Molecular Ecology (1998) **7**, 399–412

To tree or not to tree

PETER E. SMOUSE

Center for Theoretical and Applied Genetics, and Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08903–0231, USA

Abstract

The practice of tracking geographical divergence along a phylogenetic tree has added an evolutionary perspective to biogeographic analysis within single species. In spite of the popularity of phylogeography, there is an emerging problem. Recurrent mutation and recombination both create homoplasy, multiple evolutionary occurrences of the same character that are identical in state but not identical by descent. Homoplasic molecular data are phylogenetically ambiguous. Converting homoplasic molecular data into a tree represents an extrapolation, and there can be myriad candidate trees among which to choose. Derivative biogeographic analyses of 'the tree' are analyses of that extrapolation, and the results depend on the tree chosen. I explore the informational aspects of converting a multicharacter data set into a phylogenetic tree, and then explore what happens when that tree is used for population analysis. Three conclusions follow: (i) some trees are better than others; good trees are true to the data, whereas bad trees are not; (ii) for biogeographic analysis, we should use only good trees, which yield the same biogeographic inference as the phenetic data, but little more; and (iii) the reliable biogeographic inference is inherent in the phenetic data, not the trees.

Keywords: homoplasy, phylogeography, population structure, spanning trees

Introduction

The molecular revolution has made it feasible to attack problems of population genetics and systematics with a vast array of variable genetic markers. While a single marker can yield idiosyncratic results, a large collection of independent markers should exhibit a coherent pattern, perhaps best viewed as a molecular manifestation of the Central Limit Theorem. We often have more genetic markers than we have individuals, however, and we are in a poor position to make any assumptions about so many markers. For any substantial number of markers, the standard assumption of independent segregation is difficult to accept, and we are certainly in no position to test it, given the limited number of individuals at our disposal. Rather than assume independent segregation for all the markers, we assume that the mutations that have given rise to the different states of each character were (evolutionarily) independent and sequential. We can then relate multicharacter genotypes to each other, either in terms of the numbers of substitutional differences that separate them (phenetic distances) or in terms of their pathways of evolutionary divergence (patristic distances).

The comparison of multicharacter genotypes over geographical space, using phenetic methods, is a form of biogeographic analysis. Traditional population genetics and systematics both have voluminous literatures on the subject. The practice of tracking biogeographic divergence along a phylogenetic tree, an enterprise for which Avise *et al*. (1987) have coined the term phylogeography, has added an overtly evolutionary perspective to biogeographic studies. The practice of intraspecific phylogeography has spread so rapidly (e.g. Moritz *et al*. 1987; Avise 1989; Merriwether *et al*. 1991; Quattro *et al*. 1991; Vigilant *et al*. 1991; Excoffier *et al*. 1992; Maddison *et al*. 1992; Excoffier & Smouse 1994; Joseph & Moritz 1995; Crandall & Templeton 1996) that we find ourselves celebrating its 10th anniversary with a special issue of *Molecular Ecology*. At the risk of casting a pall over the celebration, I must point out that something is amiss. In many cases, the genetic data on which the enterprise depends are phylogenetically ambiguous, and that has implications for the biogeographic analyses that follow. There are two obvious sources of such ambiguity.

Correspondence: P. E. Smouse. Tel.: +01-732-932-1064; Fax: +01-732-932-8746; E-mail Smouse@AESOP.Rutgers.EDU

(i) Recurrent mutation. In our never-ending quest for highly informative (highly polymorphic) markers within a single species, we are (almost by default) selecting genetic markers with recurrent mutation rates high enough to yield multiple mutations over the phylogenetic timescale of interest. The net consequence is a large amount of mutational homoplasy in the raw data, multiple evolutionary substitutions that are identical in state but not identical by descent (Templeton 1983; Hudson 1989; Excoffier & Smouse 1994; Bandelt *et al*. 1995). We have always known that allozyme alleles could be homoplasic, but the same is obviously true for RFLP markers, single nucleotide sites, RAPD markers, minisatellites, and microsatellites; homoplasy is almost inherent in the exercise. Homoplasy is a serious problem with many molecular data sets, and it leads to phylogenetic ambiguity. Many of our molecular data sets are less than compellingly phyletic.

(ii) Recombination. Within a single species, even closely linked genes will become decoupled over evolutionary time, and multicharacter genetic profiles cannot be related to each other in a strictly phyletic fashion. In a sense, all character states become homoplasic. The phylogenetic tree derived from all the data (species tree) will be some average that does not conform well to any particular gene tree. Many authors have made the obvious distinction between gene trees and a species trees, but what we should do about the duality, other than acknowledge it, remains unclear.

The usual strategy is to concentrate on one or a few loci, deriving a separate gene tree for each; this strategy has the virtue of avoiding the complication of recombination among them. To the extent that the data are convincingly tree-like, the tree will tell the phylogenetic story for the gene, but even for closely linked markers within a single gene, intragenic recombination can occur often enough (over evolutionary time) to confound a strictly phyletic interpretation. Strobeck & Morgan (1978) and Morgan & Strobeck (1979) showed that if intragenic recombination occurs no more often than point mutation within the gene, surely a plausible assumption, then a very substantial fraction of the allelic variation encountered will be due to recombination. As an unavoidable consequence, individual molecular characters can become highly homoplasic. A striking example is provided by the β-globin region in humans (Long *et al*. 1990), the evolution of which has as much to do with recombination as it does phylesis. To avoid recombinational complications, many have turned to nonrecombining organellar genomes that are uniparentally inherited, mtDNA or cpDNA (Cann *et al*. 1987; Birky 1988; Harris & Ingram 1991; Ward *et al*. 1991; Soltis *et al*. 1992; Kolman *et al*. 1995; Mason-Gamer *et al*. 1995). If the mutation rate is high enough and the time frame is large enough, mutational homoplasy will remain a problem, and many molecular character sets are chosen precisely because the mutation rate is high enough to generate substantial variation.

The more homoplasic ambiguity we encounter in the data, the larger the degree of phylogenetic uncertainty we have. Notwithstanding this difficulty, the usual practice is to use some convenient tree-making algorithm to choose a tree, and then to proceed with biogeographic analysis. Where there exists a considerable range of alternatives from among which to choose, the choice itself can become an issue. The 'African Eve' exchange (Maddison 1991; Vigilant *et al*. 1991; Hedges *et al*. 1992; Templeton 1992) is a case in point.

There are data sets which are compellingly tree-like and, for them, the choice of tree is easy, and the resulting biogeographic interpretation is entirely plausible. Many of our data sets are not credibly tree-like, however, and to the extent that a tree is not inherent in the raw data, there will be an element of extrapolation involved in the resulting phylogeny. To the same extent that phylogenetic extrapolation is uncertain, any derivative population structure or biogeographic analyses are also uncertain. Attention has recently been devoted to minimizing the impact of that phylogenetic ambiguity (Excoffier *et al*. 1992; Crandall & Templeton 1993; Crandall *et al*. 1994; Excoffier & Smouse 1994; Bandelt *et al*. 1995), but there are important questions that need attention. Is there any necessity to impose a phylogenetic tree for population analysis? Is there anything to be gained by doing so? Basically, we need to decide whether to tree or not to tree.

The objectives of this study are to explore the informational aspects of converting a multicharacter data set into a phylogenetic tree, and then to explore what happens when that tree is used for population analysis. In the process of exploration, using a simple example, I will address two proximal questions: (i) how much phylogenetic ambiguity is inherent in the molecular data set, and how can we choose among multiple candidate trees; and (ii) how does our choice of tree impact on subsequent biogeographic analyses? Buttressing the example with results from the literature, I will argue that while some trees are harmless, in biogeographic context, others are problematic. Inasmuch as we can extract the reliable biogeographic inference from the phenetic data themselves, I will argue that we might be better served by using tree-free methods for biogeographic analysis.

