New Zealand Journal of Botany, 1993, Vol. 31: 231–237
0028–825X/93/3103–0231 $2.50/0

Hadamard conjugation: a versatile tool for modelling nucleotide sequence evolution

MICHAEL D. HENDY
MICHAEL A. CHARLESTON
Department of Mathematics
Massey University
Private Bag 11 222
Palmerston North, New Zealand

Abstract Hadamard conjugation has proved to be a useful tool in examining some of the properties of the patterns of nucleotide sequences arising from the evolution of the taxa they represent. It has a considerable advantage in that the formulae are independent of the phylogenetic structure under consideration, and can be given for any number of taxa. Hadamard conjugation is outlined and four applications are introduced. The applications are the theoretical examination of tree building methods, the generation of sample sequences under various models for simulation studies, the identification of some phylogenetic invariants, and the closest tree method for inferring phylogenetic trees and their edge lengths.

Keywords nucleotide sequences; evolution; Hadamard conjugation; phylogenetic invariants; closest tree algorithm

INTRODUCTION

The set of aligned homologous sequences representing a set $S = \{t_1, t_2, ..., t_n\}$ of n taxa is a consequence of the evolution of those taxa and should reflect, to some degree, their evolutionary history. Each site of the sequences individually labels the taxa by their character states. In this paper we will assume that the character states are either four nucleotides {A, C, G, T/U} or one of two states, purines and pyrimidines, for example.

The character states at a site split (partition) the taxon set into subsets of taxa with common character state. We refer to such a partition as a **site partition**. (A partition of S is a collection of subsets such that each taxon belongs to precisely one subset.) If all character states at a site are the same, then the corresponding partition is trivial, just S itself. If only two character states occur at a site, the site partition will comprise two subsets. We refer to partitions with only one or two subsets as **bipartitions**, and those with at most four subsets as **quadripartitions**. There are 2^{n-1} possible bipartitions and 4^{n-1} possible quadripartitions. The site partitions for two state characters are bipartitions, and those for four state characters are quadripartitions.

A phylogenetic tree linking the n taxa is a tree with n endpoints, these being labelled by the n taxa. When we do not specify the placement of the common ancestor, the tree is unrooted, with every internal point linked by edges to at least three other points, and the tree has at most $2n-3$ edges. An edge e of T defines a bipartition of the set S of taxa comprising the subset of taxa to the left of e and the subset of taxa to the right of e. We will refer to this as an **edge bipartition**. A collection of bipartitions is **compatible** if it is a set of edge bipartitions of a common tree T. Hadamard conjugation provides a link between the probabilities of character state changes on the edges of T and the expected frequencies of each of the site bipartitions in the consequent character sequences, for several models of character state changes.

A simple model of sequence evolution on a phylogenetic tree T, Cavender's model (Cavender 1978), assumes that there are only two character states (R and Y, say), that the state changes occur independently at different sites and on different edges of T, and that all changes across an edge e of T at different sites are governed by a common probability p_e that the states at the endpoints differ. (This is a symmetric model in that it assumes that the changes R to Y and Y to R are equally likely.

B93009
Received 19 January 1993; accepted 5 July 1993

Under this assumption, we do not need to know the distribution of the character states at the root.) The model is described by $(T, \mathbf{p})$, where T is the phylogenetic tree and $\mathbf{p}$ is a vector of the probabilities p_e for the edges e of T. We linearise $(T, \mathbf{p})$ to form a vector $\mathbf{q}$ of 2^{n-1} components called the **edge length spectrum** (Hendy et al. 1992). For $i > 0$, the q_i's that do not correspond to an edge of T are equal to 0, while those that do are greater than 0. The q_i's that correspond to the edges in T are calculated by

$$q_i = \frac{-1}{2}(1 - \ln(2p_i)),$$

and are additive on the edges of the tree. Under the Cavender model, the expected number of character state changes occurring on edge e_i is q_i.

