_{\sigma} s(\sigma; T, \mathbf{p}) = 1$ is essentially the

only linear relationship between the components of $\mathbf{s}$ (except for exceptional values of $\mathbf{p}$).

We now give bounds on $s(\sigma_0; T, \mathbf{p})$, the probability of observing a uniform colouring on $(T, \mathbf{p})$.

Choose a pendant vertex v_1 of T as root, and direct all edges of T away from v_1 . Let p_1 be the weight of the pendant edge incident with v_1 and for each internal vertex $v \in IV(T)$ let $\{v(1), v(2)\}$ be the two edges of T incident with, and directed away from v .

6.16 Lemma

$$q_1 \prod_{v \in IV(T)} q_{v(1)} q_{v(2)} \leq s(\sigma_0; T, p) \leq q_1 \prod_{v \in IV(T)} (q_{v(1)} q_{v(2)} + p_{v(1)} p_{v(2)}),$$

where $q_e = 1 - p_e$.

Proof:

The left hand inequality is just the statement that a sufficient condition for all the pendant vertices of T to be the same colour is that there are no colour changes on any of the edges of T . The right hand inequality is proved by induction on n . For $n \leq 3$ the result holds. Suppose the result is true for all $n \leq k$, $k \geq 3$. For $T \in BPT(k+1)$, choose a pendant vertex v_1 . Then since every $T \in BPT(n)$, $n \geq 4$ has at least two binary edges (definition 2.17), we can choose two pendant vertices v_2, v_3 , both distinct from v_1 , which have pendant edges incident with a common internal v of T . Deleting v_2 and v_3 and their incident edges, and labelling v and the other pendant vertices from the set L_k gives a tree $T' \in BPT(k)$. If σ_1 is the partition of T' corresponding to all labels other than the label on v , and p' is the restriction of p to T' , we have

$$s(\sigma_0; T, p) = q_{v(1)} q_{v(2)} s(\sigma_0; T', p') + p_{v(1)} p_{v(2)} s(\sigma_1; T', p').$$

But by lemma 6.13 (2), $s(\sigma_1; T', p') \leq s(\sigma_0; T', p')$, which establishes the induction step. §

6.17 Corollary

If each edge of $T \in \text{BPT}(L)$ has edge weight $\geq p$, then

$$s(\sigma_0; T, p) \leq q(q^2 + p^2)^{n-2}, \text{ where } n = |L|, q = 1 - p.$$

Proof:

By lemma 6.13 (1) we may assume each edge has weight p . The result now follows from the previous lemma. §

6.18 Lemma

For $k > 0, C > k/2$, let $p(n) = C \ln(n)/n$, $S(n) = n^k(1 - 2p(n) + 2p^2(n))^{n-2}$.

Then $S(n) \sim 0$.

Proof: Choose n sufficiently large so that $x(n) = 2p(n) - 2p^2(n)$ lies between 0 and 1. Then since $\ln(1-y) \leq -y + y^2/2$ for $y \in (0,1)$,

$$\ln(S(n)) \leq k \ln(n) + (n-2)(-x(n) + x^2(n)/2) \rightarrow -\infty \text{ as } n \rightarrow \infty, \text{ as required. §}$$

The last theorem in this section is an illustration of the use of some of the above results, and is referred to in section eight. We show that if the number of sites grows no faster than a polynomial function of the data, and the edge lengths of the underlying trees don't go to zero too quickly then with increasing certainty, all partitions (including the uniform partition σ_0) will occur at most once.

6.19 Theorem

For $(T(n), p(n)) \in \text{WBT}(n)$, suppose

(1): $\min\{np_e(n)/\ln(n) : e \in E(T(n))\} \rightarrow \infty$, as $n \rightarrow \infty$,

(2): c grows as a polynomial function of n .

Let A_n be the event that $X_\sigma(c) \leq 1$ for all $\sigma \in \Omega(n)$.

Then $P(A_n; T(n), p(n)) \sim 1$.

Proof:

Let $Z(X) = \sum_{\sigma \in \Omega(n)} (X_\sigma^2 - X_\sigma)$,

Thus on $(T, p) \in \text{WBT}(n)$, $E[Z(X)] = \sum_{\sigma \in \Omega(n)} E[X_\sigma^2] - \sum_{\sigma \in \Omega(n)} E[X_\sigma]$.

Now $\sum_{\sigma \in \Omega(n)} E[X_\sigma] = c$, the number of sites being sampled,

while, $\sum_{\sigma \in \Omega(n)} E[X_\sigma^2] = \sum_{\sigma \in \Omega(n)} (\text{Var}[X_\sigma] + E[X_\sigma]^2)$

$= \sum_{\sigma \in \Omega(n)} cs(\sigma)(1-s(\sigma)) + c^2 s(\sigma)^2$, by lemma 6.2 (3a).

Thus $E[Z(X)] = c + (c^2 - c) \sum_{\sigma \in \Omega(n)} s(\sigma)^2 - c \leq c^2 \sum_{\sigma \in \Omega(n)} s(\sigma)^2$

By corollary 5.29,

$\sum_{\sigma \in \Omega(n)} s(\sigma; T, p)^2 = s(\sigma_0; T, p')$, where $p_e' = 2p_e(1-p_e)$, and from corollary

6.17, $s(\sigma_0; T, p') \leq \alpha^{n-2}(1-p')$, where $\alpha = 1-2p'+2p'^2$, and $p' = 2p(1-p)$, and

$p = \min\{p_e : e \in E(T)\}$.

Applying this to $T(n), p(n)$ gives

$E[Z(X)] \leq c^2 \alpha^{n-2}(1-p'_n) \leq c^2 \alpha^{n-2}$, where $\alpha = 1-2p'_n+2p_n'^2$, and

$p'_n = 2p_n(1-p_n)$, $p_n = \min\{p_e(n) : e \in E(T(n))\}$.

Now since $np'_n / \ln(n) > np_n / \ln(n) \rightarrow \infty$ as $n \rightarrow \infty$, if $c = O(n^k)$, then by lemma

6.18, $\lim_{n \rightarrow \infty} E[Z(X); T(n), p(n)] = 0$.

Since $Z(X) \geq 0$, Markov's inequality, (lemma 6.2 (2)) gives
 $P(Z(X) \geq 1; T(n), p(n)) \leq E[Z(X); T(n), p(n)]$. Thus $P(Z(X) \geq 1; T(n), p(n)) \sim 0$. But
 A_n , is precisely the event that $Z(X) < 1$, which gives the required result. §

6.20 Summary

Cavender's model is at the same time simple yet powerful, being based on assumptions of independence. These reduce much of the theory in section seven to the analysis of the partition frequencies $\mathbf{s}$. Cavender's model guarantees that these partition frequencies characterize not only the edge weights on a tree (lemma 6.13 (3)) but also the tree itself (lemma 6.13 (4), theorem 6.9), allowing for the consistent recovery of trees from the sequence data they generate. Indeed not all of the $\mathbf{s}$ vector is necessary to characterize the tree (theorem 6.14). It would be interesting to know just how much of $\mathbf{s}$ characterizes T and/or its edge weights. For example one might consider just the partitions induced by deleting edges of T .

§7: Consistency

"Nothing general is known about what the conditions for consistency (of parsimony) might be..." J. Felsenstein 1988.

7.1 Introduction

Many of the methods used to build phylogenetic trees are based on the principle of parsimony. One version of this principle, sometimes called Wagner parsimony, is to fit data to that tree (or trees) requiring the fewest number of mutations on its edges. Another version, sometimes called compatibility, or the "clique method", is that data should be fitted to the tree(s) having the largest number of sites which fit the tree with at most one edge change (i.e. without duplications). A major motivation for these principles is the belief that changes (mutations) at DNA sites are rare, so that only changes that are absolutely necessary to explain the variation in the data should be allowed. Underlying this belief of rare events is the inherent assumption of a probability model, which we take to be Cavender's model. In this case, Felsenstein [1978 (2)] has shown that with four taxa Wagner parsimony and compatibility are not always consistent. Consistency is the desirable statistical property that as the number of DNA sites sampled tends to infinity, the probability that the chosen tree (derived from data generated under Cavender's model) is the underlying tree that produced the data, tends to 1.

Felsenstein's example requires an interplay of short and long edges to achieve inconsistency. We first exploit the essentials of why his counterexample works to show that two general classes of tree-building methods will fail to be consistent in general on four taxa. Then concentrating on parsimony, and extending the analysis to trees on n taxa, we show that even if all edges are the same length consistency can fail. In the process we show that Wagner parsimony can fail to be consistent under conditions where compatibility will be consistent. It is shown that imposing a molecular clock does not improve matters, extending work by Hendy and Penny [1988].

A second major result in this section is the establishing of sufficient conditions for the consistency of compatibility on all trees. The corresponding question for Wagner parsimony is also considered, though it appears more difficult.

We begin by making precise what we mean by a method for choosing trees, and then derive some general results before concentrating on parsimony.

Selection Procedures

7.2 Definition

A *selection procedure* φ , is a function, $\varphi: \Delta_2(n,c) \rightarrow 2^{BPT(n)}$, where $\Delta_2(n,c)$ is as defined in 4.9.

7.3 Remark

Certain statistical selection procedures (for example the one described by Cavender, [1978]) may not select a tree (corresponding to the $\{\emptyset\}$ option) if the data significantly rejects all trees, suggesting that the model itself is in error. At the other extreme all trees may be selected if there is insufficient data to reject any tree.

Convergence and Consistency

7.4 Definitions

Let $X(c)$ be a sequence space of length c stochastically generated by (T_0, p) . A selection procedure φ *converges* to $T \in BPT(n)$ on (T_0, p) if $\lim_{c \rightarrow \infty} P(\varphi(X(c)) = \{T\}; T_0, p) = 1$.

A selection procedure φ is *consistent* on (T_0, p) if φ converges to T_0 .

A selection procedure is *always consistent* if it is consistent on (T, p) for all $T \in BPT(n)$ and all $p \in (0, 0.5)^{2n-3}$.

7.5 Remark

An alternative definition of consistency, (implicitly suggested in

[Felsenstein, 1978 (2)]) is $P(\lim_{c \rightarrow \infty} \varphi(X(c)) = \{T_0\}; T_0, p) = 1$.

This says that with probability 1, there exists a natural number c_0 such that if $X(c)$ are the partition frequencies for the initial segment of length c of a collection of infinitely-long, aligned sequences, randomly generated under Cavender's model, then $\phi(X(c)) = \{T\}$, for all $c > c_0$. Thus the selection procedure is guaranteed to eventually converge to the true tree and remain so from that point on. While the results in this section will work under either definition, we will take consistency to be the former version since under the latter version only the existence c_0 can be proved using probabilistic arguments--never a bound for it. However the former version allows the assignment of lower bounds on relevant probabilities for any fixed length of sequence, given (T_0, p) . This is illustrated in section eight.

7.6 Definition

- (1): A selection procedure is *linear* if to each partition σ and to each tree $T \in \text{BPT}(n)$, a real-valued weight $W(\sigma, T)$ is assigned such that:
- (a): for any two trees $T_1 \neq T_2$, there exists a $\sigma \in \Omega(n)$ with $W(\sigma, T_1) \neq W(\sigma, T_2)$,
 - (b): the trees selected are those that minimize $\sum_{\sigma} W(\sigma, T) X_{\sigma}$.
- (2): A selection procedure is *central* if it depends only on those partitions σ with $2 \leq |\sigma| \leq n-2$ and $\sigma = \sigma_0$.

7.7 Examples : (Compatibility and Wagner Parsimony)

Wagner parsimony and compatibility are both linear, and the former procedure is also central in the above sense. For Wagner parsimony $W(\sigma, T) = w(\sigma, T)$ (the number of steps required to fit σ to T , as in section five), while for compatibility, $W(\sigma, T) = -1$ if σ is induced by deleting an edge of T , and $W(\sigma, T) = 0$ otherwise.

Note that if we index $\Omega(n)$ to obtain a vector $W(T)$, then for compatibility $\langle W(T_1), W(T_2) \rangle = 2n-3-d(T_1, T_2)/2$, where d is the symmetric difference metric from section two. Among other major methods of tree reconstruction, those using dissimilarity are sometimes linear. For example in selecting a tree in $BPT(4)$, one approach, based on the four point condition (see Bandelt and Dress [1986] for a description of this property) is to choose $T = (ij)(kl)$ if $d(i,j)+d(i,k) = \min\{d(x,y)+d(z,w) : \{x,y,z,w\}=\{1,2,3,4\}\}$.

Since δ is linearly related to X (example 4.5), this procedure is linear. Statistically-based methods such as maximum likelihood (Felsenstein, [1973]) are generally neither linear nor central.

7.8 Theorem

For data stochastically generated by (T_0, p) , any linear procedure converges to that tree T (if any) which strictly minimizes $\sum_{\sigma} W(\sigma, T) s(\sigma; T_0, p)$. Such a tree exists except on an algebraic set of values for p .

Proof:

First suppose at least two trees T_1, T_2 minimize $\sum_{\sigma} W(\sigma, T) s(\sigma; T_0, p)$. Let $\lambda_{\sigma} = W(\sigma, T_1) - W(\sigma, T_2)$. By part (a) of the definition, $\lambda_{\sigma} \neq 0$ for some σ .

Now by theorem 6.15, $\sum_{\sigma} \lambda_{\sigma} s(\sigma; T_0, p) = 0$ only on an algebraic set for p .

Otherwise we may suppose $\sum_{\sigma} W(\sigma, T) s(\sigma; T_0, p)$ is strictly minimized by a tree T_1 . For $T \in BPT(n)$, and any vector v indexed over $\Omega(n)$,

let $Z(T, T_1, v) = \sum_{\sigma} (W(\sigma, T) - W(\sigma, T_1)) v_{\sigma}$, so that T_1 is selected precisely if $Z(T, T_1, X) > 0$ for all $T \neq T_1$.

$$\begin{aligned} \text{Now } P(Z(T, T_1, X) > 0; T_0, p) &= P(Z(T, T_1, X/c) \geq 0; T_0, p) \\ &\geq P(|Z(T, T_1, X/c) - Z(T, T_1, s)| < Z(T, T_1, s); T_0, p), \end{aligned}$$

where $s = s(T_0, p)$.

Since $Z(T, T_1, s(T_0, p)) > 0$, and $E[Z(T, T_1, X/c)] = Z(T, T_1, s)$ lemma 6.2 (3b) gives $\lim_{c \rightarrow \infty} P(Z(T, T_1, X) > 0) = 1$. The result now follows from lemma 6.2 (1). §

7.9 Theorem

Linear and central methods are not always consistent, even on four taxa.

Proof:

For the tree T as in fig. 7.1 set $p_1 = p_3 = \alpha$, $p_2 = p_4 = p_5 = \beta$, with $\alpha^2 = \beta(1 - \beta)$ so that $s_{12} = s_{13}$. Interchange pendant vertices 2 and 3 on T , leaving all the edges and their weights unchanged, to obtain a weighted tree T' , with central partition probabilities $s'_{12}, s'_{13}, s'_{14}, s_{1234}'$ the same as those for T . Thus for a central procedure ϕ ,

$$P(\phi(X) = \{T\}; T, p) = P(\phi(X) = \{T\}; T', p) \leq 1 - P(\phi(X) = \{T'\}; T', p).$$

In particular $P(\phi(X) = \{T\}; T, p)$ and $P(\phi(X) = \{T'\}; T', p)$ cannot both tend to 1 as c tends to infinity. Thus ϕ cannot be always consistent.

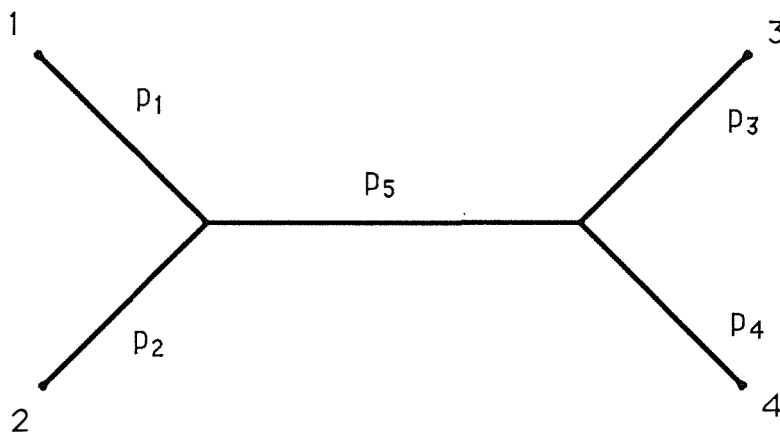


Figure 7.1

Suppose now there is a linear selection procedure φ which is always consistent.

Let $\delta(T_1, T_2, \sigma) = W(T_1, \sigma) - W(T_2, \sigma)$. Then by theorem 7.8,

$$\lim_{c \rightarrow \infty} P(\sum_{\sigma} X_{\sigma} \delta(T_1, T_2, \sigma) < 0; T_1, p) = 0 \text{ if } \sum_{\sigma} s(\sigma) \delta(T_1, T_2, \sigma) > 0$$

For $i=1, \dots, 4$, let $T_1 = (12)(34)$ and $T_2 = (13)(24)$. Assign weight p to the edge of T_1 , and T_2 incident with the pendant vertex labelled i , and assign weight ϵ on all the other edges, giving a edge weight vector $p(i)$.

If σ_i denotes the partition induced by the colouring $\{\{i\}, L_4 - \{i\}\}$, then

$$\sum_{\sigma} s(\sigma; T_1, p(i)) \delta(T_1, T_2, \sigma) = \delta(T_1, T_2, \sigma(i))p + O(\epsilon).$$

By choosing ϵ sufficiently small, the consistency condition for T_1 requires that $\delta(T_1, T_2, \sigma(i)) \leq 0$. Repeating the argument for T_2 gives $\delta(T_2, T_1, \sigma(i)) \leq 0$.

But $\delta(T_2, T_1, \sigma(i)) = -\delta(T_1, T_2, \sigma(i))$, hence $\delta(T_1, T_2, \sigma(i)) = 0$ for $i=1, 2, 3, 4$.

Taking $p = \epsilon j$, we have $\sum_{\sigma} s(\sigma; T, p) \delta(T_1, T_2, \sigma) = \delta(T_1, T_2, \sigma_0) + O(\epsilon)$, and by a similar argument, $\delta(T_1, T_2, \sigma_0) = 0$. It follows that any consistent linear method is central contradicting the first part of the theorem. Hence there cannot be a consistent linear selection procedure on $n \geq 4$ taxa. $\S$

The next result shows that if a selection procedure is always consistent on four taxa it can be turned into a selection procedure which is always consistent in general. We apply this result shortly.

7.10 Lemma

Suppose a selection procedure φ is always consistent on four taxa. For $n \geq 4$ taxa define

$$\varphi^*(X) = \langle \cup_{S \in [n]_4} \varphi(X[S]) \rangle, \text{ where } \langle Q \rangle \text{ is given in definition 3.10, and } X[S]$$

is the aggregate of X relative to S .

Then φ^* is always consistent on n taxa.

Proof:

By proposition 3.11, $\langle \cup_{S \in [n]_4} T[S] \rangle = \{T\}$, so that

$P(\varphi^*(X(c)) = \{T\}) \geq P(\cap_{S \in [n]_4} \varphi(X[S](c) = T[S]; T[S], p[S])$ which tends to 1 as $c \rightarrow \infty$ by lemma 6.2 (1) and the consistency of φ , as required. §

7.11 Remark

While linear selection procedures are in general inconsistent there exist selection procedures which are always consistent and which are quadratic in the sense that the procedure chooses that tree T minimizing

$\sum_{\sigma} W(\sigma, T) X_{\sigma} + \sum_{\sigma_1, \sigma_2} W(\sigma_1, \sigma_2, T) X_{\sigma_1} X_{\sigma_2}$ for suitable weights $W(\sigma, T)$,

$W(\sigma_1, \sigma_2, T)$. Such procedures arise from the quadratic invariants described in section six. An explicit example of a quadratic selection procedure is "closest tree" (Hendy [1988]. For a proof that it is quadratic in the above sense see Hendy [1989]).

Consistent recovery of trees from dissimilarities

One problem which arises from attempts to consistently infer phylogenies from dissimilarities is that frequently the number of sites, c , which could have changed is unknown. Indeed even the number c^* which actually changed may also be unknown, since the dissimilarity matrix gives only bounds on c^* , as illustrated in example 4.5. We now show that using just the dissimilarity matrix, and without knowledge of c or c^* , (and without assuming a molecular clock) there exist procedures for consistently recovering trees. This is not entirely obvious in view of the loss of information involved in converting sequences to dissimilarities, as described in section four.

By lemma 7.10 it suffices to consider only trees with four taxa.

Consider the truncated vectors $\hat{X}$ and $\hat{s}(T, p)$ defined by $X = [\hat{X}, X_{\sigma_0}]^t$, $s = [\hat{s}(T, p), s(\sigma_0)]^t$. Write the dissimilarity matrix as a vector δ , so that c^* and δ determine $\hat{X}$ as in example 4.5.

Given δ , let $l(\delta) = \max_{ij} \{\delta_{ij}\}$. For each tree T on the four taxa, assign edge weights $p(T)$ and a value $\gamma = \gamma(\delta, T)$ for c^* , with $l(\delta) \leq \gamma \leq 2l(\delta)$ so that for the associated partition frequencies $\hat{X} = \hat{X}(\delta, \gamma(T))$ we minimize $\Delta(T) = \|\hat{X}/\gamma - (1-s_0(T, p(T)))^{-1} \hat{s}(T, p(T))\|$. Finally let φ be the procedure which selects that tree (or trees) that minimizes $\Delta(T)$.

