Phylogeny Estimation and Hypothesis Testing using Maximum Likelihood

For: Prof. Partensky

Group: Jimin zhu
Rama Sharma
Sravanthi Polsani
Xin Gong
Shlomit klopman

April 7, 2003
# Table of Contents

Introduction.................................................................................................................................................. 3

Understanding Maximum Likelihood through Coin Tossing................................................................. 4

The Phylogenetic Tree.................................................................................................................................. 9

Methods for constructing phylogenetic trees............................................................................................. 10

DNA the basis of Molecular Phylogenetics............................................................................................... 14

Maximum Likelihood methods in DNA substitution................................................................................. 19

Likelihood Ratio tests................................................................................................................................. 29

Conclusion.................................................................................................................................................. 30

Bibliography............................................................................................................................................... 31
**Introduction**

Phylogeny is defined as the study of the relationships between life forms. It is a part of a larger field called systematics, which also includes the field of taxonomy. The field of phylogenetics is rather recent and has received a huge push forward because of the development of stronger and faster computers.

Phylogenetics help paleontologists, place organisms in proper context through out evolution. By using molecular Phylogenetics scientists trace the relationships between organisms by studying the similarities of differences in DNA and protein sequences. Today the question phylogenetics is addresses biological fields such as ecology, developmental biology, population genetics and epidemiology.

A realistic and major obstacle that the field of phylogenetics is struggling with is reaching one all accepted answer to the process of evolution. The evolutionary biologist is often uncertain which method of analysis should be used to explain the data. Moreover, when the same data is examined by different phylogenetic methods the outcomes may be different and even contradicting. Evolutionary biologists are left asking the question which conclusion do I believe in?

Today more than ever researchers are in agreement that arriving at a correct species lineage is a matter of statistic relationships. In this paper we will review a strong statistical method – the Maximum likelihood Method for establishing the most likely phylogenetic tree of a given data set. The Maximum Likelihood method was first described in 1922, by English statistician R. A. Fisher. This method depends on a complete and specified data set and a probabilistic model that describes the data. By using this method the researcher ends up with the most likely explanation
(phylogenetic tree) for his data, but he can also learn about the evolutionary process through hypotheses testing. Recent advances in DNA substitution models have made Maximum likelihood a practical and dependable tool for phylogenetic analysis.

**Understanding Maximum Likelihood through Coin Tossing**

The aim of the Maximum Likelihood method is to find the parameter value(s) that make the observed data most likely. Basically, to choose the value of parameter that maximizes the probability of observing the data.

**Probability:** Knowing parameters → Prediction of outcome

**Likelihood:** Observation of data → Estimation of parameters

We toss a coin a $n$ times and record the number of heads. The probability distribution that describes this kind of scenario is the binomial probability distribution. The probability of observing $h$ heads out of $n$ tosses can be described as:

$$Pr[h \mid p, n] = \binom{n}{h} p^h (1-p)^{n-h}$$

The above equation can be considered in two parts. The second part involves the joint probability of obtaining $h$ heads (and therefore $n-h$ tails) if a coin is tossed $n$ times and has probability $p$ of landing heads on any one toss (and therefore probability $1-p$ of landing tails).

Because we have assumed that each of the $n$ trails is independent and with constant probability the joint probability of obtaining $h$ heads and $n-h$ tails is simply the product of all the individual probabilities.

For example: if we obtained 2 heads and 3 tails in 5 coin tosses, then it will simply be $p^2 (1-p)^3$. 
The first half of the binomial distribution function is concerned with the fact that there is more than 1 way to get, h heads and n-h tails if a coin is tossed n times. Every one of the permutations is assumed to have equal probability of occurring - the binomial coefficient \( \frac{n!}{h!(n-h)!} \).

