Reçu le 19 juin 1992

A parsimonay analysis of eukaryotic small subunit ribosomal RNA ("18 S") sequences

BY

Pascal ARROTIN and Vincent DEMOULIN*

(Département de Botanique, B22, Université de Liège, B-4000 Liège, Belgium)

(5 figures)

Using the principles of FELSENSTEIN's DNA parsimony analysis program (package PHYLIP 3.2), a program with the capacity to handle data sets of the magnitude of ss rRNA has been developed. With this program we applied an heuristic approach to 83 ss rRNA sequences from the Antwerp data bank, compared along 4230 positions of their alignment.

Branch and bound can also be applied to a smaller number of sequences.

We suggest a heuristic tree for a large number of species might be more interesting than a proven minimal tree for a smaller set of species. Comparisons are also made with earlier studies using distance matrix methods; caution is urged for deciding on the evolutionary position of widely divergent groups, whatever the tree building technique used.

Key Words : eukaryotes; small subunit ribosomal RNA sequences; phylogeny; parsimony analysis.

Introduction

Nucleic acid sequence data are presently produced at an exponential rate and very often used to reconstruct species phylogenies. Many such studies transform sequence data into a distance matrix following FITCH and MARGOLIASH's (1967) seminal paper. This reduces computational problems but entails a loss of information (STEEL *et al.*, 1988) and has been criticized especially by FARRIS (1981, 1985, 1986). While those criticisms have not been generally accepted (FELSENSTEIN, 1984, 1986), many authors are interested in using the full character state information and apply the parsimony technique which "correctly reports the evidential meaning of character distributions" (SOBER, 1985).

Parsimony is much more difficult to apply to large data sets than distance methods and not without problems (FELSENSTEIN, 1983; HENDY and PENNY, 1989; PENNY *et al.*, 1990), but it seems difficult to admit that properly used it is not superior to distance treatments (SMITH, 1989). An introduction to those various techniques and computationally even more difficult ones can be found in FELSENSTEIN (1988) and SWOFFORD and OLSEN (1990).

The small subunit ribosomal RNA is the molecule whose sequence has been most used for phylogenetic reconstructions (OLSEN, 1988; BAVERSTOCK and JOHNSON, 1990). However since it is about 2300 bases long in eukaryotes and hundred of sequences are available, it is difficult to treat even a part of those data using parsimony methods.

We have thus decided to upgrade some of the programs in the package PHYLIP of FELSENSTEIN to handle full sequences of ss rRNA. The PHYLIP package has the advantage of being widely and freely distributed and to be usable on IBM-PC. Until now it has been noted (FINK, 1986; PLATNICK, 1987, 1989) that it is rather slow and cannot handle large data sets. The version 3.0 of the package PAUP of Swofford can handle data of the type we are interested in. It is however until now available for Mac only and its performance on a 4230 positions alignment is unknown. We thus believe it would be interesting for the numerous PC owners who are used to PHYLIP, to be able to perform DNA parsimony analysis of large data sets.

Material and Methods

The data set included all 90 sequences of eukaryotic cytoplasmic small subunit ("18 S") ribosomal RNA available in spring 91 from the Antwerp data bank built up by R. DE WACHTER and coworkers. Reports on the structure of this data bank and availability of files are in Nucl. Ac. Res. (for example in NEEFS *et al.*, 1990). To take care of every possibility, this extensive collection of sequences covers 4230 positions.

* Author to whom correspondence should be adressed. Tél 32.41.56 38 53, Fax 32.41.56 38 40.

The sequence of *Halobacterium cutirubrum* was usually the prokaryotic outgroup.

We start with PHYLIP version 3.2, available from J. FELSENSTEIN, Department of Genetics, SK-50, University of Washington, Seattle, WA 98195. Our first aim was to upgrade the heuristic DNAPARS and branch and bound DNAPENNY programs, which are devised for a maximum of 25 species and 1500 sites, so as to be applicable to a large set of rRNA. We also devised a program for testing the parsimony of a given tree inspired from DNAMOVE and smaller programs for printing trees or consulting files.

