J Mol Evol (1981) 17:368–376

Journal of
Molecular Evolution

© Springer-Verlag 1981

Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach

Joseph Felsenstein

Department of Genetics, University of Washington, Seattle, Washington 98195, USA

Summary. The application of maximum likelihood techniques to the estimation of evolutionary trees from nucleic acid sequence data is discussed. A computationally feasible method for finding such maximum likelihood estimates is developed, and a computer program is available. This method has advantages over the traditional parsimony algorithms, which can give misleading results if rates of evolution differ in different lineages. It also allows the testing of hypotheses about the constancy of evolutionary rates by likelihood ratio tests, and gives rough indication of the error of the estimate of the tree.

Key words: Evolution – Phylogeny – Maximum likelihood – Parsimony – Estimation – DNA sequences

Introduction

As DNA sequences accumulate, there will be an increasing demand for statistical methods to estimate evolutionary trees from them, and to test hypotheses about the evolutionary process. Most evolutionary trees constructed from DNA or protein sequence data have been produced by parsimony methods (Edwards 1963; Edwards and Cavalli-Sforza 1964; Camin and Sokal 1965). These methods implicitly assume that change is improbable a priori (Felsenstein 1973, 1979). If the amount of change is small over the evolutionary times being considered, parsimony methods will be well-justified statistical methods.

Most data involve moderate to large amounts of change, and it is in such cases that parsimony methods can fail. When amounts of evolutionary change in different lineages are sufficiently unequal, it can be shown (Felsenstein 1978b) that parsimony methods make an inconsistent estimate of the evolutionary tree, converging to the wrong tree with increasing certainty as more sequences are considered for the same set of species. The compatibility approach to estimating evolutionary trees (Le Quesne 1969; Sneath et al. 1975; Estabrook and Landrum 1975) will suffer from the same difficulty (Felsenstein 1978b).

A third approach estimates the tree from information on the pairwise similarity of the sequences (Fitch and Margoliash 1967), without attempting to make full use of the information available in the original sequences. Of these methods, the least-squares approach of Chakraborty (1977) is of particular interest in having an explicit statistical justification. Colless (1970) has shown that simple clustering methods based on pairwise similarities can give inconsistent estimates of an evolutionary tree if rates of evolution are sufficiently unequal in different lineages.

A fourth approach involves methods which try to make explicit and efficient use of all of the sequence data by formulating a probabilistic model of evolution and applying known statistical methods. Neyman (1971) and Holmquist (1972) have stated probabilistic models of DNA evolution, and Neyman explicitly discussed statistical estimation methods using sequence data on

By acceptance of this article, the publisher and/or recipient acknowledges the U.S. government's right to retain a nonexclusive, royalty-free licence in and to any copyright covering this paper

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately-owned rights

0022–2844/81/0017/0368/$01.80

three species. Kashyap and Subas (1974) later extended Neyman's results to estimate evolutionary trees with many species, by looking at overlapping subsets of three species at a time. The problems which have prevented examination of multispecies sequence data are the difficulty of the computations and the paucity of DNA sequence data available for analysis. Felsenstein (1973) gave an algorithm for evaluating the likelihood of an evolutionary tree with protein data, but this could not be developed into a practical maximum likelihood method for protein sequences because of the computational burden. Ferris et al. (1979) have used a likelihood method for inferences regarding rates of loss of tetraploid expression at enzyme loci. Kaplan and Langley (1979) have used a probabilistic model of DNA evolution to obtain maximum likelihood estimates of divergence times based on restriction enzyme fragment maps. They have not extended their method beyond two species, though they state that it can be so extended.

This paper addresses the problem of inferring evolutionary trees (phylogenies) from DNA sequences under a simple probabilistic model of DNA evolution. An algorithm for computing the likelihood of a given tree is developed from the more general algorithm stated previously (Felsenstein 1973). It is quite feasible computationally, in contrast to the situation with protein sequence data. Even with this algorithm in hand, we would still be faced with the daunting prospect of searching for the maximum likelihood tree by making small alterations in the tree while repeatedly evaluating its likelihood. For the particular case in which rates of base substitution are allowed to differ among lineages, an iterative method of altering the tree is developed which guarantees a continued increase in the likelihood. This forms the basis of a computer program which makes maximum likelihood estimates of an evolutionary tree from DNA (or RNA) sequences.

