Volume 10 Number 1 1982 Nucleic Acids Research

A strategy for sequence phylogeny research

David Sankoff[1], R.J.Cedergren[2] and W.McKay[1]

[1]Centre de recherche de mathématiques appliquées, and [2]Departement de biochimie, Université de Montréal, C.P. 6128, Montréal H3C 3J7, Canada

Received 30 September 1981

ABSTRACT

Minimal mutation trees, and almost minimal trees, are constructed from two data sets, one of phenylalanine tRNA sequences, and the other of 5S RNA sequences, from a diverse range of organisms. The two sets of results are mutually consistent. Trees representing previous evolutionary hypotheses are compared using a total weighted mutational distance criterion. The importance of sequence data from relatively little-studied phylogenetic lines is stressed. A procedure is illustrated which circumvents the computational difficulty of evaluating the astronomically large number of possible trees, without resorting to suboptimal methods.

INTRODUCTION

Progress in the phylogenetic inference of early molecular evolution is made through the accumulation of data from diverse evolutionary lines, coupled with the refinement of formal inference techniques and their application to these data. In this paper we illustrate how these two lines of investigation may interact in an efficient research strategy. Specifically, we discuss how uncertainties in some of the results of earlier phylogenetic reconstruction exercises is due to the lack of key RNA sequence data, how the incorporation of these into analytical inference schemes necessitates, and suggests the direction for, improvements in the formal methodology of reconstruction.

PROKARYOTE EVOLUTION

The study of molecular evolution has led to some conclusive results in the classification of prokoryotes into evolutionary taxa (1,2,3,4,5). For example, much is now known about the phylogenetic divergence within Bacilli and within the enterobacteria. The relationship between blue-green algae on one hand and plant, algal and Euglena chloroplasts on the other has also been well established, while any analogous evolutionary relationship between mitochondrial RNA and any known prokaryote line must have very

remote origins.

At the level of the early divergence of the prokaryotes into the major families now recognized, however, there is little consensus. Attesting to this is the number of rather contradictory classifications recently published, as well as the restraint shown by Fox et al (3) in not favouring any particular evolutionary hypothesis for subgrouping among the Bacilli, enterobacteria, spirochaetes, cyanobacteria and the micrococci in their comprehensive work on prokaryotic evolution. Hori and Osawa (2) give a more detailed hypothesis at early stages of evolution, but provide explicit error estimates to indicate the uncertainty inherent in their model.

It becomes clear then that the study of early prokaryote evolution will be little served by the availability of further 5S RNA sequences from among the Bacilli, say, or from plant chloroplasts. What is needed is information on phylogenetic lines from which no homologous sequence data it is yet available. This consideration prompted our recent choice of Rhodospirillum rubrum (6) as the source of potentially critical data, as well as a number of other organisms which are currently being investigated.

As for the choice of molecules to be sequenced, it is crucial for present purposes to opt for those for which as many sequences as possible, from as diverse a set of organisms as possible, have previously been reported. (This is a different context from the study of the pre-Darwinian origin of the genetic code, for example, in which it is equally important to have data on as many different tRNA molecules as possible, within the same organism - cf. Cedergren et al 1981.) The RNAs which satisfy best this criterion are 5S RNA (6,7,8) and phenylalanine tRNA (6,9,10) for which we have 19 and 9 distinct prokaryotic sequences, respectively.

OPTIMAL TREES

Much of the most insightful work on evolutionary inference involves a rather uninhibited eclecticism in the choice of data sources and in ways of evaluating and combining them. In contrast, we adopt a conservative, methodologically constrained approach in the hope that the disadvantages of the current sparseness of strictly comparable data will eventually be more than compensated for by the indisputable relationship between data and evolutionary hypothesis.

