Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 3339–3343, April 1994
Evolution

A discrete Fourier analysis for evolutionary trees

M. D. HENDY*, D. PENNY†, AND M. A. STEEL‡

*Department of Mathematics and †Molecular Genetics Unit, Massey University, Palmerston North, New Zealand; and ‡Department of Mathematics, University of Canterbury, Christchurch, New Zealand

Communicated by Roy J. Britten, October 25, 1993

ABSTRACT **Discrete Fourier transformations have recently been developed to model the evolution of two-state characters (the Cavender/Farris model). We report here the extension of these transformations to provide invertible relationships between a phylogenetic tree *T* (with three probability parameters of nucleotide substitution on each edge corresponding to Kimura's 3ST model) and the expected frequencies of the nucleotide patterns in the sequences. We refer to these relationships as spectral analysis. In either model with independent and identically distributed site substitutions, spectral analysis allows a global correction for all multiple substitutions (second- and higher-order interactions), independent of any particular tree. From these corrected data we use a least-squares selection procedure, the closest tree algorithm, to infer an evolutionary tree. Other selection criteria such as parsimony or compatibility analysis could also be used; each of these criteria will be statistically consistent for these models. The closest tree algorithm selects a unique best-fit phylogenetic tree together with independent edge length parameters for each edge. The method is illustrated with an analysis of some primate hemoglobin sequences.**

Spectral analysis, by which Fourier transformations allow invertible calculations between scientific models and their predicted data, has many applications in science (1). As a *predictive* tool it allows the generation of expected data from a model. As a *deductive* tool it both allows the parameters for the model to be estimated from observed data and can test the applicability of the model. We previously developed a spectral analysis, using a Hadamard (discrete Fourier) transformation, to analyze the evolution of genetic sequences of two-state characters under the Cavender/Farris model (2). Recent results (3, 4) allow the extension of this analysis to sequences of four-state characters evolving under Kimura's 3ST (three-substitution-type) model (5) of DNA and RNA evolution. Similar extensions to models of more character states are possible.

These Hadamard spectral analyses are based on the invertible relationship between a model of sequence evolution encoded in a vector, the *probability spectrum* $\mathbf{p}(T)$ (see Fig. 1), and properties of the consequent set of sequences, the *expected sequence spectrum* $\mathbf{s}(T)$. The model is a phylogenetic tree T with independent probabilities of nucleotide substitutions on each edge. The relation between $\mathbf{p}(T)$ and $\mathbf{s}(T)$ is described by vector functions called *Hadamard conjugations* (Eqs. **1** and **2**). An intermediate vector is the *edge lengths spectrum* $\mathbf{q}(T)$, whose positive values are additive parameters on the edges of T. $\mathbf{q}(T)$ can be calculated either from $\mathbf{s}(T)$ or from $\mathbf{p}(T)$, by applying the appropriate Hadamard conjugation (4).

As a deductive tool we express the sequence data D, as the *observed sequence spectrum* $\mathbf{s}(D)$ (the relative frequencies of character patterns), which we use as an estimate of $\mathbf{s}(T)$, for an unknown tree T. Applying the inverse Hadamard conjugation to $\mathbf{s}(D)$, we derive the *conjugate spectrum* $\gamma(D)$, which is an estimate of $\mathbf{q}(T)$. This inversion globally corrects for all multiple and parallel substitutions which are included in $\mathbf{s}(D)$. $\gamma(D)$ converges to $\mathbf{q}(T)$ as the sequence lengths grow, hence many selection procedures from $\gamma(D)$ will be statistically consistent (6).

A selection criterion can be used to determine the tree T with $\mathbf{q}(T)$ best approximating $\gamma(D)$; for example, a least-squares criterion gives the *closest tree* (7). This identifies a unique phylogenetic tree T and provides an estimate of the edge length spectrum $\mathbf{q}(T)$. Fig. 1 summarizes the relationships between the spectra. These spectra can be displayed as histograms. In Fig. 2 we display some spectra derived from a set of four primate hemoglobin pseudogene DNA sequences.