The likelihood function for the above mentioned coin toss experiment in which a binomial distribution is assumed would be:

\[
L(p|h, n) = \binom{n}{h} p^h (1-p)^{n-h}
\]

The likelihood function is obtained by multiplying the probability function for each toss of the coin. In the case of the coin toss experiment, we are assuming a Bernoulli distribution for each coin flip, so the likelihood function becomes:

\[
L(p|X_1 \cdots X_n) = \prod_{i=1}^{n} f(X_i | p) = p^{x_i}(1-p)^{1-x_i} p^{x_i}(1-p)^{1-x_i} \cdots p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n-\sum x_i}
\]

Taking the natural log of the likelihood function does not change the value of p for which the likelihood is maximized. Applying log on both sides of the equation produces the following equation:

\[
\log L(p|X_1, X_2 \cdots X_n) = \sum_{i=1}^{n} X_i \log(p) + (n - \sum_{i=1}^{n} X_i) \log(1-p)
\]
The following graphs show plots of the likelihood (L) as a function of the probability (p) for four different possible outcomes when tossing a coin ten times.

Note that for the case in which 3 heads and 7 tails were the outcome of the experiment, that the likelihood appears to be maximized at p = 0.3. Similarly the likelihood appears to be maximized when p = 0.5 for the experiment outcome with 5 heads and 5 tails. p = 0.8 for the case of 8 heads and 2 tails. p = 0.9 for the case of 9 heads and 1 tail.

The likelihood appears to be maximized when p is the proportion of the time that heads appeared in our experiment. This illustrates the brute force way to find the maximum likelihood estimation of p.
Say we toss a coin 100 times and observe 56 heads and 44 tails. Instead of assuming that \( p \) is 0.5, we want to find the Maximum Likelihood estimation for \( p \). Then we want to ask whether or not this value differs significantly from 0.50. How do we do this? We find the value for \( p \) that makes the observed data most likely.

As mentioned, the observed data are now fixed. They will be constants that are plugged into our binomial probability model:

\[
n = 100 \text{ (total number of tosses)}
\]

\[
h = 56 \text{ (total number of heads)}
\]

Imagine that \( p \) was 0.5. Plugging this value into our probability model gives:

\[
L(\ p = 0.5 \text{ given data}) = \frac{100!}{56! \cdot 44!} \cdot 0.5^{56} \cdot 0.5^{44} = 0.0389
\]

But what if \( p \) was 0.52 instead?

\[
L(\ p = 0.52 \text{ given data}) = \frac{100!}{56! \cdot 44!} \cdot 0.52^{56} \cdot 0.48^{44} = 0.0581
\]

So from this we can conclude that \( p \) is more likely to be 0.52 than 0.5. We can tabulate the likelihood for different parameter values to find the maximum likelihood estimate of \( p \):

<table>
<thead>
<tr>
<th>( p )</th>
<th>( L )</th>
<th>( p )</th>
<th>( L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.48</td>
<td>0.0222</td>
<td>0.56</td>
<td>0.0801</td>
</tr>
<tr>
<td>0.50</td>
<td>0.0389</td>
<td>0.58</td>
<td>0.0738</td>
</tr>
<tr>
<td>0.52</td>
<td>0.0581</td>
<td>0.60</td>
<td>0.0576</td>
</tr>
<tr>
<td>0.54</td>
<td>0.0739</td>
<td>0.62</td>
<td>0.0378</td>
</tr>
</tbody>
</table>
If we can graph the data across the full range of possible values for $p$:

![Maximum Likelihood Estimation](image)

We see that the maximum likelihood estimate for $p$ seems to be 0.56. It is easy to see why this makes sense in this trivial example. The best estimate for $p$ from any one sample is clearly going to be the proportion of heads observed in that sample.

In such a simple example as this, nobody would use maximum likelihood estimation to evaluate $p$. But not all problems are this simple! As we shall see, the more complex the model and the greater the number of parameters, it often becomes very difficult to make even reasonable guesses at the Maximum Likelihood Estimate.
The Phylogenetic Tree

Molecular phylogeny methods using a given set of aligned sequences reconstruct phylogenetic trees which aim at demonstrating the history of successive divergence which took place during the evolution, between the considered sequences and their common ancestor.

