Phylogenies from restriction sites: a maximum likelihood approach

by

Joseph Felsenstein

Department of Genetics SK-50, University of Washington,
Seattle, Washington 98195

Abstract. – Restriction sites data can be analyzed by maximum likelihood to obtain estimates of phylogenies. The likelihood methods of Smouse and Li (1987), who were able to compute likelihoods for up to four species under a simplified model of base change, can be extended numerically to deal with any number of species. The computational methods for doing so are outlined. The resulting algorithms are slow but take multiple gains and losses of restriction sites fully into account, unlike parsimony methods. They also allow for the failure to observe potential sites that are absent from all species. Analysis of the five-species hominoid data of Ferris et. al. (1981) confirms the pattern found by Smouse and Li with four species – that a chimpanzee-gorilla clade is favored, but not statistically significantly over other tree topologies. A large data set produced by computer simulation has also been analyzed to confirm that the method works properly. The methods used here do not allow for different rates of transitions and transversions. They can be extended to do so, but only at a cost of considerably slower computations. The present method is available in a computer program.

When restriction sites data first began to be collected, it seemed clear to everyone that they would have a limited lifespan. Surely in five or ten years nucleotide sequences would be so easy to obtain that they would make restriction sites data obsolete. But restriction sites data have been so much cheaper and easier to obtain that they have continued to be extensively used. This should impel us to develop better ways of analyzing these data. The present paper shows that the likelihood methods used by Smouse and Li (1987) to estimate phylogenies from restriction sites data can be extended to multiple species. They gave algebraic expressions for the likelihood; the present paper shows how to compute similar quantities numerically for larger numbers of species.

The first attempts to develop methods enabling use of restriction patterns for phylogenetic purposes were based, not on sites, but on presence and absence of restriction fragments. The methods estimated the product of divergence time and nucleotide substitution rate for two species. In this paper this quantity will be called branch length. Upholt (1977) computed the probability of retention of a restriction fragment as a function of the net probability of base change. Nei and Li (1979) corrected some of Upholt's formulas.

Most subsequent work has been based on sharing of restriction sites, rather than restriction fragments. Upholt (1977) and Nei and Li (1979) also presented estimation methods for this case. Gotoh et. al. (1979) and Kaplan and Langley (1979) derived methods based on maximum likelihood approaches. Kaplan and Risko (1981) showed that these methods were closely related and give similar results. Li (1981) investigated the statistical properties of Nei and Li's method by computer simulation.

These methods were restricted to two species at a time. Kaplan and Langley (1979) argued that their method could be extended to multiple species. However, Kaplan and Risko (1981) said that "computational difficulties render their method virtually unusable" for multiple species. They suggested using the divergence time estimates for all pairs of species and deriving branch lengths in a multiple species tree from these by a process of averaging, although they acknowledged that this did not result in an overall maximum likelihood estimate of the phylogeny.

Most multiple-species data for restriction sites has been treated by parsimony methods. The locations of the sites are treated as characters with two states: present and absent. The Wagner parsimony criterion, which allows change back and forth between the two states and favors that tree which minimizes the total number of state changes, has frequently been used. DeBry and Slade (1985) questioned this practice, pointing out the great asymmetry between presence and absence of restriction sites (for which see also Templeton, 1983*b*). They presented approximate likelihood calculations for four-species phylogenies, where the $4^6 = 4096$ possible states of a location for a six-base cutter have been collapsed to three states, 0, 0', and +. These indicate the presence of a site, the presence of sequences one substitution away from a site (0') and sequences more than one base away. Using their probability calculations they investigated numerically whether Wagner parsimony would be consistent, that is, whether it would converge to the correct tree as the number of locations was increased. They found that for four species, assuming a molecular clock, Wagner parsimony was not always consistent.

They suggest instead that the Dollo parsimony criterion be used. This originated in suggestions by Le Quesne (1974) and was defined formally by Farris (1977). It allows both gain and loss of a site, but constrains to one the number of times a particular site can be gained. It favors that tree which has the fewest changes given that constraint. I had suggested (1983) that this was a better parsimony method to use for restriction sites data, and DeBry and Slade found it to be consistent for all clocklike four-species cases.

However, allowing each site to arise only once must be an overreaction. Albert et. al. (1991) have argued that the overreaction is far too extreme, and that Wagner parsimony is more appropriate in realistic cases. If a site arises on one branch of the tree, a parallel gain of the site at the same location in a nearby lineage is not infinitely improbable, as the location is likely to have been only one base different from a restriction site in the immediate ancestor of both lineages. Parallel gains can be countenanced, but only if near each other on the tree. No parsimony method takes this into account.

The methods that could in principle do so are maximum likelihood methods. DeBry and Slade (1985) considered them superior in principle, but computationally impractical. Maximum likelihood was first applied to molecular sequence data by Neyman (1971). I have presented (Felsenstein, 1981) some algorithms intended to make the computation of maximum likelihood phylogenies from nucleotide sequences more practical. In this case there are four possible states, the four bases, at each site.

The difficulty with extending a similar approach to restriction sites data has been that the number of possible states is enormously larger. In the algorithms for the nucleotide sequence case, the computational effort is proportional to the square of the number of states, $4^2 = 16$. In the restriction sites case with an enzyme whose recognition sequence is 6 bases long and is not ambiguous, the presence of a site specifies one 6-base sequence out of all $4^6 = 4096$ possible 6-base sequences. A straightforward application of the computational methods for nucleic acid sequences would therefore require effort proportional to $4^{12} = 16,777,216$, or a million times as much work. This is not yet practical.

However, Nei and Tajima (1985) have made a simplification that has major effects, and it is this that turns out to be the key to the present method. They used a symmetric model of base change, the one introduced by Jukes and Cantor (1969). They noted that, under this model, the probability of change between any two r-base sequences depends only on the difference between them in the number of nucleotides matching the enzyme recognition site. This leads to an enormous collapsing of the states that have to be bookkept. In the case of a 6-base cutter, instead of dealing with 4^6 different states, we need deal with only

7 different states, as the number of bases by which a 6-base sequence can differ from the recognition site can range only from 0 to 6.

