J Mol Evol (1984) 20:175–186

Journal of
Molecular Evolution

The Alignment of Sets of Sequences and the Construction of Phyletic Trees: An Integrated Method

P. Hogeweg and B. Hesper

Bioinformatica, Padualaan 8 Utrecht, The Netherlands

Summary. In this paper we argue that the alignment of sets of sequences and the construction of phyletic trees cannot be treated separately. The concept of 'good alignment' is meaningless without reference to a phyletic tree, and the construction of phyletic trees presupposes alignment of the sequences.

We propose an integrated method that generates both an alignment of a set of sequences and a phyletic tree. In this method a putative tree is used to align the sequences and the alignment obtained is used to adjust the tree; this process is iterated. As a demonstration we apply the method to the analysis of the evolution of 5S rRNA sequences in prokaryotes.

Key words: Sequence alignment — Phyletic trees — Matrix methods — Internode sequences — Homology assessment — Tree representation — Prokaryotic 5S rRNA evolution

1. Introduction

There is a bootstrapping problem in all non-supervised pattern recognition procedures (e.g., cluster analysis): to assess the similarity (relationship) of objects, one has to assess the correspondence and similarity of parts of the objects (characters). Most procedures simply assume that the assessment of the characters has been done prior to the analysis; they start with a set of characters and character weights (including scaling of characters). It appears, however, that in, for example, taxonomic practice character selection is done implicitly on the basis of a prior assessment of the similarity of the objects (Leenhouts 1968; Hogeweg 1976). Therefore, it seems useful to include this mutual dependence explicitly in non-supervised pattern recognition procedures. To this end, we introduced iterative character weighting into numerical taxonomy (Hogeweg 1976). Although the correspondence of characters (i.e., homology) is not assessed by this method, the relative weighting of the characters is adjusted on the basis of a classification generated using previous, initially equal character weights. Thus, noisy (and inconsistent) characters are filtered out and the classification is sharpened up and allows oligothetic characterisation of clusters (Hogeweg and Hesper 1981). This method was introduced to improve phenetic classifications, but can be seen as bridging the gap between phenetic and phyletic (character weighting) methods (McNeill 1978).

Along the same lines, we introduce in this paper an iterative method for the construction of phyletic trees that are based on molecular sequence data. In the case of molecular sequences, assessment of the homology of characters takes the form of the alignment of sequences. The method presented here goes an important step further than the previous one in that the homology of characters (here, alignment of sequences) does not have to be assessed prior to the analysis. Starting with a set of unaligned sequences, it generates an alignment of the sequences as well as a phyletic tree; like the previous method it does so in an iterative way.

2. The Alignment of Molecular Sequences

In the course of evolution not only base changes, but also insertions and deletions of bases (sub-sequences, amino acids), occur. Therefore, sets of se-

Offprint requests to: P. Hogeweg

quences to be compared in phyletic studies have to be 'aligned,' i.e., correspondence of positions has to be established.

Satisfactory algorithms exist for the alignment of *pairs* of sequences (Needleman and Wunsch 1970; Sellars 1974; Waterman et al. 1976; Smith et al. 1981). These algorithms use some measure of maximal matching (or minimal difference) as an optimisation criterion, and insert gaps into the sequences to optimise for that criterion. The set of choices for the assessment of maximal matching is identical to that used for the assessment of the similarity of aligned sequences in phyletic tree construction. In addition a choice has to be made regarding the incidence of deletions/insertions: The algorithms use a 'penalty' or 'bias' factor to inhibit too frequent gap formation. Using the fact that insertion/deletion events involve sub-sequences rather than single bases, Needleman and Wunsch (1970) originally used a penalty for such events irrespective of the length of the sub-sequence involved. Contrarily, Sellars (1974) originally used only the length of the inserted sequences in the penalthy; longer sequences do after all change the molecule more (as assessed in most similarity calculations). The distinction between the two approaches is related to the question: Is mutation or selection the limiting factor in sequence divergence? Smith et al. (1981) showed that the two algorithms are equivalent when generalised so as to use both a penalty for the insertion/deletion event and a penalty based on length, i.e.,

$$\text{bias} = X,Y$$

where X is the penalty for the insertion/deletion event and Y the additional penalty for each gap position. X and Y can be 0; when $Y = 0$ the shorthand notation bias = X is used. The choice of X and Y is largely a matter of taste.

A problem in using the pairwise alignment algorithms is that they produce many different alignments with the same optimality value. For example, the localisation of gaps is not determined uniquely and can vary over at least the stretch where no matches occur (for all or none matchings).

However, most algorithms for the construction of phyletic trees based on molecular sequences use as primary data *sets* of aligned sequences, and there is no practical method for obtaining such alignments (the pairwise method is in principle extendible, but for more than two sequences the required computer time/space becomes excessive). Thus, in practice the sets of sequences are 'mind-aligned,' using algorithmic pairwise alignments only as guidelines, because pairwise alignments give incompatible results. Secondary structure can also be considered in constructing such alignments. This practice can, however, reveal convergence in function rather than common ancestry.

The alignment of sets of sequences is crucial for the results of phyletic studies based on such alignments, even more so than the similarity measure or tree construction algorithm used (see Fitch and Yasunobu 1974). Therefore, alternative alignments and their consequences for the similarity relations should be considered in the study of the evolution of a set of sequences. For this, one needs an automatic method for generating 'interesting' alignments in a flexible way using explicit criteria. 'Mind-mediated' methods are not suited for this, because minds tend to get stuck in a previously recognised pattern and never confine themselves to a set of explicit criteria, but use implicit information, which renders impossible an evaluation relative to internal and external criteria.

3. Phyletic Tree Construction Algorithms

3.1. Tree Construction

Many different algorithms have been proposed for the construction of phyletic trees. We shall mention only some points that are important for our integrated method of tree construction and alignment of sets of sequences.

