Molecular Evolution, Substitution Models, and Phylogenies

T. P. Speed, K. J. Kechris, and S. K. McWeeney

1 Molecular Evolution

1.1 Introduction

The sequencing of DNA and polypeptides has now become easy and fast. As sequence databases grow, they are not only providing information for biologists, but also introducing new challenges for computer scientists, mathematicians and statisticians. For example, interesting problems include locating genes in long strings of nucleotides or determining protein structure from amino acid sequences.

These databases also provide an opportunity to study the relationship between the sequences in an evolutionary context. Research in this area will not only give insight into the divergence of molecules, but also into the process of evolution. These notes will be an introduction to the basic mathematical and statistical methods behind the study of molecular evolution. See Li (1997) for a more thorough survey of the different approaches.

1.2 Evolution and Mutations

Before jumping into models about sequences, there will be a brief summary about evolution and the introduction of mutations at the gene level, called substitutions. The information regarding the development of an organism is stored in its DNA. DNA is inherited from the parent organisms more or less unchanged, but mutations are introduced over time in surviving lineages at fairly steady rates. Mutations occur when there are errors in DNA replication or repair. Although there are enormous numbers of possible mutations, most are not observed in extant populations because of natural selection. In population genetics, natural selection means the differential rates of reproduction. It causes substitution of selectively advantageous genes for less advantageous ones in the population. Thus, many mutations are not inherited to subsequent populations because they have affected the lifespan or fertility of the organism. Therefore, molecular evolution is dominated by mutations that are neutral from the standpoint of natural selection. When a mutant allele of a gene replaces the predominant allele in a population, this process is called gene substitution.

1.3 Substitution Models

At the simplest level, the proportion of different nucleotides p can be used to measure the evolutionary divergence between two aligned sequences. This proportion can be estimated by,

$$\hat{p} = \frac{n_d}{n} \tag{1}$$

where n is the total number of nucleotides in the sequence and n_d is the number of different nucleotides for the pair. If p is small, $\hat{p}$ is approximately equal to the number of substitutions per

site. If p is large, there may be multiple and back substitutions at a given site, so $\hat{p}$ will give an underestimate of the number of substitutions. There are a number of correction methods, based on probabilistic models, which attempt to give a more accurate estimate for p (Jukes and Cantor 1969; Kimura 1980, 81; Hasegawa et al. 1985; Tamura and Nei 1993).

1.4 Jukes-Cantor Model

The Jukes-Cantor (J-C) model is the simplest of these models. It will be derived by first introducing a simple Poisson model for substitutions and then extending it to a continuous time Markov process. The derivation will not be rigorous. Please refer to Jukes and Cantor (1969) and Taylor and Karlin (1994) for an introduction to stochastic processes.

1.4.1 Derivation

Suppose the distribution of the number of substitutions s is a Poisson random variable with mean λt. The rate of substitutions (relative to the unit of time) at a given site is λ. The probability of $s > 0$ at a site in a time period t is

$$Pr(s) = \frac{\exp^{-\lambda t}(\lambda t)^s}{s!}. \tag{2}$$

Thus, the mean number of substitutions during t units of time is λt. The probability of no changes occurring at a site is

$$Pr(s = 0) = \exp(-\lambda t) \tag{3}$$

and the probability for at least one substitution is

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

When t is small (so that multiple and back substitutions are rare), these probabilities can be approximated by,

$$Pr(s = 0) \approx 1 - \lambda t \tag{4}$$

and

$$Pr(s \neq 0) \approx \lambda t.$$

These probabilities can be seen as the infinitesimal probabilities relating to a Markov process. In the context of DNA sequences, this process is a 4-state Markov chain. Let p_{ij} be the transition probability that the next state (nucleotide) is j given that the current state is i,

$$p_{ij} = Pr(next\ state\ S_j \mid current\ state\ S_i).$$

The Markov assumption is that once the state at the current time is specified, no additional information is needed about the past for predicting the status of the next state.

$$p_{ij} = Pr(next\ state\ S_j \mid current\ state\ S_i\ \&\ any\ configuration\ of\ states\ before\ this).$$

