CABIOS

Vol. 12 no. 6 1996
Pages 543–548

SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny

N.Galtier[1], M.Gouy and C.Gautier

Abstract

SEAVIEW and PHYLO_WIN are two graphic tools for X Windows–Unix computers dedicated to sequence alignment and molecular phylogenetics. SEAVIEW is a sequence alignment editor allowing manual or automatic alignment through an interface with CLUSTALW program. Alignment of large sequences with extensive length differences is made easier by a dot-plot-based routine. The PHYLO_WIN program allows phylogenetic tree building according to most usual methods (neighbor joining with numerous distance estimates, maximum parsimony, maximum likelihood), and a bootstrap analysis with any of them. Reconstructed trees can be drawn, edited, printed, stored, evaluated according to numerous criteria. Taxonomic species groups and sets of conserved regions can be defined by mouse and stored into sequence files, thus avoiding multiple data files. Both tools are entirely mouse driven. On-line help makes them easy to use. They are freely available by anonymous ftp at `biom3.univ-lyon1.fr/pub/mol_phylogeny` or `http://acnuc.univ-lyon1.fr/`, or by e-mail to `galtier@biomserv.univ-lyon1.fr`.

Introduction

Reconstructing a phylogenetic tree from molecular data involves multiple tasks: sequence alignment, selection of sequences and sites to analyse, tree building, rooting, plotting and printing. These tasks are usually performed via successively called distinct programs that cannot be easily interconnected. We present here SEAVIEW (SEquence Alignment VIEW) and PHYLO_WIN, two computer programs dedicated to molecular phylogenetics, which allow completion of all the above tasks with a mouse–driven, graphical interface.

General display

Coloured, aligned nucleotide or protein sequences are displayed in a scrollable panel (Figures 1 and 2). A specific colour is assigned to the four nucleotides A, C, G, T and to five amino acid groups defined according to biochemical similarities. A distinct colour is used for gaps or for undetermined residues. The colour affected to each amino acid may also be redefined by the user; program SEAVIEW allows in addition to define an alternative colour scheme for amino acids and to switch between the standard and alternative schemes. The usage of both programs is explained through an on-line help mechanism.