Nei and Tajima (1985) used this principle to compute the probabilities of all possible patterns of presence and absence of a restriction site for three species. Li (1986) extended their results to the case where transitions and transversions are distinguished. In this case there are 27 relevant ways that a location can differ from the recognition site, if we count the number of transition differences separately from the number of transversion differences. I will not make use of this generalization in the present paper. Li (1986) and Smouse and Li (1987) also extended Nei and Tajima's formulas to the case of four species.

Smouse and Li (1987) used these calculations to show how to calculate likelihoods for four-species restriction site phylogenies. They actually developed a Bayesian framework, using equal prior probabilities of different tree topologies, but without a complete prior on branch lengths. In this paper, I will follow their path only as far as the likelihood method. For tractability, the results of this paper will make use of only the Jukes-Cantor model. The more general model in which transitions and transversions differ in rate could be incorporated into the present framework, but only if we could tolerate computation an order of magnitude slower.

The Model

Using the Jukes-Cantor model the probability that a base changes in time t is

$$p(\mu t) = \frac{3}{4}(1 - e^{-\frac{4}{3}\mu t}) \tag{1}$$

where μ is the rate of substitution per unit time. The quantity μt is the expected number of substitutions. We will refer to this quantity for some branch i of the tree as v_i, the branch length. We will compute p as a function of branch length, as we intend to allow departure from a molecular clock, and thus need to allow the values of μ to differ from branch to branch.

Collapsing States. – Nei and Tajima's (1985) central insight was that one could collapse states. If the recognition sequence for a given restriction enzyme is r bases long, then in a branch of length v the probability that one r-base sequence changes to another is a function only of how many bases differ in the two sequences. We can take all 4^r possible sequences at that location and group them into $r+1$ classes according to how many bases differ between them and the recognition sequence. For example the $4^6 = 4096$ different six-base sequences group themselves into one which is the recognition sequence, 18 which are one base away from it, 135 which are two bases away, and 540, 1215, 1458 and 729 which are respectively 3, 4, 5, and 6 bases away.

Under the Jukes-Cantor model all 4^r sequences of length r are equally likely, and thus the probabilities of falling into the $r+1$ classes is proportional to the numbers of different sequences in each class. As Nei and Tajima (1985) noted, the probability of being i bases away from an r-base recognition site is

$$\pi_i = \binom{r}{i} \frac{3^i}{4^r}. \tag{2}$$

Under the Jukes-Cantor model with equal rates of change at all r nucleotide sites, Nei and Tajima (1985) computed the probability that a sequence currently j bases away from the recognition site will be k bases away from it at the end of a branch of length v. This is the sum, over all relevant values of m, of the probability that $k-j+m$ of the $r-j$ bases which matched the site will change to no longer match it, and m of the j bases which did not match it will come to match it. We call that quantity $P_{jk,m}(t)$. It is the product of two binomial probabilities:

$$P_{jk,m}(t) = \binom{r-j}{k-j+m} p^{k-j+m}(1-p)^{r-k-m} \binom{j}{m} (\frac{p}{3})^m (1-\frac{p}{3})^{j-m} \tag{3}$$

None of the powers $k-j+m$, $r-k-m$, m, and $j-m$ can be negative, so that the index m can take values from the maximum of 0 and $j-k$ to the minimum of j and $r-k$. The probability that a base changes is p, and the probability that it changes to a specific alternative is $p/3$. The resulting expression for the transition probability between states j and k is:

$$P_{jk}(t) = \sum_{m=max(0,j-k)}^{min(j,r-k)} \binom{r-j}{k-j+m} p^{k-j+m}(1-p)^{r-k-m} \binom{j}{m} (\frac{p}{3})^m (1-\frac{p}{3})^{j-m} \tag{4}$$

This is a version of equation (4) of Nei and Tajima (1985) with a somewhat altered notation (their $1-p$ is my p, their q is my $p/3$).

Expressions (2) and (4) give us the equilibrium probabilities and the transition probabilities of the process of change among the $r+1$ states of the process. Nei and Tajima recognized that the states of the full 4^r-state process could be collapsed into these $r+1$ states, a great saving. Since we observe only whether or not a site exists, we are observing whether or not the nucleotide sequence at that location is or is not 0 steps away from a site.

Computing the Likelihood

We can now apply an algorithm I developed (Felsenstein, 1973) to compute the likelihood of a phylogeny. It is a cousin of the Elston-Stewart algorithm for calculating likelihoods for genetic pedigrees, under genetic models (Elston and Stewart, 1971). The central operation of their algorithm is called “peeling”; I have called the analogous operation on trees “pruning”. The algorithm works by computing downwards along the phylogeny from the tips, updating a quantity called the “conditional likelihood”.

The conditional likelihood for a given location, i, and a given node j on the tree, is the probability $L_j^{(i)}(s_j)$ that we would observe all of the data for location i found at or above node j, given that node j was in state s_j. In our case the states s_j are the counts of how many bases away from a restriction site the location is, $0, 1, ..., r$. For a node which is a tip of the tree, if location i has a restriction site, then the data are certain given that $s_j = 0$, and impossible otherwise, so that the vector of conditional likelihoods is $(1, 0, 0, ..., 0)$. If a restriction site is not observed at that tip, then this is certain given any of the states other than 0, so that the vector of conditional likelihoods is $(0, 1, 1, ..., 1)$. If we have missing data, so that we are not able to observe whether or not a site exists, the vector is $(1, 1, ...1)$.