Turning phenetic data into a tree

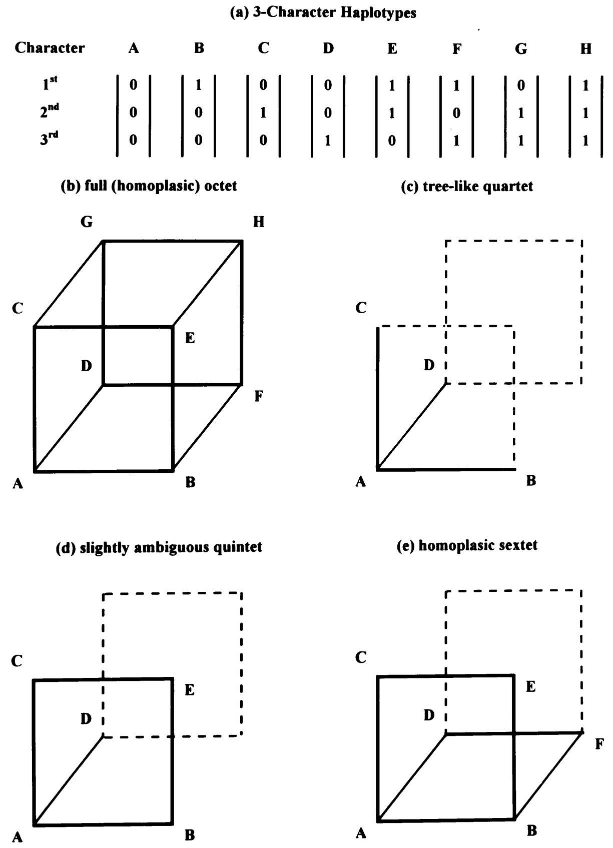
Homoplasy and the hypercube

Our problem is best understood if we have a clear sense of what we are actually doing when we make a tree. What follows is applicable to almost any type of multicharacter data set, but I will concentrate on uniparentally inherited haploid genomes, mitochondria or chloroplasts. For organelle genomes, any particular combination of multiple character states is inherited as an indivisible unit, called a haplotype. In the absence of recombination, the primary problem becomes one of mutational homoplasy. I will concentrate on haplotypic marker sets, because recombining markers can only make the situation worse. My comments thus pertain to a best-case analysis but, as we shall see, the best may not be very good.

It is also convenient to illustrate with binary characters, because they are sufficient to show the nature and scope of the problem. Multistate characters can also be accommodated, but they require additional mathematical elaboration, and they change neither the essential questions nor the answers. So, imagine a string of M binary characters, each coded (0, 1). We might be talking about a set of restriction sites, each coded as either present (1) or absent (0), a set of deletions/insertions/inversions, each coded as present (1) or absent (0), or even a sequence of polymorphic nucleotides that, by virtue of strong transitional bias, exhibit alternative states A and G (or C and T), each site coded arbitrarily as (0, 1).

For each individual in the study, we have a binary vector of length M (a string of 1s and 0s), representing the M polymorphic character states for that individual. The different states of that vector (haplotypes) could be treated simply as distinguishable alleles (as with allozyme alleles), without regard to their levels of divergence, because they are inherited as indivisible units. While that is a perfectly legitimate treatment, it sometimes makes poor use of our hard-won molecular data (cf. Epifanio *et al.* 1995). If we use the degree of divergence among the haplotypes, we can extract additional (potentially useful) information (Excoffier *et al.* 1992).

We begin by defining a measure of phenetic divergence (distance) between any two haplotypes. There are many different distance metrics in common usage, but it is convenient to formulate the problem with a Manhattan (city-block) metric. To set the notation and begin the illustration, imagine a set of M = 3 binary characters, denoted the 1st, 2nd and 3rd (Fig. 1a). With M = 3 characters, there are $2^3 = 8$ possible haplotypes, denoted by a set of 3 vectors (v_i) = (A = 000, B = 100, C = 010, D = 001, E = 110, F = 101, G = 011, H = 111), as in Fig. 1a, and collectively occupying all the vertices of a three-dimensional cube (Fig. 1b). The Manhattan distance between adjacent haplotypes (e.g. A–B, B–E, G–H) is 1; that between haplotypes at opposite corners in the same plane (e.g. A–E, B–D, F–G) is 2; and that between haplotypes at opposite vertices of the three-dimensional cube (e.g. A–H, B–G, C–F) is 3. In a perfectly tree-like (strictly phyletic) data set, we would have one more haplotype (L) than we have characters (M), and there would be no gaps and no closed loops (i.e. L = M + 1). We prefer the sort of data set represented in Fig. 1c which, although small, is phylogenetically unambiguous. Unfortunately, we all too often encounter data sets more similar to those represented in Fig. 1d,e. Many data sets have portions of the character space that even show the pattern in Fig. 1b. The illustration is a mere three-dimensional version of a much larger problem. With multiple (M > 4) characters, we can always place the L haplotypes at the vertices of an M-dimensional analogue of our cube (a hypercube). Multicharacter data sets frequently exhibit substantial homoplasy, many closed (sometimes interlocking) loops, and nontrivial phylogenetic ambiguity.


(a) 3-Character Haplotypes
(b) full (homoplasic) octet
G
H
C
D
E
F
A
B
(c) tree-like quartet
C
D
A
B
(d) slightly ambiguous quintet
C
D
E
A
B
(e) homoplasic sextet
C
D
E
F
A
B

Character	A	B	C	D	E	F	G	H
1st	0	1	0	0	1	1	0	1
2nd	0	0	1	0	1	0	1	1
3rd	0	0	0	1	0	1	1	1

Fig. 1 A homoplasic octet of three-character, binary haplotypes, and their representation as the vertices of a three-dimensional unit cube. (a) vector representation of the eight haplotypes; (b) cubic representation of the full octet; (c) a tree-like set of four haplotypes; (d) an ambiguous set of five haplotypes; (e) an ambiguous set of six haplotypes; solid lines represent single-step connecting edges between observed haplotypes.

Message 1. At the heart of our problem is the fact that many of the available data sets are homoplasic for the most informative (most polymorphic) portion of the character set, and phylogenetic ambiguity is necessarily severe.

How many characters are enough?

Keeping in mind that the entire exercise is being pursued in the interests of population structure and/or biogeographic analysis within a single species, we will have a survey of *P* populations, and we will have sampled *N* individuals. As we anticipate variation within populations, we should have $N > P$. I have commented elsewhere on the need for adequate population subsampling (Smouse *et al*. 1991; Smouse & Chevillon 1998). These *N* individuals will exhibit L differentiated haplotypes, occupying L of the 2^M vertices of our M-dimensional hypercube. The natural tendency in molecular systematics is to use a large battery of characters, so that each individual is genetically unique, but how many characters do we need to characterize *N* individuals? Is it really necessary to have many more molecular characters than individuals?

There are three cases of interest. Case (i) is that of a perfectly tree-like data set (L = M + 1); there are no closed loops and no missing intermediates. An example would be a data set consisting of haplotypes A, B, C and D, as in Fig. 1c. There are three single-step substitutional changes needed to connect those haplotypes (A–B, A–C and A–D), and each involves a separate character. The ideal outcome is a 'perfectly tree-like data set', because there is then only one acceptable candidate. Using that tree for subsequent population structure or biogeographic analyses is utterly unobjectionable; there is a very real sense in which the data are the tree. It is unfortunate that we seldom encounter this case in molecular phylogenetic practice.