The SEAVIEW program

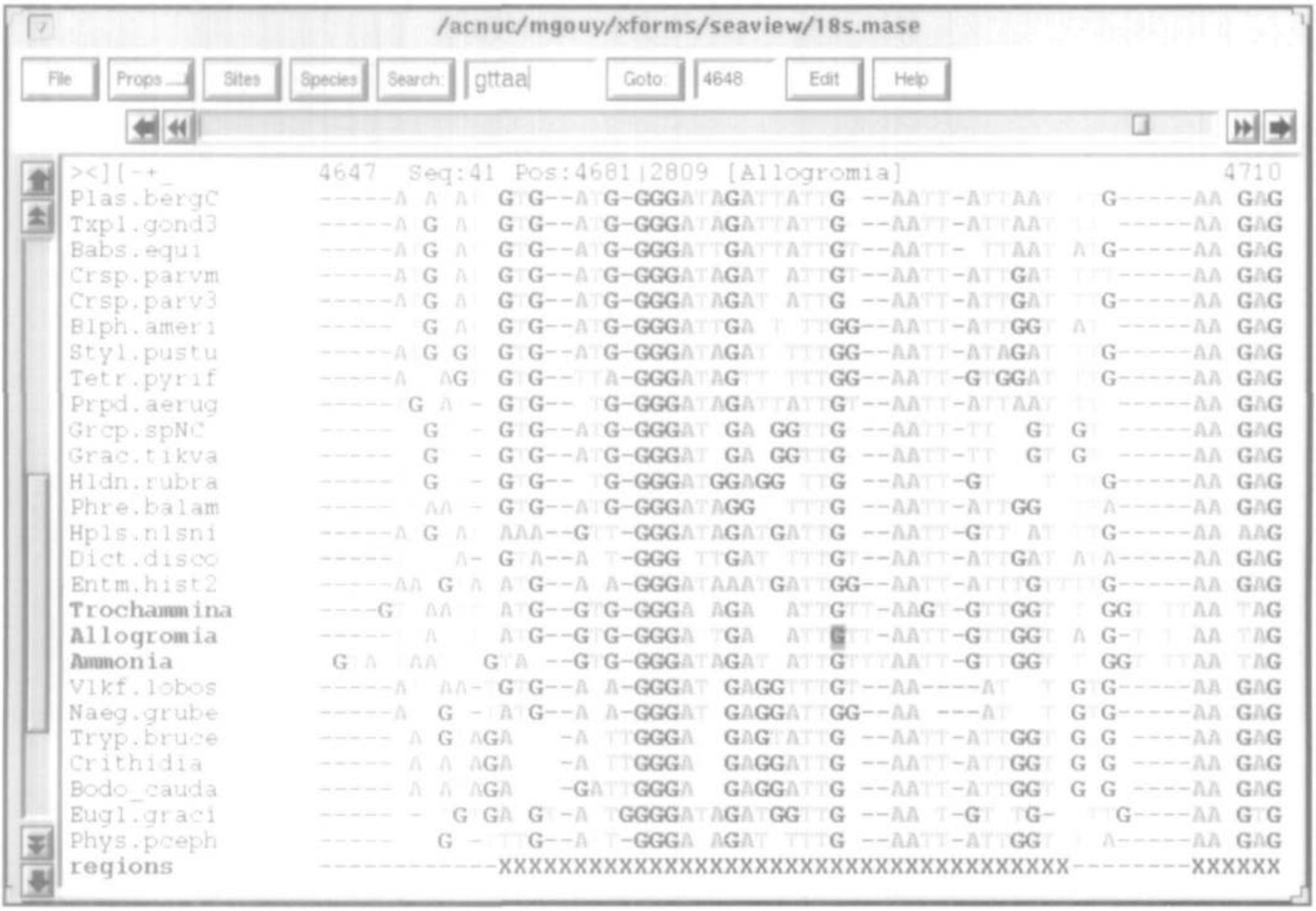
SEAVIEW (Figure 1) is a multiple sequence alignment editor adapted to nucleotide or protein data enhanced by an interface to the CLUSTALW multiple alignment program (Thompson *et al.*, 1994) and to the dot plot approach of pairwise sequence comparisons.

Alignment editing in SEAVIEW is characterized by the ability to alter in parallel the alignment of any group of sequences relatively to other members of the multiple alignment. The desired sequence group is specified by clicking or dragging on sequence names. Gaps may then be inserted in or deleted from all group members or in/from all other sequences in parallel. Sequence groups can be named and stored with the sequence data so that several groups can be alternatively employed. Sequences can be renamed and their order in the multiple alignment can be altered by mouse clicks. New sequence data can be typed by the user with optional use, for DNA data, of four neighbour keys of the keyboard for faster typing, or can be pasted to SEAVIEW from any source. Other features of interest to users of raw sequencing data allow to build the complementary strand or the reverse of a sequence. Multiple alignments can also be formatted for printout with full control of line and page sizes.

Nucleotide and protein sequences evolve at distinct rates at each site, so that frequently only part of the full set of sequence sites may be reliably aligned, fast evolving parts of the molecules being too variable in sequence and structure. Furthermore, the exact set of unambiguously alignable sites depends on the phylogenetic depth under consideration. Therefore molecular phylogeneticists frequently have to supplement a multiple alignment by one or several sets of accurately aligned sites. SEAVIEW offers a graphical solution to this problem by allowing site sets to

CNRS UMR 5558, Biométrie, Génétique et Biologie des Populations, Université Claude Bernard Lyon 1, 43, Boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

[1] *To whom correspondence should be addressed. E-mail galtier@biomserv.univ-lyon1.fr*



