Progress with Methods for Constructing Evolutionary Trees

David Penny, Michael D. Hendy and Michael A. Steel

Evolutionists dream of a tree-reconstruction method that is efficient (fast), powerful, consistent, robust and falsifiable. These criteria are at present conflicting in that the fastest methods are weak (in their use of information in the sequences) and inconsistent (even with very long sequences they may lead to an incorrect tree). But there has been exciting progress in new approaches to tree inference, in understanding general properties of methods, and in developing ideas for estimating the reliability of trees. New phylogenetic invariant methods allow selected parameters of the underlying model to be estimated directly from sequences. There is still a need for more theoretical understanding and assistance in applying what is already known.

Reconstructing evolutionary trees has been notorious for its difficulties. But these same difficulties make the field exciting because the boundaries of hard science are being extended. It is difficult to make testable predictions[1] about unique events that happened a billion years ago. However, the aim must be to make the study of evolutionary trees as objective and quantitative as other branches of science. Our coverage is in two parts: the criteria that we want programs to meet, and recent developments that help to meet the criteria. The emphasis is on sequence data because trees from this source will increasingly be the basis on which the evolution of both molecular and phenotypic characters is studied. The field has been hindered by nonstandard mathematical nomenclature[2] so we clarify our usage in Box 1. Readers should also consult recent reviews[3–7].

Much of the earlier uncertainty resulted from a lack of knowledge of the strengths and weaknesses of different approaches to tree reconstruction. We would all like programs to meet the five criteria of being **efficient** (fast), **powerful**, **consistent**, **robust** and **falsifiable**. From our current perspective we can see that the earlier methods are best considered as exploratory data analysis. Such methods have an important place. Indeed, it should be pointed out that so far most of the useful results in molecular evolution have come from the simpler methods – though to balance this, the simpler methods have almost certainly given more wrong answers than right ones. Advocates of 'sophisticated' methods have seldom provided novel results! However, results from exploratory data analysis should be considered as tentative working hypotheses and should be followed by more detailed work. For reasons we explain later, we put both simple parsimony (minimal length) and methods using genetic distances in the category of exploratory data analysis. Classes of methods for inferring trees are summarized in Box 2, and their general properties are shown in Table 1.

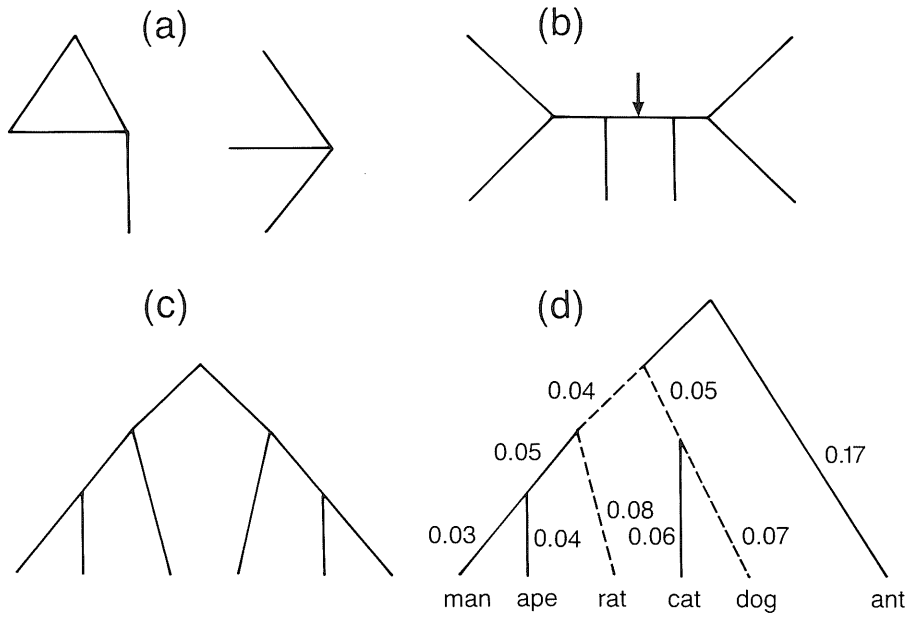
Desirable criteria for programs

Criterion 1: E*fficient* (*fast*)

The fundamental problem of the large number of trees is now well known by biologists. Most tree-building methods work by a **double optimization** procedure. The first optimization is on a single tree and evaluates an optimality criterion such as parsimony (minimum number of mutations), maximum likelihood, or minimizing sums of squares of genetic distances. Parsimony, for example, minimizes the number of

David Penny is in the Molecular Genetics Unit, and Michael Hendy and Michael Steel are at the Dept of Mathematics, Massey University, Palmerston North, New Zealand.