The Hadamard conjugation (Hendy et al. 1992),

$$\mathbf{s} = \mathbf{H}^{-1} \exp(\mathbf{Hq}), \quad (1)$$

produces a vector of probabilities $\mathbf{s}$. $\mathbf{H}$ (= $\mathbf{H}_{n-1}$; see Table 1) is a symmetric Hadamard matrix of 2^{n-1} rows and columns, $\mathbf{H}^{-1} = (1/2^{n-1})\,\mathbf{H}$, and the exponential function (exp) is applied individually to each of the 2^{n-1} components of $\mathbf{Hq}$. Each component s_i of $\mathbf{s}$ is the expected relative frequency of the i-th bipartition as a site bipartition, where the bipartitions are numbered according to the scheme described by Hendy & Penny (1993). $\mathbf{s}$ is referred to as the **sequence spectrum**. The relationship (1) is easily inverted using the natural logarithm function (ln) applied individually to each of the 2^{n-1} components of $\mathbf{Hs}$, so

$$\mathbf{q} = \mathbf{H}^{-1} \ln(\mathbf{Hs}), \quad (2)$$

which is also a Hadamard conjugation. Thus, $\mathbf{q}$, and hence $(T, \mathbf{p})$, can be recovered from $\mathbf{s}$.

Kimura's three parameter model (Kimura 1981) allows three independent parameters, p_e^1, p_e^2, and p_e^3, of nucleotide change on each edge e of T. (p_e^1 is the probability of a transition, p_e^2 and p_e^3 are probabilities of transversions, A $\leftrightarrow$ G or C $\leftrightarrow$ T, and A $\leftrightarrow$ T or C $\leftrightarrow$ G, respectively.) These can be encoded by $(T, \mathbf{p}^1, \mathbf{p}^2, \mathbf{p}^3)$ and linearised to give an edge length spectrum $\mathbf{q}$ of 4^{n-1} components (Steel et al. 1992). The Hadamard conjugations in equations (1) and (2) also apply to this spectrum, with $\mathbf{H} = \mathbf{H}_{2n-2}$, and $\mathbf{H}^{-1} = (1/4^{n-1})\,\mathbf{H}$. The corresponding sequence spectrum $\mathbf{s}$ contains the expected frequencies of each of the 4^{n-1} possible site quadripartitions. Again, these results are independent of the frequencies of the states at the root.

Table 1 The Hadamard matrix $\mathbf{H}$ is a square matrix whose entries are all 1 or −1, and with every row (and column) orthogonal to every other row and column. The Hadamard matrices we use can be easily described using Kronecker products:

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{H}_2 = \mathbf{H}_1 \otimes \mathbf{H}_1 = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_1 \\ \mathbf{H}_1 & -\mathbf{H}_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

$$\mathbf{H}_3 = \mathbf{H}_1 \otimes \mathbf{H}_2 = \begin{bmatrix} \mathbf{H}_2 & \mathbf{H}_2 \\ \mathbf{H}_2 & -\mathbf{H}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}, \text{ etc}$$