We use a small PC IBM compatible with a 80286 processor, 640 Kbytes of RAM and a hard disk. To treat data sequences that cannot be contained by 640 Kbytes, a mechanism of writing to a hard disk has been implemented. We did our best to limit the number of disk accesses required to process the data and of course preprocessed the sequences to eliminate the columns of redundant information.

The method of FITCH (1971) is used to count the number of changes of base needed on a given tree. Change from an occupied site to a deletion is counted as one change. Reversion from a deletion to an occupied site is allowed and is counted as one change.

A difference of our programs with those of J. FELSENSTEIN is that we do include the branch length in the output trees. This length is proportional to the sum of discordances of nucleotides for each position of the sequence assigned at the two nodes or at the node and the leaf of a tree branch.

We also made it possible to specify a tolerance for the score of the selected output trees so as to be able to present not only the most parsimonious tree found but also those nearly as good at a certain number of score percent.

We devised programs of heuristic parsimony PARSHEU, interactive parsimony PARSINTER, branch-and-bound PARSBAB, tree drawing PTREE and file consultation CONTENU, SELSEQ, SELTREE. P.A. ARROTIN is the author of the programs, which are freely available from V. DEMOULIN (corresponding author premanently attached to the University of Liège).

Results and Discussion

An advantage of rRNA for phylogenetic reconstructions is that some sections of the molecule are highly conservative and can provide information on the relationships of distant taxa, while other sections are quite variable and provide data for the relationships of closely related taxa. The drawback of this situation is that in the comparison of distantly related taxa the highly variable positions introduce noise that can obscure phylogenetic relationship.

Distance matrix methods of phylogeny reconstruction are especially sensitive to this situation so that they are usually applied to limited sections of the molecule that are considered relevant for the comparison performed. HENDRIKS *et al.* (1991) give an interesting table of the number of positions used in previous phylogenetic studies of eukaryote ss rRNA.

This selection of sequence sections constitutes a loss of information that could be useful for interrelating neighboring taxa, moreover it introduces an element of subjectivity, while one of the most interesting aspects of developing phylogenies from sequence data is that this can be a totally automatic process.

A further way of addressing the problem of very different sequences being treated with distance methods which is frequently used is to introduce a correction for multiple hits. GILLESPIE (1986) has shown how this additional manipulation may not be adequate.

While parsimony is not immune to the problems of very divergent sequences it is able to integrate in a single operation the comparison of closely related and widely divergent sequences. We thus did not perform any selection of position. This apparently allowed us to resolve intragroup relationships which are not always deciphered by matrix methods. One particular reason for which variable regions are not too troublesome for the comparison of rRNA sequences is that they often are insertions which do not appear when ancestral sequences are compared.

We will not discuss in detail the results of the application of our programs : more runs and further inclusions of new sequences will make those biologically more significant. We will limit ourselves to the fact we have been able to apply an heuristic search to 86 sequences (a few polymorphic sequences have been omitted) and compared on the 4230 positions, necessary for aligning them completely. This took 15 h 15 min. For a further run we limited ourselves to 83 sequences which too 9 h. The resulting tree is presented in Fig. 1.

The program informed us that the three arrangements of the human sequences were equally parsimonious. Other equally parsimonious trees are of course possible (one was discoverd by interactive search in placing *Lycopersicon esculentum* as sister group of *Zea* and *Oryza*), but applying the global rearrangement option (not to speak of a branch and bound) to such a number of sequences was prohibitive : for 49 sequences if the simple heuristic approach took 2 h the global rearrangement needs 10 h.