Box 1. Terminology for graphs and trees



Graphs (including trees) consist of **points** (**vertices, nodes**) and **edges** (**lines, internodes** or **links**); edges may have **weights** (edge **lengths**). (a) A graph consisting of two **components**, one with a **cycle**. (b) A **tree** (**connected** acyclic graph). (c) A **rooted tree** derived by inserting an identified additional point into the edge indicated by the arrow in (b). (d) A weighted rooted tree, also derived from (b). The **leaves** (**pendent** points) have been **labelled**, in this case, with names of taxa. This is an example of a **phylogenetic tree**, that is, a tree (rooted or unrooted) with leaves labelled. Often it is useful to assign nucleotides to the leaves. In addition, the edges have weights (edge lengths) so it is a weighted tree. The **path** (a series of **adjacent** edges) connecting dog to rat is indicated by a dashed line in (d). This path consists of two **pendent** edges and two **internal** edges, and passes three **internal** points. (b) and (c) are **topologies**, unlabelled trees either rooted or unrooted. The **degree** of a point is the number of edges incident at that point. A **binary tree** has points of degree one (leaves) and three (internal points). Ambiguous and unnecessary terms, best avoided, are 'network' (an unrooted tree or a connected graph with cycles) and 'branch' (either a path, an edge, or a subtree containing several edges and points – this latter being botanically and mathematically correct).

mutations needed to fit sequences onto a tree.

The second optimization step is finding the global optimum over all trees. It is here that the fundamental barrier of the large number of trees arises. Any computable problem is said to have an 'efficient' solution if an algorithm is known for which, even in the worst case, the time required increases no faster than a polynomial function of n, where n is some measure of the size of the problem. For example, if an algorithm for n taxa requires time or storage capacity proportional to, say, n or $n.\ln(n)$ or n^4, then it is classified as efficient. This contrasts with algorithms whose time requirement increases as an exponential function of n (say, 2^n or $n!$). Biologists are particularly well aware of the explosive increase of exponential growth from the potential increase of population numbers.

Is there an efficient algorithm for identifying the global optimum? We don't know of any, and it is unlikely that one will ever be found. It is an example of a large class of mathematical problems called NP-complete (NP = nondeterministic polynomial)[10]. These are interrelated in that: (a) if one problem could be solved efficiently then all could be; but, (b) no efficient algorithm is known for any of them. There is no proof that an efficient algorithm is impossible, though it is doubted that efficient algorithms exist.

NP-complete problems are framed as questions with a simple yes/no answer and it is straightforward to decide whether a suggested solution is valid. The best-known evolutionary example is finding the shortest tree: 'Is there a tree that requires only k, or fewer, changes for this data set?' It is easy to check whether a specific tree meets the requirement, but difficult in general to find such a tree. There is little comfort to biologists in knowing that mathematicians and computer scientists have, without success, spent years trying to find efficient algorithms for this class of problem. Programs that do find the global optimum for a given optimality criterion are called 'exact' methods. Because the global optimum may be so hard to find, people often resort to 'heuristic' methods (see Box 2) that find good solutions quickly, but without any guarantee of being the best.

Criterion 2: Powerful

Biologists accept that with short sequences we can get a 'sampling error', and would not be surprised if details of the optimal tree changed as longer sequences became available. But eventually, with longer sequences, we expect a program to settle on a single tree. This we call convergence[11]. But do we need sequences of 10^3, 10^5 or 10^7 base pairs long? The answer will depend on the method used for inferring trees. A method that converges with relatively short sequences we call powerful; one that converges slowly (requires very long sequences) is weak.

Different rates of convergence can result from methods using different amounts of the information in sequences. As the number of taxa increases, methods using genetic distances use only a vanishingly small fraction of the information in the data[12]. The number of distance values increases in proportion to n^2 while the number of classes in the data (bipartitions or splits) increases exponentially (see Box 3). Parsimony methods omit nucleotide positions that are constant or where only one nucleotide occurs more than once (singletons). But both types of position give information about the appropriate model of evolution. In addition, they usually occur more frequently so the estimates of their true frequencies are more accurately known.