7.12 Theorem

φ always converges to the underlying tree T_0 as $l(\delta) \rightarrow \infty$.

Proof:

First note that $l(\delta) \rightarrow \infty$ precisely if $c^* \rightarrow \infty$.

Let $\hat{s} = \hat{s}(T_0, p_0)$ and $s_0 = s(\sigma_0; T_0, p_0)$.

Let $\epsilon_0 = \min\{\|(1-s_0)^{-1} \hat{s} - \beta(1-s_0(T, p))^{-1} \hat{s}(T, p)\|; T \neq T_0, p \in (0, 0.5)^5, \beta \in \mathbb{R}\}$.

By theorem 6.14, $\epsilon_0 > 0$.

Let $E_1 = E_1(c^*)$ be the event that

$$\|\hat{X}/c^* - (1-s_0)^{-1} \hat{s}\| < \epsilon_0/4,$$

let $E_2 = E_2(c^*)$ be the event that $\Delta(T_0) < \epsilon_0/4$

and let $E_3 = E_3(c^*)$, $E_4 = E_4(c^*)$ be respectively the events that

$$\Delta(T_1), \Delta(T_2) \geq \epsilon_0/4, \text{ for the other two trees } T_1, T_2 \in \text{BPT}(4).$$

Finally let $E_5 = E_5(c^*)$ be the event that $\Delta(T_0) < \min\{\Delta(T_1), \Delta(T_2)\}$, that is,

that φ^* selects T_0 . We wish to show $\lim_{c^* \rightarrow \infty} P(E_5(c^*)) = 1$.

Since c^* lies between $l(\delta)$ and $2.l(\delta)$, as shown in example in 4.5, we have $E_1 \subseteq E_2$. Also $E_2 \cap E_3 \cap E_4 \subseteq E_5$ so that $E_1 \cap E_3 \cap E_4 \subseteq E_5$. Thus

$$P(E_5) \geq P(E_1 \cap E_3 \cap E_4). \quad (*)$$

For $i = 3, 4$ let E_i' be the complement of E_i . Thus E_3' is the event that

$$\| \hat{X} / \vartheta(T_1) - (1-s_0(T_1, p(T_1)))^{-1} \hat{s}(T_1, p(T_1)) \| < \epsilon_0 / 4. \text{ But this implies}$$

$$\| \hat{X} / c^* - \beta'(1-s_0(T_1, p(T_1)))^{-1} \hat{s}(T_1, p(T_1)) \| < \epsilon_0 \vartheta(T_1) / 4 c^* \text{ where } \beta' = \vartheta(T_1) / c^*,$$

and this is $< \epsilon_0 / 2$ since $c^* \geq l(\delta)$ (example 4.5) and $\vartheta(T_1)$ is chosen $\leq 2.l(\delta)$.

Then by the triangle inequality for Euclidean distance, and the definition of ϵ_0 , we see that $E_1 \cap E_3'$ requires

$$\| (1-s_0)^{-1} \hat{s} - \beta'(1-s_0(T_1, p))^{-1} \hat{s}(T_1, p) \| < 3\epsilon_0 / 4, \text{ which cannot occur by}$$

definition of ϵ_0 . Thus $E_1 \cap E_3' = \emptyset$, and similarly $E_1 \cap E_4' = \emptyset$, so that

$E_1 \cap E_3 \cap E_4 = E_1$. Combining this with (*) we have

$$P(E_5) \geq P(E_1). \quad (**)$$

But by lemma 6.2 (3c) $\hat{X}$ has a multinomial distribution with parameters

c^* , and $\hat{s} / (1-s_0)$. Thus $\lim_{c^* \rightarrow \infty} P(\| \hat{X} / c^* - (1-s_0)^{-1} \hat{s} \| < \epsilon) = 1$ for all

$\epsilon > 0$, by lemma 6.2 (3b). In particular, taking $\epsilon = \epsilon_0$ we have

$\lim_{c^* \rightarrow \infty} P(E_1(c^*)) = 1$, so that, by (**), $\lim_{c^* \rightarrow \infty} P(E_5(c^*)) = 1$, as

required. §

Consistency of Parsimony and Compatibility.

We now establish conditions for the consistency of Wagner parsimony and compatibility.

7.13 Definition

Let e be an edge of $T \in \text{BPT}(n)$. Then e partitions the labels of the endpoints of T into two sets, $L_1(e), L_2(e)$, corresponding to the two component subtrees of $T - e$.

Then for $\sigma \in \Omega(n)$, σ is *induced by e* , if $\sigma = L_1(e)$ or $L_2(e)$. We let $\Omega I(T)$ (resp. $\Omega P(T)$) be the set of $\sigma \in \Omega(n)$ induced by an internal (resp. pendant) edge of T , and $\Omega(T) = \Omega I(T) \cup \Omega P(T) \cup \{\sigma_0\}$, the "treelike" partitions of T .

7.14 Lemma

Under Cavender's model, for sequence data stochastically generated on $(T_0, p) \in \text{WBT}(n)$,

(1): compatibility converges to a tree $T \in \text{BPT}(n)$ if T strictly maximizes

$$\sum_{\sigma \in \Omega(T)} s(\sigma; T_0, p).$$

(2): Wagner parsimony converges to a tree $T \in \text{BPT}(n)$ if T strictly minimizes $\sum_{\sigma \in \Omega(n)} w(\sigma, T) s(\sigma; T_0, p)$.

(3): In both cases such a tree T exists except for an algebraic set of weights for p .

Proof:

This lemma follows immediately from theorem 7.8. §

7.15 Example

For $T=(12)(34)$ with edge weights as given in fig. 7.1, the condition for consistency of both types of parsimony given by theorem 7.14 is

$s_{12} > \max\{s_{13}, s_{14}\}$, which translates by theorem 5.28 into:

$$\omega_5 < \min \left\{ (\omega_1\omega_2 + \omega_3\omega_4) / (\omega_1\omega_3 + \omega_2\omega_4), (\omega_1\omega_2 + \omega_3\omega_4) / (\omega_1\omega_4 + \omega_2\omega_3) \right\},$$

where $\omega_i = 1 - 2p_i$.

Thus, for any assignment of edge weights to the pendant edges, Wagner parsimony and compatibility are consistent if the central edge weight is not too small. Setting $p_1 = p_3 = \alpha$, $p_2 = p_4 = p_5 = \beta$ the requirement becomes $\alpha^2 \leq \beta(1 - \beta)$ as given by Felsenstein [1978 (2)]. §

We now show that Wagner parsimony fails even in the case where all edges have equal length, by generalizing Felsenstein's example which achieved inconsistency by separating two long edges and two short edges by a short internal edge. We replace the long edges by a large tree, whose edges have the same length as the previous short edges.

7.16 Definitions:

(1): For $Z \in \{\{A\}, \{B\}, \{A,B\}\}$, and $T \in \text{BPT}(n)$, let $V(T, Z)$ be the set of colourings of the pendant vertices of T in which Z is the set of possible colourings of the root in any minimal colouring of T .

(2): Let $T(k)$ be a rooted binary tree with 2^k endpoints and a root (of degree 2), defined recursively by taking $T(1)$ to be two edges each incident with a root vertex and $T(j)$ for $j > 1$ to be the tree obtained by joining the roots of two copies of $T(j-1)$ to a new root by two new edges, as shown in fig. 7.2 (refer to following page). The labelling of the pendant vertices of $T(k)$ is arbitrary and unimportant in what follows.

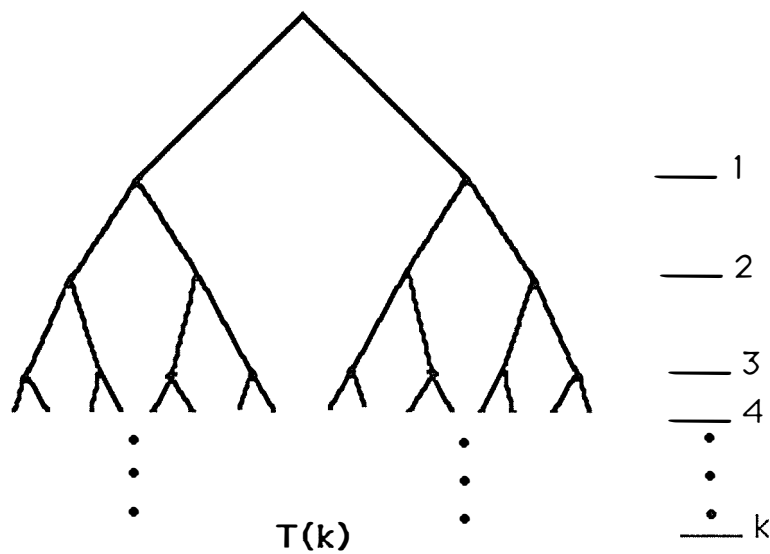


Figure 7.2

7.17 Remark

For $k \geq 5$, there are colourings in $V(T(k), \{A\})$ in which the majority of endpoints are coloured B. Indeed if we let $f(k)$ be the minimum number of endpoints of $T(k)$ that are coloured A over all $\chi \in V(T(k), \{A\})$, we have:

7.18 Theorem

$$f(k) = f(k-1) + f(k-2); f(0) = 1, f(1) = 2.$$

Thus $f(k)$ is the $(k+1)$ -th Fibonacci number, F_{k+1} , and so for each $\epsilon > 0$ there is a k , and a $\chi \in V(T(k), \{A\})$ for which the proportion of endpoints of $T(k)$ coloured A by χ is $< \epsilon$.

Proof: Refer to Appendix.

7.19 Definition

Let $RWT(n)$ denote the set of weighted rooted binary trees on label set $\{1, \dots, n\}$. For $T \in RWT(n)$, let $S(T)$, (resp. $D(T)$, $E(T)$) denote the probability that a colouring of the pendant vertices of T lies in $V(T, \{A\})$ (resp. $V(T, \{B\})$, $V(T, \{A, B\})$, where A is the actual colour of the root of T . Thus, for example, $S(T)$ is the probability that the endpoints of T are coloured so that all minimal extensions of that colouring to the vertices of T will assign the root its actual colour, A . Note that $S(T) + D(T) + E(T) = 1$. Given $T_1 \in RWT(n_1)$, $T_2 \in RWT(n_2)$, let $T_1 * T_2(p_1, p_2) \in RWT(n_1 + n_2)$ be obtained by joining the roots of T_1 and T_2 to a new root, and assigning edge probabilities p_1, p_2 respectively to the new edges.

7.20 Lemma

For $T = T_1 * T_2(p_1, p_2)$; $q_i = 1 - p_i$, $S_i = S(T_i)$, $D_i = D(T_i)$, $E_i = E(T_i)$ for $i = 1, 2$:

$$S(T) = (q_1 S_1 + p_1 D_1)(q_2 S_2 + p_2 D_2) + E_1(q_2 S_2 + p_2 D_2) + E_2(q_1 S_1 + p_1 D_1)$$

$$D(T) = (p_1 S_1 + q_1 D_1)(p_2 S_2 + q_2 D_2) + E_1(p_2 S_2 + q_2 D_2) + E_2(q_1 D_1 + p_1 S_1)$$

$$E(T) = E_1 E_2 + (S_1 D_2 + D_1 S_2)(q_1 q_2 + p_1 p_2) + (D_1 D_2 + S_1 S_2)(p_1 q_2 + p_2 q_1).$$

Proof: The four possible actual states, -- (A, A) , (A, B) , (B, A) , (B, B) -- of the roots of T_1, T_2 , have probability $q_1 q_2$, $q_1 p_2$, $p_1 q_2$, $p_1 p_2$ respectively. The results now follow from 5.18 (2), by considering when Fitch's algorithm will assign the root of T the set $\{A\}$, $\{B\}$ or $\{A, B\}$ in terms of the corresponding set assigned to the roots of T_1 and T_2 .

7.21 Corollary

$$(1): S(T) > D(T)$$

$$(2): E(T) \leq 0.5.$$

Proof:

$S(T)-D(T) = (q_1q_2-p_1p_2)(S_1S_2-D_1D_2) + E_1(q_2-p_2)(S_2-D_2) + E_2(q_1-p_1)(S_1-D_1)$, and

(1) follows immediately by induction, since $q_1 \geq p_1, q_2 \geq p_2$.

(2) can be established by standard analytic optimization. §

7.22 Theorem

Assign edge weight p to all the edges of $T(k)$.

Let $S(k) = S(T(k)), D(k) = D(T(k)), E(k) = E(T(k))$. Then

(1): if $p \geq 1/8$, $\lim_{k \rightarrow \infty} S(k) = \lim_{k \rightarrow \infty} D(k) = \lim_{k \rightarrow \infty} E(k) = 1/3$

(2): if $p < 1/8$,

$$s = \lim_{k \rightarrow \infty} S(k) = (1 - 2x + \sqrt{\Delta}) / (1 - 2p) / 2$$

$$d = \lim_{k \rightarrow \infty} D(k) = (1 - 2x - \sqrt{\Delta}) / (1 - 2p) / 2$$

$$e = \lim_{k \rightarrow \infty} E(k) = 2x$$

where $x = p / (1 - 2p)$ and $\Delta = (1 - 6x)(1 - 2x)$.

Proof:

From lemma 7.20 we have:

$$S(k+1) = (S(k)q + D(k)p)(S(k)q + D(k)p + 2E(k))$$

$$D(k+1) = (S(k)p + D(k)q)(S(k)p + D(k)q + 2E(k))$$

$$E(k+1) = E^2(k) + 2(S(k)q + D(k)p)(S(k)p + D(k)q).$$

Let s , (resp. d, e) denote the limit of $S(k)$ (resp. $D(k), E(k)$) as $k \rightarrow \infty$. The real numbers s, d, e then satisfy:

$$(1): s = (sq + dp)(sq + dp + 2e)$$

$$(2): d = (sp + dq)(sp + dq + 2e)$$

$$(3): e = e^2 + 2(sq + dp)(sp + dq).$$

We show that the only solutions to these equations are those given in the theorem. From (1) and (2) we have:

$$2e = \frac{s}{(sq+dp)} - \frac{d}{(sp+dq)} = \frac{d}{(sp+dq)} - \frac{s}{(sp+dq)}. \quad (4)$$

Let $Q = (sq+dp)(sp+dq)$. Then the second equality in (4) gives:

$$(s^2-d^2)p/Q = (s-d)(1-2p), \text{ which is } (s-d)((s+d)p/Q - (1-2p)) = 0.$$

Thus either $s = d$ or $(s+d) = \frac{(1-2p)Q}{p}$. In the first case, we have from (4),

$$2e = 1-s. \text{ Since } s+d+e=1 \text{ we have } s=d=e=1/3.$$

In case $(s+d) = \frac{(1-2p)Q}{p}$, let $u = s+d$, so that (3) becomes

$$e = (1-u) = (1-u)^2 + 2Q, \text{ which is } u^2 - u = -2Q. \text{ Replacing } Q \text{ by } \frac{up}{(1-2p)}, \text{ we}$$

have $u(u - \frac{2p}{(1-2p)}) = 0$. But $u \neq 0$ by corollary 7.21 (2). Thus, letting

$$x = \frac{p}{(1-2p)}, \text{ we have}$$

$$s+d = 1-2x = \frac{(1-2p)Q}{p}. \text{ Let } a = (sp+dq), b = (sq+dp), \text{ so that}$$

$$a+b=s+d, ab = Q. \text{ Thus we have: } a+b = 1-2x, ab = \frac{p}{x} - \frac{2p^2}{x^2}.$$

Solving the associated quadratic equation (for a and b) and noting that $a \geq b$ by 7.21 (1), then solving for s and d gives:

$$s = (1-2x + \sqrt{(\Delta)}) / \frac{(1-2p)}{2}; \quad d = (1-2x - \sqrt{(\Delta)}) / \frac{(1-2p)}{2},$$

$$\text{and so } e = 1-s-d = 2x.$$

Now for $p > 1/8$, $\Delta < 0$, so the only solution to (1), (2) and (3) is $s=d=e=1/3$.

For $p < 1/8$, there are two possible solutions so consider $\delta(k) = S(k) - D(k)$.

Then writing $p = \frac{(1-\epsilon)}{8}$, $\epsilon > 0$, and using $p+q=1$ and $S(k)+D(k)+E(k)=1$, it is

easily checked that

$$\delta(k+1) = (1-2p)(1+E(k))\delta(k) = \frac{(3+2\epsilon)(1+E(k))\delta(k)}{4}.$$

Now if $\lim_{k \rightarrow \infty} E(k) = 1/3$, we have $\delta(k+1)/\delta(k) > (1+2\epsilon/3) - \epsilon/3 = 1 + \epsilon/3$, for infinitely many k , which implies $\lim_{k \rightarrow \infty} \delta(k) = \infty$, a contradiction, since $\delta(k) < S(k) \leq 1$. This eliminates the solution $(s,d,e) = (1/3, 1/3, 1/3)$, as required. §

The previous theorem is applied shortly to study the consistency of Wagner parsimony. First we derive analogous results which are required for analysing the consistency of compatibility.

7.23 Definition

For $T \in \text{RWT}(n)$, let $Q(T)$ denote the probability that the endpoints of T are all the same colour as the root, and $P(T)$, the probability that they are all the opposite colour to the root. We now present the analogue of lemma 7.20.

7.24 Lemma

For $T = T_1 * T_2(p_1, p_2)$; $q_i = 1 - p_i$, $Q_i = Q(T_i)$, $S_i = S(T_i)$, for $i=1,2$:

$$Q(T) = (q_1 Q_1 + p_1 P_1)(q_2 Q_2 + p_2 P_2)$$

$$P(T) = (p_1 Q_1 + q_1 P_1)(p_2 Q_2 + q_2 P_2).$$

Proof:

Considering the four possible states of the roots of T_1, T_2 gives the lemma immediately. §

7.25 Corollary

(1): $P(T) < Q(T)$

(2): $Q(T) < \min\{Q(T_1), Q(T_2)\}$.

Proof:

(1) follows by induction, using lemma 7.24, while (2) follows from (1). §

7.26 Theorem

Assign edge weight p to all the edges of $T(k)$, and let

$Q(k) = Q(T(k))$, $P(k) = P(T(k))$ and $\tau(k) = P(k)/Q(k)$. Then

$$\tau = \lim_{k \rightarrow \infty} \tau(k) = \begin{cases} 1, & \text{if } p \geq 1/4, \\ (\lambda - \sqrt{\lambda^2 - 4})/2, & \text{if } p < 1/4, \text{ where } \lambda = (2p^2 - 4p + 1)/p^2. \end{cases}$$

Proof:

By lemma 7.24, $\tau(k+1) = (pQ_k + qP_k)^2 / (qQ_k + pP_k)^2 = (p + q\tau(k))^2 / (q + p\tau(k))^2$, where $q = 1 - p$, $Q_k = Q(k)$, $P_k = P(k)$.

Set $x = (p + qx)^2 / (q + px)^2$, which is:

$(x-1)(x^2 - \lambda x + 1) = 0$, where $\lambda = (2p^2 - 4p + 1)/p^2$. Now the discriminant of the quadratic factor is $\Delta = \lambda^2 - 4 = (\lambda - 2)(\lambda + 2)$ which is negative if and only if $\lambda < 2$, and this occurs precisely when $p > 1/4$. Thus the only real solution for x when $p > 1/4$ is $x = 1$. For $p = 1/4$, the cubic equation becomes $(x-1)^3 = 0$, which again has the unique solution $x = 1$.

For $p < 1/4$, and hence $\lambda > 2$, the only solution for $x \neq 1$, with $x \leq 1$ (required by corollary 7.24) is $x = (\lambda - \sqrt{\lambda^2 - 4})/2$.

Now for $\Phi(t) = (p+qt)^2 / (q+pt)^2$, we have $\Phi'(t) > 0$ on $[0, x]$ (since $q > p$), so that $\Phi(t)$ is monotone increasing and continuous on $[0, x]$ and x is the only fixed point of $\Phi(t)$ in this interval, so by an elementary argument, $\tau(k)$ converges to x , as required. §

7.27 Definition

Let $T_1(k) \in \text{BPT}(2^{k+1}+2)$ consist of two copies of $T(k)$, whose roots are identified to v_2 and v_4 of the tree $T = (v_1 v_2)(v_3 v_4)$. Assign labels 1 and 3 to v_1 and v_3 respectively, and label the rest of $T_1(k)$ arbitrarily. For $p = [p_1, p_2, p_3, p_4]^t$, let $T_1(k, p) \in \text{WBT}(n)$ be $T_1(k)$ with edge weight p_4 on all the edges of the two copies of $T(k)$ and the other edge weights as shown in fig. 7.3. For $i=1, 2, 3$ let $\omega_i = (1-2p_i)$. Let $T_2(k) \in \text{BPT}(2^{k+1}+2)$ be the tree obtained from $T_1(k)$ by performing a nearest neighbour interchange so as to join the two copies of $T(k)$.