Let $P = \{P\}_{ij}$ denote the matrix of transition probabilities for this Markov process. The following property holds (for appropriate P)

$$P(t + h) = P(t)P(h). \tag{5}$$

In the J-C model, the rate of substitution of one nucleotide i by another j is constant (λ is the same for all substitutions i to j). Extending the results from the simpler Poisson model to the 4 nucleotide case, for h small, the transition probabilities are approximated by,

$$P(h) \approx I + Qh \tag{6}$$

where Q is the infinitesimal matrix for a continuous Markov process and

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

By plugging (6) into (5) and taking the limit as $h \to 0$, P solves

$$P' = PQ. \tag{7}$$

With initial condition, $P(0) = I$, the solution of this first-order differential equation is

$$P = \exp(Qt). \tag{8}$$

By simplifying this expression,

$$P = \frac{1}{4}1'1 - \frac{Q}{4\lambda}\exp(-\lambda t).$$

Thus, the chance of a nucleotide i changing into nucleotide j in time t is $P_{ij} = \frac{1}{4}(1 - \exp(-4\lambda t))$ and there are 12 different ways this can occur. The chance of the nucleotide staying the same is $P_{ii} = \frac{1}{4} + \frac{3}{4}\exp(-4\lambda t)$. These values for P constitute the J-C Model. One variation of this model, the Kimura 2-parameter model (Kimura, 1980), takes into account that transitions (A $\leftrightarrow$ G, C $\leftrightarrow$ T) occur at a higher rate than transversions (A $\leftrightarrow$ T, G $\leftrightarrow$ T, A $\leftrightarrow$ C, C $\leftrightarrow$ G).

1.4.2 Examples and Extensions

The following example with just one site will illustrate these concepts. Figure 1 shows the nucleotide of the second position in the sequence of α-globin Alu 1 for both the orangutan and the human. Assuming the common ancestor has A, G, C and T with probability $\frac{1}{4}$, and that time t has passed from the ancestor, then the chance of the nucleotides differing at this site for both sequences is $P = 12(\frac{1}{4} \times \frac{1}{4}(1 - \exp(-8\lambda t))) = \frac{3}{4}(1 - \exp(-8\lambda t))$. If all sites are assumed to be independent, P is estimated as $\hat{P} = \frac{n_d}{n}$ where n is the total number of nucleotides in the sequence and n_d is the number of different nucleotides.

The two parameters λ and t are not identifiable. Let k be the number of nucleotide substitutions of the pair sequences per site for t units of evolutionary time. Then, $k = 3\lambda \times 2t = -\frac{3}{4}\log(1 - \frac{4}{3}P)$ and is estimated by plugging in $\hat{P}$. This is the corrected or adjusted fraction of differences under this model.

Although mutations occur at the DNA level, it is also important to study the evolution of proteins. Changes due to mutation are primarily observed through proteins because they are the molecules that directly determine the morphological characteristics and the physiological functions of the organism. The J-C model can also be applied to proteins. The model would specify that the rate of

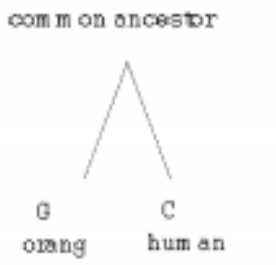

Figure 1: Second position in α-globin Alu 1.

amino acid substitutions is constant for all possible substitutions. The corrected distance k would be estimated as $\hat{k} = -\frac{19}{20}\log(1-\frac{20}{19}\hat{P})$. See Figure 2 for an example of a pairwise distance matrix for β globins. It includes both observed number of differences between the sequences and the corrected distances based on this model.

The relationship between the corrected differences and divergence times has been studied for some families of proteins (cytochrome C, hemoglobin). The graph of $\hat{k}$ against time t for these examples is approximately linear and the slope, to a scale factor, is an estimate of λ.

Substitution matrices are often used to query or align sequences. For the most part, there is no knowledge about λ and/or the divergence times t among the sequences. t is convenient to measure time in terms of the time length for .01 substitutions per site, $k = .01$. The time t such that $k = .01$ is called a PAM unit. PAM is the acronym for point accepted mutations. The PAM(1) transition matrix is $P(t)$ for $t = 1\ PAM$. In the J-C model, $t = \frac{1}{300\lambda}$ is 1 PAM. In effect, increasing PAM units is equivalent to increments in time t. By using PAM units of time, there is a uniform measurement of time for all families, regardless of their substitution rate.

