Published online 9 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W13–W17***
doi:10.1093/nar/gkr245

T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension

Paolo Di Tommaso[1], Sebastien Moretti[2,3], Ioannis Xenarios[2], Miquel Orobitg[4], Alberto Montanyola[4], Jia-Ming Chang[1], Jean-François Taly[1] and Cedric Notredame[1,*]

[1]Centre For Genomic Regulation (Pompeu Fabra University), Carrer del Doctor Aiguader 88, 08003 Barcelona, Spain, [2]Vital-IT, Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland, [3]Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and [4]Department of Computer Science and Industrial Engineering, University of Lleida, Campus de Cappont, C. de Jaume II 69, E-25001 Lleida, Spain

Received February 18, 2011; Revised March 23, 2011; Accepted April 5, 2011

ABSTRACT

This article introduces a new interface for T-Coffee, a consistency-based multiple sequence alignment program. This interface provides an easy and intuitive access to the most popular functionality of the package. These include the default T-Coffee mode for protein and nucleic acid sequences, the M-Coffee mode that allows combining the output of any other aligners, and template-based modes of T-Coffee that deliver high accuracy alignments while using structural or homology derived templates. These three available template modes are Expresso for the alignment of protein with a known 3D-Structure, R-Coffee to align RNA sequences with conserved secondary structures and PSI-Coffee to accurately align distantly related sequences using homology extension. The new server benefits from recent improvements of the T-Coffee algorithm and can align up to 150 sequences as long as 10 000 residues and is available from both http://www.tcoffee.org and its main mirror http://tcoffee.crg.cat.

INTRODUCTION

As judged by citation index, multiple sequence alignment (MSA) is one of the most widely used techniques in biology. Indeed the multiple comparisons of homologous sequences has applications in almost all fields of modern biology, from simple data monitoring up to sophisticated modeling-like structure prediction and phylogenetic reconstruction. In the past 20 years, more than 50 aligners have been published (1), and a wide diversity of choices that mostly reflects the lack of a universal method solving unambiguously the multiple sequence alignment problem. It is indeed a complex task that stands at the interface between computer science and biology. The biological problem is the definition of a mathematical formula (objective function) accurately quantifying the biological relationship between two sequences on the basis of their alignment. The computational problem is the estimation of an optimal model with respect to the objective function. In practice the objective functions described so far have difficulties accurately modeling the homology between protein sequences having <30% identity (70% in the case of nucleic acids). Yet, these functions are not only limited in accuracy but they are also difficult to optimize and it has been shown that, for the most commonly used functions, the computation of an optimal multiple sequence alignment is an NP-complete problem (2). The lack of an exact solution has prompted the development of a large number of heuristic solutions, either focused on the design of novel objective functions (3,4), the improvement of the optimization algorithm (5,6) or a trade-off between accuracy and speed (7).

T-Coffee (8) belongs to the class of aligners known as consistency based, which may be described as slow and accurate. These include ProbCons (9), PCMA (10), MAFFT (the slow accurate mode) (11), PROMALS (12) and Pecan (13). All the aligners of this class trade speed for increased precision. Over the years, they have been shown by many independent studies to outperform their simpler counterparts in terms of accuracy. While one may debate over the value of a modest but significantly increased accuracy at a sometimes prohibitive CPU cost, one should not overlook what is probably the main

*To whom correspondence should be addressed. Tel: +34 93 3160271; Fax: +34 93 3160099; Email: cedric.notredame@crg.eu