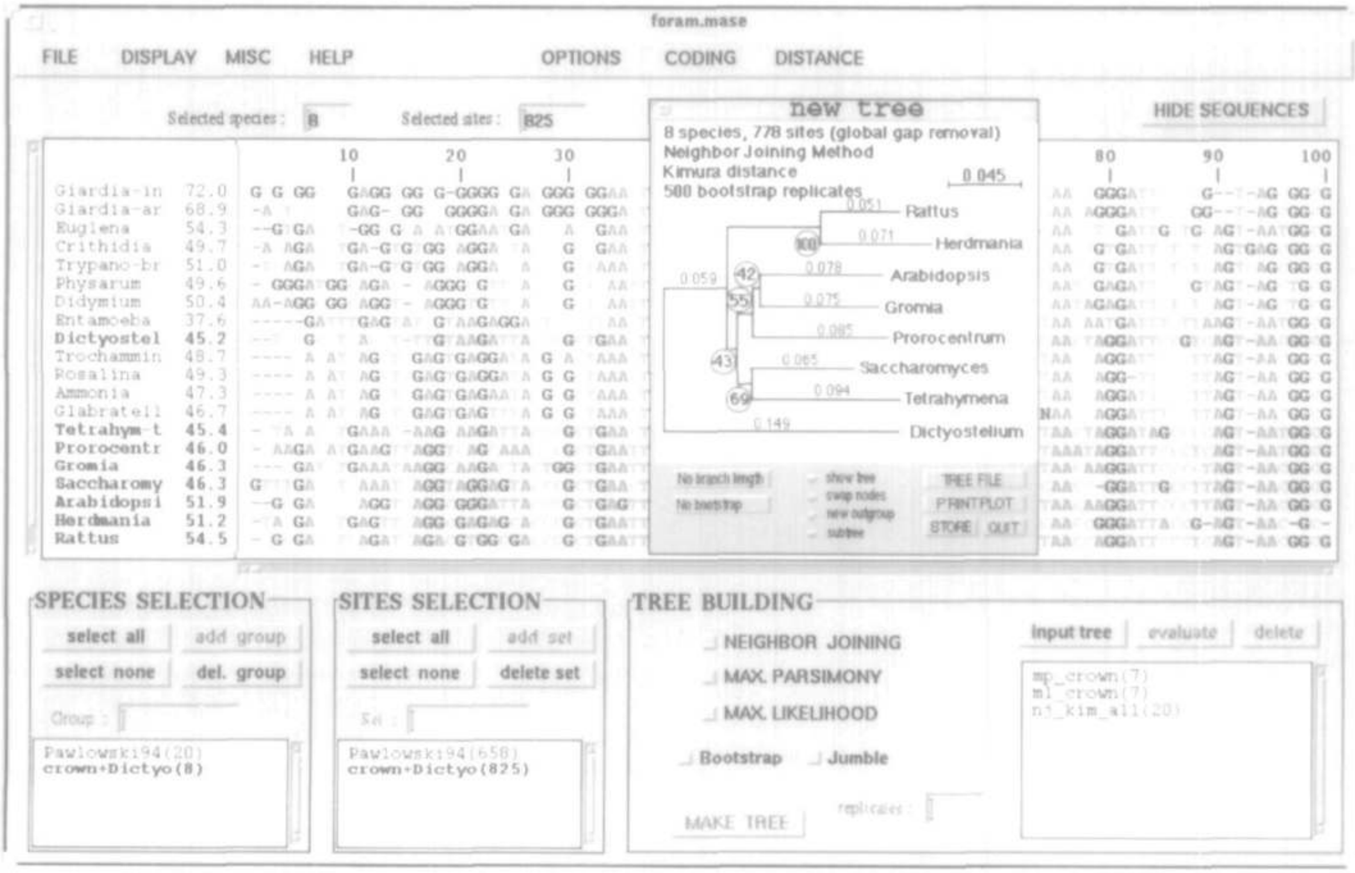
```
><][-+_         4647   Seq:41  Pos:4681|2809 [Allogromia]                  4710
Plas.bergC      -----A A A  GTG--ATG-GGGATAGATTATTG --AATT-ATTAAT  TG-----AA GAG
Txpl.gond3      -----A G A  GTG--ATG-GGGATAGATTATTG --AATT-ATTAAT TT -----AA GAG
Babs.equi       -----A G A  GTG--ATG-GGGATTGATTATTGT--AATT- TTAAT A G-----AA GAG
Crsp.parvm      -----A G A  GTG--ATG-GGGATAGAT ATTGT--AATT-ATTGAT TTT-----AA GAG
Crsp.parv3      -----A G A  GTG--ATG-GGGATAGAT ATTG --AATT-ATTGAT  TG-----AA GAG
Blph.ameri      -----  G A  GTG--ATG-GGGATTGA T TTGG--AATT-ATTGGT A  -----AA GAG
Styl.pustu      -----A G G  GTG--ATG-GGGATAGAT TTTGG--AATT-ATAGAT  TG-----AA GAG
Tetr.pyrif      -----A  AG  GTG--TTA-GGGATAGTT TTTGG--AATT-GTGGAT  TG-----AA GAG
Prpd.aerug      ------G A - GTG-- TG-GGGATAGATTATTGT--AATT-ATTAAT T  -----AA GAG
Grcp.spNC       ------ G  - GTG--ATG-GGGAT GA GGTTG --AATT-TT  GT G  -----AA GAG
Grac.tikva      ------ G  - GTG--ATG-GGGAT GA GGTTG --AATT-TT  GT G  -----AA GAG
Hldn.rubra      ------ G  - GTG-- TG-GGGATGGAGG TTG --AATT-GT   T  TG-----AA GAG
Phre.balam      ------ AA - GTG--ATG-GGGATAGG   TTTG --AATT-ATTGG   TA-----AA GAG
