PHYLOGENETICS (ARNE MOOERS, UNIVERSITY OF AMSTERDAM, 1999)

Chapter 10 of Evolutionary Analysis (Freeman and Herron, 1998, published by Prentice Hall) offers a fine overview of phylogenetics. These notes are designed to be a supplement to this chapter, and will make little sense if you haven't read it or some other equivalent introductory literature. The ideas found here can be gleaned from Molecular Systematics (Hillis et al., 1996), chapters 11 and 12 (by Swofford et al.), Molecular Evolution (Li, 1997), both published by Sinauer Associates, and Molecular Markers, Natural History and Evolution (Avise, 1994), published by Chapman and Hall. In these notes, I will use the word tree as the general term for a phylogenetic hypothesis.

1 OPTIMALITY CRITERIA VERSUS ALGORITHMS

An important distinction exists between **optimality criteria** and **algorithms**. In the context of phylogenetics, an algorithm is specific sequence of steps that leads to a single tree. An optimality criterion, on the other hand, does not specify how a tree is to be inferred, but offers the criterion for comparison among contenders. One could use a recipe to make pancakes, and just make them, or one could use trial and error, keeping the food that fit some lengthy description of what a pancake should look like (e.g. the thinest, roundest, sweetest thing made with milk, butter, eggs, sugar and flour). Algorithms are fast, but if you are a bad cook or have a poor recipe any one outcome can be sub-optimal.

UPGMA and Neighbor-Joining are two algorithms. Algorithms such as these are fast, and were popular in the 1980's. Because these algorithms do not consider alternative trees, we do not know how much better the tree produced by the algorithm is than its competitors. If the assumptions of the algorithm are not met, the single answer may be incorrect. **Maximum Parsimony** and **Maximum Likelihood** are general optimality criteria. Using optimality criteria necessitates comparing (in theory) all possible trees; this is impossible for large datasets (i.e. over 15 tips) However, there are algorithms which guarantee that an optimality criterion can be be used (e.g. the branch and bound algorithm). Most trees today are created using such a hybrid approach. Just because an optimality criterion has been specified, however, doesn't mean that it has been reached - section 10.4 of Evolutionary Analysis details an example of this. The strength of using an optimality criterion over a simple algorithm is that one has (at least a subset of) all possible trees at hand, and can evaluate different trees - how much better is the best tree than its nearest competitor, for instance. When you see a tree in the literature, one of the first questions to ask yourself is "what are the purely algorithmic steps that were involved in producing this tree, and were these algorithm(s) appropriate (i.e. unlikely to be misled by the data)?" The answer, of course, will depend on the dataset.

2 MAXIMUM PARSIMONY

The concept of parsimony in science maintains that simpler hypotheses should be preferred to more complex ones. Under the criterion of parsimony in phylogenetics, attributes shared between taxa (e.g. two species each having incisors) are assumed to have been inherited from a common ancestor rather than having evolved independently. When character conflicts occur (e.g. one character is shared by A and B but not by C, while another is shared by A and C to the exclusion of B), then we must invoke **some second process: independent evolution of the character state, ancestral polymorphism followed by differential loss, etc.** In general, parsimony methods in phylogenetics operate by choosing the tree with the minimum number of character state changes. We refer to the sum of the number of changes inferred to have occurred in a tree its length. A tree which minimizes the total number of changes (e.g. the tree which places the two species with incisors together such that the evolution of incisors only occurred once in a common ancestor, along a single branch, rather than twice independently in different parts of the tree) also minimizes the number of homoplasies (character states shared for reasons other than common ancestry). This minimizing criterion is the essence of Maximum Parsimony (which I will abbreviate MP). Researchers do not believe that evolution is necessarily parsimonious, but they do believe that MP offers an objective criterion for considering alternate trees. The major question involves how individual characters change: which charaters are most likely to evolve such that invoking maximum parsimony is most likely to describe their evolution? Do we have a priori ideas about which characters might evolve more rapidly, which sorts of change (i.e. loss of a complex character versus its gain) might be more likely, and which characters might be more likely to evolve more than once (e.g., due to natural selection)?