Most algorithms try in some sense to minimise mutational cost along the branches of a tree. Given a criterion for minimal cost, a tree topology can be evaluated relative to this criterion. Thus, one approach is exhaustive search: generate all possible tree topologies and choose the best one. The number of alternative topologies is too large to make this method practical (for more than eight compared sequences). Thus, one must resort to non-optimal tree-generating heuristics. These methods fall into two classes: matrix methods and character-based methods. The former use as primary data a triangular matrix of pairwise overall similarities (generally, but not necessarily, computed from an item-versus-character matrix), whereas the latter work with the item-versus-character matrix and consider trees of character (in case of sequences, nucleotides or amino acids) mutations and their compatibility, and sometimes generate 'internode' (ancestral) sequences. Matrix methods are the simpler of the two and are therefore used most often in practical applications, even by authors who themselves have proposed character-based methods (e.g., Dayhoff 1976). In fact, matrix methods do not need a global alignment of the set of sequences, but only pairwise alignments, because pairwise similarities are used. Character-based methods do need global alignment, since this is necessary for the construction of the

character-based trees. As mentioned above, no algorithms exist for the construction of the global alignment of sets of sequences. For this reason the method we propose for obtaining such an alignment uses matrix methods for the construction of trees. However, consistent behaviour of characters in their mutational history is used in evaluating the tree and in constructing the alignment; the consistency of a subset of the characters can be optimised by a character-weighting scheme.

3.2. Representation of Trees

Phyletic trees and phenetic trees (dendrograms) represent different things: The former represent hypothesised ancestral relations and the latter similarity between groups of objects. Nevertheless, the same algorithms, or closely related ones, can be used to construct either of them. It is therefore annoying that the representational conventions have exaggerated the gap between these types of results. We shall use a representation that can serve both purposes, a 'cumulogram'. Cumulograms use the convention of parallel branches and horizontally extended nodes, as in dendrograms (thus minimising the unwarranted suggestions about divergence that plague attempts to draw more 'lifelike' trees), but represent the similarity between groups of objects cumulatively from node to node instead of on one scale measured from the top of the dendrogram. In this way nodes can be interpreted as ancestral objects if and only if one is so inclined, and differential rates of divergence (evolution) can be represented (see Figs. 1, 2, 4).

4. An Integrated Method for the Alignment of Sets of Sequences and the Construction of Phyletic Trees

4.1. Outline of the Method

The following considerations led us to formulate an integrated method for the alignment of sequences and the construction of phyletic trees:

1. The concept of 'good alignment' seems to be meaningless without reference to ancestry: an alignment hypothesises evolutionary events (insertions, deletions) that happened at some point of divergence.

2. The criteria for alignment and tree construction are closely related, both being based in some way on maximal matching. Alternative schemes for measuring maximal matching apply in both contexts.

3. Pairwise alignments are feasible (Section 2).

4. Pairwise alignments suffice for tree construction when matrix methods are used (Section 3).

5. Phyletic trees produced by matrix methods are *binary* trees.

6. Alignments of similar sequences are the most reliable.

7. Internode sequences can be constructed that represent the ancestral sequences of the branches that they join (Section 3).

Thus, it seems good heuristically to reduce the problem of aligning sets of sequences to a series of successive pairwise alignments that follow the branches of a hypothesised (phyletic) tree [this was previously suggested by Fitch and Yasunobu (1974)] and to optimise the tree and the alignment iteratively by using the previous tree (alignment) to construct the next alignment (tree).

Such a heuristic method takes the following form:

1. All pairs of sequences are aligned independently; the match value obtained in the first step of the alignment procedure is used (possibly after a transformation, e.g., to correct for backward mutations) as the similarity value. In this way an $N \times N$ similarity matrix is constructed.

2. This similarity matrix is used for the construction of the first phyletic tree (which is then, however, based on an incompatible alignment of the sequences). Any matrix method can be used in this step, although it is advisable not to use a method that enters very dissimilar sequences 'high in the tree' (i.e. near the leaves), which would imply a very rapid change in that particular sequence after a recent splitting off.

3. The sequences are successively pairwise-aligned, following the branches of the tree, and internode sequences are constructed. First the two sequences are taken that are most similar in the tree, since their alignment is most reliable. An internode sequence is constructed as follows: (a) if two positions of the aligned sequences contain the same base (amino acid) it is placed in the new sequence; or (b) if not, the decision about which base (amino acid) to enter is postponed. If a sequence is joined to an internode, it is aligned to that internode. The problem of aligning sets of sequences is reduced to that of aligning pairs of sequences by assuming that the relative alignments of the sequences that diverged later in evolution are fixed, since they evolved from the ancestral sequence. At 'open' positions of an internode sequence (i.e. positions for which no decision was made) the similarity calculation used in the alignment algorithm backtracks to the higher sequences (internodes) and uses the highest similarity value available. From the thus aligned sequences/internodes a new internode is constructed in the same way as before. At the open positions the algorithm backtracks to the higher nodes; if a corresponding base (amino acid) is found, it is placed into the new sequence and intermediate internode

sequences: apparently (on the hypothesis of minimal mutations) this base (amino acid) was the one common to the sequences until they diverged in various directions. If there are several common bases (amino acids), as can happen when two 'undecided' internodes are joined, the decision is again postponed until there is a best choice. This local backtracking method of internode determination gives a minimal mutation solution if a binary similarity criterion is used. Otherwise it gives a solution close to the minimal solution for nucleotides, because only limited backtracking is possible, there being only four bases and a gap. In the case of amino acids, only binary similarity can be used. Other internode-generating algorithms can be substituted, however.

4. Thus, a consistent alignment of the entire set of sequences (consistent in the sense that one and the same set of insertions/deletions is used for each pairwise comparison of sequences) is obtained by making N pairwise alignments and by constructing internode sequences.

5. The number of mutations along the branches is calculated and the tree is plotted with these branch lengths (the topology is not changed).

6. The mutational frequency at each position of the aligned sequences is calculated. These frequencies can be used as character weights, e.g.,:

$$W_i = N_i/(M_i + 1)$$

where W_i is the weight at the i-th position, N_i the number of different nucleotides (amino acids) occurring at that position and M_i the number of mutations (along the branches of the tree) occurring at that position.

7. The aligned sequences are used in the next round of the iterative procedure. A similarity matrix is calculated from them (possibly using character weights); this similarity matrix is based on a consistent alignment, whereas the first one was not.

8. The process is repeated from step 2 until convergence or until some stopping rule is satisfied.

Thus a series of phyletic tree estimates and corresponding alignments is produced. Each of them can be evaluated on the basis of various internal and external criteria. Internal criteria could include, for example, the number of mutations and the profuseness of gap formation. External criteria could include phyletic trees based on other data about the same species and secondary structures of the sequences under consideration (including the generated internode sequences).