Computing the Likelihood of a Tree

A simple general model of the evolution of DNA sequence data would involve a probabilistic model of the process of branching which leads to the evolutionary tree, as well as a model of the process of change in DNA sequence along this tree. There are a number of reasons for not attempting a model of the first process. Such a model would have to involve both speciation and extinction processes, as well as the process by which the species under study have been selected from among those potentially available. This last process seems impossible to model adequately. For that reason I have taken the evolutionary tree to be the unknown entity being estimated, and have not attempted to use a probabilistic model of branching to place prior probabilities on the form of the tree.

Maximum likelihood estimation is the method of statistical inference most readily applicable to data of this sort. It involves finding that evolutionary tree which yields the highest probability of evolving the observed data. Note that although the likelihood of a tree is the probability of the data given the hypothesis, it is taken as a function of the hypothesis (the tree) rather than a function of the data. This means that the likelihoods for different trees do not sum to unity. Note also that the likelihood of a tree is not the probability that the tree is the correct one.

Given that we are attempting maximum likelihood estimation, our problem reduces to computing the probability of a particular set of sequences on a given tree and maximizing this probability over all evolutionary trees. The probability of obtaining a given set of sequences at the tips of a given tree can be computed if we have a model specifying the probability that sequence S_1 changes to sequence S_2 during evolution along a segment of the tree of length (in time or other units) t. Computation is enormously facilitated if we can assume that changes at different sites in the sequence are probabilistic events which are independent. This is a restrictive assumption, but practical computation does not appear feasible without it. In particular, deletion and insertion events, which usually involve adjacent sites, cannot be adequately modeled by assuming independence of events at different sites. DNA sequence changes which are constrained to avoid termination codons or to avoid particular classes of amino acids also will involve some violation of independence.

Given that we are willing to assume independence of evolution at different sites, it turns out that the probability of a given set of data arising on a given tree can be computed site by site, and the product of the probabilities taken across sites at the end of the computation. We therefore concentrate our efforts on computing this probability for a single site. For this we make use of the probabilities $P_{ij}(t)$, where i and j take values 1, 2, 3, and 4 corresponding to the four bases A, C, G, and T. $P_{ij}(t)$ is the probability that a lineage which is initially in state i will be in state j after t units of time have elapsed. We compute the $P_{ij}(t)$ later in this paper. For the moment we concentrate on obtaining an expression for the likelihood of the tree, namely the probability of the data given the tree.

We assume that after speciation two lineages evolve independently, and that the same stochastic process of base substitution applies in all lineages. It is possible to write a general expression for the likelihood of a tree, but it will be more useful to present the expression for a particular case, the tree in Fig. 1, since the general pattern will be clear from that expression. The lengths of the segments of the tree are given by the quantities v_i. If we knew the states (bases) at a particular site at points 0, 6, 7 and 8 on this tree, and these were s_0, s_6, s_7, and

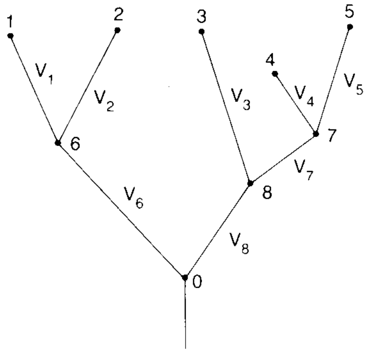

Fig. 1. The tree used in the discussion of computing the likelihood. The v's are the lengths of the segments

s_8, the likelihood of the tree would be the product of the probabilities of change in each tree segment, times the prior probability π_{s_0} of state s_0, so that it would be

$$L = \pi_{s_0} P_{s_0 s_6}(v_6) P_{s_6 s_1}(v_1) P_{s_6 s_2}(v_2) P_{s_0 s_8}(v_8) \\ P_{s_8 s_3}(v_3) P_{s_8 s_7}(v_7) \cdot P_{s_7 s_4}(v_4) P_{s_7 s_5}(v_5) \quad , \tag{1}$$

where s_i is the state at point i on the tree. In practice we do not know s_0, s_6, s_7, and s_8, so the likelihood will be the sum over all possible assignments of bases to those forks on the tree:

$$L = \sum_{s_0} \sum_{s_6} \sum_{s_7} \sum_{s_8} \pi_{s_0} P_{s_0 s_6}(v_6) P_{s_6 s_1}(v_1) \\ P_{s_6 s_2}(v_2) P_{s_0 s_8}(v_8) P_{s_8 s_3}(v_3) \\ \cdot P_{s_8 s_7}(v_7) P_{s_7 s_4}(v_4) P_{s_7 s_5}(v_5) \tag{2}$$

This expression will have 256 terms, and in general the expression for n species will have 2^{2n-2} terms, which can easily be a very large number.

