Reconstructing Evolutionary Trees Through the Use of Markov Chains

Katherine Williams
Introduction

One of the universal questions concerning people of all cultures and faith is that of how humans came to exist. Our origins have been suggested to come from different places varying from divine creation to simple amino acid evolution. The field of molecular genetics has taken an interest in solving this puzzle biologically through the use of genetic analysis and evolutionary tree construction.

Historically, biologists have focused on Darwin’s theory of evolution to try and explain how different species came about. His principle of “survival of the fittest” was based on how the advantages and disadvantages of various morphological traits affected the success and ability of different individuals to survive. As mutations of individual’s genes resulted in the appearance of new physiological traits which endured better, the populations of future generations contained higher frequencies of those traits that more successfully survive. Greater and greater diversions from the original form resulted in what could be classified as a new species. The better surviving species are so different from the original, non-mutated forms, that a speciation event is said to occur. These alterations of traits tried to explain the evolution of species through time based on the morphology. Constructing evolutionary trees simply through the relationship of physical qualities leaves too much room open, however, for speculation and guesswork (Balding et al, 2001, p.327).

Nowadays, due to the recent developments and discoveries in genome projects of various species, the DNA sequences of several taxa can be established. DNA relationships are much more reliable than those of physical attributes since the actual path
of mutation can be viewed whereas conjecture comes heavily into play for how various morphological characteristics came about. Once the DNA sequences have been successfully aligned, the species can then be arranged into an evolutionary tree based on genetic differences. Such a phylogeny will then possibly represent the true evolution of different species and how they derived from common ancestors at the root of the tree.

One drawback to this method is that as many species are now extinct and their DNA is not fully available; finding the completed genome sequences of every species on the planet is not entirely feasible, let alone discovering the sequences of every species that has ever existed (which are most likely unobtainable). However, a mathematical model can be created anyway to present a method by which a partial evolutionary tree can be constructed. The purpose would be to find the evolutionary tree which best fits the DNA data available. “Filling in” the tree by morphological traits would then most likely come into play. Once the best phylogeny is constructed, it is also valuable to approximate the branch lengths of the tree. This helps us to understand the rate of evolution among species.

The motivation behind this attempt is to find a way in which to examine evolution and how current species today came about. It also helps us to understand how different species relate to one another and compare in structure and function. For example, by finding common ancestors to humans, further genetic analysis and testing can find ways to induce mutations/diseases among the species with the shared DNA. Studying the effects on the common species can help us to find a possible cure for diseases such as Parkinson’s disease and Huntington’s disease in humans.
General Background

In order to assess the relationships between species and their DNA, a brief introduction of evolutionary trees and genetic analysis is needed. Evolutionary trees are best explained with an example:

Here, the root of the tree is the common ancestor of contemporary taxa 1, 2, and 3. Taxa 1 and 2 are more closely related to each other than to taxon 3. The node at which they split is called a speciation event (or internal node) (Lange, 2002, p.200). This is the point from which two new species bifurcate. As a result, these taxa most likely have more similar DNA sequences. All 3 groups ultimately derived from the root.

When the descent path is unknown for the example species discussed above (1,2, and 3), there are actually three different possible trees:
The general rule is that given \( n \) contemporary taxa (species that still exist in the current
time), there are \( T_n = \frac{(2n-3)!}{2^{n-2} (n-2)!} \) different possible trees (Lange, 2002, p.165). This is
relatively simple for small numbers of species, but the possibilities quickly rise when
comparing greater numbers. For example, \( T_5 = 105 \) and \( T_6 = 945 \) possible trees.
Obviously, more information (morphology and DNA sequencing) is needed to try and
lower these numbers into something more feasible to work with. In order to see how this
can be done, a basic understanding of genetics is necessary.

Genetics is a much broader topic and only the essentials will be mentioned. The
most important aspect applicable to the model concerns the composition of DNA. DNA
is made up of a sugar-phosphate backbone and four different types of bases: adenine (A),
guanine (G), cytosine (C), and thymine (T). These four bases, also known as nucleotides,
make up the DNA sequences that vary among species.

