Nucleic Acids Research, 2003, Vol. 31, No. 13 **1–4**
DOI: 10.1093/nar/gkg522

Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments

Olivier Poirot, Eamonn O'Toole[1] and Cedric Notredame*

Information Genetique et Structurale, CNRS, 31 Chemin Joseph Aiguier, 13 402 Marseille Cedex 20, France and [1]High Performance Technical Computing Division, Hewlet Packard, BallyBrit, Galway, Ireland

Received February 14, 2003; Revised and Accepted March 17, 2003

ABSTRACT

This paper presents Tcoffee@igs, a new server provided to the community by *Hewlet Packard* computers and the *Centre National de la Recherche Scientifique*. This server is a web-based tool dedicated to the computation, the evaluation and the combination of multiple sequence alignments. It uses the latest version of the T-Coffee package. Given a set of unaligned sequences, the server returns an evaluated multiple sequence alignment and the associated phylogenetic tree. This server also makes it possible to evaluate the local reliability of an existing alignment and to combine several alternative multiple alignments into a single new one. Tcoffee@igs can be used for aligning protein, RNA or DNA sequences. Datasets of up to 100 sequences (2000 residues long) can be processed. The server and its documentation are available from: http://igs-server.cnrs-mrs.fr/Tcoffee/.

INTRODUCTION

Over the years, multiple sequence alignment methods have been established as a key component of biological sequence analysis techniques. Few procedures remain in bioinformatics that do not require at one point or another the assembly of a high quality multiple sequence alignment. One could cite in bulk: the identification of a protein signature such as a PROSITE pattern (1), the building of a domain profile (or HMM) needed for identifying the most remote members of a protein family (2), structure prediction (3) and phylogenetic analysis (4). More recently, multiple sequence alignments have also proven useful to the characterization of nsSNP (the non-synonymous single nucleotide polymorphisms) (5,6). Despite the importance of these applications, the design of an efficient and accurate algorithm for the assembly of multiple sequence alignments remains a difficult problem that has not yet been entirely solved. As a consequence, most of the available packages merely provide approximate solutions (for recent reviews on this problem, see 7 and 8). Furthermore, none of these methods is consistently better than the others. For instance, systematic benchmarking experiments carried out with established collections of reference alignments (9) have shown that each available package is better suited than the others to certain types of problems but that none is always the best (10). This situation explains why from one bioinformatics project to the next, the authors often use a different multiple sequence alignment package. Unfortunately it is usually difficult to determine which software or algorithm will work best on a given set of sequences and the only way to address this problem is through a tedious trial and error process.

This complicated situation makes it critical to ensure that non-specialists have full access to state-of-the-art resources dedicated to multiple sequence alignments. The server we introduce here offers access to the latest version of T-Coffee (11), a recent multiple sequence alignment method. The web interface makes it possible to put some emphasis on important options of this program that are normally buried in the command line syntax. Another advantage of this server is that the people who maintain it also develop the T-Coffee package. This insures that the server always runs the latest version of the package. Several tutorials are available for T-Coffee including one in the book 'Bioinformatics for dummies' (12) and another one in preparation for *Current Protocols in Bioinformatics*.

METHODS

Multiple sequence alignment assembly

T-Coffee is a multiple sequence alignment method. Given a set of pre-selected protein or DNA sequences, T-Coffee computes a multiple sequence alignment of the input sequences. To that effect T-Coffee starts by computing a collection of pair-wise alignments: for each possible pair of sequences in the dataset, the program computes the best global alignment and the 10 best local alignments [using the Sim algorithm from the Lalign package (13)]. This collection of pairwise alignments is named a library in the T-Coffee jargon. In the second step of the procedure T-Coffee assembles a multiple sequence alignment that has the highest possible level of consistency with the alignments within the library. T-Coffee is only a heuristic and the optimality of this process is not guaranteed, although the results are usually satisfactory as judged by comparison with

*To whom correspondence should be addressed. Tel: +33 ???; Fax: +33 ???; Email: cedric.notredame@europe.com

Q1

alternative optimization methods (14). To assemble its alignment, our package uses a progressive strategy similar to the one described for ClustalW. Extensive details on this procedure are available in the original publication (11).

It is important to point out that in T-Coffee, the construction of the library affords many possibilities. Although the strategy described in the original paper relies on a combination of local and global alignments, one may also use different types of libraries. For instance, the most obvious alternative would be to fill the library with multiple sequence alignments generated with various methods or various parameters setting. This is an option we now give through our new interface (combining alignments). It is also possible to construct the library using structural information rather than sequence. This is exactly the strategy adopted for the assembly of the multiple sequence alignments in the DALI Domain Dictionary (15) where libraries are produced using the DALI structure superposition algorithm. Such an option will soon be available on the server we describe here. Users are also encouraged to download and install the package locally in order to test their own recipes: T-Coffee has been designed to seamlessly turn any type of sequence alignment into the kind of libraries it needs.