advantage of consistency-based protocols: their integrative capacity. In the consistency-based framework, the considered sequences are not directly integrated in a multiple sequence alignment. They are first aligned using any suitable combination of third-party aligners. The resulting collection of alignments (named a library in T-Coffee) is then turned into a multiple sequence alignment using a position specific scoring scheme derived from the library (consistency-based progressive algorithm). In practice the way the library is computed defines most of the variations around T-Coffee. The first version was using a combination of ClustalW all against all pair-wise alignments combined with Lalign all against all local alignments, the current version uses all against all pair-wise alignments computed with a pair HMM (9). There is no limit on how many methods and what kind of methods may be combined this way. One can even use pre-existing multiple sequence alignment methods, like in the M-Coffee protocol (14) where the library is made of a collection of multiple sequence alignment produced with third-party multiple aligners. This approach is becoming increasingly popular as it makes it possible to compare and combine the output of several aligners therefore simplifying the software selection dilemma. This possibility is important in a period where concerns are growing on the non-neutrality of the alignment methods towards subsequent modeling (15). When combining multiple sequence alignments, one can either combine all existing methods or only a selected subset. For instance, since 2009, the *compara* component of ENSEMBL uses M-Coffee to combine the output of three fast aligners (MAFFT, MUSCLE and Kalign) in order to produce the MSAs needed for the computation of the reference trees. On the server, users can use check-boxes to select the methods they want to combine. Our aim is to integrate as many public methods as possible and we welcome users requests for unsupported methods. We currently have an interface with eight popular aligners.

The most recent improvement in T-Coffee has been the development of the concept of template-based multiple sequence alignment (16,17). When run as a template-based aligner, T-Coffee uses a different procedure to generate the primary library: rather than directly aligning the sequences, it associates each input sequence with a template, it then aligns every pair of templates with an appropriate aligner and projects the resulting alignments onto the original sequences.

The new server offers three template-based alignment modes: one for RNA sequences [R-Coffee (18)], one for protein with a known structure [Expresso (17)] and one for the alignment of distantly related sequences [PSI-Coffee (1)]. R-Coffee uses as templates RNA secondary structure predictions obtained by applying the RNAplfold prediction algorithm onto the considered sequences. The primary library is then produced using any user-defined combination of aligners and extended using the predicted secondary structures. Expresso uses protein data bank (PDB) 3D structures as templates. For each input sequence, putative templates are identified by a BLAST search against the sequences of the PDB and the subsequent selection of the best hit (>30% identity over >50% of the query sequence). The library is then computed by aligning every pair of templates with a structural aligner. SAP (19) is used by default although users have the possibility to select other structural aligners or to combine them. Whenever a sequence lacks a closely related structure, the standard pair-wise sequence alignment procedure (proba_pair) is used for all the pair-wise alignments involving this sequence. Once the library is compiled the alignment is produced using the standard T-Coffee algorithm.

PSI-Coffee is a novel mode of T-Coffee (manuscript in preparation). It uses protein profiles as templates rather than structures and works as follows: each sequence is BLASTed individually against NR database and the resulting BLAST alignments (i.e. one-to-all between each query and its hits) are turned into profiles (sequences with identity <30% or coverage <40% are excluded). Given a set of N sequences, the result is a collection of N profiles each embedding a distinct query sequence. The profiles are then aligned two by two (using the proba_pair pair-HMM) and the resulting alignment for the query sequences is added to the library. The rest of the procedure uses the standard T-Coffee methodology to deliver a consistency-based MSA. The principle of PSI-Coffee is very similar to that of PROMALS (20). Its main advantage is its reliance on homology extension for the computation of the library, a process shown by us and others as a source of improvement for the alignment of remote homologues (1,20). It should be noted that the homologous sequences identified by BLAST are not added to the final MSA. They are only used to increase the accuracy of the underlying alignment.

All these aforementioned alignment procedures are now available via the web-server described in this article. Its main strength is to offer the most sophisticated modes of T-Coffee for the production of highly accurate sequence alignments. Some of these modes integrate complex component such as BLAST database searches and secondary structure predictions, yet thanks to the web server, users do not need to install, maintain or integrate these resources.

WEB SERVER

The server runs on our local infrastructure. It is composed of a front-end web application and a back-end execution cluster. The front-end is based on the *Play!* framework, a lightweight Java toolkit for developing web applications (http://www.playframework.org). The server is designed in such a way that each individual T-Coffee mode can be considered as an independent plug-in. This way, one can deploy alternative versions of the package and alternative configurations very easily. The front-end submits users' alignment requests to the batch-queuing system based on Oracle Grid Engine (formerly Sun Grid Engine).

