DOI: 10.1093/nar/gkh382

3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment

Olivier Poirot[1], Karsten Suhre[1], Chantal Abergel[1], Eamonn O'Toole[3] and Cedric Notredame[1,2,*]

[1]Information Génomique et Structurale UPR2589-CNRS, CNRS, 31, Chemin Joseph Aiguier, 13 402 Marseille Cedex 20, France, [2]Swiss Institute of Bioinformatics, Lausanne University, Chemin des Boversesses, 1066 Epalinges, Switzerland and [3]hp High Performance Technical Computing Division, Hewlett Packard, BallyBrit, Galway, Ireland

Received February 14, 2004; Accepted March 16, 2004

ABSTRACT

This paper presents 3DCoffee@igs, a web-based tool dedicated to the computation of high-quality multiple sequence alignments (MSAs). 3D-Coffee makes it possible to mix protein sequences and structures in order to increase the accuracy of the alignments. Structures can be either provided as PDB identifiers or directly uploaded into the server. Given a set of sequences and structures, pairs of structures are aligned with SAP while sequence–structure pairs are aligned with Fugue. The resulting collection of pairwise alignments is then combined into an MSA with the T-Coffee algorithm. The server and its documentation are available from http://igs-server.cnrs-mrs.fr/Tcoffee/.

INTRODUCTION

The assembly of an accurate multiple sequence alignment (MSA) is a key step in many sequence analysis procedures. 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 and homology modeling (3) and phylogenetic analysis (4). More recently, MSAs have also proven useful to characterize nsSNPs (non-synonymous Single Nucleotide Polymorphisms) (5,6).

The success of such applications depends very much on the MSA quality, hence the importance of accuracy when computing an alignment. In practice, structurally correct alignments are considered to be a good starting point for most MSA applications (with maybe the exception of phylogenetic reconstruction), and established collections of reference structural alignments are widely used to benchmark and train existing MSA packages (7,8). However, when state-of-the-art packages are applied to sets of distantly related sequences, they deliver alignments that are only partly correct from a structural point of view (8), thus suggesting that sequence-based alignment procedures can still be greatly improved. In the current situation, the best way to produce a high-quality MSA remains the assembly of a multiple structural alignment. Unfortunately, few examples exist where enough related structures are available to carry out such a task.

An elegant alternative to the use of many structures is to mix sequences and structures, in the hope that the 3D information contained within the structures will help deliver a better alignment of the other sequences. Such a mix also constitutes a realistic solution, considering the increasing proportion of sequences without a known structure and the decreasing proportion of protein families not associated with at least one structure. However, the problem of combining sequences and structures has not yet been extensively addressed, and only a handful of methods are available that allow the seamless combining of sequences and structures (9) while appropriately using 3D information.

Here we present 3DCoffee@igs, a web server especially designed to combine sequences and structures by seamlessly integrating in T-Coffee (10) the three types of alignment methods needed for this purpose: sequence–sequence, sequence–structure and structure–structure alignment methods. When using one or more structures, the alignments thus produced are more accurate than similar alignments based on sequence information alone, as judged by the comparison with reference structure-based alignments (O.O'Sullivan, K.Suhre, D.Higgins and C.Notredame, submitted for publication). The inclusion of a threading method (sequence–structure alignment) makes it possible to use as little as one structure.

METHODS

Standard T-Coffee sequence alignment assembly

We use T-Coffee to mix sequences and structures. Given a set of sequences, the regular T-Coffee procedure involves the

*To whom correspondence should be addressed. Tel: +33 491 164 606; Fax: +33 491 164 549; Email: cedric.notredame@europe.com computation of a collection of pairwise alignments where 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 (11)]. This collection of pairwise alignments is named a library. The second step of the procedure involves the assembly of an MSA with a high level of consistency with the alignments contained in the library. Since T-Coffee uses a heuristic, the optimality of this process is not guaranteed, although the results are generally satisfactory as judged by comparison with alternative optimization methods (12). The assembly procedure is very similar to that described for ClustalW (13); extensive details are available in the original publication (10).

3D-Coffee protocol

The 3D-Coffee protocol takes advantage of the method-independent manner in which T-Coffee uses its libraries. Rather than filling the library with sequence-based pairwise alignments, 3D-Coffee compiles it using three types of pairwise methods: sequence–sequence, structure–structure and structure–sequence (threading) alignment procedures. From among the vast variety of structure comparison algorithms, we selected SAP (14) for the structure–structure alignments and Fugue (15) for the structure–sequence comparisons. A full validation of these choices is detailed in O. O'Sullivan, K. Suhre, D. Higgins and C. Notredame, submitted for publication. Our main criterion was the relatively high accuracy of these two methods and their ease of integration within the T-Coffee framework. It is nonetheless worth pointing out that any method with similar characteristics (i.e. able to deliver a sequence alignment) could easily be added to the procedure we describe here.