The algorithmic construction of phylogenetic trees based on molecular data is generally carried out in one of two types of ways, reflecting the phenetics/cladistics distinction prevalent in the wider field of numerical

systematics (11). In one approach, a distance matrix (or alternatively, a similarity matrix) is calculated as a first step, summarizing the overall differences (or resemblances) between all pairs of data sequences. A hierarchical clustering algorithm is then applied to the matrix, yielding a tree-like structure (dendogram, phenogram, hierarchy, etc.). These techniques, which aim at a 'best fit' between the matrix and the tree, are relatively rapid and can accommodate many sequences since the computation time and space need not exceed a constant times the square of the number of sequences, and may often require less.

The other approach is to try to construct a tree which is the most likely according to some phylogenetically easily interpretable criterion applied to each sequence term separately. The usual criterion is that of 'parsimony' or 'minimal mutations' which simply ensures that the evolutionary hypothesis implicit in the optimal tree involves, as little as possible, identical mutations at identical site occurring repeatedly in different evolutionary lines. (It does not stem from a belief that 'evolution takes the shortest path possible' as is sometimes mistakenly suggested.) We adopt this approach since it is biologically more directly intepretable, providing a 'best fit' between the data sequences and the tree, rather than between a matrix of distances (which represents much less information than the complete sequences) and the tree.

Unfortunately the construction of the minimal mutation tree is far more difficult than hierarchical clustering. The amount of computation time necessary grows exponentially with the number of data sequences, so that even extremely sophisticated programming cannot accommodate large data sets.

To circumvent this difficulty, a frequent approach has been to use stepwise and/or iterative algorithms which seek optimality by adding sequences one by one in an optimal way to a partial tree, or adjusting a tree by local changes, each of which decreases the number of mutations implied. Neither of these approaches guarantees optimality, and experience shows that with reasonably large data sets, they tend to produce a number of different trees, each depending on some arbitrary initial decision on the order in which the sequences are incorporated into the tree. To avoid this suboptimal behaviour, we have devised a procedure which has been programmed in Fortran and implemented on a CYBER 173 Computer. This is not guaranteed, in the mathematical sense, to produce the optimal tree, but will do so whenever this tree does not involve some thoroughly

startling grouping, from the biological point of view, such as the grouping of some chloroplast RNA and some Bacillus RNA more closely than the chloroplast groups with other chloroplasts and the Bacillus with other Bacilli. The procedure can generally be carried out in reasonable computation time. If as occasionally occurs, required computation time is excessive, this will be known early in the procedure. The key to the procedure is to impose restrictions on possible optimal trees, restrictions which simultaneously represent fairly certain biological knowledge and reflect clearcut patterns in the data. For example, in the $tRNA^{Phe}$ data to be discussed in the next section, it is biologically obvious, and unmistakable in the sequence similarities, that spinach chloroplast and bean chloroplast should be grouped together, and that Euglena chloroplast groups with these at a higher level, and that none of the other organisms are more closely related to any of these than they are to each other. If a restriction is imposed to the effect that the optimal tree must be consistent with this fact, the number of possible trees to be examined is still very large, but has been reduced to a tiny fraction of what it was without this restriction. If enough such restrictions can be imposed, the tree optimization problem can always be reduced to manageable proportions. If not, then we must have recourse to restrictions of a more equivocal nature with a concomitant reduction in our confidence in the results.

Each tree satisfying the set of restrictions is evaluated for its mutational 'cost' using a dynamic programming algorithm (12). The cost ratio of purine-purine or pyrimidine-pyrimidine transitions versus purine-pyrimidine or pyrimidine-purine transversions, versus base insertion or deletion is taken to be 0.45: 0.77: 1.0 (cf. 13, 5).

PHENYLALANINE tRNA

For the nine sequences in the $tRNA^{Phe}$ data set, only the restriction mentioned above as well as a grouping of the two Bacillus sequences were imposed a priori. The sequences were aligned as described in (14). The optimal tree (Figure 1) had length 35.68. Note that the position of the root is not determined by the optimality criterion. We have arbitrarily placed it to reflect the majority consensus on prokaryote evolution. Two trees which were almost as good (length 36.14) are schematized in Figure 2.

