Hadamard Conjugation
an analytic tool for phylogenetics

Mike Hendy

Allan Wilson Centre

for Molecular Ecology and Evolution

Massey University
New Zealand

m.hendy@massey.ac.nz
http://awcmee.massey.ac.nz

Outline:

1. Theory

2. Applications

1. Theory

    - Hadamard Matrices
    - Diagonalising Stochastic Matrices
    - Hadamard conjugations

2. Applications

- Sequence Analysis
- Phylogenetic Invariants
- Proof of invertibility
- Maximum Parsimony
- Corrected Parsimony
- Maximum Likelihood

Hadamard Matrices (J. Hadamard, 1893)
Definition: An $n \times n$ matrix, $A = [a_{ij}]$, $a_{ij} \in \{-1, 1\}$, is *Hadamard* (order n)

$$\iff A^T A = nI.$$

Properties: It can be easily shown that:

1. $\forall B = [b_{ij}], n \times n$, with $b_{ij} \in [-1, 1]$,

$$|detB| \leq |detA| = n^{n/2}.$$

2. Rows (and columns) of A are orthogonal.

3. $A^{-1} = \frac{1}{n} A^T$.

4. Hadamard matrices of order n can only exist for $n = 1$, 2 or $4m | m \in \mathbb{Z}^+$.

5. $\exists$ Hadamard matrices of order n for:

   (a) $n = 1$, 2;

   (b) $2m$, if $\exists$ Hadamard matrix of order m;

   (c) $4m$ if $4m - 1$ is prime.

Hadamard Conjecture: $\exists$ Hadamard matrix order $n = 4m$, for all $m \in \mathbb{Z}$.
Currently known for $m = 1, 2, \cdots, 106$.
Does there exists a Hadamard matrix of order 428?

Sylvester Matrices

(A family of Hadamard matrices.)

Definition:

$$H_0 = [1]; \quad H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and for $n \geq 1$

$$H_{n+1} = H_1 \otimes H_n = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

Hence

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

$$H_3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix},$$

etc.

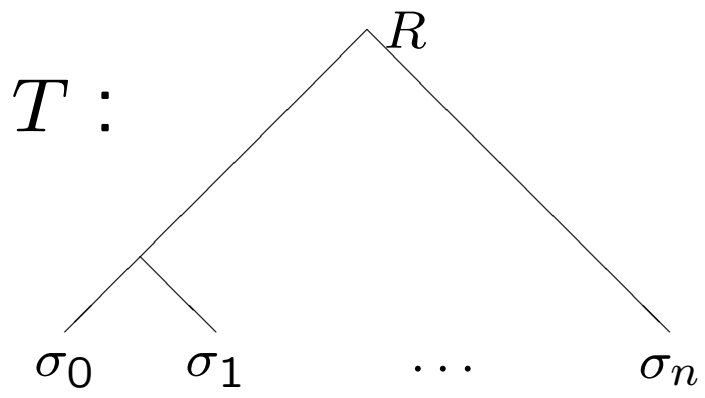
Substitution Models:
Given the set

$$\Sigma = \{\sigma_0, \sigma_1, \cdots, \sigma_n\}$$

of homologous aligned (without gaps) nucleotide sequences, the common problem of phylogeny is to discover the evolutionary relationship (as a phylogenetic tree) connecting these sequences.

The method (class of methods) called *Maximum Likelihood* requires us to propose a model of nucleotide substitutions across the edges of a putative phylogenetic tree $T = (V, E)$.

The set leaves $L \subseteq V$ at the tips of the T represents the $n + 1$ sequences, the remaining vertices are putative ancestors descending from the root R which represents the common ancestor of all the sequences.



