BIOINFORMATICS APPLICATIONS NOTE
Vol. 17 no. 6 2001
Pages 573–574

SOAP, cleaning multiple alignments from unstable blocks

A. Löytynoja and M. C. Milinkovitch*

Evolutionary Genetics, Free University of Brussels (ULB), cp 300, Institute of Molecular Biology and Medicine, rue Jeener & Brachet 12, B-6041 Gosselies, Belgium

Received on January 19, 2001; revised on February 27, 2001; accepted on March 8, 2001

ABSTRACT

Summary: SOAP is a stand-alone, multi-platform program to test the stability of a multiple alignment of molecular sequences.

Availability: The software is available free of charge at http://dbm.ulb.ac.be/ueg/SOAP/.

Contact: apl@dbm.ulb.ac.be; mcmilink@ulb.ac.be

Optimal multiple alignment of molecular sequences is to homology estimation what an exact search is to tree inference: an NP-complete problem. Hence, it is, and will always be, computationally impractical for more than a few sequences. Heuristics have therefore been developed in the form of progressive alignment methods (Feng and Doolittle, 1987), and the program CLUSTALW (Thompson *et al.*, 1994) is one of the most widely used profile-based variant of this method. The algorithm implemented in CLUSTALW: (i) constructs a distance matrix of all $N(N-1)/2$ pairs of sequences followed by approximate conversion of similarity scores to evolutionary distances; (ii) builds, on the basis of this distance matrix, a guide tree with NJ clustering, and (iii) progressively performs pairwise alignments of sequences and profiles at nodes in order of decreasing similarity.

Different portions of a given nucleic acid or amino acid sequence fragment can evolve at very different rates (both in terms of insertion/deletion and substitution of characters) such that the efficiency with which homology is inferred across sequences can vary greatly. Furthermore, different alignments produced through the use of even slightly different heuristic alignment parameters (such as gap/substitution costs) can yield significantly different results when subjected to phylogenetic analyses. Various methods for testing the reliability and consistency of alignments have been developed (e.g. Mevissen and Vingron, 1996; Notredame *et al.*, 1998, and references therein). The complexity of these elegant methods has prevented their frequent use for testing the reliability of alignments used in phylogenetic analyses. Recently, rough estimation of alignment stability has been assessed through comparison of alignments produced under a few different gap opening/extension penalties (Gatesy *et al.*, 1993). This conservative approach assumes that confidence in homology assessment increases with stability to variation in alignment parameters. Very few authors (e.g. Milinkovitch *et al.*, 1998) have applied the procedure as it is extremely tedious and error-prone when performed manually. Hence, we implemented it in a computer program, SOAP, which produces and compares CLUSTALW alignments.

SOAP can be used to define the interval for two alignment parameters (gap opening and gap extension penalties) and to produce an alignment for each of the possible combinations of parameters within that defined parameter space (Figures 1 and 2). As it can also compare alignments including different sets of sequences, the stability of alignment to removal/addition of taxa/sequences can also be investigated. The sequences are aligned by the standard CLUSTALW algorithm, and they can either be produced locally or be distributed on several CPUs on a remote server. After the alignments have been produced, SOAP identifies the ‘unstable-hence-unreliable’ aligned columns by comparing a chosen set of alignments against a user-defined reference alignment. The reference alignment is shown in a scrollable window with color-coded character states, and the instability of each column (= the proportion of non-reference alignments differing, for that position, from the reference alignment) is indicated on a histogram header (Figure 2). A consensus level can be chosen and the sites showing less than the defined threshold level of stability are automatically highlighted. The alignments being compared can be changed at any time by selecting/deselecting them in a separate window. Any column (aligned position) and row (sequence) can also be manually selected/deselected. The full reference alignment can be saved in Nexus or in plain text formats with the highlighted positions and taxa defined in a separate Paup-command-block or as a stability vector,

*To whom correspondence should be addressed.

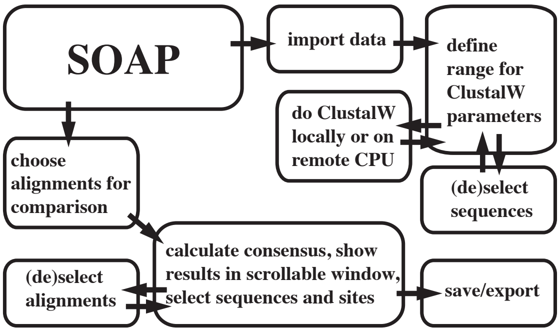
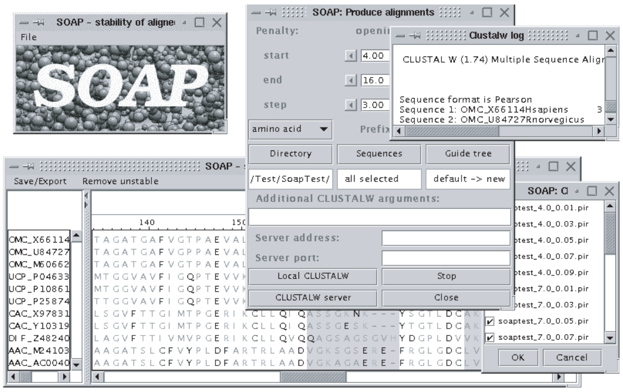