In practice, given a sequence dataset, the program starts by identifying the sequences associated with a structure and those that are not. It then considers all the possible pairs and applies the appropriate methods to these pairs. For instance, given a pair of structures, the program will successively make a global pairwise alignment, a local pairwise sequence alignment and a structure-based sequence alignment with SAP. If only one of the two sequences has a known structure, Fugue will be used instead of SAP.

The resulting pairwise alignments are compiled into a list of weighted pairs of aligned residues found in the individual alignments. Each pair receives a weight equal to the average level of identity within the pairwise alignment where it occurred. When two or more alignments contribute the same pair, their respective weights are added to yield the final weight. The collection of weighted residue-pairs constitutes the T-Coffee library.

T-Coffee uses the library to assemble a standard progressive alignment in a ClustalW-like manner. The program starts by computing the distance matrix of the sequences and uses it to estimate a guide tree. The guide tree controls the order in which the sequences are included one by one into the MSA. Each sequence is incorporated using the library in place of a substitution matrix. A recent modification of the T-Coffee algorithm (to be described elsewhere) has made it possible to significantly reduce the time complexity of the algorithm, down to $O(N^2L^2)$ from the previously reported $O(N^3L^2)$, N being the number of sequences and L their average length. However, in 3D-Coffee, SAP is the limiting factor, with a time complexity in the order of $O(L^3)$.

USING THE TCOFFEE@IGS SERVER

3D-Coffee is a new service that is available through the previously presented Tcoffee@igs server (17). It is maintained by IGS (Information Génomique et Structurale) and runs on a dedicated Alpha ES45 quadriprocessor server. It supports the analysis of a maximum number of 100 sequences with a maximum of 2000 residues each.

The 3D-Coffee service is provided in two versions, a regular and an advanced version. The regular version requires limited input from the user while the advanced version offers more possibilities such as uploading personal PDB structures and controlling the methods used to compute the library.

Tcoffee@igs server

The homepage of the server (http://igs-server.cnrs-mrs.fr/Tcoffee/) contains pointers to the four types of computation performed:

(i) The *Make a Multiple Alignment* section opens to the standard computation of a T-Coffee MSA, using the default parameters of the program, as described in (10).

(ii) The *Evaluate a Multiple Alignment* section provides an alignment evaluation using the CORE method as described in (17).

(iii) The *Combine Multiple Alignments* section makes it possible to combine several alignments into one. The advanced section of each server offers extra control on the library computation (choice of the methods) as well as a larger number of output options. These servers have all been previously described in (16).