Fortunately, a considerable economy can be realized by moving the summation signs rightwards in (2), obtaining

$$L = \sum_{s_0} \pi_{s_0} \{\sum_{s_6} P_{s_0 s_6}(v_6) [P_{s_6 s_1}(v_1)] [P_{s_6 s_2}(v_2)]\} \\ \{\sum_{s_8} P_{s_0 s_8}(v_8) [P_{s_8 s_3}(v_3)] [\sum_{s_7} P_{s_8 s_7}(v_7) \\ (P_{s_7 s_4}(v_4)) (P_{s_7 s_5}(v_5))]\} \quad . \tag{3}$$

Notice that the pattern of parentheses in expression (3) bears an exact relationship to the topology of the tree, since it is {[][]} {[][( )( )]}. There is one P for each segment of the tree. The expression can be evaluated by working outwards from the innermost parentheses. The correspondence between the parentheses and the form implies that this corresponds to starting at the tips of the tree and moving downward. We can restate this process in terms of conditional likelihoods: We define $L_s^{(k)}$ as the likelihood based on the data at or above point k on the tree, given that point k is known to have state s for the site under consideration. If point k is a tip, then $L_s^{(k)}$ will be zero for all s except that actually observed, for which $L_{s_k}^{(k)} = 1$. This enables us to start the computation by computing for each tip k a set of four $L_s^{(k)}$.

This evaluation of expression (3) is then exactly equivalent to the following algorithm. We work our way down the tree from the tips (in computer science parlance, we perform a postorder tree traversal). For point k, whose immediate descendants are i and j, we can compute for all four values of s_k

$$L_{s_k}^{(k)} = (\sum_{s_i} P_{s_k s_i}(v_i) L_{s_i}^{(i)}) (\sum_{s_j} P_{s_k s_j}(v_j) L_{s_j}^{(j)}) \quad . \tag{4}$$

If this process is continued until we reach the bottom fork on the tree, it can be seen that all of the terms in (3) have been computed. For the bottom fork, point 0 in our example, we will then have computed the four conditional likelihoods $L_{s_0}^{(0)}$ given the possible states of the site at point 0. The overall likelihood of the tree for the site under consideration is then

$$L = \sum_{s_0} \pi_{s_0} L_{s_0}^{(0)} \quad . \tag{5}$$

completing the calculation of (3). This algorithm was stated earlier for a more general case (Felsenstein 1973). I have dubbed it "pruning", since it in effect removes two tips from the tree at each step. The pruning procedure is closely analogous to the "peeling" algorithms widely used in pedigree analysis in human population genetics (Elston and Stewart 1971; Cannings et al. 1976), and to methods long used for evaluating polynomials in numerical analysis (Dahlquist et al. 1974, p. 14).

The π's must be the prior probabilities of finding each of the four bases at point 0 on the tree. Since we are assuming an evolutionary steady state in base composition, they reflect the overall base composition in the group under study. We will specify the $P_{ij}(t)$ in such a way that the probabilistic process leads to maintenance of this same base composition, which we assume is given from external evidence.

The Base Substitution Probabilities

We have not yet specified how the quantities $P_{ij}(t)$ are to be computed. These are the probabilities of transition from one base to another over a segment of length t. We shall assume that these probabilities reflect a Markov process, a process in which the probability of a base changing may depend on its current identity, but not on its past history. Kaplan and Langley (1979) stated a simple Markov process model of base substitution. Ours will be similar in spirit but different in detail.

We assume that in a small interval of time of length dt, there is a probability u dt that the current base at a site is replaced. The quantity u is the rate of base substitution per unit time. If a base is replaced, its replacement is A, C, G, or T with probabilities π_1, π_2, π_3, or π_4. Note that this means that a base could be replaced by the same base, so that not all substitutions are observable even in principle. Note also that this model makes no distinction between transitions and transversions. If we let δ_{ij} be 0 if $i \neq j$ and 1 if $i = j$ (the Kronecker delta function), then we are in effect assuming that for infinitesimal dt

$$P_{ij}(dt) = (1 - u\,dt)\,\delta_{ij} + u\,dt\,\pi_j \quad . \tag{6}$$

From this it can be shown in straightforward fashion that for arbitrary t,

$$P_{ij}(t) = e^{-ut}\,\delta_{ij} + (1 - e^{-ut})\,\pi_j \quad . \tag{7}$$

This follows almost immediately once one observes that e^{-ut} is the probability that the site does not change at all over a length of time t, and that if it does change the probability that it ends up in state j is π_j. This model of base substitution seems a useful compromise between realism and tractability.