There are two types of trees in terms of phylogenetic trees, rooted and unrooted. The difference between them in the biological meaning is that for the rooted, the ancestor of all taxa considered and/or known common ancestor; on the opposite, the unrooted tree is usually unknown common ancestor, but is a measure of similarity between them. Most of phylogenetic methods construct unrooted trees.

A phylogenetic tree is a data structure, characterized by its topology (form) and its length (sum of its branch lengths) that stores information regarding the relationship of several species, which is a measure of homology. But computationally it is usually thought of in terms of an identity or similarity score $d_{i,j}$ between two entities (taxa, sequences, etc.).
Therefore, a phylogenetic tree is simply an arrangement of the data inherent within a multiple sequence alignment into a tree. This arrangement is useful to biologists because it organizes the species into their projected evolutionary history. Due to the construction method of a phylogenetic tree (as fig below), the node (C) represents a common ancestor. The distance (a and b) from the leaves to this common ancestor is a measure of the evolutionary distance (time) between the leaves (A and B).

The evolutionary distance between a & b is \[ p = \frac{(a+b)}{n}, \]  (n total number of species).

The tree structure shows that two sequences are related, how they are related in the context of other sequences, and how distantly they are related.

**Methods for constructing phylogenetic trees**
There are three major methods for constructing phylogenetic trees, parsimony, Maximum Likelihood and distance matrices. Parsimony and Maximum Likelihood are the methods directly based on the sequences, while distance matrices is the method indirectly based on sequences. It is expected that if sequences have strong phylogenetic relationship, different methods should show the same phylogenetic tree. However, it not always true.

There is at present no statistical method which allows comparisons of trees obtained from different phylogenetic methods, nevertheless many studies have been made to compare the relative consistency of the existing methods. A node is said to be "consistent" if the elements it contains are found 95% - 100% of the time. The consistency depends on many factors; among these are the topology and branch lengths of the tree.

Distance matrix methods convert sequence data into a set of discrete pairwise distance values, arranged into a matrix. A tree is fit to this matrix using a cluster analysis method, which makes an estimation of the distance for each pair as the sum of branch lengths in the path from one species to another. This method is easy to perform, quick to calculate and proper for similar species.

Parsimony determines the minimum number of changes (substitutions) required to transform a species to its nearest neighbor.

Maximum Likelihood method was first used by Edward & Cavalli –Sforza in 1964, but since computers were not as fast as they are today they found the problem too computationally difficult and ended up using parsimony. Neyman (1971) applied Maximum Likelihood method to molecular sequences (amino acids and nucleotides) using a simple model that assumed that substitutions occur independently among sites. It was Felsenstein in 1981 that implemented a general Maximum Likelihood method approach to nucleotide sequence data.
The data for molecular phylogenetic problems are the DNA nucleotides: A C T G. For every site, one of four possible nucleotides could be present and if we have four different sequences, there are \(4^4 = 256\) possible site patterns. This data can be described using a multinomial distribution.

\[
Pr[n_1, n_2, \ldots, n_r | p_1 \ldots p_r] = (n_1, n_2, \ldots, n_r) \prod_{i=1}^{r} p_i^{n_i}
\]

where \(n_1, n_2, \ldots, n_r\) is the number of ways that \(n\) objects can be grouped into \(r\) classes; \(ni\) is the number of observations of the \(ith\) site pattern, and \(pi\) is the probability that site pattern \(i\) occurs. A Maximum Likelihood estimate of \(pi\) is:

\[
p_i = \frac{n_i}{n}
\]

that is, the probability of the \(ith\) object equals the proportion of the time it was observed divided by the total number of objects.