Before concluding this section, it is important to clarify that the previous extensions are based on the truth of the substitution model. These models are overly simplified versions of a very complex and variable process. They are based on many assumptions, such as the constancy of rates and uniformity at sites, of which many cannot be tested. Nevertheless, these models and the derived substitution matrices have served well for exploring the relationship among sequences.

2 Phylogenetic Methods

2.1 Introduction

Determining the evolutionary lineage of organisms is one of the major problems in the study of evolution. In the past, phylogeny was inferred by examining fossil records and morphological characters. In the 1960's, when molecular techniques were introduced to the field, scientists used the evolution of the organisms' macro-molecules to reconstruct their evolutionary history. This work is called molecular systematics. It is based on the assumption that sequences from different species have descended from some ancestral gene in a common ancestral species. Thus, divergence between sequences is a result of speciation. These genes are essentially the same and are called orthologues. The assumption may not hold because of gene duplication, another mode of evolution. Genes which

DISTANCES between protein sequences
Calculated over: 1 to 147
Below diagonal: observed number of differences
Above diagonal: estimated number of substitutions per 100 amino acids
Correction method: Jukes-Cantor

	hum	mac	bov	pla	chi	sha
hum	—	5	17	27	37	108
mac	7	—	18	27	36	102
bov	23	24	—	32	46	110
pla	34	34	39	—	34	106
chi	45	44	52	42	—	98
sha	91	88	91	90	87	—

Figure 2: β globins: Corrected pairwise distances.

have diverged from a common ancestor by gene duplication are called paralogues. They are different genes in the same organism. Figure 3 presents a UPGMA tree (see next sections for tree methods) based on orthologues of the same gene (β hemoglobin). Figure 4 presents a Neighbor-Joining tree based on paralogues, the $\alpha, \beta, \gamma, \delta, \epsilon$ Hemoglobin chains and Myoglobin.

2.2 Constructing Trees

Phylogenetic trees are often used to represent the history of molecules. The root of the tree is regarded as the common ancestor of all the sequences. Some tree-building methods infer the location of the root but others give no information. Internal nodes in the tree represent divergence points. The length of each edge of the tree is determined by some measurement of distance and represents the amount of evolutionary divergence between sequences. These lengths may be informative, but do not necessarily correspond to evolutionary time periods, because proteins evolve at different rates in different organisms.

There are a number of procedures for building trees. The two main approaches are (i) distance-based methods, which work from pairwise distances between the sequences (e.g. UPGMA and Neighbor-Joining), (ii) character-based methods, which work directly from the multiply aligned sequences (e.g. parsimony methods and likelihood approaches).

2.3 Distance Methods

There are many ways of defining distance, d_{ij}, between a pair of sequences i and j in a data set. The simplest, when dealing with aligned sequences, takes the number of sites that differ between the two sequences. Another measure may be defined by a model of substitution of residues over evolutionary time (see the previous section). In the methods described below, it is also assumed that the distances are additive. When this is true, the distance between pairs of leaves is the sum