We nonetheless consider the 83 sequences tree more interesting than others for which the technique gives a better probability, even certainty, to find the most parsimonious tree but are based on a smaller number of sequences. The reason is parsimony obviously gives results closer to the true evolutionary history when the number of sequences treated increases (HENDY and PENNY, 1989; PENNY *et al.*, 1990). For this reason we believe that with what should be the objective of molecular phylogenies, that is provide suggestions for thought, a heuristic tree of a large number of species might be better than a minimal tree for a smaller set. An example of this situation is that if the branch and bound is applied to the subtree of the seed plants, the most parsimonious tree (score 1643) (Fig. 2) is biologically impossible and among the 10 best trees it is only the third one (score 1647, Fig. 3) which gives a reasonable position to the gymnosperm *Zamia*.

One may however note that choosing a more closely related outgroup gives a more acceptable solution. With a green alga (*Chlamydomonas*) as outgroup the most parsimonious tree (score 712) is the same as the third one with the bacterium as outgroup. The con-

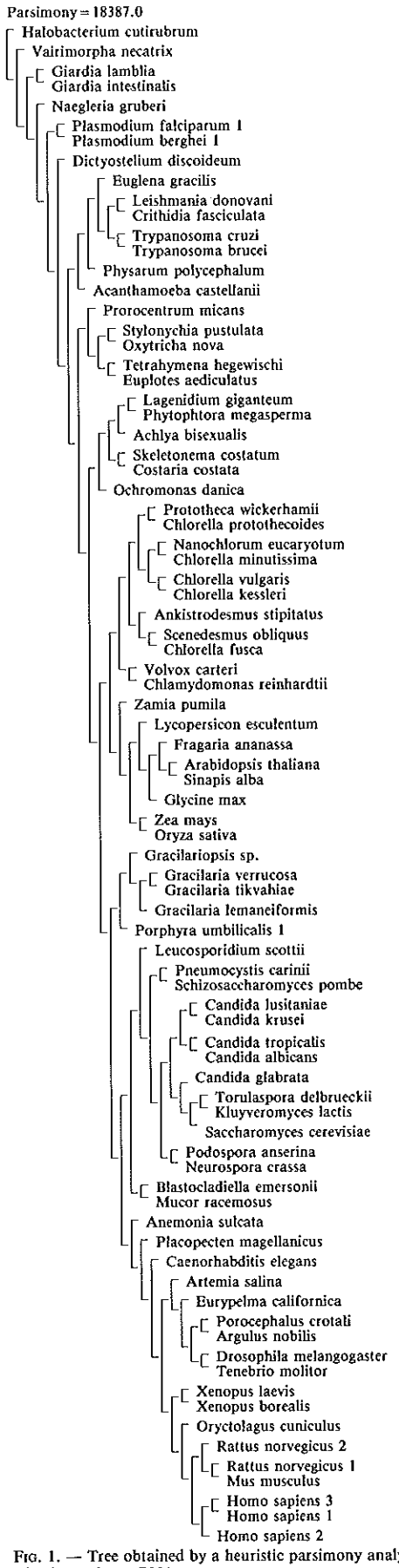

FIG. 1. — Tree obtained by a heuristic parsimony analysis for 83 sequences of eukaryotic ss rRNA.

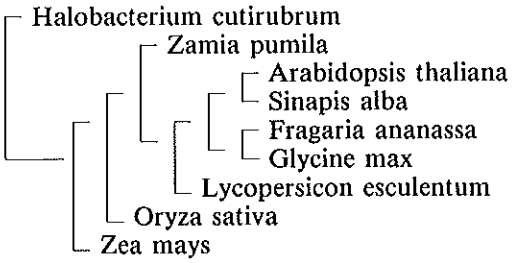

FIG. 2. — *Most parsimonious (score 1643) tree obtained by branch and bound for seed plants.*

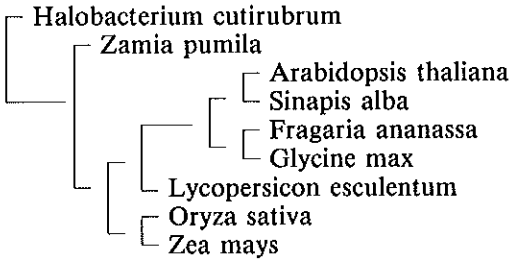

FIG. 3. — *Third best tree (score 1647) obtained by branch and bound for seed plants.*

figuration found by the heuristic approach, which was not among the ten best trees with the first outgroup is the second best tree in this case, almost as good as the first one : score 713.