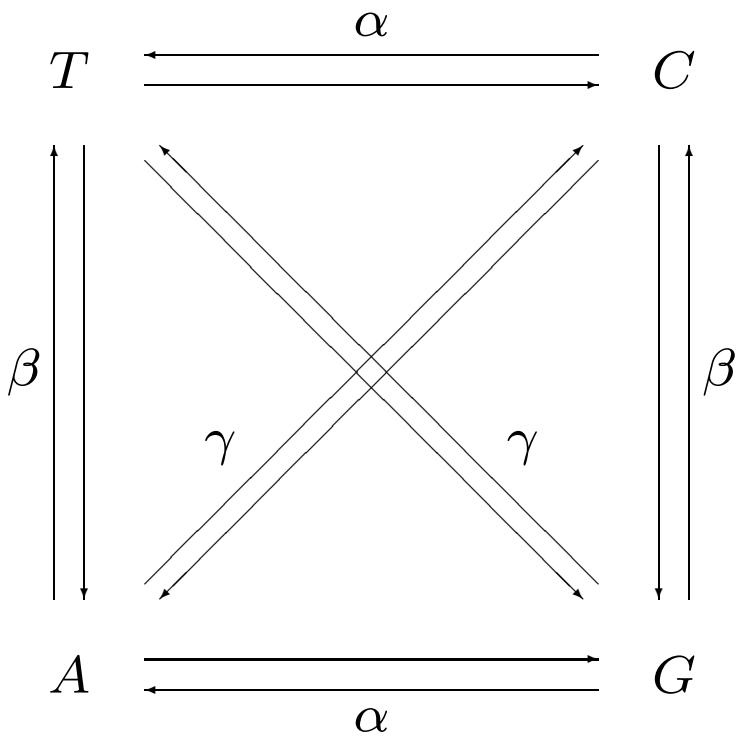
We need to specify the distribution π of nucleotides at the root R and some model parameters at each edge e of T.

These parameters are often described by a *rate matrix* Q (common to all edges) together with an edge length (time) t_e for each edge e.

The simplest substitution rate model for the four DNA (or RNA) nucleotides is that of Jukes and Cantor (JC), where each possible substitution is postulated to occur at the same common rate α.

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}.$$

Other rate matrices commonly used include Kimura's two substitution types model (K2ST) and the five parameter model of Hasegawa, Kishino and Yano (HKY85).



Kimura's 3-substitution type model.

The rate matrix for Kimura's 3ST model is

$$Q = \begin{bmatrix} -K & \alpha & \beta & \gamma \\ \alpha & -K & \gamma & \beta \\ \beta & \gamma & -K & \alpha \\ \gamma & \beta & \alpha & -K \end{bmatrix},$$

where $K = \alpha + \beta + \gamma$. The α substitutions are transitions, the β and γ substitutions are transversions.

The rate matrix for Kimura's 3ST model is

$$Q = \begin{bmatrix} -K & \alpha & \beta & \gamma \\ \alpha & -K & \gamma & \beta \\ \beta & \gamma & -K & \alpha \\ \gamma & \beta & \alpha & -K \end{bmatrix},$$

where $K = \alpha + \beta + \gamma$.

Setting $\beta := \gamma$ gives Kimura's 2ST model, setting $\gamma := \beta := \alpha$ gives the Jukes Cantor model.

For the time period t_e across an edge e of T, the probability of observing the nucleotide substitution $\mathtt{X} \to \mathtt{Y}$ is the $(\mathtt{X}, \mathtt{Y})$ entry of the stochastic matrix

$$P_e = \exp(Qt_e),$$

where matrix exponentiation of a 4×4 matrix M is obtained by the power series

$$\exp(M) = I_4 + \sum_{j \geq 1} \frac{M^j}{j!}.$$

The computations are made much easier as we find the rows of the 4×4 Hadamard matrix

$$H := H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

are the eigenvectors of Q (and hence of P), so H diagonalises both. Thus

$$HQH = -2 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \alpha+\gamma & 0 & 0 \\ 0 & 0 & \beta+\gamma & 0 \\ 0 & 0 & 0 & \alpha+\beta \end{bmatrix},$$

so

$$HP_eH = \mathsf{Exp}(HQHt_e) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & p_e(\alpha) & 0 & 0 \\ 0 & 0 & p_e(\beta) & 0 \\ 0 & 0 & 0 & p_e(\gamma) \end{bmatrix},$$

where

$$p_e(\alpha) = \exp(-2(\alpha+\gamma)t_e), p_e(\beta) = \exp(-2(\beta+\gamma)t_e),$$

$$p_e(\gamma) = \exp(-2(\alpha+\beta)t_e).$$

For a vector $\mathbf{v} = [v_i]$ and a matrix $M = [m_{i,j}]$ we will define the functions Exp and its inverse Ln to mean applying the usual exponential (exp) and natural logarithm (ln) functions respectively to each component of the vector or matrix. Thus

