J Mol Evol (1990) 31:330–342

Journal of Molecular Evolution

Iterative Character Weighting Based on Mutation Frequency: A New Method for Constructing Phyletic Trees

A. van Ooyen and P. Hogeweg

Bioinformatics Group, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands

Summary. In this paper we present an iterative character weighting method for the construction of phyletic trees. An initial tree is used to calculate the character weights, which are the number of mutations normalized so that the possible range is corrected for. The weights obtained are used to adjust the tree; this process is iterated until a stable tree is found. Using data generated according to a model tree, we show that the trees constructed by the iterative character weighting method converge to the true underlying tree. Using biological data, the trees become closer to the systematic classification of the species concerned, and patterns conflicting with the phylogenetic pattern can be singled out. The method involves a combination of minimal length methods and similarity methods, whereby the strict parsimony criterion is relaxed.

Key words: Character weighting — Phyletic trees — Parsimony — Tree similarity

Introduction

Protein and DNA/RNA sequences are preeminently suitable data with which to construct evolutionary trees (phyletic trees). The minimum number of mutations is often used as a criterion for selecting the "best" tree. The only way to be sure of finding the minimal tree is to generate all possible trees and select the one with the minimum number of mutations. Because for more than eight compared sequences the number of possible trees "explodes," studying all trees becomes impossible, and one must resort to less successful tree-generating methods, which will not necessarily produce the minimal tree.

Parsimony fails to converge to the "true" tree if mutation rates are very unequal in different lineages and not very low in most characters, as Felsenstein (1981) has demonstrated. He suggested, therefore, the use of methods that do not strictly minimize mutational cost. Such methods have to use the information value of characters, as characters are not equally informative due to back mutations, convergence (possibly as a result of functional constraints), etc. In the context of taxonomic pattern recognition, Hogeweg (1976a,b) has demonstrated that iterative character weighting enables us to filter out noisy and inconsistent characters, and to single out conflicting patterns. In this method relative weights are assigned to characters according to the degree of character difference between the clusters. When using molecular data, one usually evaluates characters along the branches of a tree, instead of on the basis of their distribution in clusters. Farris (1969) was the first to apply an iterative weighting procedure for inferring phylogenies. In this procedure, the weights are allotted to characters according to the degree of compatibility with a tree. Here we introduce a modified procedure, which uses as character weights the number of mutations normalized so that the possible range in the number of mutations is corrected for. In this method an initial tree is used to calculate the weights and the weights obtained are used to adjust the tree; this process is iterated. Trees are constructed by a method that can correct for unequal mutation rates in different lineages [i.e., the present-day-ancestor method (Klotz et al. 1979; Klotz and Blanken 1981; Blanken et al. 1982)].

We show that in the case of artificial data the

Offprint requests to: A. van Ooyen

constructed trees converge to the true underlying tree and that in the case of biological data the trees converge to one or more patterns. The method is a tool for singling out conflicting patterns. The number of mutations tends to decrease in successive trees, although the minimal tree is usually not reached. One of the advantages of the method is that the minimum mutation criterion can be relaxed, which was Felsenstein's suggestion.

Methods

Introduction

The mutation rate can differ greatly both between molecules and within molecules. For example, cytochrome c has an extremely low mutation rate, whereas fibrinopeptides have a high mutation rate. Within molecules, both highly conserved sites and variable sites (e.g., third positions in coding sequences) occur. Moreover, even if the expected mutation rate is the same at all positions, the actual amount of change will vary because of chance. If the mutation rate is either too low or too high (relative to the time scale of speciation events), information about the phylogeny cannot be preserved. Therefore, molecules mutating slowly should be used to assess phylogenetic relationships among distantly related species, and molecules mutating rapidly should be used to assess phylogenetic relationships among closely related (sub)species.

In our method for constructing phyletic trees, we use parts of sequences that mutate rapidly and parts of sequences that mutate slowly in the same way as we use sequences with different rates of change. Then positions with a low mutation frequency to a large extent define the overall structure (main groups) of the tree and positions with a high mutation frequency to a large extent define the detailed structure (subgroups). For this purpose positions with a small number of mutations should be weighted more highly than positions with a high number of mutations. However, the number of mutations cannot be calculated without a phyletic tree. In order to solve this bootstrapping problem, we use an initial tree based on equally weighted characters to determine the number of mutations per position. The latter are then used to calculate character weights. The weights obtained are used to adjust the tree. This process is iterated until a stable tree is found, i.e., until two successive trees are the same.

Outline of the Method

The general outline of the procedure is as follows:

1) A similarity matrix is constructed using equally weighted characters. For simplicity, we assume that the sequences are already aligned (see Discussion).

2) The similarity matrix is used to construct a phyletic tree. Any tree construction method that works with a similarity matrix can be used in this step, although it is preferable to use the present-day-ancestor method (Klotz et al. 1979; Klotz and Blanken 1981; Blanken et al. 1982) (see Discussion).

3) Each character is evaluated according to its degree of compatibility with the given tree, i.e., the number of mutations is determined at each position. In order to obtain the number of mutations, internode sequences are constructed so as to minimize the total number of mutations in the given tree. This is done by using the back-tracking algorithm described in Hogeweg and Hesper (1984) and in Konings et al. (1987), which is equivalent to the first pass of Fitch's (1971) parsimony procedure. It gives a minimal mutation solution only if a binary similarity criterion is used. To calculate the weights, we normalize the number of mutations obtained so as to correct for the possible range. The weights are used to adjust the similarities.

4) The process is repeated from step 2 until a stable tree is found or until an identical tree topology recurs.

The method differs from other weighting schemes (Farris 1969; Penny and Hendy 1985) in that it combines minimum mutation methods and similarity methods. The weighting formula used here differs from that of Farris (1969) in that it incorporates information content. Instead of using pairwise character compatibility in order to weight positions, as do Penny and Hendy (1985), we use the degree of character compatibility with the tree and try to increase it during the iteration.

Implementation of the Method

The specific implementation of the general scheme used in this paper is described in this section.

The similarity between two sequences is computed as

$$s(a, b) = \left[\sum_{i=1}^{n} \delta_{a(i),b(i)} \cdot W_i\right] \Big/ \left(\sum_{i=1}^{n} W_i\right) \qquad (1)$$

where, s(a, b) is the similarity between sequence a and b, n is the total length of the sequence expressed as the number of nucleotides (amino acids), $\delta_{a(i),b(i)}$ is 1 (same state at corresponding positions of the sequence) or 0 (different states), and W_i is the weight at the ith position (initially, $W_i = 1$ for all i) [see Eq. (2)].

Trees are constructed by the present-day-ancestor method (Klotz et al. 1979; Klotz and Blanken 1981; Blanken et al. 1982). In this method it is assumed that there is a tree topology in the data, and, if this assumption is warranted, we are able to find the correct tree topology (an unrooted tree), notwithstanding variable mutation rates in different lineages. For this purpose, the pairwise distances are converted to correct for their shared distance from a putative common present-day-ancestor. Any ancestor can be chosen. If there is a tree topology in the data all ancestors will yield the same correct tree topology (and therefore the minimal length tree). On the other hand, if there is no tree topology in the data, each ancestor can yield a different tree. There can be as many different tree topologies as there are ancestors.