Case (ii) is that where we have more characters than we need for the number of haplotypes recovered (L < M + 1). Some of the characters are redundant, and there will be gaps in the connecting network. An example would be a data set consisting of haplotypes A = (000) and F = (101) in Fig. 1a. The 2nd character is monomorphic within the data set, but the 1st and 3rd characters are polymorphic; they are also perfectly correlated, being either both 0 or both 1. What this implies is that either the two characters have experienced simultaneous evolutionary substitutions, possibly indicating a recombination event, or that a mutationally intermediate haplotype is absent from the data set. That missing haplotype may now be extinct, or we may just have missed it in our (usually less than exhaustive) sampling. The missing haplotype might be either B or D, and although there is some ambiguity of pathway, we can say that the path between A and F requires two steps. In the absence of other haplotypes that might confuse the issue, whether we invoke B or D does not affect the tree or subsequent analysis. This case occurs most often when the number of characters is quite large, relative to the number of individuals sampled, i.e. when $M > N$. This is a common problem, for example, with DNA sequence data from the hypervariable region of the mitochondrial D-Loop. Finding the optimal tree can be a challenge, because the hypercube is sparsely occupied, and the critical information on the links between the observed haplotypes is contained in a large number of missing haplotypes. As long as the proper links are unambiguous, there is no problem, but difficulty arises wherever there are myriad alternative connections that change the nature of the tree. The number of alternative trees can be huge, and there is sometimes not much to choose from among very divergent candidates. The dictates of good practice suggest that we should avoid informational overkill by using a smaller number of cladistically informative characters from the outset, but we have limited ability to choose among characters a priori, on the basis of their probable cladistic behaviour. The alternative is somehow to remove the homoplasy *post hoc*, and Wills (1996) has recently attempted this, in connection with the 'African Eve' story. How well a *post hoc* strategy will perform in routine practice is a matter that needs further exploration.

Case (iii) is that where there are more haplotypes than can be accounted for by unique evolutionary substitutions (L > M + 1), frequently encountered with RFLP haplotypes. An example would be a data set consisting of haplotypes A, B, C, D and E from Fig. 1d. The A–D link is obvious enough, but there is a closed loop (A–B–E–C–A) that implies two substitutions (homoplasy) for either the 1st or 2nd character. We can devise four trees, very different, but all equally good. A more complicated example is provided by the haplotypic array in Fig. 1e, where there are two homoplasies and two closed loops. The array in Fig. 1b has five homoplasies. For more characters and haplotypes, the problem can become overwhelming. An obvious finesse is to increase the number of characters, but the empirical reality is that the homoplasy already present in a data set of M characters cannot decrease. My own experience has been that an expanded character set will generally contain additional homoplasies. We convert a homoplasic Case (iii) situation into an even more homoplasic Case (ii) situation.

The point of phylogeographic analysis is to use an unambiguous tree to describe how the individuals are connected, and the molecular characters represent evolutionary replication. On the other hand, if we are forced to use the operational taxonomic units (OTUs) to convert the character arrays into a tree, because the data are phylogenetically ambiguous, the individuals become the replication, and we cannot credibly characterize more characters than we have individuals. Moreover, there is an element of circularity involved in using the OTUs to order character changes into a tree, and then using the tree to order the OTUs. We really should not be doing both.

Message 2. Using a vast number of characters to make a tree for small numbers of individuals is informational overkill. For biogeographic analyses, we would do better to characterize more individuals for fewer characters.

Converting hypercubic data into a tree

By virtue of employing orthogonal axes and a Manhattan metric, distances between haplotypes obey the Pythagorean Theorem, and we conveniently define the squared phenetic distance between the *i*th and *j*th haplotypes (υ_i and υ_j) as

$$(\delta^2_{ij})_P = (\upsilon_i - \upsilon_j)\,(\upsilon_i - \upsilon_j) = \Delta_{ij} \qquad (1)$$

where Δ_{ij} is the number of (0 vs. 1) site-by-site differences between the two haplotypes. Our distance metric $(\delta^2_{ij})_p$ is thus both Manhattan and Euclidean, which is analytically ideal for what follows. For example, haplotypes A and B in Fig. 1a differ only for the 1st character, so $(\delta^2_{AB})_P = 1$; similarly, haplotypes C and H differ for both the 1st and 3rd characters, so $(\delta^2_{CH})_P = 2$, etc. It is convenient to represent the total collection of squared distances between all possible pairs of L haplotypes in the form of an L × L phenetic distance matrix, henceforth denoted $\boldsymbol{D}_p = (\delta^2_{ij})_p$. The matrix $\boldsymbol{D}_p$ makes no assumptions about treeness of the data, being a simple tally of the differences between pairs of haplotypes. The data are explicitly embodied in $\boldsymbol{D}_p$, and we should always remember that $\boldsymbol{D}_p$ is only as tree-like as are the raw data.

It is not absolutely essential to use a Manhattan or a Euclidean metric, and many other choices are available, but the choice of metric is not the issue here. Once a metric (describing the degree of divergence) has been chosen to represent the divergence between haplotypes, the pairwise interhaplotypic distances contain all the information available. Whether we use the raw data or a phenetic distance matrix derived from those data, a derivative tree represents an imposition of additional assumptions on the data. To the extent that 'the tree' is an extrapolation beyond the data, analysis of the tree is unavoidably analysis of those extraneous assumptions.

So how do we make/choose a tree? For purposes of exposition, I will concentrate on the six-haplotype data set in Fig. 1e, which is homoplasic enough to illustrate, yet simple enough to permit exhaustive examination. We have six haplotypes and seven connecting (single-step) edges. To make a tree, we have to cut both the (A–B–E–C–A) and (A–B–F–D–A) loops. To do that, we must remove two edges, while maintaining connectedness of all six haplotypes. There are precisely 15 ways to do that, each yielding a different tree (Fig. 2). These trees are represented as straight-line segments, portrayed in two dimensions, but it is important to remember that the trees are actually wending their way through five-dimensional space in city-block fashion, with each step at right angles to all steps that come before and all steps that follow. By careful construction, the interhaplotype distances, measured along any particular tree (patristic distances) are both Manhattan and Euclidean, facilitating subsequent comparison of the trees with the data and with each other.

The trees listed in Fig. 2 are minimum spanning trees (MSTs). Spanning trees have OTUs (haplotypes) as both nodes and branch tips. Steiner trees, with which most of us are more familiar, have OTUs only as branch tips, with the nodes representing (now) extinct ancestral OTUs. For interspecific work, the ancestral intermediates are commonly absent from the data set, and Steiner trees capture the essence of the situation. For intraspecific work, the intermediates are often still present, and our trees resemble strawberry plants, spreading across the landscape as an articulating network; MSTs have their attractions. I will use MSTs, because they are sufficient to illustrate the point, and because the critical messages do not depend on the particular tree-building algorithm in any event. It is possible to compute the exact number of