We can frame all these questions in general mathematical terminology: From the set of all possible trees, find all trees (T) such that:

$$L(T) = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j diff(x_{k'j}, x_{k''j}) \text{ is minimized.} \quad (1)$$

Here L(T) is the length of the tree, B is the number of branches, N is the number of characaters, k' and k" are the the two nodes incident to the branch k (the beginning and end of the branch k), $x_{k',j}$ and $x_{k",j}$ represents the optimal character states at the ends of branches (optimal in that, in those states, the minimum number of changes are inferred), and diff(y,z) is the cost of changing from state y to state z along any branch. w_j is the weight assigned to character j. The cost is the amount the tree changes in length when a change occurs (e.g. a single change for a simple character might have cost = 1, while a more complex character may be given diff(y,z)=2). The weight is the relative importance assigned to a character. The minimal length tree is that which has the smallest sum of costs over all branches and all characters.

By inspecting this equation, we can start to appreciate what is involved in inferring a

tree under MP. If we believe that all characters j are equally likely to be have arisen only once, then we can give them all the same weight $w_j = 1$ (for j= 1 to N). If we believe that some character j is twice as likely to undergo homoplasy, we would place *less* emphasis on it: we would give it a weight $w_j = 0.5$, so that it contributes less to the total tree length. If we believe that another character is very likely to have arisen only once (i.e. homoplasy is very unlikely), we would give it a high weight (e.g. $w_j = 5$). If we weighted the character "precence/absence of a placenta" with weight 5, then the tree ((echidna, kangaroo),(horse,hippopotamus)), which places the placental mammals together and the marsupials together, would have a cost of 5 for this character - one change along the branch between the placental mammals and the marsupials; the tree ((echidna, horse),(kangaroo,hippopotamus)) would have a cost of 10 (changes along two branches). What would $L(T)_j$ be on the tree ((echidna,horse),(mouse,(kangaroo,hippopotamus)))?

The other important part of the equation is $diff(x_{k',j}, x_{k'',j})$. If all types of changes are equally likely, then the cost would be equal for all characters along all branches (i.e. $diff(x_{k',j}, x_{k'',j}) = 1$ for all $x_{k'}$, $x_{k''}$, and all j). Some sorts of changes might bear a higher cost, however. For instance, consider a character that has three states - e.g. no fenestrae (extra holes besides eyesockets) in the skull, one fenestrum, and two fenestrae. We might have reason to believe that the two fenestrae state evolved from the one fenestrum state. This would mean, if we assumed it was as probable to go from having no holes in the head to one as from one to two, that, for character j = number of holes in head, if k' = no holes and k" = two holes, $diff(x_{k',j}, x_{k'',j}) = 2$, as we assume that no holes first evolved into one hole, and then into two. In this example, we say that holes in the head is an "ordered" character. It need not be so - if we have a trait such as feather colour (e.g. white, red or blue), we might believe that changes among any of the colours are equally likely, and so $diff(x_{k',j}, x_{k'',j}) = 1$ regardless of the states k' and k".

Assumptions about character transitions are often summarized in a character **step matrix**. This is a representation of the costs associated with different sorts of transitions along a branch. So, for example, for character j = number of holes in the head, the matrix might look like:

Character state	no holes	one hole	two holes
no holes	0	1	2
one hole	1	0	1
two holes	2	1	0

This is a step matrix for an ordered, fully reversible character, and the entries represent $diff(x_{k',j}, x_{k'',j})$ for the the row k' and column k" for character j. For an unordered character, the entries would all = 1.