Hpls.nlsni      -----A G A  AAA--GTT-GGGATAGATGATTG --AATT-GTT AT  TG-----AA AAG
Dict.disco      ------   A- GTA--A T-GGG TTGAT TTTGT--AATT-ATTGAT A A-----AA GAG
Entm.hist2      -----AA G A ATG--A A-GGGATAAATGATTGG--AATT-ATTTGTTTTG-----AA GAG
Trochammina     ----GT AA T ATG--GTG-GGGA AGA  ATTGTT-AAGT-GTTGGT T GGT TTAA TAG
Allogromia      ------ A  T ATG--GTG-GGGA TGA  ATTGTT-AATT-GTTGGT A G-T T AA TAG
Ammonia          G A AA   GTA --GTG-GGGATAGAT ATTGTTTAATT-GTTGGT T GGT TTAA TAG
Vlkf.lobos      -----A  AA-GTG--A A-GGGAT GAGGTTTGT--AA----AT  T GTG-----AA GAG
Naeg.grube      -----A  G - ATG--A A-GGGAT GAGGATTGG--AA ---AT  T GTG-----AA GAG
Tryp.bruce      ----- A G AGA   -A TTGGGA  GAGTATTG --AATT-ATTGGT G G ----AA GAG
Crithidia       ----- A A AGA   -A TTGGGA  GAGGATTG --AATT-ATTGGT G G ----AA GAG
Bodo_cauda      ----- A A AGA   -GATTGGGA  GAGGATTG --AATT-ATTGGT G G ----AA GAG
Eugl.graci      ----- -  G GA GT-A TGGGGATAGATGGTTG --AA T-GT TG-  TTG----AA GTG
Phys.pceph      -----  G -TTTG--A T-GGGA AGAT TTTG --AATT-ATTGGT T A-----AA GAG
regions         ------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX--------XXXXXX
```

