SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building

Manolo Gouy,*[1] Stéphane Guindon,[2,3] and Olivier Gascuel[2]

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, Université Lyon 1, Université de Lyon, Villeurbanne, France
[2]Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS UMR 5506, Université Montpellier II, Montpellier, France
[3]Department of Statistics, University of Auckland, Auckland, New Zealand

*Corresponding author: E-mail: mgouy@biomserv.univ-lyon1.fr.

Associate editor: Sudhir Kumar

Abstract

We present SeaView version 4, a multiplatform program designed to facilitate multiple alignment and phylogenetic tree building from molecular sequence data through the use of a graphical user interface. SeaView version 4 combines all the functions of the widely used programs SeaView (in its previous versions) and Phylo_win, and expands them by adding network access to sequence databases, alignment with arbitrary algorithm, maximum-likelihood tree building with PhyML, and display, printing, and copy-to-clipboard of rooted or unrooted, binary or multifurcating phylogenetic trees. In relation to the wide present offer of tools and algorithms for phylogenetic analyses, SeaView is especially useful for teaching and for occasional users of such software. SeaView is freely available at http://pbil.univ-lyon1.fr/software/seaview.

Key words: SeaView, multiple sequence alignment, molecular phylogeny, PhyML, graphical user interface.

Multiple alignment and phylogenetic tree reconstruction from molecular sequence data are key tasks for many molecular evolution analyses. They involve the sequential use of several programs that perform part of the complete procedure and often require series of tedious and error-prone data reformatting to transfer sequences and trees between these programs. The computer programs SeaView and Phylo_win pioneered the use of graphical user interfaces for performing multiple sequence alignment and phylogenetic tree reconstruction (Galtier et al. 1996). These programs have been widely used but were lacking access to recently developed methods for maximum-likelihood tree estimation. We present here SeaView version 4, a program that allows its users to perform the complete phylogenetic analysis of a set of homologous DNA or protein sequences, from network-based sequence extraction from public databases to tree building and display using up-to-date alignment and maximum-likelihood tree-building algorithms (fig. 1).

SeaView can read and write the most widely used file formats defined for holding aligned or unaligned protein or nucleotide sequence data: Fasta (Pearson and Lipman 1988), interleaved Phylip (Felsenstein 1993), Clustal (Higgins et al. 1992), multiple sequence file of the Genetics Computer Group package, Nexus (Maddison et al. 1997), and Mase (Faulkner and Jurka 1988). The last two formats allow for much useful information besides sequence and name, that is, trees, species, and sites selections, sequence annotations. SeaView can also import sequence data from the major public sequence databases using a network access (Gouy and Delmotte 2008) to daily (for GenBank and EMBL) or weekly updated (for UniProt/SwissProt) databases. Imported sequences can be identified by name, accession number, or keyword and named either with their database identifier or using the species name of their organism of origin. SeaView can also

directly import from nucleotide databases most feature table elements (e.g., coding sequence, rRNA, and non-coding RNA) and select those whose annotations contain a user-given character string (fig. 2).

Nucleotide sequences can be translated to protein using any user- or database-assigned genetic code, so operations such as alignment and tree building can be performed at the nucleotide or the protein levels. Unaligned protein-coding DNA sequences can be translated to protein, aligned, and displayed back as DNA sequences, a procedure that yields more realistic coding sequence alignments than would result from nucleotide-level alignment. Protein-coding DNA sequences can also be displayed by assigning the same color to all synonymous codons of the corresponding amino acid. Alignments can alternatively be displayed in reference mode, that is, where only residues that differ from the homologous one in a reference sequence are shown. Several sequence alignments can be handled simultaneously and copy/paste and concatenation operations can be performed between them. As far as display is concerned, SeaView accepts large sequence numbers (tens of thousand) and long sequences. SeaView is able to handle any number of sequence and site sets. Such sets can be named and saved in the Nexus or Mase file formats for subsequent use, by tree-building algorithms for instance.

