Expresso: Automatic incorporation of Structural Information in Multiple Sequence Alignments using 3D-Coffee

Fabrice Armougom[1], Sébastien Moretti[1], Olivier Poirot[1], Stéphane Audic[1], Pierre Dumas[2], Basile Schaeli[2], Vladimir Keduas[1], Cedric Notredame [1*]

[1] Laboratoire Information Génomique et Structurale,
CNRS UPR2589
Institute for Structural Biology and Microbiology (IBSM),
Parc Scientifique de Luminy
163 Avenue de Luminy
FR- 13288, Marseille cedex 09
France

[2] Laboratoire de systèmes périphériques
Ecole Polytechnique Fédérale de Lausanne
CH 1015 Lausanne,
Switzerland

*To whom correspondence should be addressed
Email: cedric.notredame@europe.com

F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, Nucleic Acids Res., 1 July 2006; 34: W604 - W608.

Abstract

Expresso is a multiple sequence alignment server that aligns sequences using structural information. The user only needs to provide sequences. The server runs BLAST to identify close homologues of the sequences within the PDB database. These PDB structures are used as templates to guide the alignment of the original sequences using structure based sequence alignment methods like SAP or Fugue. The final result is a multiple sequence alignment of the original sequences based on the structural information of the templates. An advanced mode makes it possible to either upload private structures or specify which PDB templates should be used to model each sequence. Providing the suitable structural information is available, Expresso delivers sequence alignments with accuracy comparable to structure based alignments. The server is available on http://www.tcoffee.org/.

Introduction

Over the past years, multiple sequence alignments (MSAs) have become one of the most widely used tool in biology along with database search methods. MSAs are needed for profile analysis, phylogenetic reconstruction, structure prediction and a wealth of minor but important applications such as PCR primer design or sequence reconciliation. The ever-growing reliance on MSAs is even more pronounced now that hundreds of complete genomes are being made available. This window opened on evolution provides an ideal context for MSAs to fulfill their potential as key tools in functional genomics.

Unfortunately, MSA packages are not yet accurate enough to deliver all their promises and the sharp increase in the number of methods recently published (25 novel programs over the last 5 years) illustrates well the expectation for improvement within the community. MSAs are not always good enough for large scale analysis and while immense progress has been made to accurately align multiple sets of sequences with more than 40% average identity, recent benchmarks published with the MAFFT 5 package (1) reveal that state of the art methods still fail to reliably align distantly related sequences. In the so-called “Twilight zone”(2), sequences with less than 20 % identity cannot be aligned with more than 30 % average accuracy (as judged by comparison with reference alignments). So far, the most convincing solution to this problem has been to supplement sequences with structural information (3).

The reason why structure-based MSAs are more accurate is not so much a consequence of better algorithms but rather an effect of structures evolutionary stability. Structures evolve slower than the sequences(4) and even when sequences have diverged beyond recognition it is often possible to establish homology (i.e. common ancestry) on the basis of 3D folds comparisons(3). The increasing availability of structural data (5) means that relying on structure based methods for sequence analysis has become much more realistic than it used to be. However, sequences are still being determined much faster than structures, thus creating a context where methods able to efficiently combine sequences and structure into accurate MSAs are needed. To the best of our knowledge, only 6 algorithms have been designed that are able to make use of secondary (6,7) or tertiary (8-10) structure information. In the context of this work, we used 3D-Coffee (11) for its ability to combine the output of several methods into one unique model. 3D-Coffee is based on the T-Coffee algorithm (12), a heuristic method that uses a progressive algorithm to compute an MSA having a high consistency with a

collection of pre-computed pairwise alignments (the library). In 3D-Coffee, the principle is the same except that the library's pairwise alignments are derived from structural superpositions using methods like Sap (13), Lsqman (14) or possibly any alternative structure alignment package (For a review see (15)). When using combinations of structures and sequences, 3DCoffee can also incorporate structure-sequence (threading) alignment methods like Fugue (16) to ease the diffusion of structural information onto the sequences.