One of the convenient properties of this Markov process model of base substitution is known as *reversibility*. This means that the process of base substituion will look the same whether followed forward or backward in time. Reversibility requires that for all i, j, and t

$$\pi_i P_{ij}(t) = P_{ji}(t)\,\pi_j \quad . \tag{8}$$

which is easily proven using (7). The Markov process model used by Kaplan and Langley (1979) was also reversible.

In equation (7), the probability of change from base i to base j depends on t only through their product ut. If we were to double u and halve t, there would be no change in the P_{ij}, so that the probabilities of various data sets would not be affected. Thus all we can infer is the product ut, not u or t individually. If u is the same for all lineages and all times, then ut should be proportional to the elapsed time t. In the absence of external evidence about u or about t, we can adopt the convention that u = 1, and thus measure t in units of expected numbers of substitutions. If u is allowed to differ from segment to segment on the treee, this will have the same consequences as letting t differ from elapsed time. Except in the case where u is to be assumed constant, we will not assume that t is elapsed time, but rather that it is time measured on a molecular clock which may run at different rates in different segments of the tree. Thus we will usually not require that the total length of segments from the bottom fork up to each tip be the same. The tips need not be contemporaneous on this molecular clock whose units are expected fractions of bases substituted.

The Pulley Principle

The reversibility of our Markov process and the absence of constraints on segment lengths can be used to establish an interesting and useful property of the estimation of evolutionary trees under this model. Consider the last two steps of our algorithm for calculating likelihoods. They involved (in our example) forks 0, 6, and 8 in the expression for the likelihood of the tree at one site:

$$L = \sum_{s_0} \pi_{s_0} \left(P_{s_0 s_6}(v_6)\, L_{s_6}^{(6)}\right)\left(P_{s_0 s_8}(v_8)\, L_{s_8}^{(8)}\right) \tag{9}$$

A short derivation (Appendix 1) can be used to show that L is unaffected if we add a length x to v_6 and subtract the same amount from v_8. In other words, L depends on v_6 and v_8 only through their sum $v_6 + v_8$. Since the likelihood L is our only basis for comparing evolutionary trees, this means that the tree whose likelihood is being calculated could have its root anywhere between points 6 and 8. The root of the tree is a sort of pulley, so that if all parts of the tree to one side of the

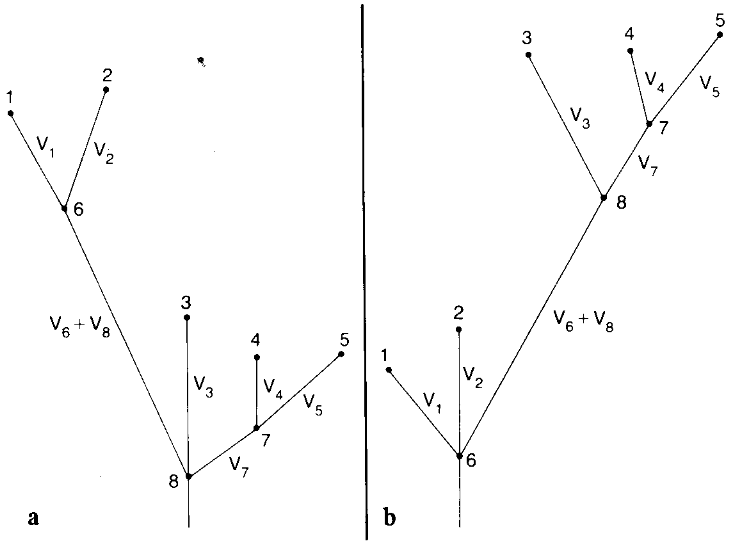

Fig. 2. Two trees whose likelihood will be equivalent to that in Fig 1 under the assumption of this paper

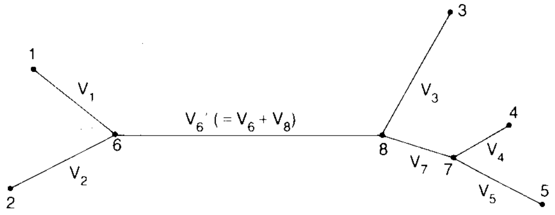

Fig. 3. The unrooted tree whose likelihood is equivalent to those in Fig 1 and 2

root are moved down, and all parts to the other side moved up by the same amount, the likelihood remains unaltered. The two trees in Fig. 2 cannot be distinguished from the tree in Fig. 1, for the same v's.

In fact, the argument can be applied repeatedly, and shows that there is no information whatsoever about the placement of the root of the tree. Fig. 3 shows the unrooted tree, which is what we are in effect estimating. The root of the tree can be placed anywhere on that tree without affecting the likelihood. We are estimating not a single rooted tree but an equivalence class of rooted trees, namely all those compatible with a given unrooted tree. This I dub the Pulley Principle, and it will apply whenever the Markov process of base change is reversible and the v_i are unconstrained by any requirement that the tips be contemporaneous.

