

mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet.

Sylvain Léonard, Agnel Praveen Joseph, Narayanaswamy Srinivasan, Jean-Christophe Gelly, Alexandre de Brevern

▶ **To cite this version:**

Sylvain Léonard, Agnel Praveen Joseph, Narayanaswamy Srinivasan, Jean-Christophe Gelly, Alexandre de Brevern. mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet.. Journal of Biomolecular Structure and Dynamics, 2014, 32 (4), pp.661-8. 10.1080/07391102.2013.787026 . inserm-00926338

HAL Id: inserm-00926338
https://inserm.hal.science/inserm-00926338

Submitted on 6 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

mulPBA : an efficient multiple protein structure alignment method based on a structural alphabet.

Sylvain Léonard [1,2,3,4,+], Agnel Praveen Joseph [1,2,3, 4,+,#], Narayanaswamy Srinivasan [5], Jean-Christophe Gelly[1,2,3,4] & Alexandre G. de Brevern [1,2,3,4,*]

[1] *INSERM, U665, DSIMB, F-75739 Paris, France.*

[2] *Univ Paris Diderot, Sorbonne Paris Cité, UMR_S 665, F-75739 Paris, France*

[3] *Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.*

[4] *Laboratoire d'excellence, GR-Ex, F-75739 Paris, France.*

[5] *Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.*

[#] *Present address: NCBS, Bangalore, India.*

* Corresponding author: de Brevern Alexandre G., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

E-mail: alexandre.debrevern@univ-paris-diderot.fr

[+] Both authors contributed equally

mulPBA : an efficient multiple protein structure alignment method based on a structural alphabet.

The increasing number of available protein structures requires efficient tools for multiple structure comparison. Indeed, multiple structural alignments are essential for analysis of function, evolution and architecture of protein structures. For this purpose, we propose a new web server called mulPBA (multiple Protein Block Alignment). This server implements a method based on a structural alphabet to describe the backbone conformation of a protein chain in terms of dihedral angles. This 'sequence-like' representation enables the use of powerful sequence alignment methods for primary structure comparison, followed by an iterative refinement of the structural superposition. This approach yields alignments superior to most of the rigid-body alignment methods and highly comparable to the flexible structure comparison approaches.

We implement this method in a web server designed to do multiple structure superimpositions from a set of structures given by the user. Outputs are given as both sequence alignment and superposed 3D structures visualized directly by static images generated by PyMol or through a Jmol applet allowing dynamic interaction. Multiple global quality measures are given. Relatedness between structures is indicated by a distance dendogram. Superimposed structures in PDB format can be also downloaded and the results are quickly obtained. mulPBA server can be accessed at www.dsimb.inserm.fr/dsimb_tools/mulpba/.

Keywords: amino acid; structural alphabet; Protein Blocks; progressive sequence alignment strategy; semi-global alignment, anchor-based alignment; protein folds; structural comparison; Protein Data Bank.

Introduction

It has been an important requirement to compare protein structures for the interpretation of functional, dynamic and evolutionary properties. Multiple structure comparisons are essential to obtain a simultaneous comparison of a group of structures, which is also a critical step in many modeling and threading approaches (Akutsu & Sim 1999; Panchenko, Marchler-Bauer & Bryant 1999; Dunbrack 2006). Most of the methods of structure comparison identify structural equivalences by comparing local structural fragments (Holm & Sander 1993; Shindyalov & Bourne 1998; Ortiz, Strauss & Olmea 2002; Krissinel & Henrick 2004; Ye & Godzik 2005; Konagurthu, Whisstock, Stuckey & Lesk 2006; Menke, Berger & Cowen 2008; Ilinkin, Ye & Janardan 2010). These fragment-based approaches do not require a priori knowledge of the conformation of the fragments.

A more recent group of methods attempt to classify local protein structures into a limited set of local backbone conformations before carrying out comparisons. These methods are based on libraries of local backbone structures that represent the frequently occurring regular backbone conformations. A library of local backbone conformations that can be used to abstract a complete protein backbone is called a Structural Alphabet (SA). Abstraction of structures in terms of SA helps to encode 3D information into a 1D sequence (Offmann, Tyagi & de Brevern 2007; Joseph et al. 2010; Joseph, Bornot & de Brevern 2010). Classical amino acid sequence alignment strategies can be adopted for comparison. A few methods have been developed for comparing protein structures based on structural alphabets (e.g., (Guyon, Camproux, Hochez & Tuffery 2004; Friedberg et al. 2007; Tung, Huang & Yang 2007; Ku & Hu 2008; Wang & Zheng 2008; Yang 2008)). When compared to the methods based on similarity of 3D structural measures, the approaches based on structural alphabets are significantly faster. A widely