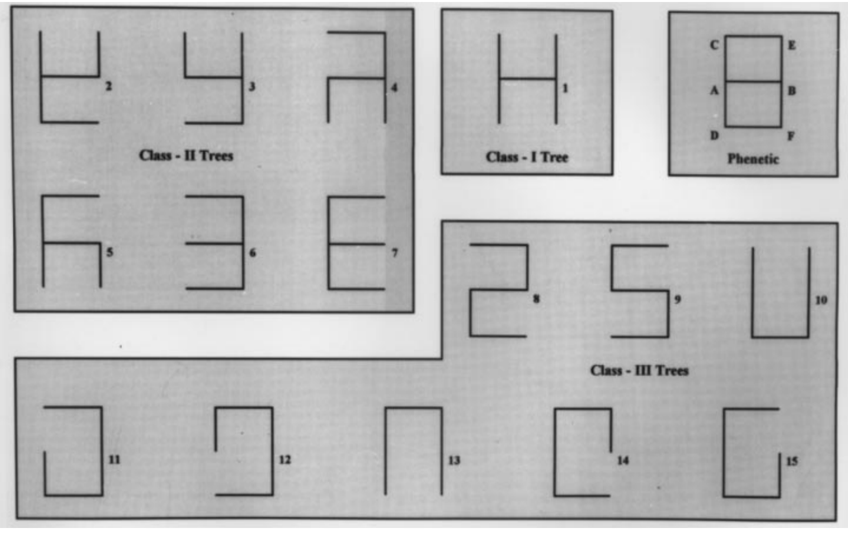


Fig. 2 The 15 minimum spanning trees (MSTs) that can be extracted from the sextet of haplotypes in Fig. 1d, each depicted in two-dimensional form; the connected sextet from which the trees are derived is shown in the upper-right corner; the trees are broken into classes, based on topological similarity.

competing MSTs from such a molecular data set, and (in principle) to delineate them, using Kirchoff methods (Prim 1957; Gibbons 1985). Recall that the haplotypes shown in Fig. 1c yield only one MST, while those shown in Fig. 1d yield four, and those shown in Fig. 1e yield 15. Those shown in Fig. 1b yield 384 MSTs; the greater the degree of homoplasy, the greater the number of MSTs. If the degree of homoplasy is substantial (L > M), the number of MSTs can become vast. As an empirical illustration, Excoffier *et al*. (1992) reported L = 56 haplotypes and M = 34 RFLP sites in a collection of human mtDNA genotypes. There were in excess of 10^9 (equally) maximum parsimony MSTs. Those trees were topologically diverse and, at least in terms of total tree length, there was nothing to choose among them. I shall have more to say about choosing among equally parsimonious trees later.

A homoplasy is created wherever the same character has changed more than once, yielding parallel edges for our M-dimensional hypercube. Breaking a loop is tantamount to transforming one such character, mutating twice, into a pair of independent characters, changing once each. That amounts to adding orthogonal (right-angle) axes to our character space, one per homoplasy. Making a tree is thus tantamount to imposing additional dimensions on the data. If there are m = L – (M + 1) homoplasies in the data set, we have to add m dimensions to our character set. Those extra dimensions were a real part of the evolutionary process, but their identities have long since become obscured. All we have are the phenetic data. Any tree we construct represents an extrapolation from the data. Derivative analyses are, by default, analyses of that extrapolation. There are multiple sets of m loops we can cut, sometimes a huge number, and subsequent analyses depend on the tree chosen. If we insist on analysing a tree, we cannot escape the questions of which extrapolations we choose to analyse, and why we have chosen that set?

Message 3. Converting the phenetic data into a tree represents an extrapolation from the data; alternative trees are alternative extrapolations. Analysis of any tree is an analysis of the corresponding extrapolation.

Choosing among equally parsimonious trees

Setting aside, for the moment, the question of whether we should even be using a tree for subsequent analysis, it has to be said that some of these equal-length trees are better than others. Return to the illustration (Fig. 2). Each of the possible trees yields a derivative patristic distance matrix. A tree and its patristic distance matrix represent a 1:1 mapping. Each of these matrices has M dimensions, where M is the number of dimensions in our augmented character space (M = M + m), defined by breaking the m loops. A patristic distance matrix defines and is defined by its tree; the augmented data are perfectly (and artificially) tree-like. By inserting an additional orthogonal axis for each broken loop, we preserve the Manhattan/Euclidean geometry of our patristic distance metric during tree construction. Three example MSTs (one from each topological class) and their matching patristic distance matrices ($\boldsymbol{D_1}$, $\boldsymbol{D_3}$ and $\boldsymbol{D_{10}}$) are presented in Table 1, for contrast with the phenetic distance matrix, $\boldsymbol{D_p}$.

As total tree length (L – 1) is not a discriminating criterion among MSTs, we must invoke a 2nd criterion if we wish to narrow the search any further. It is possible to compare competing MSTs by means of the sum of patristic distances between each of the L (L – 1)/2 pairs of haplotypes, divided by the total number of those haplotypes

$$Q_t = \sum_{i=1}^{\mathrm{L}-1} \sum_{j>i}^{\mathrm{L}} (\delta^2_{ij})_t/\mathrm{L} \tag{2}$$

where the subscripts i and j index the different haplotypes. Basically, we sum the elements of the patristic distance matrix, below the diagonal, and divide by the total number of haplotypes (L = 6 in this case). While all MSTs are of the same total length (L – 1), they vary considerably in their Q_t values.

It develops that Q_t is also the sum of squared deviations of the OTUs from the centroid of the M-dimensional augmented character space containing the tth tree, which will be useful when we return to population structure analysis. If we divide Q_t by its degrees of freedom (L – 1), we convert the sum of squared deviations into a corresponding variance. To be specific,

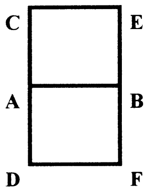
$$\mathrm{Var}(\boldsymbol{D_t}) = Q_t \div (\mathrm{L} - 1) \tag{3}$$

Excoffier & Smouse (1994) have suggested ranking trees on the basis of Var($\boldsymbol{D_t}$) or (equivalently) Q_t and choosing that tree with the minimum value, an optimization strategy they labelled molecular variance parsimony. The sum of patristic distances between all pairs of OTUs serves as a powerful 2nd level sorting criterion.

For the trees in Table 1, Q_1 = 4.833, Q_3 = 5.167 and Q_{10} = 5.833. Note that the sum of squares for the phenetic distance matrix, representing the raw data, is Q_p = 4.167. The patristic distance between the ith and jth haplotypes, measured along any particular tree, cannot be shorter than the phenetic distance between them, because broken loops lengthen some of the patristic distances. Hence $(\delta^2_{ij})_P \le (\delta^2_{ij})_t$ for all pairs of haplotypes (i and j), and it follows that $Q_p \le Q_t$ for all trees (t = 1, …, T). The MST which least increases the variation among haplotypes is that tree which represents minimal extrapolation from the raw data. As a practical matter, the trees with smallest Q_t are topologically the most tightly packed (most articulated).

