BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 2983–2985
doi:10.1093/bioinformatics/btq572

Structural bioinformatics

Advance Access publication October 10, 2010

Pre-calculated protein structure alignments at the RCSB PDB website

Andreas Prlić[1,*], Spencer Bliven[2], Peter W. Rose[1], Wolfgang F. Bluhm[1], Chris Bizon[3], Adam Godzik[4] and Philip E. Bourne[5,*]

[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, Mailcode 0505 La Jolla, CA 92093-0505, [2]Bioinformatics Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, [3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, NC 27517, [4]Joint Center for Structural Genomics, Bioinformatics Core, University of California at San Diego, La Jolla, CA 92093 and [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, Mailcode 0743, La Jolla, CA 92093-0743 USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: With the continuous growth of the RCSB Protein Data Bank (PDB), providing an up-to-date systematic structure comparison of all protein structures poses an ever growing challenge. Here, we present a comparison tool for calculating both 1D protein sequence and 3D protein structure alignments. This tool supports various applications at the RCSB PDB website. First, a structure alignment web service calculates pairwise alignments. Second, a stand-alone application runs alignments locally and visualizes the results. Third, pre-calculated 3D structure comparisons for the whole PDB are provided and updated on a weekly basis. These three applications allow users to discover novel relationships between proteins available either at the RCSB PDB or provided by the user.

Availability and Implementation: A web user interface is available at http://www.rcsb.org/pdb/workbench/workbench.do. The source code is available under the LGPL license from http://www.biojava.org. A source bundle, prepared for local execution, is available from http://source.rcsb.org

Contact: andreas@sdsc.edu; pbourne@ucsd.edu

Received on August 12, 2010; revised and accepted on October 1, 2010

1 INTRODUCTION

At its core, the *RCSB PDB Protein Comparison Tool* contains a new implementation of the two structure alignment algorithms Combinatorial Extension (CE) (Shindyalov and Bourne, 1998) and FATCAT (both rigid body and flexible versions) (Ye and Godzik, 2003).

Both the CE and FATCAT algorithms detect aligned fragment pairs (AFPs) during the alignment process. These AFPs are based on similarities in local geometry. There is a difference in how initial AFPs are combined in order to calculate an optimal alignment. CE applies the process of 'Combinatorial Extension' to find possible continuous alignment paths leading to an optimal alignment. The resulting alignment is a 'rigid-body' based alignment. In contrast to this, FATCAT allows the introduction of 'twists' into the alignment with the consequence that different regions of a protein structures can undergo different geometric transformations. This is required in order to be able to deal with protein flexibility.

A protein that undergoes significant domain re-arrangement during iron binding is transferrin. It consists of two domains that can move relative to each other. FATCAT in its flexible mode can easily detect an alignment between the apo and holo forms that covers 95% of both protein chains. (e.g. PDB ID 1IEJ chain A and PDB ID 1BTJ chain A). However, using rigid body superposition only a partial alignment is possible (both CE and FATCAT using the rigid mode only). One drawback of the flexible mode is that if distantly related proteins are being aligned, sometimes twists between unrelated regions can be introduced (e.g. alignment of PDB ID 1CDG chain A and PDB ID 1TIM chain A), in which case it is better to run the alignments in rigid mode.

A limitation of both CE and FATCAT in their original versions is that they compute sequence order-dependent alignments. A number of difficult to detect relationships between proteins have been published, some of which require sequence order independence for a correct alignment (Andreeva *et al.*, 2006). An algorithm that can detect such order-independet alignments is Triangle Match (Bachar *et al.*, 1993; Nussinov and Wolfson, 1991). Dali in its early versions also could detect permutated proteins; however, this feature seems to have been lost in its recent implementations, (Holm and Sander, 1993). We have recently improved CE to be able to detect circularly permutated alignments (Bliven *et al.*, manuscript in preparation). This implementation is available as an option of the RCSB Protein Comparison Tool.

2 APPROACH

The CE and FATCAT algorithms have been re-implemented in the Java programming language, which is indicated by a lower-case j in front of the new names, jCE and jFATCAT. Several components were added to these implementations.