used SA, named Protein Blocks (PBs) (de Brevern, Etchebest & Hazout 2000; Dudev & Lim 2007; Zimmermann & Hansmann 2008; Rangwala, Kauffman & Karypis 2009; Joseph et al. 2010; Suresh, Ganesan & Parthasarathy 2012) was used for 3D to 1D approximation. A curated online protein block sequence database, PDB-2-PB was recently published (Suresh et al. 2012). PBs were used to develop an efficient method for comparing two protein structures (Tyagi, Gowri, Srinivasan, de Brevern & Offmann 2006; Tyagi et al. 2006; Tyagi, de Brevern, Srinivasan & Offmann 2008). The structures were translated into PB sequences followed by the alignment of the PB sequences. The alignment was carried out with the use of an anchor-based dynamic programming algorithm which first identifies all high scoring and structurally favorable local alignments (anchors) and then aligns the segments between them to obtain a global alignment. This improved PB based structure alignment approach (iPBA) outperformed other established methods when tested on benchmark datasets (Gelly, Joseph, Srinivasan & de Brevern 2011).

We have extended the iPBA approach to the comparison of multiple structures (Joseph, Srinivasan & de Brevern 2012). A progressive strategy similar to that used in CLUSTALW (Thompson, Higgins & Gibson 1994) was adopted. PB sequence alignment determines the residue equivalences for the 3D structural fit and the fitted structures are optimized by structure based iterative refinements (Joseph et al. 2012). The web-server provides a good platform for multiple structure comparisons. Different measures for determining the quality of alignments are incorporated. A dendogram is also displayed to indicate the relative structural divergence. The proposed development also offers a user-friendly interface to view and analyse the 3D superposition along with the access to downloadable alignment files (both sequence and structural alignment).

Methods

The server can be used to compare multiple protein structures. Figure 1 presents the different steps of the method.

Input. The user can either provide the coordinates in the standard PDB format or enter the PDB code (Figure 1.1). The identifiers of chains to be compared should also be given.

Pre-Processing. The atomic coordinate sets are first translated into sequences of Protein Blocks (Figure 1.1). PBs constitute a library of 16 pentapeptide conformations (labeled from *a* to *p*) each described by a series of Φ, Ψ dihedral angles. Representing backbone conformations in terms of PBs provided a reasonable approximation (de Brevern et al. 2000) with a root mean square deviation (*rmsd*) of 0.42 Å (de Brevern 2005).

Computing pairwise alignment. The pairwise alignments are obtained using iPBA which performs an anchor based alignment by finding structurally conserved regions, identified as local alignments (Gelly et al. 2011; Joseph, Srinivasan & de Brevern 2011). The structurally conserved regions are defined for residues with Cα atoms within 3 Å (Tyagi et al. 2006). A combination of local (Huang 1991) and global (Needleman & Wunsch 1970) dynamic programming algorithms is used for the alignment (Figure 1.2). The PB substitution matrix was generated using substitution frequencies obtained from alignments of domain pairs in PALI (Balaji, Sujatha, Kumar & Srinivasan 2001) with no more than 40% sequence identity (Gelly et al. 2011).

Computing multiple alignments. A progressive multiple sequence alignment strategy similar to CLUSTALW (Thompson et al. 1994) was used. A guide tree was used to guide the assembly of sequences based on the degree of similarity (Figure 1.3). The alignment of two sequences (or groups of sequences) is carried out using dynamic

programming based on average ‘sum of pairs’ scores. The structurally conserved regions in the pairwise alignments are given higher weights during the progressive alignment, similar to the idea used in DBCLUSTAL (Thompson, Plewniak, Thierry & Poch 2000) (Figure 1.4). Optimizations of gap penalties and anchor weights have been carried out extensively to have the best multiple alignment scores (see section below) on a large dataset. Optimization was done with a dataset composed of 330 protein families with more than 2 members, from HOMSTRAD database (Mizuguchi, Deane, Blundell & Overington 1998) and 200 domain families from the recent version of PALI dataset V 2.8a (Balaji et al. 2001; Gowri, Pandit, Karthik, Srinivasan & Balaji 2003).

3D structural alignment. PROFIT (version 3.1) (Martin & Porter 2010) performs least squares fit of protein structures based on the residue equivalences in a given sequence alignment. The multiple PB sequence alignment is translated to amino acid sequence alignment which is given as input for PROFIT (Figure 1.5).