For two-state character sequences $\mathbf{q}(T)$ contains one parameter for each edge of T. For four-state character sequences $\mathbf{q}(T)$ contains three Kimura parameters (5) for each edge, corresponding to Kimura's 3ST model of evolutionary distance between sequences, but not constrained by a molecular clock. $\mathbf{q}(T)$ also contains many zero values, which relate to partitions which are not edge bipartitions of T and so the corresponding components of $\gamma(D)$ have expected value zero. Some of these are dependent on T, and hence are *invariants* as defined by Cavender and Felsenstein (9) and Fu and Li (10), and some are independent of T, and hence are invariants of the model. The remaining components of the conjugate spectrum, other than γ_0, have positive expected values, identifying the edges of T. Under a Poisson process these values can be interpreted as expected numbers of nucleotide substitutions on each edge. We refer to them as *edge lengths*. From the positive values of $\mathbf{q}(T)$ we can determine the probability spectrum $\mathbf{p}(T)$. The determination of the probability spectrum $\mathbf{p}(T)$ from $\mathbf{s}(T)$ is independent of the process of substitution with the edge lengths being interpreted as additive parameters.

HADAMARD CONJUGATION

The main tools in relating the spectra are the Hadamard conjugations (Eqs. **1** and **2**). (In mathematics conjugation is the successive application of three transformations, with the third being the inverse of the first.) For each power $m = 2^t$ of 2, we define an $m \times m$ Hadamard matrix $H = H_t$. The conjugations relate two vectors $\mathbf{x}$ and $\mathbf{y}$ of m components:

$$\mathbf{y} = H^{-1}(\ln(H\mathbf{x})), \qquad [1]$$

provided $H\mathbf{x} > 0$, and its inverse

$$\mathbf{x} = H^{-1}(\exp(H\mathbf{y})). \qquad [2]$$

The natural logarithm (ln) and the exponential (exp) functions are applied to each component of the vectors $\mathbf{x}$ and $\mathbf{y}$ separately. The Hadamard matrices are symmetric orthogonal matrices with entries 1 and -1 so that $H^{-1} = (1/m)H$.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

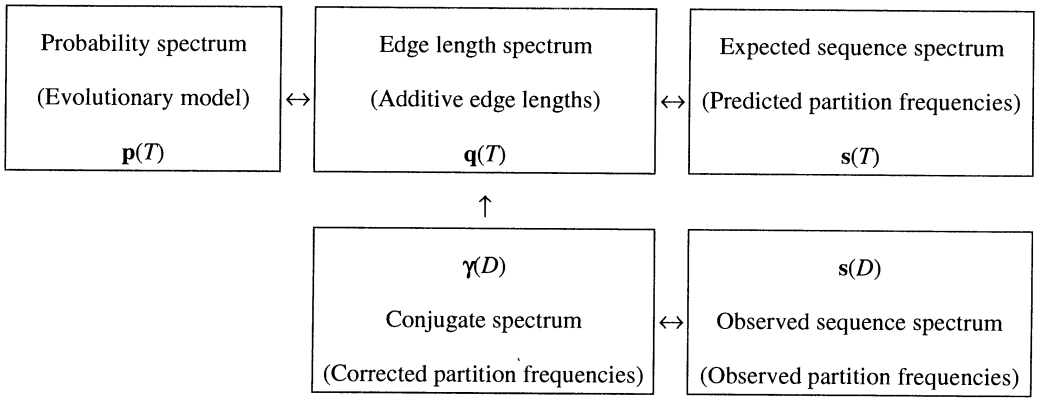

FIG. 1. The interrelationships between the various spectra. The relationships represented by the double arrows are invertible Hadamard conjugations. Thus the spectra on the same level are equivalent, as they represent the same information. Those on the upper row are dependent on a tree T, while those on the bottom are derived from a set of sequence data D. The mapping from $\mathbf{s}(D)$ to $\gamma(D)$ corrects for all parallel, multiple, and higher-order substitutions. The data fit a model exactly if $\gamma(D) = \mathbf{q}(T)$. The vertical arrow can represent any fitting algorithm. We advocate a least-squares best fit, the closest tree algorithm, which selects the tree T for which the distance $|\gamma(D) - \mathbf{q}(T)|$ is minimal.

They can be defined recursively, using the Kronecker product of matrices:

$$\mathrm{H}_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathrm{H}_2 = \mathrm{H}_1 \otimes \mathrm{H}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

$$\mathrm{H}_t = \mathrm{H}_1 \otimes \mathrm{H}_1 \otimes \mathrm{H}_1 \otimes \ldots \otimes \mathrm{H}_1 \ (t \text{ terms}) = \mathrm{H}_1 \otimes \mathrm{H}_{t-1}$$

$$= \begin{pmatrix} \mathrm{H}_{t-1} & \mathrm{H}_{t-1} \\ \mathrm{H}_{t-1} & -\mathrm{H}_{t-1} \end{pmatrix}.$$

In a data set D of n sequences, there are 4^{n-1} possible patterns of nucleotides at any position (site) relative to the nth sequence. (In the appendix we describe their indexing.) We refer to these patterns as *quadripartitions*. We construct the observed sequence spectrum $\mathbf{s}(D)$ where the entry s_k is the proportion of sites in the sequence with quadripartition Q_k.