These sequences are then transferred over to another form of genetic information
known as messenger RNA (mRNA) through the process of transcription. Messenger
RNA also contains the bases A, G, and C; but in place of thymine, mRNA uses uracil
(U). Different arrangements of mRNA sequences produce various amino acids through
the process of translation, which thereby create various proteins that ultimately make up a
unique organism. This process has been greatly simplified, yet the main message
remains that the fundamental source of the genetic information lies in the DNA.

It is important to note that DNA consists of both coding fragments (exons) and
non-coding fragments (introns). It is only the exons that contain the regions responsible
for eventually coding the amino acids. Introns are regions of DNA with unverified
function.
Some examples of this genetic coding for amino acids follow: the sequence GGC codes for the amino acid glycine while AGC codes for the amino acid serine. These three nucleotide long sequences are called codons. There are 64 different possible codons, most of which ultimately code for 20 different amino acids.

There is another interpretation of the 3-nucleotide codon which results in something known as a STOP code. This can be interpreted as a "period" in the DNA sequence which tells the coding to stop at this codon, and begin again at the next three nucleotides. As a result, those sequences do not code for amino acids and are not included in the resultant protein.

<table>
<thead>
<tr>
<th>First position (5'-end)</th>
<th>Second position</th>
<th>Third position (3'-end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU { Phe }</td>
<td>UUC { Leu }</td>
<td>UUG { Cys }</td>
</tr>
<tr>
<td>UUA { Leu }</td>
<td>UCA { Ser }</td>
<td>UAG { STOP }</td>
</tr>
<tr>
<td>AUG { Ser }</td>
<td>UCU { Tyr }</td>
<td>UGC { STOP }</td>
</tr>
<tr>
<td>CUU { Leu }</td>
<td>CCA { Ser }</td>
<td>CGA { Arg }</td>
</tr>
<tr>
<td>CUA { Pro }</td>
<td>CCG { Ser }</td>
<td>CGG { Arg }</td>
</tr>
<tr>
<td>CUG { Ile }</td>
<td>ACC { Thr }</td>
<td>AGU { Ser }</td>
</tr>
<tr>
<td>AUA { Met }</td>
<td>ACA { Thr }</td>
<td>AGG { Arg }</td>
</tr>
<tr>
<td>AUG { Met }</td>
<td>ACG { Thr }</td>
<td></td>
</tr>
<tr>
<td>GUU { Val }</td>
<td>GCU { Asp }</td>
<td>GGU { Gly }</td>
</tr>
<tr>
<td>GUC { Val }</td>
<td>GCC { Asp }</td>
<td>GGC { Gly }</td>
</tr>
<tr>
<td>GUA { Val }</td>
<td>GCA { Asp }</td>
<td>GGA { Gly }</td>
</tr>
<tr>
<td>GUG { Val }</td>
<td>GCG { Asp }</td>
<td>GGG { Gly }</td>
</tr>
</tbody>
</table>

**Figure 1: The genetic code**

---

1. Figure from *The Nature of the Genetic Code*. Codons in this figure correspond to the codons found in the mRNA sequence.
Consequently, several codons may code for the same amino acid. For instance, the sequences AGU and UCG also code for serine. However, if a single base is mutated in a codon to another base, the resulting amino acid may change. For example, if the U in GUC, which codes for valine, is mutated into a C, the following codon is GCC, which codes for alanine. (Griffiths et al, 2000, p.318)

These single site mutations over the course of several generations may eventually result in the evolution of an entirely new species in time. These changes of bases are one of the causes of the branching in evolutionary trees. Those species with more closely related DNA sequences are then assumed to be closer together on a tree and more likely share a common ancestor.

In addition, in order to analyze and compare the genetic data of different species, their genetic sequences must first be aligned. Not all species have the same number of nucleotides in their genome. For instance, the human genome is estimated to be comprised of about 3000 million base pairs whereas Drosophila melanogaster (fruit flies) are thought to have about 180 million base pairs (Freeman, 2002, p.328). How can we compare these sequences when their sizes differ by almost 17 fold? In order to help alleviate this problem, gaps are introduced into the shorter sequences so as to maximize alignment of the available bases. Now the question remains, how do we bring all this information together to form a possible evolutionary tree?

**Phylogeny Construction Methods**

In general, there are three currently accepted models of ways to construct evolutionary trees: maximum likelihood methods, maximum parsimony methods, and
distance methods. Each has their advantages and drawbacks, which will briefly be discussed here.