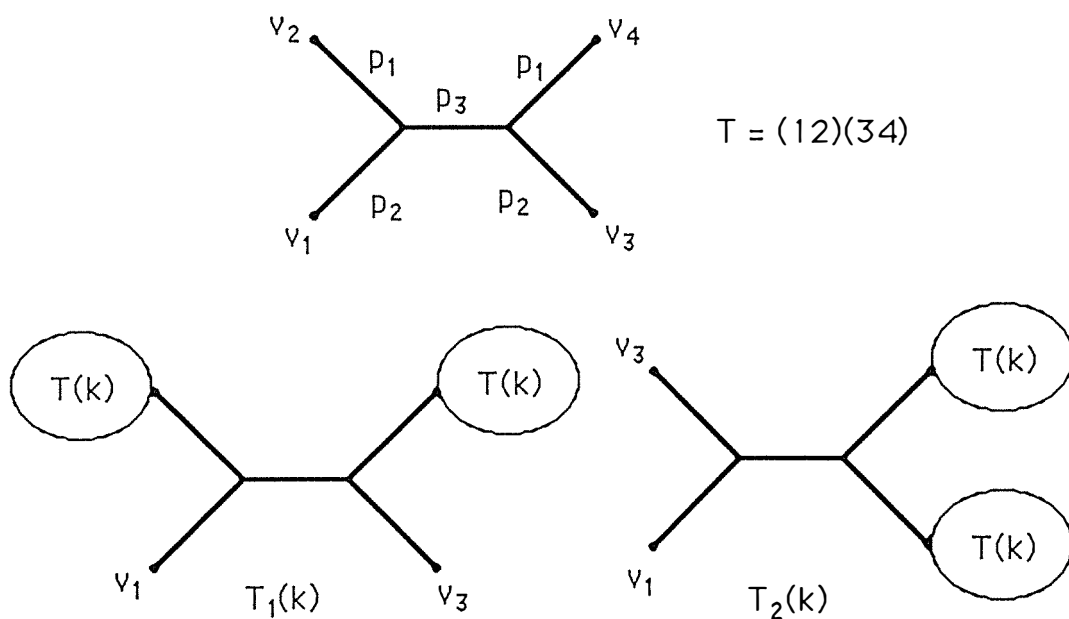


Figure 7.3

7.28 Theorem

Under Cavender's model, for data derived from $T_1(k,p)$,

(1): Wagner parsimony favours $T_2(k)$ over $T_1(k)$ precisely if

$$\omega_3(\omega_1^2 + \omega_2^2)(S^2(k) + D^2(k)) - 2\omega_1\omega_2(S^2(k) - D^2(k)) + 2\omega_3(\omega_2^2 - \omega_1^2)S(k)D(k) > 0.$$

(2): Compatibility favours $T_2(k)$ over $T_1(k)$ precisely if

$$\omega_3(\omega_1^2 + \omega_2^2)(Q^2(k) + P^2(k)) - 2\omega_1\omega_2(Q^2(k) - P^2(k)) + 2\omega_3(\omega_2^2 - \omega_1^2)Q(k)P(k) > 0.$$

[where $S(k)$, $D(k)$, $Q(k)$, $P(k)$ are as in theorems 7.22 and 7.26, with $p = p_4$]

Proof:

(1): Denote the state of the v_1 by A, and let B be the other colour.

For $\sigma \in \Omega(n)$ let $\lambda_\sigma = w(T_1(k), \sigma) - w(T_2(k), \sigma)$, and let χ_2 and χ_4 be the

induced colouring of the two copies of $T(k)$ attached to v_2, v_4 respectively.

By Fitch's algorithm (5.18(2)) there are precisely two cases for which

$$\lambda_\sigma \neq 0.$$

Case One: If $\chi_2, \chi_4 \in V(T(k), \{B\})$ and v_3 is coloured A, then $\lambda_\sigma = 1$. Denote the set of such σ as Ω_1 .

Case Two: If $\chi_2 \in V(T(k), \{A\})$ and $\chi_4 \in V(T(k), \{B\})$, v_3 is coloured B, then $\lambda_\sigma = -1$. Denote the set of such σ as Ω_2 .

Now the probability of case one arising on $T_1(k,p)$ is,

$$S^2(k)s_{13} + D^2(k)s_{1234} + S(k)D(k)(s_{123} + s_{134}), \text{ where } s_\sigma = s(\sigma_0; T, p), \text{ with}$$

$T = (v_1v_2)(v_3v_4)$ and p given in fig. 7.3.

The probability of case two arising from $T_1(k,p)$ is,

$$S^2(k)s_{12} + D^2(k)s_{14} + S(k)D(k)(s_1 + s_{124}).$$

Then letting $Z(k) = \sum_{\sigma \in \Omega(n)} \lambda_{\sigma} X_{\sigma} / c$, we have

$$E[Z(k)] = \sum_{\sigma \in \Omega(n)} \lambda_{\sigma} s(\sigma) = \sum_{\sigma \in \Omega_1} s(\sigma) - \sum_{\sigma \in \Omega_2} s(\sigma).$$

$$= S^2(k)s_{13} + D^2(k)s_{1234} + S(k)D(k)(s_{123} + s_{134}) \\ - (S^2(k)s_{12} + D^2(k)s_{14} + S(k)D(k)(s_1 + s_{124})).$$

$= s^t u(k)$, where

$$u(k) = [-S(k)D(k), -S^2(k), S^2(k), S(k)D(k), -D^2(k), -S(k)D(k), S(k)D(k), D^2(k)]^t$$

$$\text{and } s^t = [s_1, s_{12}, s_{13}, s_{123}, s_{14}, s_{124}, s_{134}, s_{1234}].$$

Now, letting $\omega_i = (1-2p_i)$, and using the notation (and ordering of $\Omega(4)$ and $[4]^+$) of example 6.10 we have

$$r = [1, \omega_1\omega_2, \omega_2^2\omega_3, \omega_1\omega_2\omega_3, \omega_1\omega_2\omega_3, \omega_1^2\omega_3, \omega_1\omega_2, \omega_1^2\omega_2^2]^t \text{ and}$$

$$r = H^t s.$$

Now since $s = 1/8 H r$, where $H = [(-1) | \sigma \cap X |]_{\sigma, X}$ (as in theorem 5.28) we

have $E[Z(k)] = 1/8 s^t u(k) = r^t H^t u(k)$, and it is easily checked that

$$H^t u(k) = 2[0, D^2(k)-S^2(k), (S(k)+D(k))^2, 0, 0, (S(k)-D(k))^2, D^2(k)-S^2(k), 0]^t$$

giving:

$$E[Z(k)] = 0.25(\omega_3(\omega_1^2 + \omega_2^2)(S^2(k) + D^2(k)) - 2\omega_1\omega_2(S^2(k) - D^2(k)) + \\ 2\omega_3(\omega_2^2 - \omega_1^2)S(k)D(k)).$$

The result now follows from lemma 7.14.

(2): Let $Z^*(k) = \sum_{\sigma \in \Omega(T_1(k))} X_{\sigma} / c - \sum_{\sigma \in \Omega(T_2(k))} X_{\sigma} / c = X_{\sigma_1} / c - X_{\sigma_2} / c$ say,

where $\sigma_1 = \{1, v_2\}$, with v_2 the labels on the copy of $T(k)$ whose root is identified with v_2 , and $\sigma_2 = \{1, 3\}$.

Since $E[Z^*(k)] = s(\sigma_1) - s(\sigma_2)$, the result now follows by a similar argument to part (1), and from lemma 7.14. §

7.29 Corollary

(1): Wagner parsimony is inconsistent on $T_1(k,p)$ if

$$p < D^2(k)/(S^2(k)+D^2(k)).$$

Thus Wagner parsimony is inconsistent on $T_1(2,p) \in \text{WBT}(10)$ for $p > p_1 \approx 0.39076$, and on $T_1(k,p)$ for $p > p_2 \approx 0.116413$, and k sufficiently large (dependent on p).

(2): Compatibility is inconsistent on $T_1(k,p)$ if

$$p < P^2(k)/(P^2(k)+Q^2(k)).$$

Thus compatibility is inconsistent on $T_1(2,p)$ for $p > p_3 \approx 0.436154$, and on $T_1(k,p)$ for $p > p_4 \approx 0.245122$, and k sufficiently large (dependent on p).

Proof:

Setting $p_i = p$, for $i=1,\dots,4$, so that $\omega_i = (1-2p) = \omega$ (say), the expression in the statement of theorem 7.28 for Wagner parsimony becomes $\omega(2\omega^2(S^2(k)+D^2(k))-2\omega(S^2(k)-D^2(k)))$ which is positive precisely if $\omega > (S^2(k)-D^2(k))/(S^2(k)+D^2(k))$, or equivalently, $p < D^2(k)/(S^2(k)+D^2(k))$.

For compatibility, the analogous requirement is

$$p < P^2(k)/(Q^2(k)+P^2(k)) = 1/(1+\tau(k)^{-2}), \text{ (where } \tau(k) \text{ is defined in theorem 7.26).}$$

Now for any real-valued function, f defined on positive integers,

(in particular $f(k) = E[Z(k)]$, and $E[Z^*(k)]/Q^2(k)$), if

$\lim_{k \rightarrow \infty} f(k) > 0$, then for some integer k_0 we also have $f(k_0) > 0$.

Thus the values p_2 and p_4 are obtained by solving the equations

$$p = d^2/(s^2+d^2), \text{ and } p = 1/(1+\tau^{-2}), \text{ respectively, where } s, d \text{ are given by}$$

theorem 7.22, and τ is given by theorem 7.26.

The values p_1 , and p_3 are obtained by solving $p = D(2)^2 / (S^2(2) + D^2(2))$ and $p = P^2(2) / (Q^2(2) + P^2(2))$, where $S(2)$, $D(2)$, $P(2)$ and $Q(2)$ given (in terms of p) by lemmas 7.20 and 7.24. §

7.30 Remark

An example of a $T_1(2, p)$ is given in fig. 7.4. For $p > 0.4$, (resp. $p > 0.25$) Wagner parsimony (resp. compatibility) will be inconsistent.

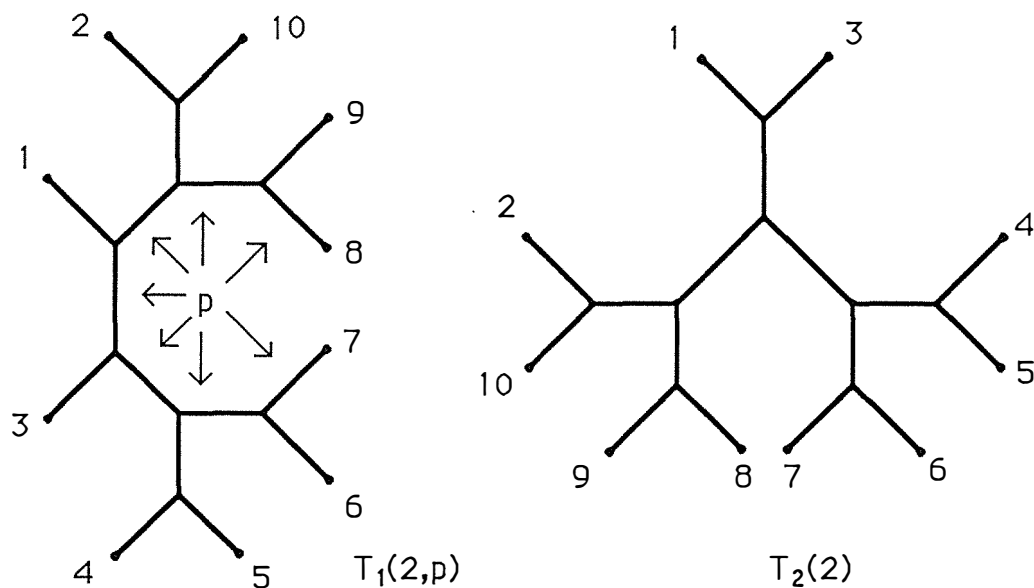


Figure 7.4

Sufficient Conditions

In view of the above results it is natural to ask whether parsimony methods will sometimes fail for equal edge weights below those prescribed in corollary 7.29. We show that in the case of compatibility this is not so--indeed if the edge weights are "sufficiently small and sufficiently equal" (in a sense to be made precise) compatibility is consistent.

7.31 Lemma

For any $T \in \text{WBT}(n)$, and $C \subseteq E(T)$, if $p(C) = \sigma$, then

$$s(\sigma; T, p) \geq s(\sigma_0; T, p) \prod_{e \in C} (p_e / (1 - p_e)).$$

Proof:

We use induction on $|C|$. If $|C| = 0$, the result holds, so suppose it holds for all C with $|C| = k-1 \geq 0$, and let $|C| = k$, $C' = C - \{e\}$ for $e \in C$, and suppose $p(C') = \sigma'$. If $e = [v_1, v_2]$, consider the conditional probabilities,

$\alpha = P(\sigma \mid \chi(v_1) \neq \chi(v_2); T, p)$, $\beta = P(\sigma \mid \chi(v_1) = \chi(v_2); T, p)$, where $\chi(v_i) \in \{A, B\}$ is the colour of v_i , $i=1,2$, (where, by convention, A denotes the colour of the pendant vertex of T labelled 1).

$$\text{Then } s(\sigma; T, p) = p_e \alpha + (1 - p_e) \beta, \quad s(\sigma'; T, p) = (1 - p_e) \alpha + p_e \beta.$$

Thus, since $(1 - p_e)^2 \geq p_e^2$, we have $s(\sigma; T, p) \geq s(\sigma'; T, p) p_e / (1 - p_e)$.

Applying the inductive hypothesis we see that the result holds for $|C| = k$, as required. $\S$

7.32 Definition

For real numbers $\lambda > 0$ and $L \geq 1$, and integer $n \geq 3$, let $B(n, \lambda, L)$ be the set of trees $T \in \text{WBT}(n)$ whose edge lengths are all $\leq \lambda$ and whose set of ratios of edge lengths is bounded above by L .

Let $B(\lambda, L) = \cup_n B(n, \lambda, L)$, and $B(\lambda) = \cup_L B(\lambda, L)$.

7.33 Theorem

For $C > 1$, choose $0 < \lambda_0 < 1$ so that

$$(1): C^4\lambda_0^3 + 2C^2\lambda_0 + 1 - C \leq 0, (2): 2C\lambda_0(1 + C\lambda_0) \leq 1, (3): C \geq 1/(1 - \lambda_0).$$

[Such a choice is possible since for $C > 1$ fixed, the inequalities all become strict as λ_0 tends to zero.]

Suppose $(T, p) \in B(\lambda)$, $\lambda \leq \lambda_0$. Then

$$(1): \text{if } \sigma \neq \sigma_0, s(\sigma; T, p) \leq s(\sigma_0; T, p)C\lambda,$$

$$(2): \text{if } w(\sigma, T) > 1, s(\sigma; T, p) \leq s(\sigma_0; T, p) \cdot 2C^2\lambda^2(1 + C\lambda).$$

Proof:

(1): We first note that by the first condition on λ_0 , and $\lambda \leq \lambda_0$ we have:

$$1 + C^2\lambda \leq 1 + C^2\lambda_0 \leq C - C^2\lambda_0 - C^4\lambda_0^3 < C, \text{ thus}$$

$$(a): 1 + C^2\lambda < C.$$

We prove the theorem by induction on $|T|$, the number of pendant vertices of T . For $|T| = 2$, if $\sigma \neq \sigma_0$, then $s(\sigma; T, p) = p$, where p is the edge weight on the unique edge of T , while $s(\sigma_0; T, p) = (1 - p)$. Now by condition (3), $C\lambda_0/(1 + C\lambda_0) \geq \lambda_0 \geq \lambda \geq p$, giving $p \leq (1 - p)C\lambda$, as required.

Suppose the results holds for all trees with $|T| < k$, where $k \geq 3$, and suppose $T \in B(\lambda)$, $|T| = k$.

Case One: Suppose $w(\sigma, T) = 1$.

If σ is induced by an internal edge e of T , let T_1, T_2, T_3, T_4 be the four pendant subtrees of T having endpoints incident with e , as in fig. 7.5(a) (refer to following page).

For $i = 1, \dots, 4$, let s_i (resp. d_i) be the probability that the original endpoints of T_i are all the same colour, (resp. the opposite colour) to the new endpoint incident with e .

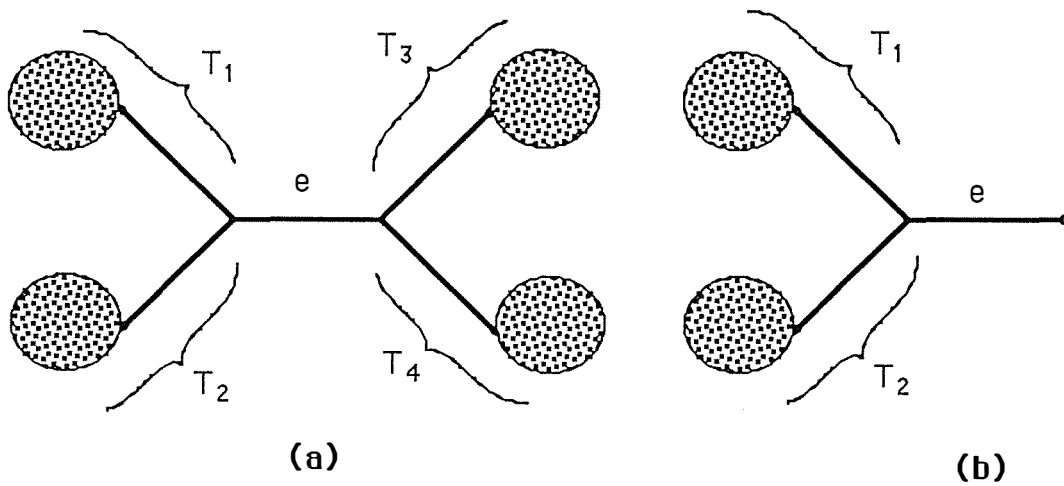


Figure 7.5

Then $s(\sigma; T, p) = p_e \alpha + (1-p_e) \beta$, where

$\alpha = (s_1 s_2 s_3 s_4 + d_1 d_2 d_3 d_4)$, $\beta = (s_1 s_2 d_3 d_4 + d_1 d_2 s_3 s_4)$, so that

$s(\sigma; T, p) \leq s_1 s_2 s_3 s_4 (p_e + C^4 \lambda^4 + 2(1-p_e) C^2 \lambda^2)$, by the induction hypothesis.

Now $s(\sigma_0; T, p) = (1-p_e) \alpha + p_e \beta$, and $s(\sigma_0; T, p)$ is maximized when $p_e = 0$,

(lemma 6.13 (1)), so that $s(\sigma_0; T, p) \leq s_1 s_2 s_3 s_4$. Thus, since $p_e \leq \lambda$, we have

$s(\sigma; T, p) \leq s(\sigma_0; T, p) (\lambda + C^4 \lambda^4 + 2C^2 \lambda^2) \leq s(\sigma_0; T, p) C \lambda$, since

$1 + C^4 \lambda^3 + 2C^2 \lambda \leq C$.

Similarly, if σ is induced by a pendant edge e , represent T as in fig. 7.5(b).

For $i=1,2$ let s_i (resp. d_i) be the probability that the original endpoints of

T_i are all the same colour, (resp. the opposite colour) to the new endpoint

incident with e . Then

$s(\sigma; T, p) = p_e s_1 s_2 + (1-p_e) d_1 d_2 \leq s_1 s_2 (\lambda + C^2 \lambda^2) \leq s(\sigma_0; T, p) C \lambda$, by (a).

Case Two: $w(\sigma, T) > 1$.

For σ with $w(\sigma, T) = k \geq 2$, there exists by Menger's theorem (refer Harary [1969]) two disjoint paths between differently coloured endpoints of T so that we can represent T as in fig. 7.6.

For $i=1, \dots, 4$, let $\sigma_i(z)$, $z \in \{a, b\}$, denote the colouring induced by σ on the pendant subtrees T_i with its distinguished pendant vertex (circled in fig. 7.6) coloured z . Let $\sigma_5(x, y)$ denote the colouring of the central tree T_5 with its two distinguished endpoints coloured x, y .

For $i=1, \dots, 4$, let $f_i(x) = s(\sigma_i; T_i, p_i)$, and $f_5(x, y) = s(\sigma_5; T_5, p_i)$, where p_i is the restriction of p to the edges of T_i .

Thus $s(\sigma; T, p) = f_1(a)f_2(a)f_3(a)f_4(a)f_5(a, a) + f_1(b)f_2(b)f_3(a)f_4(a)f_5(b, a) + f_1(a)f_2(a)f_3(b)f_4(b)f_5(a, b) + f_1(b)f_2(b)f_3(b)f_4(b)f_5(b, b)$.