Weights are computed using the following formula:

$$W_i = (MAX_i - MIN_i)/(NMUT_i - MIN_i + 1) \qquad (2)$$

where W_i is the weight at the ith position, $NMUT_i$ is the number of mutations (along the branches of the tree) at that position, MAX_i is the maximum number of mutations possible at that position, and MIN_i is the minimum number of mutations possible at that position. The value of MAX_i and MIN_i depend only on the distribution of nucleotides (amino acids) at the ith position, and are constants of a data set. Under the condition that internodes are constructed so as to minimize mutational cost, MIN_i and MAX_i are computed as

$$MIN_i = NSTATES_i - 1 \qquad (3a)$$

$$MAX_i = NOTU - NFSTATE_i \qquad (3b)$$

where $NSTATES_i$ represents the number of different states at the ith position, NOTU is the number of OTUs included in the data set, and $NFSTATE_i$ is the number of OTUs in which the most frequent state at the ith position occurs.

Annotations to the weighting Eq. (2) include the following.

1) The actual number of mutations is corrected for the possible range of mutations, so that positions having a distribution of states that require a high number of mutations (relative to MIN_i) will get higher weights if the actual number of mutations is low, than positions having a distribution of states for which $MAX_i - MIN_i$ is small.

2) Such a range normalization of the mutational frequency can be regarded as a way of measuring the information content of a position. The qualitative behavior of $MAX_i - MIN_i$ as a function of the distribution of states is equivalent to that of information measures.

3) A position can get zero weight if $MAX_i = MIN_i$. Only positions at which two or more states occur with a frequency greater than one will get a weight not equal to zero. Only these positions are relevant for the construction of the minimal tree.

4) Positions not bearing information about the phylogeny may nevertheless have a high value for $MAX_i - MIN_i$. Suppose, for example, that at a certain position the bases A and G are completely interchangeable as far as their function is concerned. Provided that the mutation rate is sufficiently high, one expects in a data set an equal number of As and Gs, yielding a high value of $MAX_i - MIN_i$. However, the majority of the other characters are not expected to be compatible with this spurious character, so on a tree this position requires a high number of mutations, which will then result in a low weight, although $MAX_i - MIN_i$ is large. Generally, a set of informative characters compatible with a phyletic tree is generated during the iteration. A character that got a high weight because it was compatible with the given tree will have a greater share in the construction of the next tree. This character and its associated tree will be supported by characters that are compatible with the new tree.

5) In this paper, all initial trees are constructed with equally weighted characters, which include characters for which $MAX_i = MIN_i$. If we were to use the value of $MAX_i - MIN_i$ as initial character weight, unreliable characters (as mentioned under 4) would get too high a weight. Then they would have such a strong influence on the initial tree that all successive trees would be biased.

6) Without further adjustments, Eq. (2) can be applied to binary characters as well as to multistate characters. This weighting formula can thus be applied to both nucleotide data (4 states) and protein (20 states) data.

Evaluation of the Method

The method proposed above is investigated closely with the help of artificial data as well as biological data. Two artificial data sets are generated according to a specific tree (the true historical tree) in order to test whether the method improves the initially reconstructed tree so that it better approximates the true tree. To measure the deviation between trees, two methods are used: the partition metric of Robinson and Foulds (1981) and the tree cohesion method introduced here. Minimal trees are generated by the branch and bound algorithm (Hendy and Penny 1982).

Generation of Artificial Data: Simulation Model

There are two concepts of evolution with respect to the significance of selection: evolution as a process predominantly determined by positive selection (adaptation) and evolution as a process caused by random fixation (initiated, for example, by isolation) of selective neutral or nearly neutral mutations (neutral theory; this does not exclude negative selection). The model of evolution used in this paper is a minimal one, as it is based on the neutral theory. Neither positive nor negative selection is incorporated, and positions evolve independently of other positions. Because in our model the tree generation precedes the sequence generation, we assume, in accordance with the neutral theory, that the emergence of a new species is caused by isolation rather than by adaptation of the genome. In fact, in all existing tree construction methods it is assumed that selection does not play a significant role in speciation.

Tree Generation. The method used for the generation of trees is essentially the same as that described in Raup et al. (1973). Starting with one ancestral lineage, the model generates new lineages with a constant probability by creating new branches from already existing lineages. Each lineage also has a constant probability of becoming extinct.

Sequence Generation. The tree generated in this way is used to generate a set of sequences by passing an ancestral sequence through the tree. The sequence is duplicated at each branching point. The newly formed sequence will subsequently evolve independently of its ancestor. Every sequence is subjected to mutation until it reaches a terminal point of the tree. Only point mutations take place. Insertions and deletions are not allowed, so the sequences need not be aligned to compute similarities. We start with a random ancestral nucleotide sequence (a gene), in which each base has an equal chance of being included. At every time-step each nucleotide at every position in the sequence has a constant and equal probability of being replaced by three of the other nucleotides. Although the expected number of replacements per unit time is equal at all positions, the actual number of mutations will differ due to chance. The sequences of all the terminal nodes of the tree are used to reconstruct the true tree. The sequences of the extinct species also are used, so an unequal mutation rate in different lineages is incorporated. This makes the reconstruction of the true tree more difficult.

Generated Trees. Two specific trees, as generated by the model, are used to check the method. Tree 1 (spanning 49 species) was generated using a branching probability of 0.008, an extinction probability of 0.003, and a simulation time of 500 time-steps. On the basis of this tree, sequences of 200 nucleotides were generated using a mutation probability of 0.001. A tree of only 11 species was generated, so the obtained trees could be compared with the minimal tree. The tree of 11 species was generated using a braching probability of 0.045, an extinction probability of 0.015, and a simulation time of 100 time-steps. On the basis of this tree (henceforth referred to as tree 2), sequences of 150 nucleotides were generated using a mutation probability of 0.01. In these artificial data sets the mean and the variance of the dissimilarity (tree 1: 0.46 and 0.027, respectively; tree 2: 0.47 and 0.028, respectively) are of the same order of magnitude as in biological data sets.

Biological Data

We apply our method not only to simulated data, but also to biological data. By comparing the results obtained with biological data and artificial data, we may highlight processes that are not included in our minimal description of molecular evolution but that do take place in real evolution. As protein data, we use the 40 N-terminal residues of the small subunit of ribulose-1,5-biphosphate carboxylase (henceforth referred to as RBC-SSU) of 16 angiosperm species (see also van Ooyen 1987). The small subunits are encoded in the nucleus and together with the large subunits (encoded in the chloroplast) form an active enzyme, which in the chloroplast is involved in catalyzing the CO_2 fixation reaction. The sequence data were taken from Martin et al. (1983), Martin and Dowd (1984a,b,c), and Smeekens et al. (1986). As nucleotide data, we use hemoglobin A sequences of 11 mammalian taxa taken from table 2 of Penny et al. (1982). Although this table lists only positions that are relevant for the construction of minimal trees, we use the same data so that we can compare our results with those of Penny et al. (1982).

Tree Similarity

Two methods are used to measure the (topological) deviation of a reconstructed tree from the underlying model tree. The first method is the partition metric of Robinson and Foulds (1981)

for comparing trees. By deleting an internal link in a binary tree we partition the set of objects in the tree into two subsets. From a binary tree spanning n objects we can derive n − 3 such partitions, there being n − 3 internal links to delete. The difference between two trees is now measured as the number of partitions that can be derived from either of the trees but not from both. For n objects the partition metric takes even values between zero (identical trees) and 2n − 6 (no partitions in common). Small differences between trees are usually statistically significant because of the large number of possible trees.

There are two drawbacks to this method: (1) major and minor partitions are equally weighted, whereas humans, when comparing trees, weight overall similarity more highly than similarities in detailed structure; and (2) a single replacement high up in the tree affects many lower partitions.