Fig. 1. SEAVIEW main window with colours replaced by shades of grey. Here and in Figure 2, bold names indicate selected sequences, and selected sites appear on dark background.

be specified by the user through mouse clicks and drags. Several such site sets may be defined and stored in an alignment file, and one of them can be visualized with selected sites appearing on a dark background. Site set coordinates get automatically updated when the alignment is changed by gap insertions. Subsets of a multiple alignment in terms of site sets and/or of sequence sets can be extracted from the full data file.

SEAVIEW does not contain any multiple alignment algorithm but does offer a graphical interface to the CLUSTALW multiple alignment program if this program is installed on the user's computer (Thompson *et al.*, 1994). The user may define a set of sequences and a series of contiguous sequence sites on which the CLUSTALW algorithm can be applied. If this procedure is applied to several but not to all sequences of the multiple alignment, gaps inserted by CLUSTALW in the longest processed sequence are propagated by SEAVIEW to other sequences of the alignment, thus maintaining the relative alignment of processed and unprocessed sequences.

Many genes, and particularly ribosomal RNAs, can vary extensively in length when organisms separated by large evolutionary distances are compared (De Rijk *et al.*, 1996; Van de Peer *et al.*, 1996). Alignment algorithms derived from dynamic programming methods such as CLUSTALW often fail to correctly align complete genes with extensive length variations. The standard procedure in such cases is to first locate and align highly conserved regions and then to try and align variable regions spanning between conserved ones. The dot plot method (Li and Graur, 1991) is a very efficient way of visually identifying conserved regions between two sequences even when substantial length differences occur. SEAVIEW implements the dot plot method between two sequences as an interactive tool for manual alignment: the dot plot between two sequences, typically a previously aligned sequence and another to be aligned, is computed ignoring all gaps present in the sequences; the dot plot is then drawn in a window with coordinates of matching regions changed to take gaps into account; the user selects a diagonal of the dot plot which indicates a local similarity between the two sequences; corresponding sequence regions appear on the screen; if the user decides that the two similar regions are indeed homologous, he/she may



Selected species : 8

Selected sites : 825

8 species, 778 sites (global gap removal)
Neighbor Joining Method
Kimura distance
500 bootstrap replicates

Group :

Pawlowski94(20)
crown+Dictyo(8)

Set :

Pawlowski94(658)
crown+Dictyo(825)

mp_crown(7)
ml_crown(7)
nj_kim_all(20)

Figure 2. PHYLO_WIN main window and tree editor with colours replaced by shades of grey

click on a button to have the two regions aligned to one another through insertion of gaps just before one of them; after several such operations all local similarities judged as homologous become located on the main diagonal of the plot; another button allows to propagate all gap insertions to the multiple alignment itself. SEAVIEW use in our laboratory as shown this dot plot-guided alignment to be also useful for assembling overlapping data from several sequencing gels.

The PHYLO_WIN program

PHYLO_WIN is an entirely mouse-driven interface for molecular phylogenetic purposes (Figure 2). Its basic functions are:

- displaying a sequence alignment.
- allowing an easy selection of sequences and sites to analyse.
- reconstructing phylogenetic trees according to numerous methods
- drawing and printing the reconstructed trees.
- saving species groups, site sets and trees into a file together with the data.

Any subset of the data may be selected as target of a phylogenetic analysis by choosing both sequences and sites. User-defined species groups and site sets can be stored. The lists of stored species groups and site sets are displayed so that any of them can be easily recalled. These lists may be saved to and loaded from files together with the sequence data and can be exchanged both ways with program SEAVIEW. For coding nucleotidic sequences, first and/or second and/or third codon positions can be selected. Several basic statistics including nucleotide or amino acid frequencies, number of variable and informative sites, transition/transversion ratios and observed substitution matrices can be computed for the currently selected data set. Nucleotidic sequences can be recoded into R's for purines and Y's for pyrimidines, so that only transversion-type changes are taken into account in subsequent phylogenetic analysis.

Tree–making methods

The distance-based neighbor joining method (Saitou and Nei, 1987), the maximum parsimony method (Fitch, 1971) and the maximum likelihood method for nucleotide sequences (Felsenstein, 1985; Olsen *et al.*, 1994) are implemented. Various distance estimation methods are provided. For nucleotide sequences, these are: observed divergence, Jukes and Cantor (1969) distance, Kimura (1980) two-parameter distance, Tajima and Nei (1984) distance, Galtier and Gouy (1995) distance, logdet distance (Steel, 1993; Lake, 1994). For protein sequences, observed divergence and the Poisson correction may be used. For protein-coding sequences, synonymous and non-synonymous substitution rates (Ks, Ka: Li *et al.*, 1985; Li, 1993) may be computed. Gaps are handled either globally (any gap-containing site is ignored) or pairwise (only those sites with a gap in one of the two currently compared sequences are ignored). Distance matrices may be output to a file. The maximum parsimony algorithms are those of programs DNAPARS and PROTPARS from the PHYLIP package (Felsenstein, 1989, use of C code granted by Joseph Felsenstein). Up to ten equally most parsimonious trees are recovered. The maximum likelihood algorithm for nucleotide sequences is that of fastDNAml program (Olsen, 1994, use of C code granted by Gary Olsen). The evolutionary model used in this program is described by Yang (1994). The assumed transition/transversion ratio and the number of branches crossed by moving subtrees during tree rearrangements (G option in PHYLIP) can be set.