Finding the Maximum Likelihood Tree

Our interest in the Pulley Principle is that it allows us to regard any given segment of the unrooted tree as containing the root. This in turn allows us to alter the length of that segment in an optimal fashion. We are interested in doing this because we still face a large computational task. We have a computationally feasible method for evaluating the likelihood of a given tree, but this leaves us with the task of finding the maximum likelihood tree. Consider the problem of finding values of the v_i which maximize the likelihood of the tree given a particular topology. We could do this by direct search, but this would require a very large number of computations of the likelihood, one for each set of v_i's tried. At a minimum, many hundreds of likelihood computations would be needed.

It is this problem which the Pulley Principle helps us to solve. It allows us to construct an algorithm which alters one of the v_i at a time, each one being altered to that value which results in the highest likelihood. This process continues until none of the v_i can be altered in a way which substantially improves the likelihood. At each stage one v is changed to the value which gives the greatest possible likelihood, given that only that v can be varied. Thus at each step the likelihood of the tree increases. This process cannot fall into an endless loop, since the likelihood can never decrease. We describe this iterative method below.

Searching Among Tree Topologies

There still remains the problem of examining many different tree topologies. It would in principle be possible to simply look at each possible unrooted tree topology, iterate the branch lengths to their optimal values for each one, and then pick that topology which has yielded the highest likelihood. This strategy is rendered impractical by the astronomical number of possible topologies for even moderate numbers of trees. Edwards and Cavalli-Sforza (1964) found that the number of unrooted bifurcating trees with n labelled tips was $(2n\text{-}5)!/[(n\text{-}3)!\ 2^{n-3}]$, which for 10 tips is more than 2 million topologies. I have given elsewhere (Felsenstein 1978a) a method for computing the number of rooted multifurcating trees. Since the number of rooted trees with n labelled tips is the same as the number of unrooted trees with n+1 labelled tips, we have immediately that there are over 12 million unrooted multifurcating trees with 10 tips. For 20 tips the number exceeds 18×10^{21}.

Obviously some less ambitious search strategy must be employed. The strategy I have found useful is to build the tree up by successively adding species to it, starting with a two-species tree. When the k-th species is being added to the tree, there will be 2k-5 segments from which it could arise. Each of these is tried and the maximum likelihood within the resulting topology evaluated, by the iteration technique presented below. The placement yielding the highest likelihood is accepted. If the tree now has more than four species, before the next species is added local rearrangements are carried out in the tree to see if any of these improves the likelihood of the tree. If any does, it is accepted and the local rearrangement process continues until a tree is found which no local rearrangement can improve.

This strategy of searching among possible topologies is not guaranteed to find the best topology, but I have found its performance satisfactory in practice. It depends on the order in which the species are added to the tree, so that if it is repeated with a different ordering of the species, a different result may be obtained. With extremely self-consistent data the same tree will result from all orderings of the input data. With less self-consistent data, different results will be obtained. This is in a sense an advantage, as it allows us to explore different regions of the likelihood surface, taking as our final estimate the result with highest likelihood. With n tip species this search strategy will examine at least $2n^2 - 9n + 8$ different topologies.

Finding Optimum Segment Lengths

Within each topology we examine, we must adjust the v_i to their maximum likelihood values. As already indicated, this is done by adjusting each of the v_i in turn according to a method which guarantees that each of the

v_i changes to that value which maximizes the likelihood of the tree, given the current values of the other v's. We now derive this iteration method.

Consider segment 7 of the tree in Fig. 3, the segment connecting nodes 7 and 8. We can consider the root to be located in this segment, and use the pruning algorithm given above to compute the sets of conditional likelihoods at points 7 and 8. Suppose that we take the root to be immediately to the right of point 8. The likelihood of the tree for one site is then given by

$$L = \sum_{s_0} \sum_{s_8} \sum_{s_7} \pi_{s_0} (P_{s_0 s_8}(0) L_{s_8}^{(8)}) (P_{s_0 s_7}(v_7)\, L_{s_7}^{(7)}) \tag{10}$$

$$= \sum_{s_0} \pi_{s_0} L_{s_0}^{(8)} (\sum_{s_7} P_{s_0 s_7}(v_7) L_{s_7}^{(7)}) \ ,$$

since $P_{s_0 s_8}(0) = \delta_{s_0 s_8}$.