The algorithm allows us to compute the conditional likelihood at a location for an interior node of the tree, given the conditional likelihoods of the two immediate descendants of that node, and given the branch lengths. A brief consideration should establish that the formula for a node j whose immediate descendant nodes are k and l, with branch lengths v_k and v_l, is (Felsenstein, 1973)

$$L_j^{(i)}(s_j) = \left[\sum_{s_k=0}^{r} P_{s_j,s_k}(v_k) L_k^{(i)}(s_k)\right] \left[\sum_{s_l=0}^{r} P_{s_j,s_l}(v_l) L_l^{(i)}(s_l)\right] \tag{5}$$

since the probability of everything above node j (at site i) is the product of probabilities, each the weighted sum of conditional likelihoods for one descendant node, weighted by the probabilities of changing to the different states in that descendant. The transition probabilities P are given by equation (4).

One can start at the tips of the tree, where one can write down the conditional likelihoods by inspection, then use equation (5) to compute the conditional likelihoods for nodes successively further down the tree, until one computes them for the root (for unrooted trees one simply chooses an arbitrary placement of the root). When one reaches the root (node 0) the overall likelihood at the location is simply the weighted sum of the conditional likelihoods at the root, each weighted by the prior probability of that state:

$$L^{(i)} = \sum_{j=0}^{r} \pi_j L_0^{(i)}(s_j) \tag{6}$$

This permits us to rapidly calculate the likelihood of any given site on a given tree. The overall likelihood of the tree is the product of these over sites

$$L = \prod_{i=1}^{S} L^{(i)} \tag{7}$$

since each location is assumed to evolve independently (all of the locations are assumed not to overlap). By comparison, explicit expressions were given for the likelihoods $L^{(i)}$ in the three- and four-species cases by Smouse and Li (1987). The present quantities are not algebraically explicit, but they are readily computed numerically and apply to any number of species.

Exclusion of absent sites

One property of the data that is not taken into account in the above calculation is that locations at which recognition sequences are not present in any of the species may be omitted from the data entirely (Smouse and Li, 1987). This will frequently be the case. One could in principle infer the number of such locations from the physical distances between the observed sites, but this seems fraught with difficulty. An attractive alternative is to compute, for each location, the probability of observing the data, given that at least one species has a recognition sequence present. This is

$$L_+^{(i)} = \frac{L^{(i)}}{1 - L_-^{(i)}} \tag{8}$$

where $L_-^{(i)}$ is the probability of the site being absent in all species. If all the sites in our data have the same length r of their recognition sequence, $L_-^{(i)}$ will have the same value for all i. To correct for the omission of absent locations, we can compute this value, by adding to the data a fictional location 0 at which all species lack the recognition sequence. As we proceed down the tree computing likelihoods, we compute the conditional likelihoods at this fictional location as well. At the root of the tree we use equation (6) to compute $L^{(0)}$, the likelihood for the whole tree at this fictional location, which is also L_-. Then we use (7) and (8) to compute the overall likelihood as

$$\begin{aligned} L_+ &= \prod_{i=1}^{S} L_+^{(i)} = \prod_{i=1}^{S} \left(\tfrac{L^{(i)}}{1-L_-^{(i)}}\right) \\ &= \left(\prod_{i=1}^{S} L^{(i)}\right) \Big/ [1 - L^{(0)}]^S \end{aligned} \tag{9}$$

where S is the number of locations. If there are sites with several different recognition sequence lengths in the data, we can easily modify this procedure to compute a separate L_- for each length and apply the appropriate factor $1 - L_-$ in (9) the correct number of times. Smouse and Li (1987) argued that the correction for the absent locations would usually have an insignificant effect on the choice of tree topologies, although it could have a substantial effect on estimated branch lengths. This may or may not be true in general, but the present method allows us to make the correction with relatively little computational effort.

Iterating the branch lengths

So far, we have shown how the likelihood can be computed for a given tree and a given set of branch lengths. There remain two daunting problems, finding the maximum likelihood estimates of the branch lengths and searching among all possible tree topologies for the one with the highest likelihood.

We wish to find the value of the length of one branch which maximizes the likelihood. We do this without allowing the other branch lengths to change. We then proceed through the tree, maximizing the likelihood successively with respect to each branch length. The likelihood can never decrease during this process (except as a rounding-off error in the computations), and it is bounded above by the unknown true maximum likelihood. It can be proven that this sequence of likelihoods must converge. This is not quite the same thing as saying that the collection of branch lengths converges, but I have never known it to do otherwise. When each branch has its maximum likelihood length and none are changing, we are at a stationary point on the likelihood surface. In practice I have never known this to be anything but a maximum.

To estimate the length of a branch we make use of the EM-algorithm (Dempster, Laird, and Rubin, 1978). Our approach is related to the counting method developed by Nei and Tajima (1983) for estimating divergence of pairs of sequences, which was also an EM-algorithm. First we apply the pruning algorithm outlined above to prune the tree until it contains only a single branch. We are then in possession of the location-by-location conditional likelihoods $L^{(i)}$ and

$L^{(i)'}$ at each end of this branch of length v. The equation for the overall likelihood of the tree is then

$$L = \prod_{i=1}^{S} \sum_{j=0}^{r} \sum_{k=0}^{r} \pi_j L_j^{(i)} \sum_{m=max(0,j-k)}^{min(j,r-k)} P_{jk,m}(v) L_k^{(i)'} \tag{10}$$

the product over i being over all locations and the summation over m being as described for equations (3) and (4).

The EM-algorithm can be constructed heuristically by considering the j, k, m term in the summation for location i in equation (10) as proportional to the probability of the data at that location, given that there were in reality $k - j + m$ changes away from matching the restriction site and m changes towards matching it. If we could observe and count all the changes, then a maximum likelihood estimate of p, the probability of change per nucleotide in that branch of the tree, would simply be the observed fraction of nucleotides that changed. The tactic of the EM-algorithm is to use our best current estimate of p to compute the probabilities of each of the possible different values of m, given the data, and then use those to obtain an improved estimate of p. This process is continued until the estimate of p converges. Dempster, Laird, and Rubin proved that it converges to the maximum likelihood estimate.