Table 1 The phenetic data of Fig. 2, and an example trio of minimum spanning trees that can be extracted from those same data, their patristic distance matrices, their sums of squares, Q$_t$, and their cophenetic correlations with the phenetic distance matrix

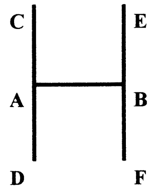
Phenetic Data



$Q_p = 4.167$

A	B	C	D	E	F	
0	1	1	1	2	2	A
1	0	2	2	1	1	B
1	2	0	2	1	3	C
1	2	2	0	3	1	D
2	1	1	3	0	2	E
2	1	3	1	2	0	F

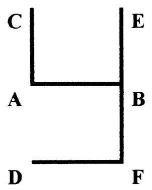
Tree 1



$Q_1 = 4.833$ $r_{p,1} = 0.577$

A	B	C	D	E	F	
0	1	1	1	2	2	A
1	0	2	2	1	1	B
1	2	0	2	3	3	C
1	2	2	0	3	3	D
2	1	3	3	0	2	E
2	1	3	3	2	0	F

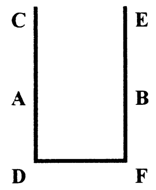
Tree 3



$Q_3 = 5.167$ $r_{p,3} = 0.547$

A	B	C	D	E	F	
0	1	1	3	2	2	A
1	0	2	2	1	1	B
1	2	0	4	3	3	C
3	2	4	0	3	1	D
2	1	3	3	0	2	E
2	1	3	1	2	0	F

Tree 10



$Q_{10} = 5.833$ $r_{p,10} = 0.357$

A	B	C	D	E	F	
0	3	1	1	4	2	A
3	0	4	2	1	1	B
1	4	0	2	5	3	C
1	2	2	0	3	1	D
4	1	5	3	0	2	E
2	1	3	1	2	0	F

It is also possible to compare the details of each tree ($t = 1, \ldots, T$) with the raw data, in point by point fashion, by computing the cophenetic correlation (r_{pt}) between $\boldsymbol{D}_p$ and each of the $\boldsymbol{D}_t$ values (Sokal & Rohlf 1962; Sokal *et al.* 1986; Barrantes *et al.* 1990; Smouse & Long 1992). The computed values for the three example trees are presented in Table 1. Excoffier & Smouse (1994) have shown that trees with small Q_t values are highly correlated with the data, whereas trees with larger Q_t values are generally less so. All else being equal, trees with small Q_t are preferable, and among a set of equally parsimonious MSTs. In fact, we can use the cophenetic correlation (r_{pt}) as a gauge of the 'inherent treeness' of the data set. None of the example trees in Table 1 is credibly tree-like, because $r_{pt} < 0.58$ for the best of them, a consequence of the inherent homoplasy in the example data set. A practitioner who finds a tree with $r_{pt} > 0.90$ or 0.95 could be reasonably confident that the data were reliably tree-like (nonhomoplasic). Any tree with $r_{pt} < 0.90$ should probably be viewed with a degree of healthy skepticism.

There is no guarantee that we have the 'true tree', even if Q_t is almost as small as Q_p and r_{pt} is close to unity. A close fit of model (tree) to data should not be interpreted as phylogenetic truth. On the other hand, phylogenetic truth is generally unknowable, and a close fit to the data is about the best we can do; most of us will accept it. Other optimality criteria could be used, of course, and other tree-making algorithms employ them, but the details of alternative optimization algorithms are not the issue here. The central idea with all of the available algorithms is to choose that tree which is as closest to the data, given the constraints imposed by the algorithm/model.

Message 4. We want a tree that is as true to the data as we can manage. Thus, good trees are those that mimic the phenetic data closely, whereas bad trees are those that do not conform closely to the data.

Frequency weighting – a 3rd level criterion

The use of Q_t provides us with a telling 2nd level criterion for choosing among MST topologies, but any given topological type can be represented by different trees. To illustrate, our simple example (Fig. 2) exhibits three different topological classes: (a) Tree 1, (b) Trees 2–7, and (c) Trees 8–15. Within a topological class, all trees have the same Q_t value. It would be useful to have a 3rd level sorting criterion to distinguish among topological ties. Excoffier *et al*. (1992), Crandall & Templeton (1993), Crandall *et al*. (1994), and Excoffier & Smouse (1994) have all argued that the frequencies of different haplotypes provide useful information about tree structure. The basic tenet, traceable to Watterson & Guess (1977), is that (under neutrality), the most frequent haplotypes in the population are probably the oldest, everything else being equal. High-frequency haplotypes have probably been present for a long time, having had ample time to achieve substantial copy numbers. New variants are probably derived from the more common haplotypes and, because they are new, will be rarer. It is possible to imagine a number of scenarios for which this pattern will not hold (e.g. Slatkin & Hudson 1991), but as a first approximation, the two assertions imply that the most frequent haplotypes should cluster relatively close together, near the centroid of the phylogenetic space, whereas the rarer variants should occupy the outer fringes of the tree. Those expectations favour a frequency-weighted Q-criterion, as even those trees sharing similar topologies will have different distributions of mass.

For a sample of N individuals in the study, exhibiting L different haplotypes, we have n_i replicates of the ith haplotype. We can either increase the dimension of the distance matrices to $N \times N$, reporting the genetic distance between each pair of individuals, or (equivalently) we can leave the (L × L) interhaplotypic distance matrices as is, but weight the $(\delta^2_{ij})_p$ and $(\delta^2_{ij})_t$ by $(n_i\ n_j)$, yielding a weighted version of Q, denoted

$$\Theta_p = \sum_{i=1}^{L-1} \sum_{j>i}^{L} n_i n_j (\delta^2_{ij})_p / N \tag{4a}$$

for the phenetic distance matrix, $\mathbf{D}_p$, and by

$$\Theta_t = \sum_{i=1}^{L-1} \sum_{j>i}^{L} n_i n_j (\delta^2_{ij})_t / N \tag{4b}$$

for the tth patristic distance matrix, $\boldsymbol{D}_t$. In other words, we compute the sum of the elements below the diagonal of the $N \times N$ interindividual distance matrix, and divide by N. Any MST that connects the more common haplotypes has an Θ_t advantage. By virtue of that same frequency weighting, the placement of rare haplotypes has relatively little impact on Θ_t. Good trees generally have the common haplotypes near the centre of mass, where they define the central branching order, with the rare haplotypes being relegated to the periphery.

Again to illustrate, imagine that we had sampled $N = 100$ individuals, which had replicate numbers for the six haplotypes of $n_A = 40$, $n_B = 30$, $n_C = 15$, $n_D = 2$, $n_E = 3$ and $n_F = 10$. Each of the trees in Fig. 2 has a different Θ_t value, listed as Θ_t (total) in Table 2. It develops that most of the Class III trees (trees 10–15) disconnect haplotypes A and B, and have severely inflated Θ_t values as a consequence. The Class II trees are better, in general, but those candidates that disconnect either haplotypes A and C or haplotypes B and F (trees 2, 4, 6 and 7) have modestly elevated Θ_t values. Tree 1 remains the best, with tree 5 and tree 3 close behind; even tree 9, from Class III, is a credible contender, because it preserves the most critical links (C–A–B–F). All other trees are poorer contenders, to one degree or another.

Hudson *et al*. (1992) have shown, at least under some circumstances, that using haplotype frequency information can provide more power for detecting population subdivision. It is important to note here that we have assumed something nontrivial about the evolutionary process in order to fine-tune our choice of tree, and the refinement is only as reliable as that assumption. Frequency should thus only be used as a 3rd level criterion, but it does narrow the range of tree choices dramatically, and it certainly helps us distinguish between different members of the same topological class. It is also completely compatible with the population and biogeographic analyses that motivate the larger enterprise, and to which I will turn next. If we are to use frequency weighting, however, it is absolutely imperative that we refrain from sampling individuals differentially, based on their genotypes. We need a well-balanced and unbiased sample of the variation that is present in our P populations.