Many methods ignore information from insertions and/or deletions (which are highly informative[13]) or from other types of biochemical information. Using only subsets of taxa (such as quartets) disregards most of the information that can come from incompatibilities. Variability within taxa is again ignored. Information loss is not limited to a particular method. All omit some information and

consequently the power of the method is reduced. As yet, we can make only rather weak predictions about the relative power of methods. For large data sets, maximum likelihood and closest tree methods use more sequence information than parsimony, and these together use more than is in distance data. Because of this, we expect the power of the methods to be: maximum likelihood ≈ closest tree[14] > parsimony > distance methods. However, it is more complicated than this because distance methods use more information than parsimony when the data set is small.

Criterion 3: Consistent

We would all like an algorithm to converge to the correct tree, the tree from which the data were generated. Unfortunately, this is not universally the case and so the requirement for consistency (convergence to the correct model) is additional to that of convergence. Consistency requires reference to a model of evolution and a model has three parts (see Fig. 1 for an example): (1) a **tree** (or more generally a graph) – the putative ancestral relationships among the taxa; (2) a **mechanism** for changes to the sequences; and (3) **edge lengths** (**weights** or **probabilities of change** on edges of the tree). In the two-character-state (two-color) case the probabilities of the states being different at the ends of an edge are the observed weights. The observed and corrected edge lengths are the **p** and γ values in Box 3. ('Discrete unordered character states' are 'colors' in mathematics and we will use this simpler expression.)

A frequent assumption is that changes occurring in the sequence are approximately 'independent and identically distributed'[15]. This means that changes to the sequence anywhere on the tree, and at any site, are independent, and also that all positions (nucleotide or amino acid) have the same chance of changing (identically distributed). We call this the **standard model**. Some mechanisms also assume the same rate along each edge of the tree – the molecular clock – and this imposes additional constraints on edge lengths.

The role of models of evolution

There has been confusion over the need for 'assuming' a model of evolution. You may want your programs to be 'free' of assumptions. The wording has made it appear as if assumptions could be added or omitted at will. At one level, programs are free of assumptions. An algorithm is simply a defined mathematical procedure or set of rules – a numerical recipe. As such, the output from an algorithm is quite independent of any model of evolution and many algorithms are applied to both biological and non-biological data.

However, biologists require an algorithm to give trees close to historical reality. We need to ask, are there possible mechanisms of evolution that will result in an algorithm converging to the correct tree? To take a ludicrous example, a method might join taxa based only on the alphabetical order of their scientific names. But there is no known mechanism of evolution that would relate evolutionary trees to Latin binomials. We must use our biological knowledge to select algorithms

Box 2. Classification of tree-building methods

Tree-building methods vary in many properties. They can be **exact**, or **heuristic** (quick and dirty). Exact methods consider all trees and so guarantee (given sufficient time) to find the best tree(s) for the optimality criterion being used. This of course does *not* guarantee that the tree is correct, just that it is the best possible for the data and optimality criterion used. A branch and bound algorithm usually greatly reduces the time required to find an exact solution but the reduction varies with the data. Heuristic methods run quickly and give an answer that, based on previous experience, is hoped to be 'close' to the correct solution.

Some methods use **sequences** directly; others use **distances** or dissimilarities. Distances are a subclass of dissimilarities where the values are **metric**; that is, for any three taxa i, j and k, $d_{i,i} = 0$, $d_{i,j} = d_{j,i} > 0$, and $d_{i,j} \leq d_{i,k} + d_{j,k}$; this last requirement is the triangle inequality. There is a loss of information in converting sequences to distances which gets worse as more taxa are used.

Methods may either optimize an **optimality criterion** (or objective function), or follow an algorithm without using an optimality criterion (**algorithmic methods**). Cluster analysis methods use distance data but no optimality criterion. Frequently-used optimality criteria are **parsimony** (for sequences or distances), **sum of squares** of differences between tree and observed distances, **maximum likelihood**, **closest tree**, and statistical tests based on **invariants**.

Methods that use an optimality criterion can be further subdivided into those that evaluate the objective function over all taxa, or evaluate the criterion on **subsets** of taxa (usually four taxa or **quartets**) and then look for a tree compatible with all subsets. Several quartet methods, including evolutionary parsimony[8], are described by Li *et al.*[9]

Some properties of the different classes are shown in Table 1.