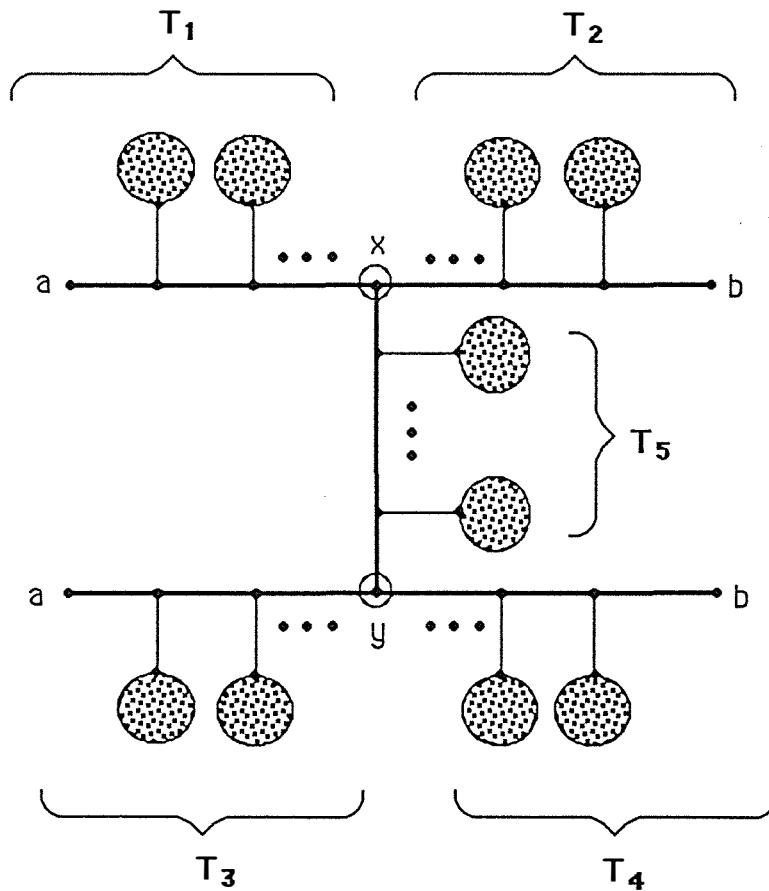


Figure 7.6

Now for any assignment of x, y , at least one of $w(\sigma_1, T_1)$, and $w(\sigma_2, T_2)$ is ≥ 1 , and at least one of $w(\sigma_3, T_3), w(\sigma_4, T_4)$ is ≥ 1 . Furthermore, if $x \neq y$, $w(\sigma_5(x, y), T_5) \geq 1$. By induction

$s(\sigma; T, p) \leq s_1 s_2 s_3 s_4 s_5 (C^2 \lambda^2 + C^3 \lambda^3 + C^3 \lambda^3 + C^2 \lambda^2)$, where s_i is the probability that the endpoints of T_i are all the same colour. Furthermore $s(\sigma_0; T, p) \geq s_1 s_2 s_3 s_4 s_5$, thus $s(\sigma; T, p) \leq s(\sigma_0; T, p) 2(C^2 \lambda^2 + C^3 \lambda^3) \leq s(\sigma_0; T, p) C \lambda$, since $C \lambda + C^2 \lambda^2 \leq C \lambda_0 (1 + C \lambda_0) \leq 1/2$.

(2): This follows by induction on $|T|$ by using part (1) and repeating the argument used above to establish case two of part (1), since in the last step of this proof we have

$$s(\sigma; T, p) \leq s(\sigma_0; T, p) 2C^2 \lambda^2 (1 + C \lambda). \quad \S$$

7.34 Theorem

For all $L \geq 1$ there exists $\lambda_0 > 0$, such that compatibility is consistent on $B(\lambda, L)$ for all $\lambda \leq \lambda_0$.

Proof:

By lemma 7.14, compatibility is consistent for data generated by the Cavender model on tree (T_0, p) precisely if $T = T_0$ strictly maximizes $\sum_{\sigma \in \Omega(T)} s(\sigma; T_0, p)$. This is clearly equivalent to saying $T = T_0$ strictly maximizes $\sum_{\sigma \in \Omega_0(T)} s(\sigma; T_0, p)$, where $\Omega_0(T) = \Omega(T) - \{\sigma_0\}$.

Choose $\lambda_0 > 0, C > 1$ so that:

$$(1): L^{-1} > 2C^2 \lambda_0 (1 + C \lambda_0), (2): C^4 \lambda_0^3 + 2C^2 \lambda_0 + 1 - C \leq 0, (3): 2C \lambda_0 (1 + C \lambda_0) \leq 1,$$

$$(4): C \geq 1/(1 - \lambda_0).$$

Then if $p = \max\{p_e: e \in E(T_0)\}$, $p \leq \lambda_0$, we have

(a): if $\sigma \in \Omega_0(T_0)$, (Ω_0 defined as in this proof above) then lemma 7.31 gives

$$s(\sigma; T_0, p) \geq s(\sigma_0; T_0, p)p_e/(1-p_e) \geq p_e s(\sigma_0; T_0, p) \geq (p/L)s(\sigma_0; T_0, p).$$

(b): if σ is not in $\Omega_0(T_0)$, so that $w(\sigma, T_0) > 1$, then by theorem 7.33, (with $\lambda = p$), we have $s(\sigma; T_0, p) \leq s(\sigma_0; T_0, p) \cdot 2C^2p^2(1+Cp)$.

Let $s^0_{\min} = \min\{s(\sigma; T, p): \sigma \in \Omega_0(T_0)\}$, $s^1_{\max} = \max\{s(\sigma; T, p): \sigma \in \Omega(n) - \Omega(T_0)\}$.

The condition $L^{-1} > 2C^2\lambda_0(1+C\lambda_0)$ implies $L^{-1} > 2C^2p(1+Cp)$ since $p < \lambda_0$, so that eliminating $s(\sigma_0; T_0, p)$ from (a) and (b), we have

$$s^0_{\min} > s^1_{\max}.$$

Thus

$$\begin{aligned} & \sum_{\sigma \in \Omega_0(T_0)} s(\sigma; T_0, p) - \sum_{\sigma \in \Omega_0(T)} s(\sigma; T_0, p) \\ &= \sum_{\sigma \in \Omega_0(T_0) - \Omega_0(T)} s(\sigma; T_0, p) - \sum_{\sigma \in \Omega_0(T) - \Omega_0(T_0)} s(\sigma; T_0, p), \\ &\geq ks^0_{\min} - ks^1_{\max} = k(s^0_{\min} - s^1_{\max}) \text{ where} \end{aligned}$$

$$k = |\Omega_0(T) - \Omega_0(T_0)| = |\Omega_0(T_0) - \Omega_0(T)| \text{ (since } |\Omega_0(T)| = |\Omega_0(T_0)| = 2n-3).$$

Now if $T \neq T_0$ we have $k \geq 1$ so that

$$\sum_{\sigma \in \Omega_0(T_0)} s(\sigma; T_0, p) > \sum_{\sigma \in \Omega_0(T)} s(\sigma; T_0, p), \text{ as required. } \S$$

7.35 Example

For $L=1$, taking $C = 2.18$, $\lambda_0 = 0.12$, it is easily checked that conditions (1) to (4) in the previous theorem are satisfied. Thus if the edge weights are all equal and less than 0.12, compatibility is consistent on all trees. In particular, by corollary 7.29 (1) there are weighted trees (such as $T_1(k, p)$ for $0.117 < p < 0.12$, and k sufficiently large) on which compatibility is consistent but Wagner parsimony is not.

A considerably weaker result for Wagner parsimony is the following:

7.36 Theorem

For all n and for all L , there exists a λ_0 , for which Wagner parsimony is consistent on $B(n, \lambda_0, L)$.

Proof:

By theorem 7.33 (2), and lemma 7.31, given n and L we can choose λ_0 so that if $(T, p) \in B(n, \lambda_0, L)$ then

$$\min\{s(\sigma; T_0, p) : \sigma \in \Omega I(T_0)\} > \sum_{\sigma \in \Omega(n) - \Omega(T)} w(\sigma, T_0) s(\sigma; T_0, p),$$

since $w(\sigma, T) > 1$ if $\sigma \in \Omega(n) - \Omega(T)$. Then

$$\sum_{\sigma} w(\sigma, T_0) s(\sigma; T_0, p) <$$

$$\sum_{\sigma \in \Omega I(T_0)} s(\sigma; T_0, p) + \sum_{\sigma \in \Omega P(T_0)} s(\sigma; T_0, p) + \min\{s(\sigma; T_0, p) : \sigma \in \Omega I(T)\}. (*)$$

Now if $T \neq T_0$, then $w(\sigma, T) \geq 1$ for all $\sigma \in \Omega I(T_0)$ and $w(\sigma, T) \geq 2$ for at least one $\sigma \in \Omega I(T_0)$, while $w(\sigma, T) = 1$ for $\sigma \in \Omega P(T_0)$. Thus

$$\sum_{\sigma \in \Omega(n)} w(\sigma, T) s(\sigma; T_0, p) \geq$$

$$\sum_{\sigma \in \Omega I(T_0)} s(\sigma; T_0, p) + \min\{s(\sigma; T_0, p) : \sigma \in \Omega I(T_0)\} + \sum_{\sigma \in \Omega P(T_0)} s(\sigma; T_0, p) >$$

$$\sum_{\sigma \in \Omega(n)} w(\sigma, T_0) s(\sigma; T_0, p), \text{ by } (*).$$

Thus $\sum_{\sigma \in \Omega(n)} w(\sigma, T) s(\sigma; T_0, p) > \sum_{\sigma \in \Omega(n)} w(\sigma, T_0) s(\sigma; T_0, p)$ for all $T \neq T_0$ so that Wagner parsimony is consistent on (T, p) by lemma 7.14, as required. §

7.37 Remark

In view of the previous two theorems it is tempting to ask whether Wagner parsimony is consistent on $B(\lambda, L)$ for sufficiently small $\lambda = \lambda(L)$. One approach might be to show that any colouring of a given weight on a tree in $B(\lambda, L)$ is more probable than any other colouring of greater weight on that tree for $\lambda = \lambda(L)$ sufficiently small, thus generalizing the result for colourings of weight one. However, the next theorem gives a counterexample to this suggestion, and sheds more detail on the subtle relationship between weight of colourings and their probability.

7.38 Theorem

Even for $L=1$, there are no constants $\delta, \lambda, k > 0$ for which the following statement holds for all $(T, p) \in B(\lambda)$:

" $w(\sigma, T) < w(\sigma', T) + k$ implies $s(\sigma; T, p) \geq \delta s(\sigma'; T, p)$."

Proof:

Consider the two partitions σ_1, σ_2 of the caterpillar tree, J_{2n+2k} , defined by the colourings:

$(A, A, B, B, A, A, B, B, A, A, \dots)$ and $(A, A, A, \dots [2k \text{ times}], B, A, B, A, B, A, \dots, B, A)$, respectively as in fig. 7.7 (refer to following page), where in all cases $\{x, x'\} = \{A, B\}$.

Then $w(\sigma_1, J_{2n+2k}) = n+k-1$, while $w(\sigma_2, J_{2n+2k}) = n$.

Let J_{2n+2k}^* denote the tree obtained from J_{2n+2k} by deleting pendant vertex labelled $2n+2k$ and its associated pendant edge e and distinguishing the vertex v incident with e .

For σ_1 as before, let σ_1^+ (resp. σ_1^-) denote the colouring of v and the pendant vertices of J_{2n+2k}^* in which the pendant vertices are coloured as in J_{2n+2k} under σ_1 , and v is assigned the same colour (resp. the opposite colour) as the pendant vertex labelled $2n+2k-1$. For σ_2 , define σ_2^+ and σ_2^- analogously. These definitions are also illustrated in fig. 7.7.

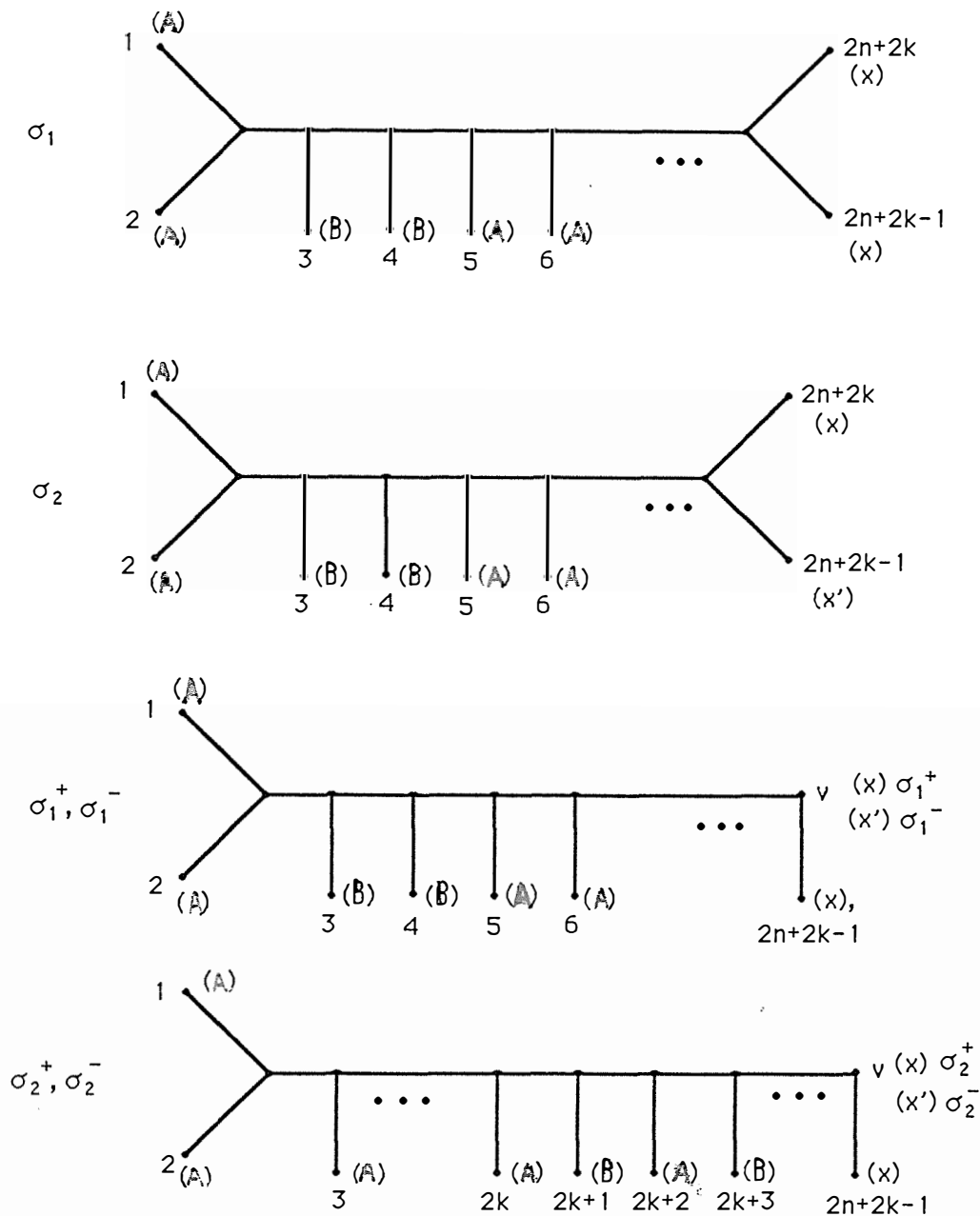


Figure 7.7

Assign edge change probability p to all the edges of J_{2n+2k}^* , and let $s_1(n)$, (resp. $d_1(n)$) denote probability of the colouring σ_1^+ (resp. σ_1^-) on the weighted tree (J_{2n+2k}^*, p_j) . Define $s_2(n), d_2(n)$ for σ_2 similarly. Then a straightforward probability argument gives:

$$[s_1(n+1), d_1(n+1)]^t = C[s_1(n), d_1(n)]^t$$

$$[s_2(n+1), d_2(n+1)]^t = D^2[s_2(n), d_2(n)]^t, \text{ where}$$

$$C = \begin{bmatrix} (pq^3 + p^2q^2) & (q^4 + pq^3) \\ (p^3q + p^2q^2) & (p^3q + p^2q^2) \end{bmatrix} = \begin{bmatrix} pq^2 & q(p^3 + q^3) \\ p^2q & p^2q \end{bmatrix}$$

$$D = \begin{bmatrix} pq & q^2 \\ pq & p^2 \end{bmatrix} \text{ so that } D^2 = \begin{bmatrix} pq^2 & pq^2 \\ p^2q & p(p^3 + q^3) \end{bmatrix} \text{ with } q=1-p.$$

C and D^2 have characteristic equations: $x^2 - pqx - p^2q^2(\alpha - pq) = 0$ and $x^2 - p(\alpha + q^2)x - p^2q^2(pq - \alpha) = 0$, respectively, where $\alpha = p^3 + q^3$. For $i=1,2$, let $\lambda_1^+ \geq \lambda_1^-$ denote the eigenvalues of C , and let $\lambda_2^+ \geq \lambda_2^-$ denote the eigenvalues of D^2 .

Then

$$\lambda_1^\pm = 0.5(pq \pm \sqrt{(pq)^2 + 4p^2q^2(\alpha - pq)}),$$

$$\lambda_2^\pm = 0.5(p(\alpha + q^2) \pm \sqrt{(p(\alpha + q^2))^2 + 4p^2q^2(pq - \alpha)}).$$

It is easily checked that for $0 < p < 0.5$, and $i=1,2$, $\lambda_i^+ \neq \lambda_i^-$, so that C and D^2 are similar to diagonal matrices. Thus for $i=1,2$,

$$s_i(n) = a_i(\lambda_i^+)^n + b_i(\lambda_i^-)^n \text{ and}$$

$$d_i(n) = c_i(\lambda_i^+)^n + d_i(\lambda_i^-)^n, \text{ with } a_i, b_i, c_i, d_i \text{ independent of } n \text{ (but with } a_2, b_2, c_2, d_2 \text{ depend on } k).$$

Now

$$s(\sigma_1; J_{2n+2k}, p_j) = s_1(n-1)(q^2p + p^2q) + d_1(n-1)(q^3 + p^3), \text{ while}$$

$$s(\sigma_2; J_{2n+2k}, p_j) = s_2(n-1)(q^2p + p^2q) + d_2(n-1)(q^2p + p^2q), \text{ where } q = 1-p.$$

Thus $s(\sigma_1; J_{2n+2k}, p) = a'_1(\lambda_1^+)^n + b'_1(\lambda_1^-)^n$

$$s(\sigma_2; J_{2n+2k}, p) = a'_2(\lambda_2^+)^n + b'_2(\lambda_2^-)^n,$$

for new constants a'_i, b'_i , so that

$$s(\sigma_1; J_{2n+2k}, p) / s(\sigma_2; J_{2n+2k}, p) \sim (a'_1/a'_2)(\lambda_1^+/\lambda_2^+)^n \text{ since } \lambda_1^+ > \lambda_2^+.$$

Now $\lambda_1^+/\lambda_2^+ = (q + \sqrt{(q^2 + 4q^2(\alpha - pq))}) / (\alpha + q^2 + \sqrt{((\alpha + q^2)^2 + 4q^2(pq - \alpha))})$, so

that $\lim_{p \rightarrow 0} \lambda_1^+/\lambda_2^+ = (1 + \sqrt{5})/2$, the "golden ratio". Thus we can

certainly choose $p_0 > 0$ so that for all $p < p_0$, $\lambda_1^+/\lambda_2^+ > 1$, (in fact we can take $p_0 = 0.5$, though this is not necessary for the proof).

Thus as $n \rightarrow \infty$, $s(\sigma_1; J_{2n+2k}, p) / s(\sigma_2; J_{2n+2k}, p) \rightarrow \infty$, even though

$$w(\sigma_1, J_{2n+2k}) = w(\sigma_2, J_{2n+2k}) + (k-1), \text{ as required. } \S$$

7.39 Remark

We now offer an intuitive explanation for the motto "heavier colourings can be more probable, even on trees with equal edge weights," which was formally demonstrated above.

Recall that $s(\sigma; T, p) = \sum_{\{C \subseteq E(T): p(C) = \sigma\}} p^{|C|} (1-p)^{|E(T)| - |C|} = (1-p)^{|E(T)|} \sum_{k \geq w(\sigma, T)} |\{C \subseteq E(T): p(C) = \sigma, |C| = k\}| \epsilon^k$, where $\epsilon = p/(1-p)$, so that $0 < \epsilon < 1$.

Regard $s(\sigma) = s(\sigma; T, p)$ as a polynomial in ϵ , ordered in increasing powers of ϵ . Then although $w(\sigma_1, T) > w(\sigma_2, T)$, (so that $s(\sigma_1)$ begins with higher powers of ϵ than $s(\sigma_2)$), if the first nonzero coefficients of $s(\sigma_1)$ are relatively much larger than those of $s(\sigma_2)$ then we may still have $s(\sigma_1) > s(\sigma_2)$ for suitable ϵ (though clearly for T fixed, as ϵ tends to zero, we will force $s(\sigma_1) < s(\sigma_2)$).

Thus $w(\sigma_1, T) > w(\sigma_2, T)$ and $s(\sigma_1, T) > s(\sigma_2, T)$ suggests many more ways to fit σ_1 to T with $\approx w(\sigma_1, T)$ edge changes than to fit σ_2 to T with the same or a lesser number of edge changes. For example, take σ_1, σ_2 and $T = J_{2n+2k}$ as in the previous theorem, and consider just the first non-zero coefficients k_1, k_2 of $s(\sigma_1), s(\sigma_2)$, respectively. For $i=1,2$ choose $C_i \subseteq E(T)$ so that $p(C_i) = \sigma_i$, (where p is defined in 5.2). Then it is easily checked that C_2 has no weakly-connecting trees and so by theorem 5.9, C_2 is strictly minimal, so that $k_2=1$.

Now the weakly-connecting trees of C_1 consist of the $n+k-2$ pairs of adjacent internal vertices of T which are both adjacent to pendant vertices of the same colour.