Both character weights and character transitions costs can be specified for every character in a dataset. Y ou will manipulate both in the practical exercise. It is important to remember that, given character conflict (homoplasy), changing these parameters will change L(T), and so is likely to affect which tree is chosen by the MP optimality criterion. Some purists feel that setting $w_j = 1$ for all j and treating all

characters as unordered (diff($x_{k',j}$, $x_{k'',j}$) = 1 for all $x_{k',j}$, $x_{k'',j}$) is somehow inherently superior to other schemes. While it may better represent the assumptions of the researcher who chose the characters to study, it is only superior in that, because it involves the simple concept of equality, it is more comfortable. For molecular characters, for instance, it may make better sense to weight different codon positions differently, and to posit that certain types of changes are more likely than others (e.g. second versus third codon positions and **transitions** versus **transversions**). The practise of changing parameter values after looking at an inferred tree, in order to get a "better" answer, is, however (common and) dubious.

Well before equation 1 was created, different weighting and cost functions were given specific names, recognizing venerable scientists. So Fitch parsimony assumes that (diff($x_{k',j}$, $x_{k'',j}$) = 1 for all characters and states. Wagner parsimony assumes that characters are measured on an interval scale, as in the holes-in-head example above. Dollo parsimony assumes that characters can only arise once, but can be lost multiple times. Finally Camin-Sokal parsimony, probably the first parsimony method used formally (by Camin and Sokal in a paper published in 1965) assumes that evolution is irreversible, so that characters can never revert to previous states (what would the cost matrix for such a set of assumptions look like?). These names are now little more than shorthand for a particular set of costs and weights in the generalized parsimony function.

Maximum Parsimony seems to be a simple and straightforward strategy for inferring evolutionary trees, and it is the method preferred by most practising systematists who work with morphological characters. However, the greatest increase in systematics research in the past decades has come from another set of characters, namely amino acid and DNA sequences. Here the characters might be thought of as positions, and the character states A, C, T, or G for sequence data or one of the 20 amino acids for protein data. Because the states are the same for most of life, we can look for generalities. This search has led to new methods for inferring relationships, namely the construction of explicit genetic models which allow us to infer trees using distance and maximum likelihood methods.

3 GENETIC MODELS

We can be even more explicit than equation (1) about how we think characters change. Mathematical models can be created which specify the probability of change from one state to another, and these can be used to analyse the distribution of characters among taxa. Because we are so ignorant of how morphological characters change through time, explicit process models for their evolution are in their infancy, and are not commonly used to help assess which trees might be most likely. Molecular characters, however, have been the subject of extensive modelling. Because the character states (4 nucleotides, 20 amino acids) and the processes (e.g. 61 codons of three nucleotides with third base position redundancy) are nearly universal, and the amount of data so exhaustive, generalities are possible. So, for example, if fixed mutations in a stretch of DNA occur at random, than there is a 1/3 chance that, when two mutations occur sequentially at the same site, the second will revert the site back to its original state (e.g. if the original nucleotide was a T, and the first mutation produced a C, the second has a 1/3 chance (T,G,A) of being back to T), and so we will never know that anything happened (so-called replacement substitutions). For many genes, transitions are expected to be much more common than transversions, for purely physico-chemical reasons. We can use this combination of chance and chemistry to produce a set of relative probabilities of observing differences between taxa, and use these to help us choose between alternative trees (would you prefer a tree in which a group shared many transitions or many transversions?).