While this equation provides a model for the question that Edward & Cavalli–Sforza raised in 1964, it is difficult to compute practically. First of all, the equation cannot estimate tree topology, it ignores the other biological interesting parameters. To make it feasible to the general model on molecular phylogenetic estimation, especially for a tree structure, an equation has been introduced.

\[
L(\tau, \nu, \Theta | X_1 \cdots X_n) = \prod_{i=1}^{r} pr[X_i | \tau, \nu, \Theta]
\]

Assuming independence among sites; where \(\tau\) is a tree; \(\nu\) is a vector containing the lengths of the branches and is either \(\nu = (v_1 \ldots v_{2^{(n-2)}})\) for rooted trees or \(\nu = (v_1 \ldots v_{2^{(n-3)}})\) for unrooted trees \((n\) is the number of sequences); and \(r\) is the total number of site patterns possible for \(s\) sequences. The multinomial coefficient \((n/n_1, \ldots, n_r)\) is a constant and is usually disregarded when calculating
the likelihood of a tree. Also, to speed up computation of the likelihood, the product is taken only over observed site patterns.

In the application of Maximum Likelihood one will have to check each of the \((2n-5)! / (2^{(n-3)})(n-3)!\) Possible unrooted bifurcating trees in turn, and for rooted trees \((2n-3)! / (2^{(n-2)})(n-2)!\), in order to find the Maximum Likelihood tree. Table 1. shows the number of unrooted trees and rooted trees that in to be visited depending on the number of sequences.

<table>
<thead>
<tr>
<th>Number of sequences</th>
<th>Number of unrooted trees</th>
<th>Number of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>10395</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
<td>135135</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
<td>2027025</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
<td>34459425</td>
</tr>
</tbody>
</table>

The procedure of visiting all possible trees and calculating the likelihood for each is computationally expensive. However, there are many shot cuts that can substantially speed up the procedure. An efficient method to calculate the likelihood by taking advantage of the tree topology when summing over all possible assignment of nucleotides to internal node was introduced by Felsenstein.

Basically, with the Maximum Likelihood method, the bases (nucleotides or amino acids) of all sequences at each site are considered separately (as independent), and the log-likelihood of having these bases is computed for a given topology by using a particular probability model. This log-
likelihood is added for all sites, and the sum of the log-likelihood is maximized to estimate the branch length of the tree. This procedure is repeated for all possible topologies, and the topology that shows the highest likelihood is chosen as the final tree. Maximum Likelihood is usually consistent and is extended to allow differences between the rate of transition and transversion.

**DNA the basis of Molecular Phylogenetics**

The DNA molecule is a polymer. The monomer units of DNA are called nucleotides, and the polymer is known as a "polynucleotide." Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base that is attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations (the first letter of their name): Adenine, Guanine, Cytosine, Thymine. Adenine and guanine are purines. Purines are the larger of the two types of bases found in DNA. Structures are shown below:
The 9 atoms that make up the fused rings (5 carbon, 4 nitrogen) are numbered 1-9. All ring atoms lie in the same plane. Cytosine and thymine are pyrimidines. The 6 atoms (4 carbon, 2 nitrogen) are numbered 1-6. Like purines, all pyrimidine ring atoms lie in the same plane.

The deoxyribose sugar of the DNA backbone has 5 carbons and 3 oxygens. The carbon atoms are numbered 1', 2', 3', 4', and 5' to distinguish from the numbering of the atoms of the purine and pyrimidine rings. The hydroxyl groups on the 5'– and 3'– carbons link to the phosphate groups to form the DNA backbone. Deoxyribose lacks an hydroxyl group at the 2'–position when compared to ribose, the sugar component of RNA.
A nucleotide is a nucleoside with one or more phosphate groups covalently attached to the 3'- and/ or 5'-hydroxyl group(s). The DNA backbone is a polymer with an alternating sugar-phosphate sequence. The deoxyribose sugars are joined at both the 3'-hydroxyl and 5'-hydroxyl groups to phosphate groups in ester links, also known as "phosphodiester" bonds. Below is an example a DNA backbone with the following sequence: 5'-d(CGAAT).