Whithout wanting to give too much importance to our tree at the moment we however like to point out that the unorthodox branching order of dicotyledons is the same as the one obtained on the basis of several protein sequences (MARTIN *et al.*, 1985).

Finally we would like to compare our results with some obtained by distance methods.

First we have used the same 49 sequences as those treated by HENDRIKS *et al.* (1991) which constituted the most inclusive study of eukaryotic ss rRNA. It should be noted ESCHBACH *et al.* (1991) obtained the same tree with a neigborliness technique.

Eight trees of parsimony from 14896 to 14899 have been obtained heuristically. The best one (Fig. 4) can be compared to Fig. 3 of HENDRIKS *et al.* (1991). Methodologically the most interesting fact is that we can resolve the branching order of some groups of relatively closely related organisms like seed plants; for this specific group the pattern is that which is supported by the larger sample and was discussed above.

We can also note that several groups well supported by morphology and cytology (green algae and plants, ciliates, heterokont fungi and algae, ascomycetes, metazoa) are recognized by the two techniques. There are, however, two undisputed groupings that appear in our trees and are not found by the matrix method : the link of basidiomycetes (*Leucosporidium*) to ascomycetes and the insects (*Drosophila, Tenebrio*). This may indicate that our tree is more reliable for the disputed relationships, for which we however believe caution will always be necessary.

For distnat relationships we note differences in the position of *Acnthamoeba, Porphyra, Dictyostelium,*