In the second method, called tree cohesion, these shortcomings are remedied. Instead of counting the number of identical partitions, we measure the similarity between partitions using Watanabe's (1969) entropic measure of similarity and cohesion. The tree cohesion has the following properties: (1) major partitions are more highly weighted than minor partitions, as Watanabe's C has higher values for partitions dividing the set of OTUs into equal subsets than for partitions dividing the set into unequal subsets; (2) a partition need not match perfectly a partition of the other tree in order to increase the similarity between the two trees; and (3) each tree topology has its own maximum value, which is determined by comparing the tree with itself. Chained trees have the highest maximum value. Note that the tree cohesion measures method similarity and the partition metric dissimilarity. For further details, see the Appendix.

Cluster Analysis on Trees

Having a measure for comparing trees, we studied trees by means of cluster analysis in which trees are treated as objects. Dendrograms of trees were generated using UPGMA (Sneath and Sokal 1973) as clusterings criterion and the partition metric as similarity measure.

Results

As a result of weighting, the present-day-ancestor trees become more similar to each other. This holds for trees constructed on the basis of artificial data as well as for trees constructed on the basis of biological data. Thus, weighting causes convergence in tree structure. The trees constructed on the basis of artificial data converge to the true underlying model tree. The trees constructed on the basis of biological data are improved by weighting in the sense that they become closer to the systematic classification of the species concerned. Furthermore, the pattern reflecting phylogeny is separated from other patterns (i.e., convergence to more than one pattern). It turned out that the best tree is not the most parsimonious one (this statement holds for artificial and biological data). We illustrate the information used in minimum mutation methods and in our method by studying a specific tree. This reveals that our method remedies some of the weaknesses of minimum mutation methods.

Artificial Data

Tree 1 (49 Species)

If the whole sequence was used to reconstruct the true tree, the initial trees were already very similar to the true tree and could hardly be improved by iterative character weighting. Hence, only small parts (stretches of 20 nucleotides) of the sequence were used. This also enabled us to investigate whether weighting brings about convergence in trees based on different stretches of the sequence. Most of the trees were constructed using the first 20 positions of the sequences. Other parts were less extensively studied using only a few species as present-day-ancestor. The results are summarized in Table 1.

Convergence. In most cases the stable trees (which are often obtained within three iteration cycles) are more similar to the true tree than are the initial trees, as measured by the two similarity measures. This holds for trees constructed on the basis of different parts of the sequence as well as for trees generated using different ancestors.

From Table 1 we can make the following observations about the convergence to the true tree.

1) Although the improvement in terms of the number of partitions that perfectly match partitions of the true tree is usually rather small, the trees show that these are partitions of the higher order structure of the tree (see also Fig. 1). This is reflected by the tree cohesion, which then increases by a considerable amount, as similarities in the higher order structure make a larger contribution to the total similarity. The tree cohesion can also increase as a result of partial improvements; these improvements do not contribute to the partition metric.

2) Partitions involving the major structure of the tree sometimes improve at the expense of partitions involving the detailed structure of the tree. This happens when the tree cohesion measure detects improvement while the number of partitions that perfectly match partitions of the true tree does not increase.

3) Generally, the number of mutations drops after weighting. The decrease is mostly in the range of one to four, but can sometimes be rather substantial: as many as eight mutations are observed. Sometimes the number of mutations increases as a result of weighting (slightly in most cases). If the increase was rather substantial, the weighting procedure was found to generate trees that were less similar to the true tree than were the initial trees. When biological data are used, a large increase in the number of mutations should be regarded as an indication that the weighting procedure has picked out an alternative pattern that does not represent phylogeny.

Table 1. Results of weighting using artificial data (tree 1, 49 species)

Part of sequence	Present-day-ancestor	Tree cohesion		Partition metric		Number of mutations	
		Before	After	Before	After	Before	After
1–20	1	0.36; 0.39	0.41; 0.43	42	36	94	90
	3	0.36; 0.35	0.36; 0.35	42	40	94	93
	6	0.40; 0.42	0.42; 0.41	42	38	92	89
	8[a]	0.35; 0.35	0.38; 0.37	46	44	94	94
	9	0.39; 0.40	0.39; 0.39	46	46	94	94
	11	0.33; 0.35	0.35; 0.33	44	42	97	96
	18	0.40; 0.38	0.42; 0.43	32	30	91	89
	19	0.35; 0.37	0.35; 0.37	48	46	95	93
	21	0.35; 0.36	0.35; 0.36	42	44	94	94
	24[b]	0.35; 0.35	0.35; 0.35	46	40	94	96
	28	0.38; 0.37	0.41; 0.43	40	34	97	89
	32	0.37; 0.38	0.39; 0.39	46	38	95	92
	33	0.36; 0.35	0.36; 0.36	40	40	96	93
	35	0.35; 0.37	0.38; 0.40	44	42	95	95
101–120	1	0.38; 0.40	0.39; 0.40	38	34	86	86
	24	0.36; 0.37	0.36; 0.38	46	46	89	93
	28	0.35; 0.38	0.35; 0.39	46	48	94	97
	32	0.36; 0.38	0.33; 0.37	50	58	93	102
	35	0.36; 0.36	0.37; 0.36	46	44	90	92
171–190	1	0.33; 0.34	0.34; 0.37	50	52	98	101
	24	0.31; 0.35	0.35; 0.38	48	52	97	100
	32	0.30; 0.29	0.37; 0.35	50	50	107	100
	35	0.31; 0.32	0.30; 0.32	56	56	101	107

Results are shown of different present-day-ancestors and different parts of the whole sequence (stretches of 20 nucleotides). The deviation of the reconstructed tree (before and after weighting) from the true underlying model tree is measured by the tree cohesion measure and the partition metric of Robinson and Foulds (1981). Note that the tree cohesion measures similarity and the partition metric dissimilarity. For tree 1, the maximum value of the tree cohesion is 0.460; this is reached if the reconstructed tree and the true tree are identical. The partition metric can range from 0 (identical trees) to 92 (2n − 6; no partitions in common). Number of mutations before and after weighting are shown. The true tree has 90 mutations for part 1–20, 90 mutations for part 101–120, and 99 mutations for part 171–190. These are not the number of mutations that actually occurred in the simulation, but the numbers that are the minimum required on the true tree. The actual numbers are higher

[a] The final, stable tree (the fifth tree after the initial tree) is not the best reconstruction of the true tree; the third tree is much better (0.395; 0.391, and 40) and of all the six generated trees (initial + five iterated trees) has the lowest number of mutations (93)

[b] No stable tree is found; cycling occurs between the tree shown in the column After weighting and a tree of 95 mutations. The latter tree bears little resemblance to the true tree (0.343; 0.341, and 46)

4) It appears that the true tree is not the most parsimonious one. Although we did not generate the minimal tree(s) (because of the large number of species), we know that there are trees that have fewer mutations than has the true tree: for part 101–120 we found a tree with 86 mutations, whereas the true tree has at least 90 mutations (see Table 1).

Tree 2 (11 Species)

Convergence. The seven different initial trees obtained using each of the 11 species in turn as ancestor differ from each other on average by 4.67 pmu (partition metric units), whereas the five trees obtained by weighting using the seven trees as initial tree proposals, differ from each other on average by 3.20 pmu. The true tree is found among the five trees obtained by weighting; it is the tree having the fewest mutations (which is one more than the minimum). Comparison of the dendrogram of trees before and after weighting (Fig. 2) once again shows a marked convergence in tree structure after weighting. Note that the scale is identical for the two dendrograms, and that all trees tend to converge to the true tree. The average dissimilarity between the seven unweighted trees and the true tree is higher than between the true tree and the five trees obtained by weighting, being 4.29 pmu and 2.80 pmu, respectively. However, the tree to which most of the ancestors converge (see Fig. 3) differs from the true tree in that species 7 is wrongly placed between group (9,11) and (1,10) instead of being chained to 5, as in the true tree (causing only one not matching partition).