Many models have been published, each fancier than the last. Let's consider how the simplest one is derived. This will give you insight into how genetic distances and Maximum Likelihood trees are constructed. To do this, we need to remind ourselves of three tenants of basic probability theory. The first is that if p_1 is the probability that event 1 occurs, then $1-p_1$ is the probability that it does not occur. The second is that if event 1 occurs with probability p_1 and event 2 with probability p_2, then the probability that both occur is p_1p_2. Third, if we want to know the probability that either event 1 OR event 2 occur, this is $p_1 + p_2$. If p is the probability that something occurs, 1-p is the probability that it doesn't occur at all then the probability that either something occurs or that it doesn't is $p+(1-p) = 1$, which it must be. The Jukes-Cantor (1969) model assumes that substitutions occur randomly among all four types of nucleotide, and so there is only one parameter to estimate - the substitution rate α. Now consider the probability $p_{A(t)}$, the probability that a given site is in state A at time t. If the site started in state A, the $P_{A(0)}$ is 1. At time 1, the probability that the state is still in state A is $P_{A(1)} = 1-3\alpha$, e.g. if $\alpha = 1/10$ is the chance that it A changes to C, and the same for changes to G or T, then there is a 3/10 chance that it changes to something (event 1 or event 2 or event 3), and a 7/10 chance that it does not change (= $1-3\alpha$). The probability at time = 2 is

$$p_{A(2)} = (1-3\alpha)p_{A(1)} + \alpha(1-p_{A(1)}) \qquad (2)$$

The plus sign signifies an 'event 1' or 'event 2' situation. The first term is the probability that the site was in state A at time 1 times the probability that it remained in state A between time 1 and time 2 (which is just 1-3α). The second term is the probability that the the site changed in time period 1, [= 1- p(site did not change in time period 1)] times the probability that it changed back to A in time period 2 (which is just α). This equation has all the basic structure needed for the Jukes-Cantor model. It is a recursive equation, and we can write, generally

$$p_{A(t+1)} = (1-3\alpha)p_{A(t)} + \alpha(1-p_{A(t)}) \qquad (3)$$

We can then solve for $p_{A(t+1)} - p_{A(t)}$, or $\Delta p_{A(t)}$

$$\Delta p_{A(t)} = -3\alpha p_{A(t)} + \alpha(1-p_{A(t)}) = -4\alpha p_{A(t)} + \alpha \qquad (4)$$

Furthermore, if we consider the process to be continuous rather than discrete, then Δp can be considered the rate of change $dp_{A(t)}/dt$. Equation 4 then becomes a first order linear equation, and we can solve for $p_{A(t)}$. The answer is

$$p_{A(t)} = \frac{1}{4} + (p_{A(0)} - \frac{1}{4})e^{-4\alpha t} \qquad (5)$$

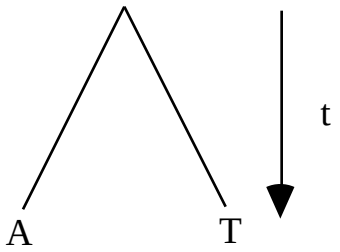
If we know that the site was A, $p_{A(0)} = 1$, and the equation becomes $1/4 + 3/4e^{-4\alpha t}$. Indeed, because the same parameter is estimated for all types of substitutions, we can write this equation in more general form: the probability that a site will remain the same over time t, regardless of whether it was an A, C, G, or T, is

$$p_{ii} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \qquad (6)$$

This is a probability, i.e. the chance that a site will be the same. At the limit (when the rate is very high or alot of time has passed), it will be 1/4 (the second term will become very small). The probability can be converted to an expected observation - at the limit, for instance, 25/100 base pairs should be the same, due to chance back mutations alone. If $\alpha t = 0.2$ (e.g. α is 1×10^{-9} substitutions/site/year and 200 million years have passed) the number of similar sites should be 58/100.