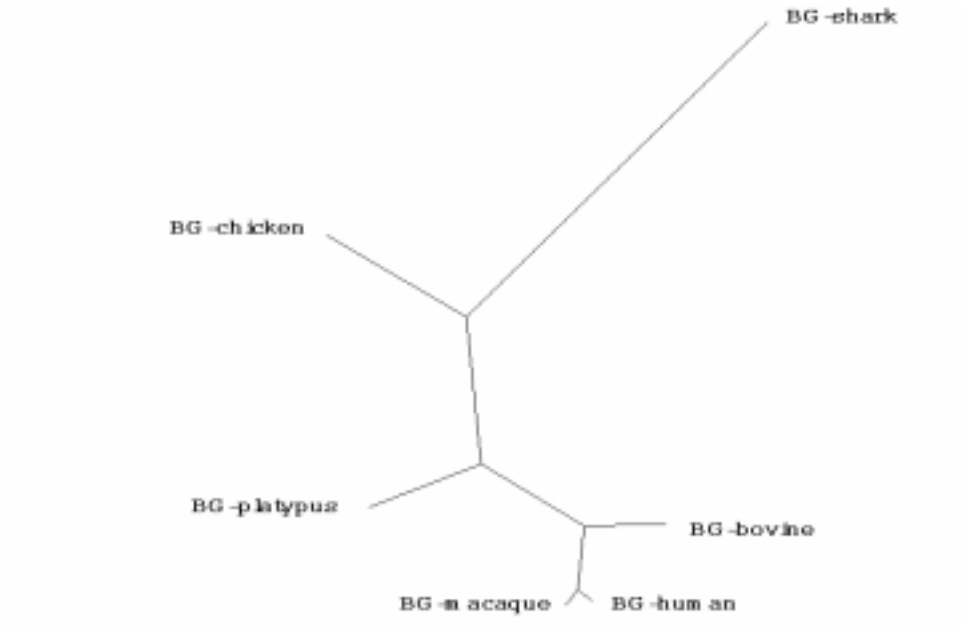

Figure 3: UPGMA tree of β hemoglobin.

of the edges on the path that connects them. Additivity may not hold if multiple substitutions have occurred at any of the sites.

2.3.1 UPGMA

A simple distance-based clustering procedure is the Unweighted Pair Group Method using Arithmetic Averages (UPGMA). At each step, sequences which are 'close' are clustered together and a new node is created. It assembles the tree upwards, each new node is added above the others.

The algorithm, as outlined in Durbin et al. (1998), begins with a distance matrix $D = \{d_{ij}\}$ between all sequences and each sequence i is assigned to its own cluster C_i. At each step, a new cluster C_k is formed by the union of clusters C_i and C_j which minimize $d_{C_iC_j}$. The distance between pairs of clusters is,

$$d_{C_iC_j} = \frac{1}{\mid C_i \mid\mid C_j \mid} \sum_{p \in C_i \ q \in C_j} d_{pq}. \tag{9}$$

$\mid C \mid$ is the number of leaves in the cluster C. A new node k is placed in the tree, with daughter nodes i and j, at height $d_{ij}/2$. Clusters C_i and C_j are removed from the set of clusters and C_k is added with distances d_{kl} between all other clusters l. This procedure is repeated until only two clusters remain, C_i and C_j, and the final node, the root, is placed at height $d_{ij}/2$. See Figure 3 for an example of a tree constructed by this method.

2.3.2 Neighbor-Joining

The underlying assumption of the UPGMA method is that the sequences diverge at a constant rate. The edge lengths can be seen as times measured by a molecular clock. When there are

sequences from organisms with different substitution rates, UPGMA will incorrectly reconstruct the tree topology. A more recent method called Neighbor-Joining (NJ) (Saitou and Nei, 1987) will produce a more accurate tree if this assumption does not hold. Figure 4 presents is a tree built by the NJ method.

The NJ method beings with a starlike tree with no interior nodes. Unlike UPGMA, the algorithm keeps track of the nodes rather than the clusters of sequences. All sequences i are assigned as leaf nodes. Pairs i and j are 'neighbors' if they minimize,

$$D_{ij} = d_{ij} - (r_i + r_j) \tag{10}$$

where

$$r_i = \frac{1}{\mid L \mid -2} \sum_{k \in L} d_{ik}.$$

L is the set of leaves in the current tree. The number of leaves is $\mid L \mid$. A parent node k, to neighbors i and j, is added with edges to those leaves with lengths $d_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j)$ and $d_{jk} = d_{ij} - d_{ik}$, respectively. The leaves i and j are removed from L and k is added as a leaf. This procedure continues until there are only two leaves, i and j left. The length of the edge between them is d_{ij}. The term r is the average distance (scaled by a constant) to all other leaves. Thus, two sequences may have a relatively large distance between them, but are still neighbors because both are also 'far' from all the other sequences (the r's are large). By subtracting the r's, NJ can reconstruct a tree topology which has different sized edges.

This procedure is closely related to the method of Minimum Evolution. The minimum evolution criteria is to choose the tree with the smallest sum of branch lengths. The NJ method is a fast approximate method for finding a tree that satisfies this criteria. It guarantees that a short tree is found, but not the shortest (Gascuel, 1994).

2.4 Parsimony Methods

Parsimony methods were among the first methods for inferring phylogenies. The central idea is that the preferred evolutionary tree requires the smallest number of evolutionary changes to explain the differences observed among the taxa under study. As an example, we will look at the 4 taxa case, in which there are 3 unrooted trees (Figure 5).