Table 2 AMOVA for the 15 candidate trees of Fig. 2, and a comparison with the phenetic data, when haplotypic replication is unequal ($n_A = 40$, $n_B = 30$, $n_C = 15$, $n_D = 2$, $n_E = 3$ and $n_F = 10$)

Distance	Weighted Θ criteria			Variances		Φ_{st}
matrix	Θ (total)	Θ (within)	Θ (among)	s^2_w	s^2_p	value
Phenetic	49.83	40.45	9.38	0.426	0.096	0.184
Class I Tree						
Tree – 1	51.13	42.95	8.18	0.452	0.080	0.150
Class II Trees						
Tree – 2	57.33	50.15	7.18	0.528	0.063	0.107
Tree – 3	52.93	42.95	9.98	0.452	0.102	0.184
Tree – 4	62.83	55.35	7.48	0.583	0.064	0.099
Tree – 5	52.63	42.05	10.58	0.443	0.110	0.199
Tree – 6	64.03	54.95	9.08	0.578	0.085	0.128
Tree – 7	58.23	48.95	9.28	0.515	0.090	0.149
Class III Trees						
Tree – 8	72.03	65.05	6.98	0.685	0.053	0.072
Tree – 9	54.55	42.05	12.50	0.443	0.134	0.233
Tree – 10	87.03	67.65	19.38	0.712	0.207	0.225
Tree – 11	100.83	85.05	15.78	0.895	0.152	0.146
Tree – 12	85.63	61.15	24.48	0.644	0.274	0.298
Tree – 13	83.83	63.15	20.68	0.665	0.225	0.253
Tree – 14	87.63	72.35	15.28	0.762	0.153	0.167
Tree – 15	87.03	62.55	24.48	0.658	0.273	0.293

Before leaving the question of how to make and compare trees, one final comment is in order. There are many optimization criteria other than total tree length (L), the sum of pairwise distances between haplotypes (Q_t), or even the weighted sum of pairwise distances (Θ_t), and my choices above are not intended to be proscriptive. Each of the standard tree-making algorithms in common usage has its own particular criterion, and each algorithm makes assumptions about the nature of the evolutionary process, assumptions that we can check only poorly. In practice, we construct the best tree we can from the data, using whatever criterion is explicit (or implicit) in our favourite tree-making algorithm, and we then proceed with the population structure or biogeographic analyses of interest. Any inference we choose to draw is contingent on the assumptions we have made. The important point is not so much the choice of that criterion as it is the preference we have for a tree that mimics the data as closely as possible, given the constraints.

Message 5. If we are going to make a tree from phenetic data, especially if we are going to use it for subsequent population or biogeographic analysis, we want a tree that is as 'true to the data' as we can make it.

Analysing biogeography

Phenetic or patristic distances?

At this juncture, it is timely to remind ourselves that our initial motivation for making a tree was to facilitate a geographical analysis of the pattern of variation within the species, couched in overtly evolutionary terms. Having discovered that there are competing trees we can use, and having chosen one, we move on to biogeographic analysis. There are many multivariate methods we can use, in pursuit of biogeographic pattern, but the choice of method is not the issue here. Suffice that virtually all of these analyses can be accessed either from the raw character data on the N individuals, or equivalently from an $N \times N$ distance matrix among those N individuals. The question of larger interest is whether we should analyse the phenetic data themselves, embedded in the phenetic distance matrix ($\boldsymbol{D}_p$), or whether we should analyse the tree, represented by a patristic distance matrix ($\boldsymbol{D}_t$)? I will pursue that question in the context of population structure analysis, with which my colleagues and I have some experience, but the answer is fairly general.

Population structure analysis

The first task is to determine the degree to which populations have different haplotypes or different frequencies of shared haplotypes, a problem conveniently pursued with the use of a molecular analysis of variance (AMOVA). AMOVA, which is a generic analogue of analysis of variance, begins with an $N \times N$ interindividual distance matrix, particularly convenient for molecular data

(Excoffier *et al*. 1992). AMOVA has been used for inversion polymorphisms (Etges *et al*. 1998), for multiallelic codominant allozyme data (Peakall *et al*. 1995), for multilocus dominant/recessive RAPD data (Huff *et al*. 1993), for haplotypic RFLP data (Excoffier *et al*. 1992), and it can easily be extended to microsatellites (Michalakis & Excoffier 1996) and multistate sequence data. It can be used for just about any type of genetic data imaginable. The basic idea is to partition the variation among individuals, Θ_t (total), into separate components for variation within and divergence among populations. We can easily extend the model to multiple nesting levels, should we have several layers of subdivision (e.g. Excoffier *et al*. 1992; Huff *et al*. 1993; Epifanio *et al*. 1995; Peakall *et al*. 1995). We can even use cross-classified analysis where individuals can be stratified in more than one way (Brown *et al*. 1996).

For our immediate purposes, it is sufficient to compute the sums of squares of interindividual distances, Θ_t (total), for a collection of N individuals in P populations, and partition it into within-population and among-population components. Using the classic ANOVA-type extraction procedures, we estimate the corresponding variance components, and compute the fraction of the total variance accounted for by divergence among populations, defined as

$$\Phi_{st} = \frac{s^2{}_p}{s^2{}_w + s^2{}_p} \tag{5}$$

a haplotypic analogue of the F statistic used to describe population structure for allozyme loci (Weir & Cockerham 1984). We can treat the haplotypes as simply different, as for multiallelic allozymes (Peakall *et al*. 1995), or we can use phenetic distances (e.g. Huff *et al*. 1993; Epifanio *et al*. 1995; Brown *et al*. 1996). We can even use the patristic distances for any specified tree (Excoffier *et al*. 1992; Excoffier & Smouse 1994). I will use AMOVA to compare the population structure analysis derivable from the phenetic data with that obtained by using each of the phylogenetic trees.

This is best illustrated with an extension of our contrived example. Earlier, I listed a set of unequal replicate numbers for the six haplotypes. Imagine that the $N = 100$ individuals were actually sampled from $P = 5$ populations (20 individuals each), with the observed haplotypic replicate numbers shown in Fig. 3. The five sampled populations are drawn from two ill-defined subspecies, populations (a) and (b) from subspecies I, populations (d) and (e) from subspecies II, and population (c) from a bridging intermediate type. Our suspicion is that there has been enough gene flow between the two subspecies, through the intermediate type, to blur the edges, and we hope to use molecular data to shed some evolutionary light on the taxonomic situation.

For the phenetic distance matrix, and for each of the 15 separate patristic distance matrices, we conducted a singly nested AMOVA; the results are tallied in Table 2. As pointed out earlier, Θ_t (total) – which ignores the geographical pattern, favours tree 1, but trees 5 and 3 are also strong contenders, as is tree 9. The Θ_t (within)-values yield a slightly different order, with trees 5 and 9 both better than trees 1 and 3. A close examination of Fig. 3 will show that all four trees share the C–A–B–F pathway, both

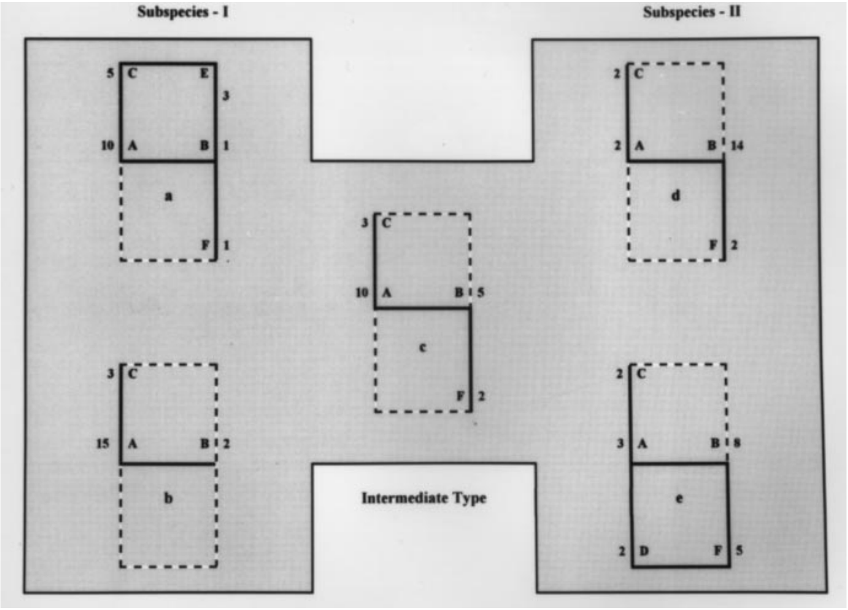


Fig. 3 Schematic representation of the population distribution of the example sextet of haplotypes in Fig. 1e, with sample haplotype replicate numbers attached; populations (a) and (b) are found in subspecies I; populations (d) and (e) are found in subspecies II; population (c) represents an intermediate type.

