Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet.

Manoj Tyagi, Priyanka Sharma, C. S. Swamy, Frédéric Cadet, N. Srinivasan, Alexandre De Brevern, Bernard Offmann

Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet

M. Tyagi[1], P. Sharma[1], C.S. Swamy[2], F. Cadet[1], N. Srinivasan[1,2], A.G. de Brevern[3] and B. Offmann[1,*]

[1]Laboratoire de Biochimie et Génétique Moléculaire, Bioinformatics Team, Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France.

[2]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India.

[3] INSERM, U726, Equipe de Bioinformatique et Génomique Moléculaire (EBGM), Université Paris 7 – Denis Diderot, case 7113, 2, place Jussieu, 75251 Paris Cedex 05, France.

* Corresponding author: bernard.offmann@univ-reunion.fr

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in Nucleic Acids Research following peer review. The definitive publisher-authenticated version 2006 Jul 1;34(Web Server issue):W119-23 is available online at: http://nar.oxfordjournals.org/cgi/content/full/34/suppl_2/W119

ABSTRACT

Encoding protein 3D structures into 1D string using short structural prototypes or structural alphabets opens new front for structure comparison and analysis. Using the well-documented 16 motifs of Protein Blocks (PBs) as structural alphabet, we have developed a methodology to compare protein structures that are encoded as sequences of PBs by aligning them using dynamic programming which uses a substitution matrix for PBs. This methodology is implemented in the applications available in Protein Block Expert (PBE) server. PBE addresses common issues in the field of protein structure analysis such as comparison of proteins structures and identification of protein structures in structural databanks that resemble a given structure. PBE-T provides facility to transform any PDB file into sequences of PBs. PBE-ALIGNc performs structure comparison between two protein structures based on the alignment of their corresponding PB sequences. PBE-ALIGNm is a facility for mining SCOP database for similar structures based on the alignment of PBs. Besides, PBE provides an interface to a database (PBE-SAdb) of preprocessed PB sequences from SCOP culled at 95% and of all against all pairwise PB alignments at family and superfamily level. PBE server is freely available at http://bioinformatics.univ-reunion.fr/PBE/.

INTRODUCTION

The central paradigm of protein science suggests that protein functions are directly controlled by protein structures. With the increasing number of solved protein structures, structure comparison methods are becoming increasingly important. A number of semi or fully automated structure comparison methods have been developed based on

methodologies like alignment of secondary structure elements (1-3), environmental profiles (4) and distance measure matrices (5).

Most of these methods use regular secondary structure information in their algorithms. By analyzing local protein structures, many groups have found recurring short structural motifs also called structural alphabet (SA) spanning structural space (6-8). These short motifs represent local structure variations in protein space upon which backbone model of most proteins can be built. They have been shown to be informative to analyze protein structures (9) and have been used in structure prediction (10), backbone reconstruction (11,12) and loop modeling (13).

We present a web based service called Protein Block Expert (PBE) for protein structure comparison and analysis using a SA of 16 pentapeptide structural motifs known as "Protein Blocks" (PBs) (14,15). A protein structure can be encoded into sequence of PBs by sliding an overlapping window of five residues. Hence, simplified 1D representation of protein structure can be used just like amino acid sequence analysis to find similarity, dissimilarity and relationship among proteins in terms of structure. PBE is similar to classical sequence alignment (16,17). Its concept is related to SA-Search (18) web server, but differs greatly as it uses a genuine SA substitution matrix derived on the basis of aligned homologous proteins present in the large *Phylogeny and ALIgnment of homologous protein structures* (PALI) database (19,20). Applications and validation of such a matrix have been shown (Tyagi et al., *submitted*). PBE is not only a service to find structural similarities between proteins or a mining tool for recognizing the fold of a protein structure, it also provides an interface to a database to study proteins in terms of

PBs at the levels of super-family and family. PBE provides the following features to the user :

- A tool to encode protein structure into PBs sequence.
- Structure comparison between a pair of proteins using PB description using both local and global alignment algorithms.
- Mining a databank based on SCOP for proteins with similar fold.
- Access to a database of preprocessed PB sequences and pairwise alignments at family and super-family level based on SCOP.

PBE is freely accessible at http://bioinformatics.univ-reunion.fr/PBE/

PBE-T : ENCODING PROTEIN STRUCTURE INTO PROTEIN BLOCKS

Proteins Blocks (PBs) are a set of 16 structural motifs of five residue long representing local structural features of protein (14,15). Each of the PBs is represented by a vector of eight ϕ,ψ dihedral angles associated with five consecutive Cα atoms and are denoted by the letters *a, b, …,p.* Encoding of protein 3D structure into sequence of PBs as implemented in our server is a two step process. First, protein backbone is encoded into sequence of (ϕ,ψ) angles calculated from backbone atomic positions. Second, an overlapping window of five Cα atoms *i.e.* vector of eight (ϕ,ψ) angles is moved along the backbone. PBs for each window is assigned on the basis of smallest dissimilarity measure called root mean square deviation on angular values or *rmsda* (21) between observed ($\phi;\psi$) values in the window and the standard dihedral angles for various PBs. PBs have been used in several prediction methods (22-24). PBE-T allows easily this encoding. It takes a structure or a structural model and gives its direct transcription in terms of PBs.