From these results it follows that

1. Anacystis groups with the chloroplasts
2. mycoplasma groups with the Bacilli

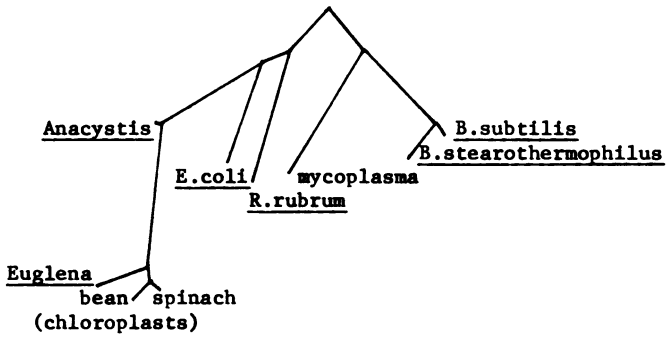


Fig. 1: Minimal mutation tree for 9 prokaryote phenylalanine tRNA sequences. Branch lengths proportional to weighted mutational distance between two end points.

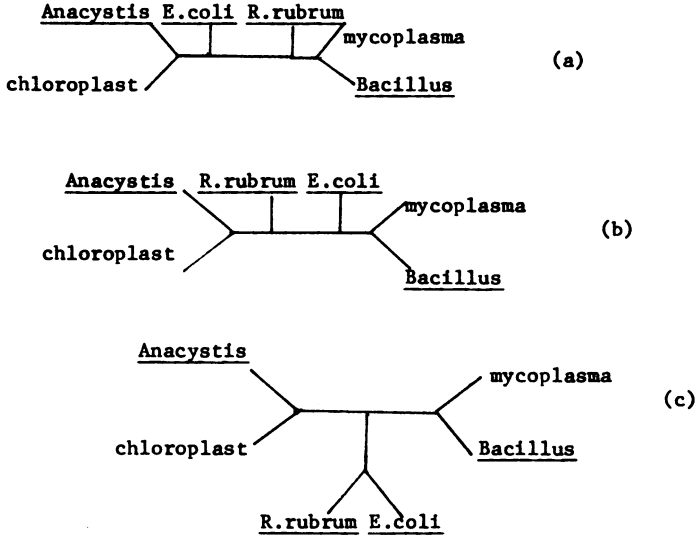


Fig. 2: Schematic representation of minimal mutation tree (a), total length 35.68, and two next best trees (b) & (c), total length 36.14 each. Tree root and branch lengths not indicated.

3. Anacystis-chloroplast does not group with Bacilli-mycoplasma

5S RNA

For the 19 5S RNA sequences, the restrictions summarized in Figure 3 were assumed. (cf. 2, Figure 4)

As will be discussed later, the positions of P. fluorescens, Clostridium and T. aquaticus have been placed elsewhere in some evolutionary theories.