Given a library, it is also possible to evaluate the consistency between a multiple alignment and every pair of aligned residues contained in this library. This measure of consistency indicates the 'support' of the library for the alignment. It can be measured for the complete alignment or at the local level for every individual residue. The local measure is named the CORE index (consistency of overall residue evaluation). Some properties of this index were recently characterized using an established collection of reference alignments (9). These analyses indicate that residues with a core index of ≥ 5 (on a scale 0–9) have 90% chance of being correctly aligned as judged from their reference structural alignments. This information on the alignment reliability can conveniently be used in order to remove from the alignment the portions that are incorrectly aligned. This may help enhance the sensitivity of protein domain profiles or the accuracy of phylogenetic tree reconstructions.

Various recent studies have shown that T-Coffee is one of the most accurate multiple sequence alignment packages available today, for protein and nucleotide sequence alignments alike (10,16). These studies all show that T-Coffee is notably more accurate than its close relative ClustalW. This increased accuracy comes at a price and T-Coffee is N times more expensive in terms of CPU time than ClustalW with a time complexity in the order of $O(N^3L^2)$ (L being the length of the sequences and N being the number of sequences). With datasets <30 sequences this difference is barely noticeable, but high performance hardware is needed for datasets >100 sequences.

THE Tcoffee@igs SERVER

The IGS (Information Génetique et Structurale) has developed the most powerful T-Coffee server to date. This server runs on an Alpha ES45 quadriprocessor, kindly provided by HP. This powerful server supports the analysis of a maximum number of 100 sequences with a maximum of 2000 residues each. The homepage of the server (igs-server.cnrs-mrs.fr/Tcoffee/) contains pointers to the three types of calculation performed: computing, evaluating or combining multiple sequence alignments. The fourth section points to the online T-Coffee documentation. For each option, one has a choice between the regular and the advanced mode. Regular gives access to a very straightforward interface where the user simply needs to paste the data and retrieve the results. Advanced offers more possibilities for setting various computation and output parameters. It should be pointed out that the Swiss EMBnet node also maintains an alternative T-Coffee server: *www.ch.embnet.org*.

Computing a multiple alignment

Sequences must be pasted using the FASTA format. In the regular mode, the library is computed using global and local alignments. In the advanced mode, the user is free to change this and can also request the incorporation of a ClustalW multiple sequence alignment within the library (by checking the clustalw_aln box). As time goes on (and upon request), new methods will be incorporated in the method section of the advanced part.

When the computation is finished, T-Coffee returns a table similar to the one shown in Figure 1. This table gives a pointer to the multiple alignment in its raw form or in a colored format. The table also gives access to the phylogenetic tree. This tree is not a guide tree (i.e. a pre-estimation of the phylogenetic tree computed from unaligned sequences) but a genuine phylogenetic tree computed with the multiple sequence alignment produced by T-Coffee. The procedure involves feeding ClustalW (17) with the T-Coffee alignment and running ClustalW from the command line with the following parameters:

$$\text{clustalw -infile} = \text{tcoffee\_alignment -tree -bootstrap} = 100 \text{ -tossgaps}$$

In this mode (-tree), ClustalW computes the phylogenetic tree associated with an alignment without re-computing the alignment. The -tossgaps parameter causes ClustalW to remove from the alignment every column that contains a gap before computing a Neighbor Joining tree (18) on the remaining columns. The -bootstrap flag indicates that 100 bootstrap cycles are performed to assess the reliability of the tree (cf. ClustalW documentation).

In the table of Figure 1, html and pdf point to a color-coded evaluated version of the T-Coffee multiple alignment (Fig. 2). In this colored representation, blue portions and green portions correspond to inconsistent bits, unlikely to be correctly aligned, while the yellow, orange and red bits correspond to the most consistent portions, much more likely to be correctly aligned. In this alignment the level of conservation column by column is also indicated using the ClustalW notation of a '$*$' for completely conserved column, a ':' for highly conserved ones and a '.' for the less conserved ones. T-Coffee can also produce an ASCII version of the evaluated alignment (request the score_ascii output from the advanced menu). In the score_ascii format, residues are recoded using a 0–9 index (0 corresponds to blue bits and 9 to the red ones). This ASCII

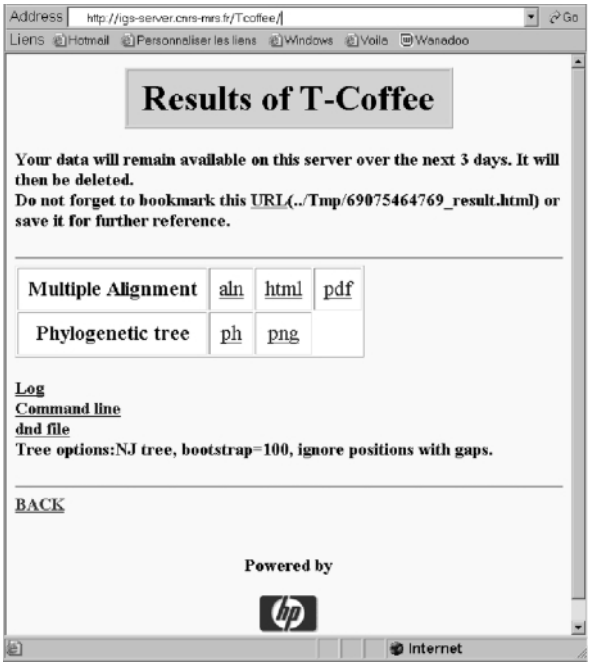

Figure 1. Typical output of a standard T-Coffee computation. The results can be retrieved at the indicated URL for up to 3 days. ph, a pointer to the phylogenetic tree in Newik format; png, a graphic display of this same tree. Command line indicates the exact command line used by T-Coffee to compute the alignment.