To begin with, the maximum likelihood method is somewhat unintuitive and concerns finding the tree that makes the data the most likely to occur. Its basis is that given a set of genetic data, $D$, and a possible evolutionary tree, $T$, the likelihood is the probability that the genetic data, $D$, occurs when assuming a given tree $T$. Or in mathematical terms, $L_D = \Pr(D | T)$. The tree that gives us the highest probability is the one that is the maximum likelihood estimate and corresponds best with the data available to us (Page & Holmes, 1998, p.193). Three main assumptions are made for this model (Halloran & Geisser, 1999, p.95):

1) Each site in the sequence evolves independently.
2) Different lineages evolve independently.
3) Each site undergoes substitution at an expected rate that is chosen from a series of rates with a given distribution.

Additional parameters may be added, (including the substitution models discussed later), in order to specify the tree’s emphasis. In general, the maximum likelihood method gives a mathematically sound explanation for the evolution of different species, yet may not be realistic biologically speaking. Such computations for every possible tree are also time consuming on a computational level. A specific tree may also give different maximum likelihood values, making the value of such computations circumspect. For instance, there may actually be a higher likelihood value for a given tree than those found. What is the real likelihood value of a tree if there are several, known and/or unknown? (Page & Holmes, 1998, p.201).

Maximum parsimony methods help to solve these problems by introducing the concept that “evolution is parsimonious.” This implies that evolution occurs in as few
evolutionary steps as possible. So, the optimal tree would be one in which the number of changes between ancestors and descendents is minimized (Halloran & Geisser, 1999, p.92).

For example, given the four sequences: 1) ATATT, 2) ATCGT, 3) GCAGT, and 4) GCCGT, we wish to find the most parsimonious tree connecting them. The three possible unrooted trees are:

![Possible unrooted trees of the 4 given sequences](image)

In order to find the most parsimonious tree, the sequences must be looked at site by site. First we will look at tree X. The two possible trees for the first nucleotide site look like the following, where the slash marks indicate an evolutionary change from one node nucleotide to another:

![Two possible paths of evolution for site 1 of tree X](image)

The first tree obviously has fewer steps in evolution (only 1 change) compared to the second tree which has 5 nucleotide substitutions. As a result, the first tree is the most parsimonious. The remaining nucleotide sites are examined similarly until all sites have
been accounted for. For each site, the number of evolutionary changes are counted and
totaled together to give a parsimony score (also known as the tree's length) for the total
tree. After this is done, the following totals are found:

Tree X: 5, Tree Y: 6, Tree Z: 7.

Tree X has the lowest length of 5, so it is the most parsimonious, and therefore the most
likely tree.

The main advantages of this method lie in the fact that it is very easy to
understand and doesn't make many assumptions about the actual biological process of
evolution with one exception. It only assumes that evolution happens in as few steps as
possible. The main disadvantage, however, is that several different trees can result in the
same parsimony score and are deemed to be phylogenetically uninformative. One tree
may be just as likely as several other trees (Page and Holmes, 1998, pp.187-189). How
are we to know which one is correct?

The last method concerns the distance method. A cluster analysis technique that
makes use of this method will be discussed in more detail later. However, the main
concept this technique uses is to determine the “distance” or genetic differences between
two different sequences of DNA and then apply a “clustering” model to construct the tree
from the data. The distances are acquired through the use of Markov matrix modeling.
In order to discuss this topic further, we must first explain the different types of
applicable Markov models.

Markov Substitution Models

Although there are several different ways to approach this problem (allele
frequency studies, Monte Carlo methods, etc.), we will attempt to find a possible solution
through the use of Markov chains in a nucleotide substitution model. This form of modeling works well with our problem as it only looks at the current state of a system to determine the probability of change from one form to another. In other words, in DNA sequences, the probability that a nucleotide will mutate into another depends only upon the base currently present at a specific site, not on what may have been present there earlier due to evolution.

To clarify this idea, an example serves as the best tool. Given a sequence ATTGCTA, if this sequence evolved into ATTGGTA and we want to know the probability that it will mutate again to the sequence ATTGATA; it does not matter that the sequence used to originally be ATTGCTA. All we care about is the present state of information (that the sequence is now ATTGGTA). According to the Markov model, the probability of a nucleotide substitution at a specific site is independent of what the sequence may have been beforehand. This works well for our purposes because we are largely unable to know what sequences might have been present at specific sites in earlier species. We don’t know all the genetic makeups of every species ever to have inhabited the planet. The only information we have is what is currently available to us in present species or archaeological findings.