Multiple alignment scores: Different kinds of scores mainly derived from earlier works were employed (Gelly et al. 2011; Joseph et al. 2011; Joseph et al. 2012):

(1) $N_{rms:}$ The percentage of alignment columns with less than 30% of elements as gaps and *rmsd* less than 3.0Å.

(2) N_{gdt}: The percentage of aligned *positions* with less than 30% gaps and maximum distance less than a given cut-off. A weighted average of the number of columns associated with the distance cut-offs of 3.0Å, 4.0Å, 5.0Å and 6.0Å was calculated in a similar way as that of GDT score (Zemla 2003; Zemla et al. 2007).

(3) $N_{3.5}$: The *average* number of aligned residue pairs that are within a distance of 3.5Å, counted for different combinations of pairwise comparisons in the multiple alignments.

Output for multiple alignments. With the Jmol applet (JMol), users can have a 3D analysis of the superposed structures and also choose different visual structure representations. Images of the aligned structures rendered in PyMol are also provided. The residue equivalences in the 3D alignment are given as a complete sequence alignment. The corresponding PBs are also shown in the alignment. A structural distance based dendogram is provided to identify outliers in the alignment. Users can download the coordinates of the aligned structures in PDB format and PyMol scripts are also given for local analysis of the superposition. Raw output file with sequence alignment and quality scores is also downloadable in text format (see sup S1 for more details).

Discussion

The quality of alignments generated by mulPBA was compared with other popular methods available. An average gain of 84.7% in alignment quality was obtained across the different measures (N_{rms}, N_{gdt} and $N_{3.5}$), with respect to the alignments in the HOMSTRAD dataset (Mizuguchi et al. 1998). This databank is used as a reference set that encompasses more than 300 protein families superimposed. A similar comparison was also carried out with MUSTANG software (Konagurthu et al. 2006) that is used in the PALI database (Balaji et al. 2001). For more than 300 protein families, 85% of the alignments were improved with mulPBA while the other cases are quite close to MUSTANG results.

Assessments have also been carried out on a small dataset of 50 non-trivial cases randomly chosen from the twilight set in the SABMARK dataset (Van Walle, Lasters & Wyns 2005). Alignments generated by methods like MUSTANG (Konagurthu et al. 2006), MultiProt (Lupyan, Leo-Macias & Ortiz 2005) and 3DCOMB (Wang, Peng & Xu 2011) were used for comparison.

About 48 (96%) cases of alignments were of better quality than MUSTANG and 44 (88%) were better when compared to MultiProt. The difference was less striking with respect to the recent 3DCOMB methodology with 29 (58%) cases of better alignment quality. Figure 2 gives the number of cases where mulPBA have better N_{rms}, N_{gdt} and $N_{3.5}$ scores when compared to MUSTANG, MultiProt and 3DCOMB. mulPBA clearly shows high improvements when compared to widely used approaches like MUSTANG and MultiProt. In the SABmark dataset, about 6 of the alignments generated with mulPBA had large decline in the alignment quality (scores > 5) with respect to 3DCOMB. Most of the cases involved inherent flexibility of structures where the equivalences reflected in the PB alignments were not captured efficiently in the 3D structural fit. In a few of these cases, the structures involve long and multiple helices. Hence the PB sequences are characterized by long stretches of low complexity (series of PB '*m*') and this resulted in wrong residue equivalences in the alignment. Currently, 3DCOMB needs to be locally installed and no webserver is available for the community. Figure 3 shows the improvement of protein superimposition quality (N_{rms}) in regards to the sequence identity of the proteins. Figure 3a shows the quality of mulPBA with all the alignment of HOMSTRAD, while Figure 3b shows the same alignment compared to MultiProt. These representations underline the interest of mulPBA when proteins share a low sequence identity.

Figure 4 shows an example of non-trivial alignment of 5 related structures with Rossmann fold (SCOP Ids: 1gd1o1, 1gpba_, 4mdha1, 5ldha1, 6ldha1 and 8adha2)). They have been superimposed with different available servers like MASS (Dror, Benyamini, Nussinov & Wolfson 2003), MATT (Menke et al. 2008), SALIGN (Madhusudhan, Webb, Marti-Renom, Eswar & Sali 2009) and POSA (Ye & Godzik 2005). The values of N_{rms}, N_{gdt} and $N_{3.5}$ underline the difficulty of this superimposition. Both MASS and MATT give relatively lower alignment scores (N_{rms}, N_{gdt} and $N_{3.5}$). POSA has quality scores closest to that of mulPBA (Figure 4), with mulPBA having slightly better scores.