We examine the sequences to look for informative sites, i.e., those sites that will favor one tree over the other.

	Site				
	1	2	3	4	...
1	A	G	G	A	...
2	A	G	G	G	...
3	A	A	C	A	...
4	A	A	C	G	...

Site 1 is uninformative, as there is no change in any of the sequences. At site 2, the most parsimonious explanation is that there is a change from G to A which favors the first tree in Figure 5. At site 3, once again the most parsimonious explanation is one change from G to C and again the

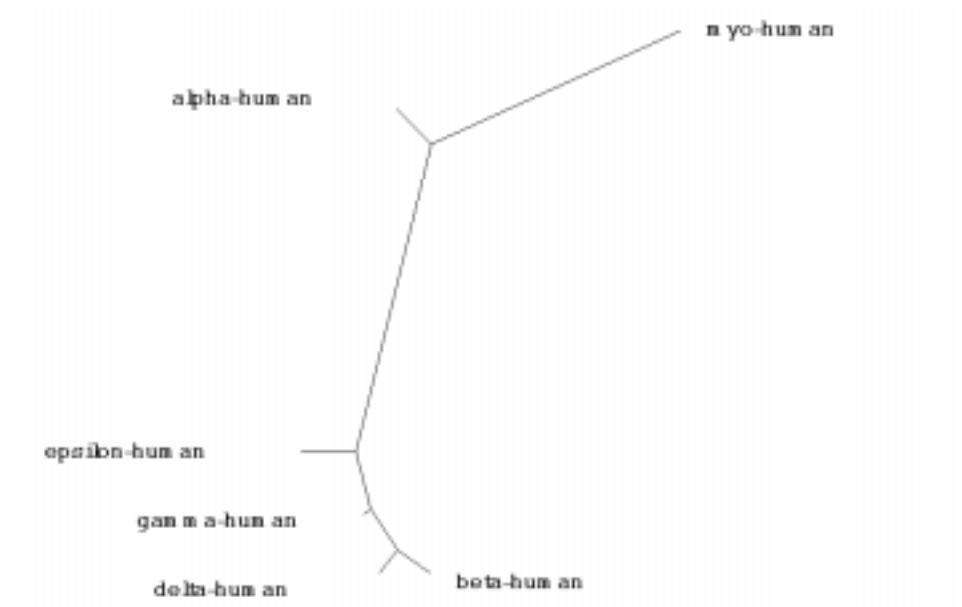

Figure 4: Neighbor-Joining tree for human globins.

most parsimonious tree is the first tree. For both sites 2 and 3, the other 2 trees would require 2 changes. At site 4, the 2nd tree is the most parsimonious, only requiring once change from A to G. Based on these sites, the first tree is the most parsimonious with 2 informative sites favoring it. As the number of taxa being investigated increases, so will the number of trees that must be considered. Regardless of the complexity, the main idea is still to infer the minimum number of substitutions/changes for a given tree.

Although parsimony makes no explicit assumptions, there is the critical assumption of the parsimony criterion that a tree that requires fewer substitutions/changes is better than a tree that requires more (Li 1997). This can be contrasted with likelihood methods that make explicit assumptions about the rate of evolution and patterns of nucleotide substitution.

2.5 Likelihood Methods

Likelihood methods for phylogenies were first introduced by Edwards and Cavalli-Sforza (1964) for gene frequency data. Neyman (1971) applied likelihood to molecular sequences and this work was extended by Kashyap and Subas (1974). Felsenstein (1973,1981) brought the maximum likelihood framework to nucleotide-based phylogenetic inference. As an example, we will compute the likelihood for a given tree using DNA sequences, but these procedures can be used for other characters as well. This example is borrowed from Felsenstein (1981).