The Appendix shows that if there were only one location, for which our current estimate of p is $\hat{p}$ and at which the state at one end of the branch is j and that at the other end k, and if in addition to the $j - k$ changes there are known to be m more changes away from the restriction site, balanced by m changes back, the EM algorithm for the estimate of p would be

$$\hat{p}' = \frac{(k - j + 2m) + (j - m)\left(\frac{2\hat{p}/3}{1-\hat{p}/3}\right)}{r}. \tag{11}$$

This is readily rationalized as follows: out of r bases, $k - j + m + m$ are seen to have changed, and of the $j - m$ that do not match the recognition sequence but appear not to have changed, the probability that they have actually changed to a different base that also does not match the recognition sequence is estimated to have been $(2\hat{p}/3)/(1-\hat{p}/3)$.

The overall EM algorithm is obtained by taking the weighted average of the right-hand-side of (11) over all values of j, k, and m, and the average of that over all locations. The weights for j, k, and m can be computed from the i-th term of (10). The probability that this location goes from state j to state k with a given value of m is simply the fraction of the likelihood in the i-th term contributed by that possibility, which is

$$\Big(\pi_j L_j^{(i)} P_{jk,m}(v) L_k^{(i)'}\Big) \Big/ \Big(\sum_{j=0}^{r}\sum_{k=0}^{r} \pi_j L_j^{(i)} \sum_m P_{jk,m}(v) L_k^{(i)'}\Big). \qquad (12)$$

Unobserved locations in the EM iteration

If we assume that restriction site locations that do not have a restriction site in any species are unobserved, we can alter the EM iteration to take this into account. This is done by using the fictional location 0 mentioned above, and estimating from the other locations how many of these fictional locations must have existed in the data, and basing the EM iteration both on the observed locations and on these unobserved ones. If $R^{(i)}/L^{(i)}$ is the average of the fraction on the right-hand side of equation (11) over all states j, k, and m for location i, weighted by the expression (12), then if there is no correction for unobserved locations, the EM iteration will be:

$$\hat{p}' = \frac{\sum\limits_{i=1}^{S} \left(\frac{R^{(i)}}{L^{(i)}}\right)}{S}. \qquad (13)$$

The probability that a location will go unobserved on the current tree is given by $L^{(0)}$. The number of unobserved locations per observed location will then be $L^{(0)}/(1-L^{(0)})$, and the total number of unobserved locations can then be estimated to be S times this.

Adding to the numerator and denominator of (13) the terms that would be contributed by this many unobserved locations leads to the EM iteration taking unobserved locations into account:

$$\hat{p}' = \frac{\sum_{i=1}^{S} \left(\frac{R^{(i)}}{L^{(i)}}\right) + S\left(\frac{L^{(0)}}{1-L^{(0)}}\right)\left(\frac{R^{(0)}}{L^{(0)}}\right)}{S\left(1 + \frac{L^{(0)}}{1-L^{(0)}}\right)} \tag{14}$$

which simplifies to

$$\hat{p}' = \left(\frac{\sum_{i=1}^{S} \left(\frac{R^{(i)}}{L^{(i)}}\right)}{S}\right)\left(1 - L^{(0)}\right) + R^{(0)}. \tag{15}$$

Equation (9) can be used to compute the likelihood.

Finding the maximum likelihood tree

The main problem with the above EM algorithm (whether with or without the correction for unobserved locations) is that it is slow. In order to speed up its convergence I have resorted in practice to extrapolating the changes in $\hat{p}$ to speed the convergence of the algorithm. If the extrapolation factor is (as is the default in my program) 1000, then the new value of $\hat{p}$ in each iteration is given by

$$\hat{p}'' = \hat{p} + 1000(\hat{p}' - \hat{p}). \tag{16}$$

This speeds the convergence but is dangerous: it removes the guarantee that the algorithm will move us upwards on the likelihood surface. In practice the user may have to adjust this initial extrapolation factor, which we expect to be data-dependent. In the program described below I have also made use of Aitken's δ^2 method of extrapolation. This continually adjusts the extrapolation factor to better values.

The branch length itself, μt, can be found from $\hat{p}$ by solving equation (1) for μt, obtaining the well-known result:

$$\hat{\mu t} = -\frac{3}{4}\ln(1 - \frac{4}{3}\hat{p}). \tag{17}$$

The above algorithms search for the best length of one branch. Unless the extrapolation factor has misled it, each iteration should increase the likelihood of the tree. But, as we have noted, this is not the same as a simultaneous EM iteration of all branch lengths on the tree. In practice, the successive improvement of each branch length seems to converge to a reasonable result. If we reach a point at which the likelihood is not improved by a change in any one branch length, this is a stationary point on the surface of trees of that topology (the coordinates being the branch lengths). In practice it seems always to be a maximum, although we have no proof that this is always so.

Multiple Restriction Enzymes

Smouse and Li (1987) discuss an additional complication: that our data may represent digests with multiple restriction enzymes, rather than with one as assumed here. There are, as they note, possible complications arising from the overlap of recognition sequences. They also note that as the absent locations for one restriction enzyme include locations that have recognition sequences for another, there is a lack of independence which is very complex.

The different enzymes are all used on the same piece of DNA. We can make an approximate treatment for the cases where there are or are not unobserved locations. When there are no unobserved locations and there are N different restriction enzymes employed, we assume that the recognition sequences of all of these enzymes are sufficiently different that no two are near each other. In such a case if there is a restriction site in one species at a given point on the DNA, we approximate by saying that this excludes having a restriction site for any other enzyme in any species at that location, on the grounds that in the amount of time available none of the other species is likely to have departed so much from the recognition sequence as to have a recognition sequence for another enzyme. Thus for each enzyme there are a series of possible patterns of "+" and "-", and all the patterns that have a "+" anywhere are mutually exclusive alternatives. With n species there are of course $2^n - 1$ different such patterns for each enzyme, for a total of $N(2^n - 1)$.