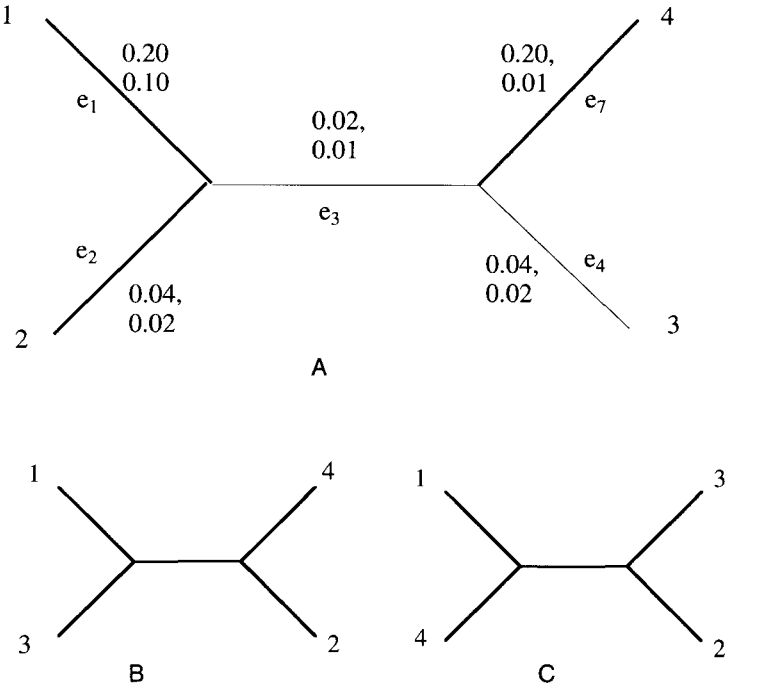


Fig. 1 With the edge length probabilities on the tree of Fig. 1A for a sequence length of 1000 we use equation 1 to calculate the expected numbers of sites with various partitions. (These are the components, rounded to the nearest whole number, of 1000 × **s**.) For each edge e, the figures are the probabilities of a transition and of a transversion between the endpoints of e.

We calculate that the expected number of sites with the partition $\{\{1, 2\}, \{3, 4\}\}$ is 25 (19 of these are transitions and 6 transversions)}, with the partition $\{\{1, 3\}, \{2, 4\}\}$ is 13 (11 transitions and 2 transversions), and with the partition $\{\{1, 4\}, \{2, 3\}\}$ is 39 (35 transitions and 4 transversions). For parsimony on four taxa these are the only "informative" sites, hence parsimony will incorrectly favour the tree of Fig. 1C over the generating tree of Fig. 1A.

APPLICATIONS

Consistency of tree building

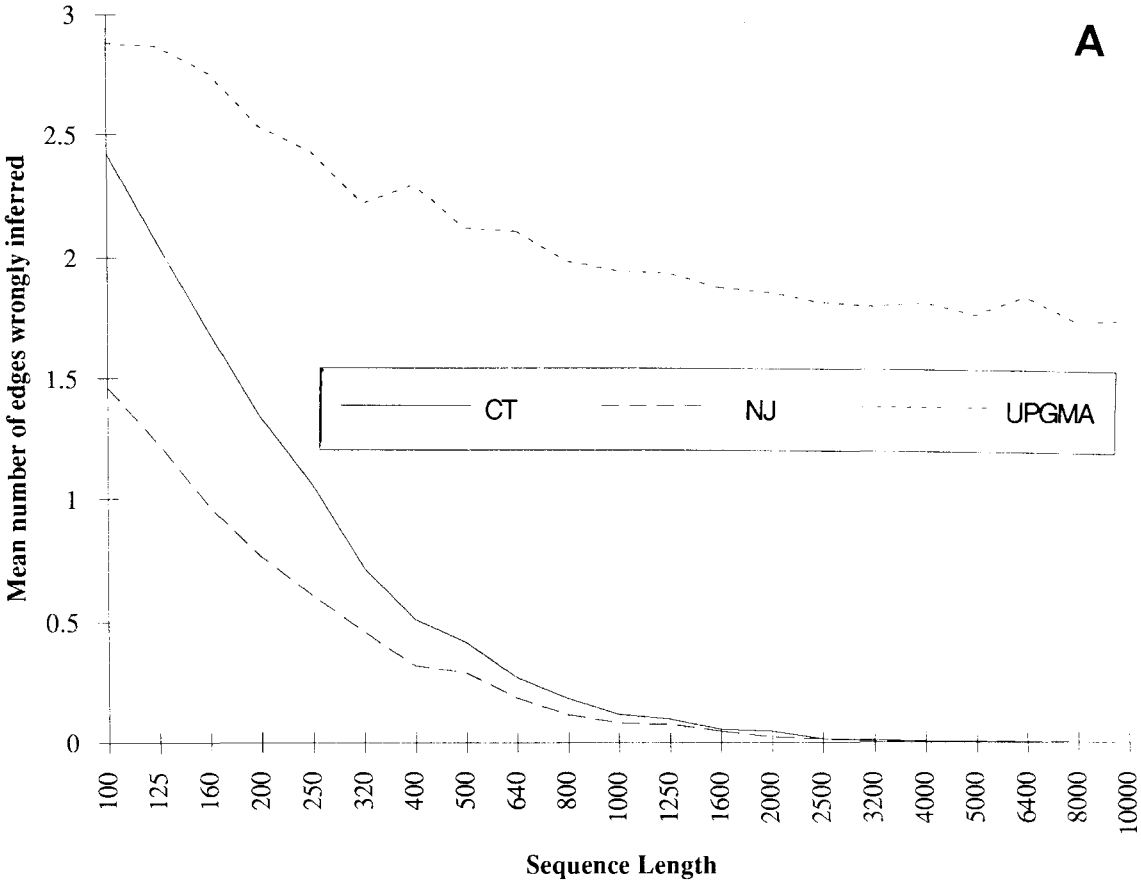
For the models above, the tree T and the initial probabilities **p** can be precisely calculated from knowing the site bipartition probabilities **s** accurately. Tree reconstruction methods such as Neighbour-joining (Saitou & Nei 1987), Closest Tree (Hendy & Penny 1993), and Maximum Likelihood (Felsenstein 1981), which will guarantee to produce T accurately from the site probabilities, are called **consistent** methods. It is known that Maximum Parsimony is not consistent (Felsenstein 1978; Hendy & Penny 1989). Figure 1 gives an example using Kimura's three parameter model on four character states, illustrating this inconsistency. (In this case, setting the two transversion probabilities as the same for each edge, we reduce it to his two parameter model.) However, if parsimony was applied to the components of the conjugate spectrum, where the partition frequencies had been adjusted to remove the effects of parallel and reverse changes, the method would be consistent (Steel et al. in press).