Thus the weakly-connecting forests of T consist of all collections of weakly-connecting trees, no two of which are adjacent to a common vertex. Since the $n+k-2$ weakly-connecting trees are arranged in a line, the number of weakly-connecting forests is precisely the number of ways of selecting a subset of $n+k-2$ ordered objects so that no two are consecutive. But this is enumerated in example 2.2.23 of Goulden and Jackson [1983, p.43], in terms of $\{0,1\}$ sequences, as the $(n+k-1)$ -th Fibonacci number, F_{n+k-1} . Thus, by theorem 5.9 (3), $k_1 = F_{n+k-1}$, (compared with $k_2=1$) so that as n grows the first nonzero coefficient of $s(\sigma_1)$ greatly exceeds the corresponding coefficient of $s(\sigma_2)$. §

7.40 Conjecture

Wagner parsimony is consistent on $B(\lambda)$ for λ sufficiently small.

Consistency under a molecular clock

Hendy and Penny [1988] give an example of a tree (a $T_1(2)$) with edge weights arbitrarily small and subject to a molecular clock for which Wagner parsimony and compatibility are inconsistent. The authors observe that for this to happen one edge length must be asymptotically proportional to the square of the length of another edge, as both edge lengths tend to zero. To find a "best possible" condition under which we might hope for parsimony and compatibility to be consistent under a molecular clock we might try and rule this out.

In fact we will see that even assuming a molecular clock, inconsistency can occur even in the "best possible" condition that does not in itself constrain the set of possible trees. Specifically suppose the Cavender model is subject to a molecular clock. The first attempt at a "best possible constraint" on the edge lengths might be to assume that they are all equal. However unlike the clock-free case, this assumption constrains the tree topology and *a fortiori* constrains the number of taxa to be a power of two.

Similar topological constraints follow from supposing that the edge length ratios are bounded above by a constant which is independent of the number of taxa. On the other hand, if the ratio of the edge lengths of rooted trees with n pendant vertices is bounded above by some function $K(n)$, there are no constraints on the possible trees, precisely if $K(n) \geq n-1$. This suggests the following definition.

7.41 Definition

Call a rooted binary tree with n pendant vertices *balanced* if

(1): it has additive edge lengths--i.e. the length of all the paths from the root to pendant vertices are the same (this is the *height* of the tree).

(2): the ratio of the edge lengths is bounded above by $n-1$.

Let $BE(\lambda)$ be the set of balanced (rooted) trees, with edge lengths bounded above by λ , and let $BH(\lambda)$ be the set of balanced (rooted) trees of height λ .

Then for the counterexample by Hendy and Penny [1988], if the tree $T_1(2)$ is balanced, consistency is regained for sufficiently small edge lengths, or sufficiently small height. We now show that for any given bound on the edge lengths or the height there are always trees for which both types of parsimony fail to be consistent.

7.42 Theorem

(1): For all λ , Wagner parsimony and compatibility are inconsistent on some trees in $BE(\lambda)$.

(2): For all λ , Wagner parsimony and compatibility are inconsistent on some trees in $BH(\lambda)$.

Proof:

Let $RT_1(k, \lambda)$ be the tree obtained by rooting $T_1(k)$ (definition 7.27) at the midpoint of its central edge, and assigning length (as in 6.4 (4)) λ to all the edges except the two pendant edges incident with the central edge which are assigned length $\lambda(k+1)$. Then since $RT_1(k, \lambda)$ has $2^{k+1}+2$ pendant vertices, $RT_1(k, \lambda)$ is balanced for all $k \geq 1$.

(1): Regarding the tree as unrooted, and using the notation of theorem 7.28, we have for Wagner parsimony,

$$4E[Z(k)] = \omega_3(\omega_1^2 + \omega_2^2)(S^2(k) + D^2(k)) - 2\omega_1\omega_2(S^2(k) - D^2(k)) + 2\omega_3(\omega_2^2 - \omega_1^2)S(k)D(k).$$

As in the proof of corollary 7.29 it suffices to show

$\lim_{k \rightarrow \infty} E[Z(k)] > 0$ since then $E[Z(k_0)] > 0$ for some integer k_0 .

Now $\omega_1 = e^{-2\lambda}$, $\omega_3 = e^{-4\lambda}$, $\omega_2 = e^{-2\lambda(k+1)}$, so that $\lim_{k \rightarrow \infty} \omega_2 = 0$.

Thus, letting $s = \lim_{k \rightarrow \infty} S(k)$, $d = \lim_{k \rightarrow \infty} D(k)$,

$\lim_{k \rightarrow \infty} 4E[Z(k)] = 2\omega_1^3(s^2 + d^2 - 2\omega_1 s d) > 2\omega_1^3(s - d)^2 \geq 0$, and the result follows by theorem 7.28.

For compatibility, again using the notation of theorems 7.28, and 7.26

$$\lim_{k \rightarrow \infty} \frac{4E[Z^*(k)]}{Q^2(k)} = 2\omega_1^3(1+\tau^2-2\omega_1\tau) > (1-\tau)^2 \geq 0, \text{ and the result}$$

follows by theorem 7.28.

(2): We have $\cup_k RT_1(k, \lambda/(k+2)) \subset BH(\lambda)$. For $T \in RT_1(k, \lambda/(k+2))$,

$$\omega_1 = e^{-2\lambda/(k+2)}, \omega_3 = e^{-4\lambda/(k+2)}, \omega_2 = e^{-2\lambda(k+1)/(k+2)}.$$

$$\text{Thus } \lim_{k \rightarrow \infty} \omega_1 = \lim_{k \rightarrow \infty} \omega_3 = 1, \lim_{k \rightarrow \infty} \omega_2 = e^{-2\lambda}.$$

For Wagner parsimony we have

$$\begin{aligned} \lim_{k \rightarrow \infty} 4E[Z(k)] &= (1+e^{-4\lambda})(s^2+d^2) - 2e^{-4\lambda}(s^2-d^2) + 2sd(e^{-4\lambda}-1) \\ &= (s-d)^2 + (3d^2+2sd-s^2)e^{-4\lambda}. \end{aligned}$$

Now if the second term in this sum is strictly positive we have

$$\lim_{k \rightarrow \infty} E[Z(k)] > 0, \text{ as required. If the second term is } \leq 0 \text{ we have}$$

$$4E[Z(k)] \geq (s-d)^2 + (3d^2+2sd-s^2) = 4d^2 > 0, \text{ by theorem 7.22, as required.}$$

$$\text{For compatibility, } \lim_{k \rightarrow \infty} \frac{4E[Z^*(k)]}{Q^2(k)} = (1-\tau)^2 + (3\tau^2+2\tau-1)e^{-4\lambda}$$

which by a similar argument is strictly positive, since $\tau > 0$ by theorem 7.26. §

7.43 Remarks

(1): The previous theorem shows that imposing a molecular clock in no way restores consistency for parsimony-based methods. Indeed comparing theorems 7.42 and 7.34, a molecular clock is more problematic for the consistency of compatibility than the assumption of equal edge weights.

(2): From example 7.35, compatibility can be consistent under conditions which lead to the inconsistency of Wagner parsimony. It is not known whether the converse can occur, or whether compatibility is consistent for sets of edge lengths of greater Lebesgue measure than Wagner parsimony. It is worth noting that because compatibility uses fewer partitions to evaluate each tree than Wagner parsimony, the former method may be less robust (i.e. more sensitive to errors) when applied to relatively short sequences.

(3): The above results hold an important message for taxonomists -- while compatibility may be consistent for a particular set of taxa, the method is not necessarily consistent on a subset of those taxa. Equivalently, although compatibility may fail to be consistent for a collection of taxa, if that set is expanded to a larger set of taxa, consistency may be regained, so that if compatibility selects tree T , the phylogeny of the original set S of taxa can be consistently recovered as $T[S]$. This approach would succeed if the larger set of taxa is chosen so that all edges of the underlying tree have (approximately) equal (and small) edge weights, though knowing which additional taxa to choose to achieve this goal is clearly a problem. We illustrate this relativity of consistency by the following example.

7.44 Example

Consider the rooted binary tree $T(k+2) \in \text{RWT}(2^{k+2})$ with edge lengths of λ on each edge, except for the two edges incident with the root, which are assigned edge length $\lambda/2$, as illustrated on the following page in fig.

7.8(a) (where the shaded subtrees are copies of $T(k)$). Thus the edge lengths of $T(k+2)$ are subject to a molecular clock, while if $T^*(k+2)$ denotes the weighted binary tree obtained from $T(k+2)$ by suppressing the root, $T^*(k+2)$ has equal edge lengths of λ on each edge. Select a subset $S = \{s_1, s_2, s_3, s_3, s_5, s_6\}$ of $\{1, \dots, 2^{k+2}\}$ of size 6 as indicated in fig 7.8(a). Then the tree $T^*(k+1)[S]$ has induced edge lengths as indicated in fig. 7.8(b). In the notation of definition 7.27, $T^*(k+2)[S]$ is a $T_1(1, p)$ where

$p = (p_1, p_2, p_3, p_4)$, $p_1 = p_3 = 0.5(1-\omega)$, $p_2 = 0.5(1-\omega^{k+1})$, $p_4 = 0.5(1-\omega^k)$, with $\omega = e^{-2\lambda}$.

Then in the notation of theorem 7.28,

$P(1) = p_4^2 = 0.25(1-\omega^k)^2$, and $Q(1) = (1-p_4)^2 = 0.25(1+\omega^k)^2$, $\omega_1 = \omega_3 = \omega$, $\omega_2 = \omega^{k+1}$, so that by theorem 7.28, the condition for compatibility to be consistent on $T_1(1, p)$ is that

$$f(k, \omega) = \omega(\omega^2 + \omega^{2k+2})((1-\omega^k)^4 + (1+\omega^k)^4) - 2\omega^{k+2}((1+\omega^k)^4 - (1-\omega^k)^4) + 2\omega(\omega^{2k+2} - \omega^2)(1-\omega^k)^2(1+\omega^k)^2 > 0.$$

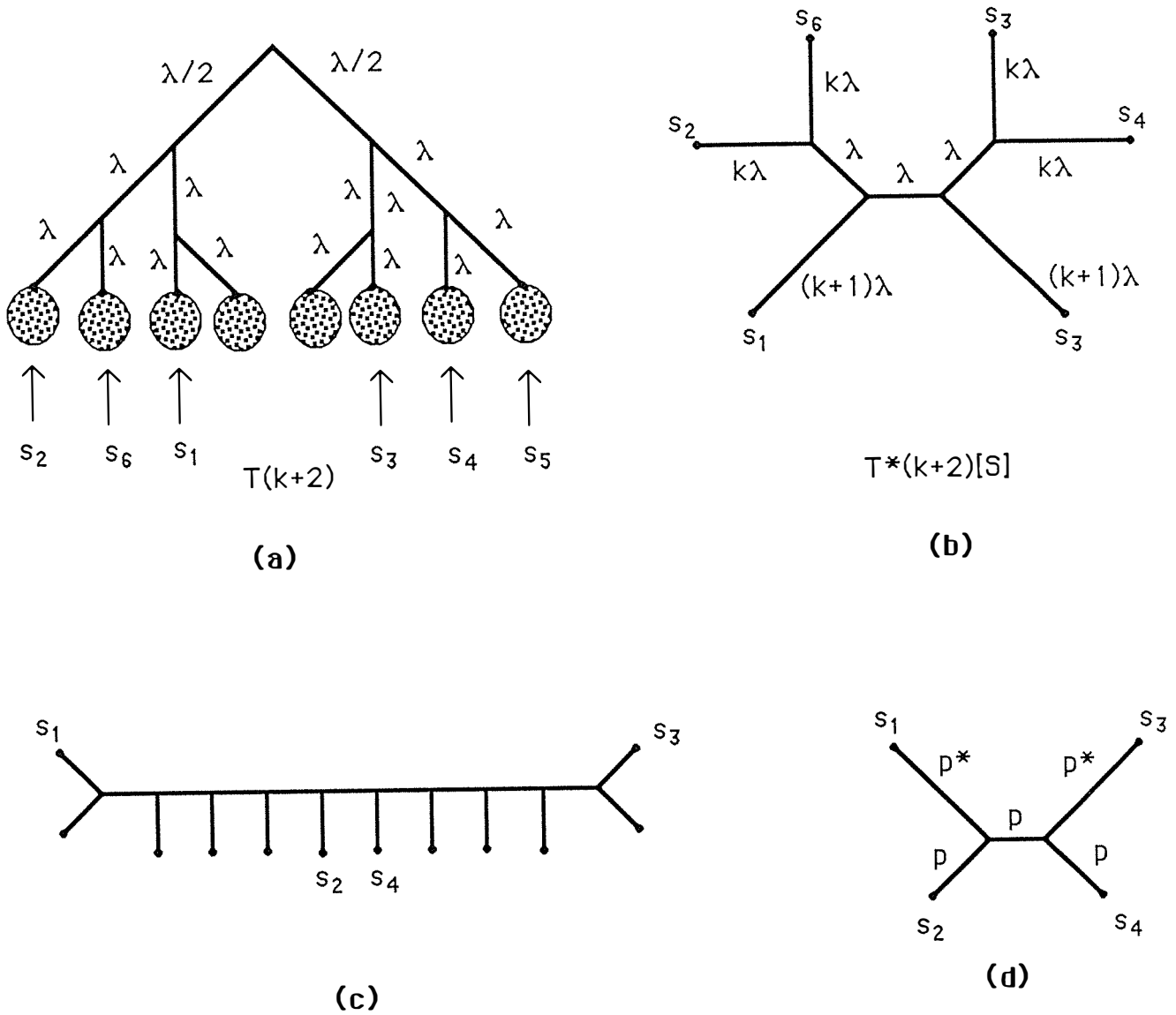


Figure 7.8

Letting $y = \omega^{2k}$, we have $f(k, \omega) = 4\omega^2 y (5\omega - 4 - y(4 - 2\omega - \omega y))$.

Since $\lim_{k \rightarrow \infty} \omega^k = 0$, we see that if $\omega > 0.8$, (that is if

$p_1 = 0.5(1 - e^{-2\lambda}) < 0.1$) we can choose k so that $f(k, \omega) > 0$, implying that compatibility will be inconsistent for the taxa set S . But this range for ω contains the range for which we have established that compatibility will be consistent on the weighted parent tree $(T^*(k+2), p_{1j})$.

In particular taking $k = 7$, and $\lambda = 0.05$ (so that $\omega \approx 0.9048$) we have:

- (a): $p < 0.1$ (so that compatibility is consistent on $T^*(k+2), p_1j$) and
- (b): $f(7, \omega) > 0$ (so that compatibility is not consistent on the weighted subtree induced by S).

Thus for sequence data derived under Cavender's model from the bifurcating tree $T(9)$ on 512 endpoints (!) with an expected number of change of 0.1 on each edge, compatibility will consistently recover the original phylogeny of the 512 taxa, but consistency will fail if applied to some of the subsets of the taxa.

Examples of this phenomena using considerably less taxa are possible if edge lengths on the parent tree are not required to satisfy a molecular clock. Indeed for the caterpillar tree J_{12} on 12 taxa, with weight p on each edge, and $S = \{s_1, s_2, s_3, s_4\}$ as shown in fig. 7.8(c), the weighted subtree $T[S]$ induced by S has edge weights as shown in fig. 7.8(d), where $p^* = 0.5(1 - (1 - 2p)^5)$.

Then by Felsenstein's criterion (refer to example 7.15), compatibility will fail to be consistent on $T[S]$ if $(p^*)^2 > p(1 - p)$.

Taking $p = 0.1$ (so that compatibility is consistent on T) this condition is satisfied. §

7.45 Summary

In this section we have derived some necessary and sufficient conditions for selection procedures to be consistent under Cavender's model. Necessary conditions have been given both for general classes of procedures (theorem 7.9) and for parsimony and compatibility (theorems 7.28 and 7.42). We see that for both methods consistency can fail in conditions that might be expected to be most favourable for consistency (i.e. equal edge weights or a molecular clock). Basically both methods fail because they undercount the true amount of change which is likely to occur on the tree. Regarding sufficient conditions, we showed that (without assuming a molecular clock) phylogenies can be consistently recovered from dissimilarities (theorem 7.12). Perhaps most importantly nontrivial sufficient conditions for the consistency of compatibility have been established (theorem 7.34). The problems involved in extending this approach to parsimony have been discussed, and clearly there is an interesting and important open problem in settling conjecture 7.40.

§8: Confidence Intervals

"The question of how to obtain confidence intervals and carry out statistical tests is in a relatively primitive state ... but is of greater practical importance to the molecular evolutionist," J. Felsenstein 1988.

8.1 Introduction

Consistency is certainly a desirable property for a selection procedure, but by itself, under finite sampling, it is essentially useless for putting confidence intervals on the trees so constructed. In this section we consider three questions related to the construction of trees:

- (1): Among selection procedures, what sort of confidence intervals, which converge to the true tree (as $c \rightarrow \infty$) are possible, independent of the underlying tree (T_0, p)?
- (2): Can these confidence intervals be described efficiently, that is in polynomial time?
- (3): How fast must the data grow as a function of the number of taxa in order to retain a given level of accuracy in reconstructing the underlying tree T_0 ?

The construction of confidence intervals in the case of parsimony with four taxa has been solved by Cavender [1978] and Felsenstein [1985]. In general however there are problems with these and other methods which have been proposed for constructing confidence intervals from sequence data or dissimilarities, as detailed in Felsenstein [1988]. Constructing confidence intervals using maximum likelihood ratios is also theoretically difficult for two reasons, as is frequently pointed out by J. Felsenstein (see for example Felsenstein [1988]). These difficulties are that the hypotheses being decided between are not nested within each other, and that likelihood ratio results are asymptotic (in the length of the sequences).

We begin by answering the first question and show that under Cavender's model only selection procedures that build (arbitrarily large) sets of trees can have pre-set confidence intervals, and describe such a procedure. The desire to always select a single "best tree" is thus incompatible with the desire to be confident that one has always selected the correct tree, amongst those chosen.

Confidence Intervals

8.2 Definition

A family of selection procedures, φ_α , $\alpha \in (0,1)$ has *convergent confidence intervals (CCI)* if for all α :

(a): $\varphi_{\alpha'}(X) \subseteq \varphi_\alpha(X)$ if $\alpha \leq \alpha'$.

(b): $P(T \in \varphi_\alpha(X); T, p) \geq 1 - \alpha$, for all $(T, p) \in \text{WBT}(n)$, $p \in (0, 0.5)^{2n-3}$.

(c): φ_α converges (as defined in 7.4).

8.3 Lemma

If $\{\varphi_\alpha: \alpha \in (0,1)\}$ is a family of CCI selection procedures, then φ_α is consistent for each $\alpha \in (0,1)$.

Proof:

From property (c), there exists a number c_0 (dependent on T, p, ϵ and α):

$$P(|\varphi_\alpha(X)| = 1; T, p) \geq 1 - \epsilon/3, \text{ and } P(|\varphi_{\epsilon/3}(X)| = 1; T, p) > 1 - \epsilon/3,$$

for all for all $c > c_0$.

Now if $0 < \epsilon/3 < \alpha$ and $|\varphi_\alpha(X)| = |\varphi_{\epsilon/3}(X)| = 1$ and $T \in \varphi_{\epsilon/3}(X)$, then by

property (a), $\varphi_\alpha(X) = \{T\}$. Thus for $0 < \epsilon/3 < \alpha$, and $c > c_0$,

$$P(\varphi_\alpha(X) = \{T\}; T, p) \geq P(T \in \varphi_{\epsilon/3}(X) \& |\varphi_\alpha(X)| = 1 \& |\varphi_{\epsilon/3}(X)| = 1; T, p) \geq 1 - \epsilon$$

by lemma 6.2 (1), and property (b). Since this holds for all ϵ : $0 < \epsilon/3 < \alpha$ (for

appropriate choice of c_0), $\lim_{c \rightarrow \infty} P(\varphi_\alpha(X) = \{T\}; T, p) = 1$, as required. §

While condition (b) succeeds in replacing the asymptotic nature of consistency by actual probabilities, condition (c) does not give any control over the size of $\varphi(X)$ for finite values of c . The reason, as shown next, is that no such control is possible -- in particular there can be no CCI procedure that always selects exactly one tree.

Indeed even if we were only to apply the procedure if the data was in some sense "good" enough we still cannot have control on $|\varphi(X)|$. This last consideration amounts to considering conditional probabilities, and motivates the following definition, (where $P(A|B)$ is the conditional probability of A given B).

8.4 Definition

A family φ_α , $\alpha \in (0,1)$ of selection procedures has *bounded confidence intervals (BCI)* if

(a): $\varphi_{\alpha'}(X) \subseteq \varphi_\alpha(X)$ if $\alpha \leq \alpha'$.

(b): $P(T \in \varphi_\alpha(X) \text{ \& } |\varphi_\alpha(X)| < k \mid E; T, p) \geq h(\alpha, n, k)$, for all (T, p) and some event E (possibly dependent on n, k, α), and some function h with

$\lim_{\alpha \rightarrow 0} h(\alpha, n, k) = 1$ for at least one positive integer $k < b(n)$.

(Here $(T, p) \in WBT(n)$, and $p \in (0, 0.5)^{2^{n-3}}$)

8.5 Remark

This definition is intended to generalize the particular example of a BCI procedure satisfying condition (a) and $P(T \in \varphi_\alpha(X) \mid |\varphi_\alpha(X)| = 1) \geq 1 - \alpha$. Here E is the event $|\varphi_\alpha(X)| = 1$. As another example we might take E to be the event that for some $T_1 \in BPT(n)$, $X_\sigma \neq 0$ precisely if $\sigma \in \Omega(T_1)$.