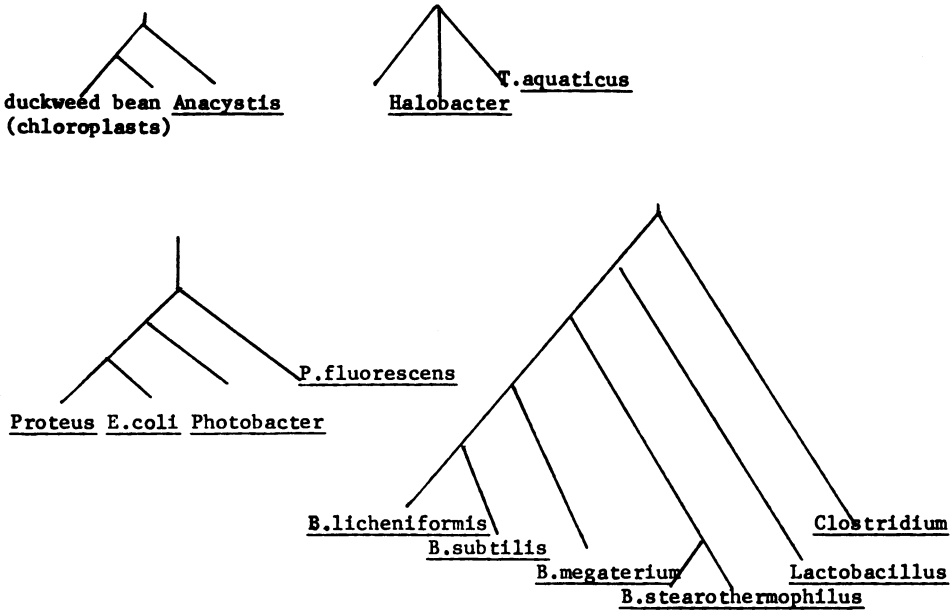


Fig. 3: Restrictions on optimal trees for 5S RNA phylogeny. All trees tested must contain the above subtrees.

In these sequence data, however, the above subgroupings are all clearly indicated. In particular, the grouping of T. aquaticus with Halobacter seemed unavoidable in preliminary data manipulation, although the two sequences are not very similar.

The alignment of the 5S sequences is less a matter of consensus than that of tRNA. We used the one in Figure 4, constructed as a reconciliation of the alignments in (15,2,16) as well as of secondary structure considerations.

The optimal tree, length 277.0, appears in Figure 5. The seven next best trees, with their lengths, are schematized in Figure 6. From these results it seems clear that

1. among the eubacteria, after the presumed remote origin of T. aquaticus, the Bacilli diverge from the rest as an early evolutionary event.
2. Probably S. griseus, but possibly Mycobacter, also branches off at an early stage.
3. R. rubrum groups with the Anacystis-chloroplast group or possibly with the enterobacteria.

duckweed chloroplast	UAUUCUGGUG-UCC-UAGGCGUAGAGGAACCACACCAAUC-CAUCCCGAA
broadbean chloroplast	UAUUCUGGUGCUCC-UAGGCGUAGAGGAACCAAACCAAUC-CAUCCCGAA
Escherichia coli	-UGCCUGGCGGCCG-UAG-CGCGGUGGUCCCAC-CUGACCCCAUGCCGAA
Pseudomonas fluorescens	-UGUUCUUUGACGAGUAGUAGCAUUGG-AACAC-CUGAUCCCAUCCCGAA
Bacillus stearothermophilus	---CCUAGUGACAA-UAG-CGGAGAGGAAACAC-CCGUUCCCAUCCCGAA
Bacillus subtilis	---UUUGGUGGCGA-UAG-CGAAGAGGUCACAC-CCGUUCCCAUACCGAA
Bacillus megaterium	---UCUGGUGGCGA-UAG-CGAAGAGGUCACAC-CCGUUCCCAUACCGAA
Anacystis nidulans	--UCCUGGUGUCUA-UGG-CGGUAUGGAACCACUCUGACCCCAUCCCGAA
Photobacter	-UGCUUGGCGACCA-UAG-CGUUAUGGACCCAC-CUGAUCCCUUGCCGAA
Clostridium pasteurianum	---UCCAGUGUCUA-UGA-CUUAGAGGUAACAC-UCCUUCCCAUUCCGAA
Thermus aquaticus	--AAUCCCCGCCCU-UAG-CGGCGUGGAA-CAC-CCGUUCCCAUUCCGAA
Halobacterium cutirubrum	--UUAAGGCGGCCA-UAG-CGGUGGGGUUACUC-CCGUACCCAUCCCGAA
Mycobacterium smegmatis	GUUCACAUCCGCCA-GGA-CGCGGCGAUUACACCCGGUAUCCAGCCCGAA
Proteus vulgaris	-UGUCUGGCGGCCA-UAG-CGCAGUGGUCCCAC-CUGAUCCCAUGCCGAA
Bacillus licheniformis	---UUUGGUGGCGA-UAG-CGAAGAGGUCACAC-CCGUUCUCAUGCCGAA
Bacillus stearothermophilus	---CCUAGUGGUGA-UAG-CGGAGGGGAAACAC-CCGUUCCCAUCCCGAA
Rhodospirillum rubrum	UGGCCUGGUGGUCA-UUG-CGGGCUCGAAACAC-CCGAUCCCAUCCCGAA
Streptomyces griseus	-GUUUCGGUGGUCA-UAG-CGUGAGGGAAACGC-CCGGUUACAUUCCGAA
Lactobacillus vividescens	---UGUUGUGAUGA-UGG-CAUUGAGGUCACAC-CUGUUCCCAUACCGAA

