Physics 498BIN, Fall 2001

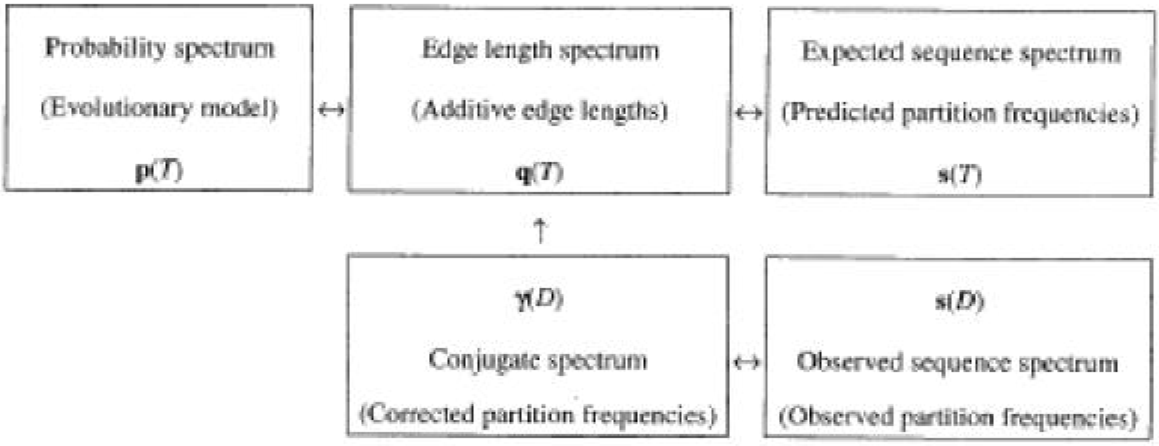
Discrete Fourier analysis for phylogenetic trees

MARTIN PH. STEHNO
Department of Physics, University of Illinois at Urbana-Champaign

ABSTRACT Discrete Fourier transformations (DFTs) provide a useful tool to assign a phylogenetic tree (PGT) to an observed frequency of nucleotide patterns in DNA sequences of species. The advantage of this sort of spectral analysis is that it allows global correction for multi-substitution processes [1].

SPECTRAL ANALYSIS OF PGTS

Two spectras characterize a PGT, the *probability spectrum* p(T), and the *expected sequence spectrum* s(T). After labelling the edges of the tree in an appropriate way, they are can be related by two steps of transforms using vector functions called *Hadamard conjugations*. The intermediate vector is called the *edge length spectrum*. The transformation scheme is given in Fig. 1.


Probability spectrum
(Evolutionary model)
p(T)
↔
Edge length spectrum
(Additive edge lengths)
q(T)
↔
Expected sequence spectrum
(Predicted partition frequencies)
s(T)
↑
γ(D)
Conjugate spectrum
(Corrected partition frequencies)
↔
s(D)
Observed sequence spectrum
(Observed partition frequencies)

(Hendy et al., Proc. Natl. Acad. Sci. USA 91 (1994))

Fig.1. Scheme of transformations.

This scheme can be used in two ways. Starting with a *probability distribution* we can calculate the *edge length spectrum* and the *expected sequence spectrum*. On the other hand, given a data set D, we can take the *observed sequence spectrum* s(D) (the relative frequencies of character patterns) as an estimate for s(T). From this we calculate a *conjugate spectrum* γ(D) (the 'corrected partition frequencies') [1, 4]. This will correct for all parallel, multiple, and higher order substitutions. We find the corresponding tree, that is the tree for which $|\gamma(D) - q(T)|$ is minimal, using a fitting algorithm (e.g. least-squares best fit or 'closest tree algorithm'). Having found the correct tree one is able to reconstruct the *probability spectrum* and *expected sequence spectrum*.

HADAMARD CONJUGATION

A conjugation consists of three transformations that are successively applied. The third transformation is the inverse of the first. The m m – Hardamard matrix H_t is defined as

$$H_t = H_1 \otimes H_{t-1},$$

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The Hadamard transform[1] is

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

and its inverse

$$\mathbf{x} = H^{-1} (\exp(H\mathbf{y})), \qquad H\mathbf{x} > 0.$$

In order to apply the Hadamard transform to our problem, we have to label the tree edges e_i. We define a binary coding of the four nucleotides by elements of the Klein 4-group, see Table 1.

Nucleotide	Code	Substitution	
A	(0, 0)	-	-
G	(1, 0)	A-G	C-T
C	(0, 1)	A-C	G-T
T	(1, 1)	A-T	C-G

Table 1. Binary coding

The substitutions have been chosen to agree with the definition of the Kimura 3ST (or 'three substitution-type') model. This model has 3 parameters: p_1 is the rate of *transitions* of type A-G, T/U-C, p_2 the rate of *transversions* of type A-T/U, G-C, and p_3 the rate for *transversions* of type A-C, T/U-G [2].