Figure 5 shows the output of the webserver. The output for protein superimposition is both interactive and non-interactive. Important alignment quality measures (alignment score from PB substitutions, N_{rms}, N_{gdt}, RMSDcore and $N_{3.5}$) are also given (Figure 5.1). The color - code allows an easy interpretation of the quality of the superimposition. The complete alignment is presented as aligned amino acid sequences along with their PB assignments and residue numbers (Figure 5.2). A dendrogram gives the structural relatedness between the proteins (Figure 5.3). As one protein could be far away from the other protein folds, leading to poor global scores, the user could also test without this outlier. Importantly, all outputs can be downloaded as simple flat files, which can be an independent file or a global archive (Figure 5.4).

Jmol applet gives an interactive view of the alignment with different rendering (Figure 5.5). PyMol figures are also provided (Figure 5.6) and a downloadable PyMol script can be easily used locally to render nice pictures. Hence the user will have access to all the important data (superimposed structures, precise alignments with residue numbers) and will also have a visual display of the results.

Conclusion

The ability to represent the complete backbone conformation of a protein chain as a sequence of characters followed by the use of sequence alignment techniques mainly distinguishes mulPBA from other structure comparison tools (Joseph et al. 2012). In terms of alignment quality and the efficiency in detecting structural relatives, mulPBA has been quite successful among the wide range of methods available. The web tool also provides an interface for the visualization and analysis of the alignments. Hence, mulPBA can be of great use to the general scientific community. Future improvements of the approach would focus on the optimization of speed as the current approach is very simple. In the same way, we would like to improve the quality of the superimposition using methodologies we developed locally (Gelly & de Brevern 2011).