CUUGGUGGUUAAACUCUACUGCG--GUGA-CGAU-ACUGUAGGGG--AGGUCCUGCGGAAAAAUAGCU-CGACGCCAGA-AU
CUUGGUGGUUAAACACUACUGCG--GUGA-CAAU-ACUGUAGGGG--AGGUCCUCCGGAAAAAUAGCU-CGGCGCCAGA-AU
CUCAGAAGUGAAACGCCGUAGCG--CCGA-UGGU-AGUGU-GGGG--UCUCCCCAUGCGAGAGUAGGG-AACUGCCAGGCAU
CUCAGAGGUGAAACGAUGCAUCG--CCGA-UGGU-AGUGU-GGGG--UUUCCCCAUGUCAAGAUCUCG--ACCAUAGAGCAU
CACGGAAGUUAAGCUCUCCAGCG--CCGA-UGGU-AGUU--GGGGCCAGCGCCCCUGCAAGAGUAGGU-CGUUGCUAGGC--
CACGGAAGUUAAGCUCUUCAGCG--CCGA-UGGU-AGUC--GGGG-GUUUCCCCCUGUGAGAGUAGGA-CGCCGCCAAGC--
CACGGAAGUUAAGCUCUUUAGCG--CCAA-UGGU-AGUU--GGGA-CUUUGUCCCUGUGAGAGUAGGA-CGUUGCCAGGC--
CUCAGUUGUGAAACAUACCUGCG--GCAA-CGAU-AGCUCCCGGG--UAGCCGGUCGCUAAAAUAGCU-CGACGCCAGGUC-
CUCAGUAGUGAAACGUAAUAGCG--CCGA-UGGU-AGUGU-GGGG--UCUCCCCAUGUGAGAGUAGCA-CAUCGCCAGGCAU
CAGGCAGGUUAAGCUCUAAUGUG--CUGA-UGGU-ACUGCAGGGG--AAGCCCUGUGGAAGAGUAGGU-CGACGCUGGGU--
CACGGAAGUGAAACGCGCCAGCG--CCGA-UGGUCACUGG-GACC-GCAGGGUCCUGGA-GAGUAGGUGCUGGUGCGGGGAU
CACGGAAGAUAAGCCCGCCUGCGUUCCGGUCAGU-ACUGG-AGUG--CGAGCCUCUGGGAAAUCCGGU-UCGCCGCCUACU-
CCCGGAAGCGAAAGCCGCGAACCCGCCGA-UGGU-AGCU--CGGG--UAUCCCCCGGCAAGAGUUACG-CAU-GUGAACA--
CUCAGAAGUGAAACGUUGUAGCG--CCGA-UGAU-GGUGU-GGGG--UCUCCCCAUGUGAGAGUAGGG-AACUGCCAGGCAU
CACGGAAGUUAAGCUCUUCAGCG--CCGA-UGGU-AGUU--GGGG-GCUUCCCCCUGUGAGAGUAGGA-CGCCGCCAAGC--
CACGGAAGUUAAGCCCUCCAGCG--CCGA-UGGU-AGUU--GGGGCCAGCGCCCCUGCAAGAGUAGGU-CGCUGCUAGGC--
CUCGGCCGUGAAAGAGCCCUGCG--CCAA-UGGU-ACUG--CGUC--UUAAGGCGUGGGAGAGUAGGU-CGCCGCCAGGCCU
CCCGGAAGCUAAGCCUUACAGCG--CCGA-UGGU-ACUGCAGGGG--GGACCCUGUGGGAGAGUAGGA-CGCCGCCGAACU-
CACAGAAGUUAAGCUCAAUAGCG--CCGA-AAGU-AGUU--GGAGGAUCUCUUCCUGCGAGGAUACGA-CGUCGCAAUGC--

Fig. 4: Alignment of 19 5S RNA sequences.