![DNA backbone structure](image)

Here are some features of the 5'-d(CGAAT) structure:

- **Alternating backbone of deoxyribose and phosphodiester groups**

- **Chain has a direction (known as polarity), 5' - 3' from top to bottom**

- **Oxygens (in red) of phosphates are polar and negatively charged**

- **A, G, C, and T bases can extend away from chain, and stack atop each other**

- **Bases are hydrophobic**
DNA is normally a double stranded macromolecule. Two polynucleotide chains, held together by weak thermodynamic forces. Shown below in the structure of a double stranded DNA molecule – also called a double helix:

The features of the double helix are:

- Two DNA strands form a helical spiral, winding around a helix axis in a right-handed spiral
- The two polynucleotide chains run in opposite directions
- The sugar-phosphate backbones of the two DNA strands wind around the helix axis like the railing of a spiral staircase
- The bases of the individual nucleotides are on the inside of the helix, stacked on top of each other like the steps of a spiral staircase.
Within the DNA double helix, A forms 2 hydrogen bonds with T on the opposite strand, and G forms 3 hydrogen bonds with C on the opposite strand. Below are examples of base pairs as found within DNA double helix (dA-dT, dG-dC):
The dA-dT and dG-dC base pairs are the same length, and occupy the same space within a DNA double helix. Therefore the DNA molecule has a uniform diameter. The dA-dT and dG-dC base pairs can be found in any order within a DNA molecule.

The DNA helix axis is most apparent from a view directly down the axis. The sugar-phosphate backbone is on the outside of the helix where the polar phosphate groups (red and yellow atoms) can interact with the polar environment. The nitrogen (blue atoms) containing bases are inside, stacking perpendicular to the helix axis. Below is a view down the helix axis. Now that we are familiar with the structure of the DNA we can discover how Maximum Likelihood is used to learn about the process of evolution.

**Maximum Likelihood methods in DNA substitution**

Sequences diverge from a common ancestor because mutations occur. Some fractions of those mutations are fixed into the evolving population by selection and by chance, resulting in the substitution of one nucleotide for another at various sites. In order to reconstruct an evolutionary tree, we must make some assumptions about that substitution process.

DNA substitution consists two basic elements, composition and process. Composition (or equilibrium frequency) is defined as the proportion of the four nucleotides in a sequence when the probability of the nucleotide’s substitution approaches the equilibrium state.

The process can be described by a matrix of numbers, describing how the nucleotides change from one to another. This process can be described as a Binomial distribution. Because there are a total of \( r = 4^s \) site patterns possible for \( s \) species. The probability of the nucleotide change from one to
another is very small and the number of total site patterns could be very large. The Poisson distribution which is a limited case of the Binomial distribution can explain this process. Therefore, all current implementations of likelihood estimation assume a time-homogeneous Poisson process to describe DNA substitutions.

This Poisson process is time-homogeneous because the following assumptions are made. The occurrence of any nucleotide substitution in the time interval (a, b) is independent of the occurrence of any nucleotide substitution in the time interval (c, d), where (a, b) and (c, d) do not overlap. The probability of a nucleotide substitution in the time interval (t, t+h) is independent of t. The probability of a nucleotide substitution occurring in a small time interval is proportional to the length of the interval.

Suppose the distribution of the number of substitutions $s$ is a Poisson random variable with mean $\lambda$. $t$ during $t$ units of time. The rate of substitutions (relative to the unit of time) at a given site is $\lambda$.

The probability of $s > 0$ at a site in a time period $t$ is:

$$P_r(s) = \frac{\exp(-\lambda t)(\lambda t)^s}{s!}.$$ 

Thus, the probability of no changes occurring at a site is:

$$P_r(s = 0) = \exp(-\lambda t)$$

and the probability for at least one substitution is:

$$P_r(s \neq 0) = 1 - \exp(-\lambda t).$$

Therefore, the transition probabilities with two-character states can be described as follows (3):

$$P(ij) = \begin{pmatrix}
\pi_i + (1 - \pi_i)e^{-\lambda t} & \pi_i - \pi_j e^{-\lambda t} \\
\pi_j - \pi_i e^{-\lambda t} & \pi_j + (1 - \pi_j)e^{-\lambda t}
\end{pmatrix}$$
Where, $\pi_i$ and $\pi_j$ are the equilibrium frequencies of states i and j (i and j are nucleotides). $\lambda$ is the rate of change from i to j and t is the arbitrary interval of time.