Figures legend

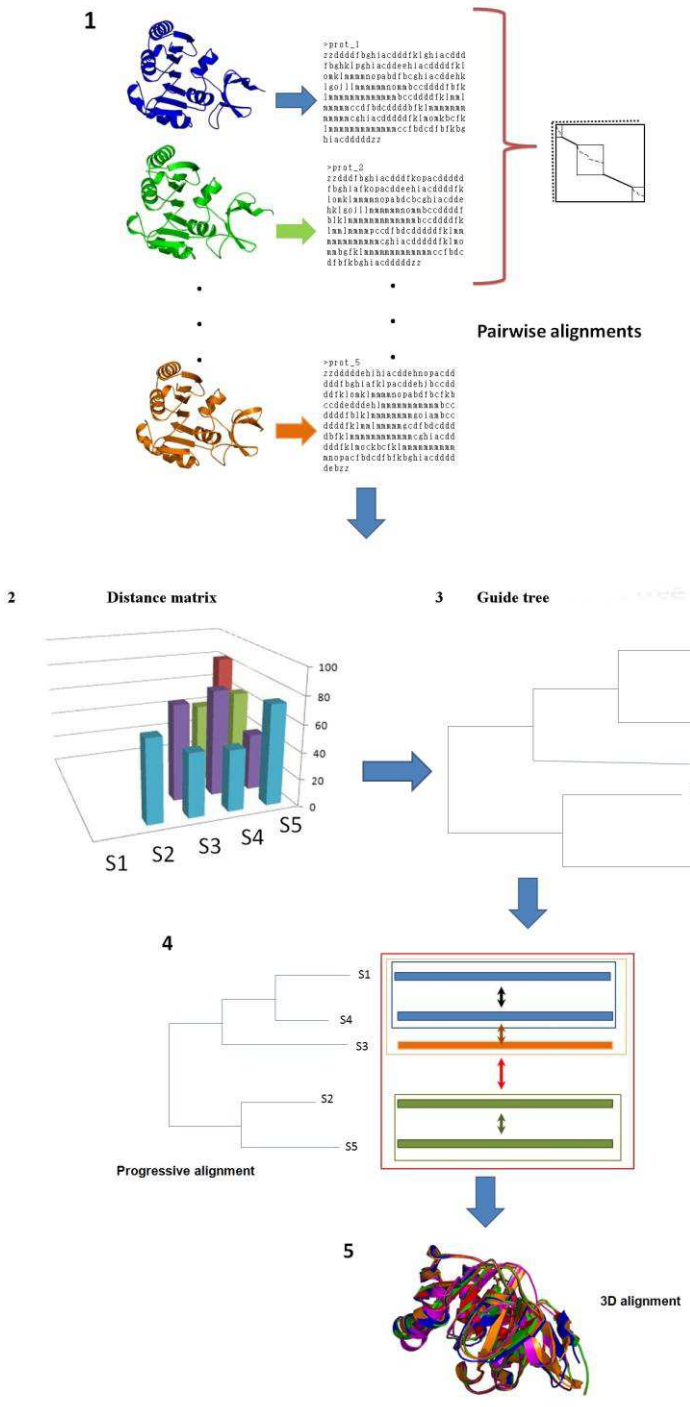


Figure 1. *mulPBA approach*. (1) After uploading the protein structures (PDB files), the protein 3D structures are encoded as series of 1D Protein Blocks; each PB sequence is then aligned to the others. (2) These pairwise alignments are used to generate a distance matrix which helps to (3) build a dendogram. (4) This tree guides a progressive alignment leading to the final multiple structural alignment. (5) The multiple PB alignment is then translated into a 3D alignment.

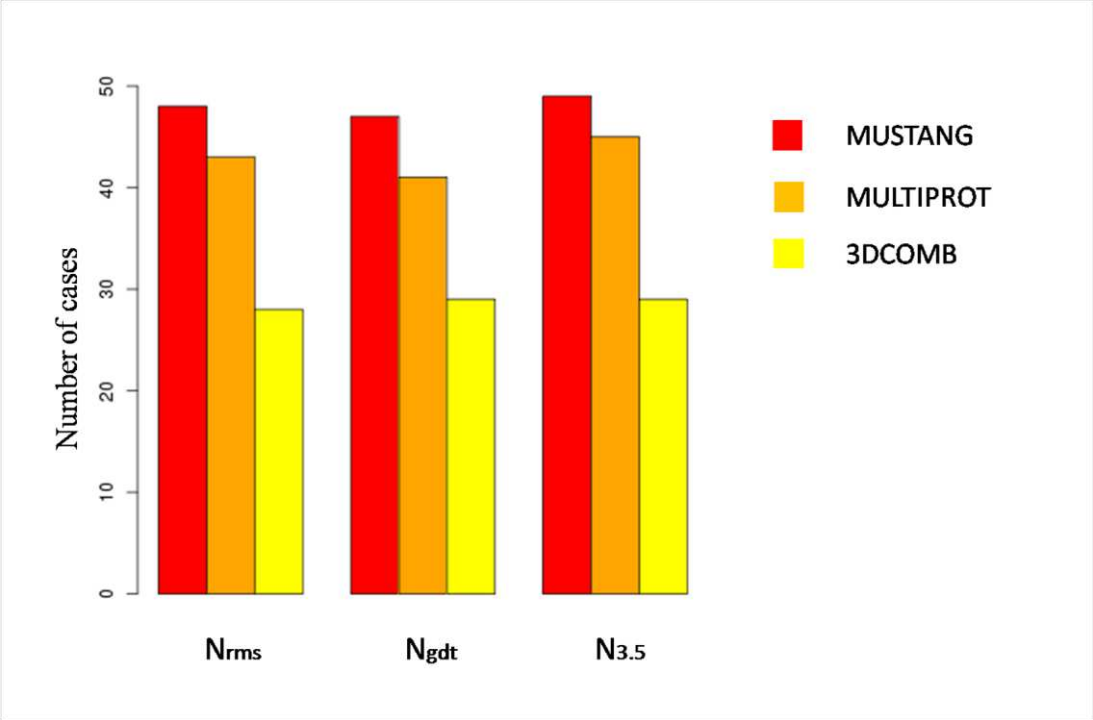


Figure 2. *Comparison of different MSTAs*. Out of 50 difficult cases, the number of alignments where mulPBA gives better N_{rms}, N_{gdt} and $N_{3.5}$ scores, are given.