3D-Coffee has been available via the web server 3DCoffee@igs for more than two years (17). The original implementation made it possible to combine sequences and structures using the most advanced T-Coffee options through a simple web interface. Although it provides access to most of the T-Coffee inline functions, this server requires the user to explicitly specify which structural template is to be associated with each sequence. This specification, made through a cumbersome procedure of sequence renaming, was complicated and impractical for non specialists.

The novel version of 3D-Coffee@igs is named Expresso because it makes it possible for non-specialists to rapidly and automatically benefit from the strength of 3D-Coffee. The term Expresso also conveys the notion of aroma extraction and concentration, a notion that resonates with the way structures are "expressed" within the MSA. In Expresso, we implemented an automated identification of suitable structural templates via a BLAST search against the PDB database. 3D-Coffee uses the selected structures to assemble a genuine structure-based MSA during a process that merely looks like a standard sequence alignment procedure from the user's point of view. Providing the appropriate structural information is available, Expresso is significantly more accurate than regular homology based methods and its alignments are often indistinguishable from reference structure based alignments (11) .

Methods

Selection of the Structural Template

The core idea of Expresso is to reliably identify structures that can be used as templates for the sequences (source) one wishes to align. The rationale is that any alignment carried out on the templates can easily be transposed onto the source sequences as long as the source and the template are highly homologous. The most basic and important step in Expresso is a BLAST search of the source sequences against PDB, in order to identify suitable templates. A BLAST match is considered a suitable template if it displays a minimum of 60% sequence identity with the source sequence and a minimum coverage of 70% (i.e. 70% of the source sequence residues matched). These rather conservative criteria were chosen to limit the template selection to close homologues whose alignment with the source is entirely non-ambiguous. No effort is made to identify structures with special conformations, or resolutions, although this could easily be added to the pipeline. However, whenever the automatic procedure appears inappropriate, the user can explicitly declare the source-template association using the advanced mode of the server.

Integration of the Structural Template

Once every sequence with a structural homologue has been assigned its template, 3DCoffee undertakes the library computation step. It applies a collection of pre-defined pairwise alignment methods on every pair of sequences. The methods are either sequence-based alignment (e.g. lalign) or structure-based (e.g. SAP). When using structural methods, a structure based alignment of the templates is first computed. The two source sequences are

then aligned to their respective templates, and the induced pairwise alignment of the two sources is integrated within the library (Figure 1). The accuracy of this delicate process relies on a high level of identity between the source and the template sequence, hence the stringency of the original BLAST search.

Alignment Computation

Once the library assembly step is finished, the MSA is assembled in a progressive fashion, using the standard T-Coffee algorithm. The default mode of the server for running T-Coffee is:

t_coffee <seq> -in Mslow_pair, Msap_pair, Mlalign_id_pair –template_file SCRIPT_blast.pl

Where **SCRIPT_blast.pl** is a stand-alone script that BLASTs every source sequence against PDB in order to identify suitable structural templates (if they exist).

Using Expresso

Default Mode

The server can be accessed at http://www.tcoffee.org/, by clicking on the Expresso link, either advanced or regular. To use the regular mode, one simply needs to cut and paste FASTA sequences. No special precaution is needed to name the sequences.

Advanced Mode

The advanced mode of the server offers many more possibilities and guides the user with a series of bulleted points:

1-*Cut and paste your sequences*.

2-*Upload your PDB structures* - should be used when some of the structures are not in the public domain. When uploading a PDB template, the associated source sequence is automatically generated using the SEQRES field. PDB files must follow the standard PDB format and the server requires a TITLE, a HEADER, an ATOM and a SEQRES section.

3-*Select the methods*. The default selection corresponds to 3DCoffee. Further structure alignment methods will soon be added, along with new multiple sequence alignment packages. Users are welcome to suggest the incorporation of any public domain method.