A simplified model assumes that the probability of any nucleotide changing to any other nucleotide is equal. Let i be at some site at time $t = 0$. Let $\pi 0 = \pi 1 = \frac{1}{4}$, the probability at time $t$ there will still be no change $P(ii)$ and change $P(ij)$ are $P(ii) = \frac{1}{4} + \frac{3}{4}(e^{\lambda t})$ and $\frac{1}{4} - \frac{1}{4} (e^{\lambda t})$, respectively.

Fig 1 is a plot that shows this simple case (let $\lambda = 0.5$) by Mathematica:

(* no change vs time t*)

\[ P_{ii} = \text{Plot}[1/4 + (3/4)*e^{(-0.5*t)}, \{t, 0.001, 10\}, \text{Frame} \to \text{True}, \text{PlotRange} \to \text{All}] \]

(* change from i to j vs time t*)

\[ P_{ij} = \text{Plot}[1/4 - (1/4)*e^{(-0.5*t)}, \{t, 0.001, 10\}, \text{Frame} \to \text{True}, \text{PlotRange} \to \text{All}] \]
When \( t \) is very close to zero, the probability that the site has not changed, \( P_{ii} \), is very close to 1, while \( P_{ij} \), the probability that the nucleotide at that site has changed from \( i \) to \( j \) is very close to 0. As time goes on, both probabilities approach \( \frac{1}{4} \) (equilibrium frequency). The time required for that to happen depends on \( \lambda \).

With the assumption above and assuming independence among sites, the likelihood of two DNA sequences, \( L(A, B \mid t_1, t_2) \), can be represented as follows:

\[
L = \frac{1}{16^n + n^2} \cdot (1 + 3e^{-4\lambda t_1})^{n_1} \cdot (1 - e^{-4\lambda t_1})^{n_2}
\]

where \( \lambda \) is the rate of change. \( n_1 \) sites remain same and \( n_2 \) sites change, \( t \) is the sum of \( t_1 \) and \( t_2 \), and can also be considered as branch length form node A to B.

Suppose we have two nucleotide sequences, or simply, that only the nucleotides C and G, are present, for instance, that sequence might be:

- Sequence A: CCGGCGCG
- Sequence B: CGGCGCGGC

The maximum likelihood \( L(A, B \mid t_1, t_2) \), of these two sequences can be calculated (using Mathematica) as follows:

(*Likelihood vs. interval time \( t \) assume \( \lambda \) is known*)

\[
Clear[t1, t2]; \lambda = 0.007; n1 = 8; n2 = 3;
Plot[\left(1 + 3e^{-4\lambda t} \right)^{n_1} \cdot \left(1 - e^{-4\lambda t} \right)^{n_2}, t, 0, 100, Frame \rightarrow True, PlotRange \rightarrow All]\]
The graph above shows that t or (t1 + t2) at a Maximum Likelihood of 1.4E-11 the distance between sequence A and sequence B is 17 arbitrary distance units.

If the number of n1 and n2 are changed to 3 and 8, respectively. The graph shown above becomes the graph below, and the curve is not as sharp as before. The Maximum Likelihood shifts to the right dramatically, and the t that will produce the best maximum Likelihood is unclear (probably the sequence is too short).

It makes sense that the more identical the sites are the shorter the distance between them. It also makes sense that when λ is smaller the distance between two sequences is larger. The comparison
of the Maximum Likelihood with small $\lambda$ and large $\lambda$ is shown in the following two graphs. With a small $\lambda$ (0.002) we see the curve shift to the right compared to the curve with large $\lambda$ (0.007).

$\lambda = 0.007$:

$\lambda = 0.002$:
The algorithm used to find the Maximum Likelihood for an arbitrary number of sequences is more complicated. A primary tree-building program that uses the Maximum Likelihood method is called PAUP* is Available at http://www.sinauser.com. PAUP is used for reconstruction of phylogenetic trees based on nucleic acid alignments.