Substituting into (10) the expression (7) with u = 1, this becomes after some simplification

$$L = e^{-v_7} \sum_s \pi_s L_s^{(8)} L_s^{(7)} + (1 - e^{-v_7}) [\sum_{s_8} \pi_{s_8} L_{s_8}^{(8)}] [\sum_{s_7} \pi_{s_7} L_{s_7}^{(7)}] \ . \tag{11}$$

This is the factor of the full likelihood which corresponds to one site. There is one factor like this for each site in the DNA, so that the full likelihood is of the form

$$L = \prod_i (A_i q + B_i p) \ , \tag{12a}$$

where $q = e^{-v_7}$, $p = 1-q = 1-e^{-v_7}$, and A_i and B_i are the terms

$$A_i = \sum_s \pi_s L_s^{(8)} L_s^{(7)} \tag{12b}$$

and

$$B_i = (\sum_{s_8} \pi_{s_8} L_{s_8}^{(8)}) (\sum_{s_7} \pi_{s_7} L_{s_7}^{(7)}) \tag{12c}$$

for the i-th DNA site. We want to find the value of v_7 which maximizes the likelihood. This is equivalent to finding the value of p which maximizes (12a) and then solving for $v_7 = -\ln(1-p)$.

Taking logarithms in (12a),

$$\ln L = \sum_i \ln (A_i q + B_i p) \tag{13}$$

and equating the derivative of this expression to zero,

$$\frac{d \ln L}{dp} = \sum_i \frac{B_i - A_i}{(A_i q + B_i p)} = 0 \ . \tag{14}$$

Since if there are K sites in all,

$$K = \sum_i 1 = \sum_i \frac{A_i q + B_i p}{A_i q + B_i p} = \sum_i \frac{B_i - (B_i - A_i) q}{A_i q + B_i p} \ . \tag{15}$$

We can use (14) to eliminate the terms containing q in the numerator of (15), obtaining

$$K = \sum_i \frac{B_i}{A_i q + B_i p} \tag{16}$$

This condition must be satisfied if the derivative of lnl is zero. Multiplying both sides of (16) by p, we can turn this into the iteration formula

$$p^{(k+1)} = \frac{1}{K} \sum_i \frac{B_i p^{(k)}}{A_i q^{(k)} + B_i p^{(k)}} \ . \tag{17}$$

where $q^{(k)} = 1-p^{(k)}$. The result is an equation with p on both sides, which must be satisfied when the likelihood has reached a relative maximum. Equation (17) is an iterative version of that equation.

The iteration (17) is a specific case of the general EM algorithm of Dempster et al. (1977), which is guaranteed never to go downhill on the likelihood surface. We use it by proceding through the evolutionary tree, iterating the p_i one after another. Each one of the p_i is iterated until (17) converges, before moving to the next one. We iterate until we can made a complete pass through the tree, iterating all of the p_i, without any of them changing substantially. We presume that this represents a maximum of the likelihood within the given topology. We can then use (13) to evaluate the likelihood of the resulting tree. In fact, we are only guaranteed that we have increased the likelihood and have arrived at a stationary point. It could be a saddle-point rather than a maximum. I have yet to encounter such a case in practice. If one of the p_i iterates to zero, this indicates that we could not find a stationary point of the likelihood within the topology, and rearrangement of the tree is indicated.

A Computer Program

The above algorithm has been incorporated into a computer program, written in PASCAL by Mark Moehring. The program computes the estimates in terms of the p_i rather than v_i, as the former seem more meaningful. The program is part of a package of programs for numerical estimation of evolutionary trees. This package will be supplied on request, written in standard ANSI format on a magnetic tape supplied by the recipient. It must be acknowledged that this computer program is quite slow, and could be effectively used only by someone who had free computer time available. The other programs in the package do not share this difficulty.

Extensions

There are many natural directions in which the present scheme can be extended. It is straightforward to incorporate into the current algorithm the case in which some bases are not known unambiguously in the original data. If (say) site 3 could be either an A or a G, then it is merely necessary ot note that, by the definition of the conditional likelihood at that site $L_1 = L_3 = 1$ and $L_2 = L_4 = 0$. The result will be a correctly computed likelihood.

Allowing some sites to be "hot spots" is also straightforward. If each site has probability x of having substitution rate u_1 and 1-x of having substitution rate u_2, then if L(u) is the likelihood for the whole tree at a site given substitution rate u, the overall likelihood will be $x\ L(u_1) + (1\text{-}x)\ L(u_2)$. The iteration method can be appropriately altered to correspond to this model. It would also be quite easy to allow different substitution rates at the three positions of each codon.