8.6 Lemma

Let $F: 2^X \rightarrow \mathbb{R}^+$. Then

$$\sum_{X \in \mathcal{X}} \sum_{\{A: X \in A \subseteq X, |A| \leq k\}} F(A) \leq k \sum_{\{A: A \subseteq X, |A| \leq k\}} F(A).$$

Proof:

Reversing the order of summation,

$$\sum_{X \in \mathcal{X}} \sum_{\{A: X \in A \subseteq \mathcal{X}, |A| \leq k\}} F(A) = \sum_{\{A: A \subseteq \mathcal{X}, |A| \leq k\}} |A| F(A) \leq \sum_{\{A: A \subseteq \mathcal{X}, |A| \leq k\}} k \cdot F(A). \quad \S$$

8.7 Theorem

(1): CCI selection procedures exist.

(2): BCI selection procedures do not exist.

Proof:

(1): Define a selection procedure φ_α as follows:

Let s be any function from $2^{\text{BPT}(n)}$ to $\text{BPT}(n)$ such that $s(P) \in P$ for all $P \subseteq \text{BPT}(n)$.

For each tree $T \in \text{BPT}(n)$, let

$$\epsilon(T, X) = \min\{\|X/c - s(T, p)\| : p \in [0, 0.5]^{2n-3}\} \text{ and let}$$

$$\varphi_\alpha(X) = \begin{cases} T: \epsilon(T, X) \leq 1/\sqrt{c\alpha} \end{cases}; \text{ if this set is } \neq \emptyset$$

$$\left\{ s(\{T: \epsilon(T, X) = \min\{\epsilon(T', X): T' \in \text{BPT}(n)\}) \right\}, \text{ otherwise.}$$

We show φ_α is a CCI. Condition (a) is clearly satisfied.

For (b), if T_0, p_0 is the underlying tree, $p_0 \in (0, 0.5)^{2n-3}$,

$$E[\|X/c - s(T_0, p_0)\|^2] = c^{-2} E[\sum_{\sigma} (X_{\sigma} - cs(\sigma; T_0, p_0))^2] =$$

$$c^{-2} \sum_{\sigma} E[(X_{\sigma} - cs(\sigma; T_0, p_0))^2] = c^{-2} \sum_{\sigma} \text{Var}[X_{\sigma}]. \text{ Now since } X \text{ has a}$$

multinomial distribution, by lemma 6.2 (3)(a), $\text{Var}[X_{\sigma}] = cs(\sigma)(1-s(\sigma))$, so

$$\text{that } E[\|X/c - s(T, p)\|^2] = (1 - \sum_{\sigma} s^2(\sigma))/c \leq 1/c. \quad (*)$$

Now $P(T_0 \in \varphi_\alpha(X)) \geq P(\epsilon(T_0, X) \leq 1/\sqrt{c\alpha})$

$\geq P(\|X/c - s(T_0, p_0)\|^2 \leq 1/c\alpha) = 1 - P(\|X/c - s(T_0, p_0)\|^2 > 1/c\alpha) \geq 1 - \alpha$, by

Markov's inequality (lemma 6.2 (2)) establishing condition (b).

For (c), let $\delta = \min\{\|s(T, p) - s(T_0, p_0)\| : p \in [0, 0.5]^{2n-3}, T \neq T_0\}$. Then $\delta > 0$ by lemma 6.13 (4). By Markov's inequality (lemma 6.2 (2)) and (*) we have,

$P(\|X/c - s(T_0, p_0)\| \geq \delta/2) \leq 4/\delta^2 c$. Now if for any tree $T \neq T_0$, $T \in \varphi_\alpha(X)$, then

for some p , $\|X/c - s(T, p)\| \leq 1/\sqrt{c\alpha}$ and for $c > 4/\delta^2 \alpha$, the triangle

inequality gives $\|X/c - s(T_0, p_0)\| \geq \delta/2$. Thus $\lim_{c \rightarrow \infty} P(\exists T \neq T_0 : T \in \varphi_\alpha(X)) = 0$.

It thus remains to show that $\lim_{c \rightarrow \infty} P(T_0 \in \varphi_\alpha(X)) = 1$. In view of the

definition of φ_α we need only check that

$\lim_{c \rightarrow \infty} P(\epsilon(T_0, X) < \min\{\epsilon(T, X) : T \neq T_0\}) = 1$.

Now $\lim_{c \rightarrow \infty} P(\|X/c - s(T_0, p_0)\| < \delta/2) = 1$, while if $\|X/c - s(T, p)\| \leq \delta/2$, for any $T \neq T_0$ and some p , the triangle inequality, together with the definition of δ gives $\|X/c - s(T_0, p_0)\| \geq \delta/2$, an event which has probability tending to zero as $c \rightarrow \infty$, as required. §

(2): Suppose a BCI procedure exists. Given $T \in \text{BPT}(n)$, let $T(\epsilon) \in \text{WBT}(n)$ have edge weight ϵ on all internal edges and $0.5 - \epsilon$ on all pendant edges. We first note from lemma 6.13 (5) that for any event E , $P(E; T(\epsilon)) > 0$, so that we may form all conditional probabilities.

Let $E \in \Delta_2(n, c)$, $E \neq \emptyset$. Then for any $T \in \text{BPT}(n)$, lemma 6.13 (6), and the continuity of $P(\sigma; T, p)$ as a function of p imply that $\lim_{\epsilon \rightarrow 0} P(E; T(\epsilon))$ is positive and independent of T . Thus for any tree $T^0 \in \text{BPT}(n)$, any event E and any $\delta > 0$ we can choose $\epsilon > 0$ so that $|P(E; T(\epsilon)) - P(E; T^0(\epsilon))| < \delta$.

Now $\sum_{T \in \text{BPT}(n)} P(T \in \varphi_\alpha(X) \text{ \& } |\varphi_\alpha(X)| \leq k\frac{1}{b(n)}E; T(\epsilon)) \leq$
 $\sum_{T \in \text{BPT}(n)} P(T \in \varphi_\alpha(X) \text{ \& } |\varphi_\alpha(X)| \leq k\frac{1}{b(n)}E; T^0(\epsilon)) + b(n)\delta =$
 $\sum_{T \in \text{BPT}(n)} \sum_{A: T \in A, |A| \leq k} P(\varphi_\alpha^{-1}(A) \text{ \& } E; T^0(\epsilon)) + b(n)\delta \leq$
 $k \sum_{A: |A| \leq k} P(\varphi_\alpha^{-1}(A) \text{ \& } E; T^0(\epsilon)) + b(n)\delta$, by lemma 8.6.
 $= kP(|\varphi_\alpha(X)| \leq k\frac{1}{b(n)}E; T^0(\epsilon)) + b(n)\delta$. Now if φ_α has BCI, we have
 $P(T \in \varphi_\alpha(X) \text{ \& } |\varphi_\alpha(X)| < k\frac{1}{b(n)}E; T(\epsilon)) \geq h(\alpha, n, k)P(E; T(\epsilon))$
 $\geq h(\alpha, n, k)P(|\varphi_\alpha(X)| < k\frac{1}{b(n)}E; T(\epsilon))$.
 Thus $\sum_{T \in \text{BPT}(n)} P(T \in \varphi_\alpha(X) \text{ \& } |\varphi_\alpha(X)| < k\frac{1}{b(n)}E; T(\epsilon)) \geq$
 $h(\alpha, n, k) \sum_{T \in \text{BPT}(n)} P(|\varphi_\alpha(X)| < k\frac{1}{b(n)}E; T(\epsilon))$
 $\geq h(\alpha, n, k)b(n)(P(|\varphi_\alpha(X)| \leq k\frac{1}{b(n)}E; T^0(\epsilon)) - \delta)$. Combining this with the previous
 calculation we have, letting $\beta = P(|\varphi_\alpha(X)| \leq k\frac{1}{b(n)}E; T^0(\epsilon))$,
 $k\beta + b(n)\delta \geq h(\alpha, n, k)b(n)(\beta - \delta)$, giving $h(\alpha, n, k) \leq (k\beta + b(n)\delta) / (b(n)(\beta - \delta))$. Let δ
 $\rightarrow 0$ (choosing $\epsilon = \epsilon(\delta) > 0$ as $\delta \rightarrow 0$). Then since β does not tend to zero by
 lemma 6.13 (5), we have $h(\alpha, n, k) \leq k / b(n)$ for all α , so that φ is not BCI. §

We now address the question, raised in the construction of a CCI procedure
 in the previous theorem, of how to locate the closest point in **s**-space.

Approximate methods

Suppose the edge weights are assumed to be small, so that the product of
 any two edge probabilities can be neglected. One way to encapsulate this
 notion formally is to regard the edge weights p_i as indeterminants and
 work with the algebra **A** over **R** generated by $\{p_i\}_i$, subject to the formal
 identities $p_i p_j = 0$ for all i, j .

Then for any tree T we can solve exactly for the closest point in $\mathbf{s}$ -space, and the "best fit tree" is given by a combinatorial condition similar to that for compatibility. After presenting this result we show that such procedures are basically unsound for the following reason: even on four taxa, no matter how small the edge weights really are, inconsistency can still arise using this method.

8.8 Lemma

In the algebra A , $s(\sigma; T, p) = \begin{cases} p_e, & \text{if edge } e \text{ induces } \sigma \\ 1 - \sum_{e \in E(T)} p_e, & \text{if } \sigma = \sigma_0 \\ 0, & \text{otherwise.} \end{cases}$

Proof:

The lemma follows immediately from

$$s(\sigma; T, p) = \sum_{p(E)=\sigma} \prod_{e \in E} p_e \prod_{e \in E(T)-E} (1-p_e). \quad \S$$

8.9 Theorem

Assume that X satisfies $\max\{X_\sigma: \sigma \neq \sigma_0\} \leq c^{(n-2)}/2^{(n-1)}$.

$$\text{Let } \alpha_1(T) = \sum_{\sigma \in \Omega(n)-\Omega(T)} X_\sigma, \quad \alpha_2(T) = \sum_{\sigma \in \Omega(n)-\Omega(T)} X_\sigma^2.$$

(1): For any tree, T , the edge weights p that minimize the distance

$\|X/c - s(T, p)\|$ in $\mathbf{s}$ -space under A (i.e. with $s(\sigma)$ replaced by its value in A)

are given by:

$$p_e = \alpha_1(T)/c^{(2n-2)} + X_{\sigma(e)}/c,$$

where $\sigma(e)$ is the partition induced by deletion of e .

(2): The trees in $BPT(n)$ which minimize (over $BPT(n)$) the minimum distances given in (1) are precisely those trees T which minimize the sum $\alpha_1(T)^2/(2n-2) + \alpha_2(T)$ over $BPT(n)$.

(By comparison recall that compatibility choses that tree minimizing $\alpha_1(T)$ over $BPT(n)$).

Proof:

(1): By lemma 8.8, in A,

$\partial/\partial p_e \|X/c-s\|^2 = -2(X_{\sigma(e)} - cp_e) + 2(X_{\sigma_0} - c(1 - \sum_{e \in E(T)} p_e))$, where $\sigma(e)$ is the partition corresponding to deletion of the edge e .

The linear system, $\partial\|X/c-s\|^2/\partial p_e = 0$ for all $e \in E(T)$, can be written

$(I+J)p = c^{-1}(X^\# - (X_{\sigma_0} - c)j)$, where $X^\#$ is the vector $[X_{\sigma(e)}]_{e \in E(T)}$, and J is the square matrix having 1 in each position. Since $(I+J)^{-1} = I - (2n-2)^{-1}J$, the solution to this linear system is:

$$p_e = (c - X_{\sigma_0} + (2n-2)X_{\sigma(e)} - \sum_{e \in E(T)} X_{\sigma(e)})/c(2n-2).$$

But $c = \sum_{e \in E(T)} X_{\sigma(e)} + X_{\sigma_0} + \alpha_1(T)$. Thus $p_e = \alpha_1(T)/c(2n-2) + X_{\sigma(e)}/c$.

In particular $p_e > 0$, and since $\alpha_1(T) < c$, $p_e < (2n-2)^{-1} + (1-(n-1)^{-1})/2 \leq 0.5$, as required for a feasible solution.

Now $\partial^2 \|X/c-s\|^2/\partial p_e \partial p_{e'} = 4c$ if $e=e'$; $2c$ if $e \neq e'$, so that the Jacobian of $\|X/c-s\|^2$ is $2c(J+I)$ which is positive definite, so the solution for p minimizes $\|X/c-s\|^2$, completing the proof of (1).

(2): The minimum distance in A is $\sum_{\sigma \in \Omega(n)-\Omega(T)} X_\sigma^2 +$

$\sum_{e \in E(T)} (X_{\sigma(e)} - cp_e)^2 + (X_{\sigma_0} - c(1 - \sum_{e \in E(T)} p_e))^2$, which from part (1) is

$\sum_{\sigma \in \Omega(n)-\Omega(T)} X_\sigma^2 + (2n-3)\alpha_1(T)^2/(2n-2)^2 + \alpha_2(T)/(2n-2)^2$, giving (2). §

8.10 Corollary

Let φ_A be the procedure which selects that tree which has an edge weight minimizing the Euclidean distance from s to X/C , under A . Then φ_A is not consistent, even on four taxa, no matter how small the edge weights might be. Thus φ_A differs from another linearized procedure, "closest tree" described by Hendy [1988].

Proof: In view of the previous theorem, we see that φ_A is a central procedure, and the result now follows by theorem 7.2. Closest tree, on the other hand, is consistent by lemma 6.2 (3b). §

8.11 Remarks

(1): Although computationally appealing as indicated by Hendy and Penny [1988], closest tree does not give a family of CCI procedures in the way Euclidean distance in s -space does in theorem 8.7.

(2): While φ_A is not consistent, it may nevertheless be useful in giving an initial iteration value p_1 for numerical methods aimed at finding the value of p which minimizes the Euclidean distance from X/C to $s(T,p)$. This value for p_1 is given by theorem 8.9 (1).

(3): The restriction on X in theorem 8.9 is very mild, for unless $X_\sigma \ll c$ for all $\sigma \neq \sigma_0$, there would be no justification for assuming the edge weights were small, and thereby working in A .

Exact Solutions: (n=4)

Consider the case of four taxa. If the value of p which minimizes $f(p) = \|X/C - s(T,p)\|$ lies in $(0,0.5)^5$, p is given by the solution of the nonlinear system of eight equations in five variables: $\partial f(p)/\partial p_e = 0$. The next theorem shows that this system can be reduced to finding the solution of two equations, each in one variable, thereby allowing faster numerical methods. Furthermore we show that if there is a minimum in $(0,0.5)^5$, it is unique. Of course if the minimum lies on the boundary of $[0,0.5]^{2n-3}$ further analysis is required.

8.12 Theorem

If $p_0 \in (0,0.5)^5$ is a critical point of $f(p) = \|X/C - s(T,p)\|$ then

(1): p_0 can be found by solving two algebraic equations, each in one variable.

(2): p_0 is unique.

Proof:

(1): Using the notation of example 6.10, since $r = H^t s$ and H is Hadamard, the value of p which minimizes $\|X/C - s\|$ is the same value which minimizes $\|r - r^0\|$, and hence $\|r - r^0\|^2$, where $r^0 = H^t(X/C)$.

Let $\omega_j = (1-2p_j)$ and $x = \|r - r^0\|^2 = \sum_i (r_i - r_i^0)^2$.

Then $dx/d\omega_j = 2\sum_i (r_i - r_i^0) \partial r_i / \partial \omega_j$
 $= 2\sum_{i \in K(j)} (r_i - r_i^0) r_i / \omega_j$, where K is the matrix in the proof of lemma 6.11, and $K(j) = \{i: K_{ij}=1\}$.

Thus $dx/d\omega_j = 0$ for all j precisely if $\sum_{i \in K(j)} (r_i - r_i^0) r_i = 0$ for all j .

Let $u_i = (r_i - r_i^0)r_i$, for $i \geq 1$, so that $r_i = (r_i^0 \pm \sqrt{(r_i^0)^2 - 4u_i})/2$. In fact the conditions $r_i > 0$, exclude the negative roots. For if $u_i < 0$, then the negative root gives $r_i < 0$, while if $u_i > 0$, the condition $u_i = (r_i - r_i^0)r_i$ and $r_i > 0$, implies $r_i > r_i^0$, which again requires the positive root.

Then since $u_1 = 0$, letting K_1 be K with the top row deleted, these conditions become $K_1^t u = 0$. Now $\ker K_1^t = [a, b, -b, -b, b, a, -a]^t$, which together with $r_i = (r_i^0 + \sqrt{(r_i^0)^2 - 4u_i})/2$ gives the following conditions on a, b , from which u and hence r can be derived:

$$\begin{aligned} f(a) &= (r_8^0 + \sqrt{(r_8^0)^2 + 4a}) - \\ &\quad 0.5(r_2^0 + \sqrt{(r_2^0)^2 - 4a})(r_7^0 + \sqrt{(r_7^0)^2 - 4a}) = 0; \\ g(b) &= (r_4^0 + \sqrt{(r_4^0)^2 + 4b})(r_5^0 + \sqrt{(r_5^0)^2 + 4b}) - \\ &\quad (r_3^0 + \sqrt{(r_3^0)^2 - 4b})(r_6^0 + \sqrt{(r_6^0)^2 - 4b}) = 0. \end{aligned}$$

(2): Since $f(t)$, and $g(t)$ are monotone increasing in t , there exists at most one solution to the system $f(a) = 0$, $g(b) = 0$, establishing the required uniqueness. §

8.13 Example

Consider $X = [X_1, X_{12}, X_{13}, X_{123}, X_{14}, X_{124}, X_{134}, X_{1234}]^t = [189, 58, 1, 8, 5, 7, 47, 1403]^t$, taken from the EMBL data bank of nucleotide sequences (Hamm and Cameron [1986]), and derived from 18S RNA ribosomal sequences of length $c=1718$ for nematode (1), brine shrimp (2), xenopus (3) and mouse (4).

Applying theorem 8.12 for the tree $T = (12)(34)$ we find $a = -0.00200$, $b = -0.00175$, giving a minimum distance in s -space of 0.00214 (approx). §

Efficiency (I)

Two problems arise from using the CCI procedure φ_α described in theorem 8.7. Firstly, as we have seen, approximate solutions, are statistically unsound. Secondly, finding $\varphi_\alpha(X)$ involves a search among all trees in $BPT(n)$, and so is not efficient.

In fact most of the more widely-used selection procedures have been shown to be NP-complete. These include, parsimony methods (Foulds and Graham [1982]; Day, Johnson and Sankoff [1986]), as well as compatibility (Day and Sankoff [1986]) and dissimilarity methods (Day [1987]). Of course by theorem 8.7, there is no family $\{\varphi_\alpha: \alpha \in (0,1)\}$ of selection procedures which is both efficient in constructing $\varphi_\alpha(X)$, and CCI, because $|\varphi_\alpha(X)| = b(n)$ for α sufficiently small, and $b(n)$ is not bounded above by a polynomial function of n . We now demonstrate the existence of CCI families having the property that for every tree $T \in BPT(n)$ it can be efficiently decided whether $T \in \varphi_\alpha(X)$.

8.14 Definitions

Let Q be a quartet spectrum on label set L . The *restriction* of Q , denoted $(Q)_{res}$ is the set of trees $T \in BPT(L)$ with $\psi(T) \cap Q = \emptyset$.

The *dual* of Q , denoted $\delta(Q)$, is the minimal set defined by the rule: if $(xy)(zw) \in Q$, and $(xz)(yw) \in Q$, then $(xw)(yz) \in \delta(Q)$.

Note that for any such Q , $(Q)_{res} \cap \langle Q \rangle = \emptyset$, and for any binary tree T ,

$(\psi(T)) = \emptyset$, by theorem 3.8 (2), with $k=4$. Furthermore, clearly,

$(Q)_{res} \subseteq \langle \delta(Q) \rangle$, and taking $Q = \psi(T)$, so that $\delta(Q) = \emptyset$, and $\langle \delta(Q) \rangle = BPT(L)$,

we see that this containment is, in general, strict.

Now given any family of CCI selection procedures φ_α on sequence spaces on four taxa, consider the following families $\varphi^*_\alpha, \varphi^{**}_\alpha$ defined on sequence spaces on n taxa. Define $g(\alpha, n) = \alpha/nC_4$, and let $Q_\alpha = \cup_{S \in [L]_4} \text{BPT}(S) - \varphi_{g(\alpha, n)}(X[S])$. Finally, let $\varphi^*_\alpha(X) = (Q_\alpha)_{\text{res}}$ and $\varphi^{**}_\alpha(X) = \langle \delta(Q_\alpha) \rangle$, so that $\varphi^*_\alpha(X) \subseteq \varphi^{**}_\alpha(X)$.

8.15 Theorem

φ^*_α , and φ^{**}_α are CCI selection procedures, and for each $T \in \text{BPT}(n)$ it can be efficiently (in n) decided whether $T \in \varphi^*_\alpha(X), \varphi^{**}_\alpha(X)$.

Proof:

One can efficiently decide whether $T \in \varphi^*_\alpha(X), \varphi^{**}_\alpha(X)$, since

Q_α can be efficiently (in n) constructed and one can then efficiently decide whether or not $T \in (Q_\alpha)$ and $T \in \langle \delta(Q_\alpha) \rangle$.