Figure 2. Colored output of T-Coffee. The first lines indicate the average CORE index associated with every sequence. In the rest of the alignment, red residues correspond to highly reliable portions of the multiple alignment. The Cons line is a consensus, it indicates the average reliability value for every column.

version can be used for automatically filtering unreliable portions of your alignment (a small package will soon be made available on our server for this purpose).

Evaluating a multiple alignment

It is possible to use the Tcoffee@igs server in order to evaluate pre-computed multiple sequence alignments. In this case, the user simply needs to input a pre-computed alignment in any of the following formats: ALN (the ClustalW output format), MSF or FASTA. T-Coffee automatically computes the corresponding library and outputs a colored version of the alignment. The final colored output is similar to the one shown in Figure 2 and discussed in the previous section.

Combining alternative multiple sequence alignments into a consensus multiple alignment

The third option of the T-Coffee server makes it possible to use the package in order to combine several alternative alignments into one. The color-coded evaluated version makes it possible to identify the portions of the consensus alignment that occurred in many of the input alignments (red) and those that are poorly represented (blue). In the current configuration, it is possible to combine up to six multiple alignments (two being the minimum). Upon request to the authors, this number will be increased.

It is worth noting that the server is very flexible with the nature and the state of the input sequences. While the program assumes that two sequences with the same name coming from two different alignments are indeed the same sequence, T-Coffee does not require these sequences to come in the same order or to be perfectly identical. If a discrepancy occurs, T-Coffee aligns the conflicting sequences and reconstructs a master sequence based on the consensus given by the reconciliation alignment (see online documentation for further details). While they can be deceptive when working with data of poor quality, these facilities make it easier to work with structural data (where the sequences agreement is not always perfect) or local multiple alignments obtained from a database search where identical sequences may slightly differ on their extremities. It is also possible to combine alignments that do not contain the same number of sequences.

CONCLUSION AND FUTURE DEVELOPMENTS

In this paper we describe Tcoffee@igs, a new multiple sequence alignment server. Tcoffee@igs makes it possible for non-specialists to use the T-Coffee package in a simple and intuitive fashion in order to produce high quality multiple sequence alignments. This server also makes it possible to evaluate existing multiple sequence alignments or to combine them into a consensus alignment.

Future developments will include enhanced abilities regarding the mixing of sequences and structures. New modules are under development for this server that will smoothly allow the combination of sequences and structures within the framework of a multiple sequence alignment.

While the most widely accepted means of validating a new multiple sequence alignment method is its benchmarking with a collection of reference alignments, it has become obvious over the years that the only efficient (albeit highly empirical) means of carrying out such an evaluation is the feedback of biologists: how good was the method in their hand and how well did it fare on proteins or nucleic families they know well? A web interface reduces the technological barrier between biologists and the use of bioinformatics tools, making it easier

to collect such feedback. We strongly encourage users of this server to let us know about their impressions, good or bad!

REFERENCES

1. Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.
2. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
4. Phillips,A., Janies,D. and Wheeler,W. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.*, **16**, 317–330.
5. Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
6. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
7. Notredame,C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.
8. Duret,L. and Abdeddaim,S. (2000) In Higgins, D. and Taylor, W. (eds), *Bioinformatics, Sequence, Structure and Databanks.* Oxford University Press, Oxford.

Q3

9. Notredame,C. and Abergel,C. (2002) In Andrade, M. (ed.), *Bioinformatics Methods for Genome Analysis*, in press.

Q4

10. Lassmann,T. and Sonnhammer,E.L. (2002) Quality assessment of multiple alignment programs. *FEBS Lett.*, **529**, 126–130.
11. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
12. Claverie,J.M. and Notredame,C. (2003) *Bioinformatics for Dummies.* Wiley Publishing, Inc.

Q5

13. Huang,X. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, **12**, 337–357.
14. Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.
15. Dietmann,S., Park,J., Notredame,C., Heger,A., Lappe,M. and Holm,L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.*, **29**, 55–57.
16. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
17. Thompson,J., Higgins,D. and Gibson,T. (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**, 4673–4690.
18. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

OXFORD UNIVERSITY PRESS

JOURNAL: NAR MS NUMBER: 00259 ARTID: gkg522

TO: AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the editing of your manuscript. Please answer the queries by marking the required corrections at the appropriate positions in the text.

Q1	Please supply tel and fax numbers	
Q2	Is the O correct here, if so what does it mean?	
Q3	Please supply article title and page range for ref 8	
Q4	Please supply article title and page range for ref 9 and update if possible	
Q5	Please supply publisher’s location	