4.2. Annotations to the Method

The scheme outlined above involves the same problems as those inherent in pairwise alignments of sequences and matrix methods of phyletic tree construction. Accepting these difficulties, we have yet another problem. As mentioned, the pairwise alignment algorithms do not produce a unique solution. Several alignments are equivalent in the sense that the same number of gaps is formed and the same overall similarity is obtained. Such alternative solutions are not equivalent when more nodes are entered. Nevertheless, the proposed algorithm requires that a choice be made before the next node is considered. This type of difficulty always occurs when local strategies are used to attain a global goal (as is the case for most phyletic tree construction algorithms). This does not hold for locally equivalent solutions only; the globally optimal solution may require locally suboptimal solutions. This fact has often been used to criticise such local methods. However, the opposite argument can also be put forward: global optimisation has as a drawback that locally non-optimal solutions are generated, whereas the data warrant more detail locally than globally (Hogeweg 1976b). Nevertheless, in the present case it seems necessary to allow limited a posteriori correction of the arbitrary choice of the positions of the gaps formed in the pairwise alignment. This is because the alignment algorithms treat two different evolutionary processes in the same way. Insertions of bases (amino acids) are incorporated as insertions of gaps into the opposite sequence. One supposed insertion should not lead to several gaps inserted in different positions; this could happen if a correction were not made. Therefore, when a gap is formed in aligning two sequences it is swapped so as to match an open gap (i.e. a gap accessible by the backtracking algorithm) if such a gap occurs in the region of arbitrary choice. Note, moreover, that the alignment algorithms have to be adjusted also to allow for the existence of gaps in the sequences to be aligned. When a gap is formed opposite an existing gap the penalty based on the length of the gap should be zero (i.e. bias = X, 0).

The method is likely to 'postpone' the inclusion of differences inherent in the data set since it maximises the similarity near the leaves of the tree first, both with respect to the alignment and with respect to overall similarity (in most tree construction algorithms). Thus the branch lengths obtained will be biased towards short branches near the 'leaves' and long branches near the 'root'; this structure hypothesises relatively recent divergence of similar species and relatively early divergence of groups of dissimilar species. From an evolutionary point of view this bias is reasonable if we assume that early diverged lineages might have become extinct, i.e. that the relatively large internode distances near the root of the tree include 'missing nodes' of extinct lineages. Whether this assumption can cover quantitatively the bias induced by the algorithm should be studied

by simulation experiments. It is interesting to note that this bias conforms to the optimisation of a classification: classes are maximally separated (Hogeweg 1976, Hogeweg and Hesper 1981). In our case this optimisation is a side effect of a method for minimising mutational cost.

4.3 Implementation of the Method

The method outlined above has been implemented as part of BIOPAT, a program system for bioinformatic pattern analysis (Hogeweg and Hesper 1972) in which pattern recognition methods can be integrated to provide a maximal heuristic benefit. The method was implemented by adding just one program (for step 3 of the iteration) to the system. Thus, in fact, it is the outline of the method that has been implemented rather than one particular specification of it, since it can be used in combination with a wide variety of methods for tree construction, representation and evaluation. So far we have used mainly the following set of criteria in our research:

1. A binary similarity criterion of bases/amino acids was used [i.e., the same (1) or not the same (0)] for the alignment and the calculation of the similarity matrix for tree construction. Thus we included a minimum of assumptions.

2. Maximum match value, as obtained in the alignment algorithm (i.e. including the penalty used there for gap formation), was used without transformation in the similarity matrix. In this way the alignment algorithm can be considerably speeded up, since only the first step is needed and the backtracking to construct the actual alignment can be omitted. Therefore, no match matrix has to be constructed and no search is necessary when we work in a few arrays containing maximum column and cumulative row values.

3. A relatively high gap penalty (=5) was used in the alignments for the calculation of the first similarity matrix; thus fewer incompatibilities arise and the most similar sequences found contain few gaps, both of which factors are beneficial to the method.

4. Trees were constructed mainly by two different methods:

(a) Group averaging (=UPGMA of Sneath and Sokal 1973); this method assumes equal mutation rates; it has been hailed as a very good criterion for phyletic tree construction (Tateno et al. 1982; Nei et al. 1983), although it was originally introduced as a criterion for phenetic classification. Because the similarity between clusters is calculated as the average similarity of all object pairs, it uses more information (see also Cornish-Bowden 1983) and is therefore less sensitive to 'aberrant' sequences than are methods based on the comparison of 'most recent ancestors' (e.g., median averages, the modified Fitch criterion and methods based on common ancestor construction). In the iteration we used most recent ancestors (to a variable depth, dependent on the variability of the base under consideration). The combination of the alignment algorithm and tree construction based on group averages generates trees that represent compromises between the advantages and disadvantages of these approaches: a limited amount of unequal mutation rate is allowed in a tree in which the overall similarity of evolutionary lines is optimised.

(b) Present-day ancestor method (Klotz et al. 1979, 1981; Blanken et al. 1982): This method assumes that there is a tree topology in the data and, if this assumption is warranted, it finds the correct tree toplogy, notwithstanding variable mutation rates. To this end the pairwise distances are converted to correct for their shared distance from a hypothesised common ancestor. Any node can be chosen as a common ancestor. If an arbitrary node is chosen the result is an unrooted tree. By choosing either a very distant node or a node known to be of separate lineage, one obtains a rooting that is acceptable for our method, which desires similar sequences to be joined first. The method is sensitive to the peculiarities of the chosen ancestor because the global data transformation is relative to the ancestor. In our iterative method this sensitivity is decreased because of the local optimisation of the alignment, which is dependent on the generated tree but not on the structure of the common ancestor.

5. Character weighting was not used in the experiments reported here.

5. Behaviour of the Method: 5S rRNA Sequences of Prokaryotes

5.1. Introduction

Sankoff et al. (1982) studied the phylogenetic relationships among 5S rRNA sequences of 19 prokaryotes. They argued that for real evolutionary problems one need not resort to non-optimal search algorithms, because usually enough prior information is available to reduce the problem to one of finding the best tree composed of at most eight previously established subtrees; the minimal mutation tree can then be found by exhaustive search. Accordingly, they used in their study four preassigned subtrees (Bacilli, including *Lactobacillus* and *Clostridium*; the Enterobacteria; chloroplasts and *Anacystis*; and *Halobacter* and *Thermus aquaticus*) and

three 'loose' species (*Mycobacter, Rhodospirillum* and *Streptomyces*) in an exhaustive search. All possible trees were evaluated with respect to minimal mutational cost (using values of 0.45 for transitions, 0.77 for transversion and 1 for base/gap differences). The results for the eight best trees differed by less than three mutations. Sankoff et al. also used their evaluation criterion to compare their results with previously proposed trees, using their own alignment and choice of sequences.