Regarding CCI properties suppose $\alpha_1 \leq \alpha_2$, so that since φ is a CCI procedure, $\varphi_{g(\alpha_2, n)}(X[S]) \subseteq \varphi_{g(\alpha_1, n)}(X[S])$, hence $Q_{\alpha_1} \subseteq Q_{\alpha_2}$.

Now if $A \subseteq B$ then $(B)_{\text{res}} \subseteq (A)_{\text{res}}$ so that $\varphi^*_{\alpha_2}(X) \subseteq \varphi^*_{\alpha_1}(X)$, as required.

Now $T \in \varphi^*_\alpha(X)$ precisely if $T[S] \cap \text{BPT}(S) - \varphi_{g(\alpha, n)}(X[S]) = \emptyset$ for all $S \in [L]_4$.

Now $P(T[S] \cap \text{BPT}(S) - \varphi_{g(\alpha, n)}(X[S]) = \emptyset) \geq P(T[S] \notin \varphi_{g(\alpha, n)}(X[S])) \geq 1 - g(\alpha, n)$.

Thus by lemma 6.2 (1), $P(T \in \varphi^*_\alpha(X)) \geq 1 - \alpha$, as required.

Convergence follows from the observation that

$(\cup_{S \in [L]} \text{BPT}(S) - T[S])_{\text{res}} = \{T\}$. The argument for φ^{**} is similar. §

Efficiency (II):

We now consider the third question of how fast the number of sites must grow as a function of the taxa, so as to accurately reconstruct phylogenies.

8.16 Definition

Let φ be a selection procedure, and let $C \subseteq \cup_n \text{WBT}(n)$.

We say φ *efficiently recovers* C if:

(1): For each $T \in \text{BPT}(n)$ and partition frequencies X it can be efficiently decided whether $T \in \varphi(X)$.

(2): There exists a function $f: \mathbb{R}^+ \times \mathbb{N} \rightarrow \mathbb{R}^+$ polynomial in n , such that for all $\epsilon > 0$, and all $(T, p) \in C \cap \text{WBT}(n)$

$P(\varphi(X(c)) = \{T\}; T, p) > 1 - \epsilon$, for all $c \geq f(\epsilon, n)$.

8.17 Remark

The motivation for calling a selection procedure efficient under these conditions is as follows. Consider a biologist building phylogenies for progressively larger sets of species, in the hope that these match the unknown underlying tree. Suppose a selection procedure, φ , of order $O(n^{Sc^{\dagger}})$ which efficiently reconstructs the (unknown) class of trees being sampled is used, and suppose the associated function f is $O(n^u)$. Then for fixed ϵ , the number of steps required to build a tree in $\text{BPT}(n)$ which has probability $> 1 - \epsilon$ of being the tree which produced the data, does not grow exponentially in n (it is $O(n^{S+tu})$), and thus the accurate construction of large trees may be feasible.

Since $b(n+2)$ is asymptotically proportional to $n! 2^n \sqrt{n}$, it is perhaps surprising that any infinite set of binary trees can be efficiently recovered. Clearly, $\cup_n \text{WBT}(n)$ cannot be efficiently recovered, by theorem 8.7 (2), though trivially any finite subclass can be.

We now show that the class of balanced trees of height λ , $BH(\lambda)$, (definition 7.41) which contains all rooted trees (with suitable edge weights) can be efficiently recovered. For convenience we assume a molecular clock. That is we assume the edge lengths on the underlying tree are additive as in definition 6.4. The general case (without a clock) is essentially the same, though the calculations are messy, and tend to obscure the result. For the remainder of this section recall (from definition 6.4 (5)) that $s(T \mid q) = s(T, p)$, for $p_e = 0.5(1 - e^{-2q_e})$.

8.18 Definition

For any vector $x \in (\mathbb{R}^+)^4$, indexed over $\Omega(3)$ ordered as $\{1\}, \{1,2\}, \{1,3\}, \{1,2,3,4\}$ and the rooted tree $T = (1)(23)$, let $q = q(x) = [q_1, q_2, q_2 - q_1] \in [0, \infty]^3$ (where $q_e = \infty$ is formally taken to mean $p_e = 0.5$) denote the edge lengths on the edges of T , as illustrated in fig. 8.1(a), which minimize the Euclidean distance $\Delta(T, x, q) = \|s(T \mid q) - x\|$. Let $\Delta(T, x)$ denote this minimum distance.

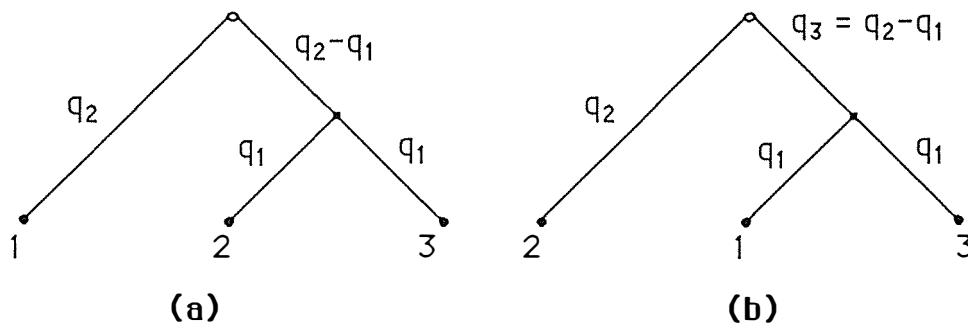


Figure 8.1

8.19 Lemma

In the notation of definition 8.18, suppose $1 \geq x_{123} \geq x_1, x_{12}, x_{13}$, $q=q(x)$ and let

$\omega = [\omega_1, \omega_2]$, where $\omega_i = e^{-2q_i}$, for $i=1,2$. Then

$$\omega = \begin{cases} [x_1 + x_{123} - x_{12} - x_{13}, x_{123} - x_1], & \text{if } x_1 \geq (x_{12} + x_{13})/2 \\ \beta[1,1], & \text{if } x_1 \leq (x_{12} + x_{13})/2, \text{ where } \beta = (3x_{123} - x_1 - x_{12} - x_{13})/3. \end{cases}$$

Proof:

A simple application of theorem 5.28 gives for $s = s(T | q)$

$$s_1 = (1 + \omega_1 - 2\omega_2)/4$$

$$s_{12} = s_{13} = (1 - \omega_1)/4 \quad (1)$$

$$s_{1234} = (1 + \omega_1 + 2\omega_2)/4.$$

Let $\Delta = \Delta^2(T, x, q)$. Then $\Delta = \sum_{\sigma \in \Omega(3)} (x_\sigma - s_\sigma)^2$ so that

$$\partial \Delta / \partial \omega_j = -2 \sum_{\sigma} (x_\sigma - s_\sigma) \partial s_\sigma / \partial \omega_j \text{ and for } j, k \in \{1, 2\},$$

$$\partial^2 \Delta / \partial \omega_j \partial \omega_k = 2 \sum_{\sigma} (\partial s_\sigma / \partial \omega_j) (\partial s_\sigma / \partial \omega_k), \text{ since } \partial^2 s_\sigma / \partial \omega_j \partial \omega_k = 0. \text{ Thus}$$

the Jacobian matrix $[\partial^2 \Delta / \partial \omega_j \partial \omega_k]_{jk}$ of Δ is diagonal with leading entries

$1/2, 1$, so that any critical points of Δ are minima.

Solving the system $\partial \Delta / \partial \omega_j = 0$ for $j=1,2$ gives the unique solution

$$\omega_1 = x_1 + x_{123} - x_{12} - x_{13}, \quad \omega_2 = x_{123} - x_1.$$

For this to be a feasible solution we require $\omega_1 \geq \omega_2 \geq 0$, that is

$1 \geq \omega_1 \geq \omega_2 \geq 0$. Since $1 \geq x_{123} \geq x_1$, by hypothesis, these conditions translate

into the given condition $x_1 \geq (x_{12} + x_{13})/2$. If this does not hold then (since

there is only one critical point for Δ), Δ is minimized subject to $q_1 = q_2$,

that is $\omega_1 = \omega_2 = \omega$ (say).

Solving the equation $\partial\Delta/\partial\omega = 0$ gives $\omega = (3x_{123} - x_1 - x_{12} - x_{13})/3$, which lies between 0 and 1 by hypothesis and so is feasible, as required. §

8.20 Lemma

For $T_1 = (3)(12)$, with edge lengths $q = [q_1, q_2, q_3]$, $q_3 = q_2 - q_1$ as in fig. 8.1(b), and $T = (1)(23)$, let $\Delta_1 = \Delta(T, s(T_1 | q))$.

Then $\Delta_1 = e^{-2q_1}(1 - e^{-2q_3})/\sqrt{6}$

Proof:

Let $y_i = e^{-2q_i}$ for $i=1,2$. Then for $s^* = s(T_1 | q)$ we have

$$s_1^* = s_{13}^* = (1 - y_1)/4$$

$$s_{12}^* = (1 + y_1 - 2y_2)/4,$$

$$s_{123}^* = (1 + y_1 + 2y_2)/4.$$

Applying lemma 8.19 with $x_\sigma = s_\sigma^*$ we find that $x_1 \leq (x_{12} + x_{13})/2$ since

$y_1 \geq y_2$, and $x_{123} \geq x_1, x_{12}, x_{13}$. Thus the value for q' which minimizes

$\Delta(T, s^*, q')$ is given (in terms of ω) by $\omega = \beta[1,1]$ where

$$\beta = (3x_{123} - x_1 - x_{12} - x_{13})/3 = (y_1 + 2y_2)/3. \text{ Substituting this value gives}$$

$$\Delta_1^2 = (y_1 - y_2)^2/6, \text{ as required. §}$$

Consider the following selection procedure, ϕ_1 . For each triple $S = \{i, j, k\}$ choose that rooted binary tree (or trees) T on S which minimizes $\Delta(T, X[S])$. If all of the trees so produced are consistent with exactly one rooted tree T_1 , then select T_1 , otherwise select Φ .

8.21 Theorem

For $T_0 \in \text{RBT}(n)$ with edge lengths q subject to a molecular clock, let δ be the minimum edge length on T_0 , and q the height of T_0 (i.e. the expected number of changes from the root to the endpoints). Then

$$P(\varphi_1(X) = \{T_0\}; T_0, q) \geq 1 - 9^n C_3 e^{4q} / 2C\delta^2(1-\delta)^2$$

Proof:

For each $S \in [n]_3$ consider the event $E(S)$ that $T = T_0[S]$ strictly minimizes $\Delta(T, X[S])$. Then the event that φ_1 selects T_0 is precisely the event

$$\cap_{S \in [n]_3} E(S).$$

Let $\epsilon(S) =$

$$\min\{\|s(T_0[S] \mid q[S]) - s(T_1 \mid q_1)\| : q_1 = [q_1, q_2, q_2 - q_1] \in [0, \infty]^3, T_1 \in \text{RBT}(S), T_1 \neq T_0[S]\}$$

By lemma 8.20, $\epsilon(S) \geq e^{-2q_1(1-e^{-2q_3})} / \sqrt{6}$ for appropriate values of

$q_1, q_2, q_3 = q_2 - q_1$ (depending on S). Thus $\epsilon(S) \geq 2e^{-2q\delta(1-\delta)} / \sqrt{6}$, since $q_1 \leq q$,

$(1-e^{-t}) \geq t^2/2$, (for $t > 0$) and $q_3 \geq \delta$, and this holds for all $S \in [n]_3$.

Let $E_1(S)$ be the event that $\|s(T_0[S] \mid q[S]) - X[S] / C\| < e^{-2q\delta(1-\delta)} / \sqrt{6}$.

Since $e^{-2q\delta(1-\delta)} / \sqrt{6} \leq \epsilon(S)/2$ we have, by the triangle inequality for

Euclidean distance, that $E_1(S) \subseteq E(S)$ so that $P(\cap_{S \in [n]_3} E(S)) \geq P(\cap_{S \in [n]_3} E_1(S))$,

and thus $P(\cap_{S \in [n]_3} E(S)) \geq 1 - \sum_{S \in [n]_3} (1 - P(E_1(S)))$, by lemma 6.2 (1). (*)

The complement of $E_1(S)$ is precisely the event that

$$\|s(T_0[S] \mid q[S]) - X[S] / C\|^2 > e^{-4q\delta^2(1-\delta)^2} / 6, \text{ so that by Markov's inequality}$$

$$(\text{lemma 6.2 (2)}), \quad 1 - P(E_1(S)) \leq (6e^{4q} / \delta^2(1-\delta)^2) \gamma$$

where $\gamma = E[\|s(T_0[S] \mid q[S]) - X[S] / C\|^2]$.

Writing $s(T_0[S] \mid q[S])$ as $[s_1, \dots, s_4]$, $X[S]$ as $[X_1, \dots, X_4]$ we have

$$\gamma = E[\sum_i (X_i - cs_i)^2 / c^2] = c^{-2} \sum_i E[(X_i - cs_i)^2] = c^{-2} \sum_i \text{Var}[X_i].$$

Now X has a multinomial distribution with parameters c and s so that

$\text{Var}[X_i] = cs_i(1-s_i)$, by lemma 6.2 (3a). Thus by proposition 6.5 ((2)

$$\gamma = (1 - \sum_i s_i^2) / c = (1 - s(\sigma_0; T[S] \mid 2q)) / c \leq 3/4c,$$

since $s(\sigma_0; T[S] \mid 2q) \geq 1/4$, by lemma 6.3 (2).

Thus $\gamma \leq 3/4c$. Combining this with (*), gives the result. §

8.22 Corollary

For any $\lambda \in \mathbb{R}^+$, ϕ_1 efficiently recovers $BH(\lambda)$ (in the sense of 8.16).

Proof:

For each $T \in \text{RBT}(n)$, it can certainly be efficiently decided whether $T \in \phi_1(X)$.

We must now construct a function f , as in definition 8.16. Let

$(T, q) \in BH(\lambda)$. Since the ratio of the edge lengths of any tree in

$BH(\lambda) \cap \text{WBT}(n)$ is bounded above by $n-1$, the shortest edge length on such a

tree is at least $\lambda / (n-1)$. By the previous theorem (taking $\delta = \lambda / n$, for

convenience, and $q = \lambda$) we have

$$P(\phi_1(X(c)) = \{T\}; T, q) \geq 1 - 9n^2 \cdot n C_3 \cdot e^{4\lambda} / 2c\lambda^2(1-\lambda/n)^2 \text{ so that we can take}$$

$f(\epsilon, n) = 9n^2 \cdot n C_3 \cdot e^{4\lambda} / 2\lambda\epsilon$ in definition 8.16, which is polynomial in n (of

degree 5), as required. §

8.23 Remarks

(1): Consider a family of trees $T(k) \in \text{RBT}(2^k)$, (definition 7.16) and assign all edges of $T(k)$ length λ/k . Then $\{(T(k), (\lambda/k)_j) : k=1,2,\dots\} \subset \text{BH}(\lambda)$, so that φ_1 efficiently recovers this class, yet the probability that there is a partition with more than one occurrence tends to zero, by theorem 6.19. This suggests the usefulness of aggregation-based approaches such as quartet (and in this case triplet) methods.

(2): The degree of n (namely 5) for the function f constructed in the previous corollary is due to the coarse nature of φ_1 . It is likely that functions of lower degree in n satisfying definition 8.16, could be constructed by using a more subtle procedure, which does not require the structure of all nC_3 triples to be known in order to select some tree. The value of corollary 8.22 lies in demonstrating that polynomial bounded procedures (in the sense of 8.16) exist even though $\text{BPT}(n)$ grows much faster than any polynomial function.

A χ^2 test: (Molecular clock)

We have seen that Markov's inequality gives a method for putting confidence intervals on selected trees by using Euclidean distance in s -space. One drawback with this approach is that the confidence intervals are not very tight unless c is large, as Markov's inequality exploits only the variance of a distribution. In general, if X has a multinomial distribution, with parameters c, λ , there is little one can say about the distribution of the Euclidean distance $\|X/c - \lambda\|^2$. However by lemma 6.2 (4d), a tight distribution exists (approximately) on a centralized variant, obtained by replacing $\sum_i (X_i/c - \lambda_i)^2$ by $\sum_i (X_i/c - \lambda_i)^2 (c/\lambda_i)$.

8.24 Definitions

For $S \in [n]_3$, $S = \{1, 2, 3\}$, $T = (1)(23)$, choose $q = [q_1, q_2, q_2 - q_1]^3$ as in fig. 8.1 (a)

so as to minimize:

$$\chi(T, S, X, q) = \sum_{\sigma \in \Omega(3)} (X[S](\sigma) - cs(\sigma; T | q))^2 / cs(\sigma; T | q).$$

Let $\chi(T, S, X)$ be this minimum value.

Let χ_3 be a chi-square random variable with three degrees of freedom, and for $\alpha > 0$ choose $\beta(\alpha)$ so that $\text{let } P(\chi_3 > \beta(\alpha)) < \alpha / n_{C_3}$.

Given a set Q of rooted binary trees on sets in $[n]_3$ define the *restriction* of Q , $(Q)_{res}$ in the analogous way as for unrooted binary trees trees on sets in $[n]_4$ (definition 8.14). Thus

$$(Q)_{res} = \{T \in \text{RBT}(n) : T[S] \cap Q = \emptyset \text{ for all } S \in [n]_3\}$$

(where for $S = \{i, j, k\}$, $T[S] = (i)(j, k)$, $(j)(i, k)$ or $(k)(i, j)$).

Finally let $Q^1_\alpha = \cup_{S \in [n]_3} \{T \in \text{BPT}(S) : \chi(T, S, X, q) > \beta(\alpha)\}$ and define

$$\varphi_\alpha(X) = (Q^1_\alpha)_{res}. \text{ Thus } \varphi_\alpha(X) \subseteq \text{RBT}(n).$$

8.25 Lemma

φ_α has convergent confidence intervals.

Proof:

A straightforward argument similar to theorem 8.15. §

8.26 Theorem

Write $[s_1, s_{12}, s_{13}, s_{123}]$ as $[s_1, s_2, s_3, s_4]$ and $X[S]$ as $[x_1, x_2, x_3, x_4]$. Then the s_i values that give $\chi(T, S, X)$ are given as follows: Let $\gamma = \sqrt{(x_2^2 + x_3^2)/2}$.

Case 1: If $x_1 \geq \gamma$, $x_4 \geq x_1$, then

$$s_1 = x_1 / (x_1 + x_4 + 2\gamma)$$

$$s_4 = x_4 s_1 / x_1$$

$$s_2 = s_3 = (1 - s_1 - s_4) / 2.$$

Case 2: If $x_1 \leq \gamma$, $x_4 \geq x_1$

$$s_1 = s_2 = s_3 = \beta / (1 + 3\beta) \text{ where } \beta = x_4^{-1} \sqrt{(x_1^2 + x_2^2 + x_3^2)/3}$$

$$s_4 = 1 - 3s_1.$$

Case 3: If $x_4 \leq x_1 \leq \gamma$

$$s_1 = s_4 = a_1 / 2(a_1 + a_2)$$

$$s_2 = s_3 = a_2 / 2(a_1 + a_2); \text{ where } a_1 = \sqrt{(x_1^2 + x_4^2)}, a_2 = \sqrt{(x_2^2 + x_3^2)}$$

Case 4: If $x_4 \leq x_1 \leq \gamma$

$$s_1 = s_2 = s_3 = s_4 = 1/4.$$

(Cases other than 1 arise because an edge length required to minimize $\chi(T, S, X, q)$ is negative and this is disallowed. Not surprisingly, in these cases, the resulting $\chi(T, S, X, q)$ values are large -- the trees are the wrong way around. This is illustrated in example 8.27.)

Proof:

As in lemma 8.19 we have

$$\begin{aligned} s_1 &= (1+\omega_1-2\omega_2)/4 \\ s_{12} &= s_{13} = (1-\omega_1)/4 \\ s_{1234} &= (1+\omega_1+2\omega_2)/4. \end{aligned} \quad (1)$$

We wish to minimize $\chi = \sum_i (x_i^2/c^2 s_i - 2x_i + cs_i)$ subject to $0 < \omega_2 \leq \omega_1 < 1$.

These conditions imply $1 \geq s_4 \geq s_1 \geq s_2 = s_3 \geq 0$.

We first solve $\partial\chi/\partial\omega_j = 0$ subject to $s > 0$.

$$\begin{aligned} \partial\chi/\partial\omega_j &= \sum_i (c - x_i^2/c^2 s_i^2) \partial s_i / \partial \omega_j \\ &= -c^{-2} \sum_i (x_i^2/s_i^2) \partial s_i / \partial \omega_j, \text{ since } \sum_i \partial s_i / \partial \omega_j = 0. \end{aligned}$$

$$\text{Now } \partial\chi/\partial\omega_2 = 0 \text{ precisely if } x_1^2/s_1^2 - x_4^2/s_4^2 = 0 \quad (2)$$

$$\text{and } \partial\chi/\partial\omega_1 = 0 \text{ precisely if } x_1^2/s_1^2 - x_2^2/s_2^2 - x_3^2/s_3^2 + x_4^2/s_4^2 = 0. \quad (3)$$

Now if the s_i 's are positive, (1) implies

$$x_1 s_4 = x_4 s_1 \quad (4)$$

Furthermore, since $s_2 = s_3$, (1) and (2) give:

$$x_1 s_2 = s_1 \sqrt{(x_2^2 + x_3^2)/2} \quad (5)$$

Finally, by (1), $s_2 = (1 - s_1 - s_4)/2$ so that letting $\gamma = \sqrt{(x_2^2 + x_3^2)/2}$, (4) and

$$(5) \text{ give, } s_1 = x_1 / (x_1 + x_4 + 2\gamma).$$

Now for $j, k \in \{1, 2\}$, $\partial^2\chi/\partial\omega_j\partial\omega_k = c^{-2} \sum_i (x_i^2/s_i^3) (\partial s_i/\partial\omega_j) (\partial s_i/\partial\omega_k)$,

since $\partial^2 s_i / \partial\omega_j \partial\omega_k = 0$.