Parsimony = 14896.0

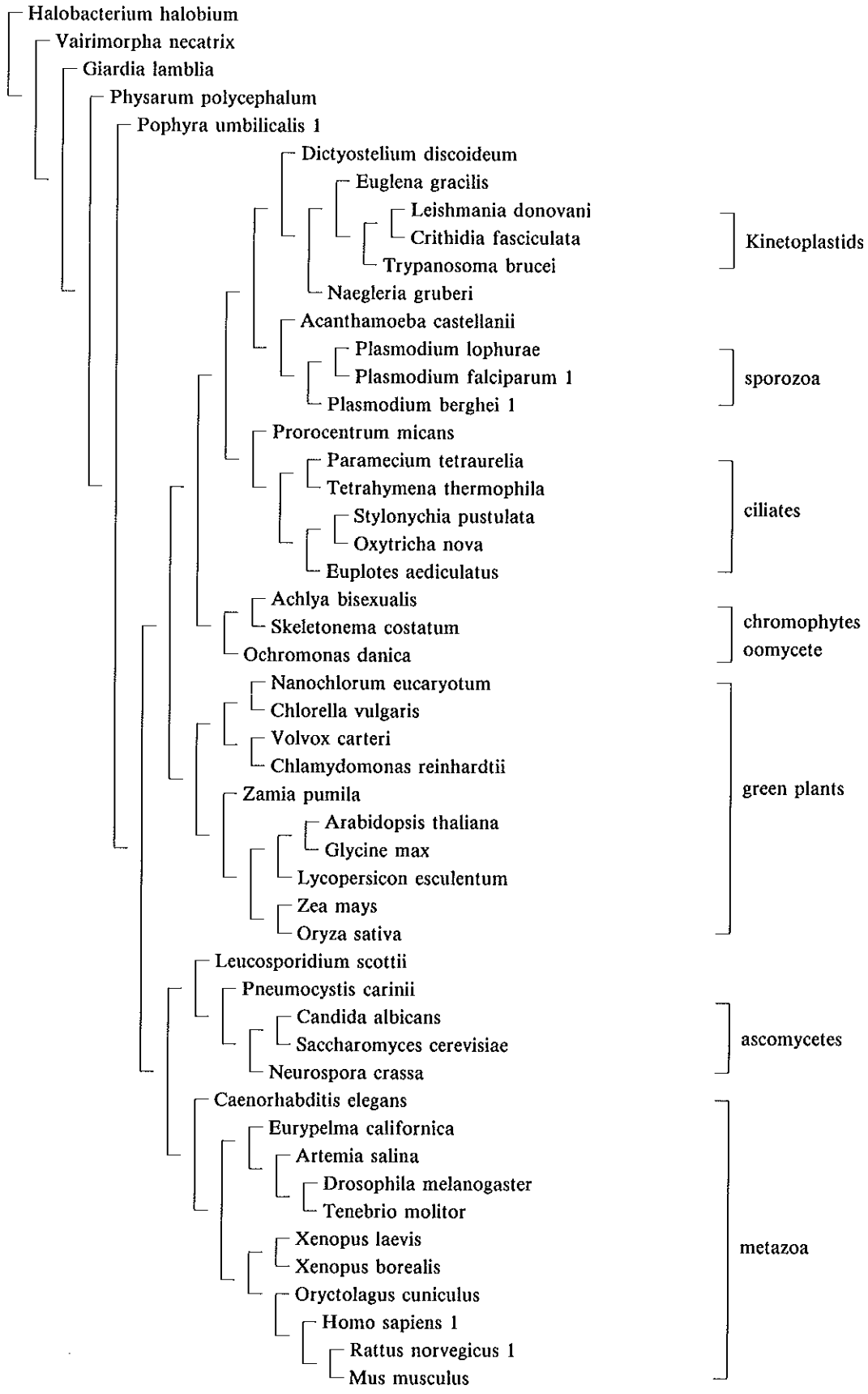

FIG. 4. — *Best tree obtained by a heuristic parsimony analysis for the 49 sequences used by* HENDRIKS *et al.* (1991).

Naegleria, Physarum, Vairimorpha and *Giardia*. The position of those sequences is constant in our 8 trees but is not always found in other trials for example when partially different sets of 50 and 57 sequences were treated with global rearrangement (this took 15 and 24 h respectively) nor in the treatment of 83 sequences. Some features were however constant in our trees like te fact *Vairimorpha* is the earliest branch followed by *Giardia*. The positions of the red algae and of *Dictyostelium* are especially variable, coming out very early, or later in the tree. HENDRIKS *et al.* (1991) also noted variation in the position of some of those organisms, induced in their case by changes in outgroups. It can also be noted that these authors obtained resolutions similar to ours for the green plants and *Leucosporidium* by using all positions in a more limited sample.

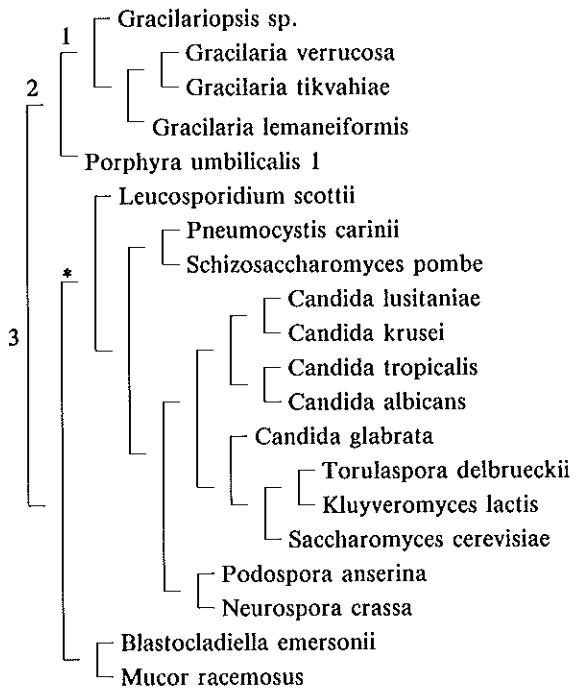

FIG. 5. — *Part of the tree in Fig. 1. on which rearrangements of the group denoted by a star (asco- and basidiomycetes) have been tried.*