Bootstrap analysis

A bootstrap analysis (Efron, 1982; Felsenstein, 1985) can be performed with any of the above described methods. As the implemented maximum parsimony and maximum likelihood algorithms are order-dependent, sequence input order is randomly drawn for each bootstrap replicate. This 'jumble' option (J option in fastDNAml program) can also be invoked for simple analysis without bootstrap. A bootstrap consensus tree may be constructed. It is defined by the $n - 3$ best supported compatible internal branches, where n is the number of compared taxa.

An original function of PHYLO_WIN is the ability to evaluate the bootstrap support of a combination of adjacent branches. Two non-overlapping subsets of the compared taxa are defined by selecting adjacent branches and/or unselecting taxa. The proportion of bootstrap trees including at least one internal branch separating the defined two subsets is computed. This function is useful to check the support of phylogenetic relationships among a subset of taxa whatever the behaviour of the remaining taxa, without loss of phylogenetic information. An example is given in the *Results* section.

Tree editor

Trees are drawn within separate windows. Branch lengths and bootstrap values may be superimposed. The location of the root can be set, branch swapping is allowed, and any magnified subtree can be viewed. Trees can be stored; a list of currently stored trees is displayed. Stored trees may be re-viewed without repeating the analysis, and may be evaluated according to four criteria, namely maximum parsimony (minimum required number of steps), maximum likelihood (likelihood of the topology after branch lengths optimization), minimum evolution (total length of the tree after least-squares estimation of branch lengths) and least-squares criterion (residual sum of squares after least-squares estimation of branch lengths). Alternative topologies may therefore be compared according to those criteria. Trees can be read from/saved to text files, and printed on PostScript printers.

Input and Output

The main sequence file format in SEAVIEW and PHYLO_WIN is an improved version of the MASE (Faulkner and Jurka, 1988) format called MASE+. In a MASE+ file, aligned sequences, names and comments are stored together with optional information about groups of species, sets of sites, and phylogenetic trees. SEAVIEW and PHYLO_WIN are compatible with most programs commonly used for phylogenetic analysis: the CLUSTAL (Thompson *et al.*, 1994), FASTA (Pearson and Lipman, 1988), PHYLIP (Felsenstein, 1989) formats and the MSF format of the GCG package (GCG 1994) are allowed. The capability of the MASE+ format of holding sequence comment data, which can be edited through SEAVIEW, can be employed to store important auxiliary data such as sequence accession numbers, full scientific names of organisms, and literature references, a very useful feature.

Results

Pawlowski *et al.* (1995) studied the phylogenetic location of phylum Foraminifera among eucaryotic lineages comparing large subunit ribosomal RNA sequences of 20 eucaryotic species. The infered neighbor-joining tree is shown (Figure 3). 1000 bootstrap replicates were performed. The authors addressed the question of whether Foraminifera diverged before the large evolutionary radiation of Animalia (*Rattus* and *Herdmania*), Plantae (*Arabidopsis*), Fungi (*Saccharomyces*), Ciliophora (*Tetrahymena*) and Dinoflagellata (*Prorocentrum*)—the so-called eucaryotic tree 'crown'. No straightforward answer could be deduced from the bootstrap analysis: three internal branches separate the Foraminifera lineage from the 'crown', each of them being associated to weak bootstrap values (a, b, c in Figure 3). The authors therefore tested the questioned relationship using a reduced five-species data set on which Li's (1989) branch length significance test could be employed. The test