Comparison with Minimal Trees. The minimal tree, which has two mutations less than the true tree, differs from the true tree by 6 pmu, whereas the tree to which most of the ancestors converge after weighting (the converged tree) differs from the true tree by only 2 pmu. The five (minimal + 1) trees do not resemble the true tree as much as does the converged tree. The minimal tree differs from the

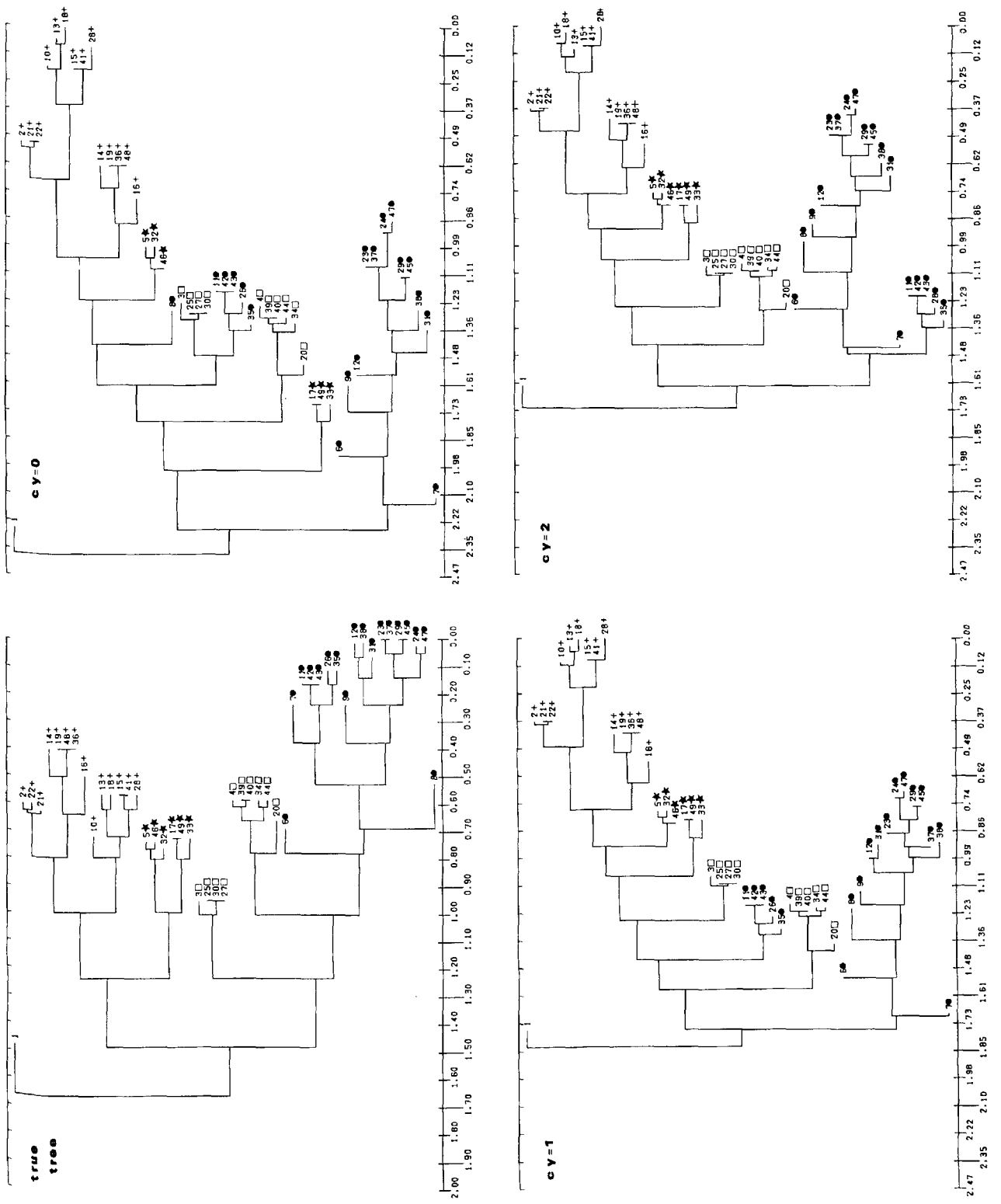


Fig. 1. True tree (tree 1, 49 species) and successive trees generated by the iterative weighting procedure (present-day-ancestor method, ancestor = 1) using the first 20 positions of the generated sequences. Groups are distinguished in the true tree in order to highlight improvements. Horizontal axis: dissimilarity level at which the species are grouped. **a** True tree rerooted on species 1. Number of mutations = 90. **b** Cycle (cy) = 0. Initial present-day-ancestor tree. Number of mutations = 94. Dissimilarity to true tree according to partition metric = 42. Tree cohesion similarity measures: TC(true tree, $T_{cy=0}$) = 0.36. TC($T_{cy=0}$, true tree) = 0.39. TC(true tree, true tree) = 0.46. **c** Cy = 1. Present-day-ancestor tree after one iteration. Number of mutations = 92. Partition metric = 38. Tree cohesion = 0.39; 0.41. **d** Cy = 2. Number of mutations = 90. Partition metric = 36. Tree cohesion = 0.41; 0.43. Stable tree.