In Table 1, following Evans and Speed (3), we define a binary coding of the four nucleotides by elements of the Klein 4-group. The code for a nucleotide substitution is then taken as the difference (modulo 2) of the codes for the two nucleotides. This classifies the substitutions into three types, with type I being the transitions, type II being the transversions A $\leftrightarrow$ C and G $\leftrightarrow$ T, and type III being the transversions A $\leftrightarrow$ T and C $\leftrightarrow$ G.

The sequence evolution model is described by a tree T, with three probability parameters for each edge. These parameters are the probabilities of substitutions of types I, II, and III, acting independently and identically on all sites. The edges e_i of a tree T are labeled so their corresponding edge bipartition is B_i. (The indexing conventions for edges, bipartitions, and quadripartitions are given in the appendix.) For each edge e_i of T, we specify three probabilities p_1^i, p_2^i, and p_3^i [equivalent to Kimura's (5) parameters P, Q, and R] of the nucleotide substitutions of types I, II, and III, respectively, along e_i. Then $p_0^i = 1 - p_1^i - p_2^i - p_3^i$ is the probability of no substitution along e_i. Provided the probabilities of substitution are not large (it is sufficient to require $p_j^i < 0.25$, for $j > 0$), we can calculate expected sequence $\mathbf{s}(T)$ (4).

For each edge, e_i, let $\mathbf{p}^i = \begin{pmatrix} p_0^i \\ p_1^i \\ p_2^i \\ p_3^i \end{pmatrix}$. We calculate $\mathbf{E}^i = \begin{pmatrix} E_0^i \\ E_1^i \\ E_2^i \\ E_3^i \end{pmatrix}$, where

$$\mathbf{E}^i = \mathrm{H}^{-1}(\ln(\mathrm{H}\mathbf{p}^i)) \qquad [3]$$

with $\mathrm{H} = \mathrm{H}_2$. The entries E_j^i are additive parameters, with $E_0^i = -(E_1^i + E_2^i + E_3^i)$ the only negative value. Under a Poisson process of substitution along edge e_i, $-E_0^i$ is the expected total number of nucleotide substitutions, and E_j^i, for $j = 1, 2$, or 3, are the expected numbers of substitutions of type j. The Hadamard conjugation (Eq. 3) is equivalent to the formulae derived by Kimura (5). For example, if $p_1^i = 0.2$, $p_2^i = p_3^i = 0.05$, we find $E_1^i = 0.2908$, $E_2^i = E_3^i = 0.0558$, which under a Poisson model, would be the expected numbers of substitutions along edge e_i. We can invert Eq. 3 to recover $\mathbf{p}^i$,

$$\mathbf{p}^i = \mathrm{H}^{-1}(\exp(\mathrm{H}\mathbf{E}^i)). \qquad [4]$$

We form the *edge length spectrum* $\mathbf{q}(T)$ (Fig. 2) from these E_j^i parameters, where for each edge e_i of T, $q_i = E_1^i$, $q_{mi} = E_2^i$, $q_{i+mi} = E_3^i$, all the other components of $\mathbf{q}(T)$ are set to zero except for q_0, which is set to -1 times the sum of the E_j^i values over all edges of T. $\mathbf{q}(T)$ contains all the information, as its positive values identify the edges of T and from these the edge substitution probabilities of each type can be determined. If we apply the conjugation (Eq. 2) using the Hadamard matrix $\mathrm{H} = \mathrm{H}_{2(n-1)}$ it can be shown (4) that the observed sequence spectrum is

$$\mathbf{s}(T) = \mathrm{H}^{-1}(\exp(\mathrm{H}\mathbf{q}(T))), \qquad [5]$$

with kth component, s_k, the expected frequency of the kth quadripartition Q_k occurring in the generated sequences. If the sequence is of length c, then cs_k will be the expected number of sites with quadripartition Q_k. This can be used in simulation studies to generate sample sequences (11).