SeaView relies on external programs to perform multiple sequence alignments. Two programs are initially available: ClustalW version 2 (Larkin et al. 2007) and Muscle (Edgar 2004). These programs are run with their default parameter values that have been chosen by their authors to perform well in most cases. When special parameter values are needed, they can be specified once using SeaView's user interface and reused for subsequent alignment operations. Alignment can be applied to all or to selected sequences
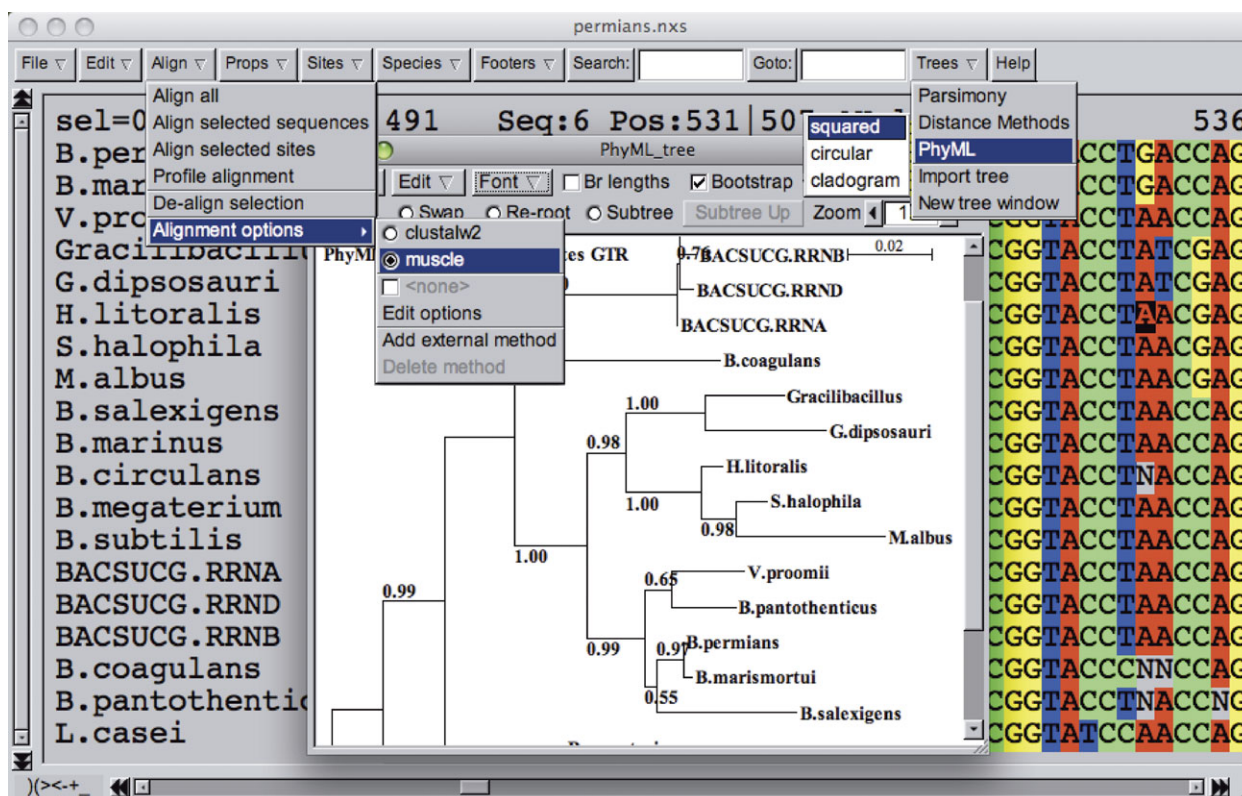
FIG. 1. The two major SeaView window types display sequence data and phylogenetic trees. Two of the displayed menus, two pilot multiple alignment algorithms and tree-building methods. Tree display tools allow printing, copy to clipboard, rerooting, zooming in and out, restricting display to a subtree, and the use of three alternative (squared, circular, and cladogram) tree-drawing formats.

or part of sequences. Profile alignment that aims at adding more sequences to a preexisting alignment can be done with both Muscle and ClustalW. SeaView is also able to drive any external sequence alignment program provided this program reads and outputs Fasta-formatted sequence data and can be run by a command line of the form "program_ name arguments." SeaView communicates with external alignment programs through a list of arguments that is initially defined by the user. This definition is made by entering once in a dialog box the list of arguments suitable for run-

ning this program, replacing the input file name by "%f.pir" and the output file name by "%f.out." The external alignment algorithm becomes directly usable after that step. For example, SeaView's interface to T-Coffee (Notredame et al. 2000) corresponds to the following argument list