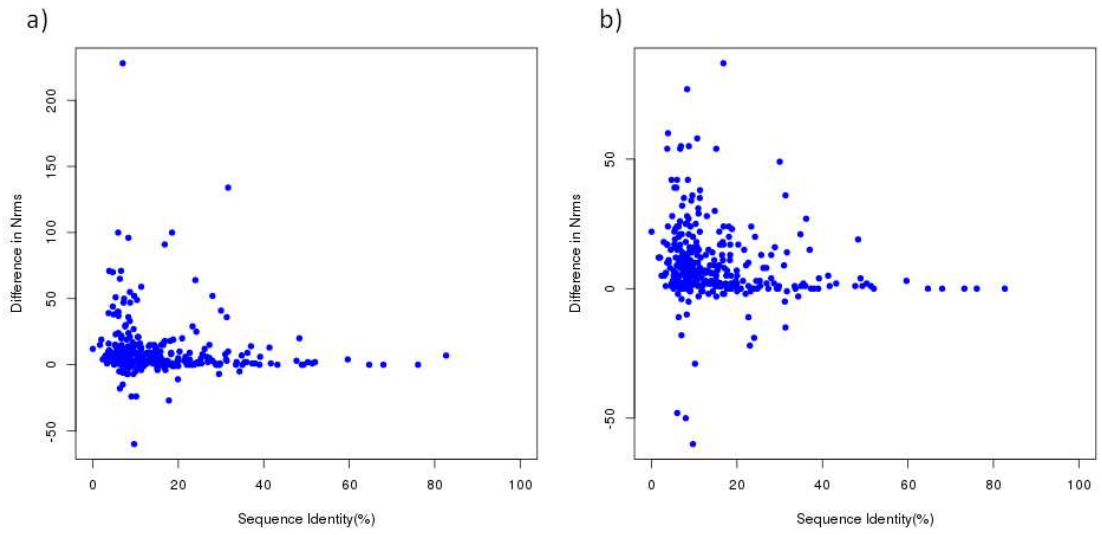


Figure 3. *Comparison of different MSTAs in regards to protein sequence identity*. N_{rms} vs. Sequence identity (a) for the alignments of Homstrad and (b) for the same proteins superimposed by Multiprot.

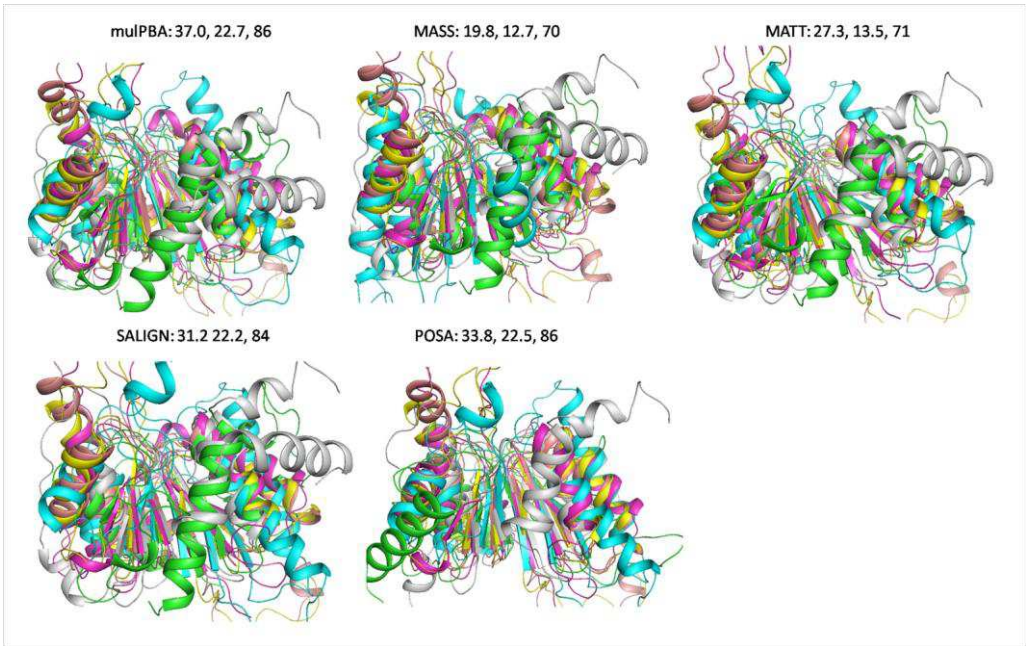


Figure 4. *Multiple structure superimpositions of 5 structures having Rossmann fold*. The 3D superposition and corresponding N_{rms}, N_{gdt} and $N_{3.5}$ scores are given.