Using equations (1) and (2) we can analyse tree building methods for consistency, either to prove analytically that a method is consistent under a particular model of character change, or to produce some counter-example to show that it is inconsistent. Such a counter example will comprise the vector probabilities **s** obtained from a $(T, \mathbf{p})$, so that the tree building method under scrutiny does not produce the generating tree T.

Simulation

In actual studies of data derived from nucleotide sequences we can obtain only a finite sample, and hence estimate **s** only from the observed frequencies **f** of the site partitions. There are a number of potential difficulties to be considered. The model imposes some idealised conditions from which the observed data may deviate to a lesser or greater extent, or there may be errors in the data as read or transcribed. Also, the limitation of sequence lengths to not more than about 10^3 characters means sampling error cannot be ignored. Theoretical and simulation studies are possible to measure the effects of violations of the hypothesised model, and sampling and data errors.

We can calculate the expected bipartition frequencies **s** for a given set of edge lengths **q** and then sample from this distribution to effectively generate a set of aligned character sequences of any desired length. This is often much faster than



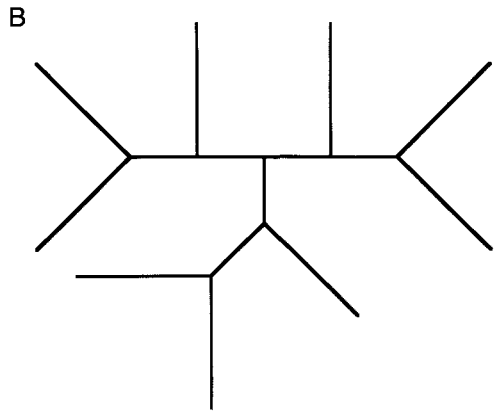


Fig. 2 This graph (A) shows the mean number of edges within the tree produced by three phylogenetic methods that differ from the edges in the generating tree (B), when edge lengths were randomly chosen from uniform distributions with certain bounds. The number of edges that differ is equivalent to half the symmetric tree difference of the two trees (Steel 1988). The methods included are UPGMA (Sokal & Michener 1955), Closest Tree (Hendy & Penny 1989), and Neighbour-joining (NJ) (Saitou & Nei 1987). Note that UPGMA does not converge to the correct tree as more data are added. In this example, CT does not perform as well as NJ when the sequence length is less than about 3000, but we find that it is superior when longer sequences are used.

mimicking the evolutionary process by "growing" sequences along a tree. The large computational overhead of calculating **s** is offset by the relative speed of sampling from it. We can therefore carry out repeated trials with the same set or different sets of edge lengths, and study the ability of tree reconstruction methods to deliver the generating tree.