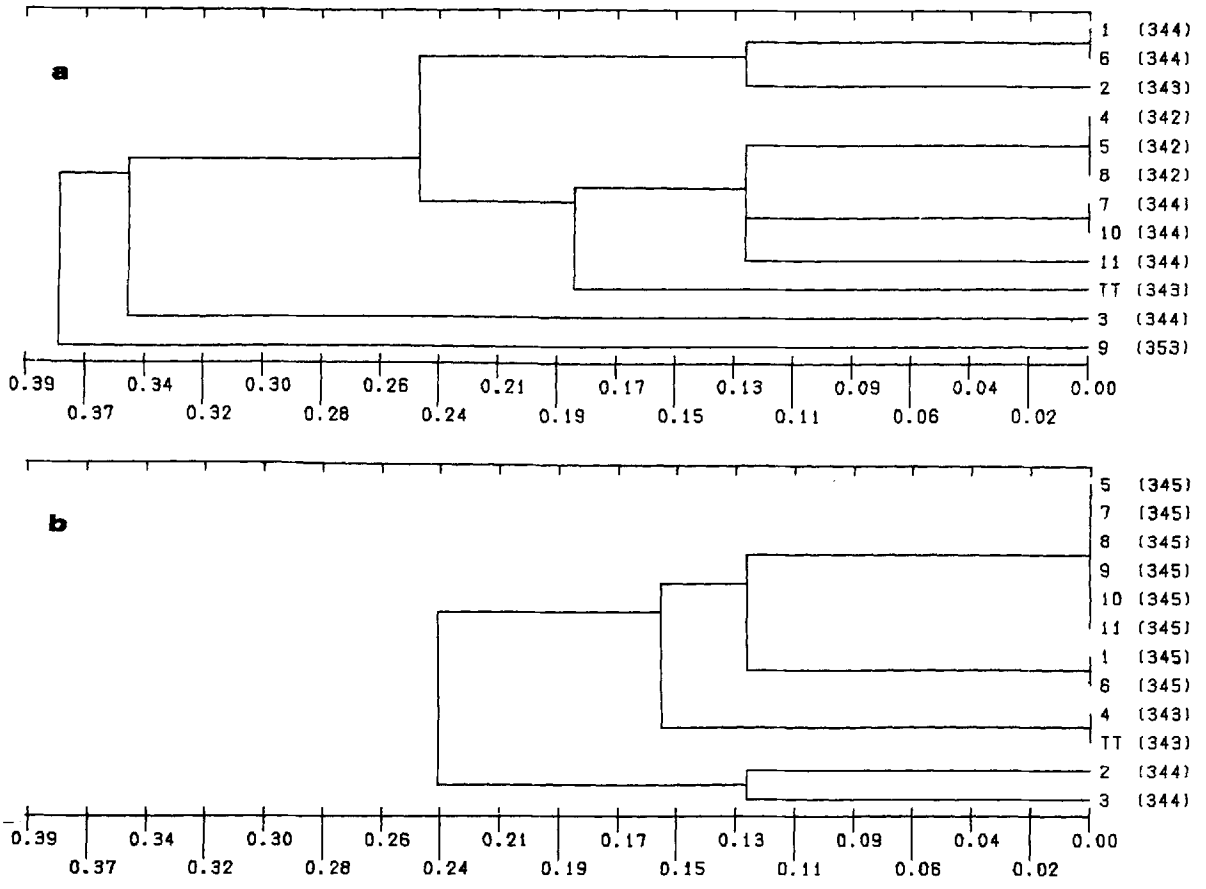
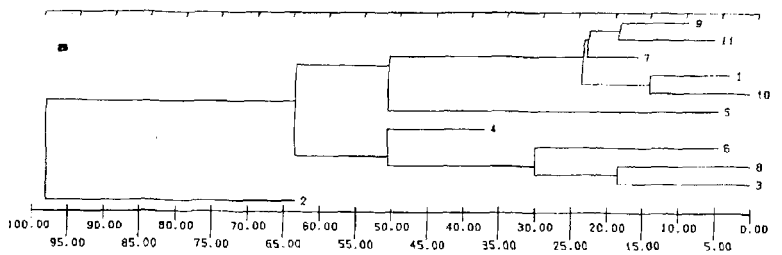


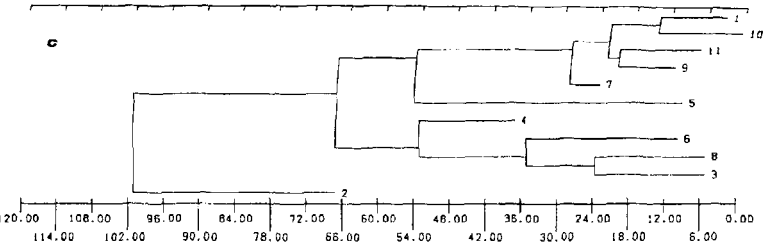
Fig. 2. Dendrograms showing the similarity of the trees constructed on the basis of artificial data (tree 2, 11 species), (**a**) before weighting and (**b**) after weighting. The trees are represented by the species number of the present-day-ancestor used to generate the tree. The true tree is also included in the dendrogram (denoted as TT). The figures in parentheses represent the total number of mutations of the tree. Horizontal axis: dissimilarity level at which the trees are grouped.

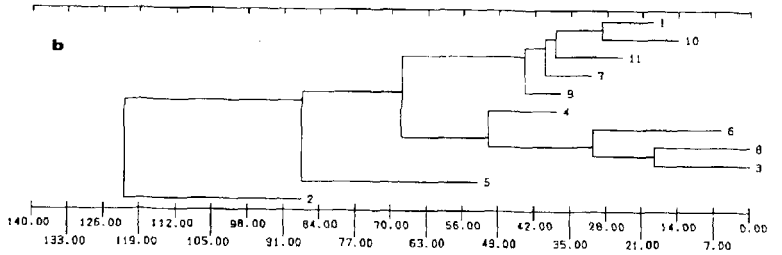
converged tree (and the true tree) in that species 5 is joined to 2 instead of grouped in cluster (1,10,11,7,9), and in that species 9 and 11 do not form a separate cluster before they are joined to species 1 and 10. The grouping of species 2 and 5 occurs in all the minimal and near-minimal trees and in most of the initial present-day-ancestor trees, but not in the true tree and the weighted trees.

We will illustrate some aspects of our method and minimum mutation methods by comparing the minimal tree and the converged tree.

Localness. Besides global information (minimizing mutational cost over the whole tree obviously implies the pursuit of a global goal), minimum mutation methods use mainly local information. Once an internodal sequence is definitely established (in a way that minimizes mutational cost) the previous nodes are (and can be) discarded while new objects are added (see Fig. 4) (see also Cornish-Bowden 1983). This information is used, however, by most similarity methods, because similarity between groups (objects) is calculated as an average of the similarities of all pairs of objects. This point can be demonstrated by comparing the converged tree and the minimal tree with respect to the location of species 5, and relating this to character distribution. See, for example, position 56 of the sequence. When the minimum number of mutations criterion is used, this position does not play any part in the distinction between group (1,10,11,9,7) from group (3,8,6,4), in which groups species 5 could be placed. In both cases it adds exactly one mutation to the minimal length tree. Similarity methods like the present-day-ancestor method can distinguish both groups, for the state at the 56th position of the sequence of species 10 is the same as the state at that position of the sequence of species 5. As a result, the similarity between species 5 and group (1,10,11,9,7) (the group in which 5 should be placed) is greater than between species 5 and group (3,8,6,4) (see further Fig. 3). In placing an object in a minimal tree, one considers only objects in the local neighborhood; other objects are, as it were, masked. This phenomenon and the one that will be discussed in the next paragraph are both involved in the placing of species 2 and 5.







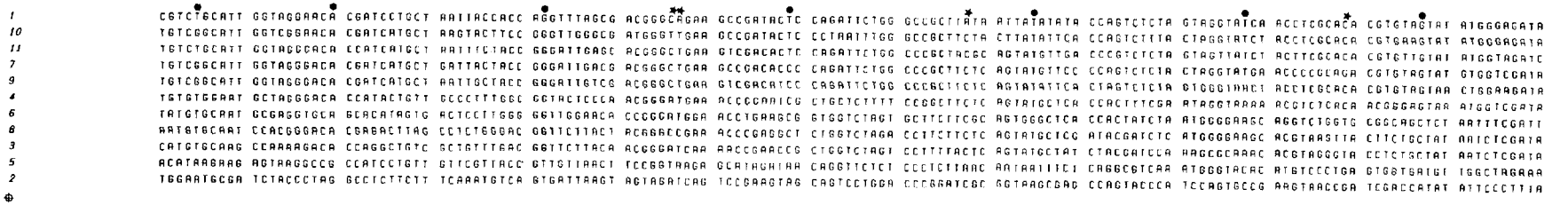


Fig. 3. Results of weighting on simulated tree, tree 2. The tree to which most ancestors converge by weighting (**a**), the minimal tree (**b**), and the true tree (**c**) are shown. The trees are represented with branch lengths corresponding to the number of substitutions between nodes as calculated by the back-tracking algorithm. The minimal tree and the converged tree are rerooted on species 2. The minimal tree is shown together with the sequences. At the positions denoted by a ■, species 2 and 5 share a character state that is not present in other species. On the basis of these positions, species 2 and 5 should be joined together for the purpose of minimizing mutational cost. In our method these positions get a low weight because $MAX_i - MIN_i$ is small. At the positions denoted by a *, species 5 has a character state also present in a species of group (1,10,11,9,7) but not of (3,8,6,4). The state of species 5 in (1,10,11,9,7) is masked by other species with a different character state. Therefore, with the minimal mutation criterion we cannot decide on the basis of these positions whether to place 5 in (1,10,11,9,7) or in (3,8,6,4). With a similarity method, on the other hand, 5 would be placed in (1,10,11,9,7), which is according to the true tree.