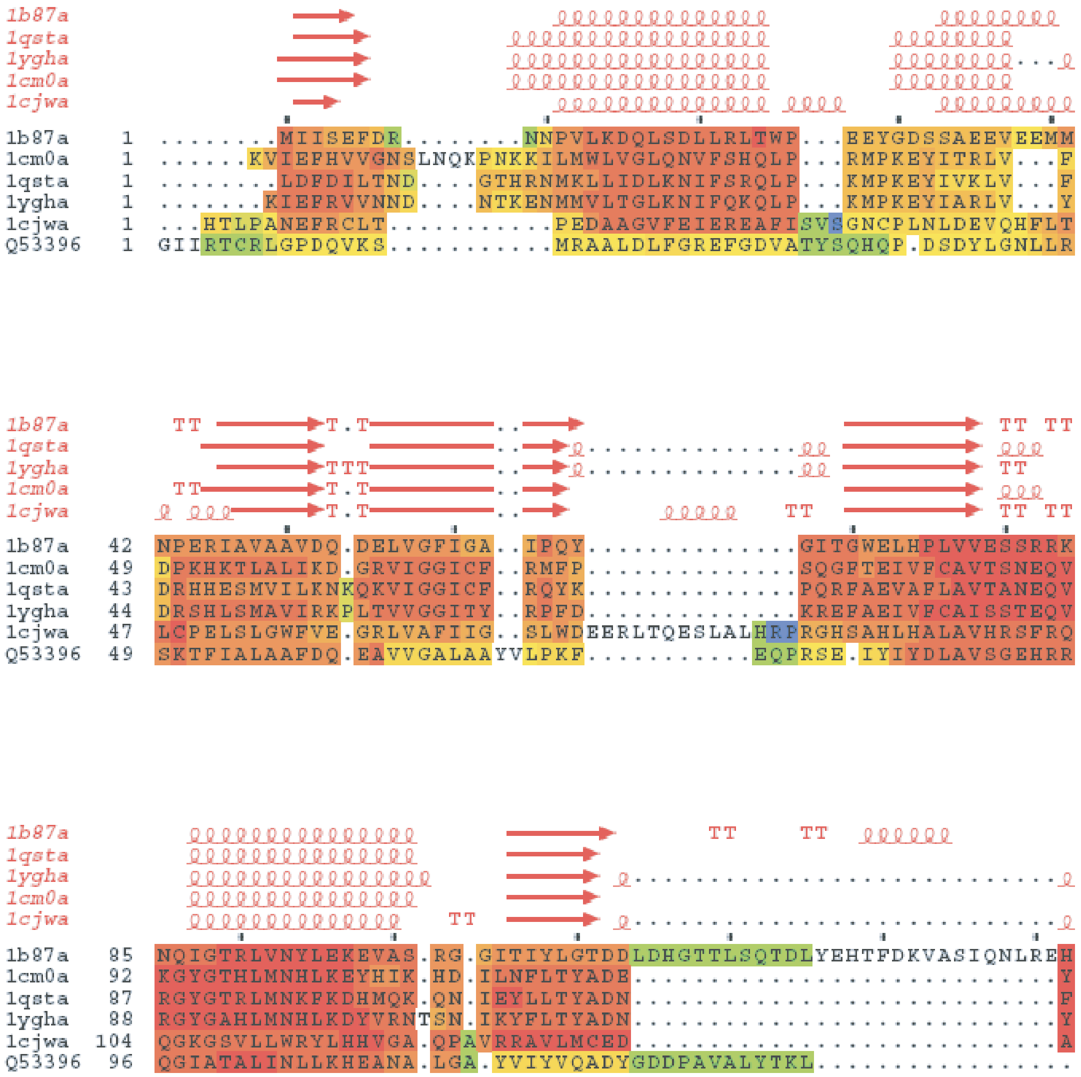
(iv) The last section, *Align Structures (3D-Coffee)*, is new and described in the next paragraph.

Align structures and sequences with 3DCoffee::regular

The 3DCoffee::regular server inputs a set of sequences in FASTA format. Among the sequences, those with a 3D structure must be named according to their PDB identifier. If the PDB file contains several chains, the chain index (letter or number) must be added to the name (1ppt**A**). If the sequence provided in the FASTA file is a subsequence of the indicated chain, T-Coffee aligns the provided sequence with its full PDB counterpart and makes sure that only the appropriate 3D information is used for alignment computation. This comparison also handles slight sequence discrepancies between the PDB and the user-provided sequence. In the regular mode of 3D-Coffee, the handling of the structures is entirely under T-Coffee control, which uses the FASTA information to gather the structures and chop them to the relevant portion. For users familiar with the stand-alone version of T-Coffee, we give the corresponding command line:

```
t_coffee-in seq.fasta Msap_pair Mfugue_pair
  Mslow_pair Mlalign_id_pair
```