This sampling approach has been used to compare the performance of a number of different tree reconstruction methods with varying sequence length, over all the 11 possible shapes of unlabelled binary trees with 10 taxa. Some results from this work are shown in Fig. 2. The graph in Fig. 2A shows the relatively poor performance of the inconsistent method UPGMA, relative to the consistent methods, and illustrates the effect on tree

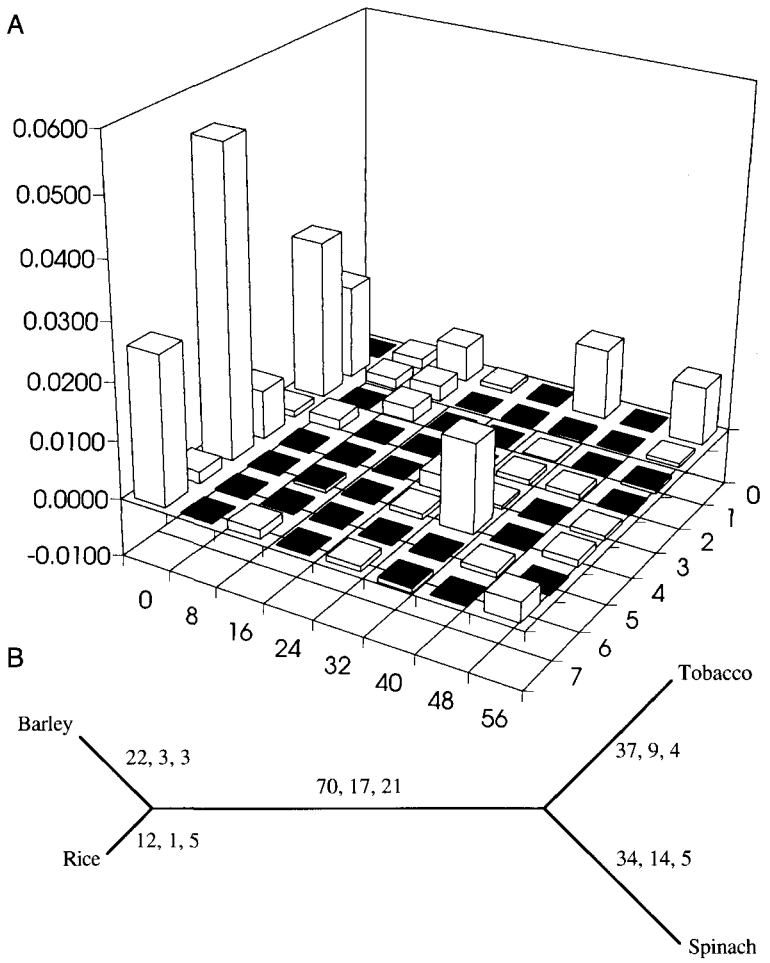


Fig. 3 **A**, The conjugate spectrum derived from four DNA sequences encoding subunits of *atp* synthetase (a chloroplast encoded protein) obtained from Lockhart et al. (1992). The vector has been drawn as an 8×8 array, with the 0-th entry omitted. The entries, which can relate to edges of a tree, form the leading row, column, and diagonal. The expected values of the off-diagonal elements are 0 as they are tree invariants. **B**, The closest tree, together with the best fit (least squares) numbers of nucleotide changes, transitions followed by the two types of transversions, A $\leftrightarrow$ G or C $\leftrightarrow$ T, and A $\leftrightarrow$ T or C $\leftrightarrow$ G, for each edge of the tree. The root would probably be placed on the long internal edge. The rates for the edges to spinach and tobacco are very similar.

reconstruction of sampling error arising from short character sequences.

The Closest Tree algorithm

A further application of the Hadamard conjugation is the **Closest Tree** algorithm for estimating T. If the relative bipartition frequencies $\mathbf{f}$ closely estimate $\mathbf{s}$, then as multiplying by matrices is linear, and the natural logarithm function is almost linear in its range of application, the resultant **conjugate spectrum**

$$\gamma = \mathbf{H}^{-1} \ln(\mathbf{Hs}) \qquad (3)$$

should closely estimate $\mathbf{q}$. For any particular tree T, we can find $\mathbf{q}(T)$ so that the distance between γ and $\mathbf{q}(T)$ is minimal. The closest tree procedure (Hendy 1991) selects the tree T for which this distance is minimal. This procedure returns the least squares best fit $\mathbf{q}$ vector for the closest tree T, from which the edge change probabilities $\mathbf{p}$ can be derived. However the q values, which are additive, may be more useful. If the changes are modelled by a Poisson process, then these are the expected numbers of changes (per site) on each edge. In Fig. 3 we illustrate the closest tree for a set of four chloroplast encoded genes together with the estimates of the numbers of changes of the transitions and the two types of transversions.