For these studies Sankoff et al. used an alignment of the sequences that was justified as a reconciliation of previously published alignments and based on secondary-structure considerations. Thus, in trying to avoid non-optimal methods of tree construction they nevertheless relied on a not entirely known method for alignment, although the results depend just as crucially on the alignment as on the tree construction algorithm.

We used their set of sequences to study the behaviour of the integrated method of tree construction and alignment proposed here.

5.2. Calculation of an Initial Pairwise Distance Matrix

For the calculation of the initial pairwise distance matrix we used a relatively high penalty for insertion/deletion events and no penalty for the length of the gaps, i.e., bias = 5,0 (in Figs. 1, 2, 4 indicated as bias = 5). This was done to ensure that sequences considered first in the alignment (i.e. the most similar sequences) were easy to align (i.e. contain few gaps; see Section 4.3).

5.3. Tree Construction by Group Averaging

The initial tree (Fig. 1a) generated by group averaging from the pairwise distances has little merit with respect to the number of mutations; it gives 437 mutations, compared with the 423 computed by our algorithm for the optimal tree of Sankoff et al. (1982) using their alignment. Treewise alignment was done using a penalty for gap formation of the form bias = X,Y (X penalty for the event, Y penalty for length); Y = 1 was chosen in all cases because the length of the insertion or deletion shows up as that many mutations; X was varied in the reported experiments between 0.5 and 3.0.

The results for bias = 1,1 are shown in Fig. 1. The first iteration step immediately gives a large improvement: the number of mutations reduces to 424 and the tree topology changes so that the relatively dissimilar sequences of *Clostridium, Streptomyces* and *Rhodospirillum,* which did not show clear affinities in the initial tree, are joined to the same groups as the ones used or generated by Sankoff et al. However, the tree topology differs from theirs; in particular, *Mycobacter,* a very aberrant species wherever it is placed, clusters with *Thermus aquaticus* and is close to *Halobacter,* an Archeaebacterium. Another difference is that the Bacilli and Enterobacteria are relatively close, whereas they were remote in the tree of Sankoff et al. However, in view of the number of mutations, this generated tree cannot be discarded. Continued iteration stabilises when the tree shown in Fig. 1c is reached, which has 427 mutations and is therefore somewhat less optimal than the previous one. The difference is in the relative position of *Streptomyces* and *Clostridium*: although in the alignment generated on the previous tree, the former is more similar to the Bacilli s.s. than the latter is, its placement according to this similarity generates a less optimal alignment (because of local peculiarities of the sequences).

For bias = 2,1 the same stable tree is obtained; the numbers of mutations are 441 and 436 for the initial and stable tree, respectively, but the alignment has the merit of having formed fewer gaps: the table of sequences has length 127, as compared to 132 for both the bias = 1,1 alignment and the alignment of Sankoff et al., and resembles the alignment proposed by Hori and Osawa (1979). This tree is also stable for bias = 0.5,1 (number of mutations, 416; expansion to length 142), but is not obtained starting from the initial tree, because the liberal insertion of gaps moulds the generated alignment to this tree [at the expense of much expansion (length = 143); there are 431 mutations]. Therefore the tree changes only slightly during the iteration and converges to a tree with 429 mutations, many more than for the stable tree obtained with a larger event penalty.

For bias = 3,1 no decrease in expansion is obtained relative to bias = 2,1, but there are more mutations: gap insertion is postponed for too long.

5.4. Tree Generation by the Present-Day Ancestor Method

5.4.1. Ancestor: Halobacterium cutirubrum. If the only Archeaebacterium is chosen as the ancestor, the 'correct' rooting should be obtained. The generated tree conforms to the tree of Sankoff et al. with respect to the relatedness of the Enterobacteria with the chloroplast group, and to the stable tree generated by group averaging with respect to the position of *Mycobacter. Rhodospirillum* and *Streptomyces* are intermediate between the chloroplast/Enterobacteria group and the Bacilli. Iterating (while aligning with bias = 1,1) gives the tree shown in Fig. 2: *Anacystis* and *Rhodospirillum* are on the Enterobacteria branch and the chloroplasts are seen as being