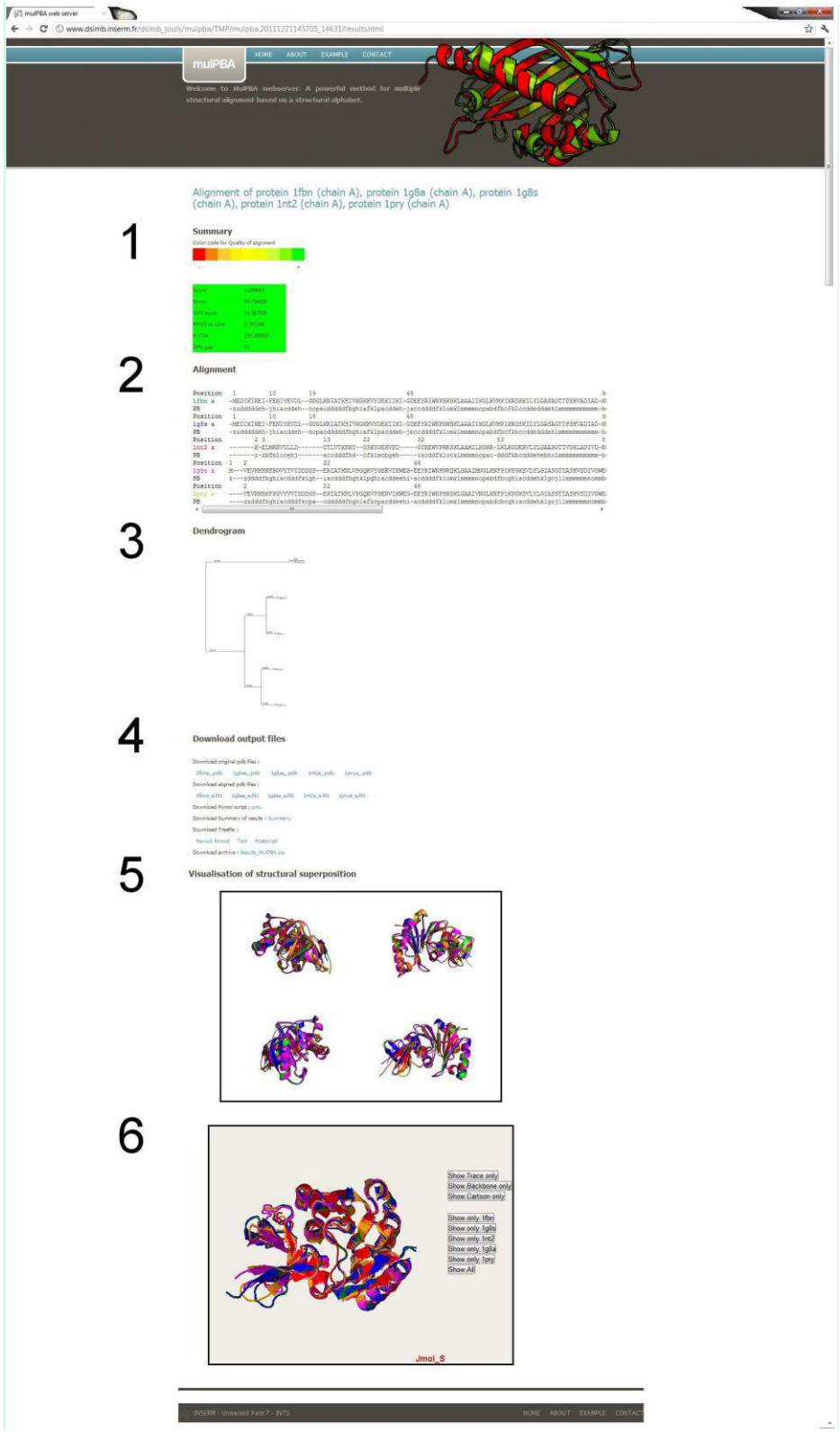


Figure 5. *Outputs:* (1) Indication of quality of the alignment based on different scores, (2) The structure based sequence alignment along with PBs and residue numbers, (3) The tree used to build the multiple structure alignment, (4) The web server output including PyMol pictures (and script file) along with summary information and the coordinates of the superimposed structures, (5) The superimposed 3D structures rendered with PyMol and (6) 3D interactive visualization of the final alignment in Jmol.

References

Akutsu, T. & Sim, K. L. (1999). Protein Threading Based on Multiple Protein Structure Alignment. *Genome Inform Ser Workshop Genome Inform 10*: 23-29.

Balaji, S., Sujatha, S., Kumar, S. S. & Srinivasan, N. (2001). PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res 29*: 61-65.

de Brevern, A. G. (2005). New assessment of a structural alphabet. *In Silico Biol 5*: 283-289.

de Brevern, A. G., Etchebest, C. & Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins 41*: 271-287.

Dror, O., Benyamini, H., Nussinov, R. & Wolfson, H. (2003). MASS: multiple structural alignment by secondary structures. *Bioinformatics 19 Suppl 1*: i95-104.

Dudev, M. & Lim, C. (2007). Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. *BMC Bioinformatics 8*: 106.

Dunbrack, R. L., Jr. (2006). Sequence comparison and protein structure prediction. *Curr Opin Struct Biol 16*: 374-384.

Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z. & Godzik, A. (2007). Using an alignment of fragment strings for comparing protein structures. *Bioinformatics 23*: e219-224.