To begin the Markov model, a transition matrix $Q = \{q_{ij}\}$, representing the rate of change between one nucleotide $i$, to another, $j$, needs to be created. We will start with the simplest model (also known as the Jukes-Cantor model) which assumes that the probability of change from one base to another is equal for all possible mutations. It will also assume that nucleotide frequencies are equal among all four bases. The resultant Markov rate matrix is:
The derivations of the subsequent formulas are somewhat tedious, so only outcomes of importance will be discussed. This substitution matrix gives the probability that a nucleotide will change from \( i \) to \( j \) during a time interval \( t \), as \( P(t) = \{ p_{ij}(t) \} = e^{Qt} \) (Yang, 2000, pp. 329-330). We then wish to find the genetic distance between two nucleotide sites through the substitution matrix above. Jukes and Cantor define this distance \( d \) between two sites of species \( x \) and \( y \) by the mean number of accumulated changes \( 2 \times 3\alpha t \) where

\[
d_{x,y} = 6\alpha t = \frac{3}{4} \log \left[ 1 - \left( \frac{\# AA + \# CC + \# GG + \# TT}{4} \right) \right]
\]

where \#AA indicates the number of times there is an A in species \( x \) matched with an A in species \( y \), and so forth (Halloran, 1999, p.91).

The problem with the Jukes-Cantor model is that although it may seem as if the transition rate between bases would be constant between all four bases, there are in fact two different kinds of mutations which lead to different rates. This is due to the fact that the bases A and G have similar structures; these are called purines. C and T also have related structures and are called pyrimidines. Because their structures are alike, it is easier to mutate between \( A \leftrightarrow G \) and between \( C \leftrightarrow T \), (both called transition-mutations), than between \( A \leftrightarrow T \) or \( C \) and \( G \leftrightarrow T \) or \( C \), (both called transversion-mutations). The following graph represents the conditions on the rate matrix. (Yang, 2000, p.177)
The Markov matrix $Q$ gives the resultant rate matrix.

$$
Q = \begin{array}{cccc}
A & G & C & T \\
A & -(\alpha + \beta) & \alpha & \beta & \beta \\
G & \alpha & -(\alpha + \beta) & \beta & \beta \\
C & \beta & \beta & -(\alpha + \beta) & \alpha \\
T & \beta & \beta & \alpha & -(\alpha + \beta) \\
\end{array}
$$

This model is also known as the Kimura two-parameter model. The Kimura model can be solved similarly to the Jukes-Cantor model. If we let:

$$
P_1(t) = \text{Fraction of transitions} \\
P_2(t) = \text{Fraction of transversions}
$$

The resulting set of differential equations is:

$$
\frac{dP_1(t)}{dt} = 2\alpha - 4(\alpha + \beta)P_1(t) - 2(\alpha - \beta)P_2(t) \\
\frac{dP_2(t)}{dt} = 4\beta - 8\beta P_2(t), \text{ with } P_1(0) = P_2(0) = 0
$$

Solving this system gives:

$$
P_1(t) = \frac{1}{4} (1 - 2e^{-4(\alpha + \beta)t} + e^{-8\beta t}) \\
P_2(t) = \frac{1}{2} (1 - e^{-8\beta t})
$$
As a result, the mean number of accumulated changes for this model and the resultant distance is then:

\[ d = 2(\alpha + \beta)t \]

\[ = \frac{1}{2} \ln\left(1 - 2P_1(t) - P_2(t)\right)\sqrt{1 - 2P_2(t)} \]

**Reconstruction of Evolutionary Trees**

Once the distances between species are found, the next step is to apply those distances to create a possible working evolutionary tree. As discussed earlier, one way to approach this problem is through “cluster analysis.” This is most easily explained with an example, (with information taken from Nei’s *Molecular Evolutionary Genetics*). The following table gives the mitochondrial distances between primates. (These distances were found using the Jukes-Cantor method).