$$\mathrm{Exp}(\mathbf{v}) = [\exp(v_i)], \quad \mathrm{Ln}(\mathbf{v}) = [\ln(v_i)],$$

and

$$\mathrm{Exp}(M) = [\exp(m_{i,j})], \quad Ln(M) = [\ln(m_{i,j})].$$

Thus on diagonalising $P_e = \exp(Qt_e)$ becomes

$$P_e = H^{-1}(\mathrm{Exp}(HQt_eH))H^{-1} \qquad (1)$$

which can be directly inverted The arguments of the log function must be positive) to give:

$$Qt_e = H^{-1}(\mathrm{Ln}(HP_eH))H^{-1} \qquad (2)$$

Equations **??** and **??** are examples of *Hadamard conjugation*.

The independent parameters can be taken as the three edge lengths (distances)

$$q_e(\alpha) = \alpha t_e, q_e(\beta) = \beta t_e \text{ and } q_e(\gamma) = \gamma t_e,$$

which are the expected numbers of each type of substitution across e, or as

$$p_e(\alpha), p_e(\beta) \text{ and } p_e(\gamma),$$

the probability of observing each type of substitution between the endpoints of edge e.

We do not need to refer to rates and time independently.

Hadamard Conjugation on an $X-$tree:
For an $X-$tree T (a tree with leaf set $X = \{0, 1, \cdots, n\}$) we can include Qt_e and P_e for each edge e of T, in suitably defined matrices $Q(T)$ and $P(T)$ of 2^n rows and columns.

The entries of $P(T)$ refer to the probabilities of each the 4^n possible site difference patterns that can occur among the sequences which evolved on T. These probabilities are *independent* of the location of the root R and the nucleotide distribution there.

$Q(T)$ is a sparse matrix (most entries 0). The leading entry is $-K$, where

$$K = \sum_{e \in E} q_e(\alpha) + q_e(\beta) + q_e(\gamma).$$

The 3 edge-lengths for each edge e of T occur in corresponding positions of the leading row ($q_e(\alpha)$), column ($q_e(\beta)$) and diagonal ($q_e(\gamma)$). All other entries are 0.

The positive entries of $Q(T)$ define the edges of T.

The matrix $Q(T)$ is called an *edge length spectrum* for T.
The matrix $P(T)$ is called the *expected sequence spectrum* of $Q(T)$.

These spectra are related by the Hadamard conjugations

$$P(T) = H^{-1}(\mathrm{Exp}(HQ(T)H))H^{-1} \qquad (3)$$

and

$$Q(T) = H^{-1}(\mathrm{Ln}(HP(T)H))H^{-1}, \qquad (4)$$

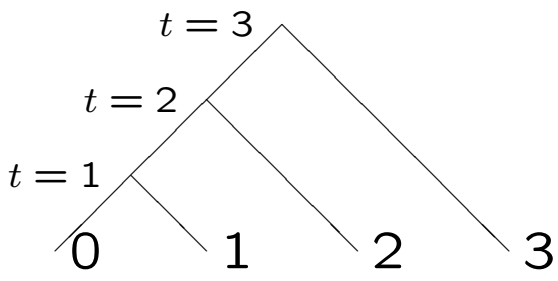
where $H = H_n$ is the $n-$th Sylvester matrix, and $H^{-1} = 2^{-n}H$.

Note that all these matrices have 4^n entries. Using Fast Hadamard Multiplication the computations (3) and (4) can be achieved using $O(4^n)$ space and $O(n4^n)$ time. However these computational resources grow exponentially, and equations (3) and (4) are not practical for large values of n. ($n > 30$.)