Using T-Coffee web server

The web server can be accessed either from http://www.tcoffee.org or http://tcoffee.crg.cat. It is compliant with all major web browsers (Mozilla Firefox 3+, Google

Chrome, Internet Explorer 7+, Safari 5+, Opera 10+). Users do not require any login although it is advisable to provide an email when submitting large jobs of over 100 sequences. Starting from the index, users can choose the most suitable mode for their sequences:

(1) T-Coffee: advisable for large data sets of protein or nucleic acids.
(2) M-Coffee: advisable for large data sets of protein or nucleic acids when one wants to compare the output of alternative aligners.
(3) R-Coffee: for RNA sequences with a conserved secondary structure.
(4) Expresso: advisable for protein sequences with known 3D structures.
(5) PSI-Coffee: advisable for very challenging protein datasets.
(6) Accurate: is an experimental mode, yet unevaluated, that attempts to automatically combine the best modes.
(7) Combine: is similar to M-Coffee but allows users to combine pre-computed multiple sequence alignments.
(8) Core: makes it possible to evaluate the consistency of any multiple sequence alignment.
(9) iRMSD: provided a data set contains at least two structures, the iRMSD returns an evaluation of the considered alignment that takes into account the quality of the implied structural superposition.

Computing a multiple sequence alignment

Once an alignment mode is selected the server will display the alignment form submission page. In the simplest case, the user needs to enter the sequences to be aligned in the displayed text box. All the MSA computation modes take as input sequences in FASTA format, while the two alignment evaluation modes Core and iRMSD take, as input, multiple sequence alignments in ALN format (ClustalW output format). Sequences can also be uploaded using the 'Upload a file' link just below the text box on the submission form. A total of 150 sequences can be entered and each sequence can contain up to 10 000 residues for the T-Coffee mode and up to 2500 for the other modes.

On all services, advanced alignment settings are available by clicking the link 'Show more options'. By accessing this section, one can change T-Coffee alignment defaults and control advanced details. These vary depending on the selected T-Coffee mode, but in general they allow two important controls: the selection of the methods used to produce the primary library and some extra controls on the output formats.

Once sequences have been properly entered and the appropriate parameters selected, the 'Submit' button at the bottom of the page must be clicked in order to send the alignment request to the server. An identification number is assigned to each request and used as a unique reference. The alignment process can take from a few seconds up to several minutes, depending on the alignment complexity and the server load. If the browser is closed while waiting, users will nonetheless be able to access the result of their computation by reopening the server page and clicking on the 'History' link. This link displays the history as stored in a local cookie. This information will be lost if this cookie is deleted or when accessing from another machine. As an alternative the page displayed while waiting can be bookmarked and later revisited. Users may also provide their email to be informed of job completion.

When computation is finished, the server displays a summary page (Figure 1). It includes the computed MSA (or a link to that MSA for data sets larger than 1 MB). The box 'Result files' contains all the files produced by T-Coffee during the alignment process as well as the sequences input file. The provided link 'Download them all' allows users to download all the output files in a single zip archive. The 'Send result' box makes it possible to post-process the alignments and send them to third-party services-like ProtoGene (21) or the SIB MSA hub 'MyHits'. One can also re-run a job by clicking the link in the 'Replay' box. It will regenerate the submission page where users may either modify the alignment parameters or change the query sequences.

Output interpretation

The graphic colored output indicates the level of consistency between the final alignment and the library used by T-Coffee. The main score is the total consistency value. A value of a 100 means full agreement between the considered alignment and its associated primary library. It also means that the library is self-consistent. High values have been shown to reflect higher accuracy (22,23). Users are nonetheless advised not to compare these values between alignments of the same sequences computed with different strategies for generating the primary library. For instance M-Coffee usually returns higher values than T-Coffee but these differences cannot be interpreted in terms of relative accuracy. The individual sequence score is more informative as it allows an estimate of the relative fit of the sequences within the MSA. For instance, any sequence having a consistency score lower than the other sequences should be considered as suspicious. Likewise the residue color scheme reflects the primary library support for the alignment of the considered residue on a scale between 0 (blue, poorly supported) and 9 (dark red, strongly supported). While the two aforementioned studies have suggested that residues with a consistency score higher than 5 (i.e. yellow/orange/red) are quite likely to be correctly aligned, this scoring scheme is probably most useful to identify the highly unreliable stretches (blue) that are rarely aligned in a biologically meaningful way.