Gelly, J. C. & de Brevern, A. G. (2011). Protein Peeling 3D: new tools for analyzing protein structures. *Bioinformatics 27*: 132-133.

Gelly, J. C., Joseph, A. P., Srinivasan, N. & de Brevern, A. G. (2011). iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res 39*: W18-23.

Gowri, V. S., Pandit, S. B., Karthik, P. S., Srinivasan, N. & Balaji, S. (2003). Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res 31*: 486-488.

Guyon, F., Camproux, A. C., Hochez, J. & Tuffery, P. (2004). SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res 32*: W545-548.

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

Huang, X., Miller, W (1991). A time-efficient linear-space local similarity algorithm. *Advances in Applied Mathematics 12*: 337 - 357.

Ilinkin, I., Ye, J. & Janardan, R. (2010). Multiple structure alignment and consensus identification for proteins. *BMC Bioinformatics 11*: 71.

JMol (2012). Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ *http://www.jmol.org/*

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L. S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadié, H., Schneider, B., Cadet, F., Srinivasan, N. & de Brevern, A. G. (2010). A short survey on protein blocks. *Biophys Rev 2*: 137-145.

Joseph, A. P., Bornot, A. & de Brevern, A. G. (2010). *Local Structural Alphabet*, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Joseph, A. P., Srinivasan, N. & de Brevern, A. G. (2011). Improvement of protein structure comparison using a structural alphabet. *Biochimie 93*: 1434-1445.

Joseph, A. P., Srinivasan, N. & de Brevern, A. G. (2012). Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies. *Biochimie*: in press.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins 64*: 559-574.

Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr 60*: 2256-2268.

Ku, S. Y. & Hu, Y. J. (2008). Protein structure search and local structure characterization. *BMC Bioinformatics 9*: 349.

Lupyan, D., Leo-Macias, A. & Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics 21*: 3255-3263.

Madhusudhan, M. S., Webb, B. M., Marti-Renom, M. A., Eswar, N. & Sali, A. (2009). Alignment of multiple protein structures based on sequence and structure features. *Protein Eng Des Sel 22*: 569-574.

Martin, A. & Porter, C. (2010). *http://www.bioinf.org.uk/software/profit/*.

Menke, M., Berger, B. & Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol 4*: e10.

Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci 7*: 2469-2471.

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

Offmann, B., Tyagi, M. & de Brevern, A. G. (2007). Local Protein Structures. *Current Bioinformatics 3*: 165-202.

Ortiz, A. R., Strauss, C. E. & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci 11*: 2606-2621.

Panchenko, A., Marchler-Bauer, A. & Bryant, S. H. (1999). Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins Suppl 3*: 133-140.

Rangwala, H., Kauffman, C. & Karypis, G. (2009). svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinformatics 10*: 439.

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

Suresh, V., Ganesan, K. & Parthasarathy, S. (2012). PDB-2-PB: a curated online protein block sequence database. *J. Appl. Cryst. 45*: 127-129

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (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*: 4673-4680.

Thompson, J. D., Plewniak, F., Thierry, J. & Poch, O. (2000). DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res 28*: 2919-2926.

Tung, C. H., Huang, J. W. & Yang, J. M. (2007). Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol 8*: R31.

Tyagi, M., de Brevern, A. G., Srinivasan, N. & Offmann, B. (2008). Protein structure mining using a structural alphabet. *Proteins 71*: 920-937.

Tyagi, M., Gowri, V. S., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins 65*: 32-39.

Tyagi, M., Sharma, P., Swamy, C. S., Cadet, F., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res 34*: W119-123.

Van Walle, I., Lasters, I. & Wyns, L. (2005). SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics 21*: 1267-1268.

Wang, S., Peng, J. & Xu, J. (2011). Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics 27*: 2537-2545.

Wang, S. & Zheng, W. M. (2008). CLePAPS: fast pair alignment of protein structures based on conformational letters. *J Bioinform Comput Biol 6*: 347-366.

Yang, J. (2008). Comprehensive description of protein structures using protein folding shape code. *Proteins 71*: 1497-1518.

Ye, Y. & Godzik, A. (2005). Multiple flexible structure alignment using partial order graphs. *Bioinformatics 21*: 2362-2369.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res 31*: 3370-3374.

Zemla, A., Geisbrecht, B., Smith, J., Lam, M., Kirkpatrick, B., Wagner, M., Slezak, T. & Zhou, C. E. (2007). STRALCP--structure alignment-based clustering of proteins. *Nucleic Acids Res 35*: e150.

Zimmermann, O. & Hansmann, U. H. (2008). LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J Chem Inf Model 48*: 1903-1908.