Thus, letting $A_i = (x_i^2/s_i^3)/16c^2$, the Jacobian matrix for χ is

$$J = \begin{bmatrix} A_1 + A_2 + A_3 + A_4 & -2A_1 + 2A_4 \\ -2A_1 + 2A_4 & 4A_1 + 4A_4 \end{bmatrix}$$

Now for $s > 0$, the A_i 's are positive so that $\partial^2 \chi / \partial \omega_1^2 > 0$ and $\det J > 0$, so that J is positive definite, and thus the unique critical point $s^* = [s_1^*, s_2^*, s_3^*, s_4^*]$ of χ in the positive s -quadrant corresponds to a minimum of χ .

Now $\omega_1 = 1 - 4s_2$, $\omega_2 = 1 - 2s_1 - 2s_2$, and we require $0 \leq \omega_2 \leq \omega_1 < 1$. For s^* we have $\omega_1 < 1$, and $\omega_2 \leq \omega_1$ precisely if $s_1^* \geq s_2^*$ which holds if and only if $x_1 \geq x_4$, while $0 \leq \omega_2$ precisely if $x_4 \geq x_1$. This gives the result for case 1.

For case 2, if $x_1 < x_4$ and $x_4 \geq x_1$, then for s^* , $\omega_1 > \omega_2 \geq 0$. Since s^* is the only critical point of χ in the positive s -quadrant, χ is maximized in case 2 (subject to $\omega_1 \leq \omega_2$) on the plane $\omega_2 = \omega_1$. In this case $s_1 = s_2 = s_3 = (1 - \omega_1)/4$ and $s_4 = (1 + 3\omega_1)/4$.

Then $\partial \chi / \partial \omega_1 = 16c^{-2}(x_1^2 + x_2^2 + x_3^2)/(1 - \omega_1)^2 - 3x_4^2/(1 - 3\omega_1)^2$

which equals zero when $s_1 = s_2 = s_3 = \beta/(1 + 3\beta)$, where

$\beta = x_4^{-1} \sqrt{(x_1^2 + x_2^2 + x_3^2)/3}$ and $s_4 = 1 - 3s_1$. Clearly these s values are all positive and $\partial^2 \chi / \partial^2 \omega_1 > 0$ in the positive s -quadrant, so that the critical point corresponds to a minimum, as required.

In case 3, if $x_1 \geq x_4$ and $x_4 < x_1$, a similar argument applies -- in this case χ is minimized subject to $\omega_2 = 0$. In case 4, χ is minimized subject to the conditions of cases 2 and 3, but these imply $\omega_1 = \omega_2 = 0$, which confine the set of feasible s values to one point.

Example 8.27

For the EMBL data X on four taxa, listed in example 8.13, let us construct $\phi_{0.005}(X)$ by the above procedure.

We require $P(\chi_3 > \beta) < 0.00125$, so that we can take $\beta = 16.7$, (Burington and May [1970, p.387]). Then we have the following table:

S	T_1	$\chi(T, X[S])$	T_2, T_3	$\chi(T_i, X[S])$ ($i = 2, 3$)
$\{1, 2, 3\}$	$(1)(23)$	2.5	$(2)(13), (3)(12)$	115.9
$\{1, 2, 4\}$	$(1)(24)$	1.7	$(2)(14), (4)(12)$	105.2
$\{1, 3, 4\}$	$(1)(34)$	0.4	$(3)(14), (4)(13)$	336.2
$\{2, 3, 4\}$	$(2)(34)$	1.2	$(3)(24), (4)(23)$	117.5

Thus $\phi_{0.005}(X) = (Q)_{res}$ where Q is the set of all 8 trees in the fourth column of the above table. Thus $\phi_{0.005}(X) = \{T\}$ where T is the tree shown in fig. 8.2.

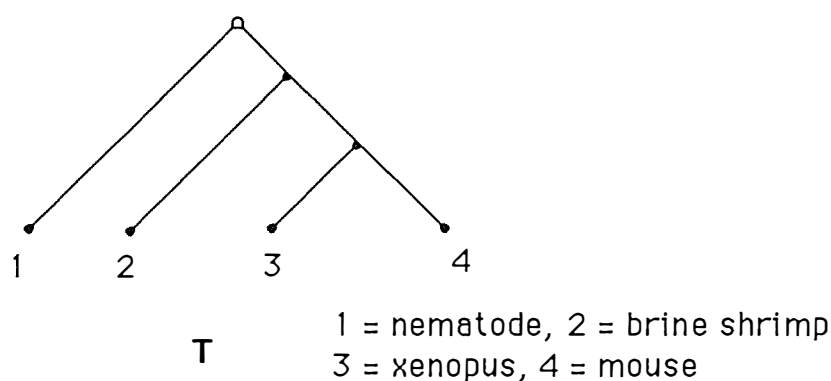


Figure 8.2

8.28 Summary

The theme of this section has been the construction of confidence intervals on phylogenies by constructing such intervals on subsets of the taxa of size three or four. These smaller cases often have simple analytic properties (theorem 8.12 and 8.26) and the confidence intervals generated can then be combined by lemma 6.2 (1) to give confidence intervals on the parent phylogeny. This lemma is sufficient to prove the existence of procedures with desirable theoretical properties (theorem 8.15, lemma 8.25 and corollary 8.22) and its appeal lies in the absence of any independence assumptions regarding the events involved. A more subtle approach would be to consider the dependence between these events arising from the constraints imposed on edge lengths of subtrees by the requirement that they must all fit together on some parent tree in their span. More specifically, although a binary tree with n endpoints is defined by $n-3$ subtrees of size four (theorem 3.12), and each of these subtrees has 5 edge lengths, the resulting $5(n-3)$ edge lengths are subject to a number of linear constraints since the parent tree has precisely $2n-3$ edge lengths (linear combinations of which give the $5(n-3)$ edge lengths).

Appendix

Proof of theorem 7.18

Let $g(k)$ be the minimum number of endpoints of $T(k)$ that are coloured A over all $\chi \in V(T(k), \{A, B\})$. We prove the following two statements simultaneously by induction on k :

$$(1): g(k) = f(k-1)$$

$$(2): f(k) = f(k-1) + f(k-2); f(0) = 1, f(1) = 2.$$

The results hold for $k \leq 2$ so suppose both results are true for $k \leq m-1$, $m \geq 3$.

Firstly let $\chi_1 \in V(T(m), \{A, B\})$ realize $g(m)$.

By Fitch's algorithm (refer 5.18 (2)), the two trees $T(m-1)$ whose roots are joined to the root of $T(m)$ must have colourings χ_1, χ_2 both in $V(T(m-1), \{A, B\})$ or with one in $V(T(m-1), \{A\})$ and the other in $V(T(m-1), \{B\})$.

In the first case, $g(m) \geq 2g(m-1)$. By the induction hypothesis on (1), we have $g(m-1) = f(m-2)$, and by the induction hypothesis on (2) we have that $f(m-1)$ is the m -th Fibonacci number, so that $f(m-1) < 2f(m-2)$. Combining these we have for the first case that $g(m) > f(m-1)$.

In the second case, we can suppose χ_1 realizes $f(m-1)$ and that χ_2 assigns colour B to every pendant vertex so that $g(m) = f(m-1)$.

Thus the second case achieves a lower value for $g(m)$, and hence establishes the induction step for (1) that $g(m) = f(m-1)$.

We now establish the induction step for (2). Let $\chi_1 \in V(T(m-1), \{A\})$ realize $f(m-1)$, $\chi_2 \in V(T(m-2), \{A\})$ realize $f(m-2)$. Assign colour B to all the pendant vertices of one copy of $T(m-2)$, colour another copy of $T(m-2)$ by χ_2 , join their roots to a third root to obtain a colouring for $T(m-1)$ in $V(T(m-1), \{A, B\})$.

Then join this tree to another copy of $T(m-1)$ coloured by χ_1 , and again join their roots to obtain a colouring χ of $T(m)$, which by Fitch's algorithm is in $V(T(m),\{A\})$.

In this way, $f(m) \leq f(m-1) + f(m-2)$. (*)

Conversely, let χ realize $V(T(m),\{A\})$. By Fitch's algorithm, the induced colourings χ_1, χ_2 of the two copies of $T(m-1)$ joined to the root of $T(m)$ are either both in $V(T(m-1),\{A\})$ or one is in $V(T(m-1),\{A\})$ and the other is in $V(T(m-1),\{A,B\})$.

In the first case, $f(m) \geq 2f(m-1)$, which combined with the previous inequality, (*) gives $f(m-1) < f(m-2)$, a contradiction, since by the inductive hypothesis on (2), $f(m-1)$ and $f(m-2)$ are respectively the m -th and $(m-1)$ -th Fibonacci numbers, so that $f(m-1) > f(m-2)$.

In the second case we have:

$f(m) = f(m-1) + g(m-1)$, and since $g(m-1) = f(m-2)$, by the inductive hypothesis on (1), the result follows. §

		a								
		2	3	4	5	6	7	8	9	10
n	4	1.67								
	5	1.80								
	6	1.86	2.31							
	7	1.89	2.52							
	8	1.91	2.64	2.97						
	9	1.92	2.71	3.22						
	10	1.93	2.75	3.37	3.63					
	11	1.94	2.79	3.48	3.90					
	12	1.95	2.81	3.55	4.09	4.29				
	13	1.95	2.83	3.61	4.22	4.57				
	14	1.96	2.85	3.65	4.32	4.78	4.96			
	15	1.96	2.86	3.69	4.40	4.94	5.25			
	16	1.96	2.88	3.72	4.46	5.07	5.47	5.62		
	17	1.97	2.89	3.74	4.51	5.17	5.65	5.92		
	18	1.97	2.89	3.76	4.56	5.25	5.80	6.16	6.29	
	19	1.97	2.90	3.78	4.59	5.31	5.91	6.35	6.59	
	20	1.97	2.91	3.79	4.62	5.37	6.01	6.51	6.84	6.95

Values of $\psi(n,a)$ for $4 \leq n \leq 20$.

$\psi(n,1) = 1$, $\psi(n,a) = 0$, $2a > n$.

Table Two (Refer p.93)

REFERENCES

- AIGNER, M. (1979)
Combinatorial Theory. Springer-Verlag New York, 483p.
- ANDERSON, I. (1974)
A first Course in Combinatorial Mathematics. Oxford: Clarendon Press, 122p.
- BANDELT, H-J. and DRESS, A. (1986)
Reconstrucing the shape of a tree from observed dissimilarity data.
Advances in Appl. Math. 7 (3) : 309-343.
- BENDER, E.A. (1974)
Asymptotic Methods in Enumeration. SIAM. Review. 16 (4) : 485-515.
- BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975)
Discrete Multivariate Analysis: Theory and Practice. MIT. MIT Press, 557p.
- BOURQUE, M. (1978)
Arbres de Steiner et reseaux dont varie l'emplacement de certains sommets. Thesis, Ph.D., Department d'Informatique et de Recherche Operationnelle, Universite de Montreal, Quebec, Canada.
- BURINGTON, R.S and MAY, D.C. (Jr), (1970)
Handbook of Probability and Statistics with Tables. 2nd ed. McGraw-Hill Inc. 462p.
- CARTER, M., HENDY, M.D., PENNY, D., SZEKELY, L.A., WORMALD, N.C. (1988)
On the Distribution of lengths of evolutionary trees. (manuscript).
- CAVALLI-SFORZA, L. L. and EDWARDS, A.W.F. (1969)
Phylogenetic Analysis: models and estimation procedures. Evolution. 21: 550-570.
- CAVENDER, J.A. (1978)
Taxonomy with confidence. Math Biosci. 40 : 271-280.
- ____ (1981)
Tests of phylogenetic alternatives under generalized models.
Math Biosci. 54 : 217-229.

- ____ and FELSENSTEIN, J. (1987)
Invariants of Phylogenies in a simple case with discrete states.
J. Classification 4 : 57-71.
- COLONIUS, H. and SCHULZE, H.H. (1981)
Tree structure for proximity data. British Journal of Mathematical and Statistical Psychology 34 : 167-180.
- CONSTANTINESCU, M. and SANKOFF, D. (1986)
Tree enumeration modulo a consensus. J. Classification 3 : 349-356
- DAY, W.H.E. (1983)
Distribution of distances between pairs of classifications. p.127-131
 In Numerical Taxonomy. Edited by Felsenstein, J. Springer-Verlag
 Berlin Heidelberg, 1983
- ____ (1985)
Optimal Algorithms for Comparing Trees with labeled leaves. Journal of Classification 2 : 7-28.
- ____ (1986)
Analysis of quartet dissimilarity measures between undirected phylogenetic trees. Syst. Zool. 35 (3) : 325-333.
- ____ (1987)
Computational Complexity of Inferring Phylogenies From Dissimilarity Matrices. Bulletin of Math. Biol. 49 (4) : 461-467.
- ____, JOHNSON, D. S. and SANKOFF, D. (1986)
The Computational Complexity of Inferring Rooted Phylogenies by Parsimony. Math Biosci. 81 : 33-44.
- JOHNSON, N.L. and KOTZ, S. (1969)
Discrete Distributions. Boston. Houghton Mifflin Company. 328p.
- ____ and SANKOFF, D. (1986)
Computational complexity Of Inferring Phylogenies by Compatibility.
Syst. Zool. 35 (2) : 224-229.
- DIAMOND, J.M. (1988)
DNA-based phylogenies of the three chimpanzees. Nature 332 : 685-686.
- ECK, R.V. and DAYHOFF, M.O. (1966)
Atlas of Protein Sequence and Structure National Biomedical Research Foundation, Silver Spring, Maryland.

EDWARDS, A.W.F. (1972)

Likelihood. Cambridge: Cambridge University Press. 235p.

ERDOS, P. and SZEKERES, G. (1935)

A Combinatorial Problem in Geometry. Composito Math. 2 : 464-470.

ESTABROOK, G.F., McMORRIS, F.R. and MEACHAM, C.A. (1985)

Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool. 34 (2) : 193-200.

FARRIS, J.S. (1973)

A Probability Model for Inferring Evolutionary Trees. Syst. Zool. 22: 250-256.

FELSENSTEIN, J. (1973)

Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters.

Syst. Zool. 22 : 240-249

____ (1978)

The number of evolutionary trees. Syst. Zool. 27 : 27-33.

____ (1978)

Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401-410.

____ (1983)

Methods for Inferring Phylogenies: A Statistical View.

p. 315-334. In Numerical Taxonomy. Edited by Felsenstein, J. Springer-Verlag Berlin Heidelberg, 1983.

____ (1985)

Confidence limits on phylogenies with a molecular clock.

Syst. Zool. 34 : 152-161.

____ (1987)

Estimation of Hominoid Phylogeny from a DNA Hybridization Data Set.

J. Mol. Evol. 26 : 123-131.

____ (1988)

Phylogenies from molecular sequences: Inference and Reliability.

Annual Review of Genetics 22, (in press).

FARRIS, J.S. (1973)

A probability model for inferring evolutionary trees. Syst. Zool. 22 :250-256.

- FITCH, W.M. (1971)
Towards defining the course of evolution: Minimal change for a specific tree topology. Syst. Zool. 20 :406-416.
- FOULDS, L.R. and GRAHAM, R.L. (1982)
The Steiner problem in phylogeny is NP-complete.
Adv. Appl. Math. (3) : 43-49.
- FOULDS, L.R. and ROBINSON, R.W. (1984)
Enumeration of Phylogenetic Trees without points of degree two.
Ars Combinatoria 17A : 169-183.
- GORDON, A.D. (1986)
Consensus Supertrees: The Synthesis of Rooted Trees Containing Overlapping Sets of Labeled Leaves. J. Classification 3 : 335-348.
- GOULDEN, I.P. and JACKSON, D.M. (1983)
Combinatorial Enumeration. John Wiley and Sons, Inc., 569p.
- GRAHAM, R.L. and FOULDS, L.R. (1982)
Unlikelihood that Minimal Phylogenies for a realistic Biological Study Can Be Constructed in Reasonable Computational Time.
Math Biosci. 60 : 133-142.
- HALL, M.H. (Jr). (1986)
Combinatorial theory. 2nd ed. John Wiley & Sons. 440p.
- HAMM, G.H. and CAMERON, G.N. (1986)
The EMBL data library. Nucleic Acids Res. 14 : 5-9.
- HARARY, F. (1969)
Graph Theory. Addison-Wesley Publishing Company. 274 pages.
- HARARY, F. and MOWSHOWITZ, A. (1975)
Enumeration of end-labeled trees. Fibonacci Quart. 13 no. 3, 252-254.
- HARDING, E. F. (1971)
The probabilities of rooted tree-shapes generated by random bifurcation.
Adv. Appl. Prob. 3 : 44-77.
- HARTIGAN, J.A. (1973)
Minimum mutation fits to a given tree. Biometrics 29 : 53-65.

HENDY, M.D. (1988)

The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. (in press).

____ (1989)

A Combinatorial Description of the Closest Tree Algorithm for Finding Evolutionary Trees. SIAM. J. Discrete Maths. (in press).

____ and PENNY, D. (1982)

Branch and Bound Algorithms to Determine Minimal Evolutionary Trees. Math. Biosci. 59 : 277-290.

____ and _____. (1988)

A framework for the quantitative study of evolutionary trees. Syst. Zool. (in press).

____, LITTLE, C.H.C. and PENNY, D. (1984)

Comparing trees with pendant vertices labelled. SIAM. J. Appl. Math. 44 :1054-1065.

____, STEEL, M.A., PENNY, D. and HENDERSON, I.M. (1988)

Families of Trees and Consensus. p.355-362. In Classification and Related Methods of Data Analysis (Edited by BOCK, H.H.) Elsevier Science Publishers B.V. (North-Holland).

HERSTEIN, I.M. (1975)

Topics in algebra. Xerox Corp. 388p.

JOHNSON, N.L. and KOTZ, S. (1969)

Discrete distributions. Houghton Mifflin Company, Boston.

MARCZEWSKI, E. and STEINHAUS, H. (1958)

On a Certain Distance of Sets and the Corresponding Distance of Functions. Colloquium Mathematica 6 : 319-327.

MEACHAM, C.A. (1983)

Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. p.304-314. In Numerical Taxonomy Edited by Felsenstein, J. Proceedings of a NATO Advanced Study Institute. NATO advanced Study Inst. Series G (Ecological Sciences), no.1 Springer-Verlag, New York.

MEIR, A., MOON J.W. and MYCIELSKI, J. (1983)

Hereditarily finite sets and identity trees. J. Comb. Theory B. 35 : 142-155.

- MOON, J. W. (1970)
Counting Labelled Trees (University of Alberta). Canadian Mathematical monographs no.1.
- MOORE, G.W. GOODMAN, M. and BARNABAS, J. (1973)
An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets.
J. theor. Biol. 38 : 423.
- OTTER, R. (1948)
The number of trees Annals of Mathematics 49 : 583-599.
- PALMER, E. M. and SCWENK, A.J. (1979)
On the Number of Trees in a Random Forest.
J. Comb. Theory B. 27 : 109-121.
- PENNY, D. (1982)
Towards a Basis for Classification: the Incompleteness of Distance Measures, Incompatibility Analysis and Phenetic Classification.
J. theor. biol. 96 : 129-142.
- ____, FOULDS, L.R. and HENDY, M.D. (1982)
Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences.
Nature 297 : 197-200.
- ____ and HENDY, M.D. (1985)
The use of tree comparison metrics. Syst. Zool. 34 (1) : 75-82.
- RESTLE, F. (1959)
A Metric and an Ordering on Sets. Psychometrika 24 : 207-220.
- ROBINSON, D. F. (1971)
Comparison of labelled trees with valency three.
J. Comb. Theory 11 : 105-119.
- ____ and FOULDS, L.R. (1979)
Comparison of weighted labelled trees. p.119-126. In Lecture notes in Mathematics. Vol. 748. Springer-Verlag, Berlin.
- ____ and ____ (1981)
Comparison of phylogenetic trees. Math Biosci. 53 : 131-147.
- SMITH, T.F. and WATERMAN, M.S. (1980)
How alike are two trees? Am. Math. Monthly 87 : 552-553.

SOKAL, R.R. and ROHLF, F.J. (1981)

Taxonomic congruence in the Leptopodomorpha re-examined.
Syst. Zool. 30 : 309-325.

STEEL. M.A. (1988)

Distribution of the Symmetric Difference Metric on Phylogenetic Trees.
SIAM J. Discrete Maths. 1 (4) : 541-551.

STEEL. M.A., HENDY, M.D. and PENNY, D. (1988)

Loss of information in genetic distances. Nature 336 (6195) : 118.

WATERMAN, M.S. and SMITH, T.F. (1978)

On the Similarity of Dendograms. J.theor. Biol. 73 : 789-800.

ZUCKERKANDL, E. and PAULING, L. (1962)

Molecular disease, evolution, and genetic heterogeneity. p.189-225. In
Horizons in Biochemistry, edited by M. Marsha and B. Pullman New York:
Academic Press. (189-225).