We have a set of aligned sequences with m sites. To compute the probability of a tree, we must have a phylogeny with branch lengths and an evolutionary model giving the probabilities of change along the tree. This model will allow us to compute the transition probabilities $P_{ij}(t)$, the probability that state j will exist at the end of a branch of length t, if the state at the start of the branch

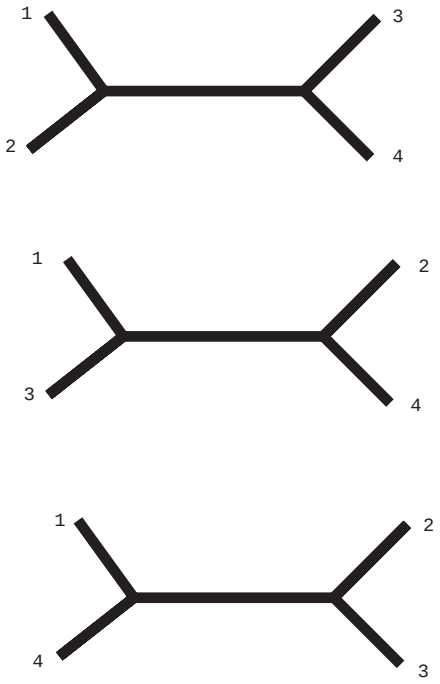

Figure 5: In the 4 taxa case, there are 3 possible trees

is i. We must remember the t is a measure of branch length, not time. There are two central assumptions: (1) evolution at different sites on the tree is independent (and identically distributed) and (2) evolution in different lineages is independent. This gives us the Markov property down trees (cf. pedigrees).

The first assumption allows us to express the probability as a product, with one term for each site:

$$L = Pr(D|T) = \prod_{i=1}^{m} Pr(D^{(i)}|T)$$

where $D^{(i)}$ is the data at the i-th site. Therefore, we need to know how to compute the probability at a single site. If we have the tree and the data at a site, the probability is the sum, over all possible nucleotides that may have existed at the interior nodes of the tree, of the probabilities of each scenario.)

$$Pr(D^{(i)}|T) = \sum_x \sum_y \sum_z \sum_w Pr(A, C, C, C, G, x, y, z, w|T)$$

Based on the Markov property (2) above, we can decompose the probability on the right side of the equation into a product:

$$\begin{aligned} Pr(A, C, C, C, G, x, y, z, w|T) &= Pr(x)Pr(y|x, t_6) \\ &\quad \times Pr(A|y, t_1) \times Pr(C|y, t_2) \times Pr(z|x, t_8) \times Pr(C|z, t_3) \\ &\quad \times Pr(w|z, t_7) \times Pr(C|w, t_4) \times Pr(G|w, t_5) \end{aligned}$$

If we assume that evolution has been proceeding for a long time according to the model of substitution that we are using, then the Pr(x) is the equilibrium probability of base x under that model.

However, there are still a large number of terms in the previous equation, such that for each site we sum $4^4 = 256$ terms. The number of terms will rise exponentially with the number of species. On a tree with n species, there are n-1 interior nodes, each of which can have one of 4 states. If we need 4^{n-1} terms, then for 10 taxa, this would be 262,144 terms!

In order to make the computation more realistic, Felsenstein (1973, 1981) introduced a "pruning" method, which is a version of the "peeling algorithm" introduced by Hilden (1970), Elston and Stewart (1971) and Heuch and Li (1972). As we saw in a previous week, these in turn are analogous to the HMM forward-backward algorithms, and all are special cases of the idea of dynamic programming over a directed acyclic graph (of Week 14). The method can be derived simply by trying to move the summation signs as far right as possible and enclose them in parentheses where possible.

$$\begin{aligned} Pr(D^{(i)}|T) &= \sum_x Pr(x)\Big(\sum_y Pr(y|x,t_6)Pr(A|y,t_1)Pr(C|y,t_2)\Big) \\ &\times \Big(\sum_z Pr(z|x,t_8)Pr(C|z,t_3)\Big(\sum_w Pr(w|z,t_7) \\ &\times Pr(C|w,t_4)Pr(G|w,t_5)\Big)\Big) \end{aligned}$$

The pattern of parentheses and terms for tips in this expression is

$$(A,C)(C,(C,G))$$

which corresponds exactly to the structure of the tree. The flow of computations is from the inside of the innermost parentheses outwards, with a flow of information down the tree).