SOAP
import data
define range for ClustalW parameters
do ClustalW locally or on remote CPU
(de)select sequences
choose alignments for comparison
calculate consensus, show results in scrollable window, select sequences and sites
(de)select alignments
save/export

Fig. 1. General architecture of SOAP. CLUSTALW alignments are produced locally or on a server; for each aligned position in the reference alignment, SOAP calculates the stability among all selected alignments; the consensus alignment can be saved/exported in various commonly used formats.



start 4.00
end 16.0
step 3.00
amino acid
/Test/SoapTest/
all selected
default -> new
Server address:
Server port:

```
CLUSTAL W (1.74) Multiple Sequence Align

Sequence format is Pearson
Sequence 1: OMC_X66114Hsapiens          3
Sequence 2: OMC_U84727Rnorvegicus
```

```
OMC_X66114  TAGATGAFVGTPAEVAL
OMC_U84727  TAGATGAFVGPPAEVAL
OMC_M60662  TAGATGAFVGTPAEVAL
UCP_P04633  MTGGVAVFIGQPTEVVK
UCP_P10861  MTGGVAVFIGQPTEVVK
UCP_P25874  TTGGVAVFIGQPTEVVK
CAC_X97831  LSGVFTTGIMTPGERIKCLLQIQASSGKNK---YSGTLDCAK
CAC_Y10319  LSGVFTTGIMTPGERIKCLLQIQASSGESK---YTGTLDCAK
DIF_Z48240  LAGVFTTIVMVPGERIKCLLQVQQAGSAGSGVHYDGPLDVVK
AAC_M24103  AAGATSLCFVYPLDFARTRLAADVGKSGSERE-FRGLGDCLV
AAC_AC0040  AAGATSLCFVYPLDFARTRLAADVGKAGAERE-FRGLGDCLV
```

test_4.0_0.01.pir
test_4.0_0.03.pir
test_4.0_0.05.pir
test_4.0_0.07.pir
test_4.0_0.09.pir
test_7.0_0.01.pir
test_7.0_0.03.pir
☑ soaptest_7.0_0.05.pir
☑ soaptest_7.0_0.07.pir

Fig. 2. The SOAP main windows.

respectively. The reference alignment *minus* the currently selected subset of sites and sequences can be exported in NBRF/PIR, Molphy or Phylip formats.

SOAP is written in a rather ‘conservative’ Java 2 such that the version for MacOS is easily modified to compile under Java 1.1.8. The graphical user interface relies on the recent JFC/Swing library, and supports each platform specific ‘Look & Feel’. SOAP also integrates a CLUSTALW (Thompson *et al.*, 1994) multiple alignment procedure that is written in C. The alignment algorithm itself of CLUSTALW was left unmodified while some code was changed to allow porting to Java. The slightly modified CLUSTALW (originally version 1.74) code was recompiled as a shared object file and is called through the Java Native Interface (JNI). The binary code for the modified CLUSTALW object file is provided for Linux, PPC MacOS, and MS Win32 platforms, and compiling the code on other UNIX-like computers is straightforward. As an alternative to running all tasks locally, SOAP can be requested to send the computation-intensive tasks of producing alignments on a remote server. The server/client connection is created with sockets, and the native CLUSTALW program is executed on the server with Java Runtime.exec() method. In that case, SOAP works as a client and a simple Java wrapper is provided for the UNIX-like computer acting as the server.

ACKNOWLEDGEMENTS

We are grateful to Daniel Van Belle for helpful discussions and advice during the development of the program. This work is supported by the National Fund for Scientific Research Belgium (FNRS), the Free University of Brussels (ULB), the ‘Communauté Française de Belgique’ (ARC 98/03-223), the Defay Fund, the Van Buuren Fund, and the Finnish Cultural Foundation.

REFERENCES

Feng,D.-F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.

Gatesy,J., DeSalle,R. and Wheeler,W. (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phyl. Evol.*, **2**, 152–157.

Mevissen,H.T. and Vingron,M. (1996) Quantifying the local reliability of a sequence alignment. *Protein Eng.*, **9**, 127–132.

Milinkovitch,M.C., Bérubé,M. and Palsbøll,P.J. (1998) Cetaceans are highly specialized artiodactyls. In Thewissen (ed.), *The Emergence of Whales: Evolutionary Patterns in the Origin of Cetacea*. Plenum, New York, pp. 113–131.

Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.

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