Unequal Evolution Time. Minimum mutation methods tend to cluster species that have been mutating for a long time without initiating new lineages (the branch lengths are then relatively long). For most characters these species have unique states, which are not relevant for the minimal length tree. As a result of chance, they will have some characters in common that are not present in other species, so when mutational cost is minimized they will be preferentially joined together. Our method corrects for this undesirable effect, because $MAX_i - MIN_i$ of positions displaying such a character distribution is low. These positions carry relatively little information, so at these positions we relax the parsimony criterion and allow a higher number of mutations. As mentioned, a separate cluster of species 2 and 5 is found in all minimal and near minimal trees. These species have indeed had a long evolution without initiating new lineages. The sequences (see Fig. 3) show that for a number of positions (for example position 20) both species share a character state that is not present in other species, as a result of which $MAX_i - MIN_i$ is low.

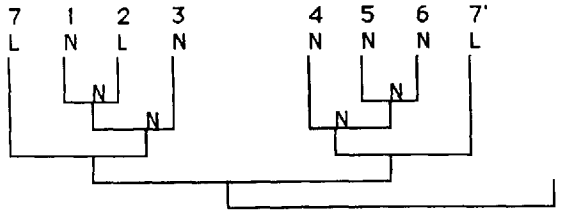


Fig. 4. Tree illustrating localness of minimal mutation methods. Two possible placings of species 7 (7 and 7') are shown. Species 7 has amino acid L at a certain position. Both placings add exactly one mutation to the minimal length tree, although species 2 has the same amino acid as species 7. This information is discarded. A similarity method would join species 7 to group (1,2,3).

Biological Data RBC-SSU

Convergence to Each Other. The use of each of the 16 plant species as present-day-ancestor yielded 16 different initial trees, which differ from each other on average by 20.2 pmu, whereas the 15 trees obtained by weighting differ on average by 18.8 pmu. As with the artificial data, we see a reduction in the number of trees and an increased similarity. Comparison of the dendrogram of trees before and after weighting (Fig. 5) shows that the similarity structure between the trees becomes more pronounced as a result of weighting. The dendrogram of the initial trees has no clearly distinguishable clusters, whereas the dendrogram of the weighted trees shows two distinct main clusters. Thus, surprisingly, the trees do not converge to one pattern, as with the simulated data, but to several patterns. In the model used to generate the artificial data, the evolution of the sequences was not constrained, so the sequences reflect only their evolutionary history. In real evolution, on the other hand, sequences are subject to constraints that give rise not only to a phylogenetic pattern, but also to alternative, functional patterns in the data.

Convergence to Systematics. The pattern reflecting systematic classification is filtered out (trees in cluster 2, see Fig. 5) and the trees are improved, i.e., in better agreement with the classification known from systematics. The grouping of *Rumex, Silene,* and *Spinacia* (all members of the Caryophyllidae), which occurs in all of the trees in cluster 2 (and in the tree constructed with ancestor = *Vigna*), appears in only 2 initial trees, whereas it appears in as many as 10 weighted trees. The genus *Solanum* (species *Solanum* and *Lycopersicon*), occurring in almost all trees in cluster 2, appears in eight weighted trees and in only five initial trees. The trees in both of the smaller clusters of the dendrogram (clusters 1 and 3) do not make sense with regard to systematics. The grouping of the Caryophyllidea occurs only in one tree, and the genus *Solanum* is not grouped (*Solanum* and *Nicotiana* are connected before *Lycopersicon* is joined on). In the trees in cluster 1 even *Lactua* and *Helianthus,* both of which belong to the Compositae, are separated in almost all trees, whereas they show a clear affinity to each other in all the trees in other clusters. The trees in cluster 1 show the grouping of *Brassica* and *Vigna,* whereas in most of the trees in the other clusters *Vigna* shows a similarity to *Vicia* (both species belong to the Papilionaceae). Thus, cluster 1 (and 3) apparently comprise trees displaying an alternative pattern.

Note that the weighted tree constructed by using ancestor = *Lactua* (Fig. 6), which is the best tree with respect to the classification known from systematics, has a small number of mutations (94), but not the smallest number among the weighted trees (and is certainly not the minimum tree, as there are many trees with 92 mutations, none of which is in agreement with systematics).

Hemoglobin A

Convergence to Each Other and Systematics. Before weighting there are nine different trees (obtained by using each of the 11 mammalian species in turn as present-day-ancestor), which differ from each other on average by 6.06 pmu, whereas the eight different trees obtained by weighting differ from each other on average by 5.37 pmu. The similarity pattern has been sharpened up by weighting, though not as clearly as with the RBC-SSU data. The largest