APPLICATIONS

- Sequence Analysis - Phylogenetic Invariants
- Proof of invertibility
- Maximum Parsimony
- Corrected Parsimony
- Maximum Likelihood
- Sample sequence generation

Example



In this tree we will assume the molecular clock applies with the K2ST model with a fixed ti/tv ratio. We willset $\alpha = 0.05$, and $\beta = \gamma = 0.01$. Then the edge-length spectrum is

$$Q = \begin{bmatrix} -0.63 & 0.05 & 0.10 & 0 & 0.20 & 0 & 0.05 & 0.05 \\ 0.01 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.02 & 0 & 0.02 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.04 & 0 & 0 & 0 & 0.04 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 \end{bmatrix}.$$

$$Q = \begin{bmatrix} -0.63 & 0.05 & 0.10 & 0 & 0.20 & 0 & 0.05 & 0.05 \\ 0.01 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.02 & 0 & 0.02 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.04 & 0 & 0 & 0 & 0.04 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 \end{bmatrix}.$$

Using **??** we find

$$\begin{aligned} P &= \frac{1}{16} H(\mathrm{Exp}(HQH))H \\ &= 10^{-3} \begin{bmatrix} 658 & 40 & 68 & 11 & 133 & 43 & 15 & 37 \\ 7 & 7 & 1 & 1 & 2 & 2 & 1 & 1 \\ 15 & 1 & 15 & 1 & 3 & 2 & 3 & 2 \\ 3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 32 & 3 & 3 & 4 & 32 & 3 & 3 & 2 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 8 & 2 & 1 & 0 & 2 & 8 & 0 & 1 \\ 7 & 1 & 1 & 1 & 1 & 1 & 1 & 7 \end{bmatrix}. \end{aligned}$$

Inversion - Tree prediction

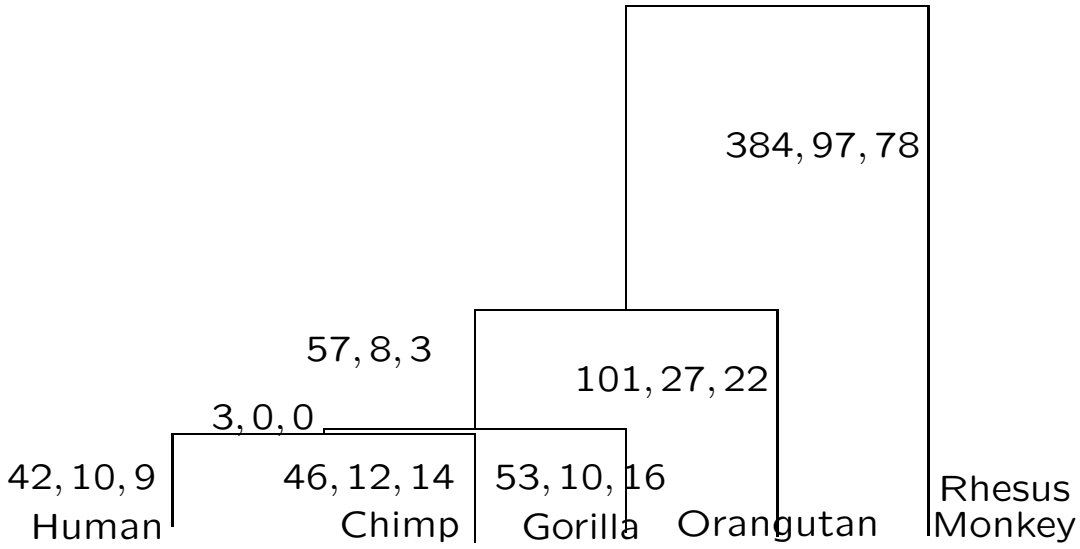
With aligned homologous sequences, we might take the observed frequencies of relative patterns $\hat{P}$ as an approximation to the expected sequence probabilities. Can we interpret

$$\hat{Q} = H^{-1}(\mathsf{Ln}(H\hat{P}H))H^{-1},$$

as an approximation to the edge length spectrum for some tree T?

Example

Using five hemoglobin $\psi-$pseudogenes of 9879 nucleotides, (Human, Chimpanzee, Gorilla, Orangutan and Rhesus Monkey) we find $\hat{Q}$ closely fits a tree.



These data fit closely to the molecular clock, and the model for a fixed transition/transversion ratio.

Phylogenetic Invariants

When the site pattern frequencies $\hat{P}$ are observed, then applying equation (4) produces $\hat{Q}$, an approximation (hopefully) for $Q(T)$, for some tree T.

It is possible for some entries in $H\hat{P}H$ to be non-positive, in which case the log function cannot be applied. In this case we must conclude that this observed data *cannot fit* any tree. This indicates that the data does not support any tree. (Usually only occurs with very short sequences.)