We saw in equation (6) the probability of the pattern observed at the i-th location, $L^{(i)}$. Let us call the same quantity, computed for the k-th pattern and the j-th enzyme Q_{jk}. Suppose that we let k index all possible patterns, 1 to $2^n - 1$, with 0 indexing the pattern that lacks restriction sites in all species. Note that the presence of pattern k for one enzyme excludes all other patterns except 0, and forces the patterns at that location to be zero for all other enzymes. If there are m_{ij} occurrences of pattern j at enzyme i, and if the total number of locations is S, the overall likelihood is:

$$L = \left(\prod_{i=1}^{N} \prod_{j=1}^{2^n-1} Q_{ij}^{m_{ij}} \right) \left(1 - \sum_{i=1}^{N} \sum_{j=1}^{2^n-1} Q_{ij} \right)^{S - \sum_i \sum_j m_{ij}} \tag{18}$$

In our models, if the length r of all restriction enzyme recognition sites is equal, then $Q_{ij} = Q_{kj} = Q_j$ for all i, j and k. If the sum of all the Q's is small, as is very likely to be the case for moderately closely related species, we can approximate

$$\begin{aligned} \left(1 - \textstyle\sum_{i=1}^{N} \sum_{j=1}^{2^n-1} Q_{ij} \right)^{S - \sum_i \sum_j m_{ij}} &= \left(1 - N \textstyle\sum_{j=1}^{2^n-1} Q_j \right)^{S - \sum_i \sum_j m_{ij}} \\ &\simeq \left(1 - \textstyle\sum_{j=1}^{2^n-1} Q_j \right)^{N\left(S - \sum_i \sum_j m_{ij}\right)} \\ &= L_{-}^{N\left(S - \sum_i \sum_j m_{ij}\right)} \end{aligned} \tag{19}$$

This implies that we can, to good approximation, treat this case by simply pretending that there are N times as many locations that are missing all sites as were observed. This approximation also carries over into the EM-algorithm iteration of branch lengths.

If there is exclusion of absent locations, then the probability of a location that is seen having pattern j for enzyme i is

$$Q_{ij} \Bigg/ \left(\sum_{i=1}^{N} \sum_{j=1}^{2^n-1} Q_{ij} \right) = \frac{Q_{ij}}{N(1 - L_{-})} \tag{20}$$

and this implies that to get correct likelihoods and correct EM-algorithm iterations of branch lengths it is only necessary to replace, in equation (9), $L^{(0)}$ by $1 - N(1 - L^{(0)})$, and that no changes are needed in (14) or (15), which remain as they would be for a single restriction enzyme.

Searching among tree topologies

The above EM algorithm searches for the maximum likelihood branch lengths for a given tree topology. Let us assume that it is successful in this search. There still remains the question of how to search among all tree topologies. In principle one could do just that, for each possible tree topology iterating the branch lengths to find their maximum likelihood values, then evaluating the likelihood of the resulting tree. The difficulty is of course that there can be far too many tree topologies for this. For example (Cavalli-Sforza and Edwards, 1967), with ten species there are 2,027,025 unrooted bifurcating tree topologies. While one could imagine the development of branch-and-bound methods for economizing in this search, these are not available, and in practice it seems necessary to search among tree topologies in a more ad hoc fashion.

The strategy which I have followed is to build up the tree by adding one species at a time. First one constructs a three-species tree using the first three species. There is only one possible unrooted tree topology. The branch lengths are estimated by the successive EM algorithm methods mentioned above. Then one adds to this tree the fourth species. If the resulting tree is to be bifurcating, this species must be added by inserting a new interior node in some branch and having the new species arise from this new node. There are 3 branches on the three-species tree, hence three four-species tree topologies to try. For each the branch lengths are iterated and the likelihood of the resulting tree evaluated. The fourth species is added wherever the resulting tree topology, once its branch lengths have been iterated, has the highest likelihood.

This process is continued. As each species is added it creates two new branches in the tree, so that there are more and more places to add the next species. By itself this strategy of addition of species will require for n species the examination of

$$1 + 3 + 5 + 7 + ... + (2n - 5) = (n - 2)^2 \tag{21}$$

different tree topologies. The strategy is identical to that of Eck and Dayhoff (1966). I have added to this successive-addition strategy a rearrangement stage. After the k-th species has been added a series of local rearrangements of the tree are examined. At each of the $k - 3$ interior branches of the tree there are two possible ways that the tree can be rearranged by exchange of branches connected to neighboring nodes of the tree. Each of these is examined, with its branch lengths being improved by EM iteration and its likelihood evaluated after that is done. If any rearrangement succeeds in improving the tree, it is accepted as the basis for further rearrangements, and so on until no local rearrangement improves the likelihood.

It will not necessarily be true that all of the tree topologies examined in this way are distinct. The number of rearranged trees that are tried if no rearranged tree is accepted will be:

$$0 + 2 + 4 + ... + (2n - 6) = (n - 3)(n - 2) \tag{22}$$

so that this addition and local rearrangement strategy requires examination of at least $(2n-5)(n-2)$ tree topologies. The computer program also has an option that tries a wider range of rearrangements, removing each possible subtree from the tree and trying to reinsert it in all possible places. It also allows the user to specify a given tree topology which will have its branch lengths estimated but which will not be rearranged.

Interval estimates and tests of different trees

The phylogenies estimated by this maximum likelihood method are point estimates. A variety of techniques have been developed for making interval estimates (such as confidence intervals) and for testing alternative phylogenies. These apply to the present case as well:

(1) Asymptotic variances of the estimate. One could use the curvatures of the likelihood surface at the estimated phylogeny to obtain an estimate of the asymptotic variances of the parameters being estimated. The parameters are the branch lengths (technically the topology is not a parameter). The problem with this approach is that it gives the variance of the estimates only in the aymptotic case where the amount of data are at least large enough that there is no ambiguity about the tree topology. We are unlikely to have this much data. It is also tedious to calculate the necessary derivates.