APPLICATIONS

Inferring Trees from Sequence Data. Some phylogenetic tree building methods, such as parsimony and compatibility analysis, attempt to find a tree T whose edge bipartitions best match these site-bipartitions. However, an exact match is not usually possible, as some pairs of the site bipartitions are *incompatible*, that is, they cannot exist together on the same tree T. (For example, bipartition 3 which groups taxa 1 and

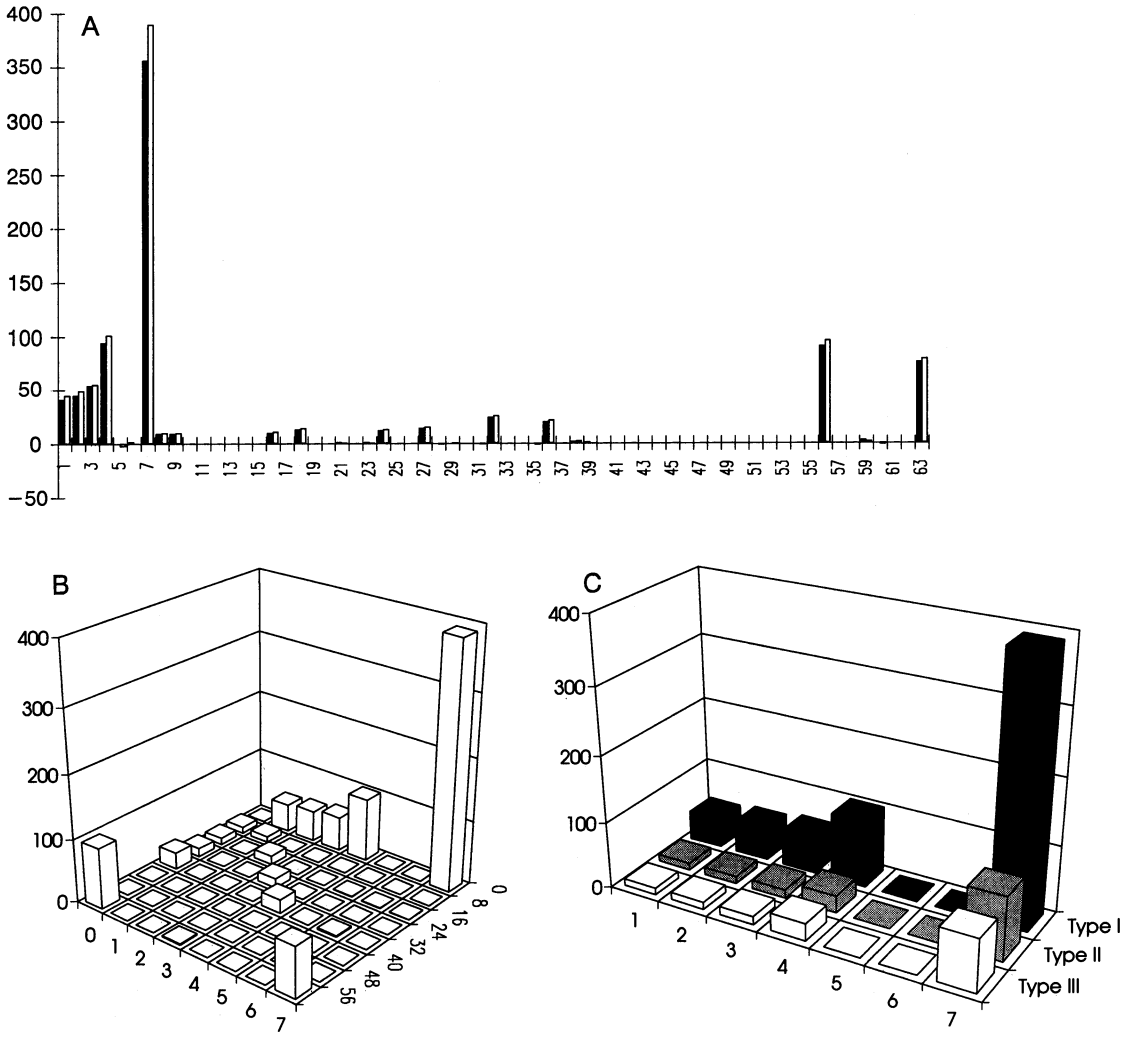

FIG. 2. Histograms representing spectra for sequences from four primate DNA sequences (8). (*A*) The observed sequence spectrum $\mathbf{s}(D)$ (black) and the conjugate spectrum $\gamma(D)$ (white), from sequences of four primates. [The values of $\mathbf{s}(D)$ are given in Table 2.] These spectra have been multiplied by the sequence length, $c = 9864$. Their 0th entries, $cs_0 = 8988$ and $c\gamma_0 = -926.9$, are not displayed. The partition indices are described in the appendix. The values of $\gamma(D)$ can be interpreted under a Poisson process as expected numbers of nucleotide substitutions of each type. They are generally, but not always, greater than the observed numbers of substitutions. (*B*) For convenience we display $\gamma(D)$ as an $m \times m$ array. The components of $\gamma(D)$ run in columns (top to bottom) beginning at γ_0 (replaced by 0), top left. Only the values which lie on the leading row, column and diagonal (radiating from γ_0) relate to bipartitions; the remaining entries are model invariants with expected values zero. (*C*) The leading row, column, and diagonal of $\gamma(D)$ (excluding γ_0) are displayed as three rows. These values are the expected numbers of transitions (type I) and the two types (II and III) of transversions. These are used to select the best fit tree T, when two corresponding entries in the leading row, column, and diagonal are invariants for T with expected value zero. We show in Fig. 3 the tree T selected by the closest tree algorithm.