<table>
<thead>
<tr>
<th>Table 1: Mitochondrial distances between selected primates</th>
<th>Human (1)</th>
<th>Chimpanzee (2)</th>
<th>Gorilla (3)</th>
<th>Orangutan (4)</th>
<th>Gibbon (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (1)</td>
<td>--</td>
<td>0.094</td>
<td>0.111</td>
<td>0.180</td>
<td>0.207</td>
</tr>
<tr>
<td>Chimpanzee (2)</td>
<td>--</td>
<td>--</td>
<td>0.115</td>
<td>0.194</td>
<td>0.218</td>
</tr>
<tr>
<td>Gorilla (3)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.188</td>
<td>0.218</td>
</tr>
<tr>
<td>Orangutan (4)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.216</td>
</tr>
</tbody>
</table>

The first step involves “linking” the two species with the closest genetic distance and calling them a new species. In this case, humans and chimpanzees are the closest together, so we call them species 1-2. Their respective species are placed on a genetic

---

2 Mitochondria are organelles in cells which contain genetic material such as DNA.
distance map with an evolutionary tree branch each. The joining of the two branches is representative of the species 1-2, (see figure below). The next step involves adjusting the average distances between the two species compared to the other remaining species. For instance, the distance between gorillas and species 1-2 is the average of the distance between humans and gorillas and chimpanzees and gorillas.

\[ d_{1,2,3} = \frac{d_{31} + d_{32}}{2} = \frac{0.111 + 0.115}{2} = 0.113 \]

This is repeated for orangutan and gibbon to create the new distance matrix:

**Table 2: Distances between combined species**

<table>
<thead>
<tr>
<th></th>
<th>Human-Chimp (1-2)</th>
<th>Gorilla (3)</th>
<th>Orangutan (4)</th>
<th>Gibbon (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Chimp (1-2)</td>
<td>--</td>
<td>0.113</td>
<td>0.167</td>
<td>0.212</td>
</tr>
<tr>
<td>Gorilla (3)</td>
<td>--</td>
<td>--</td>
<td>0.188</td>
<td>0.218</td>
</tr>
<tr>
<td>Orangutan (4)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.216</td>
</tr>
</tbody>
</table>

The process is now repeated by grouping species 1-2-3 (go back to step one), etc. For \( n \) number of species, after \((n-1)\) groupings, a possible evolutionary tree should result (Yang, 2000, pp.179-180)

*Figure 5: Proposed phylogeny from cluster analysis*
This evolutionary tree shows the relationship between these four species relative to their genetic distances according to the Jukes-Cantor model. Assuming that the genetic distances between all species are known, the relationship between every species could be displayed in such a way. As noted earlier, however, this is not possible, but the model is available and functional for current species.

Obviously the advantages of this form of tree reconstruction are the ease of computation and understanding. Clustering methods also tend to produce only one tree, (instead of several like maximum parsimony). One disadvantage to the distance method is that information is lost. For example, with the parsimony method, we could trace where the tree changes at each evolutionary step. In the distance method, we can only tell how much change has occurred without knowing exactly which changes have happened. Another problem with this technique is that it is sometimes difficult to interpret the information biologically. As in the above example, humans differ from chimpanzees by 0.094 substitutions, yet this has no real significance as 0.094 substitutions is impossible (Page and Holmes, 1998, p.186)

**Assessment**

In order to find the most effective model to approach this problem, some form of evaluative process is needed. Such qualities as consistency, computational speed, stability, adaptability, etc. are desirable in a mathematical method, yet it is difficult to acquire all in solely one technique. Clearly, the ideal way to assess accuracy would be by measuring the methods against the actual correct tree. The entire true tree is unavailable,
however, for obvious reasons. Estimating phylogenies would be completely unnecessary in the first place.

To try and overcome these obstacles, statisticians and biologists generally employ the bootstrap method as the most common form of evaluating uncertainty in phylogenetic trees. In general, nonparametric bootstrapping involves randomly resampling the population sample being studied with replacement (Miyamota and Carcraft, 1991, p.160). These tree estimates are then compared and their spread gives an idea how large the sampling error is.

Out of a given number of “bootstrap tree” estimates, if all of them share a specific characteristic (i.e. a certain arrangement of species) in common, then this aspect of the trees is given a 100% bootstrap value. It is thought that since all the trees reproduced the same feature with random resampling, then the data supports that characteristic very well and is said to be robust. (The method used is probably quite stable concerning that attribute and is likely to occur again and again in simulation). However, bootstrap trees that differ greatly over specific aspects may imply that there is insufficient data to differentiate between them and various original samples may support other trees and relationships simply due to sampling error.