(2) Likelihood ratios between trees. The preceding approach uses the asymptotic normality of the maximum likelihood estimates of the branch lengths. Even if the number of restriction sites is not large enough to ensure normality, the likelihood ratio between trees may be approximately distributed in the distribution specified by the Likelihood Ratio Test. If the true tree has $2n-3$ branches, then we can test any given tree θ against the maximum likelihood tree $\hat{\theta}$ by using the fact that, asymptotically, the log of the likelihood ratio follows

$$2\big[\ln L(\hat{\theta}) - \ln L(\theta)\big] \sim \chi^2_{2n-3} \tag{23}$$

An approximate confidence interval would be to accept all trees that generate a value of χ^2 that is below (say) the 95th percentile. This has the difficulty that one does not automatically know which trees are in the confidence interval. For any given tree topology one could find the branch lengths that maximize the likelihood; if the resulting tree were in the confidence interval as judged by its χ^2 value, one would at least know that not all trees of that topology could be excluded from consideration. The real difficulty with this procedure is that its justification is asymptotic. It assumes large enough amounts of data that there would actually be no ambiguity as to the tree topology. Thus there is some uncertainty whether the same χ^2 value should be used when more than one topology is possible.

(3) Bootstrap confidence intervals. A major disadvantage of both of the preceding approaches is that they assume that we know that the rates of evolution

at all locations are equal. This is very unlikely to be true. I have elsewhere (1985) suggested the use of the bootstrap method of resampling to construct confidence intervals for phylogenies. This assumes that the different locations are independent, but they can have different rates of evolution, as long as we are willing to assume that the prior probability of a given rate of evolution at all locations is the same, and that rates are assigned independently to different locations, which are much less restrictive assumptions than equality of rates of evolution. Bootstrap resampling involves drawing a sample of S locations with replacement. An alternative is delete-half jackknife resampling, which draws a set of $S/2$ locations at random without replacement. Phylogenies are estimated for each resampled data set, and the confidence interval is constructed from the innermost 95% of the bootstrap estimates. There are various alternative ways of deciding which ones are innermost. If our interest is in the presence or absence of a particular branch on the phylogeny, we can take it as supported with confidence if 95% of all bootstrap estimates contain the branch. Sanderson (1989) has suggested a different approach involving distance measures between trees. Hasegawa and Kishino (1989) have used bootstrapping with maximum likelihood phylogenies from nucleotide sequences, taking a Bayesian approach. The difficulty with the bootstrap confidence interval approach is that it involves making a large number of maximum likelihood estimates of phylogenies, and is thus computationally difficult.

(4) Paired-sites tests. Kishino and Hasegawa (1989) have developed a paired-sites test which is in effect a log-likelihood version of the paired sites test of Templeton (1983*a*). The test is used to compare two trees. The differences of log-likelihood between the trees are computed location by location, and a test performed that their mean (and hence their sum) is different from zero. There are a variety of possible such tests: t-tests, nonparametric tests, and even, tests based on bootstrapping. The paired sites test is much faster than constructing a bootstrap confidence interval, and the bootstrapping used in this test is far

faster computationally than the other use of bootstrapping. The difficulty is that we must know which two trees to compare. Kishino and Hasegawa present an argument that all trees that cannot be rejected by this test, compared to the maximum likelihood tree, consititute in effect a confidence interval (their argument is actually Bayesian and this statement is a modification of it). If we had a way of enumerating the trees in the confidence interval, their test would provide an effective method of constructing such an interval. Although their test was developed for nucleotide sites, it applies equally well to restriction site locations.

We thus have four approaches, each with some disadvantages. The first two assume equality of rates of evolution at all locations. The second two do not but are computationally difficult.

Availability of computer program

A computer program, RESTML, has been written to estimate maximum likelihood phylogenies from restriction sites data. It is written in a generic form of Pascal and is distributed as part of the PHYLIP package of programs for inferring phylogenies, versions 3.1 and later. It has been in distribution since April, 1988, and is available in source code. The program allows the user to search over tree topologies, with either local or global rearrangement, or to restrict the search to a given tree topology. If more than one of these user-defined tree topologies is provided, the program also performs Kishino and Hasegawa's (1989) paired-sites test. The program allows the lengths of restriction sites to be as large as 8, although it does assume that all sites have the same length. Because of the large amount of numerical computation involved, the program is slow. The data example presented below will serve to illustrate its speed: making a maximum likelihood estimate for 5 species with 134 site locations, searching among different tree topologies, took 6.6 seconds execution time on a DECstation 3100, a machine which is approximately 10 times faster than a DEC MicroVAX II and 200 times faster than an 8 MHz IBM PC/XT.

The PHYLIP package (currently in version 3.3) is distributed free by the author, written on diskettes or magnetic tapes sent by the recipient. It is also available by electronic mail or by anonymous ftp from evolution.genetics.washington.edu or anthro.utah.edu. Write to the author for information on distribution media and policies.

Some possible extensions

The present approach has a number of obvious limitations. The most important of these is probably the reliance on the Jukes-Cantor model of nucleotide substititution. This does not allow for inequalities of transition and transversion rates. It would be possible in principle to construct a version of the present algorithm using Kimura's (1980) two-parameter model. This would pose serious computational problems. In the current model, for a 6-base cutter the 4^6 different states at the nucleotide level reduce to $6 + 1 = 7$ states when the symmetries of the Jukes-Cantor model are taken into account. In the reduced state space two 6-base sequences are equivalent if they differ from the restriction enzyme recognition sequence by the same number of bases.

With the Kimura 2-parameter model, two 6-base sequences are equivalent if they differ from the restriction enzyme recognition sequence by the same number of transitions and by the same number of transversions. This means that there are many more states. If we index by i the number of transition differences and by j the number of transversion differences there are 7 possible values of i and 7 of j but not all combinations are possible, since $i + j \leq 6$. There are thus $1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$ possible states.

The algorithm used here could be extended to treat the Kimura 2-parameter model. It requires an amount of computation proportional to the square of the number of states in this Markov chain. With 4 times as many states it would take about 16 times as long to find a maximum likelihood estimate in the Kimura case as in the Jukes-Cantor case.