$$\%f.pir -outfile = f\%.out -output = fasta\_aln$$
$$-outorder = input$$

which contains the arguments expected by T_Coffee to align a Fasta-formatted file and to output its Fasta-formatted results
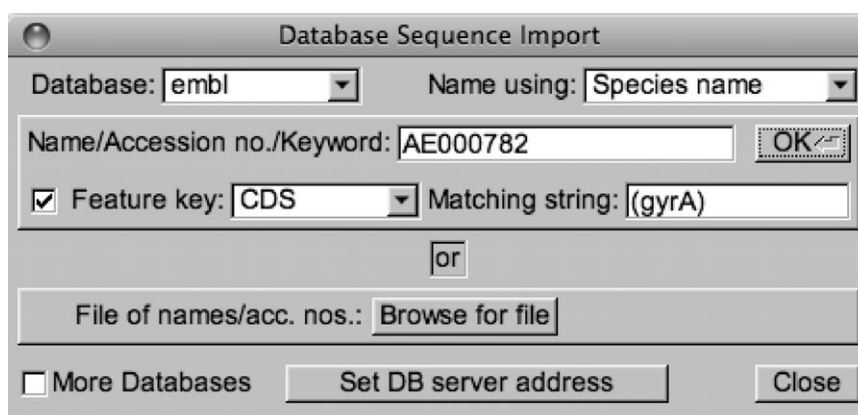
FIG. 2. Database sequence import dialog. This example will import into SeaView the single coding sequence (CDS) from EMBL's entry AE000782 (*Archaeoglobus fulgidus* complete genome sequence) containing the string (gyrA) in its database annotation and will name it *Archaeoglobus*.
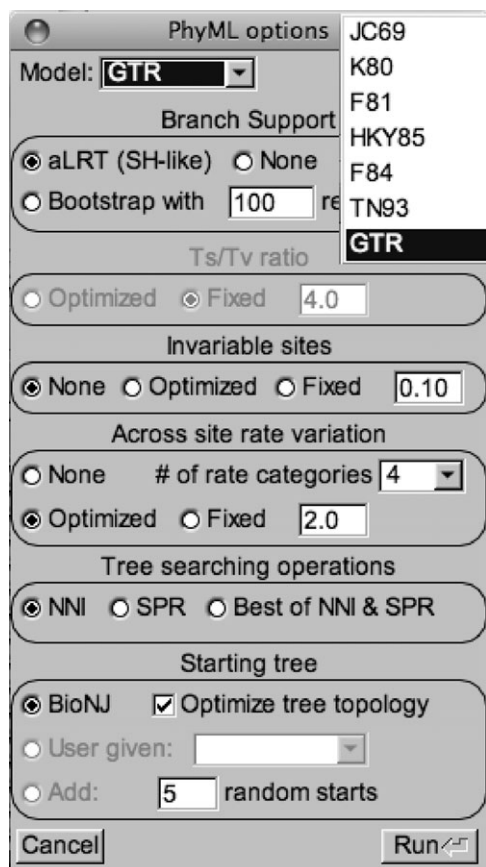
Fig. 3. SeaView dialog for setting PhyML options applied to nucleotide sequences.

without reordering sequences. Likewise, SeaView's interface to Probcons (Do et al. 2005) is straightforward:

$$\%f.pir > \%f.out$$

SeaView is also a multiple sequence alignment editor that can be used to add or remove one or several gaps in one or several sequences simultaneously. A dot-plot analysis (Maizel and Lenk 1981) can be performed between any two sequences to visually check whether alignment algorithms missed regions with high sequence similarity.

SeaView relies on PhyML version 3 (Guindon and Gascuel 2003) for maximum-likelihood phylogenetic tree reconstruction. Here again, PhyML is used as an independent program. Thus, future updates to PhyML will be accessible to SeaView users as soon as they will have installed the revised program. Tree building can be applied to all or to selected sequences and to all or selected sequence sites. Most PhyML options can be set through the graphical interface, both for nucleotide and protein-level analyses (fig. 3). Thus, branch support can be estimated either with the approximate likelihood-ratio test (Anisimova and Gascuel 2006) or by bootstrap resampling (Felsenstein 1985).

SeaView includes two distance-based tree reconstruction methods: Neighbor-Joining (Saitou and Nei 1987; Studier and Keppler 1988) and BioNJ (Gascuel 1997). These can be applied to various nucleotide and protein sequence

pairwise distances and combined with bootstrap resampling for branch-support estimation. Nucleotide-level distances are observed divergence, Jukes and Cantor, Kimura's two-parameter, Hasegawa–Kishino–Yano (see Rzhetsky and Nei 1995 for these first 4 distances), LogDet (Lake 1994), and Li's nonsynonymous (K_a) and synonymous (K_s) distances for protein-coding sequences (Li 1993). Protein-level distances are observed, Poisson and Kimura's (Nei 1987). Gap-containing sites are by default excluded from pairwise distance computations. Alternatively, sites that are gap-free in two sequences can be used to compute the distance for this sequence pair. The branch lengths of any user-given tree topology can be computed by minimization of the sum of squared differences between evolutionary and patristic distances (Rzhetsky and Nei 1993).