4-*PDB template selection*. By default no template is used in the advanced mode. Users should check the SCRIPT box to automatically fetch the templates with BLAST, or specify the source to template correspondences in the box below. The format for doing so is indicated in the corresponding section.

Figure 2 shows a typical output, computed on the HOMSTRAD thioredoxin family (18). The first alignment (Figure 2a) was computed using the standard T-Coffee protocol, while the other (Figure 2b) is an Expresso MSA computed using the regular mode. In the T-Coffee alignment, 15 % of the columns are correctly aligned (as judged by comparison with the HOMSTRAD reference alignment) while in the Expresso MSA, 49% of the columns appear to be correct. Figure 2c shows which template was selected for each sequence. When selecting the template, no attempt is made to match the source sequence name with the template name, which sometimes results in apparent discrepancies (1aaza modelled with 1de2A). While in most cases, these arbitrary choices should not affect the output, better

control can be achieved by specifying the template/sequence correspondence in the advanced mode.

Conclusion and future developments

Expresso is an improved version of the original 3DCoffee@igs server. Structures are now fetched automatically and used to guide the alignment. This procedure can result in a dramatic improvement of the sequence alignment when homologue PDB structures are available. From the user point of view, Expresso is a regular multiple sequence alignments server that seamlessly includes structural information in MSAs, allowing non specialists to benefit from the power of structure-based sequence alignment without having to address all the technical issues it implies. Future developments will involve a gradual extension of the methods available for combination in the advanced section.

We strongly encourage users to send us their feedback.

Acknowledgement

The development of the server was supported by CNRS (Centre National de la Recherche Scientifique), Sanofi-Aventis Pharma SA., Marseille-Nice Génopole and the French National Genomic Network (RNG). We thank Prof. Jean-Michel Claverie (head of IGS) for stimulating scientific discussions and material support. We also thank Prof Roger Hersch (EPFL) for useful advices on code optimization.

References

1. Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511-518.
2. Sander, C. and Schneider, R. (1991) Database of homology-derived structures and the structurally meaning of sequence alignment. *Proteins: Structure, Function, and Genetics*, **9**, 56-68.
3. Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science*, **283**, 595-602.
4. Lesk, A.M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol*, **136**, 225-270.
5. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*, **34**, D302-305.
6. Heringa, J. (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput Chem*, **23**, 341-364.
7. Simossis, V.A. and Heringa, J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res*, **33**, W289-294.
8. Zhang, Z., Lindstam, M., Unge, J., Peterson, C. and Lu, G. (2003) Potential for dramatic improvement in sequence alignment against structures of remote homologous proteins by extracting structural information from multiple structure alignment. *J Mol Biol*, **332**, 127-142.
9. Ren, T., Veeramalai, M., Tan, A.C. and Gilbert, D. (2004) MSAT: a multiple sequence alignment tool based on TOPS. *Appl Bioinformatics*, **3**, 149-158.
10. Kleinjung, J., Romein, J., Lin, K. and Heringa, J. (2004) Contact-based sequence alignment. *Nucleic Acids Res*, **32**, 2464-2473.

11. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G. and Notredame, C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, **340**, 385-395.
12. 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.
13. Taylor, W.R. and Orengo, C.A. (1989) Protein structure alignment. *J Mol Biol*, **208**, 1-22.
14. Kleywegt, G.J. and Jones, T.A. (1999) Software for handling macromolecular envelopes. *Acta Crystallogr D Biol Crystallogr*, **55 (Pt 4)**, 941-944.
15. Kolodny, R., Koehl, P. and Levitt, M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, **346**, 1173-1188.
16. 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.
17. 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**, W37-40.
18. 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.