We can use equation 6 to estimate the genetic distance - or the expected number of substitutions - between two sequences. The expected total number of substitutions per site is just $3\alpha t$. Of course, we do not know what α and t are. All we have are the observed number of substitutions. But this can be well-approximated as $1 - p_{ii}$, or one minus the probability of no substitutions. We know what p_{ii} should be from equation (6). We can therefore solve $1 - p_{ii}$ for $3\alpha t$. Before doing this, we need one more piece of information. All the equations so far have considered t as the time that has elapsed between an ancestral and a descendant sequence - i.e. we knew what the sequence was,

and we are comparing it to its descendant at some later time t. However, in reality, what we have are pairs of sequences that have both been evolving for some time t:



Therefore the time elapsed between the sequences is not t, but 2t. So $e^{-4\alpha t}$ becomes $e^{-8\alpha t}$, and the total number of substitutions becomes $2(3\alpha t) = 6\alpha t$. Now, let $p = 1 - p_{ii}$ and let us solve for $K = 6\alpha t$:

$$p = 1 - (\frac{1}{4} + \frac{3}{4}e^{-8\alpha t}) = \frac{3}{4}(1 - e^{-8\alpha t})$$

and

$$8\alpha t = -\ln(1 - \frac{4p}{3})$$

and so

$$K = 6\alpha t = -\frac{3}{4}\ln(1 - \frac{4p}{3}) \qquad (7).$$

This is the Jukes-Cantor correction for estimating the genetic distance between pairs of sequences. And we only needed 7 or so equations. If the observed difference p = 0.1 (i.e. 10% of the sites are different), K, the inferred number of substitutions, is 0.11; if p = 0.25, K = 0.30; if p = 0.4, K= 0.57. As the observed distance gets greater, so does the correction - with more time (or a higher rate - remember, α and t cannot usually be estimated independently), there is more opportunity for substitutions to have occurred which have been hidden by subsequent back mutations.

This can be represented in a **transition matrix**, which is analogous to a step matrix. Here the matrix entries srepresent the (usually instantaneous) probability of a transition from one state to another. For the common Kimura 2-parameter model (Kimura, 1980), the matrix lookes like

Character state	A	T	C	G
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$

where α is the probability of a transition, and β that of a transversion. This is matrix is read in the same order as a MP step matrix, and entries are for substitutions from the row to the column: the bold entry is the probability of a substitution from A to T. (For a symmetrical model such as this one, the reverse substitution has the same value. For other models, it needn't).

It is worth noting here a difference between step matrices in MP and substitution models such as the one above. Under MP, there is no cost (and so no change in treelength) for **no** change along a branch. Indeed, MP searches for the tree with the fewest total number of changes. Under a probabilistic model, on the other hand, there is a finite probability that no change occurs, and this observation figures into the calculation of a tree's likelihood or plausibility (see below). For instance, considering no change for a character which we expect to change often (e.g. a three-fold degenerate codon's third position) on a very long branch might be implausible, and so would have low probability. We might prefer a scenario where a change along that branch for that character did occur.

We can now use these genetc models to calculate genetic distances, and to perform Maximum Likelihood analyses.

4 DISTANCE METHODS

Distance methods refer to a large family of algorithms and optimality criteria that use as their data some measure of the similarity (or dissimilarity) among pairs of taxa, rather than traits of the taxa themselves (I may be 170cm tall, and you 180 - then our dissimilarity score might be 10cm, or 5.7%). Although morphological characters can be converted to distances, the use of distance methods in modern phylogenetics is generally confined to genetic characters. Genetic distance is a quantitative estimate of how divergent two taxa are genetically. I review some of the types of data, assumptions, and optimality criteria commonly used.

One very popular method converts differences between populations in the frequencies of allozymes to a genetic distance (e.g. Nei's genetic distance, Roger's genetic distance). The assumptions behind the various conversions vary, but all assume that the allele frequencies are changing due to genetic drift (and perhaps mutation), and that populations or closely related species whose shared alleles are at similar frequencies are historically closer to each other than groups who have very different allele frequencies.

Data on the number of polymorphic sites in a genome (e.g. those assayed using restriction enzymes when searching for RFLPs) are another source of data that are often converted to distances. Modern methods can estimate a genetic distance which includes information on the frequency of a site in a population (akin to allozymes) and some information on the number and type of DNA site substitutions (like DNA sequence data).