Table 1. General properties of classes of tree-building methods

	Cluster analysis	Objective distance	Minimal length	Quartet methods	Hadamard transformation	Maximum likelihood
Sequences or distances	Distances	Distances	Sequences	Sequences	Sequences	Sequences
Uses objective function	No	Yes	Yes	Yes	Yes	Yes
Based on a mechanism of evolution	Not applicable	No	No	Yes	Yes	Yes
Calculated over all or subsets of taxa	All	All	All	Subsets	All	All
Allows selection of subset of trees, or gives arbitrary cut-off	Cut-off	Cut-off	Cut-off (selection)	Selection	Selection	Selection
Consistent (for a given model, will converge correctly)	No	No	No	No	Yes	Yes
Estimate of upper limit of number of taxa for exact calculation	Not applicable	10?	20?	?	20+	6

In many cases, qualifications need to be put on the summary given here. For example, maximum likelihood is a general statistical approach which with trees is usually used with sequence data. The maximal number of taxa for an exact method can only be approximate because different data sets vary in attributes such as the number of parallel changes and reversions. The numbers of taxa handled will be larger when a method is used as a heuristic.

Box 3. Parameters for the Hadamard (discrete Fourier) transforms

Index	Bipartitions	Subsets of even order	Four vectors: s	r	γ	p
0	$\{\},\{t_1,t_2,t_3,t_4\}$	$\{\}$	0.6288	0.0000	−0.5091	–
1	$\{t_1\},\{t_2,t_3,t_4\}$	$\{t_1,t_4\}$	0.1575	0.3771	0.2554	0.2
2	$\{t_2\},\{t_1,t_3,t_4\}$	$\{t_2,t_4\}$	0.0164	0.1421	0.0204	0.02
3	$\{t_1,t_2\},\{t_3,t_4\}$	$\{t_1,t_2\}$	0.0173	0.2758	0.0101	0.01
4	$\{t_3\},\{t_1,t_2,t_4\}$	$\{t_3,t_4\}$	0.0709	0.2231	0.1116	0.1
5	$\{t_1,t_3\},\{t_2,t_4\}$	$\{t_1,t_3\}$	0.0191	0.3771	0.0000	–
6	$\{t_2,t_3\},\{t_1,t_4\}$	$\{t_2,t_3\}$	0.0191	0.1421	0.0000	–
7	$\{t_1,t_2,t_3\},\{t_4\}$	$\{t_1,t_2,t_3,t_4\}$	0.0709	0.4990	0.1116	0.1
			1.0000		0.0000	

There are 2^{n-1} ways of partitioning n taxa into two disjoint subsets (**bipartitions** or **splits**) and these are shown above for the four taxa ($n = 4$) t_1, t_2, t_3 and t_4. Bipartitions arise in two ways during tree building. Each character with two colors, and each edge of a tree, subdivides taxa into bipartitions. Check this for both characters and edges of trees. There are also 2^{n-1} subsets with an even number of taxa, which are shown above as 'even-ordered subsets'. There are thus 2^{n-1} bipartitions and even-ordered subsets. There is a one-to-one relationship between them so they can be interconverted. For example, even-ordered subsets are the first subset of a bipartition (if it has an even number of members), otherwise it is formed by adding t_4 to the first subset. The even-ordered subsets are equivalent to sets of non-intersecting paths on a tree. The Hadamard transform relates bipartitions to even-ordered subsets, and vice versa.

From a data set the relative frequencies of bipartitions are estimated for the entries in **s**. In the example above they are the expected values under the model in Fig. 1. The Hadamard transform, followed by a standard correction for multiple changes, gives the inferred path length entries shown in **r**. The inverse Hadamard transform gives the entries γ. The first entry is the negative sum of all others so that the sum of entries in γ is zero. Other entries in γ are the inferred number of changes, per character, along edges of every possible tree. A tree is found by selecting a set of compatible bipartitions. Removing the expected number of multiple changes from γ recovers the **p** values of Fig. 1 precisely.

The two edge bipartitions with $\gamma_i = 0$ are the invariants for the model used. There is, for example, no edge on the tree in Fig. 1 joining {1,3} to {2,4,5} and so the γ for this edge bipartition (γ_5) is zero. Figure 2 shows the **s** and γ values as spectra.

that will converge to the correct tree under the widest possible range of reasonable biological assumptions.