2 together is incompatible with bipartition 6 which groups 2 and 3.) These methods penalize each T with some penalty for each rejected bipartition, to select the phylogenetic tree with smallest penalty. This approach can lead to the problem of statistical inconsistency (12, 13). We can readily show that, assuming a molecular clock, parsimony will always be consistent for four taxa, but it can be inconsistent for five or more taxa. However, if parsimony was applied to the "corrected" $\gamma(D)$ spectrum, parsimony would be consistent in all cases.

Table 1. Klein 4-group coding of the four nucleotides and the nucleotide substitutions

Nucleotide	Code	Substitutions		Type	Chemical nature
A	(0, 0)	—	—		
G	(1, 0)	A ↔ G	C ↔ T	I	Transitions
C	(0, 1)	A ↔ C	G ↔ T	II	Transversions
T	(1, 1)	A ↔ T	C ↔ G	III	Transversions

However, we prefer the closest tree criterion, which is faster to calculate and generally produces a unique tree.

With a data set D of n sequences, we can estimate the sequence spectrum $\mathbf{s}(D)$ by the relative frequencies of occurrence of each pattern, provided evolution at each site is approximately independent and identically distributed. We invert Eq. 5 to calculate the conjugate spectrum,

$$\gamma(D) = \mathbf{H}^{-1} \ln(\mathbf{Hs}(D)). \quad [6]$$

When the sites evolve at different rates according to some known distribution across the sites this is no longer valid. However, Eq. 6 can be modified to compensate for this, by replacing the logarithm function by φ, the functional inverse of the moment-generating function for the distribution (14).

$$\gamma(D) = \mathbf{H}^{-1} \varphi(\mathbf{Hs}(D)). \quad [7]$$

The components corresponding to edges of the (unknown) generating tree have expected positive values, while the

Table 2. Frequencies of the quadripartitions from the 9879 sites of four primate hemoglobin ψ-pseudogenes

	0	8	16	24	32	40	48	56
0	**8988**	**9**	**10**	**12**	**24**			**90**
1	**41**	**9**						
2	**45**		**13**					
3	**54**			**14**				3
4	**94**				**20**			
5			1					
6	**2**				2			
7	**356**		1		1			**75**

The values of the observed sequence spectrum $\mathbf{s}(D)$ can be obtained by dividing by 9879. The blank entries indicate that the corresponding quadripartitions did not occur. Bold numbers indicate the values of the bipartitions.

remaining components, other than γ_0, have expected value 0. Under Kimura's model, these γ_{ji} represent the expected number of nucleotide substitutions of each class along edge e_i.