sap_pair, *fugue_pair*, *slow_pair* and *lalign_id_pair* are pairwise methods used to compute the T-Coffee library. Once the



```
1b87a     1  ........MIISEFDR........NNPVLKDQLSDLLRLTWP...EEYGDSSAEEVEEMM
1cm0a     1  ......KVIEFHVVGNSLNQKPNKKILMWLVGLQNVFSHQLP...RMPKEYITRLV...F
1qsta     1  ........LDFDILTND....GTHRNMKLLIDLKNIFSRQLP...KMPKEYIVKLV...F
1ygha     1  .......KIEFRVVNND....NTKENMMVLTGLKNIFQKQLP...KMPKEYIARLV...Y
1cjwa     1  ...HTLPANEFRCLT...........PEDAAGVFEIEREAFISVSGNCPLNLDEVQHFLT
Q53396    1  GIIRTCRLGPDQVKS...........MRAALDLFGREFGDVATYSQHQP.DSDYLGNLLR

1b87a    42  NPERIAVAAVDQ.DELVGFIGA..IPQY..............GITGWELHPLVVESSRRK
1cm0a    49  DPKHKTLALIKD.GRVIGGICF..RMFP..............SQGFTEIVFCAVTSNEQV
1qsta    43  DRHHESMVILKNKQKVIGGICF..RQYK..............PQRFAEVAFLAVTANEQV
1ygha    44  DRSHLSMAVIRKPLTVVGGITY..RPFD..............KREFAEIVFCAISSTEQV
1cjwa    47  LCPELSLGWFVE.GRLVAFIIG..SLWDEERLTQESLALHRPRGHSAHLHALAVHRSFRQ
Q53396   49  SKTFIALAAFDQ.EAVVGALAAYVLPKF...........EQPRSE.IYIYDLAVSGEHRR

1b87a    85  NQIGTRLVNYLEKEVAS.RG.GITIYLGTDDLDHGTTLSQTDLYEHTFDKVASIQNLREH
1cm0a    92  KGYGTHLMNHLKEYHIK.HD.ILNFLTYADE............................Y
1qsta    87  RGYGTRLMNKFKDHMQK.QN.IEYLLTYADN............................F
1ygha    88  RGYGAHLMNHLKDYVRNTSN.IKYFLTYADN............................Y
1cjwa   104  QGKGSVLLWRYLHHVGA.QPAVRRAVLMCED............................A
Q53396   96  QGIATALINLLKHEANA.LGA.YVIYVQADYGDDPAVALYTKL.................
```

Figure 1. Typical output of a standard 3D-Coffee computation. Five structures have been aligned with a sequence (Q53396). The display is the ESPript (18) output produced by the Tcoffee@igs server. The CORE index is displayed on the alignment and indicates the relative reliability of the various sections (color code: blue, unreliable; green, low reliability; red, highly reliable portion of the alignment). DSSP (19) is used to determine the secondary structures from the PDB coordinates. Blue, green and yellow portions are mostly incorrectly aligned, as judged by comparison with HOMSTRAD reference alignment (9).

computation is over, the server returns a page of links to the produced result files. An ESPript (18) post-processing step makes it possible to visualize the secondary structure elements within the used structures (Figure 1).

The returned alignment is a sequence alignment, albeit generally improved by the use of structural information. Systematic benchmarking, carried out on a subset of HOMSTRAD (O. O'Sullivan, K. Suhre, D. Higgins and C. Notredame, submitted for publication), indicates that the accuracy of mixed sequences/structure alignments increases proportionally to the amount of structural information provided.

The 3DCoffee::advanced server

The advanced server makes it possible to upload user-defined PDB structures (up to three). The sequences of the uploaded structures should not be included within the FASTA sequences. The limitation to three private structures is arbitrary and will be increased upon request. In case the file contains more than one chain, the program extracts only the first one. It is the user's responsibility to provide the correct chain.

The advanced server also makes it possible to control the computation of the T-Coffee library by selecting the methods one wishes to include. For instance, if all the sequences have a known 3D structure, it is advisable to use only sap_pair, the structure–structure alignment method, to generate a structure-based MSAs.

CONCLUSION

In this paper, we present 3D-Coffee, a major extension of the Tcoffee@igs server. This new feature of the server makes it possible to combine sequences and structures within an MSA, thus producing high-quality MSAs.

The method we present here is versatile and easy to use. It affords the possibility of seamlessly combining structure and sequence information, private and public data, without the need to install additional programs such as SAP and Fugue locally. It certainly constitutes an adequate means to efficiently use available structural data. Future plans will involve the addition of new modules, rendering easier the mapping of structural information on to sequence data.

We strongly encourage users to send us their feedback.

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. Thompson,J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
8. O'Sullivan,O., Zehnder,M., Higgins,D., Bucher,P., Grosdidier,A. and Notredame,C. (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19**, I215–I221.
9. de Bakker,P.I., Bateman,A., Burke,D.F., Miguel,R.N., Mizuguchi,K., Shi,J., Shirai,H. and Blundell,T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, **17**, 748–749.
10. 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.
11. Huang,X. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math*., **12**, 337–357.
12. Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.
13. 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.
14. Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol*., **208**, 1–22.
15. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol*., **310**, 243–257.
16. Poirot,O., O'Toole,E. and Notredame,C. (2003) Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res*., **31**, 3503–3506.
17. Notredame,C. and Abergel,C. (2003) Using multiple sequence alignments to assess the quality of genomic data. In Andrade,M. (ed.), *Bioinformatics and Genomes: Current Perspectives*. Horizon Scientific Press, Norfolk, UK, pp. 30–50.
18. Gouet,P., Robert,X. and Courcelle,E. (2003) ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res*., **31**, 3320–3323.
19. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.