BIOINFORMATICS APPLICATIONS NOTE Vol. 15 no. 4 1999
Pages 341–342

# MIAH: automatic alignment of eukaryotic SSU rRNAs

Patricia Thébault, Pierre Monestié, Annette McGrath and Desmond G. Higgins

Department of Biochemistry, University College, Cork, Ireland

Received on October 13, 1998; revised on January 17, 1999; accepted on January 22, 1999

## Abstract

***Summary:*** *MIAH is a WWW server for the automatic alignment of new eukaryotic SSU rRNA sequences to an existing alignment of 1500 sequences.*
***Availability:*** *http://chah.ucc.ie/MIAH*
***Contact:*** *des@chah.ucc.ie*

Ribosomal RNA sequences, especially the small subunit (SSU rRNA), are widely used in phyogenetic studies to estimate the relatedness of groups of organisms (e.g. Sogin *et al*., 1986). The SSU rRNA has been sequenced from thousands of different species and large alignments of these sequences are maintained at several sites (Maidak *et al.*, 1997; Van de Peer *et al*., 1998). The addition of new sequences to these alignments is very demanding. At present it must be done by a combination of manual and automatic means. Taking the eukaryotic subsection of a large reference alignment from the rRNA WWW server of the De Wachter group in Antwerp, Belgium (http://rrna.uia.ac.be/ssu/), we have recently empirically optimised parameters for the addition of new sequences to a profile (Gribskov *et al*., 1987) of this alignment through the use of sequence weighting and careful choice of parameters (O'Brien *et al.*, 1998). Here we describe the provision of a JAVA-based WWW server for the automatic alignment of new sequences to the reference alignment using these optimised weighting schemes.

An alignment (updated Jan 6 1998) of eukaryotic nuclear SSU rRNAs was obtained from the rRNA WWW server. After removal of the unaligned sequences *Babesia bovis 4* and *Polytoma oviforme*, the incomplete sequence *Butomus umbellatus* and 8 other sequences with ambiguous secondary structure annotation such as unequal numbers of bases in the two halves of a stem (*Lepidocyrtus paradoxus*, *Amphiscolops sp*., *Dekkera custeriana*, *Nitella axillaris*, *Nitellopsis obtusa*, *Coccoid haptophyte 2*, *Dientamoeba fragilis* and *Echinodorus cordifolius*) from the alignment and the removal of columns containing only gaps, the alignment contains 1570 sequences and is 5568 characters long. The alignment method we then provide is based on the construction of a profile (Gribskov *et al.*, 1987) from the existing alignment to which the new sequence is aligned by dynamic programming using Gotoh's algorithm (Gotoh, 1982). Computationally, this is similar to the alignment of 2 sequences to each other and uses scores for matching and mismatching residues and gap opening and extension penalties. Using a profile, also allows the use of sequence weighting (e.g. Thompson *et al.*, 1994) and position specific gap penalties.

The MIAH package consists of 2 parts, the client and the server. The server software is written in Java 1.14 with some native functions written in C to minimize the time taken for the large calculations necessary for the alignment calculation. The client software is written entirely in Java 1.02. The client software consists of 3 Java applets; the MIAH applet, and the Tree and Structure Viewer applets.

Sequence entry (FASTA format, Pearson and Lipman, 1988) is via the MIAH applet where alignment parameters such as gap opening and extension penalties may be set. There is a choice of four sequence weighting schemes for the construction of the profile; no weights; tree weights (Thompson *et al.*, 1994) which correct for unequal spread of taxa across the underlying phylogenetic tree of the sequences; identity weights which give most weight to sequences in the alignment which are similar to the test sequence and combination weights (O'Brien *et al.*, 1998) which are a combination of the last two. Combination weights were previously shown to work best for alignment of rRNA sequences. For convenience, the large reference alignment has been subdivided into its 4 kingdoms: Animalia, Fungi, Plantae and Protoctista and therefore a choice of 5 profiles (4 kingdoms or all together) is presented for the calculation of the alignment. The sequence is returned in a second window with gaps inserted relative to the large alignment. These gaps specify an alignment between the sequence and the reference collection.