Hypothesis Testing

The availability of maximum likelihood estimation makes available hypothesis testing by the likelihood ratio test. One could in principle test constancy of the rate of substitution. This would require some way of maximizing the likelihood under the constraint that all tips are contemporaneous. This constraint is not maintained in the current iteration method, but the likelihood evaluation method given by equations (4), (5), and (7) above could be used together with a direct search method. The likelihood ratio test of constancy of rate of evolution would have n-2 degrees of freedom if there were n tip species. Langley and Fitch (1974) have tested constancy of rate of protein evolution. Their test used as data ancestral sequences inferred by a parsimony method and thus does not constitute a likelihood ratio test of the sort carried out here. The present methodology could in principle be extended to protein data, but the computational effort would be prohibitive.

One could also test whether alternatives to the maximum likelihood topology were acceptable. One could test this in a crude way by evaluating the curvatures of the log-likelihood surface and using this to obtain an asymptotic covariance matrix of the v_i. If this indicates that one of them could be zero, this implies that alternative branching patterns in that portion of the tree may be acceptable. This is one of the great advantages of the likelihood approach (or any statistical approach) – it gives us an indication of the amount of uncertainty in our estimate.

In practice this covariance matrix is only obtained with some computational difficulty. A more limited indication of the statistical error can be obtained by obtaining only the variances of the segment lengths. These are more easily computed. Each such variance is the inverse of the curvature of the likelihood surface when all but one of the v_i are held fixed. Recall that the log likelihood as a function of the i-th segment length is given by equation (13) above. Its derivative with respect to p is given by equation (14) above. The second derivative is

$$\frac{d^2L}{dp^2} = -\sum_i \frac{(B_i - A_i)^2}{(A_i q + B_i p)^2} \quad . \tag{18}$$

The asymptotic variance of our estimate of p will be

$$-1/\frac{d^2L}{dp^2} = 1/\sum_i \frac{(B_i - A_i)^2}{(A_i \hat{q} + B_i \hat{p})^2} \tag{19}$$

The quantities A_i, B_i, $\hat{p}$ and $\hat{q}$ are needed in the iterative method of computing our estimate of p, and are presumably readily available. This variance can be readily converted into a corresponding variance of $\hat{v}$ by dividing (19) by $\hat{q}^2$.

The variance thus obtained is an underestimate of the true asymptotic variance, since the curvature (19) overestimates the true curvature which would be obtained if we allowed all the p_j to vary at once. The likelihood surface must fall off at least as quickly when only p_i can vary as it will as a function of p_j when the other p_k are allowed to vary so as to partially compensate for the effects of p_j.

A Numerical Example

As a computational example, the above computer program has been applied to some of the eukaryotic 5S RNA sequences tabulated by Erdmann (1979). A cursory examination of the sequences shows that some deletion and insertions seems to have gone on. Since these processes are not incorporated in the model, and since the computer program is quite slow, attention was confined to the five vertebrate species (trout, Xenopus, turtle, iguana, and chicken). These seem to have homologous sequences. The 3' terminal base of all but Xenopus were omitted, so that all 120 positions could be compared. The sequences used were those labeled c. (a), Re, R. T., Tu, and X. L. S. in Erdmann's table. For the purposes of this example, the frequencies of the four

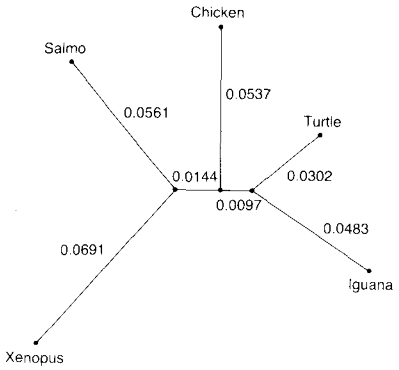

Fig. 4. The maximum likelihood estimate of the phylogeny for 5S RNA sequences from five vertebrate species

bases were all taken to be 0.25, though the result is not particularly sensitive to this assumption.

Figure 4 shows the result, with the value of $\hat{p}_i$ given next to each segment of the tree. The topology preferred is a nonstandard one. However, the lower bound variances on the segment lengths are very large, corresponding to the small amount of sequence divergence between these groups, and hence the small amount of information contained in 120 nucleotides. The segments leading to tips had 95% confidence limits which allowed them to range from near zero to about twice the estimated length shown in Fig. 4. The two interior segments could be as much as three times the length shown or could be negative in length (if the estimated length was x, the confidence limits went from about -x up to about 3x). This, being based on a lower bound on the variance, clearly implies that we cannot exclude nearby alternative topologies for the tree. The support that this data supplies for this nonstandard ancestry of turtles is very weak indeed.