To sum up, algorithms don't require assumptions, but the wrong algorithm is more likely to give a wrong result. The more we know both about the mechanisms of evolution and about the algorithms we use, the better our results should be. We can't escape models of evolution. A tree itself is part of the model!

To return to consistency. In an important development[15,16] it was shown that with only four taxa but with unequal rates of evolution, parsimony and some distance methods were not consistent, even under the standard model. With five or more taxa, parsimony is also inconsistent even with equal rates of evolution, and even with equal edge lengths for larger numbers of taxa[14,17,18]. The performance of parsimony can be improved with additional taxa that join into what otherwise would be long edges on the tree. Note that a method is inconsistent under a given model if there are any possible cases where it converges to the wrong tree. It is a separate, but important, question as to whether such cases are common. This leads us into the next criterion.

Criterion 4: *Robust*

A method may be consistent under the standard model, but be inconsistent with only small deviations from the model. Everyone would like a method that is powerful and consistent, even with sizable deviations from the model, i.e. a method that is robust. It appears that the standard model, with a tree and sites being independent and identically distributed, is the easiest for tree-reconstruction methods to handle correctly. Robustness is an area where there is almost a conspiracy of silence among those developing techniques. We have little general idea of the consistency of different approaches as real data start deviating from the simplest model.

We reported a case where a consistent method became inconsistent when the frequencies of character states (such as GC content) varied between taxa[6]. Empirical studies suggest that evolutionary parsimony also becomes inconsistent when nucleotide changes are asymmetric. An extension[19] to the method (which uses more, but not all, of the information in the data) improves performance. Allowing nucleotide positions to have different rates of evolution (with a lognormal distribution) gives more reasonable trees[20].

The problem of robustness can be studied by simulations. A model is specified, sets of data are calculated under this model, and these data sets are used to determine how often a tree-building program finds the original tree. This approach is still only of limited help because the results are not integrated into any theoretical understanding. Simulation studies have usually not separated the problem of the power of a method (rate of convergence) from consistency (converging to the correct tree), though this could easily be overcome. The real problem is that, considering all models, there are billions of combinations of parameters to be tested. By itself, testing four or five combinations of parameters is not that useful. However, simulations could be quite powerful when testing predictions from theoretical studies of algorithms. Some predictions are given in the section under convergence. It would be useful to attempt to falsify them, which is the topic of the next section.

Criterion 5: *Falsifiable*

The most fundamental criterion for a scientific method is that the data must, in principle, be able to reject the model. Hardly any tree-reconstruction methods meet this simple requirement. The worst offenders are cluster analysis methods because they (a) select a tree even if the data were not generated by a tree-like process, and (b) (because they have no optimality criterion) give no ranking of the other $|(2n-5)!!-1|$ trees.

One approach to falsifiability is by using tree comparison metrics to compare trees from different sets of data. This may be done by comparing trees derived from different data sets (say cytochrome *c*, hemoglobin α and rRNA) or by many random samples of columns from a data set[1].

There are several ways in which the standard model may be inappropriate. The data might not have been generated by a tree-like process. There are many ways a

non-tree model could sometimes arise: new viruses could arise by recombination of existing strains; many plants form hybrids and there is introgression between species; endosymbiosis as well as lateral transfer are possibilities; corrections between copies of a multigene family may give cycles in a graph; mixing gene and species phylogenies could confuse; and the problem of heterozygosity in natural populations can mislead tree reconstruction[21].

Statistical geometry[22] is an important new approach that retains information about deviations from a simple tree model. It has been useful[23] in studying RNA viruses such as the human immunodeficiency virus (HIV). Such studies with RNA viruses are particularly interesting because of the high rate of viral evolution (about a million times faster than DNA-based organisms). Experimental studies of evolution will increasingly be carried out with viruses simply because it occurs in 'real' time – on a human time scale.

The general problem of non-falsifiability remains with most methods. Goldman[24] describes the problem with current implementations of maximum likelihood, where either the tree or the mechanism of evolution has to be assumed. We prefer to give the results as observed and expected frequencies[17] of the different classes of data. This at least allows a decision that the data do *not* fit the model! This does not indicate in what way (or ways) the model fails. But it is then possible to try alternatives and see whether a variation on the model fits the data better – it could be allowing an extra edge that represents hybridization, a different distribution of rates of changes between nucleotide positions[20], changing proportions of nucleotides (GC content, for example), and so on. It is possible to keep adding more and more parameters to 'fit' the data better. Caution is required here because adding more parameters allows a fit to virtually any model, and the accuracy of each parameter is markedly reduced. The aim is to use as few as necessary; assuming a molecular clock requires only $n-1$ parameters compared with $2n-3$ for an unrooted tree[25]. There are other recent discussions of determining how many parameters are appropriate[26,27].