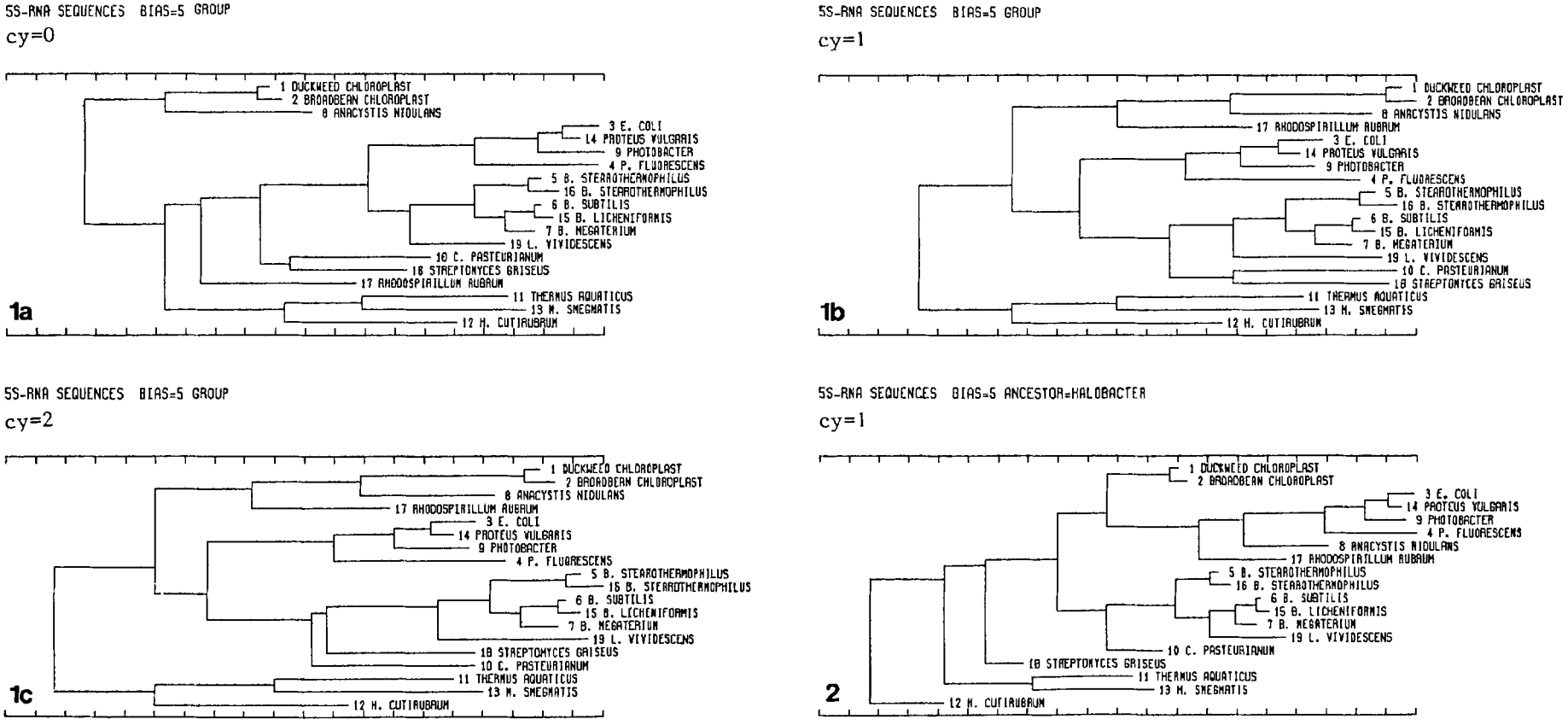


Fig. 1a-c. Successive trees generated by group averaging. **a** Cycle (cy) = 0; pairwise alignment only (bias = 5,0). Number of mutations = 437; length = 132. **b** cy = 1; treewise alignment after one iteration (bias = 1,1). Number of mutations = 424; length = 132. **c** cy = 2; treewise alignment (bias = 1,1); stable tree. Number of mutations = 427; length = 132

Fig. 2. Optimal tree generated by present-day ancestor method (ancestor = *Halobacter*). Treewise alignment with bias = 1,1 after one iteration. Number of mutations = 423; length = 132

relatively primitive. The tree has 423 mutations and a length expansion to 132, i.e. it is again of a similar optimality. However, it is not stable: *Streptomyces* has slipped down to near the root of the tree at the expense of more mutations and a larger expansion.

5.4.2. Ancestor: Mycobacterium smegmatis. *Mycobacterium smegmatis* is clearly a very dissimilar member of the data set. We experimented with it as present-day ancestor because its position varies in the various tree proposals discussed here and in the literature (e.g., Schwartz and Dayhoff 1978; Fox et al. 1980; Kuntzel et al. 1981). We again used bias = 1,1 as the penalty for gap formation. The iteration resulted in a long sequence of different trees which finally stabilised after 11 steps. Again it was the second to last iteration which was best with respect to the number of mutations and the gap expansion, with 415 mutations and length = 132. The iteration stabilised at 418 mutations and length = 133. Surprisingly, the number of mutations did not change monotonically (see Fig. 3), and showed a second optimum after six iteration steps (number of mutations, 416, length = 131). The initial tree, the two optimal trees, the worst tree and the stable tree are shown in Fig. 4. The initial tree and the optimal tree are similar in overall topology, but differ with respect to the detailed structure of some of the groups: the optimal tree 'corrects mistakes' that arose due to peculiarities of the present-day ancestor (see Section 3). The Bacilli and the chloroplast group cluster together, whereas the Enterobacteria, being relatively similar to *Mycobacterium,* are nearer the root of the tree. *Halobacter* and *Thermus aquaticus* are also near the root of the tree. The final tree and the intermediate optimal tree differ from these trees in that they show a close relationship among the Bacilli s.s. (without *Clostridium, Streptomyces* and the Enterobacteria).

As an example of the generated alignments, the one based on the most optimal tree is shown in Fig. 5. We note the following points. (1) The gaps at position 34 and at positions 74/75 are present in all alignments. Even if *Mycobacter* and *Halobacter* are not in each other's ancestry neighbourhood (i.e., the alignment algorithm cannot backtrack to the other species), they have an insertion at an identical position (here 74/75). In fact, it is unlikely for two independent events to occur at homologous positions, and therefore a close relationship between these species seems likely. (2) Unlike our algorithm, many published alignments favor multiple insertions at identical positions in distantly related species. This results in tidy-looking alignments, but does not seem to be defendable in terms of evolutionary processes. (3) Many insertions/deletions occur in stacking regions. Corresponding hairpins do not always occur at fully homologous positions [e.g., the GC-rich second hairpin around positions 102–114 (cf. Fox and Woese, 1975)].

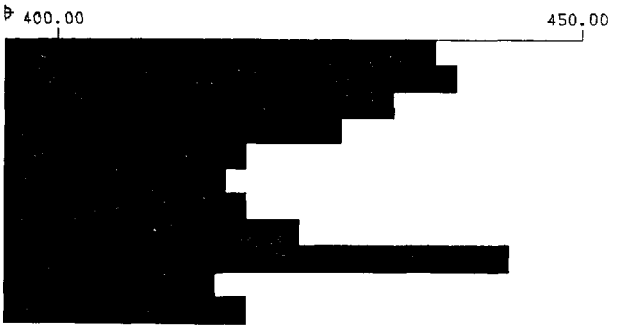


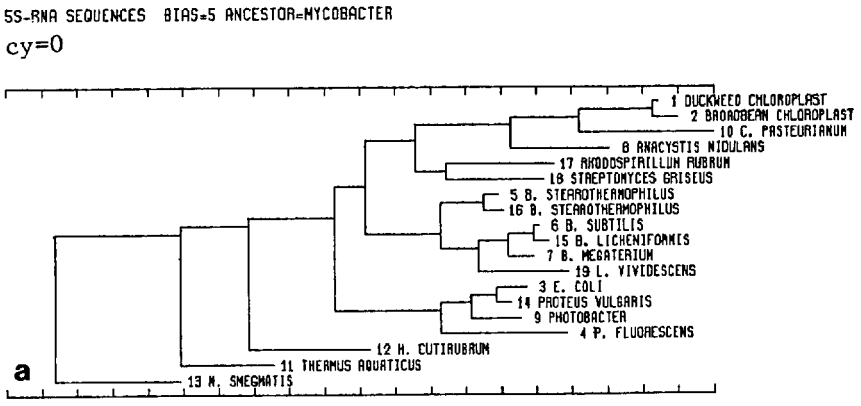
Fig. 3. Number of mutations for the successive trees generated by the present-day ancestor method (ancestor = *Mycobacter*); cycle number increases from top to bottom (see also Fig. 4)