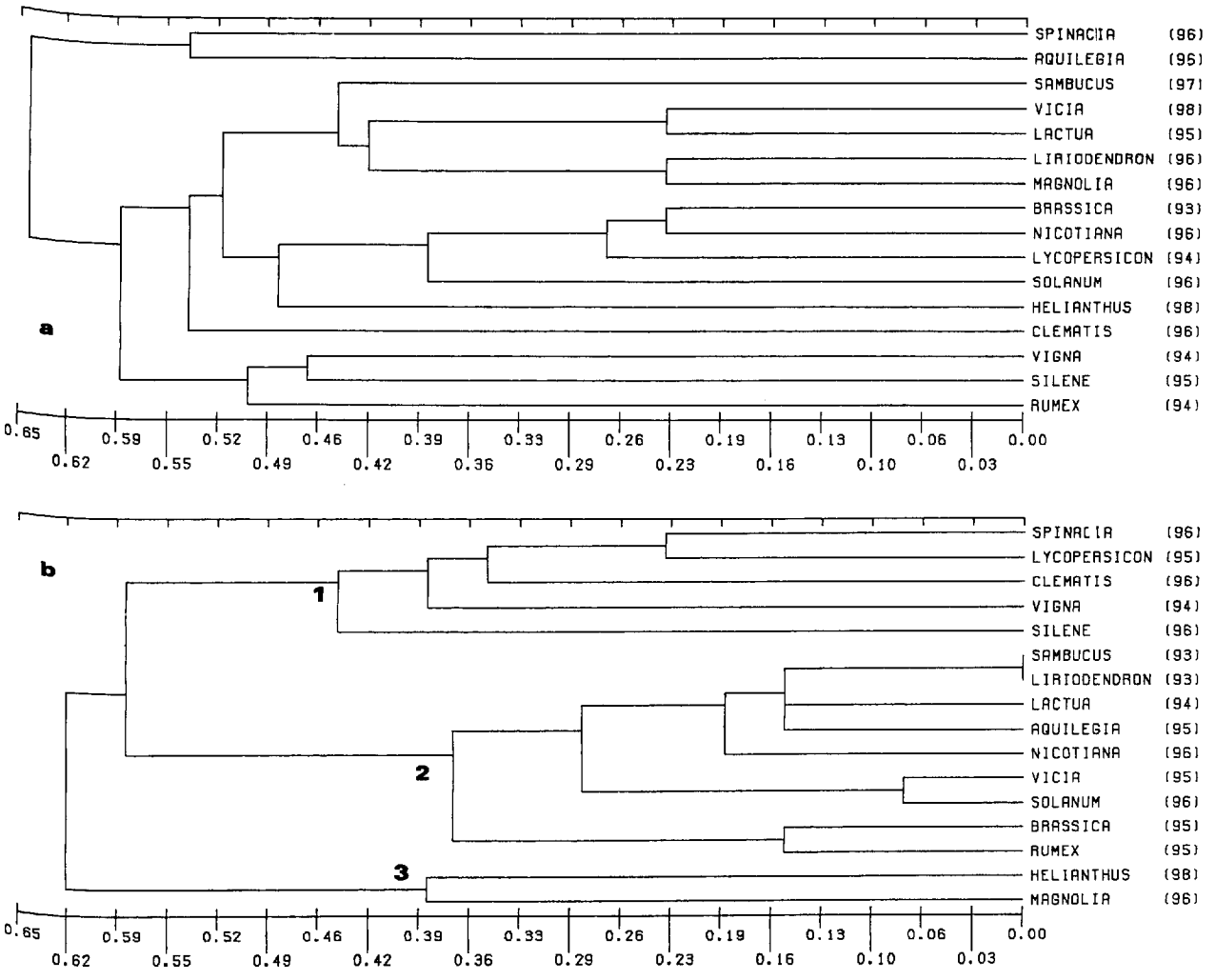


Fig. 5. Dendrograms showing the similarity of the trees constructed on the basis of the RBC-SSU data (**a**) before weighting and (**b**) after weighting. The trees are represented by the species used as ancestor. The figures in parentheses are total number of mutations of the tree. Horizontal axis: dissimilarity level at which the species are grouped. Cluster 2 consists of trees that represent the systematic classification. Both smaller clusters (cluster 1 and cluster 3) consist of trees that are not close to the systematic classification. These trees may show functional patterns rather than phylogenetic relationships.

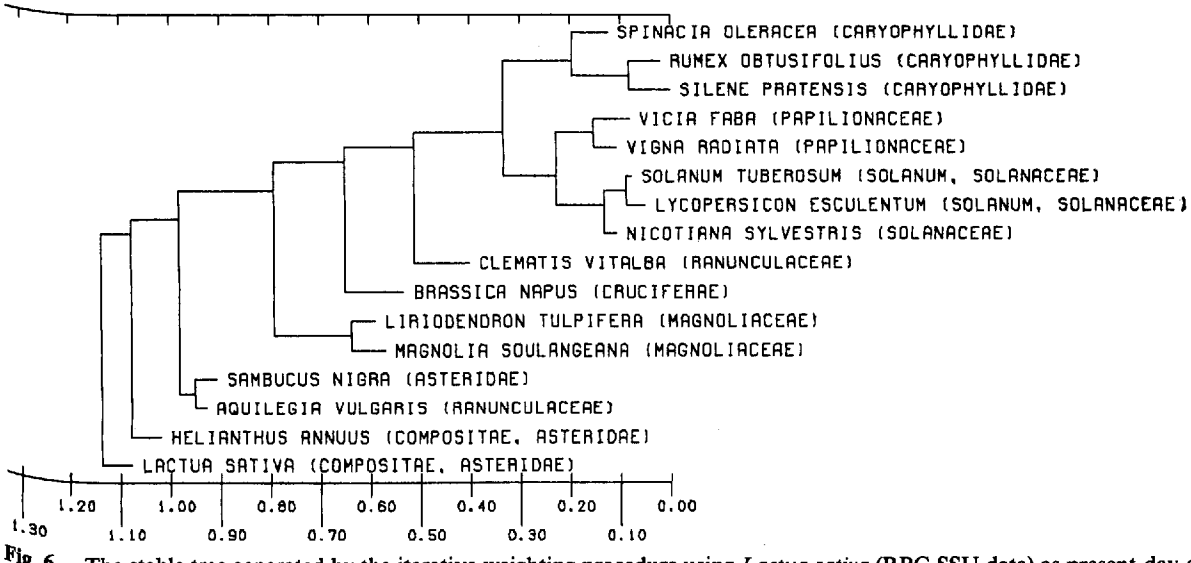


Fig. 6. The stable tree generated by the iterative weighting procedure using *Lactua sativa* (RBC-SSU data) as present-day-ancestor. The systematic classification of the species concerned is shown in parentheses. This tree has 94 mutations.

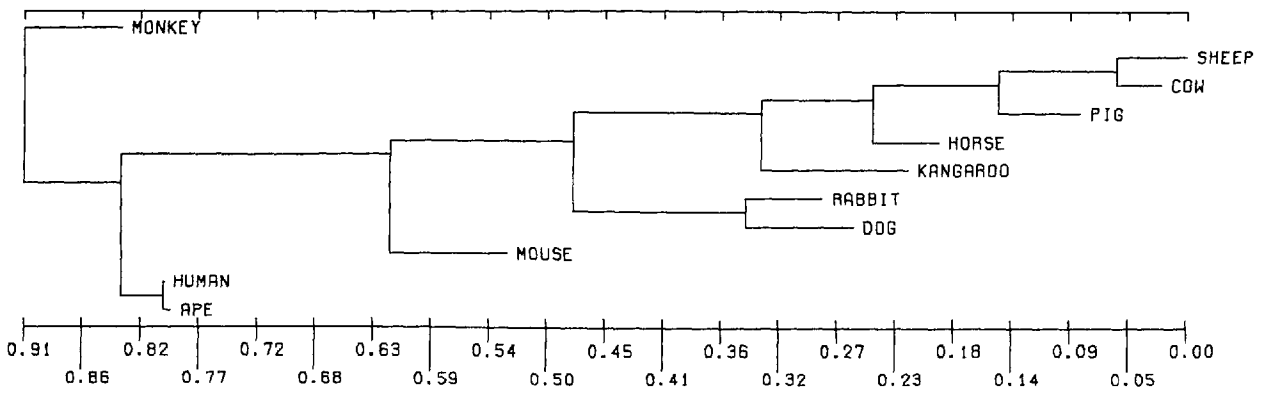


Fig. 7. The stable tree generated by the iterative weighting procedure using monkey (hemoglobin A data) as present-day-ancestor. Horizontal axis: dissimilarity level at which the species are grouped. This tree has 92 mutations.

cluster consists of trees that are in good agreement with what is known about the phylogeny of these species. In these trees (see for example the tree generated by using monkey as ancestor, Fig. 7), the kangaroo is separated from all other species, the Ungulata are joined together in the correct way, i.e., first cow and sheep (Bovidae), then pig, and then horse [i.e., the odd-toed (horse) and even-toed (pig, cow, and sheep) are separated], and the primates form a single group. The position of dog, mouse, and rabbit is less clear. The trees of the other three clusters in the dendrogram of trees can be characterized by the position of the kangaroo: the kangaroo is joined to horse, sheep + cow, and dog, respectively. In conclusion, the trees representing phylogeny are filtered out and some alternatives are generated. The minimal tree (89 mutations, Penny et al. 1982) was found among the unweighted trees, not among the weighted trees, among which a tree of 90 mutations is minimal (this tree was also found by Penny et al. 1982). The tree to which most of the ancestors converge has 92 mutations.