Analysis of Some Primate DNA Sequences. Hemoglobin ψ-pseudogene DNA sequences (8) for the four primate species human, chimpanzee, orangutan, and rhesus monkey of length 9879 were analyzed. The relative frequencies of the observed quadripartitions are given in Table 2. For the convenience of presentation this and subsequent vectors of 64 components are displayed as 8 × 8 arrays. In Table 3 we give the conjugate spectrum $\gamma(D)$, a vector of 64 entries, derived by using Eq. 5 from $\mathbf{s}(D)$ of Table 2. Only 21 values γ_k [where $k = i, mi$, and $(m + 1)i, i = 1, 2, \ldots, 7$] can relate to edge bipartitions and so assist with the determination of the underlying tree. Apart from γ_0, the remaining 42 values are "model invariants" with expected value 0. These range in value from $\gamma_{31} = -0.00008$ to $\gamma_{59} = 0.00020$, with mean -0.00002 and standard deviation 0.00007. From these 21 values we need to select those which relate to a specific tree—that is, select five consistent edge bipartitions. A simple way of selecting a tree is the *closest tree* selection (7) procedure, which finds, for each tree T, the point $\mathbf{g}(T)$ in the set of possible tree spectra for T which is closest to $\gamma(D)$. The distance $\Delta(D, T)$ from $\mathbf{g}(T)$ to $\gamma(D)$ is given as the relative error of fit. With $T_3 = (12)(34)$, $\Delta^2(D, T_3) = 5.1 \times 10^{-7}$, $T_6 = (13)(24)$, $\Delta^2(D, T_6) = 3.7 \times 10^{-5}$, $T_5 = (14)(23)$, $\Delta^2(D, T_5) = 3.8 \times 10^{-5}$. Hence the tree $T_3 = (12)(34)$, as shown in Fig. 3, fits D best. With this tree the optimal edge length parameters are given by $E_j^i = \gamma_k - \gamma_{T_3}$, where for T_3, $\gamma_{T_3} = 3.17 \times 10^{-8}$ (hence for these values this correction is negligible), and $k = i, 8i$, and $9i$ for $j = 1, 2$, and 3, $i = 1, 2, 3, 4$, and 7. Under a Poisson process these values represent the expected numbers of nucleotide substitutions of each of the three types. In Fig. 4 we show the uniformity of the ratios of these numbers on each edge of T.

Table 3. Scaled values of the conjugate spectrum $\gamma(D)$ obtained from the observed sequence spectrum

	0	8	16	24	32	40	48	56
0	**−926.9**	**9.8**	**10.9**	**12.9**	**26.0**	**−0.1**	**−0.1**	**95.7**
1	**44.7**	**9.8**	−0.1	−0.2	−0.2	0.0	−0.1	−0.4
2	**49.2**	−0.1	**14.2**	−0.1	−0.1	−0.1	0.0	−0.5
3	**55.0**	−0.1	−0.1	**15.1**	−1.2	0.0	0.0	1.9
4	**100.9**	−0.1	−0.1	−0.9	**21.6**	0.0	0.0	−1.6
5	**−2.4**	−0.1	0.5	0.0	−0.1	**−0.1**	0.0	−0.4
6	**−0.1**	−0.4	−0.1	0.0	2.1	0.0	**−0.1**	−0.3
7	**389.6**	−0.4	0.7	−0.9	−0.2	−0.1	−0.1	**78.5**

The values are multiplied by 9879, the number of sites, so are comparable to those of Table 2. Those values which relate to potential edge bipartitions, the entries of the left column, the top row and the leading diagonal are shown in bold. They are γ_k, where $k = i, 8i, 9i, i = 1, 2, \ldots, 7$.

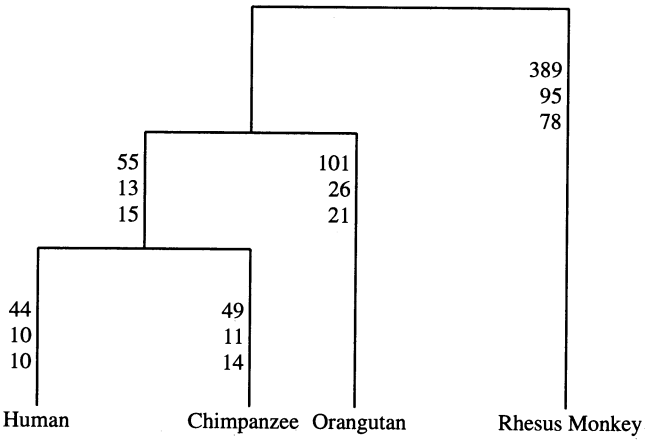

FIG. 3. Closest tree from the conjugate spectrum (Fig. 2) derived from the hemoglobin ψ-pseudogene DNA sequences (8) for the four primate species human, chimpanzee, orangutan, and rhesus monkey. The three length parameters on each edge are the expected numbers of substitutions of types I, II, and III. The tree has been arbitrarily rooted on the edge e_7, the pendant edge to rhesus monkey. This analysis is carried out without a molecular clock assumption. The sums of the expected numbers of substitutions on each edge as a single edge length parameter fit the molecular clock hypothesis with a χ^2 value of 0.65 on 4 degrees of freedom, better than a 95% confidence level.

For each additional taxon added, the number of components in the spectra is increased by a factor of 4. When we add the hemoglobin ψ-pseudogene sequence for gorilla to the set of four primate sequences above, the spectra each have 256 components. However, the number of components relating to bipartitions and hence influencing the choice of closest tree increases only from 21 to 45 [$= 3 \times (2^{n-1} - 1)$]. The closest tree from these data is given in Fig. 5, where the relative numbers of substitutions on the edge e_4 to gorilla differ from those of the other edges. This either suggests that a different process of substitution occurs on this edge or points to a possible error in the data.

