BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 19 2006, pages 2439–2440
doi:10.1093/bioinformatics/btl404

Sequence analysis

APDB: a web server to evaluate the accuracy of sequence alignments using structural information

Fabrice Armougom[1], Olivier Poirot[1], Sébastien Moretti[1], Desmond G. Higgins[2], Phillip Bucher[3], Vladimir Keduas[1] and Cedric Notredame[1,*]

[1]CNRS UPR2589, Institute for Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163 Avenue de Luminy, FR-13288, Marseille cedex 09, France, [2]The Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland and [3]Institut Suisse de Recherche et d'Experimentation sur le Cancer, Ch. des Boveresses 155 CH-1066 Epalinges, Switzerland

Received on March 31, 2006; revised on June 17, 2006; accepted on July 21, 2006

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: The APDB webserver uses structural information to evaluate the alignment of sequences with known structures. It returns a score correlated to the overall alignment accuracy as well as a local evaluation. Any sequence alignment can be analyzed with APDB provided it includes at least two proteins with known structures. Sequences without a known structure are simply ignored and do not contribute to the scoring procedure.

Availability: APDB is part of the T-Coffee suite of tools for alignment analysis, it is available on www.tcoffee.org. A standalone version of the package is also available as a freeware open source from the same address.

Contact: cedric.notredame@europe.com

1 INTRODUCTION

Structure-based sequence alignments have become a key component in the design and the improvement of sequence alignment methods. The extensive usage of structural information to align sequences results mostly from the observation that 3D folds evolve slower than primary sequences (Chothia and Lesk, 1986) and can be used to derive accurate sequence alignments, even when the sequences themselves have diverged beyond recognition. This property is often used to compute sequence alignments of remote homologues or to assemble collections of reference alignments that are used as gold standards to validate, benchmark and improve sequence alignment methods (Edgar, 2004; Thompson *et al*., 2005; Van Walle *et al*., 2005).

Detailed analysis, however, shows that structure alignment methods often disagree on distantly related proteins (Kolodny *et al*., 2005). For instance, the alignments delivered by CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1996) only agree on 70% of the positions as judged on the 1682 pairs of homologous structures contained in the Prefab Database (Edgar, 2004). These variations probably explain why established collections of structure-based alignments are sometimes inconsistent with one another. In some previous work, we argued that it may sometimes be more reliable to evaluate sequence alignments by directly using the structural information they are associated with, rather than depending on an intermediate reference alignment that constitutes a potentially distorted interpretation of the original structural signal.

In an attempt to provide such a direct measure, we developed an algorithm named APDB (Analyze PDB) (O'Sullivan *et al*., 2003) that uses the structural information independently of any structural alignment or superposition. APDB relies on the simple observation that if two similar structures are aligned correctly, the intramolecular distances between equivalent Cα (as defined by the alignment) must be similar. By simply measuring the geometric compatibility of two structures according to their alignment, APDB makes no reference to any authoritative alignment and is therefore independent from any kind of optimization, unlike similar methods like MaxSub (Siew *et al*., 2000), LSQman (Kleywegt and Jones, 1999) or TMScore (Zhang and skolnick, 2004). It also makes APDB suitable for comparing alternative alignments of the same sequences, as long as corresponding structures are available.

2 USING THE APDB SERVER

The server is available on http://www.tcoffee.org/. It only makes sense to use the APDB server if the alignment one wishes to evaluate contains at least two sequences with a known structure. These sequences must be named according to their structure PDB identifier (with the chain index appended if appropriate). Whenever there is an imperfect match between the user's and the PDB sequence, the program makes an automatic alignment based reconciliation. This process explicitly fails when the sequences are less than 80% identical. Sequences without a known structure are simply ignored and do not contribute to the scoring procedure.

The 1cvl_1tca Prefab dataset has been aligned with Muscle (Edgar, 2004) (a) and Clustalw(Thompson, *et al*., 1994) (b). The resulting alignments were evaluated with the APDB server and the following figure displays the localAPDB score. Sequences in red and orange are considered correctly aligned by APDB.

The server returns the overall APDB scores and a color-coded alignment with local APDB score (Fig. 1). The overall APDB score is an estimation of the fraction of columns likely to be correctly aligned within a pairwise structural alignment and the color code estimates the potential correctness of each individual alignment position (red: very likely; orange: possible; green/blue: unlikely).

*To whom correspondence should be addressed.

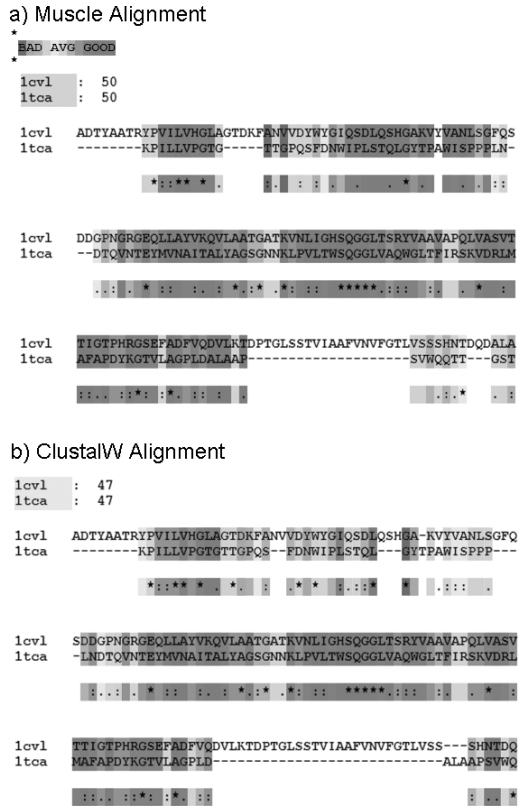

Fig. 1. Output of a standard APDB computation.

Figure 1 shows the evaluation of two alternative alignments of the same structures. The first one was produced by Muscle (3.52) and is estimated to be 43.8% accurate as compared with its Prefab reference (Edgar, 2004). The second one, produced by ClustalW (1.81), is expected to have an accuracy of 55.7% according to the Prefab criterion. The score returned by APDB is in broad agreement with these figures (Clustalw APDB: 50.3%, Muscle: 47.5%).

ACKNOWLEDGEMENTS

The authors thank Prof. Jean-Michel Claverie (head of IGS) for material support. 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).

Conflict of Interest: none declared.

REFERENCES

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **283**, 595–602.

Kleywegt,G.J. and Jones,T.A. (1999) Software for handling macromolecular envelopes. *Acta Crystallogr. D Biol. Crystallogr*, **55**, 941–944.

Kolodny,R. *et al.* (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.

O'Sullivan,O. *et al.* (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19** (Suppl. 1), i215–i221.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Siew,N. *et al.* (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Thompson,J. *et al.* (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.

Thompson,J.D. *et al.* (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Van Walle,I. *et al.* (2005) SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.

Zhang,J. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.