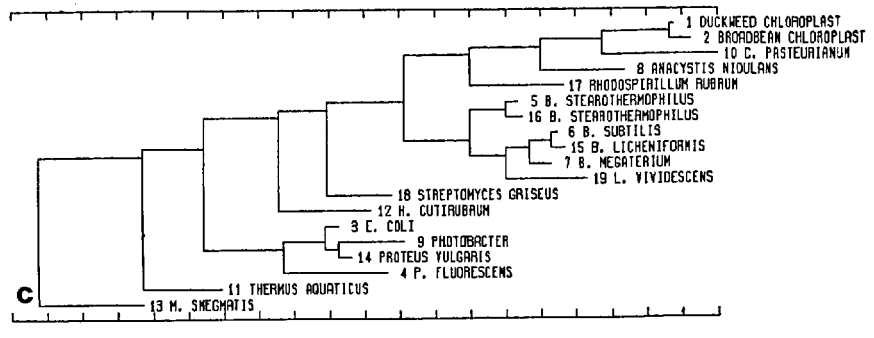
The use of non-homologous bases to produce homologous helices seems at first sight an ugly feature of the alignment. However, it points out an interesting fact: it may be 'easier' to conserve secondary structure by using non-homologous bases, since there are more possibilities for correction if such shifts are allowed. Several other observations support this idea. For one, minimal-energy foldings of homologous mRNAs have been shown to be similar but to use clearly non-homologous sub-sequences to attain this similarity (D. A. M. Konings et al., manuscript in preparation). For another, secondary-structure properties of amino acids [e.g., 'Chou Fasman parameters' (Chou and Fasman 1978)] do not cluster amino acids in groups resembling the ones formed by mutation frequencies as published by Dayhoff (1976) or McLachlan (1971). In contrast, such properties of amino acids as volume and hydrophobicity correlate strongly with mutation frequencies. From this we conclude that apparently non-local properties are involved and that non-homologous stretches may be used to maintain secondary structure through evolution.

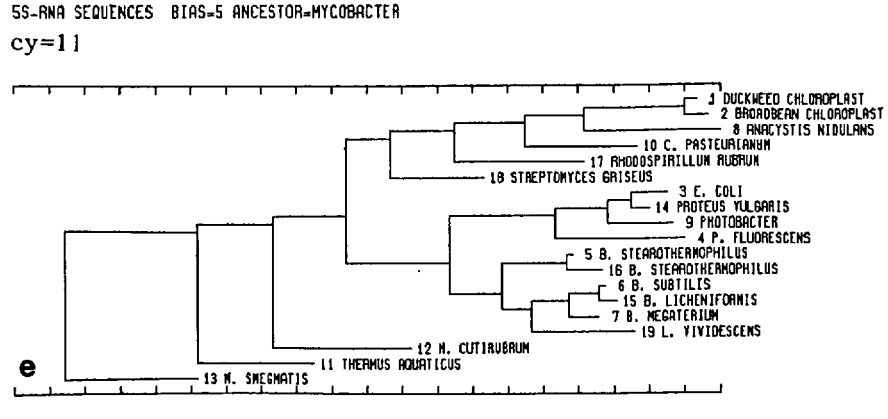
5.5. Initiation on an A Priori Tree

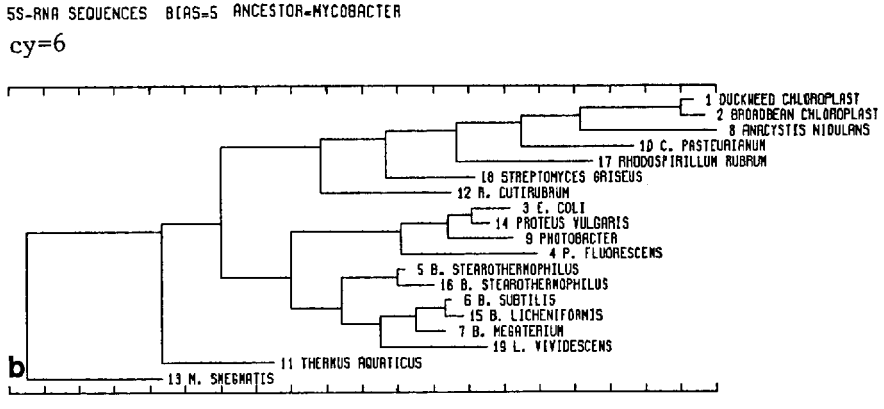
The method can be used starting with an arbitrary tree. Using the optimal tree proposed by Sankoff et al. (1982) we can evaluate both that tree and our method.

As mentioned, the number of mutations in their alignment relative to this tree is 423 (length = 132) as calculated by our binary criterion. When we align the sequences on this tree using our method we obtain 416 mutations (length = 135) for bias = 1,1. The branch lengths obtained from this alignment differ from the lengths obtained from their align-









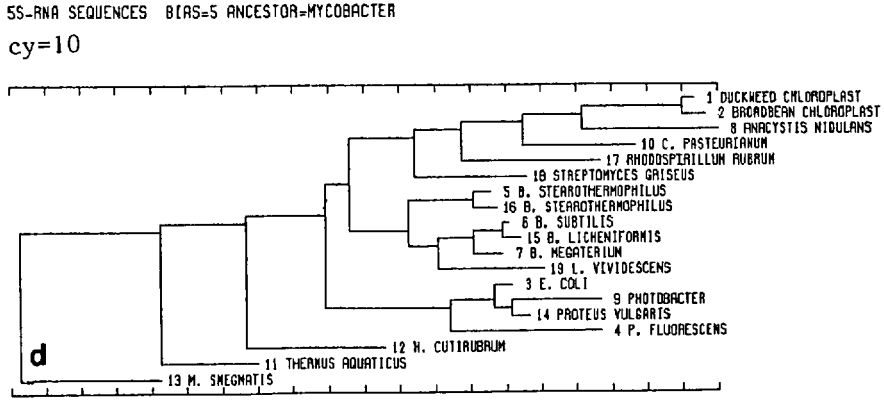


Fig. 4a-e. Five trees generated by the present-day ancestor method (ancestor = *Mycobacter*). **a** cy = 0; pairwise alignment only (bias = 5,0). Number of mutations = 427; length = 134. **b** cy = 6; treewise alignment (bias = 1,1). Number of mutations = 416; length = 131. **c** cy = 9; treewise alignment (bias = 1,1); the worst tree. Number of mutations = 443; length = 132. **d** cy = 10; treewise alignment (bias = 1,1); the most parsimonious tree. Number of mutations = 415; length = 132. **e** cy = 11; treewise alignment (bias = 1,1); stable tree. Number of mutations = 418; length = 133