The newly aligned sequence may be approximately placed in one of four existing phylogenetic trees; one for each of the 4 eukaryotic kingdoms. Trees are calculated for the reference alignment in advance by the neighbour-joining method of Saitou and Nei (1987) with Kimura's two parameter distance correction (Kimura, 1980) using the DNADIST and NEIGHBOR programs of the PHYLIP package (Felsenstein, 1993). The position of the new sequence is estimated using a least squares procedure based on the distance be-

© Oxford University Press

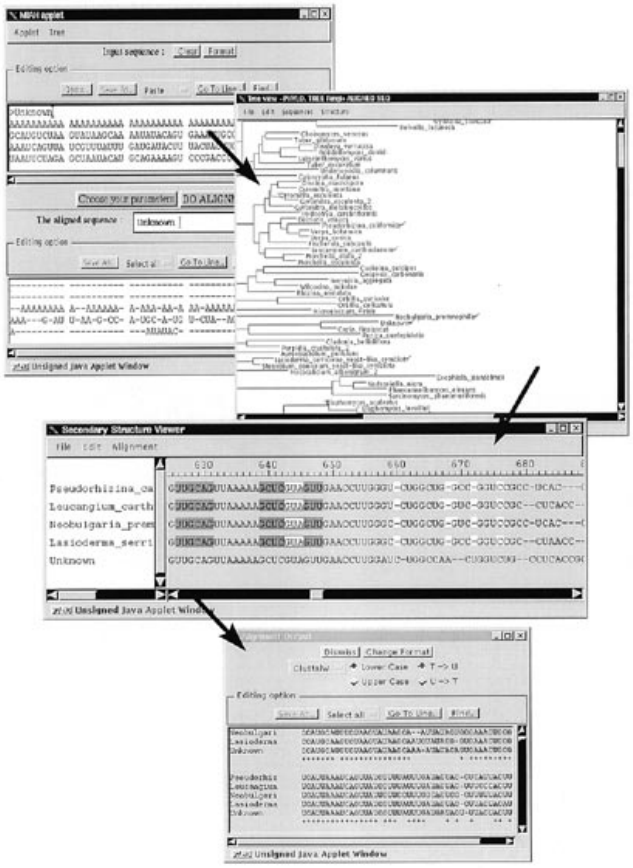

**Fig. 1.** Screenshots from the MIAH server showing parts of the main menu, treeviewer, alignment viewer and secondary structure viewer.

tween the sequence and each sequence in the reference alignment. Users are not permitted to place the newly aligned sequence into a tree where the distances between the species in the tree and the new sequence are too great. Each of the phylogenetic trees may also be viewed independently of the newly aligned sequence and the tree data may be saved in the Newick tree format. As the tree is quite large, a number of tools are provided to manipulate the graphical representation and to view the tree in subtree mode. From the Tree Viewer, sequences may be selected and retrieved. These sequences may also be viewed in the Structure Viewer and some inferences about the relative position of secondary structure elements may be drawn for the newly aligned sequence. In the Structure Viewer, the position of stems and pseudoknots for the retrieved sequences are displayed using matching blocks of colour. Information about the start and end points of specific structural elements can be gained by mouse actions. Features are provided for manipulating the appearance of the graphical display by changing the colour scheme or the sequence order. The selected sequences can also be output in an alignment viewer in a choice of output formats. The resulting alignment may also be mailed to the user as a text file for use in other programs. Figure 1.

## Acknowledgements

We wish to thank Manolo Gouy for showing us how to place a new sequence in a tree. This work was supported by a grant (BIO4-CT95-0130) from the EU Biotechnology Programme.

## References

Felsenstein,J. (1993) *PHYLIP, version 3.5c*. University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358

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

Maidak,B.L., Olsen,G.J., Larsen,N., Overbeek,R., McCaughey,M.J. and Woese,C.R. (1997) The RDP (Ribosomal Database Project) *Nucleic Acids Res.*, **25**, 109–111

O'Brien,E.A., Notredame,C. and Higgins,D.G. (1998) Optimisation of ribosomal RNA profile alignments. *Bioinformatics*, **14**, 332–341.

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

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

Sogin,M., Elwood,H. and Gunderson,J. (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc. Natl Acad. Sci. USA*, **83**, 1383–1387

Thompson,J., Higgins,D. and Gibson,T. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Applic. Biosci.*, **10**, 19–29.

Van de Peer,Y., Caers,A., de Rijk,P. and de Wachter,R. (1998) Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res.*, **26**, 179–182.