The simplest case of DNA substitution is the model called Jukes-Cantor (JC69) model. JC69 model assumes that the base frequencies are equal (\( \pi A = \pi C = \pi G = \pi T \)) and that the rate of change from one nucleotide to another is the same for all possible changes. However, the JC69 model, like several other models, is simply a special case of a general model of DNA substitution for which the instantaneous rate matrix \( Q \) has the following form:

\[
P(ij) = \begin{pmatrix}
\cdots & r_2 \pi_C & r_4 \pi_G & r_6 \pi_T \\
r_1 \pi_A & \cdots & r_8 \pi_G & r_{10} \pi_T \\
r_3 \pi_A & r_7 \pi_C & \cdots & r_{12} \pi_T \\
r_5 \pi_A & r_9 \pi_C & r_{11} \pi_G & \cdots
\end{pmatrix}
\]

The rows and columns are ordered A, C, G, and T. The matrix gives the rate of change from nucleotide \( i \) (arranged along the rows) to nucleotide \( j \) (along the columns). The \( r \) stands for rate of change and \( \pi \) stands for base frequencies. The commonly used models of DNA substitution are based of this general model. The following models differ in the settings of two parameters they are: nucleotide frequency and rate of change between two nucleotides.
<table>
<thead>
<tr>
<th>Model</th>
<th>Nucleotide frequencies</th>
<th>Rates of change</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>( \pi_A = \pi_C = \pi_G = \pi_T )</td>
<td>( r_1 = r_2 = r_3 = r_4 = r_5 = r_6 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_7 = r_8 = r_9 = r_{10} = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>K80</td>
<td>( \pi_A = \pi_C = \pi_G = \pi_T )</td>
<td>( r_3 = r_4 = r_9 = r_{10} = r_1 = r_2 = r_5 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_6 = r_7 = r_8 = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>K3ST</td>
<td>( \pi_A = \pi_C = \pi_G = \pi_T )</td>
<td>( r_3 = r_4 = r_9 = r_{10} = r_5 = r_6 = r_7 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_8 = r_1 = r_2 = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>F81</td>
<td>( \pi_A; \pi_C; \pi_G; \pi_T )</td>
<td>( r_1 = r_2 = r_3 = r_4 = r_5 = r_6 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_7 = r_8 = r_9 = r_{10} = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>HKY85</td>
<td>( \pi_A; \pi_C; \pi_G; \pi_T )</td>
<td>( r_3 = r_4 = r_9 = r_{10} = r_1 = r_2 = r_5 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_6 = r_7 = r_8 = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>TrN</td>
<td>( \pi_A; \pi_C; \pi_G; \pi_T )</td>
<td>( r_3 = r_4; r_9 = r_{10}; r_1 = r_2 = r_5 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_6 = r_7 = r_8 = r_{11} = r_{12} )</td>
</tr>
<tr>
<td>SYM</td>
<td>( \pi_A = \pi_C = \pi_G = \pi_T )</td>
<td>( r_1 = r_2; r_3 = r_4; r_5 = r_6; r_7 = r_8; r_9 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_{10}; r_{11} = r_{12} )</td>
</tr>
<tr>
<td>GTR</td>
<td>( \pi_A; \pi_C; \pi_G; \pi_T )</td>
<td>( r_1 = r_2; r_3 = r_4; r_5 = r_6; r_7 = r_8; r_9 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = r_{10}; r_{11} = r_{12} )</td>
</tr>
</tbody>
</table>

The nucleotide frequency is rather simple to understand. The models can be divided into two groups, those that assume equal nucleotide frequency (JC69, K80, K3ST, SYM), and those that assume unequal nucleotide frequency (F81, HKY85, TrN, GTR). The rate of change parameter is more complex and assigns equal or different probabilities to the changes (mutations) between two nucleotides.