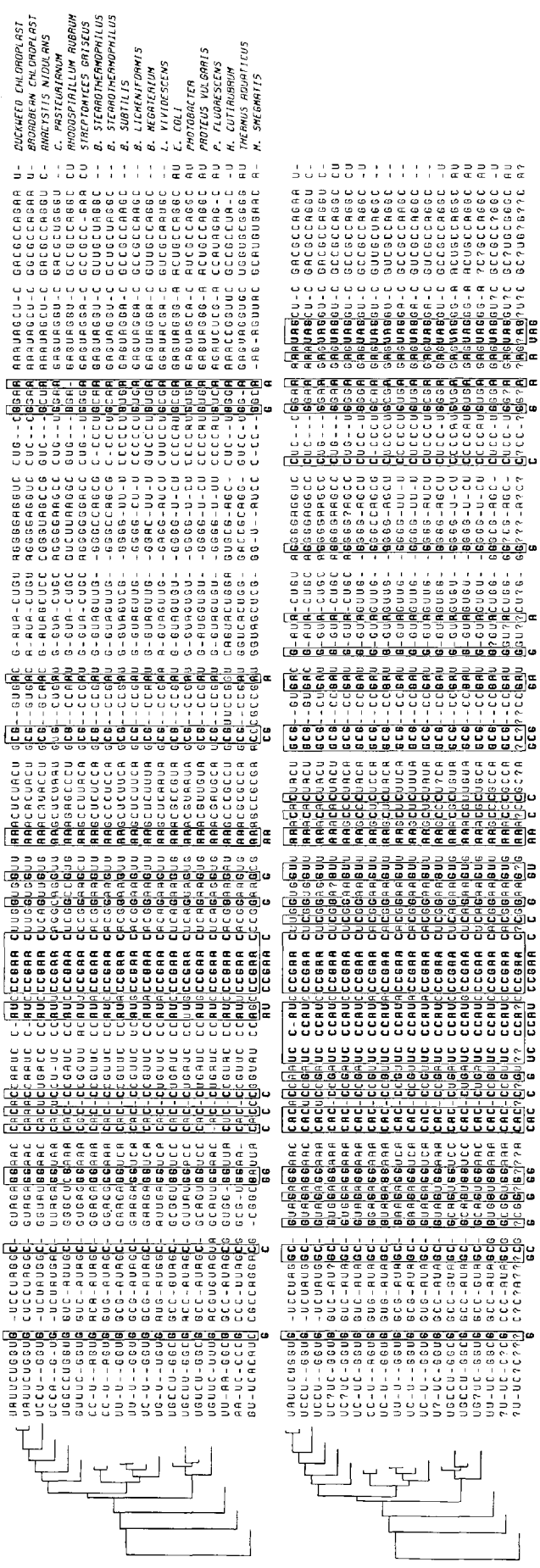


Fig. 5. Alignment generated on the most parsimonous tree (cy = 10; Fig. 4d) of the ancestor = *Mycobacter* iteration. Positions that vary in one sequence only are marked and boxed. **Top:** Alignment of present-day sequences. **Bottom:** Alignment of generated internode sequences

ment in that they exaggerate the main conclusions of their study: *Mycobacter* is closer to the Enterobacteria, and the Enterobacteria and the Bacilli differ more severely. This is caused by one of the features of the method mentioned in Section 2.3: inherent differences are entered later in the alignment. Iteration on this alignment using group averaging stabilises after the first step; the tree topology is not changed but the rooting is: the Enterobacteria (including *Mycobacter*) are remote from the other groups and determine the root. However, the different rooting results in 421 mutations and a length expansion to 134.

5.6. Conclusions

Comparing the different analyses we conclude the following:

1. Bacilli s.s. and Enterobacteria s.s. are similar, either by common descent or by convergence.

2. *Mycobacter* is remote but relatively similar to *Thermus aquaticus* and/or to the Enterobacteria.

3. *Streptomyces* and *Clostridium* are dissimilar to *Mycobacter* and fairly similar to *Halobacter*; in this respect they resemble the chloroplasts and *Anacystis* (and *Rhodospirillum*). They are, moreover, similar to the Bacilli. *Rhodospirillum* is similar to the chloroplasts and *Anacystis* but also has features in common with *Streptomyces* and *Clostridium*.

4. In none of the analyses was an affinity found among *Mycobacter* and *Streptomyces* and the Bacilli as occurs in, for example, the trees proposed by Fox et al. (1980).

Thus, if we recognise the affinity of *Mycobacter* and the Enterobacteria as well as the affinity of *Clostridium, Streptomyces* and the Bacilli, accepting the implied unequal mutation rates, the Enterobacteria s.l. and Bacilli s.l. become remote, and the similarity of the Bacilli s.s. and Enterobacteria s.s. must be attributed to convergence. Alternatively, if the remoteness of these species is taken as an indication of their early divergence, the Enterobacteria and Bacilli turn out to result from a late divergence.

Finally, we conclude that the generation of alternative alignments/trees gives more insight into evolutionary processes than does the early insertion of 'reasonable' assumptions. We think such insight is more important than the production of one definitive tree.

6. Discussion and Conclusions

In this paper we have proposed an integrated approach to the alignment of sets of sequences and the construction of phyletic trees. We think that it is meaningless to consider each of these problems in isolation. Alignment of sequences boils down to hypothesising insertion/deletion events in the course of evolutionary divergence; to hypothesise these events prior to and independently of phyletic tree construction therefore seems inappropriate. Moreover, to start evolutionary tree construction after hypothesising these (often considered major) evolutionary events seems to be begging the question. It is satisfying to find that the recognition of this interdependence has led to the development of a practical method for generating alignments of sets of sequences.