SeaView can also reconstruct maximum-parsimony phylogenetic trees using code extracted from Dnapars and Protpars programs (Phylip version 3.52, Felsenstein 1993). Parsimony computation can be combined with bootstrap resampling of sites and can be repeated a user-chosen number of times after randomly changing the input order of sequences. The parsimony score of any user-given tree can also be computed. SeaView completes parsimony analyses by computing the strict consensus of all equally parsimonious trees found.

When tree building is completed. SeaView draws the resulting phylogenetic tree on the screen (fig. 1). Plotted trees can be displayed with or without branch lengths (as when computed by parsimony), with or without branch-support values (typically, bootstrap scores or approximate Likelihood Ratio Test probabilities), binary or multifurcating, rooted or unrooted. Phylogenetic trees are initially rooted at the point in the tree that minimizes the variance of root-to-tip distances, but they can also be plotted as unrooted trees using a circular display or as cladograms containing topological but no branch-length information. The user can change the tree root and exchange the order of the two child lineages of a node. Trees can be saved to Newick, PDF, or PostScript files, and, under the Microsoft Windows and Mac OS X environments, printed or copied to the clipboard for communication with graphical tools such as Office applications. Aligned sequences can be reordered following their corresponding tree, and sequences that belong to a subtree can be selected in the corresponding alignment. Several tools such as subtree display, pattern matching in sequence names, and vertical zoom help dealing with large trees. A graphical tree editor allows topological changes by combining two basic operations: clade displacement and clade suppression.

SeaView version 4 is freely available at http://pbil.univ-lyon1.fr/software/seaview for four computer platforms (Microsoft Windows, Mac OS X, Linux, and SPARC/Solaris) and as source code.

Many software packages are available for multiple sequence alignment and phylogenetic tree reconstruction. SeaView is especially comparable with MEGA 4 that also provides an elaborate graphical user interface for multiple sequence alignment and distance or parsimony tree reconstruction

and display (Tamura et al. 2007). SeaView is less versatile than MEGA for pairwise distance computations and lacks features such as neutrality or molecular tests but is unique in being available for all major computer platforms and in allowing maximum-likelihood tree reconstruction with PhyML. Sea-View version 4 is especially valuable for teaching molecular phylogeny because of its availability at no fee for all users and because its user interface graphically expresses the conceptual steps involved in phylogenetic analyses. SeaView is also helpful for occasional users of phylogenetic tree reconstruction because it frees them from being confronted to many technical details concerning file formats and program options. SeaView thus pursues similar objectives to those of the phylogeny web server Phylogeny.fr (Dereeper et al. 2008) exploiting the user's computing resources. Because it performs, using PhyML, maximum-likelihood analyses at both nucleotide and protein levels, implements most current evolutionary models and computes statistical branch support, SeaView is also expected to be useful to seasoned phylogeneticists.

Acknowledgments

References

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol*. 55:539–552.

Dereeper A, Guignon V, Blanc G, et al. (12 co-authors). 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 36:W465–W469.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*. 15:330–340.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

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

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Felsenstein J. 1993. PHYLIP (phylogeny inference package) version 3.52. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*. 12:543–548.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Gouy M, Delmotte S. 2008. Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie* 90:555–562.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Higgins DG, Bleasby AJ, Fuchs R. 1992. CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci*. 8:189–191.

Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc Natl Acad Sci USA*. 91:1455–1459.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*. 36:96–99.

Maddison DR, Swofford DL, Maddison WP. 1997. NEXUS: an extensible file format for systematic information. *Syst Biol*. 46:590–621.

Maizel JV Jr, Lenk RP. 1981. Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci USA*. 78:7665–7669.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for multiple sequence alignments. *J Mol Biol*. 302:205–217.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*. 85:2444–2448.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol*. 10:1073–1095.

Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol Biol Evol*. 12:131–151.

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

Studier JA, Keppler KJ. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*. 5:729–731.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol Biol Evol*. 24:1596–1599.