First, the alignment algorithms were integrated into the RCSB PDB website, Berman *et al.* (2000) to provide a novel structure alignment service. Second, a stand-alone application can be run using Java Web Start technology. Third, a client-server architecture was developed for calculating large-scale comparisons using

*To whom correspondence should be addressed.

compute clouds. Finally, a software bundle is provided that allows local installation of the tool and to run custom comparisons. We describe some of these components in more detail.

Pairwise sequence and structure alignment The comparison tool allows pairwise comparison of protein sequences and 3D structures. For sequence comparison the Smith–Waterman (Smith and Waterman, 1981), Needleman–Wunsch (Needleman and Wunsch, 1970) and blast2seq (Tatusova and Madden, 1999), algorithms are provided. Support for structure comparisons includes the new implementations of CE and FATCAT and links to some of the most prominent external protein structure alignment services: the original FATCAT server (Ye and Godzik, 2004), Mammoth (Ortiz *et al.*, 2002), TM-align (Zhang and Skolnick, 2005) and Topmatch, (Sippl and Wiederstein, 2008), (Sippl, 2008). Other available structure alignment software can be found at Wikipedia http://en.wikipedia.org/wiki/Structural_alignment_software

All alignments that are run using jCE and jFATCAT are calculated server-side on the fly and cached for future retrieval using XML files. If the alignment is requested again later, it can be instantly returned by reading the XML file. A web user interface provides access to the alignment results. Alternatively, a Java Web Start client application, based on Jmol (2010), and BioJava (Holland *et al.*, 2008), provides a novel 3D visualization tool that allows the investigation of sequence–structure relationships between two aligned proteins. See Figure 1 for an example alignment.

Systematic structure alignments across the PDB Sequence database searches are a frequently used tool to identify closely related proteins within a database. However, with decreasing sequence similarity relationships between proteins become harder to detect. In order to enable identification of relationships across the PDB, even if sequence similarity is low, we are providing systematic and precomputed alignments.

The procedure providing pre-computed structure comparisons across the PDB is split into two steps.

First, the goal is to reduce the complexity of the problem by identifying representative protein chains for clusters of related proteins. BLASTClust (Altschul *et al.*, 2004) is used to cluster all protein chains by sequence similarity. We require 90% overlap between all sequences in a cluster. Therefore, a shorter fragment (e.g. a single domain) of a longer sequence (e.g. a multi-domain protein) will usually not be in the same cluster as the whole sequence. Within clusters, sequences are ranked by experimental method, resolution and release date. While the RCSB PDB website provides sequence comparisons for various levels of sequence identity within a cluster, structural comparisons are only provided based on clusters with 40% sequence identity, currently approximately 16 000 representative protein chains.

Second, the rigid version of jFATCAT is used to calculate all-against-all 3D structure comparisons across all representative protein chains. This requires a significant amount of CPU time. Specifically, a client-server architecture has been developed that allows the user to easily run a large number of jobs in parallel (for details see http://www.renci.org/publications/techreports/TR-09-03.pdf). A total of 122 million alignments were calculated on the Open Science Grid, taking approximately 102 000 CPU hours. Another 18 million alignments were calculated on the San Diego Supercomputing