We will utilize the conditional probability of a subtree, $L_k^{(i)}(s)$, which is the probability of everything observed from node k on the tree on up, at site i, conditional on node k having state s (cf. HMM backward probabilities). In the previous equation, the term

$$\Big(\sum_w Pr(w|z,t_7)Pr(C|w,t_4)Pr(G|w,t_5)\Big)$$

is one of these quantities, i.e. it is the probability of everything seen above that node, given that the node has base w. There will be four such quantities, one for each of the values of w. The key to the pruning algorithm is that once the four numbers are computed, they don't need to to be recomputed again. The algorithm is a recursion that computes $L^{(i)}(s)$ at each node of the tree from the same quantities in the immediate descendent node. The algorithm is applied starting at the node which has all of its immediate descendents being tips. Then it is applied successively further down the tree, not applying it to any node until all of its descendents have been processed. The result is $L_0^{(i)}$ for the bottom node. The evaluation of likelihood is then completed for this site by making a weighted average of these over all four bases, weighted by their prior probabilities under the model:

$$L(i) = \sum_x \pi_x L_0^{(i)}(x)$$

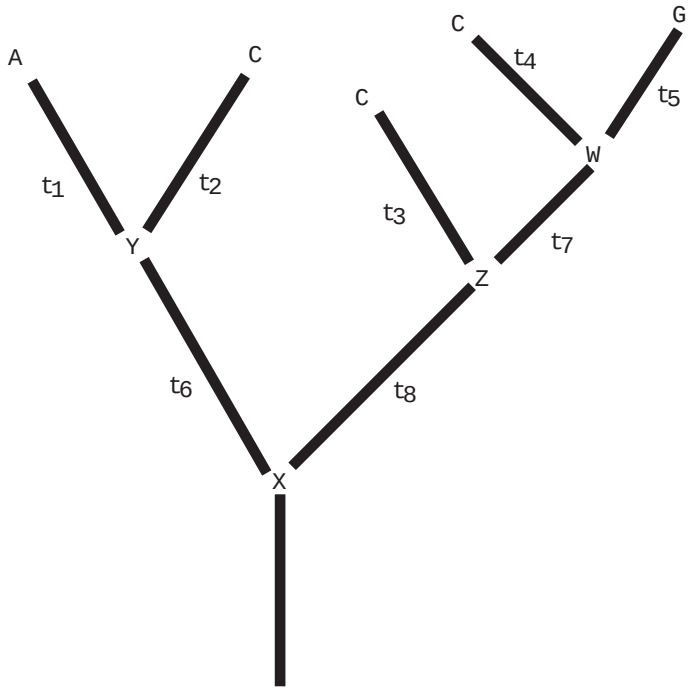

Figure 6: An example tree for Maximum Likelihood

Estimation of branch lengths (t_i) can be done by a form of the EM algorithm adapted to trees in the same way as the HMM estimation algorithm is adapted to (linear) Markov chains. Forward (above) and backward (below) probabilities can be combined to give joint probabilities of states at adjacent nodes given the data, and these become the E-terms in the EM iteration. Details can be found in Felsenstein (1981) or worked out independently by analogy with the HMM case.

References

[1] Edwards, A. & Cavalli-Sforza, L. 1964. Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification.* Systematics Association Publ. No. 6.

[2] Durbin, R. et al. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, New York.

[3] Felsenstein, J. 1973. *Syst. Zool.* 22: 240-249.

[4] Felsenstein, J. 1981. *J. Mol. Evol.* 17: 368-376.

[5] Gascuel, O. 1994. *Mol. Biol. Evol.* 6: 961-963.

[6] Hasegawa, M, Kishino, H. & Yano, T. 1985. *J. Mol. Evol.* 22: 160-174.

[7] Hillis D.M., Moritz C. & Mable, B.K. (Editors). 1996. *Molecular Systematics* 2nd ed. Sinauer Associates, Massachusetts.

[8] Jukes, T. & Cantor, C. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism.* Academic Press, New York.

[9] Kimura, M. 1980. *J. Mol. Evol.* 16: 111-120.

[10] Kimura, M. 1981. *Proc. Natl. Acad. Sci.* USA 78:454-458.

[11] Li, W. 1997. *Molecular Evolution.* Sinauer Associates, Massachusetts.

[12] Saitou, N. & Nei, M. 1987. *Mol. Biol. Evol.* 4: 406-425.

[13] Tamura, K. & Nei, M. 1993. *Mol. Biol. Evol.* 10: 512-526.

[14] Taylor, H. M. & Karlin S. 1994. *An Introduction to Stochastic Modeling.* Academic Press, San Diego.