$Q(T)$ is sparse, the corresponding entries in $\hat{Q}$ are expected to be close to 0. The only entries significantly greater than 0 should be $2n-3$ corresponding entries on the leading row, column and diagonal, which should be estimates of edge lengths, all other entries (apart from the leading entry) have expected value 0.

The entries not on the leading row, column or diagonal of $\hat{Q}$ have expected value 0, independent of the choice of the tree T. These are *model invariants*. If some of these are significantly different from 0 then we should reject a tree model for the data.

The $(2^n - 2n + 2)$ corresponding entries on the leading row, column and diagonal are indicators for the tree T. These are *tree invariants*.

For JC and K2ST there are further linear relatiuonships satisfied by the entries of $\hat{Q}$ which are also *model invariants*.

The tree T_{CT} for which the sum of squares of the invariants is minimal, is the *closest tree*.

The tree T_{CP} for which the sum of the invariants (of "informative sites") on the leading row, column and diagonal is minimal, is the *corrected parsimony* tree.
Corrected parsimony is always consistent.

Neither of these tree estimators is practical for large values of n.

Invertibility

From equations (3) and (4) we see that

$$P'(T') = P''(T'') \quad \Rightarrow \quad T' = T'' \text{ and } Q' = Q''.$$

But –

Ellen Baake, (Math., BioSci., 154 (1998), 1-21) showed that it is possible for

$$P'(T') = P''(T'') \text{ with } T' \neq T''$$

when two (not pre-specified) site rate classes are available for each tree.

2-state Sequence Analysis

The Neyman model is introduced as a simpler model of 2-state nucleotides (perhaps R and Y (purines and pyrimidines)) with a probability p_e of substitution on each edge e of T.

We obtain the relationships for the Neyman model by taking just the first columns of $P(T)$ and $Q(T)$, producing vectors $\mathbf{p}(T)$ and $\mathbf{q}(T)$ of 2^n components.

Equations (3) and (4) then become

$$\mathbf{p}(T) = H^{-1}\mathrm{Exp}(H\mathbf{q}(T)) \tag{5}$$

and

$$\mathbf{q}(T) = H^{-1}(\mathrm{Ln}(H\mathbf{p})), \tag{6}$$

where $H = H_n$ is the $n-$th Sylvester matrix, and has 2^n rows and columns.

For the Neyman model we can index each site pattern by a subset of $X^* = \{1, 2, \cdots, n\}$. Thus for $a \subseteq X^*$, p_a is the probability that at a site $a = \{i | \chi_i \neq \chi_0\}$, that is the set of taxa whose character at that site differs from the character at σ_0.

Each edge e of T can be indexed as e_a where $a \subseteq X^*$ is the set of leaves which are separated from 0 by e. We then set q_a to be the edge length of e_a. equations (5) and (6) can then be expressed as

$$p_a = 2^{-n} \sum_{b \subseteq X^*} (-1)^{|a \cap b|} \exp\Big(\sum_{c \subseteq X^*} (-1)^{|b \cap c|} q_c \Big) \tag{7}$$

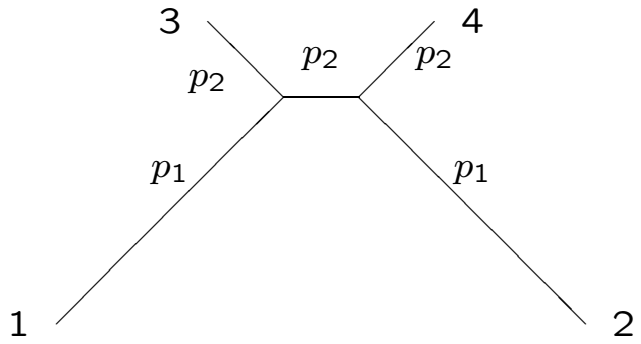
which is easily inverted to give

$$q_c = 2^{-n} \sum_{b \subseteq X^*} (-1)^{|b \cap c|} \ln\Big(\sum_{a \subseteq X^*} (-1)^{|a \cap b|} p_a \Big) \tag{8}$$

The Neyman model enables us to consider properties of phylogenetic methods such as Maximum Parsimony and Maximum Likelihood.