COMPUTATIONAL COMPLEXITY

The matrices involved in this spectral analysis are of exponential order in the number of taxa n being analyzed. This means that direct computations for all but small values of n are impractical. For $n = 10$, the Hadamard matrix for the 4-state character analysis has $2^{36} = 6.87 \times 10^{10}$ entries, and the two matrix vector products would each require 2^{36}

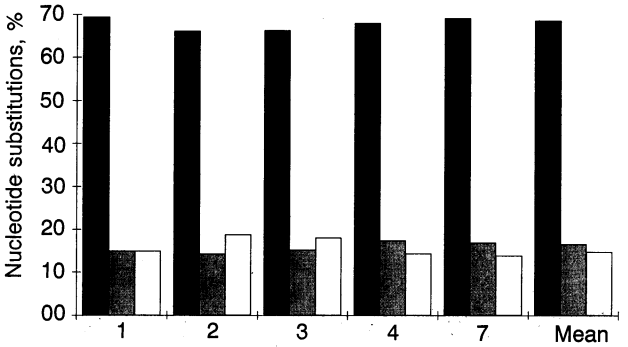

FIG. 4. Relative numbers of nucleotide substitutions of type I (dark), type II (light), and type III (white) for each edge of the closest tree T of Fig. 3 and the relative numbers overall. This illustrates that the ratios are similar on each edge, with a maximum deviation of less than 3% from the mean values, suggesting a common process for nucleotide substitutions on all edges.

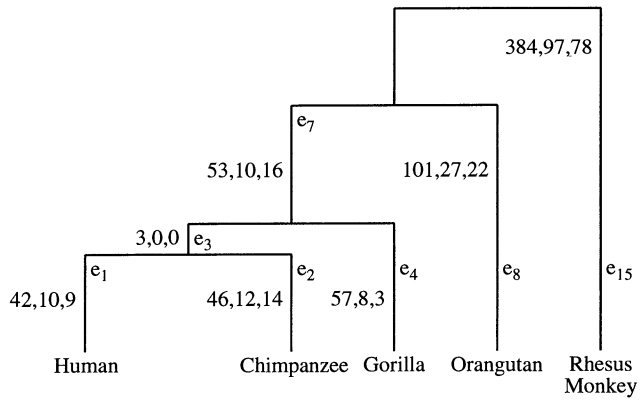

FIG. 5. Closest tree for the hemoglobin ψ-pseudogene sequences of five primates. This tree is only 17–20% closer than trees with the edge to gorilla e_4 joined to e_1 or e_2 or with a trichotomy. The expected numbers of substitutions of each type are listed with each edge. The relative numbers of substitutions on e_4 (to gorilla) differ from those on the other edges. This could result from a different process of nucleotide substitution on this edge or from errors in the gorilla sequence.

additions or subtractions. However, the fast Hadamard transformation (7) reduces each multiplication from $O(2^{4n-4})$ to $O(n2^{2n-2})$, so for $n = 10$ they each require 5.2×10^6 additions or subtractions. For the conjugations (Eqs. **5** and **6**) there are also 2^{2n-2} evaluations of the exponential or logarithm functions required. There are a number of further means of reducing the complexity, such as quadratic approximation of the conjugate spectrum (15), or by calculating only those terms of the conjugate spectrum that discriminate between the trees (unpublished results), reducing the complexity to $O(n2^{n-1})$. Using these approaches means that computation of the conjugate spectrum for $n = 20$ or more 4-state character sequences is feasible on a personal computer. In addition, the complexity of sampling may serve as a means of estimating components of the conjugate spectrum to predetermined accuracy. The application of the closest tree algorithm potentially requires the examination of all the binary trees that can link n taxa. The conjugation is computed once, and then the fitting is done to the conjugate spectrum for each potential binary tree. However, using a branch and bound algorithm (16), the closest tree algorithm can generally be computed in much less time than that for the conjugation. The least-square fits are more discriminating than are linear measures, such as those used for parsimony or compatibility analyses, and as the data are already corrected for multiple and parallel substitutions, the optimal fit may be obtained more easily.