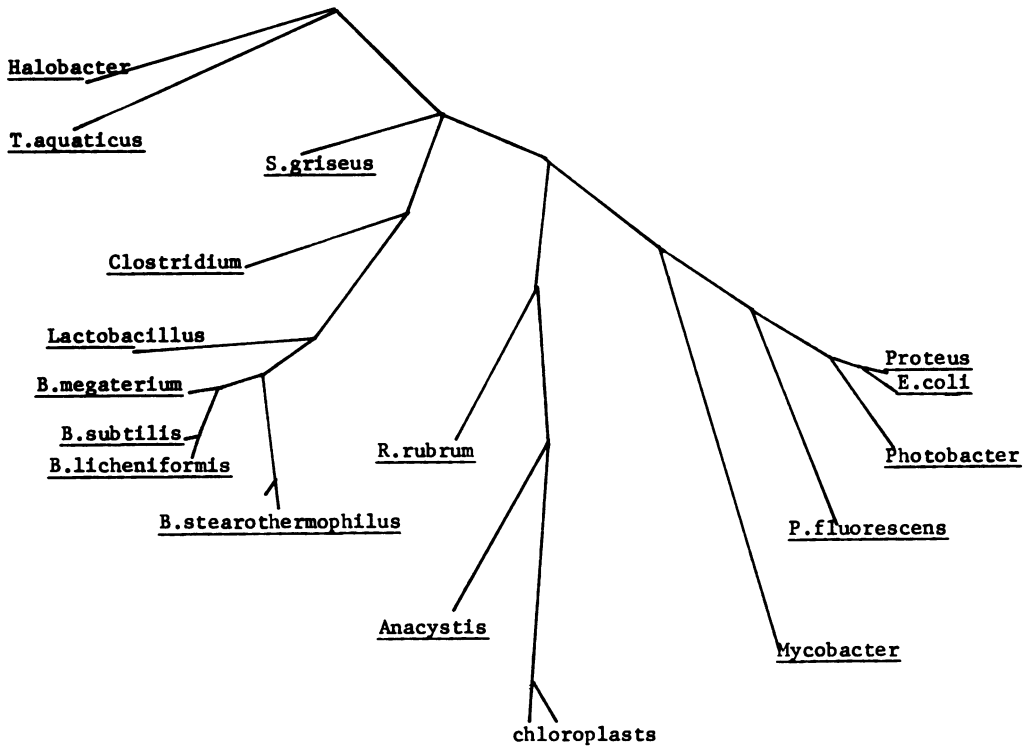


Fig. 5: Minimal mutation tree for 19 prokaryote 5S RNA sequences.

4. Both the Anacystis-chloroplast group and S. griseus seem remote from both E. coli and from Mycobacter.

Given that these trees do not differ by more than a few mutations in their total cost, it is perhaps inappropriate to infer any more definitive conclusions. Nevertheless, among the required possible trees, the field of possible evolutionary hypotheses has been drastically narrowed. Note that the 5S RNA results confirm those of tRNA with respect to the remoteness of the Bacillus and Anacystis-chloroplast groups.