Invariants

The components of γ other than γ_0, which do not correspond to edges of the tree T, have expected value zero, and hence are **invariants** of the data for T. Invariants are functions of the data whose expected values can be used to discriminate between, and estimate reliabilities of, competing phylogenetic hypotheses. A more detailed description of the invariants derived from the Hadamard conjugation is given in Steel et al. (1993, this issue).

COMPUTATIONAL COMPLEXITY

Hadamard Conjugation

The direct multiplication of **Hv**, where $\mathbf{H} = \mathbf{H}_{n-1}$ has 2^{n-1} rows and columns and **v** is a vector of 2^{n-1} components, requires $O(2^{2n})$ operations. However, we can exploit the iterative structure of **H** to perform this multiplication in $O(n2^n)$ operations (the Fast Hadamard Transform, Hendy & Penny 1993). This is still of exponential order, but it effectively doubles the number of taxa that can be analysed with the same computing resources. As $\mathbf{H}^{-1} = (1/2^{n-1})\mathbf{H}$, the Hadamard conjugation

$$\mathbf{y} = \mathbf{H}^{-1}\,\varphi(\mathbf{Hx})$$

can also be computed in $O(n2^n)$ operations, as φ (e.g., exp or ln) is computed separately on each of the 2^{n-1} components of **Hx**. For practical purposes, this allows us to calculate the conjugation for up to $n = 20$ taxa for two colours, and up to $n = 10$ for four colours, on a PC.

Simulation

A common method of generating artificial data for the study of tree reconstruction methods is to produce a common ancestral sequence, and for each of its nearest descendents and each site of the sequence, calculate using pseudo-random numbers whether the character state at that site is the same as that at the corresponding site in the ancestral sequence, and if not, to which state it changes. This process, which continues until a sequence of characters is generated for each vertex in the tree, requires $O(nc)$ operations, where c is the sequence length and n the number of taxa.

However, by calculating the vector **s** of partition probabilities (bipartition or quadripartition) and then sampling from this vector, we can generate a spectrum of "observed" partition frequencies. For small numbers of taxa, say not more than about 10, substantial savings in computation time can be made. The calculation of **s** requires $O(n2^n)$ operations for the case of two colours and $O(n4^n)$ operations for four colours. The resulting set of 2^{n-1} numbers (4^{n-1} for four colours) can be sorted in descending order in $O(2^n \ln(2^n)) = O(n2^n)$ operations, forming a vector **v**, say, and then a new vector, say **w**, is constructed from the partial sums of the components of **v**, that is, $w_0 = v_0$, $w_i = w_{i-1} + v_i$. This can then be sampled by generating pseudo-random numbers r and repeatedly testing for increasing values of i to find i such that $w_{i-1} < r \leq w_i$. The number of partitions corresponding to the i-th entry in **v** is then incremented. This process is carried out for each site, and gives the observed partition frequencies **f**.

For edge lengths that are not too large, the most common partitions are the trivial partition, in which all character states at the endpoints of T are the same, and those partitions that correspond to single changes of character state on single edges of the tree. Hence, in the vector **w** of cumulative probabilities, these $(2n-2)$ partitions will be most often chosen, requiring at most $2n-2$ comparisons of r with w_i. Some estimated numbers of operations required for this procedure are listed in Table 2.

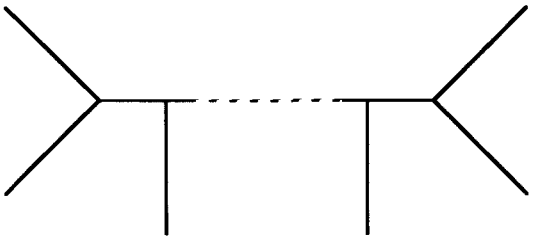


Fig. 4 The shape of the tree used in the generation of the data in Table 2. This is known as the "caterpillar" tree.