Conclusion. We have shown that spectral analysis developed for the Cavender/Farris model of the evolution of two-state character sequences can be extended to Kimura's 3ST model of four-state character sequences. The invertible relationships for both models allow spectral analysis to be used as either a deductive or a predictive tool, a feature previously not available for any evolutionary model. Extensions to more general models with the relaxation of some of the assumptions of these models are now being developed (see refs. 14 and 17, for example), although theoretical barriers to such extensions are also being realized.

APPENDIX: PARTITION INDICES

Suppose N is a set of n taxa. The components of the spectra are identified by the different ways N can be slit into two disjoint subsets. A *split*, or *bipartition*, of N is a pair of disjoint subsets (A, B) of N, where each taxon of N belongs either to A or to B. Each edge of the phylogenetic tree T for the taxa of N defines a split, the monophyletic group (below the edge) and its complement. If T were an unrooted tree where the location of the ancestral root is not specified, the subsets of taxa labeling the leaves at either side of an edge e still define a bipartition, the *edge bipartition* of e. There are $m = 2^{n-1}$ possible bipartitions, including the trivial pair $\{\varnothing, N\}$. These bipartitions are identified by the subset not containing n and are indexed as B_i in ref. 1, where $B_0 = \varnothing$, $B_1 = \{1\}, B_2 = \{2\}, B_3 = \{1, 2\}, B_4 = \{3\}, \ldots, B_{m-1} = \{1, 2, \ldots, n-1\}$.

With nucleotide sequences there are four characters; the set of nucleotides (characters) at one site defines a *site partition* of up to four subsets of the taxa. We will refer to any partition into four or fewer subsets as a *quadripartition*. To apply the spectral analysis we now need to identify the quadripartitions in a systematic way. Each quadripartition is identified as a pair of bipartitions. This is done by coding each nucleotide as a pair of binary digits as in Table 1, and then identifying separately the bipartitions formed by the first component and by the second component. This pair (B_i, B_j) of bipartitions corresponds to the quadripartition Q_{i+mj}.

For example, with $n = 5$ ($m = 16$), if the nucleotides at a site are CACTT, then the corresponding codes are 00011 for the first component and 10111 for the second; 00011 is bipartition B_7 and 10111 is bipartition B_2. (The index for the bipartition B_i is calculated as $i = 2^{a-1} + 2^{b-1} + 2^{c-1} + \ldots$, where taxa $a, b, c, \ldots$ have the component which differs from that of taxon n.) Thus CACTT gives quadripartition Q_{39}. ($39 = 7 + 2m$.) The quadripartitions are relative patterns; there are four patterns of characters for each. The site patterns GTGAA, TGTCC, and ACAGG also give quadripartition Q_{39}. The pattern of nucleotides at a site for a particular quadripartition is completely specified once the nucleotide of one taxon is given.

1. Bracewell, R. N. (1989) *Sci. Am.* **260** (6), 62–69.
2. Hendy, M. D. & Penny, D. (1993) *J. Classif.* **10**, 5–24.
3. Evans, S. N. & Speed, T. P. (1993) *Ann. Stat.* **21**, 355–377.
4. Steel, M. A., Hendy, M. D., Székely, L. A. & Erdös, P. L. (1992) *Appl. Math. Lett.* **5**, 63–67.
5. Kimura, M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 454–458.
6. Steel, M. A., Hendy, M. D. & Penny, D. (1993) *Syst. Biol.* **42**, 581–587.
7. Hendy, M. D. (1991) *Discrete Math.* **96**, 51–58.
8. Miyamoto, M., Koop, B. F., Slightom, J. L., Goodman, M. & Tennant, M. R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7627–7631.
9. Cavender, J. A. & Felsenstein, J. (1978) *J. Classif.* **4**, 57–71.
10. Fu, Y.-X. & Li, W.-H. (1992) *Math. Biosci.* **109**, 201–228.
11. Charleston, M. A., Hendy, M. D. & Penny, D. (1994) *Comp. Biol.* **1**, in press.
12. Felsenstein, J. (1978) *Syst. Zool.* **27**, 401–410.
13. Hendy, M. D. & Penny, D. (1989) *Syst. Zool.* **38**, 289–303.
14. Steel, M. A., Székely, L. A., Erdös, P. L. & Waddell, P. J. (1993) *N. Z. J. Bot.* **31**, 289–296.
15. Székely, L. A., Steel, M. A. & Erdös, P. L. (1993) *Adv. Appl. Math.* **14**, 200–216.
16. Hendy, M. D. & Penny, D. (1982) *Math. Biosci.* **59**, 277–290.
17. Steel, M. A. (1993) Research Report No. 103 (Department of Mathematics and Statistics, Canterbury University, Christchurch, New Zealand).