Since we are especially interested in the position of red algae, we also compared our results with those of BHATTACHARYA *et al.*, (1990). In the tree of that paper, also based on a distance method, there are some similarities to the figure of HENDRIKS *et al.* (1991). The red algae are, however, again differently placed : after and not before the divergence of the Dinophyte-ciliate group (a fairly well supported group, whatever the technique or sample) and that of other algae, animals and fungi. In our analysis their position varies from a sister group of green algae and ancestral to fungi and metazoa with the 83 sequences set, to a very early divergence (before *Euglena*) in analysis with smaller samples. We thus believe it is not now possible to define the position of red algae with the ssrRNA data. It is still premature to conclude that red algae are unrelated to higher fungi. On that specific point we may add that we tested by our interactive program the placement of the higher fungi as either as sister group (Fig. 5) of eiter Florideophyceae (1), the whole red algae (2) or the group comprising the red algae, chytridiomycetes, zygomycetes, and animals (3). For an initial score of 18387 the placement 1 gives 18628, 2, 18431, 3, 18434. Those are small differences, less than 1,3%, and certainly those phylogenies cannot be rejected at the light of SMITH's (1989) study.

In conclusion we believe that caution asked for by BISHOP (1988), WOLTERS and ERDMANN (1988), and SMITH (1989) must be essential before deciding on the position of many protist groups based solely on their ribosomal RNA sequence.

Better information may come from more sequences which will improve the results of parsimony analysis or from more sophisticated data treatments like a posteriori weighting (PENNY and HENDY, 1985), but maybe some widely divergent sequences have lost most of the information that can be extracted by automatic methods.

Résumé

En se basant sur les principes du programme de FELSENSTEIN "DNA parsimony analysis" (package PHYLIP 3.2), un programme a été mis au point permettant de traiter des données de l'ampleur des séquences totales de rRNA petite sous-unité. A l'aide de ce programme, une approche heuristique a été appliquée à 83 séquences de ce type provenant de la banque de données d'Anvers, comparées sur les 4230 positions de leur alignement.

Le branch-and-bound peut également être appliqué à des comparaisons impliquant un plus petit nombre de séquences.

Nous suggérons qu'un arbre heuristique obtenu avec un grand nombre d'espèces peut-être plus intéressant qu'un arbre dont la minimalité est démontrée mais que ne concerne qu'un petit nombre de séquences. Des comparaisons sont également faites avec des études antérieures utilisant des méthodes basées sur des matrices de distance et la prudence nous parait nécessaire avant de décider de la position évolutive de groupes fortement divergents, quelque soit la technique utilisée pour contruire les arbres.

Acknowledgements. — We are indebted to J. FELSENSTEIN (University of Washington, Seattle) for providing us his important program package PHYLIP and R. DE WACHTER (University of Antwerp) and his coworkers, especially J.M. NEEFS, for providing us their file of ss rRNA. The service those people provide to the scientific community by making freely available such research tools cannot be overemphasized.

References

BAVERSTOCK, P.R. and JOHNSON, A.M. (1990) Ribosomal RNA nucleotide sequence : A comparison of newer methods used for its determination, and its use in phylogenetic analysis. *Austr. Syst. Bot.* 3, 101-110.

BHATTACHARYA, D., ELWOOD, H.J., GOFF, L.J. and SOGIN, M.L. (1990) Phylogeny of *Gracilaria lemaneiformis* (Rhodophyta) based on sequence analysis of its small subunit ribosomal RNA coding region. *J. Phycol.* 26, 181-186.

BISHOP, M.J. (1988) An overview of computing with DNA and protein sequences. In *Computational Molecular Biology* (A.M. Lesk, ed), pp. 3-13. Oxford University Press, Oxford.

ESCHBACH, S., WOLTERS, J. & SITTE, P. (1991) Primary and secondary structure of the nuclear small subunit ribosomal RNA of the Cryptomonad *Pyrenomonas salina* as inferred from the gene sequence : evolutionary implications. *J. Mol. Evol.* **32**, 247-252.