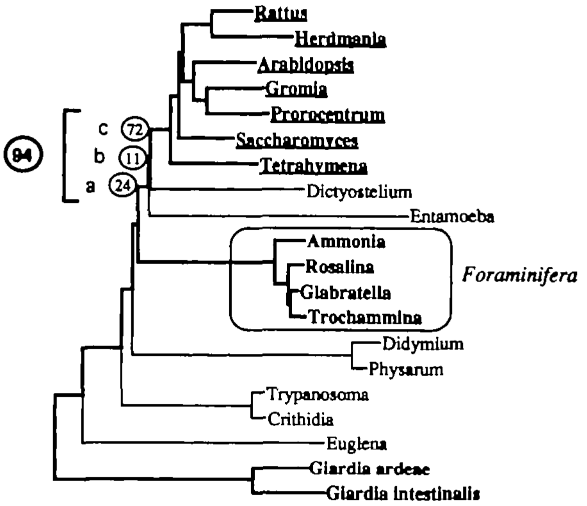


Fig. 3. Phylogenetic tree of 20 eukaryotic species recovered from partial large subunit ribosomal RNA sequences by the neighbor-joining method (Kimura's 1980 distance). 610 unambiguously aligned sites were used. 100 bootstrap replicates were performed. When all 20 species are considered, internal branches a, b, and c separate Foraminifera from the eukaryotic 'crown' (underlined species), with low bootstrap support. When the branching orders of bold species only are taken into account, 94% of bootstrap trees support the Foraminifera + Giardia versus 'crown' clustering (redrawn after Pawlowski *et al.*, 1994).

significantly supported an early divergence of phylum Foraminifera. This data-reducing strategy has drawbacks: the results may be dependent on the species chosen to represent a taxonomic group. Indeed, if genera *Giardia*, *Crithidia*, *Ammonia*, *Tetrahymena* and *Rattus* are chosen—rather than *Giardia*, *Crithidia*, *Ammonia*, *Tetrahymena* and *Prorocentrum* in the actual study—the relationship is no longer significantly supported.

We re-analyse the data using the 'branch combining' option of PHYLO_WIN. 1000 bootstrap replicates were performed with the NJ method (Kimura's two-parameter distance) using all 20 species, reproducing Pawlowski *et al.* analysis. Taxa subset S1 = 'root + Foraminifera' (*Giardia ardeae*, *Giardia intestinalis*, *Ammonia*, *Rosalina*, *Glabratella*, *Trochamina*) and S2 = 'crown' (*Rattus*, *Herdmania*, *Arabidopsis*, *Gromia*, *Prorocentrum*, *Tetrahymena*, *Saccharomyces*) were defined by first combining branches a, b, c and further removing species *Didymium*, *Physarum*, *Trypanosoma*, *Crithidia* and *Euglena* from the S1 subset. 94% of the 1000 20-species bootstrap trees included at least one internal branch separating subset S1 from subset S2, the phylogenetic location of the remaining seven species being ignored. The statistical significance of the phylogenetic relationship of interest was assessed here using the whole data set.

Discussion

SEAVIEW and PHYLO_WIN perform most usual tasks for phylogenetic studies, from sequence alignment to tree printing. User-friendly mouse-driven graphic interfaces and on-line help make them easy to use. In comparison to the sequence alignment editor of the GDE package (Smith *et al.*, 1994), SEAVIEW is original in offering a dot-plot guided alignment strategy and a full interface with the CLUSTALW program, that is, CLUSTALW-aligned sequence portions can be automatically inserted in the full multiple alignment. SEAVIEW is also noticeably faster than GDE when scrolling around sequences and usable with most if not all brands of unix workstations. Grouping several tree-building algorithms into a single environment makes possible to simultaneously deal with sequences and trees. PHYLO_WIN provides such a highly integrated environment, exclusively dedicated to molecular phylogeny. Particularly, the effects on the recovered tree of variable tree-making strategies—changes in alignment, selected sites and species, and/or tree-making methods—can be rapidly checked.

For a given gene, the data-aligned sequences, names and comments—and phylogenetic knowledge about these data—taxonomic species groups, conserved regions and inferred trees—are stored into a single MASE+ file. Multiple sequence, distance matrix, and tree files are therefore avoided. This is of special interest for ribosomal RNA sequences. In these molecules, distinct regions have distinct evolutionary rates. A careful choice of the analyzed sites is needed to remove variable regions that cannot be unambiguously aligned. The selected regions depend on the amount of divergence between the compared sequences so that a specific set of sites is required for a given group of species. A common storage of this information is thus greatly helpful for phylogenetic studies.