EVALUATING PREVIOUS HYPOTHESES

The mutational cost criterion enables us to compare conflicting evolutionary theories. From the literature on prokaryote evolution we have selected four evolutionary hypotheses (1,2,3,4). Since these do not all treat the same range of organisms, we have enlarged the trees representing each theory by adding the missing organisms in a way which seemed to us

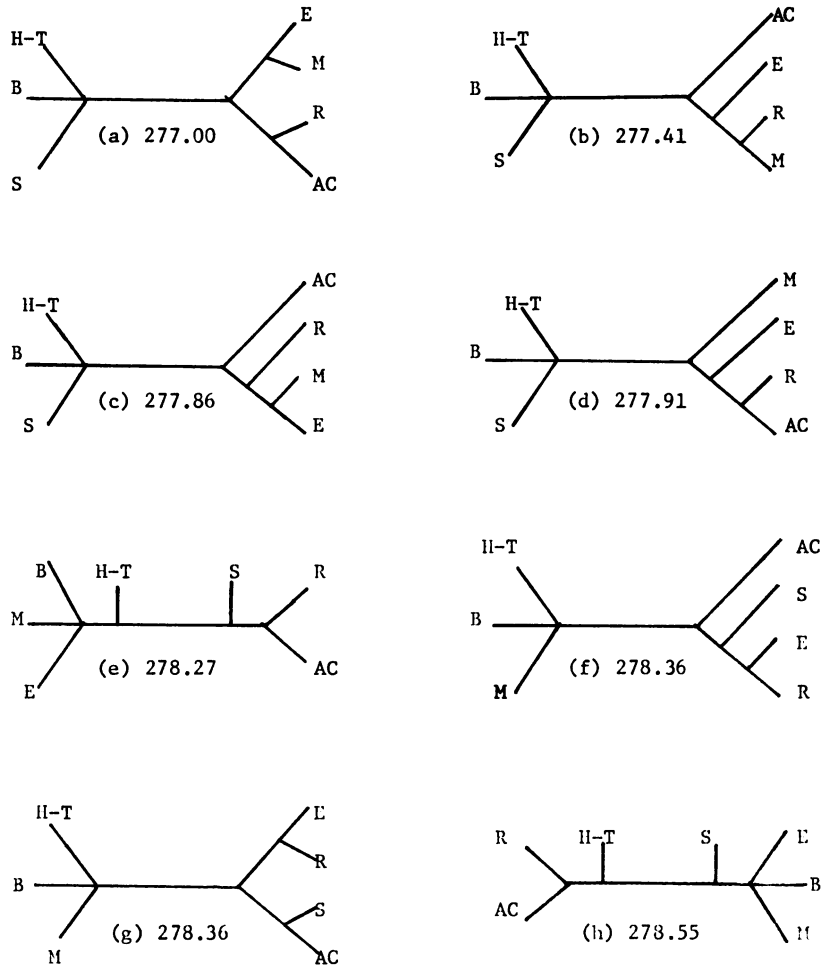
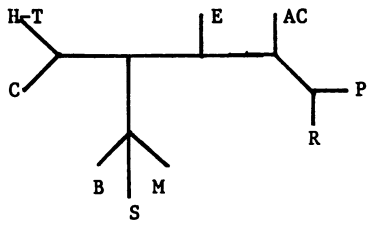
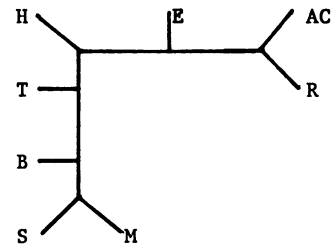


Fig. 6: Schematic representation of minimal mutation tree (a), and seven next best trees (b) - (h). Numbers indicate total mutational (weighted) length. Key: H - Halobacter, T - T.aquaticus, B - Bacilli, S - S.griseus, E - enterobacteria, M - Mycobacter, R - R.rubrum, AC - Anacystis & chloroplasts.