There can be two kinds of mutations between the four DNA nucleotides: Transition or Transversion. A transition is a mutation between two nucleotides from the same chemical/ structural group. For example:

- purine transition  \( G \leftrightarrow A \)
- pyrimidine transition  \( C \leftrightarrow T \)

A transversion can occur between any two of the four nucleotides. The list below describes in greater detail the different mutations (r1 – r12) that are used in the above models. The nucleotide pairs are formed from the general Q matrix discussed previously.
Here is a simplified table for the eight models used in DNA substitution listing their nucleotide frequencies and rates of change:

<table>
<thead>
<tr>
<th>Models</th>
<th>Nucleotide frequencies</th>
<th>Rate of nucleotide change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Transitions</td>
</tr>
<tr>
<td>JC69</td>
<td>Equal</td>
<td>Equal rates</td>
</tr>
<tr>
<td>K80</td>
<td>Equal</td>
<td>Different rates</td>
</tr>
<tr>
<td>K3ST</td>
<td>Equal</td>
<td>Different rates</td>
</tr>
<tr>
<td>F81</td>
<td>Different</td>
<td>Equal rates</td>
</tr>
<tr>
<td>HKY85</td>
<td>Different</td>
<td>Different rates</td>
</tr>
<tr>
<td>TrN</td>
<td>Different</td>
<td>Rate for transition G&lt;&gt;A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate of transition T&lt;&gt;C</td>
</tr>
<tr>
<td>SYM</td>
<td>Equal</td>
<td>Rate for transition G&lt;&gt;A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate for transition T&lt;&gt;C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTR</td>
<td>Different</td>
<td>Rate for transition G&lt;&gt;A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rate for transition T&lt;&gt;C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using these models and the Maximum Likelihood method it is possible to learn about the evolutionary process. This study is called hypotheses testing and will be explained in the following section.
Likelihood Ratio tests

All methods in the field of phylogenetics make assumptions about the process of evolution. A common assumption for example, is a bifurcating tree to describe the relationship between species. Another common assumption that is used in DNA substitution is that the nucleotide sites in a DNA are independent of each other.

The Maximum Likelihood method makes explicit assumptions about an evolutionary process and that allows the scientist not only to estimate the relationship between species, but also to learn about the process of evolution through hypothesis testing. When using hypothesis testing, we establish a hypothesis, which is referred to as the null hypothesis and an alternative hypothesis, which usually contradicts the null hypothesis.

Next, we use two competing models to try and explain our data (one model will fit the null hypothesis and the other will fit the alternative hypothesis). We compute the Maximum Likelihood given our data for each model and find their ratio as follows:

\[
Ratio = \frac{ML \text{ of null hypothesis}}{ML \text{ of alternative hypothesis}}
\]

When the ratio is less than one the null hypothesis is rejected (and the alternative hypothesis is accepted). When the ratio is greater than one the alternative hypothesis is rejected (the null hypothesis is accepted).

To help us understand species divergence through DNA substitution, we can subject different models of DNA substitution to hypothesis testing. For this purpose, we introduce 8 different DNA
substitution models. These models are subsets of the general substitution model (with the Q matrix), and are subsets of one another. This means that we can produce a hierarchy of the different models starting with the model that has only one questionable parameter all the way through to the most complex model. Using likelihood ratio tests, we can determine whether a particular parameter (also be called hypothesis) provides a significant increase in the likelihood.

For example: the first hypothesis questions equal base frequencies. The null hypothesis will state that the 4 nucleotides can be found in equal frequencies in a given sequence (model JC69). The alternative hypothesis will say the exact opposite that the nucleotide frequencies are not the same in a given sequence (model f81). To reach a decision we subject the two models to a likelihood ratio test. In this case, the ratio was less then one so the null hypothesis is rejected and we conclude that the four nucleotides appear in different frequencies in a given sequence.

**Conclusion**

The use of the Maximum Likelihood method has become a practical tool in phylogenetics because of recent advances in DNA substitution models, computer programs and faster computers. One of the strengths of the Maximum Likelihood Method of phylogenetic estimation is the ease in which hypotheses can be formulated and tested. Since the assumptions used in this method are explicit and clear statistical tests of phylogenetics can be formulated. The Maximum Likelihood method provides a uniformed framework for the evaluation of the alternative hypotheses. Likelihood ratio tests can be applied to questions for which the null distribution is difficult to determine analytically.
Bibliography