In conclusion, for the foreseeable future the five criteria appear incompatible. An example of the clash among the first three criteria is shown by comparing parsimony with the standard maximum likelihood method. Parsimony, combined with one of the three branch and bound approaches[28], is an exact method for 20 or more taxa – even though it is inconsistent. Under the standard model, maximum likelihood methods are consistent, but they are usually used heuristically (Box 1) on a subset of trees, and then usually for a small number of taxa.

Recent developments: invariants and spectral analysis

In this section we will mention some improvements to existing approaches but concentrate on the newest approach, invariants. Exact methods, such as branch and bound, can be used heuristically by taking only the best choice at any level and not completing the total search. The neighbour joining method for distances[29] appears one of the most effective. It is important to understand why it works well – is it by using more of the information when selecting the edge of the tree? Its general approach to tree building may be even more useful when combined with other optimality criteria. A new heuristic approach that appears of great value is the Great Deluge algorithm[30]. This appears to be an effective heuristic program for a wide variety of problems. It is very simple: a random walk combined with a constantly increasing 'water level'. The program moves rapidly over a wide range of solutions and is particularly effective at escaping from local minima.

The major limitation with current maximum likelihood methods is that, despite being consistent, they usually require excessive computing time even with less than ten taxa. For each tree, many different edge lengths need to be tested. Wouldn't it be desirable if some properties of the model could be inferred directly from the data, without having to optimize a large number of parameters by trial and error? It is the ability to calculate properties of the model directly from the data that is the advantage of invariant methods and spectral analysis.

What are invariants? The term is a general mathematical one for functions, defined on the expected output of a model, that will take the same value (that is, are invariant) whatever values certain 'nuisance parameters' take. In the present context, the model can be the tree and specified mechanism of evolution, and the nuisance parameters the unknown (but variable) edge weights (lengths) of the tree. Linear invariants are a subclass that are linear combinations of parameters in the data. Thus it is valid to apply the usual statistical tests of significance.

For data generated exactly by a model, phylogenetic invariants will take a fixed value, and this will generally be different for data from another tree. We cannot expect real data to fit a model exactly but statistical tests are possible that allow invariants to distinguish between phylogenies. Because invariants are calculated directly from the sequences we will refer to them as 'direct' methods. This is in contrast to standard maximum likelihood, which works in the other direction: it starts with a model and preliminary estimates of parameters, calculates some expected properties of the data, then optimizes the parameters of the model. Invariant methods estimate properties of the model from the data; maximum likelihood methods calculate expected properties of the data from a model. Perhaps combining them would be even more powerful?

Three methods were introduced in 1987. The best known is **evolutionary parsimony**[8], which operates with nucleotides on subsets of four taxa. In the same year Cavender and Felsenstein[31] published invariants for two-state characters on trees with four taxa. We introduced a Hadamard (discrete Fourier) transform[14], which at that time was limited to about eight taxa and two colors.

Evolutionary parsimony[8] has three functions corresponding to internal edges of the three unrooted trees for four taxa (these correspond to the bipartitions $\{1,2\}$, $\{1,3\}$ and $\{1,4\}$