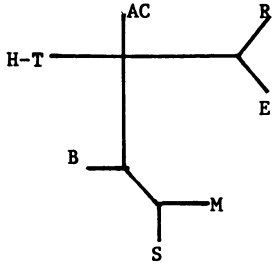
to do the least violence to the authors' hypotheses. Because of this non-comparability, however, and because these theories do not necessarily represent their authors' current views, we do not suggest that the following exercise is a conclusive test of any of the theories, and we will refer to the corresponding trees as 'S'(1), 'H'(2), 'F'(3) and 'K'(4) solely to indicate their ultimate origins in the literature. The trees are displayed



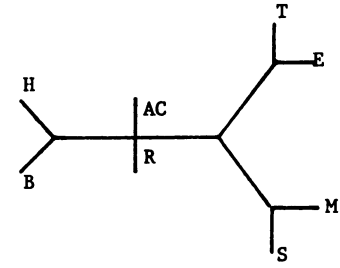
Hypothesis 'S' 287.41



Hypothesis 'H' 283.32



Hypothesis 'F' 280.23



Hypothesis 'K' 283.73

Fig. 7: Evaluation of four previous evolutionary hypotheses. Numbers indicate total mutational (weighted) distance. Subtrees in Fig.3 contained in all cases except where indicated. Key: H - Halobacter, T - T.aquaticus, B - Bacilli, S - S.griseus, E - enterobacteria, M - Mycobacter, R - R.rubrum, AC - Anacystis & chloroplasts, C - Clostridium, P - P.fluorescens.

in Figure 7 together with their lengths. It is clear that all are more costly than the tree we propose in Figure 5, and that the more they deviate from our tree, the more costly they are.

CONCLUSION

Increasing interest in sequence determination studies often seems like a rapid but random accumulation of scientific knowledge. One of the ways of introducing some order into this growth, so that our understanding of the implication of all this data increases at a comparable pace, is to focus on its evolutionary interpretations.

The choice of organism to be studied, if directed by considerations of evolutionary inference, is clearly not the further investigation of already well-explored phylogenetic lines, but the strategic choice of organism in

previously unstudied taxa. Inversely, the choice of molecular species to be sequenced should be one for which there is as much comparable information as possible.

The proliferation of sequence data eventually exceeds the capacity of rigorous minimal mutation methods. Rather than having recourse to rapid suboptimal or matrix methods, which lead to uncertain, ambiguous and non-unique results, we suggest here a way of combining reasonable degrees of biological and/or statistical certainty about the data with absolute optimization procedures. This reduces the computing problem without the disadvantages of suboptimal methods.

This study was supported by an NSERC COOP grant.

REFERENCES

1. Schwartz, R.M. and Dayhoff, M.O. (1978) Science 199, 395-403.
2. Hori, H. and Osawa, S. (1979) Proc. Natl. Acad. Sci. (U.S.A.) 76, 381-385.
3. Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magnum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. (1980) Science 209, 457-463.
4. Küntzel, H., Heidrich, M. and Piechulla, B. (1981) Nucl. Acids. Res. 9, 1451-1461.
5. Cedergren, R.J., Larue, B., Sankoff, D, and Grosjean, H. (1981) Crit. Rev. Biochem. 11, 35-104.
6. Newhouse, N., Nicoghosian, K. and Cedergren, R.J. (1981) Can. J. Biochem., in press.
7. Erdmann, V.A. (1981) Nucl. Acids. Res., 9, r25-r42.
8. Simoncsits, A. (1980) Nucl. Acids. Res., 8, 4111-4123
9. Gauss, D.H. and Sprinzl, M. (1981) Nucl. Acids. Res. r1-r23.
10. Canaday, J., Guillemaut, P., Gloeckler, R. and Weil, J.-H. (1981) in press.
11. Sneath, P.H.A. and Sokal, R.R. (1973) Principles and Practice of Numerical Classification, W.H. Freeman and Co. (San Francisco).
12. Sankoff, D. and Rousseau, P. (1975) Mathematical Programming 9, 240-246.
13. Sankoff, D., Cedergren, R.J. and Lapalme, G. (1976) J. Mol. Evol., 7, 133-149.
14. Larue, B., Cedergren, R.J., Sankoff, D. and Grosjean, H. (1979) J. Mol. Evol. 14, 287-300.
15. MacKay, R.N., Doolittle, W.F., and Gray, M.W. (1981) in press.
16. Fox, G.E., personal communication.