Maximum Parsimony

J Felsenstein (1978) gave his famous "Felsenstein zone" where under a 2-state symmetric model, on a tree with short and long edges, there could be consistency problems, with probabilities p_1 and p_2 of substitutions on the edges as indicated..



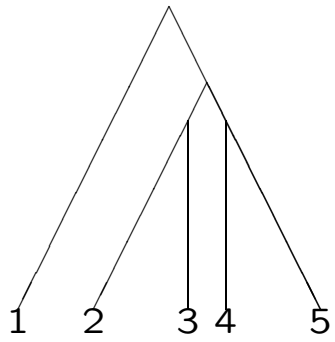
Using equation (5) we find inconsistency occurs

$$\iff p_1^2 > p_2(1 - p_2).$$

Felsenstein hinted that inconsistency might be a problem because the molecular clock was seriously violated.

Using equations (5) and (6) we were able to prove MP was always consistent on a molecular clock tree on 4 taxa.

But ...



MP is inconsistent when the internal edges are small. (An analytic expression for the boundary of this zone of inconsistency can be readily derived.)

However other methods also perform poorly in this region, UPGMA is marginally the best method, and ML and MP are poorest. Methods assuming molecular clock generally perform better on this tree.

Maximum Likelihood

Given a set X of aligned homologous sequences, we observe the frequency f_a of each site pattern $a \subseteq X^*$. The likelihood L of obtaining $F = \{f_a | a \subseteq X^*\}$ from the tree T with edge weight spectrum $Q(T)$ is

$$L = \prod_{a \subset X^*} p_a^{f_a}$$

where for each $a \subseteq X^*$,

$$p_a = 2^{-n} \sum_{b \subseteq X^*} (-1)^{|a \cap b|} \exp\left(\sum_{c \subseteq X^*} (-1)^{|b \cap c|} q_c \right).$$

We then seek the spectrum $Q(T)$ which maximises L (usually we equivalently maximise $\ln L$) under the constraint $q_c \geq 0$ for each edge e_c of T.

$$\ln L = \sum_{a \subseteq X^*} f_a \ln(p_a).$$

For each edge e_c of T,

$$\frac{\partial(\ln L)}{\partial q_c} = \sum_{a \subseteq X^*} \frac{f_a}{p_a} \frac{\partial p_a}{\partial q_c}.$$

From equation (7) we can show

$$\frac{\partial p_a}{\partial q_c} = p_{a\triangle c} - p_a,$$

(where for $a, c \subseteq X^*$, $a\triangle c = a \cup c - a \cap c$ is the symmetric difference of a and c). Hence we can show

$$\frac{\partial(\ln L)}{\partial q_c} = 0 \quad \Rightarrow \quad \sum_{a \subseteq X^*} f_a \frac{p_{a\triangle c}}{p_a} = 1.$$

Solving $\sum_{a \subseteq X^*} f_a \frac{p_{a\triangle c}}{p_a} = 1$ simultaneously for each edge e_c on a tree T gives the turning points. This will give all maximum points that are not on the boundary.

Steel (1994) showed that the likelihood function can have multiple maxima for sequence data evolving on a tree T of four leaves (not molecular clock). This puts into question the hill-climbing strategy for parameter optimisation. Rogers and Swofford (2000) conducted a simulation exercise which suggested multiple optima were uncommon. Chor et al (2000) were able to construct pathological examples with disjoint ridges of maximum likelihood points.

Yang (2000) gave the analytic solution to the maximum likelihood on an unrooted tree of three leaves, for 2−state characters evolving under the Neyman model. This point is unique.

Chor et al (2001) used Hadamard conjugation to find analytic expressions for the maximum likelihood point for a rooted tree on three leaves, for 2−state characters evolving under the molecular clock. In Chor et al (2003) they extend result this to rooted trees on four leaves. We expect (2003) to extend this to 4−state characters evolving under the Jukes and Cantor model under a molecular clock, on a rooted tree of three leaves.

Hendy and Holland (2003) have developed a branch and bound algorithm to find the maximum likelihood tree for 2-state character sequences evolving under the Neyman model with the molecular clock for all rooted trees with $n \geq 5$ leaves.