Of course, Kimura's model still has limitations – it does not allow for base frequencies that depart from equality. Unfortunately taking that into account removes all symmetries, and expands the number of states from 28 to 4096, which would imply a further slowing of the algorithm by an additional factor of 21,400. It will be some time before this is practical.

One could imagine going further and allowing for regions of different base composition and different rates of evolution in different parts of the genome. This may in some cases be practical.

Analysis of the Ferris et. al. mitochondrial data

As an example of the use of this algorithm let us analyze the mitchondrial restriction sites data of Ferris et. al. (1981). These data are given in a convenient tabular form by Templeton (1983*a*). The full data have 134 restriction sites for 19 enzymes in 5 species, but here I have followe the practice of Smouse and Li (1987) and discarded 12 locations which have one or another of the four restriction sites whose recognition sequences are subsets of the recognition sequence of another enzyme. This leaves us with 122 locations for 15 enzymes. Table 1 shows the sequences in the form required for input to RESTML. Presence of the restriction sites is given by "+" and absence by "-". In the analysis all enzymes have been assumed to be 6-base cutters. This is not true – some are 4-base cutters. At the moment the program assumes that all sites have the same length. This is done purely for computational convenience. It should not make much difference to the results as most of the enzymes are actually 6-base cutters (one of the 19 is a 4-base cutter, and two have degeneracy in their recognition sequence). Another potential problem is that the locations may not be completely independent, as assumed by the algorithm. It is probable that some overlap, although most pairs of locations will not. Smouse and Li (1987) took another kind of dependence into account by eliminating 19 of the locations from their 4-species data by dropping two of the enzymes whose recognition sites could also serve as recognition sites for another enzyme.

Figure 1 shows the phylogeny that results. It is produced in unrooted form by the program, but has been rooted on the line to the gibbon by averaging the midpoints of the human-gibbon, chimp-gibbon, and gorilla-gibbon paths. A simple midpoint rooting using the longest path on the tree would place the root on the branch connecting the clade of three African apes (human, chimp and gorilla) to the orang and gibbon. There has been little question on other morphological and molecular grounds that the root belongs on the lineage connecting the gibbon to the rest. The phylogeny estimates an amount of divergence substantially larger than estimated for nuclear DNA. Sibley and Ahlquist estimated (1984) 1.4% difference between human and chimpanzee, but the total branch length separating them in the second tree is 0.09118, which is more than sixfold higher.

Table 2 shows the likelihoods of the three trees that resolve the human-chimpanzee-gorilla clade in the three possible ways, and the results of applying the Kishino-Hasegawa test. The chimp-gorilla tree is favored, but not statistically significantly.

A simulation test

To check the possibility that some error in the algorithm or in the program implementing it biases the result, a computer simulation study was conducted. A ten-species phylogeny was simulated using a continuous-time branching process. It is shown in Figure 2a in unrooted form. The distance between tips C and B in this phylogeny is 0.250351, in units of expected fraction of nucleotide change. 10 Kb of DNA was simulated evolving along this tree according to a Jukes-Cantor model, with equal rates of change at all nucleotide sites. A program was written to recognize sites in these sequences, using five imaginary 4-base cutters, whose recognition sequences were ACGT, CAAC, GAGA, AGCT, and TTAA. The result was a data set with 9,997 sites (as the last three locations in the 10 Kb could not be the start of a restriction site), most of which were "-" for all species. A program was written to remove all the locations absent in all species, and this produced a data set with 542 locations.

Figures 2b and 2c show the results of running RESTML on these two data sets, with an initial extrapolation factor of 100. They are remarkably similar to each other and to the true tree, which suggests that the algorithms and program are correct. The similarity of the result whether or not the sites absent in all species are excluded is some evidence for Smouse and Li's (1987) inference that a correction for this exclusion would have little influence on the tree. Although the trees inferred by RESTML are not perfectly clocklike, they are quite close to being so, and midpoint rooting would in both cases place the root approximately correctly.

Acknowledgments

This work was supported by National Science Foundation grants BSR-8614807 and BSR-89-18333, and by National Institutes of Health grant 5 R01 GM41716. I am grateful to the reviewers of this paper for helpful suggestions, and to Peter Smouse for discussions of some of the issues.

Literature Cited

Albert, V. A., B. D. Mishler, and M. W. Chase. 1991. Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. *In* D. Soltis, P. Soltis, and J. Doyle (eds.), Plant Molecular Systematics. Chapman and Hall, London, *in press*.

Cavalli-Sforza, L. L. and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 32: 550-570 (also published in Amer. J. Hum. Genet. 19: 233-257).

DeBry, R. W. and N. A. Slade. 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. 34: 21-34.

Dempster A. P., Laird M. N., Rubin D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc B 39: 1-38.

Eck, R. V., Dayhoff, M. O. 1966. Atlas of Protein Sequence and Structure 1966. Silver Spring, Maryland: Natl. Biomed. Res. Found.

Elston, R. C. and J. Stewart, 1971. A general model for the genetic analysis of pedigree data. Human Heredity 21: 523-542.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's law. Syst. Zool. 26: 77-88.

Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for evolutionary trees from data on discrete characters. Syst. Zool. 22: 240-249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368-376.

Felsenstein, J. 1983. Inferring evolutionary trees from DNA sequences, pp. 133-150. *In* B. S. Weir (ed.), Statistical Analysis of DNA Sequence Data, M. Dekker, N. Y.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap Evolution 39: 783-791.

Ferris, S. D., A. C. Wilson, and W. M. Brown. 1981. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Natl. Acad. Sci. USA 78: 2432-2436.

Gotoh, O., J.-I. Hayashi, H. Yonekawa, and Y. Tagashira. 1979. An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. J. Mol. Evol. 14: 301-310.

Hasegawa, M. and H. Kishino. 1989. Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochindrial-DNA sequences. Evolution 43: 672-677.