Figures

Figure 1

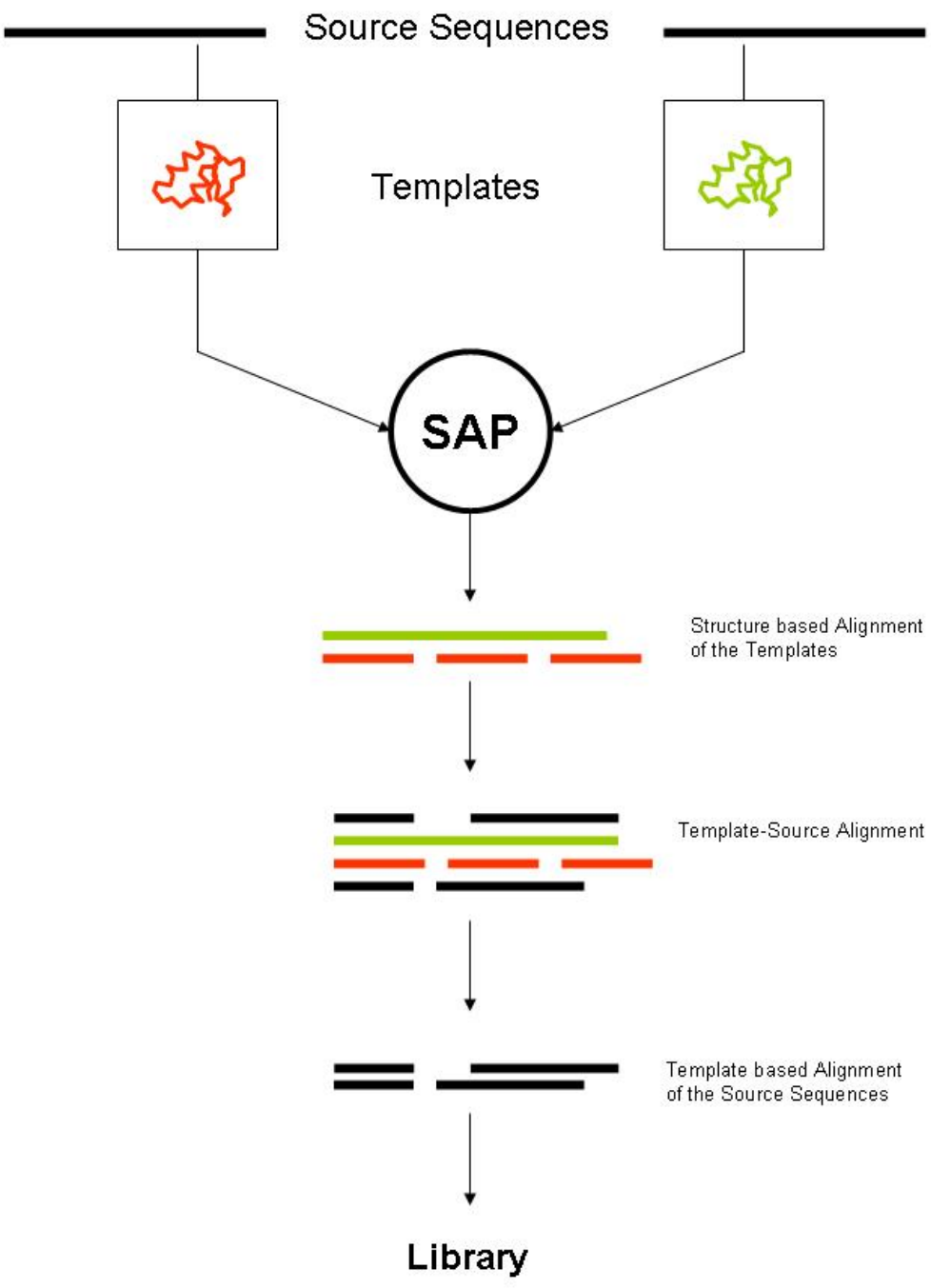
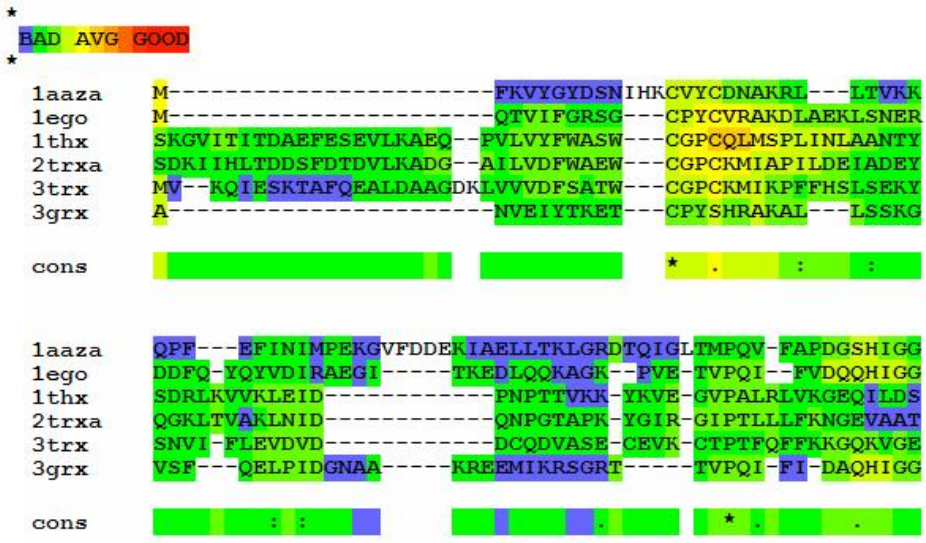