within populations and among them, but that trees 5 and 9 provide slightly clearer connections to the peripheral haplotypes D and E. On the other hand, there is nothing much to choose among these trees, and the phenetic data tell the essential story. The Φ_{st} criteria are erratic, changing substantially with small changes in the tree. Tree 9 had the largest value ($\Phi_{st} = 0.233$) and tree 1 the smallest ($\Phi_{st} = 0.150$), among the four trees, with the phenetic data in between ($\Phi_{st} = 0.184$). We should probably not use Φ_{st} to choose a tree (Excoffier & Smouse 1994), and even the results with Θ_t (total) and Θ_t (within) do not yield a compelling choice. None of these good trees creates a problem with the population structure analysis, but we might as well just let the data speak for themselves! What the data say is that the two subspecies are significantly divergent, with subspecies I favouring the A–C–E–B portion of the haplotype space, and subspecies II favouring the A–B–F–D portion of the haplotype space, although having nontrivial frequencies of haplotype C as well. The major difference is the frequency split between haplotypes A and B. The intermediate type seems to be a blend, possibly caused by gene flow, with haplotypes C, A, B and F being the only ones recovered. In terms of a possible tree, only the C–A–B–F path is compatible with all the evidence, and it is compelling. The precise connections of haplotypes E and D are difficult to discern, and the four alternative choices represent the alternative trees 1, 3, 5 and 9.

Message 6. For biogeographic analysis: (i) good trees are harmless, but they do not actually help much; (ii) bad trees mimic the data poorly and should be avoided; and (iii) the most reliable inference is that present in the phenetic data themselves.

Support from the literature

These results are entirely in accord with those obtained by Excoffier & Smouse (1994). Working with human mtDNA from $N = 672$ people from $P = 10$ populations, nested within five regional groups, a set of five restriction enzymes uncovered M = 34 RFLPs, defining L = 56 distinguishable haplotypes. As I mentioned earlier, there were in excess of 10^9 MSTs, all of the same minimal length. It was not possible to examine all 10^9 competing MSTs, but we did examine a random sample of 10 000 with AMOVA. Phenetic treatment indicated substantial population structure, measured both as differences in haplotype frequencies and partially nonoverlapping arrays of haplotypes in different populations. Good trees yielded virtually the same inference as the phenetic data. Bad trees had large Θ_t-values, both within and among populations, and they are such poor matches for the phenetic data that their inference was unreliable. The Φ_{st} criterion was extremely sensitive to small changes in the choice of tree, and it was best viewed as consequential analytical output, rather than as tree-defining input.

Even in those cases where a phenetic treatment of biogeographic pattern is productive, a patristic treatment may yield no meaningful improvement. That seems surprising at first, but a little reflection shows that it makes sense. To the extent that the data are inherently tree-like, phyletic information is explicitly embedded in the raw data and will be automatically incorporated into the phenetic distance matrix. Having expressed a preference for trees that faithfully mimic the phenetic data, it should come as no surprise that such trees yield virtually the same inference as the data themselves. In short, the inference that can be trusted the most is that derived directly from the phenetic data. A good tree will probably do no damage, but it will probably not help much, and a bad tree cannot be trusted at all.

Any inference drawn from a particular tree that is strongly divergent from the inference available from the phenetic data should be 'cause for pause'. We either have a 'bad tree' or the extraneous information (the evolutionary assumptions inherent in our choice of algorithm) are driving the analysis, notwithstanding phenetic data that tell a different story. It is possible to encounter situations like this, but the investigator should be very confident of the assumptions before overruling the data.

I have said little about the difficulty of actually discerning an optimal tree. My example was contrived to be small. There were only 15 candidate trees, and it was possible to examine them all. For the human mtDNA example just above, 10 000 random MSTs was still less than one in 100 000 of those available. There were almost surely better trees than the best we found, but there can be no guarantee that we will ever find the best tree. All tree-making algorithms are subject to the same dilemma, in the face of phenetic data that are phylogenetically ambiguous to any considerable extent. The number of tree choices is beyond our ability to enumerate, much less to analyse. We can find a good tree, but we will never know if we have found the best. In view of the subsequent inference limitations for our biogeographic survey, even were we able to identify an optimal tree, the entire exercise may be rather pointless.

Can we use the geography to choose a tree?

If we cannot use a tree to say something incremental about biogeographic pattern, can we at least use biogeographic pattern to say something incremental about a tree? Imagine a situation where the phenetic treatment yields convincing evidence of biogeographic structure or pattern, in spite of some homoplasic ambiguity in the connection pattern of peripheral haplotypes. Just to illustrate, return to the geographical array shown in Fig. 3. I have already commented that trees 1, 3, 5 and 9 are all serious contenders, but that there is not much to choose among them. A slightly different array of population samples could tip the balance in favour of one or the other, which says that we are relying on the nuances of how the rare haplotypes D and E are connected, from population to population, rather than on the strong phyletic signal inherent in the C–A–B–F pathway, which is already obvious from the phenetic data. To infer anything more is 'reaching'. In general, we should expect those features of the data set that are inherently tree-like to emerge naturally from the analysis, but they will already be obvious in the phenetic data. It is possible that one might encounter a biogeographic pattern that would provide compelling support for a particular tree, among a modest set of very good trees, but as good trees all mimic the phenetic data closely, I am not sanguine about the prospects for fine-tuning. A little gene flow and a little sampling variation go a long way toward confusing the issue. In closing this section, it is imperative to offer a cautionary note. If we insist on using biogeographic pattern to choose a tree, we are not also entitled to use that derived tree for biogeographic analysis. To do so would be to assume the conclusion, because the argument is circular.

Message 7. We might be able to use a good tree to say something incremental about biogeographic pattern. We might be able to use biogeographic pattern to say something incremental about the choice of tree. We cannot legitimately do both.

A tree is a picture

Multivariate statistical analysis of a voluminous biogeographic data set is definitely the way to proceed, but none of us thinks well in hyperdimensional space. A picture is worth a 1000 words (or molecular markers), and whatever a tree is not, it is certainly a picture. When all is said and done, we will need a picture to convey the story, so why not use a tree? Several comments are in order. First, there are myriad pictorial methods available for hyperdimensional data, most of which perform quite nicely without any assumptions about phyletic radiation. I have commented above on the fact that reticulation can be an important source of evolutionary pattern, and (mutational homoplasy aside) a strictly radiating tree can actually be misleading. Second, for the analysis of biogeographic pattern, it is obvious that ordination analyses, spatial autocorrelation analyses, and cluster analyses all have much to offer, and they all yield a picture. They all begin with either the raw data or the phenetic distance matrix, and the philosophical and practical difficulties inherent in using a tree instead are the same as those we have encountered with AMOVA. Third, most multivariate methods allow a dimensional reduction of the full data set, stressing the central tendencies and suppressing the exquisite peripheral detail. By contrast, phylogenetic trees actually increase dimensionality, beyond what is explicit in the data. While we can draw a tree in two dimensions, we should never forget that it is wandering through a hyperspace of higher dimension than are the data themselves. We are (to some extent) analysing/depicting nondata. Fourth, we need to distinguish between a tree used as a pictorial representation of an analytical result, and a tree used as an object of formal analysis. I am not averse to artful display, but we need to recognize it for what it is. A tree is one of several useful ways to draw a picture, and I have no objections, but my preference would be to analyse the data.