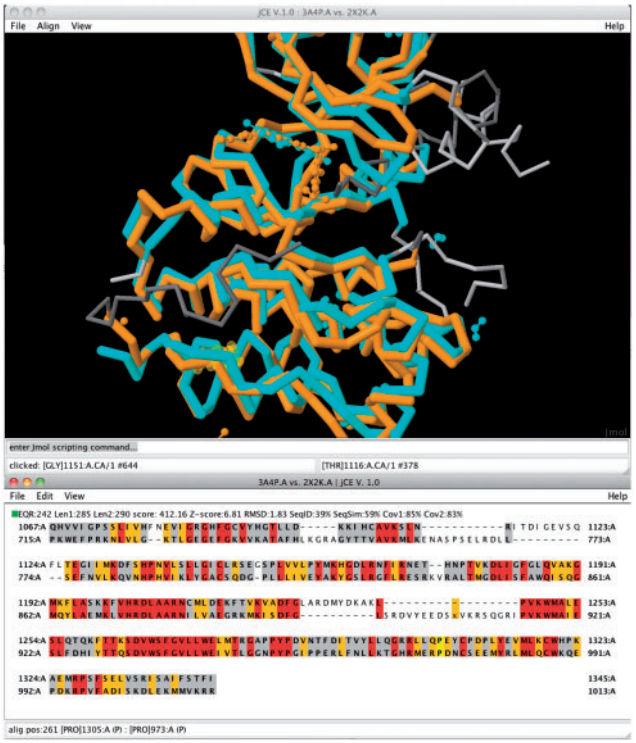

Fig. 1. A new user interface for jCE and jFATCAT structure alignments allows the investigation of sequence and 3D structure relationships. Here, the alignment of two kinases, the Hepatocyte growth factor receptor PDB ID 3A4P and the Proto-Oncogene Tyrosine-Protein Kinase Receptor RET PDB ID 2X2K. If the structures contain ligands they are also superimposed and displayed. The coloring for the sequence representation of the structure alignment represents the sequence conservation: red: identical residues, orange: similar and grey: structurally equivalent, but sequence mismatch.

Center (SDSC) Triton Cluster and local RCSB PDB servers. The alignment results were stored in ∼1 terabyte of XML files.

Weekly updates Incremental updates to the all-against-all comparisons are run weekly using in-house RCSB PDB servers at the same time the PDB itself is updated. Every week new sequence clusters are calculated and missing alignments for newly added representative chains are calculated. At present, an average weekly update requires the calculation of about 1 million structure alignments.

3 DISCUSSION

The RCSB PDB website now provides a state of the art protein comparison tool that can be used as a web service and for local access. Further, pre-calculated all-against-all comparisons of sequences and 3D protein structures, respectively, are provided on a weekly basis. Currently, the calculations are done on a whole-chain basis. We are working on another set of comparisons using domain-based comparisons of all chains with an enhanced version of jCE that includes handling of circular permutations. An alternative is to use TOPS++FATCAT (Veeramalai *et al.*, 2008), which provides a 10-fold speed up as compared to FATCAT. The domain assignment problem is non-trivial, and for newly released protein structures

results are not immediately available from classifications like SCOP (Andreeva *et al.*, 2008), or CATH (Cuff *et al.*, 2009). Hence, we are investigating whether consensus based approaches like pDomains, (Alden *et al.*, 2010), can guide which domain assignments to use for the automated calculations.

Funding: The RCSB PDB is managed by two members of the RCSB: Rutgers and UCSD, and is funded by National Science Foundation (NSF), National Institute of General Medical Sciences, Department of Energy (DOE), National Library of Medicine, National Cancer Institute, National Institute of Neurological Disorders and Stroke and National Institute of Diabetes and Digestive and Kidney Diseases. The RCSB PDB is a member of the wwPDB. This work was supported by the RCSB PDB grant NSF DBI 0829586. Computation provided in part by the Open Science Grid funded by NSF and DOE, supported by OSG Engagement under NSF award number 0753335.

Conflict of Interest: none declared.

REFERENCES

Alden,K. *et al*. (2010) dconsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment. *BMC Bioinformatics*, **11**, 310.

Altschul *et al*. (2004) BLASTClust. Available at: http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html (last accessed date October 20, 2010).

Andreeva,A. *et al*. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Andreeva,A. *et al*. (2006) SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.*, **35**, D253–D259.

Bachar,O. *et al*. (1993) A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng.*, **6**, 279–288.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Cuff,A. *et al*. (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, **17**, 1051–1062.

Holland,R.C.G. *et al*. (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J.Mol.Biol.*, **233**, 123–138.

Jmol (2010) Jmol: an open-source Java viewer for chemical structures in 3D. Available at: http://www.jmol.org/ (last accessed date October 20, 2010).

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nussinov,R. and Wolfson,H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.

Ortiz,A.R. *et al*. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, **11**, 2606–2621.

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.

Sippl,M.J. (2008) On distance and similarity in fold space. *Bioinformatics*, **24**, 872–873.

Sippl,M.J. and Wiederstein,M. (2008) A note on difficult structure alignment problems. *Bioinformatics*, **24**, 426–427.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.

Veeramalai,M. *et al*. (2008) TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model. *BMC Bioinformatics*, **9**, 358.

Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(Suppl. 2), II246–II255.

Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.