System, requirements and availability

SEAVIEW and PHYLO_WIN were written in ANSI C with two X Windows-based interface construction toolkits: XForms (Zhao and Overmars, 1995) for SEAVIEW and Vibrant (Kans, 1993) for PHYLO_WIN. The source code and executable versions for the following Unix systems are available: Sun (under SunOS 4 or Solaris 2), DEC-Alpha, IBM-RISC stations under AIX, Silicon Graphics, Hewlett Packard. Both may be installed on other Unix computers provided X Windows, XForms, and Vibrant are available. PHYLO_WIN may also be compiled on PCs or MacIntosh computers as the Vibrant libraries are available for these systems.

Acknowledgements

This work was supported by grants from the Groupement de Recherche et d'Etude sur le Génome. We thank Joe Felsenstein and Gary Olsen for freely providing us the C-

C-code of their programs, and Marc Robinson and other beta-testers for useful comments about the softwares.

References

De Rijk,P., Van De Peer,Y. and De Wachter,R. (1996) Database on the structure of large nbosomal subunit RNA. *Nucl. Acids Res.*, **24**, 92–97.

Efron,B. (1982) *The Jacknife, the Bootstrap, and Other resampling Plans*, CBMS-NFS Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia.

Faulkner,D.V. and Jurka,J. (1988) Multiple sequences alignment editor (MASE). *Trends Biochem. Sci.*, **13**, 321–322.

Felsenstein,J. (1985) Conference limit on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein,J. (1993) *PHYLIP: Phylogeny Inference Package*, version 3.5 University of Washington, Seattle, W.A.

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

Galtier,N. and Gouy,M. (1995) Inferring phylogenies from sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA.*, **92**, 11317–11321.

GCG (1994) Program Manual for the Wisconsin Package, Version 8, September 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed), *Mammalian protein metabolism*. Academic press, New-York, pp. 121–123

Kans,J. (1993) NCBI Software Development Toolkit, Version 1.8. NCBI, National Library of Medecine, Bethesda, MD.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Lake,J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Nat. Acad. Sci. USA*, **91**, 1455–1459.

Li,W.-H. (1989) A statistical test of phylogenies estimated from sequence data. *Mol. Biol. Evol.*, **6**, 424–435.

Li,W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. *J. Mol. Evol.*, **36**, 96–99.

Li,W.-H. and Graur,D. (1991) in Fundamentals of molecular evolution pp 55–56. Sinauer, Sunderland, Mass.

Li,W.-H., Wu,C.-I. and Luo,C.-C. (1985) A new method for estimating synonymous and non-synonymous rates of nucleotide substitutions considering the likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.

Olsen,G.J., Matsuda,H., Hagstrom,R. and Overbeek,R. (1994) fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci*, **10**, 41–48.

Pawlowski,J., Bolivar,I., Guiard-Maffia,J. and Gouy,M (1994) Phylogenetic position of Foraminifera inferred from LSU rRNA gene sequences. *Mol Biol. Evol.*, **11**, 929–938.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rzhetsky,A. and Nei,M. (1992) Statistical properties of the ordinary least–squares, generalized least–squares and minimum evolution methods of phylogenetic inference. *J. Mol. Evol.*, **35**, 367–375.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P.M. (1994) The Genetic Data Environment: an expandable GUI for multiple sequence analysis. *Comput. Applic. Biosci.*, **10**, 671–675.

Steel,M.A. (1993) Recovering a tree from the leaf colorations it generates under a Markov model. *Appl. Math. Lett.*, **7**, 19–23.

Swofford,D.L. (1993) *PAUP· Phylogenetic Analysis using Parsimony (PAUP)*, version 3.1 University of Illinois, Champaign

Tajima,F. and Nei,M. (1984) Estimation of evolutionary distances between nucleotide sequences. *Mol. Biol. Evol.*, **1**, 269–285.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.

Van De Peer,Y., Nicolai,S., De Rijk,P. and De Wachter,R. (1996) Database on the structure of small ribosomal subunit RNA. *Nucl. Acids Res.*, **24**, 86–91.

Yang,Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, **39**, 105–111.

Zhao,T.C. and Overmars,M. (1995) *Forms library. a graphical user interface toolkit for X*. University of Wisconsin-Milwaukee, Milwaukee.

Received on July 11, 1996; accepted on August 28, 1996