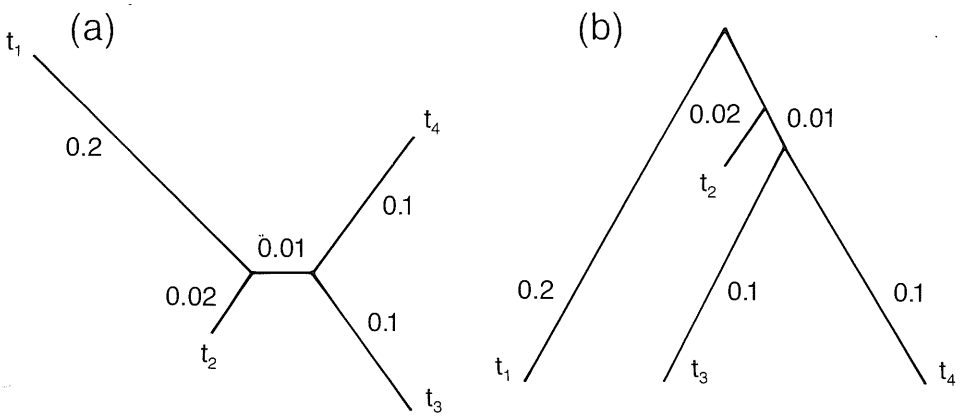


Fig. 1. Tree used for Hadamard transform calculations (Box 3): (a) the unrooted tree and (b) a rooted tree derived from it by inserting the root on the longest edge. There is a much lower rate of evolution on the lineage leading to taxon t_2. **p** values are probabilities, per character (nucleotide position), that it is in a different character state at the ends of an edge. The values of **p** for pendent edges are shown; **p** values are 0.01 for the internal edge.

– see Box 3). Values for two of these functions (the invariants) are expected to be zero, thus rejecting two trees. The third is expected to be positive. Insufficient data and/or oversimplified assumptions about mechanisms of evolution may leave the third value not significantly greater than zero. Scientifically this is important as it suggests when the data are unable to resolve a problem under a given model. If the correct tree T is ((1,2)(3,4)) then the value for {1,2} is expected to be positive while those for {1,3} and {1,4} should be zero.

Since 1987 there have been two lines of development: *ad hoc* methods for four colors (nucleotides) for small numbers of taxa[27,32,33], and a general method for up to 20 taxa with two colors[34]. The general two-color case is shown in Figs 1 and 2 and Box 3. The main points are that with two colors (see Box 3) there are 2^{n-1} **bipartitions**, and 2^{n-1} **subsets** with an even number of taxa. A simple but powerful matrix of 1's and −1's, the Hadamard matrix, interconverts the bipartitions and subsets. All calculations are invertible. The procedure is a discrete Fourier analysis so it is appropriate to refer to it as spectral analysis. The vector γ has 2^{n-1} entries. For a binary tree T, $2n-3$ of these (the edge bipartitions of T) should be positive and the remaining entries (the invariants) are expected to be zero – except the first, which is the negative sum of all the other entries, thus constraining the sum of all entries to be zero. An optimal tree for 20 taxa has been found with nucleotide data[6].

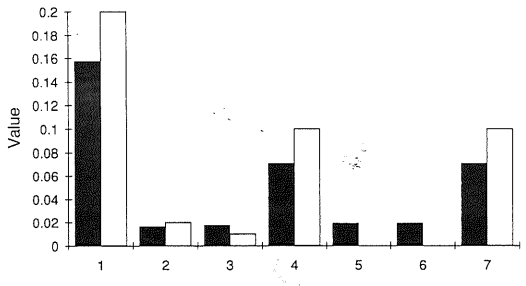


Fig. 2. Spectra derived from the model in Fig. 1. Entries for s_{1-7} and γ_{1-7} from Box 3 (s_0 and γ_0 are omitted). For each pair the **s** value is on the left (shaded) and the γ value is on the right (unshaded). In general, bipartitions that are larger in **s** are also larger in γ, and vice versa. But in this model, the observed value s_5 is larger than s_2, even though γ_5 is zero and γ_2 is positive. This results from the correction for multiple changes on different edges of the tree.

The four-color problem for evolutionary trees

Can the two approaches be unified by extending to four colors for many species? From our current knowledge, a generalized invariant method on four colors looks like the next major goal for phylogeny.

Is it possible? Almost certainly, yes! There is a well-developed mathematical theory of Fourier transforms that describes necessary and sufficient conditions for Fourier analysis. The Hadamard transform fits comfortably into this analysis. A recent manuscript[35] describes the general features of four-state invariants for the Kimura three-parameter model – but their description is quite abstract. Progress towards the goal has been made (L. Székely, P. Erdös, M.A. Steel and D. Penny, unpublished), but there are still many unresolved problems. What is the relationship between partitions of the characters and edges of the tree (it is no longer simply one-to-one)? What are the equivalents of the even-sized subsets of taxa that form the path sets?

Solving the four-color problem for evolutionary trees would be a major step to more powerful methods. Of course, it will never be enough. Insertions and deletions (indels) are highly informative[13] and should not be neglected. We may need a six-color version to include indels. Proteins would require a 20-color version, information on gene rearrangements will be important, and so on.