Jukes, T. H. and C. Cantor. 1969. Evolution of protein molecules, pp. 21-132 *In* H. S. Munro (ed.), Mammalian Protein Metabolism. Academic Press, N. Y.

Kaplan, N., and C. H. Langley. 1979. A new estimate of sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. 13: 295-304.

Kaplan, N. and K. Risko. 1981. An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. 17: 156-162.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29: 170-179.

Kluge, A. R. and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18: 1-32.

LeQuesne, W. J. 1974. The uniquely evolved character concept and its cladistic application. Syst. Zool. 23: 513-517.

Li, W.-H. 1986. Evolutionary change of restriction cleavage sites and phylogenetic inference. Genetics 113: 187-213.

Nei. M. and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269-5273.

Nei, M., and F. Tajima. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics 105: 207-217.

Nei, M. and F. Tajima. 1985. Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol. Biol. Evol. 2: 189-205.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems, pp. 1-27 *In* S. S. Gupta and J. Yackel (eds.), Statistical Decision Theory and Related Topics. Academic Press, N. Y.

Sibley, C. G. and J. H. Ahlquist. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. 20: 2-15.

Smouse, P. E. and W.-H. Li. 1987. Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. Evolution 41: 1162-1176.

Sanderson, M. J. 1989. Confidence limits on phylogenies: the bootstrap revisited. Cladistics 5: 113-129.

Templeton A. R. 1983*a*. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37: 221-244.

Templeton, A. R. 1983*b*. Convergent evolution and nonparametric inferences from restriction data and DNA sequences. pp. 151-179. *In* B. S. Weir (ed.), Statistical Analysis of DNA Sequence Data, M. Dekker, N. Y.

Upholt, W. B. 1977. Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Res. 4: 1257-1265.

Appendix

Proof of the EM-algorithm for a single site in a 2-species case.

Suppose that we have one site and two species, and the species are respectively in states j and k. Suppose that it is somehow known that state j changed to state k with $k-j+m$ of the $r-j$ bases which matched the site changing to no longer match it, and k of the j sites that did not match it changing to match it. From equation (3) we can compute the likelihood for this site:

$$L(t)=\binom{r-j}{k-j+m}p^{k-j+m}(1-p)^{r-k-m}\binom{j}{m}(\frac{p}{3})^m(1-\frac{p}{3})^{j-m} \tag{24}$$

Taking the log-likelihood and absorbing the constants into one constant K,

$$\ln L(t)=K+(k-j+2m)\ln p+(r-k-m)\ln(1-p)+(j-m)\ln(3-p). \tag{25}$$

Taking the derivative with respect to p and equating this to zero,

$$\frac{(k-j+2m)}{p}-\frac{(r-k-m)}{(1-p)}-\frac{(j-m)}{(3-p)}=0. \tag{26}$$

When multiplied by $p(1-p)(3-p)$ this becomes the quadratic equation

$$rp^2+(-k+3j-4m)p+(3k-3j+6m)=0 \tag{27}$$

If in the iteration formula, equation (11), we set $p'=p$ and solve, multiplying both sides by $3-p$, we get

$$rp(3-p)=(k-j+2m)(3-p)+(j-m)(2p) \tag{28}$$

and on collecting terms this is found to be identical to (27). The values of p to which the iteration equations converge must satisfy $p'=p$ and hence (28). As they then satisfy (27) they are also stationary points of the likelihood surface, as required.

Table 1. Restriction sites data for five hominoid species, the data of Ferris et. al. (1981) as recoded by Templeton (1983*a*) and with some sites deleted as done by Smouse and Li (1987). “+” is the presence of a site, “-” its absence.

Name	Sites					
	10	20	30	40	50	60
Gibbon	+++-----++	---+----+-	+--++++---	---+++--+-	-+--+++++-	-----+-+--
Orang	++-+-++-+-	-+++------	-++---+-+-	+--+-----+	-+-+-+--+-	--+---+-+-
Gorilla	++---+-++-	+--+-+-+++	-----++---	-+++--++--	--------+-	+-----++-+
Chimp	+----++-+-	--+++++-+-	-----+---+	--++--+-+-	+-------++	---++++---
Human	++--++--+-	--++---+--	-+-+-+++--	--++------	--+-----++	-+---++---
	70	80	90	100	110	120
Gibbon	-++--+----	-----+----	-++-+-+--+	+++---++-+	+-+-++++++	+-++-+-++-
Orang	------++--	--------+-	---+++--++	+---+---++	--+++-+--+	---+++--++
Gorilla	--++---+-+	+---+-+-++	----+----+	-+++----++	----+-++-+	-+-++-+-+-
Chimp	+-++----+-	-++-----++	+---+---++	-++-----++	-+----+--+	-+-++-+++-
Human	+-+-+--+--	---++--++-	-+--+--+++	-+---+--++	+---+-++-+	-+-+++--+-
Gibbon	-+					
Orang	+-					
Gorilla	--					
Chimp	--					
Human	--					

Table 2. Log-likelihoods for three tree topologies, their differences and site-by-site standard deviations of the difference.

Tree topology	Ln Likelihood	Difference	Standard deviation
((((C,G),H),O),Gi)	-679.43909		
((((H,C),G),O),Gi)	-683.86755	-4.42847	3.4194
((((H,G),C),O),Gi)	-683.83002	-4.39093	3.5294

Figure Captions

Figure 1. Phylogeny estimated from the data in Table 1 by RESTML. The tree is estimated as an unrooted tree but is shown here as rooted on the Gibbon lineage. Scale in nucleotide substitutions per site can be inferred from the vertical length of the line connecting Human to its immediate ancestor, which has an estimated length of 0.02853.

Figure 2. (a) True phylogeny used to create simulated restriction site data, with the phylogenies estimated by RESTML with (b) and without (c) inclusion of sites that are absent in all species. All three phylogenies have their branch lengths shown on the same scale. The total branch length between B and C in the true tree is 0.250351. The root is at the midpoint of this path.