This illustrates a strength of the statistical inference approach. Had we obtained this tree by a parsimony method we would have no way of knowing how much credence to give the fact that one tree required one more base substitution than another. A maximum likelihood estimate usually comes equipped with some indication of its own error. Cavender (1978) has made a start at computing the confidence sets for parsimony methods, but until his approach can be extended to nucleotide data and to more than four species likelihood methods seem to have the edge.

Limitations

The model used here is highly idealized, and the precision of the statistical inferences must be reduced by a factor representing one's skepticism of the assumptions involved. The absence of deletions and insertions, as well as of constraints on amino acid substituion, are particular sources of concern.

If one is fitting the present model with the intention of justifying the constancy of rate of substitution by the neutral mutation theory, then the phenomenon discovered by Gillespie and Langley (1979) will also be a source of skepticism. They showed that if $4N_e$ is large the number of substituions in a given segment of the tree has a variance greater than that of the Poisson distribution. Their work also implies a correlation in the number of substitutions in the two segments issuing from the same fork. The distribution which they computed is a distribution over independent loci. Neighboring DNA sites never separated by recombination would show numbers of substitutions, drawn from the same Poisson distribution (numbers at different loci would come from Poisson distributions with different means). It would be of interest to know where the transition from the one behavior to the other occurs. One expects on intuitive grounds that it is near the point where the two sites are sufficiently far apart that the recombination between them, r, exceeds $1/(2N_e)$.

References

Camin J H, Sokal R R (1965) Evolution 19:311–326

Cannings C, Thompson E A, Skolnick M H (1976) Adv Appl Probab 8:622–625

Cavender J A (1978) Math Biosci 40:271–280

Chakraborty R (1977) Can J Genet Cytol 19:217–223

Colless D H (1970) Syst Zool 16:289–295

Dahlquist G, Björck A, Anderson N (1974) Numerical Methods. Englewood Cliffs, Prentice Hall, New Jersey

Dempster A P, Laird M N, Rubin D B (1977) J R Stat Soc B 39:1–38

Edwards A W F (1963) Heredity 18:553

Edwards A W F, Cavalli-Sforza (1964). Phenetic and phylogenetic classification. Heywood V H, McNeill J (eds) Systematics Association Publication No 6 Systematics Association, London, pp 67–76

Elston R C, Stewart J (1971) Hum Hered 21:523–542

Erdmann V A (1979) Nucleic Acids Res 6:r29–r44

Estabrook G F, Landrum L (1975) Taxon 24:609–613

Felsenstein J (1973) Syst Zool 22:240–249

Felsenstein J (1978) Syst Zool 27:27–33

Felsenstein J (1978) Syst Zool 27:401–410

Felsenstein J (1979) Syst Zool 28:49–62

Ferris S D, Portnoy S L, Whitt G S (1979) Theor Popul Bio 15:114–139

Fitch W M, Margoliash E (1967) Science 155:279–284

Gillespie J H, Langley C H (1979) J Mol Evol 13:27–34

Holmquist R (1972) J Mol Evol 1:115–133

Kaplan N, Langley C H (1979) J Mol Evol 13, 295–304

Kashyap R L, Subas S (1974) J Theor Biol 47:75–101

Langley C H, Fitch W M (1974). J Mol Evol 3:161–177

Le Quesne W J (1969) Syst Zool 18:201–205

Neyman J (1971) Statistical decision theory and related topics Gupta S S, Yackel J (eds) Academic Press, New York, pp 1–27

Sneath P H A, Sackin J J, Ambler R P (1975) Syst Zool 24: 311–332

Received October 22, 1980/Revised April 8, 1981

Appendix

Proof of the Pulley Principle

Starting from equation (9), we can invoke the reversibility property (8) to note that

$$\pi_{s_0} P_{s_0 s_6}(v_6) = \pi_{s_0} P_{s_6 s_0}(v_6) \tag{A1}$$

and then (9) can be written as

$$L = \pi_{s_6} L_{s_6}^{(6)} L_{s_8}^{(8)} \sum_{s_0} P_{s_6 s_0}(v_6) P_{s_0 s_8}(v_8) \ . \tag{A2}$$

If the process whose probabilities are given by $P_{ij}(t)$ is a Markov process, then the Chapman-Kolmogorov equation given in any text on stochastic processes replaces the summation on the right of (A2) by $P_{s_6 s_8}(v_6 + v_8)$. This shows that the likelihood depends only on the sum $v_6 + v_8$, so that we may place the root anywhere in the segment of the tree connecting point 6 with point 8 (and in fact may move it elsewhere as well) without affecting the likelihood.