We are optimistic that over the next few years there will be continuing progress in methods for inferring and testing evolutionary trees. More powerful approaches will be found. However, a real bottleneck is developing in implementing these in the standard packages[36] so that evolutionists can evaluate them. Experience will be needed to modify and select the most effective ideas. Biologists who want to use newer approaches must insist that more support is given to those who link new ideas into workable computer packages.

Acknowledgements

Preparation of this manuscript was assisted by the Centre for Interdisciplinary Research (ZiF), University of Bielefeld, Germany.

References

1 Penny, D., Hendy, M.D. and Steel, M.A. (1991) in *Phylogenetic Analysis of DNA Sequences* (Miyamoto, M.M. and Cracraft, J., eds), pp. 155–183, Oxford University Press

2 Hendy, M.D. and Penny, D. (1984) *Syst. Zool.* 33, 245–247

3 Felsenstein, J. (1988) *Annu. Rev. Genet.* 22, 521–565

4 Sneath, P.H.A. (1989) *Syst. Appl. Microbiol.* 12, 15–31

5 Swofford, D. and Olsen, G. (1990) in *Molecular Systematics* (Hillis, D.M. and Moritz, C., eds), pp. 411–501, Sinauer Associates

6 Penny, D., Hendy, M.D., Zimmer, E.A. and Hamby, R.K. (1990) *Aust. Syst. Bot.* 3, 21–38

7 Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology,* Oxford University Press

8 Lake, J.A. (1987) *Mol. Biol. Evol.* 4, 167–191

9 Li, W-H., Wolfe, K.H., Sourdis, J. and Sharp, P.M. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 847–856

10 Graham, R.L. and Foulds, L.R. (1982) *Math. Biosci.* 60, 133–142

11 Penny, D. and Hendy, M.D. (1986) *Mol. Biol. Evol.* 3, 403–417
12 Steel, M.A., Penny, D. and Hendy, M.D. (1988) *Nature* 336, 118
13 Lloyd, D.G. and Calder, V.L. (1990) *J. Evol. Biol.* 4, 9–21
14 Hendy, M.D. and Penny, D. (1989) *Syst. Zool.* 38, 297–309
15 Cavender, J.A. (1978) *Math. Biosci.* 40, 271–280
16 Felsenstein, J. (1978) *Syst. Zool.* 27, 401–410
17 Penny, D., Hendy, M.D. and Henderson, I.M. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 857–862
18 Steel, M.A. (1990) *Bull. Aust. Math. Soc.* 41, 159–160
19 Sidow, A. and Wilson, A.C. (1990) *J. Mol. Evol.* 31, 51–68
20 Olsen, G.J. (1988) *Cold Spring Harbor Symp. Quant. Biol.* 52, 825–837
21 Pamilo, P. and Nei, M. (1988) *Mol. Biol. Evol.* 5, 568–583
22 Eigen, M., Winkler-Oswatitsch, R. and Dress, A. (1988) *Proc. Natl Acad. Sci. USA* 85, 5913–5917
23 Eigen, M. and Nieselt-Struwe, K. (1990) *AIDS* 4, S85–S93
24 Goldman, N. (1990) *Syst. Zool.* 39, 345–361
25 Degens, P.O., Lausen, B. and Vach, W. (1990) in *Trees and Hierarchical Structures* (Dress, A. and von Haeseler, A., eds), pp. 9–42, Springer-Verlag
26 Hasegawa, M. and Kishino, H. (1989) *Evolution* 43, 672–677
27 Narvidi, W.C., Churchill, G.A. and von Haeseler, A. (1991) *Mol. Biol. Evol.* 8, 128–143
28 Penny, D. and Penny, P. (1990) in *Trees and Hierarchical Structures* (Dress, A. and von Haeseler, A., eds), pp. 75–91, Springer-Verlag
29 Saito, N. and Nei, M. (1987) *Mol. Biol. Evol.* 4, 406–425
30 Dueck, G. (1990) *New Optimization Heuristics: The Great Deluge Algorithm and Record to Record Travel,* IBM Technical Report, Heidelberg
31 Cavender, J.A. and Felsenstein, J. (1987) *J. Classif.* 4, 57–71
32 Cavender, J. (1991) *Math. Biosci.* 103, 69–75
33 Sankoff, D. (1990) *Mol. Biol. Evol.* 7, 255–269
34 Hendy, M.D. and Penny D. *J. Classif.* (in press)
35 Evans, S.N. and Speed T.P. (in press)
36 Sanderson, M.J. (1990) *Syst. Zool.* 39, 414–420