Conclusions and Discussion

In this paper we have proposed a new method for the construction of phyletic trees, which makes use of the fact that mutation rates can differ within sequences. Positions are weighted according to the information content (measured as maximum minus minimum number of mutations possible) and the number of mutations required on a tree. Positions that have a relatively small number of mutations are weighted more highly. In this way, the major clusters of the tree are based mainly on the positions that mutate slowly, and the detailed structure of the tree on the positions that mutate rapidly. One of the advantages of the method is that it involves a combination of minimal length methods and similarity methods. The method allows the strict parsimony criterion to be relaxed. We found that by means of weighting (1) the pattern in the present-day-ancestor trees becomes more pronounced: the trees converge to one or more patterns, whereby patterns conflicting with the phylogenetic pattern are singled out; (2) the true tree is approximated better, especially with respect to the major structure of the tree (artificial data); (3) the trees becomes closer to the systematic classification of the species concerned (biological data); and (4) the number of mutations tends to decrease, sometimes even to the minimum number. [When myoglobin data are used (Table 1 of Penny and Hendy 1986), the minimum tree is among the weighted present-day-ancestor trees; not reported on in this paper.]

The tree-generating method used to test our procedure can generate parts in the true tree that have strong radiation. If in a radiated group high up in the tree a species having a position with a unique character state is duplicated, this position will receive a high weight, whereas without strong radiation this position would get a zero weight. This could result in an incorrect placing of the radiated group low in the tree, were it not for the present-day-ancestor method, which can correct for this phenomenon: the radiated group will be at a great distance from most of the other objects, so it will be correctly placed high up in the tree. Therefore, we prefer to use the present-day-ancestor method if the weighting procedure is used to construct trees.

The weighting formula used in this study is an ad hoc formula, and variations on this general formula are of course conceivable (e.g., nonlinear alternatives). Instead of MAX, one could use in Eq. (2) the mean number of mutations over all possible trees (given the character distribution). Sometimes the maximum value is unlikely to be reached, resulting in too high a weight, relatively. However, we do not have an explicit formula to establish the mean value. Experiments done with the mean (ob-

tained by Monte Carlo methods) instead of the max did not suggest a significantly better result.

Convergence of the present-day-ancestor trees to each other can be attributed partly to the fact that weighting reduces the number of characters. The results are, however, by no means trivial, because, the trees converge to a meaningful pattern: the true tree or a classification closer to systematics. Moreover, we observed not only convergence (to one tree), but also an enhancement of different patterns (i.e., convergence to more than one pattern).

Our future work will focus on the performance of the weighting procedure using a model that explicitly incorporates different mutation rates within and between sequences. This can be done in a natural way by subjecting the sequences to functional constraints, which can include nonlocal constraints associated, for example, with a folding pattern. If gap-generating events (insertions/deletions) are included as well as substitutions, it will be necessary to align the sequences before weighting, because then positions will no longer be defined a priori. The inclusion of gap-generating events should also make it possible to combine our iterative character weighting method with the integrated method of aligning sequences and constructing phyletic trees (Hogeweg and Hesper 1984).

Acknowledgment. We thank Miss S.M. McNab for linguistic advice.

References

Blanken RL, Klotz LC, Hinnebusch AG (1982) Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J Mol Evol 19:9–19

Cornish-Bowden A (1983) Phenetic methods of classification use information discarded by minimal length methods. J Theor Biol 101:317–319

Farris JS (1969) A successive approximations approach to character weighting. Syst Zool 18:374–385

Felsenstein J (1981) A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol J Linn Soc 16:183–196

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20: 406–416

Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59:277–290

Hogeweg P (1976a) Iterative character weighting in numerical taxonomy. Comput Biol Med 6:166–211

Hogeweg P (1976b) Topics in biological pattern analysis. Thesis, University of Utrecht

Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20:175–186

Klotz LC, Blanken RL (1981) A practical method for calculating evolutionary trees from sequence data. J Theor Biol 91:261–272

Klotz LC, Komar N, Blanken RL, Mitchell RM (1979) Calculation of evolutionary trees from sequence data. Proc Natl Acad Sci USA 76:4516

Konings DAM, Hogeweg P, Hesper B (1987) Evolution of the primary and secondary structure of the E1a mRNAs of the adenovirus. Mol Biol Evol 4:300–314

Martin PG, Dowd JM (1984a) The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxylase. III. Addition of Malvaceae and Ranunculaceae to the phylogenetic tree. Aust J Bot 32:283–290

Martin PG, Dowd JM (1984b) The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxylase. IV. Proteaceae and Fagaceae and the rate of evolution of the small subunit. Aust J Bot 32:291–299

Martin PG, Dowd JM (1984c) The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxylase. V. Magnoliaceae, Polygonaceae and the concept of primitiveness. Aust J Bot 32:301–309

Martin PG, Dowd SM, Stone SJL (1983) The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxilase II. The analysis of small subunit data to form phylogenetic trees. Aust J Bot 31:411–419

Penny D, Hendy MD (1985) Testing methods of evolutionary tree construction. Cladistics 1:266–272

Penny D, Hendy MD (1986) Estimating the reliability of evolutionary trees. Mol Biol Evol 3(5):403–417

Penny D, Foulds LR, Hendy MD (1982) Testing the theory of evolution by comparing phylogenetic trees constructed from 5 different protein sequences. Nature 297:197–200

Raup DM, Gould SJ, Schopf TJM, Simberloff S (1973) Stochastic models of phylogeny and the evolution of diversity. J Geol 81:525–542

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Smeekens S, van Oosten J, de Groot M, Weisbeek P (1986) Silene cDNA clones for a divergent cholorophyll-a/b-binding protein and a small subunit of ribulosebiphosphate carboxylase. Plant Mol Biol 7:433–440

Sneath HA, Sokal RR (1973) Numerical taxonomy. WH Freeman, San Francisco

van Ooyen A (1987) Het gebruik van moleculaire sequentie data in de plantensystematiek, toegelicht aan de hand van chloroplast proteine data van een zestiental angiospermen. Bioinformatica, internal report

Watanabe S (1969) Knowing and guessing, a quantitative study of inference and information. Wiley, New York

Received October 16, 1989/Revised May 1, 1990/Accepted May 10, 1990

Appendix

The tree cohesion measure (TC) is not symmetrical in T1 and T2, the trees for which the similarity is calculated. TC(T1, T2) and TC(T2, T1) for two strictly binary trees spanning n objects are defined as

$$TC(T1, T2) = 1/(n-3) \sum_{i=1}^{n-3} \max(c_{ij} \mid 1 \le j \le n-3)$$

$$TC(T2, T1) = 1/(n-3) \sum_{j=1}^{n-3} \max(c_{ij} \mid 1 \le i \le n-3)$$

where $n - 3$ is the number of partitions of each tree, and c_{ij} is the similarity between the ith partition of T1 and the jth partition of T2 as expressed Watanabe's (1969) C. C is defined as

$$C = \alpha \log \frac{\alpha}{(\beta + \alpha)(\beta + \delta)} + \delta \log \frac{\delta}{(\gamma + \delta)(\gamma + \alpha)}$$

$$+ \beta \log \frac{\beta}{(\alpha + \beta)(\alpha + \gamma)} + \gamma \log \frac{\gamma}{(\delta + \beta)(\delta + \gamma)}$$

where

$$\alpha = \frac{n_{11}(x_k, x_l)}{n}, \beta = \frac{n_{10}(x_k, x_l)}{n}, \gamma = \frac{n_{01}(x_k, x_l)}{n}, \delta = \frac{n_{00}(x_k, x_l)}{n}$$

where $n_{11}(x_k, x_l)$, for example, is the number of 1s that coincide on the two rows x_k and x_l, which here are partitions denoted as a sequence of 1s and 0s (of length n, total number of objects). The 1s and 0s indicate whether an object is a member of the first or of the second subset. The subsets are formed by partitioning the whole set of n objects. First or second subset depends on the orientation of the tree and is irrelevant for the similarity measure. The tree cohesion measure is mainly intended as a descriptive measure, and statistical or mathematical properties have not been explored thoroughly.