For example, if we had done a bootstrap on the primate DNA information from the earlier mentioned species, the following unrooted estimate trees might result:
Figure 6: Possible bootstrap trees from primate data. Frequencies of trees noted above.

As can be seen, all three trees agree that gibbons and orangutans most likely split from the same node in one group whereas the humans, gorillas, and chimpanzees share another group. This aspect would be given a bootstrap value of 100% since it is present in all trees. The arrangement of the humans, chimpanzees and gorillas in relation, however, vary and one way over another would not be given a high enough bootstrap value to be considered more likely than any other tree. In general, a bootstrap value of 95% is considered high enough to be well supported.

Overall, the bootstrap's main advantage is that it gives estimates of precision among the trees. (As noted earlier, accuracy is not generally measurable unless the speciation events have been induced experimentally in the lab). It compares the possible trees with one another only and how well they correspond to each other. The disadvantage to this, however, lies in the fact that a tree that the bootstrap method may deem as statistically low in sampling error or a relationship given a high bootstrap value may be completely wrong. There is no real way to tell this, however, so the bootstrap at least gives us an indication of how robust the tree estimates are (Page and Holmes, 1998, pp.219-222).
Improvements

The next step in exploring this topic involves how to enhance the techniques used to find the trees and apply more significant data to them. Doing so better represents the actual state of the information available to us and gives a more “accurate” tree. One of the many improvements that can be made on the discussed models is that biology has shown that the probability of transitions occurring is much greater than the probability of transversions.

![Diagram showing comparison of transitions versus transversions]

Figure 7: Comparison of transitions versus transversions

The models could be changed to reflect this aspect of DNA mechanics by including another rate parameter ($\gamma$) which could represent the rate between $T \leftrightarrow A$ and $G \leftrightarrow C$. This would create the rate matrix:

$$
\begin{array}{cccc}
A & G & C & T \\
A & -(\alpha + \beta + \gamma) & \alpha & \beta & \gamma \\
K = G & \alpha & -(\alpha + \beta + \gamma) & \gamma & \beta \\
C & \beta & \gamma & -(\alpha + \beta + \gamma) & \alpha \\
T & \gamma & \beta & \alpha & -(\alpha + \beta + \gamma)
\end{array}
$$

18
which is also known as the Kimura 3ST model. (Balding et al, 2001, p.465) The mathematics of this model are much more complicated, however. Increasing the number of parameters does generally improve the genetic distance estimate, yet at the same time, the variance increases.

Another improvement that could be made is to take into account the fact that several codons can code for the same amino acid, as discussed earlier. As a result, a single site mutation at a specific codon may have no effect at all on the genetic distance between two species, yet the models will give a greater genetic distance than necessary. To account for this problem, the creation of an entirely new transition matrix based upon the amino acids instead of single nucleotides could be produced. This would create a 20 x 20 matrix\(^3\) with the probabilities of transferring from one amino acid to another as the elements. Building this matrix would most likely make use of the previous models as the rates of transitions/transversions between nucleotides need to be utilized. Inferring the genetic distance from this “amino acid rate matrix” would be a more accurate portrayal of how far apart two species are from one another.

Some other more obvious improvements to these models would take into account the high likelihood that the frequencies of nucleotides aren’t equal in each species, (which was one of the assumptions we made). Some species may have more T’s present in their genetic makeup while others may have an over abundance of G’s. Also, some species have a longer genome than others. Our model does not account for the deletion of nucleotides

\(^3\) The STOP codon would not be included in this matrix as it is not ultimately part of the protein.
Conclusion

Although many enhancements can be made to these models, the hope is that the original models are close enough to the actual mechanism of evolution; that any occasional deviation here and there is very slight. The models produce a general evolutionary tree that serves the broad purpose of finding relationships between species, even though the exact distances and correlations may not be quite precise. Only when greater achievements and collaborations through the sciences of math, biology, archaeology, genetics, and chemistry will a more in depth evolutionary tree be found. Other models not mentioned here have undertaken such challenges and have incorporated new research findings into their material. Our model can at least help by serving as a basis. It gives a foundation to start with and improve upon with greater detail as new discoveries transpire.
References