FARRIS, J.S. (1981) Distance data in phylogenetic analysis. In *Advances in Cladistics* (V.A. Frunk and D.R. Brooks eds), pp. 3-23. The New York Botanical Garden, New York.

FARRIS, J.S. (1985) Distance data revisited. *Cladistics* **1**, 67-85.

FARRIS, J.S. (1986) Distance and statistics. *Cladistics* **2**, 144-157.

FELSENSTEIN, J. (1983) Parsimony in statistics : biological and statistical issues. *Annual Rev. Ecol. Syst.* **14**, 313-333.

FELSENSTEIN, J. (1984) Distance methods for inferring phylogenies : a justification. *Evolution* **38**, 16-24.

FELSENSTEIN, J. (1986) Distance methods : a reply to Farris. *Cladistics* **2**, 130-143.

FELSENSTEIN, J. (1988) Phylogenies from molecular sequences : inference and reliability. *Annual Rev. Genet.* **22**, 521-565.

FINK, W.C. (1986) Microcomputers and phylogenetic analysis. *Science* **234**, 1135-1139.

FITCH, W.M. (1971) Toward defining the course of evolution : minimum change for a specific tree topology. *Syst. Zool.* **20**, 406-416.

FITCH, W.M. & MARGOLIASH, E. (1967) Construction of phylogenetic trees. *Science* **155**, 279-284.

GILLESPIE, J.H. (1986) Rates of molecular evolution. *Annual Rev. Ecol. Syst.* **17**, 637-665.

HENDRIKS, L., DE BAERE, R., VAN DE PEER, Y., NEEFS, J., GORIS, A. & DE WACHTER, R. (1991) The evolutionary position of the rhodophyte *Porphyra umbilicalis* and the basidiomycete *Leucosporidium scottii* among other eukaryotes as deduced from complete sequences of small ribosomal subunit RNA. *J. Mol. Evol.* **32**, 167-177.

HENDY, M.D. & PENNY, D. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**, 297-309.

MARTIN, P.G., BOULTER, D. & PENNY, D. (1985) Angiosperm phylogeny studied using sequences of five macromolecules. *Taxon* **34**, 393-400.

NEEFS, J.M., VAN DE PEER, Y., HENDRIKS, L. & DE WACHTER, R. (1990) Compilation of small ribosomal subunit RNA sequences. *Nucl. Ac. Res.* **18** (suppl.), 2237-2317.

OLSEN, G.J. (1988) Phylogenetic analysis using ribosomal RNA. *Methods Enzymol.* **164**, 793-812.

PENNY, D. & HENDY, M.D. (1985) Testing methods of evolutionary tree construction. *Cladistic* **1**, 266-278.

PENNY, D., HENDY, M.D., ZIMMER, E.A. & HAMBY, R.K. (1990) Trees from sequences : Panacea or Pandora's box ? *Austr. Syst. Bot.* **3**, 21-38.

PLATNICK, N.I. (1987) An empirical comparison of microcomputer parsimony programs. *Cladistics* **3**, 121-144.

PLATNICK, N.I. (1989) An empirical comparison of microcomputer parsimony programs, II. *Cladistics* **5**, 145-161.

SMITH, A.B. (1989) RNA sequence data in phylogenetic reconstruction : testing the limits of its resolution. *Cladistics* **5**, 321-344.

SOBER, E. (1985) A likelihood justification of parsimony. *Cladistics* **1**, 209-233.

STEEL, M.A., HENDY, M.D., PENNY, D. (1988) Loss of information in genetic distance. *Nature* **336**, 118.

SWOFFORD, D.L. & OLSEN, G.J. (1990) Phylogeny reconstruction. In *Molecular Systematics* (Hillis, D.M. and Moritz, C., eds), pp. 411-501.

WOLTERS, J. & ERDMANN, V.A. (1988) Cladistic analysis of ribosomal RNAs — the phylogeny of eukaryotes with respect to the endosymbiotic theory. *Biosystems* **21**, 209-214.