Acknowledgements

I am especially grateful to David Gordon, Rod Peakall, and three reviewers for voluminous commentary on the manuscript. Any remaining errors of omission or commission remain my own responsibility. This work was undertaken while I was a sabbatical visitor at the Division of Botany and Zoology, Australian National University, and I wish to thank the faculty, students and staff at ANU for their unfailing collegiality during my visit. The work was funded by the New Jersey Agricultural Experiment Station, NJAES-32106/USDA.

References

Avise JC (1989) Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution*, **43**, 1192–1208.

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–755.

Barrantes R, Smouse PE, Mohrenweiser HW *et al.* (1990) Microevolution in lower Central America. I. Genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a taxonomy based on genetics, linguistics, and geography. *American Journal of Human Genetics*, **46**, 63–84.

Birky CWJ (1988) Evolution and variation in plant chloroplast and mitochondrial genomes. In: *Plant Evolutionary Biology* (eds Gottlieb L, Jain S), pp. 25–53. Chapman and Hall, New York.

Brown BL, Epifanio JM, Smouse PE, Kobak CJ (1996) Temporal stability of mtDNA haplotype frequencies in American shad stocks: to pool or not to pool across years? *Canadian Journal of Fisheries and Aquatic Science*, **53**, 2274–2283.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Crandall KA, Templeton AR (1993) Empirical tests and some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, **134**, 959–969.

Crandall KA, Templeton AR (1996) Applications of intraspecific phylogenetics. In*: New Uses for New Phylogenies* (eds Harvey PH, Leigh Brown AJ, Maynard Smith J, Nee S), pp. 81–99. Oxford Press, Oxford.

Crandall KA, Templeton AR, Sing CF (1994) Intraspecific cladogram estimation: Problems and solutions. In**:** *Models in Phylogeny Reconstruction* (eds Scotland RW, Siebert DJ, Williams DM), pp. 273–297. Clarendon Press, Oxford.

Epifanio JM, Smouse PE, Kobak CJ, Brown BL (1995) Mitochondrial DNA divergence among populations of American shad (*Alosa sapidissima*): How much variation is enough for mixed stock analysis? *Canadian Journal of Fisheries and Aquatic Science*, **52**, 1688–1702.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB (1998) Ecological genetics of cactophilic Drosophila. In: *Ecology and Conservation of the Sonoran Desert Flora* (ed. Robichaux R). University of Arizona Press, Tucson, in press.

Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. *Genetics*, **136**, 343–359.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction sites. *Genetics*, **131**, 479–491.

Gibbons A (1985) *Algorithmic Graph Theory*. Cambridge University Press, Cambridge.

Harris SA, Ingram R (1991) Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plasmid transmission. *Taxon*, **40**, 393–412.

Hedges SB, Kumar S, Tamura K, Stoneking M (1992) Human origins and analysis of mitochondrial DNA sequences. *Nature*, **255**, 737–739.

Hudson RR (1989) How often are polymorphic restriction sites due to a single mutation? *Theoretical Population Biology*, **36**, 23–33.

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.

Huff DR, Peakall R, Smouse PE (1993) RAPD variation within and among natural populations of outcrossing Buffalograss (*Buchloë dactyloides* (Nutt.) Englem.). *Theoretical and Applied Genetics*, **86**, 927–934.

Joseph L, Moritz C (1995) Mitochondrial phylogeography of birds from eastern Australian rainforests: first fragments. *Australian Journal of Zoology*, **42**, 385–403.

Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F (1995) Reduced mtDNA diversity in the Ngöbé Amerinds of Panamá. *Genetics*, **140**, 275–283.

Long JC, Chakravarti A, Boehm CD, Antonarakis S, Kazazian HH (1990) Phylogeny of human β-globin haplotypes and its implications for recent human evolution. *American Journal of Physical Anthropology*, **81**, 113–130.

Maddison DR (1991) African origin of human mitochondrial DNA reexamined. *Systematic Zoology*, **40**, 355–362.

Maddison DR, Ruvolo M, Swofford DL (1992) Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Systematic Biology*, **41**, 111–124.

Mason-Gamer RJ, Holsinger KE, Jansen RK (1995) Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae*). Molecular Biology and Evolution*, **12**, 371–381.

Merriwether DA, Clark AG, Ballinger SW *et al*. (1991) The structure of human mitochondrial DNA variation. *Journal of Molecular Evolution*, **33**, 543–555.

Michalakis Y, Excoffier L (1996) A generic estimate of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, **142**, 1061–1064.

Morgan K, Strobeck C (1979) Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? *Nature*, **277**, 383–384.

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics*, **18**, 269–292.

Peakall R, Smouse PE, Huff DR (1995) Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. *Molecular Ecology*, **4**, 135–147.

Prim RC (1957) Shortest connection networks and some generalizations. *Bell System Technology Journal*, **36**, 1389–1401.

Quattro JM, Avise JC, Vrijenhoek RC (1991) Molecular evidence for multiple origins of hybridogenetic fish clones (Poeciliidae: Poeciliopsis). *Genetics*, **127**, 391–398.

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555–562.

Smouse PE, Chevillon C (1998) Analytical aspects of population-specific DNA-fingerprinting for individuals. *Journal of Heredity*, in press.

Smouse PE, Dowling TE, Tworek JA, Hoeh WR, Brown WM (1991) Effects of intraspecific variation on phylogenetic inference: A likelihood analysis of mtDNA restriction site data in cyprinid fishes. *Systematic Zoology*, **40**, 393–409.

Smouse PE, Long JC (1992) Matrix correlation analysis in anthropology and genetics. *Yearbook of Physical Anthropology*, **35**, 187–213.

Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. *Taxon*, **11**, 33–40.

Sokal RR, Smouse PE, Neel JV (1986) The genetic structure of a tribal population. XV. Patterns inferred by autocorrelation analysis. *Genetics*, **114**, 259–287.

Soltis DE, Soltis PS, Milligan BG (1992) Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. In: *Molecular Systematics of Plants* (eds Soltis P, Soltis D, Doyle J), pp. 117–150. Chapman and Hall, New York.

Strobeck C, Morgan K (1978) The effect of intragenic recombination on the number of alleles in a finite population. *Genetics*, **88**, 829–844.

Templeton AR (1983) Convergent evolution and non-parametric inferences from restriction data and DNA sequences. In**:** *Statistical Analysis of DNA Sequence Data* (ed. Weir BS), pp. 151–179. Marcel Dekker, New York.

Templeton AR (1992) Human origins and analysis of mitochondrial DNA sequences. *Nature*, **255**, 737.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–1507.

Ward RH, Frazier BI, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences (USA)*, **88**, 8720–8724.

Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? *Theoretical Population Biology*, **11**, 141–160.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wills C (1996) Topiary pruning and weighting reinforce an African origin for the human mitochondrial DNA tree. *Evolution*, **50**, 977–989.

This work grew out of the author's longstanding interest in the analysis of population structure in natural populations, which has taken the predictable turn toward phylogenetic methods. The author is a theoretical population geneticist who has worked on the interface between mathematical/statistical methods and the genetic data for which they are designed. He has worked on humans and other primates, both freshwater and marine fish, forest trees and forbs, agronomic grasses, and bacteria.