Several open questions remain about the performance of the method; for example, it is not clear that a convergent, stable alignment/tree always exists (or is reached); we have not yet found a case in which one does not. Therefore, the stopping criterion of the algorithm should check for the recurrence of identical tree topology to detect cycling. Cycling may occur when the alignment of a sequence C to a pair of other sequences A and B results in a higher similarity between A and C than between A and B, whereas the alignment of the pair A and C, followed by the alignment of B to this pair gives an alignment in which B is closer to A than C is. Indeed, we have found that alignment of a sequence to a set of homologous sequences often produces better results than the alignment of the sequence to any member of this set; this can be exploited to find homologies between distantly related sequences that have attained different functions: pairwise alignment will pick up spurious similarities, whereas alignment to a set will pick up consistent homologies (P. Hogeweg et al., manuscript in preparation).

We do not know whether the derived trees are the maximally parsimonous trees for the generated alignments, although we suspect them to be among the few most parsimonous trees because of the use of a parsimony criterion in the alignment step of the iteration. Moreover, among these trees they are the ones that also possess a high average within-lineage similarity not used in parsimony criteria. For example, sequences that share rare properties but do not occur in adjacent nodes of a tree will, by such average similarity criteria, be preferentially grouped into one lineage, whereas maximum parsimony criteria cannot use this information (compare Cornish-Bowden 1983). We think that such overall similarity should be considered, because it is likely that some sequence changes can be tolerated (in the biological sense) within a set of similar sequences, but not in all contexts (e.g., because of secondary-structure constraints). This argument is in our view so strong that we prefer matrix methods using overall similarity to maximum parsimony methods even in cases in which no alignment problems exist (if there are alignment problems, we have no choice), e.g. in non-sequence data. In non-sequence data the argument holds even more strongly, because context constraints then include regulatory constraints.

These open questions (i.e. about convergence and parsimony properties of the results) should be studied using computer simulations, which should include not only variable mutation rates between sequences but also, more importantly, variable mutation rates within sequences. Moreover, ideally the sequences should be subjected to non-local constraints on their composition, e.g. a folding pattern.

Whatever the outcome of such studies will be, we conclude that the proposed method, with the criteria used, is as good as the 'mind-mediated methods' used previously for alignment. It generates alignment/tree pairs with about the same number of mutations, the same gap expansion and the same number of gap-generating events (insertions/deletions) as the various tree-generating methods (including exhaustive search) produce on the a priori alignments. In addition to the ease of use of the proposed method and its generation of reproducible results, it is superior to previous methods in the following respects:

1. It recognises and uses the mutual interdependence of tree generation and alignment of sequences (assessment of homology).

2. It can generate different alignment/phyletic tree pairs by a change of parameters and is not restricted to an implicit arbitrary choice for either the tree construction or the alignment.

3. It does not use (implicitly or explicitly) external criteria for the construction of either the alignment or the tree (except those involved in the choice of the various criteria).

4. It can therefore be used to evaluate its internal criteria (parameters). (See, for example, the discussion of the value of the gap penalty).

5. It can also be evaluated relative to external knowledge (which is not implicitly used), e.g., secondary-structure considerations.

6. Evaluation of different tree proposals can be done without bias caused by the alignment, because the method includes the generation of the alignment of the sequences. Of course, as in all other tree-generation methods (including exhaustive search), the results remain biased relative to the set of sequences (species) used in the analysis.

We conclude that to use non-optimal methods optimally one should exploit their flexibility and explicitness. This should result in the consideration

of a variety of trees and alignments instead of just one tree which is the best one under a certain constricted definition of the universe.

Acknowledgments. We thank Miss S. McNab for linguistic advice. We thank a referee for his/her valuable comments.

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

Chou FE, Fasman GD (1978) Prediction of secondary structure of proteins from their amino acid sequence. Adv Enzyol. 47: 145–148

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

Dayhoff MO (1976) Atlas of protein structure, vol 5, suppl 2. National Biochemical Research Foundation, Washington, DC

Fitch WM, Yasunobu KT (1974) Phylogenies from amino acid sequences aligned with gaps: the problem of gap weighting. J Mol Evol 5:1–24

Fox GE, Woese CR (1975) 5S RNA secondary structure. Nature 256:505–507

Fox GE, Stackebrandt RB, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457–463

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

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

Hogeweg P, Hesper B (1972) BIOPAT, program system for bioinformatic pattern analysis. Bioinformatica, Utrecht

Hogeweg P, Hesper B (1981) Oligothetic characterisation of clusters. Pattern Recognition 14:131–136

Hori H, Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenetic tree of 54 5S RNA species. Proc Natl Acad Sci USA 16:381–385

Klotz LC, Blanken RL (1981) A practical method for calculating evolutionary trees from sequence data. J Theor Biol 91:216–272

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

Kuntzel H, Heidrich M, Piechulla B (1981) Phylogenetic tree derived from bacterial, cytosol and organelle 5 S rRNA sequences. Nucleic Acids Res 9:1451–1461

Leenhouts PW (1968) A guide to herbarium taxonomy. IAPT, Utrecht (Regnum vegetabile, vol 58)

McLachlan AD (1971) Tests for comparing related amino acid sequences. Cytochrome c and Cytochrome c551. J Mol Biol 61:409–424

McNeill J (1978) Purposeful phenetics. Syst Zool 28:465–482

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 48:443–453

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data II. Gene frequency data. J Mol Evol 19:153–170

Sankoff RJ, Cedergren RJ, McKay W (1982) A strategy for sequence phylogeny research. Nucleic Acids Res 10:421–431

Schwartz RM, Dayhoff MO (1978) Origins of Prokaryotes, Eukaryotes, Mitochondria and Chloroplasts: a perspective is derived from protein and nucleic acid sequence data. Science 199:395–403

Sellars PH (1974) On the theory and computation of evolutionary distances. SIAM J Appl Math 26:787–793

Smith TF, Waterman MS, Fitch WM (1981) Comparative biosequence metrics, J Mol Evol 18:38–46

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

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data I. Distantly related species. J Mol Evol 18:387–404

Waterman MS, Smith TF, Beyer WA (1976) Some biological sequence metrics. Adv Math 20:267–287

Received September 26, 1983/Revised January 14, 1984