DNA-DNA hybridization, a new method for comparing the DNA of different taxa, was popular in the 1980's. This method compares the melting point of hybrid DNA made up of single strands from two species with the melting point of the nonhybrid species' DNA. The base pairs of hybrid DNA doesn't zip perfectly together, as base pairs at the same site may differ, and so melts at a lower temperature. The difference in temperature between intact single species DNA and hybrid DNA is a measure of a genetic distance between the two species. The advantage of this method is that it compares the entire single-copy portion of a species' DNA, and so the arguments levelled against single gene trees hold no force. The disadvantage is technical - the method is prone to large amounts of measurement error.

DNA or protein sequence data can be converted to distances. The simplest measure is percent difference (e.g. a distance of 3% means that on average 3/100 base pairs for the gene (or 3/100 amino acids for the protein) in question differ between two species), but much more complicated measures exist which make use of explicit models of character change.

If genetic distance was a perfect measure of relatedness, than building distance trees would be fairly simple. Because of measurement error and invalid assumptions, no datamatrix will be perfect. For example, based on mtDNA restriction site data, the genetic distances among *Equus przewalskii*, *E. greyvi*, *E. burchelli* and *E. zebra* are

	E. przewalskii	E. greyvi	E. burchelli	E. zebra
E. przewalskii	-	6.9	7.3	7.5
E. greyvi		-	3.3	4.5
E. burchelli			-	5.6

No tree can match this data perfectly (try to build a tree such that every path between pairs of species equals the actual distances in the table). **UPGMA** is an algorithm which assumes that genetic distances scale linearly with time, such that each lineage is evolving at the same rate and every tip is the same distance from the root. **Neighbor-Joining** does not assume this, but also build a single tree following an algorithm. Other methods evaluate individual trees, in the same way that MP considers all possible trees. Fitch-Margoliash methods search for the tree which minimizes the overall discrepancy between the observed distances and those measured on the tree (the best-fit tree). Minimum evolution methods searches for the tree which is shortest when branch lengths are optimized to best fit the data matrix (the shortest tree). These two trees need not be the same, but will be for well-behaved data.

5 MAXIMUM LIKELIHOOD

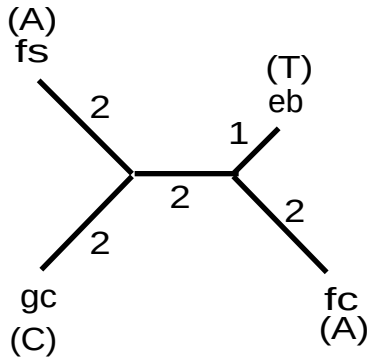
Maximum Likelihood is a general statistical technique for estimating paramaters and comparing hypotheses. It is presently sweeping through the biological sciences, and may become the method of choice for analysing many biological problems where uncertainty must be considered. Because of the computational problems associated with building trees, the use of maximum likelihood techniques in phylogenetics had to await powerful computers. ML estimating techniques are now commonplace, and so some understanding of how they work are needed.

In the context of phylogenetics, three ingredients are needed: data (e.g. DNA sequences for groups whose relationships you want to estimate), a phylogenetic hypothesis (e.g. a candidate tree that you want to evaluate) and a process model of evolution (e.g. the Jukes-Cantor model for substitutions). We can define likelihood as

$$Lik \propto P(R|Hyp, Model) \qquad (8)$$

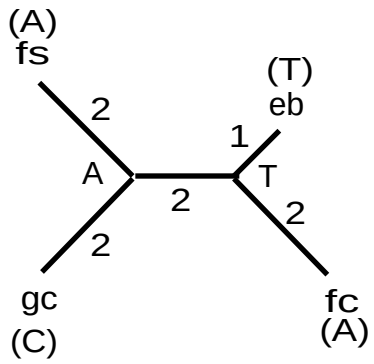
The likelihood is proportional to the probability of observing the results (data), given the hypothesis (the tree) and the model (of substitution). Maximum Likelihood, then, is an optimality criterion that states that the hypothesis (the tree) that returns the highest likelihood should be preferred, just as Maximum Parsimony is the optimality criterion that states that the most parsimonious (shortest length) tree should be preferred. The strategy is similar in both endeavours - in principle, all possible trees are created, each is evaluated, and the most parsimonious (or most likely) is kept.