Now suppose, we have a set N of n *taxa* (sequence positions, which are compared). A *bipartition* is a pair of disjoint subsets (A, B) of N. Thus a *taxon* belongs either to subset A or B. The edges of a simple PGT T for the *taxa* N with only two characters define a *bipartition*, the *edge bipartition* e. Thus each edge is fully described by the labels of the subgroups it belongs to. There are $m = 2^{n-1}$ possible *bipartitions*. Since there are four nucleotides, we have to use *quadripartitions* rather than bipartitions (four subsets for each site), but employing the scheme of Table 1, we can break the *quadripartition* Q_{i+mj} down to a first and a second component of the pair (B_i, B_j) of *bipartitions*. The indices i, j for the bipartitions are

$$i, j = 2a-1 + 2b-1 + 2c-1 + \ldots,$$

where the taxa a, b, c, … have a component, which differs from corresponding taxon on sequence n.

SEQUENCE EVOLUTION MODEL

Three parameters p_1^i, p_2^i, and p_3^i, the parameters of the Kimura 3ST model, are assigned to each edge e_i of the tree T. Together with $p_0^i = 1- p_1^i - p_2^i - p_3^i$ they form the vector **p**. From **p** we can calculate the vector (in i space)

$$\mathbf{E} = H_2^{-1} (\ln(H_2\mathbf{p})),$$

with the zero-component $E_0 = - (E_1^i + E_2^i + E_3^i))$, which can be thought of as the negative of the expected value of the total number of substitutions along e_i, and the rest of the components correspond to the number of substitutions of type 1, 2, 3 respectively.

Analogously the transform

$$\mathbf{p} = H_2^{-1} (\exp(H_2\mathbf{E})),$$

[1] Logarithm and exponential are taken for each component separately.

will give us back the probability spectrum **p**.

We obtain the *edge length spectrum* **q**(T) from the E_j's using

$$
\begin{gathered}
q0 = - \Sigma_{i,j} E_j^{\,i} \\
q_i = E_1^{\,i}, \\
q_{mi} = E_2^{\,i}, \\
q_{i+mi} = E_3^{\,i}, \\
\text{and } q_j = 0 \text{ otherwise.}
\end{gathered}
$$

This spectrum contains all the information about the edge lengths and therefore the substitution probabilities of each type. Finally, we calculate the *expected sequence spectrum*

$$\mathbf{s}(T) = H_{2(n-1)}^{-1} (\exp(H_{2(n-1)}\mathbf{q}(T))).$$

The inverse of this equation is needed to obtain the *conjugate spectrum* γ(D) from the data set D,

$$\gamma(D) = H_{2(n-1)}^{-1} (\ln(H_{2(n-1)}\mathbf{s}(D))).$$

If the sites evolve at different rates, we must replace the logarithm by the functional inverse φ of the momentum-generating function for the distribution, thus

$$\gamma(D) = H_{2(n-1)}^{-1} \varphi((H_{2(n-1)}\mathbf{s}(D))).$$

DISCUSSION

The Hadamard conjecture is an alternative to more conventional methods of PGT construction. It has been found particularly useful for analysing the performance of different phylogenetic methods under suitable conditions. [2]

Recently a new idea for a dynamical picture of phylogenetics has been proposed [5]. It is based on an analogy between *quadripartition* and a momentum space representation of perturbation theory for a many-body quantum field theory on a hypercubic lattice. The quantum field theory formalism was introduced to solve the statistical master-equations for the time-evolution of the probability functions [6]. In this context the Hadamard inversion techniques could provide a position space formulation of the theory.

BIBLIOGRAPHY

[1] M.D. Hendy, D. Penny, M.A. Steel, A Discrete Fourier Analysis for Evolutionary Trees, Proceedings of the National Academy of Sciences of the United States of America, Vol. **91**, Issue 8 (Apr. 1994)

[2] M. Steel et al., Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results, Discrete Applied. Mathematics **88** (1998) 367-398

[3] D. Penny, Comparative Analysis of Signals in Sequences, virtual web lecture at the 1st Internet-extended Bioinformatics Conference (IEC-1), April 1998

[4] F. Rodriguez et al., The General Stochastic Model of Nucleotide Substitution, J. theor. Biol. (1990) **142**, 485-501

[5] P.D. Jarvis, J.D. Bashford, Quantum field theory of phylogenetic branching, arXiv:physics/0107047, Aug. 2001

[6] Masao Doi, Second quantization representation for classical many-particle systems, J Phys. A: Math. Gen., Vol. 9, No. 9, 1976