Evaluating alignments

The web server also provides two alignment evaluation modes implemented by T-Coffee: Core and iRMSD-APDB. When running Core, users simply need to input a pre-computed alignment in ALN format. T-Coffee automatically computes the corresponding library and outputs a colored version of the alignment. When using iRMSD users need to provide a multiple sequence alignment in ALN format containing PDB identifiers.

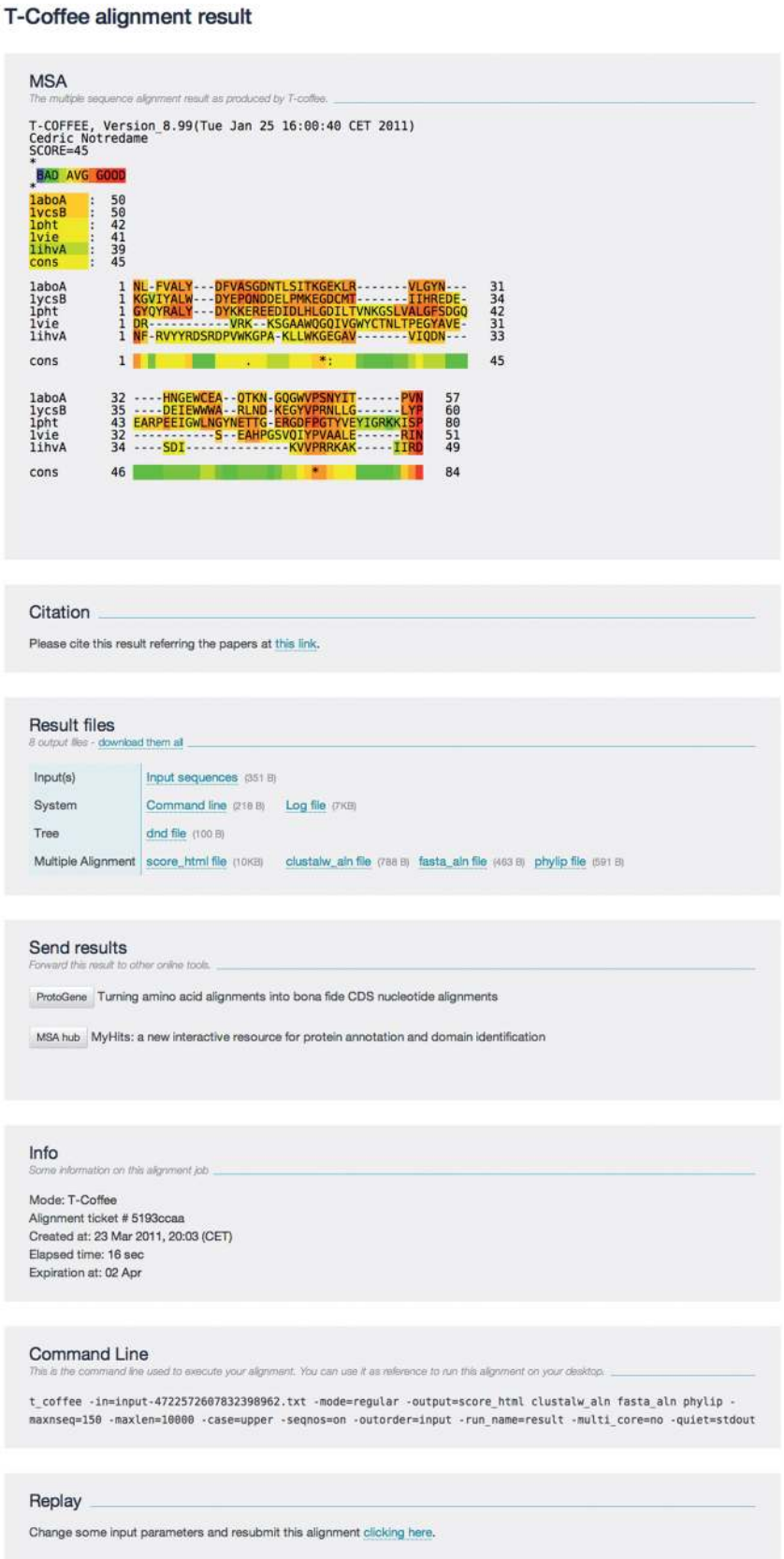

Figure 1. Sample T-Coffee result page. The top part is a color-coded alignment where sequences in red correspond to alignment portions with a strong support in the primary library. For post-processing purpose, users are advised to download the text-based version of the alignment available in the Result files section.

The iRMSD-APDB mode is similar to the server originally described (24).

CONCLUSION

In this article, we introduce the latest version of the T-Coffee web server. This new server has been completely redesigned both at the interface and server engine levels. It can be easily deployed as a single server or mirrored in a network of sites. It is designed with a pluggable component architecture that makes extension and re-configuration very easy. The user interface is more intuitive and based on the latest web standards provided by modern browser technologies. From a bioinformatics point of view, the main improvement has been the addition of the PSI-Coffee mode, a new homology extension-based mode for T-Coffee. Furthermore, the underlying T-Coffee algorithm has been parallelized (25) and significantly improved for efficiency, thus making it possible for the server to handle much larger datasets than was previously possible. Future developments will include a programmable application interface supporting REST-like web service interfaces.

FUNDING

Funding for open access charge: Plan Nacional (BFU2008-00419 to C.N. and P.D.T.); 7th Framework Programme of the European Commission, the LEISHDRUG (no 223414) and the Quantomics (KBBE-2A-222664) project; Computational resources are provided by the Center for Genomic Regulation (CRG) of Barcelona and the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the Swiss Institute of Bioinformatics; CUR of DIUE of GENCAT (to M.O. and A.M.); the Spanish ministry of education (TIN2008-05913); Consolider project (CSD 2007-00050);' Super-computacion y e-Ciencia (SYEC)'.

Conflict of interest statement. None declared.

REFERENCES

1. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**, 2455–2465.