But how is the likelihood for any particular tree evaluated? Let us take the observations of a single base in a sequence for two species of cat, a bird and a member of the horse family. At position 4 of the cyt. b gene, we may have an Adenosine (A) for *Felis concolor* (fc), an A for *Felis selvestris*, (fs) a Cytosince (C) for *Grus canadensis* (gc) and a Thymine (T) for *Equus burchelli* (eb). These are the data. Next we need an hypothesis of relationship, a tree to evaluate:



Because the branch lengths must be given (represented here with the their lengths), this is one of an infinite number of trees we could evaluate. In order to calculate the likelihood of this tree, we need to do three things:

1) Choose a set of internal states - possible states for the base at the two internal nodes



2) Calculate the probability of observing the transitions along all the branches, or, in this example, $P_{AA(2)}*P_{AC(2)}*P_{AT(2)}*P_{TT(1)}*P_{TC(2)}$. This is the probability of observing A for *Felix selvestris* remaining an A at the first internal node, changing to a C in *Grus canadensis*, changing to a T at the second internal node, and the T then remaining a T in *Equus birchelli* but changing back to an A in *Felix concolor*.
In order to do this, we need a model of evolution. If we take the Jukes-Cantor model, for instance, then equation (6) can give us $P_{AA(2)}$. The Jukes-Cantor model can also furnish us with $P_{AC(2)}$ and any other possibility and we can calculate the probability for this tree and this set of internal nodes.
3) Repeat the exercise for all other possible states for the internal nodes - the two internal nodes could be both A's, or, indeed, both G's. This last possibility is unlikely, and the associated probability would be low, but it is a possibility. Using our knowledge of probability, we can now sum over all these possible sets of internal states (the internal states could be (A,T) as shown here OR (A,A) OR (G,G), OR etc. so we sum the probabilities, using elementary statistics) and this grand total will be the likelihood of this tree for this one character.

We do not have a single observation (a single site) with which to do this, of course, but hundreds or thousands of sites (the cyt. b. gene is over 1000 bp long, for instance). We can do this simultaneously for all the sites, multiplying all the individual probabilities together. As you can appreciate, the amount of calculation effort is tremendous. This then has to be repeated for many candidate trees (e.g. we can repeat the above exercise for a tree with all braches equal to 1, and with the two cats grouped together. Would this tree return a higher likelihood for the character we chose, do you think?). It is then just a matter of computing power and algorithmic tricks to search among the possible trees, both topologies and branch lengths, to find the one that is the most probable, which becomes our Maximum Likelihood estimate.

Maximum Likelihood tree estimation is a statistical method based on explicit assumptions and models. This is a great strength, since all science methodologies should be explicit about assumptions, but also a (perceived) weakness. We know that the models we choose are only poor reflections of reality. For instance, even though the Jukes-Cantor model is sometimes still used, there is overwhelming evidence that different sorts of substitutions occur at different rates (e.g. transitions are more frequent than transversions). We very rarely if ever know what those rates are, though - we cannot observe them directly, since they all happened in the past. Maximum

Likelihood is still of limited value when considering morphological characters, as no reasonable models of trait evolution have been accepted by the community. Some practitioners claim that Maximum Parsimony makes fewer assumptions about how traits change, and is therefore preferrable. There is much ongoing research in this area - MP is much faster than ML and is easier to understand intuitively. Though it has been shown to be positively misleading under certain conditions (i.e. is almost guaranteed to give the wrong answer), we do not know how often those conditions are met with real datasets, if ever.