Table 2 Estimated number of operations required to generate bipartition spectra for small numbers of taxa, on the "caterpillar" tree (Fig. 4), with internal edge lengths shown and pendant edge lengths equal to twice the internal edge lengths. The number of characters in the sequences is 1000. The number of operations is calculated by assigning one operation to each mathematical operation, whether it be a floating-point multiplication or a comparison of two floating-point numbers. Hence, these figures are indicative of the general trend, but are not precise. In the last column the equivalent number of operations is shown for artificially "growing" data from an ancestral sequence.

	Sampling data					
	Internal edge lengths					Growing
n	0.001	0.003	0.01	0.03	0.1	data
6	2639	2789	3323	4898	9970	11000
7	3449	3664	4471	7101	16918	13000
8	5510	5807	6992	11347	30382	15000

The Closest Tree algorithm

The Hadamard conjugation requires $O(n2^n)$ operations. Methods of reconstructing phylogenetic

trees that search all possible trees have to evaluate an optimality function on up to $(2n-3)!! = (2n-3)(2n-5)...(3)(1)$ trees, which grows faster than $n2^n$. For example, with $n = 10$, $9 \times 2^9 = 4608$ operations are required for each of the multiplications by **H** in the Hadamard conjugation, and $2^9 = 512$ operations for the natural logarithm function. Thus, to infer γ from **f** would require 9728 operations. However, for this number of taxa, there are 34 459 425 possible trees!

The closest tree and parsimony methods evaluate functions which are $O(n)$ for each tree, with the possibility of eliminating some trees by branch and bound methods. The optimality functions used in maximum likelihood methods are at least of this order. We contend that the dominating factor in phylogenetic reconstruction methods that search for an optimal tree is the number of possible trees to check, even with the potential savings of branch and bound methods, and hence the Hadamard conjugation method is not prohibitively complex.

CONCLUSION

We have found Hadamard conjugation to be a very useful tool with which edge probabilities **p** and/or edge lengths **q** can be related to partition probabilities **s**. This has allowed us to perform both theoretical and empirical analyses, with both hypothetical and real data. These have already given us a deeper understanding of the relationships between evolutionary trees and the homologous character sequences from the taxa they relate. It is likely that further questions may be answered using these relationships.

REFERENCES

Cavender, J. A. 1978: Taxonomy with confidence. *Mathematical biosciences 40*: 271–280.

Hendy, M. D. 1991: A combinatorial description of the closest tree algorithm for finding evolutionary trees. *Discrete mathematics 96*: 51–58.

Hendy, M. D.; Penny, D. 1989: A framework for the quantitative study of evolutionary trees. *Systematic zoology 38*: 297–309.

Hendy, M. D.; Penny, D. 1993: Spectral analysis of phylogenetic data. *Journal of classification 10*: 5–23.

Felsenstein, J. 1978: Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology 27*: 401–410.

Felsenstein, J. 1981: Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution 17*: 368–376.

Kimura, M. 1981: Estimation of evolutionary sequences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences U.S.A. 78*: 454–458.

Lockhart, P. J.; Howe, C. J.; Bryant, D. A.; Beanland, T. J.; Larkum, A. W. D. 1992: Substitutional bias confounds inference of cyanelle origins from sequence data. *Journal of molecular evolution 34*: 153–162.

Saitou, N.; Nei, M. 1987: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution 4*: 406–425.

Sokal, R. R.; Michener, C. D. 1955: A statistical method for evaluating systematic relationships. *University of Kansas science bulletin 28*: 1409–1438.

Steel, M. A. 1988: Distribution of the symmetric difference metric on phylogenetic trees. *SIAM journal of discrete mathematics 1*: 541–551.

Steel, M. A.; Hendy, M. D.; Székely, L. A.; Erdös, P. L. 1992: Spectral analysis and a closest tree method for genetic sequences. *Applied mathematics letters 5*: 63–67.

Steel, M. A.; Székely, L. A.; Erdös, P. L.; Waddell, P. J. 1993: A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *New Zealand journal of botany 31*: 289–296 (this issue).

Steel, M. A.; Hendy, M. D.; Penny, D. in press: Parsimony can be consistent! *Systematic biology*.