2. Wang,L. and Jiang,T. (1994) On the complexity of multiple sequence alignment. *Journal of computational biology*, **1**, 337–348.
3. Morgenstern,B., Dress,A. and Wener,T. (1996) Multiple DNA and protein sequence based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.
4. Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.
5. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
6. Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515–1524.
7. Lassmann,T. and Sonnhammer,E.L. (2006) Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res.*, **34**, W596–599.
8. 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.
9. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
10. Pei,J., Sadreyev,R. and Grishin,N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
11. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
12. Pei,J., Kim,B.H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
13. Paten,B., Herrero,J., Beal,K. and Birney,E. (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, **25**, 295–301.
14. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
15. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
16. Poirot,O., Suhre,K., Abergel,C., O'Toole,E. and Notredame,C. (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res.*, **32**, W37–W40.
17. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
18. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.
19. Orengo,C.A. and Taylor,W.R. (1996) SSAP: Sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.
20. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
21. Moretti,S., Reinier,F., Poirot,O., Armougom,F., Audic,S., Keduas,V. and Notredame,C. (2006) PROTOGENE: turning amino acid alignments into bona fide CDS nucleotide alignments. *Nucleic Acids Res.*, **34**, W600–W603.
22. Notredame,C. and Abergel,C. (2003) Using Multiple Alignment Methods to Assess the Quality of Genomic Data Analysis. In Andrade,M. (ed.), *Bioinformatics and Genomes: Current Perspectives*. Horizon Scientific Press, pp. 30–50.
23. Lassmann,T. and Sonnhammer,E.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res.*, **33**, 7120–7128.
24. Armougom,F., Moretti,S., Keduas,V. and Notredame,C. (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–39.
25. Di Tommaso,P., Orobitg,M., Guirado,F., Cores,F., Espinosa,T. and Notredame,C. (2010) Cloud-Coffee: implementation of a parallel consistency-based multiple alignment algorithm in the T-Coffee package and its benchmarking on the Amazon Elastic-Cloud. *Bioinformatics*, **26**, 1903–1904.