Figure 1. Computation of a template based library

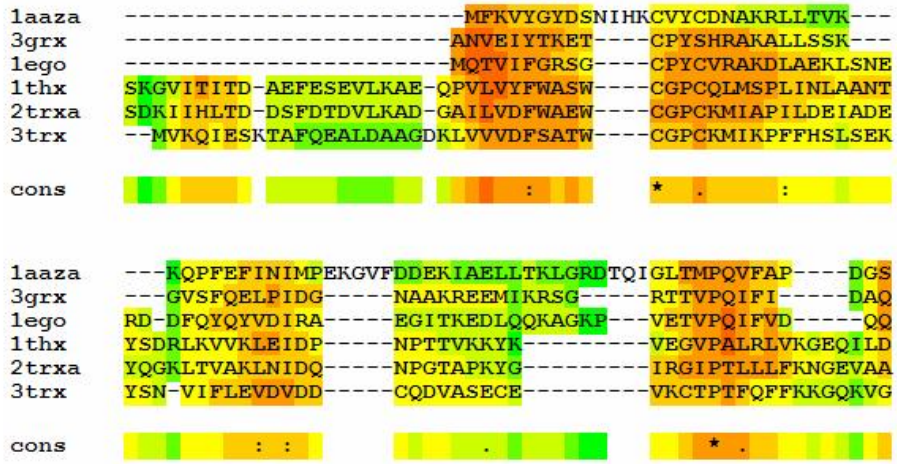
Structural templates are assigned to each original source sequence and these templates are used to generate a structure based sequence alignment. The final library alignment is generated by aligning each source sequence with its template, thus generating a template based alignment of the two sources.

Figure 2

a)



b)



c)

```
>1aaza  ➔ 1DE2A
>1ego   ➔ 1EGR
>1thx   ➔ 1THX
>2trxa  ➔ 2BTOT
>3trx   ➔ 4TRX
>3grx   ➔ 3GRX
```

Figure 2. Computation of an Expresso Alignment

a) Default T-Coffee alignment of the thioredoxin HOMSTRAD dataset. Red portions have a high reliability and are expected to be more accurate that the rest. Blue and green portions are the less consistent. Concistency is estimated from a sequence based T-Coffee library. In this MSA 15% of the columns are similar to the reference HOMSTRAD MSA.
b) Expresso alignment. Concistency is now estimated from a library computed using template based alignments. In this alignment 49% of the columns are similar to the HOMSTRAD reference MSA.
c) Automatic template assignment.