PBE-ALIGNc : PROTEIN STRUCTURE COMPARISON USING PROTEIN BLOCKS

Analysis of sequence of PBs using classical amino acid sequence alignment algorithms allows us to explore possibility of finding structural similarities between two proteins using reduced complexity of protein structure. Protein Block Expert (PBE) server has been designed and implemented to fulfill this requirement. It allows user to compare two proteins using simple dynamic programming (DP) algorithm by aligning two PB sequences using our PB substitution matrix.

The substitution table used in our study was derived by re-encoding in terms of PBs the structurally aligned homologous proteins present in the PALI database (19,20). The detailed description of calculation, discussion on PB substitution matrix and proposed applications are reported elsewhere (Tyagi et al., *submitted*).

Indeed, local structural similarities between two uploaded protein structures are found using PB sequence alignment. This approach has already been successfully benchmarked and compared to standard flexible alignment methods like DALI (5) or rigid body superposition methods like STAMP (25) where more than 75% of structurally equivalent residues in our PB alignment method overlapped with those identified with these standard methods. Moreover, careful inspection of aligned coordinates from PB-ALIGNc after aligning PBs ~~indeed~~ shows identical and in some instances, more favorable rmsd values than DALI ~~for example~~. These results are expected to improve by using more robust dynamic programming algorithm combined with optimized gap penalty. Interestingly, in the same study we have shown how PB alignment method is able to pick up subtle similarities at local level between two proteins which may be missed by standard alignment methods (i.e. possible ?). Hence PB alignment is providing both local and global flavors of dynamic programming algorithms.In PBE-ALIGNc, the user is required to upload two protein structures in PDB format. After transforming the 3D structures into 1D PB sequences and the latter are aligned using DP algorithm. If the uploaded protein structures have more than one chain, option to select any one of the possible pair for alignment is presented. Once the selection has been done the selected pair is aligned. The output displays the aligned PB sequences along with the information like length of proteins, alignment length, best fit superposition *rmsd* value using ProFit

program based on McLachlan algorithm (26). PB alignment is transformed into amino acid alignment to define equivalent regions required by ProFit and further iterations are done to obtain best fit *rmsd* value. The server provides the possibility to download the initial PB and corresponding amino acid alignments in Fasta format as well as the superimposed coordinates between the two structures. As PBE requires only backbone atoms to generate PB sequence and is independent of residues, the user can upload anonymous protein structures by changing all residues to any one kind and giving only coordinates of backbone atoms in PDB format. Hence newly solved structures can be easily analyzed without making them public.

PBE-ALIGNm : MINING SCOP DATABASE FOR PROTEINS WITH SIMILAR FOLD

Database of protein domains based on SCOP (27) classification has been used. Protein structures were extracted from SCOP 1.65 via the ASTRAL (28) server using a sequence identity cutoff of 95% (SCOP95) with 9392 domains. SCOP95 culled from ASTRAL server was a big enough databank to cover protein space. These domains were encoded into PB sequences and are made available for user to query at family and superfamily level in PBE-SAdb database. Further, an extensive all-against-all pairwise PB sequence alignments between all 7195 domains were generated using dynamic programming and our PB substitution matrix. Protein domains in SCOP95 having any chain breaks were not considered for PB sequence alignment process. Pairwise alignments within each seven major class from SCOP95, which amounts 5405433 alignments, are featured in PBE-SAdb database where option is provided to the user to

view/download pairwise PB alignments at the level of family or super-family. Each PB alignment in the generated databank had raw score given by DP algorithm. To remove the dependence of this value on the length of the two proteins, the score was normalized by dividing it with the length of the alignment including gaps. This normalized score from global alignment algorithm is used to rank alignments during the following analysis.

In the first study we analyzed the efficiency of our method to discriminate between various SCOP classes or in other words what is the confusion between classes based on PB sequence alignment. This question is important since 1D representation of protein structure using PB sequence lacks in topological information, which can create confusion due to identical linear sequence of regular structures in two proteins having different topologies. A dataset of 1500 protein domains was selected randomly from SCOP at 95% keeping the relative proportion of seven major classes same as in original databank. All-against-all PB sequence alignment for this dataset was performed. A jackknife approach was adopted to perform comprehensive analysis. Each time one domain was selected and was queried against the databank to find top 10 ranking PB alignments against the given query and statistics was calculated for true hits at each rank position. Appearance of same SCOP CLASS among top 10 ranks was considered a true hit. Analysis of the distribution of true hits shows that 85.9% of them are at first rank and a hit rate of 98.2 % is achieved when first 10 ranking alignments are considered (data not shown). It should be noted that the value increases from 85.9% to 93% when same analysis is performed on the 7267x7267 pairwise alignment.

A confusion matrix between seven SCOP classes is also calculated taking into account only top hit for each query. Matrix is populated simply based on criteria if query

protein and first rank protein have same class or not. Table 1 shows the generated matrix with first four classes shown in different color scheme. Among all the four classes, alpha plus beta class was most confused class with only in 76.2 % cases finding itself at first rank. Beta class was most well behaved class with accuracy of 94.4% followed by alpha beta and alpha class. Low accuracy rate of multi-domain and membrane class can be attributed to very low number of proteins present in given experiment.

In a second study, we assessed how well a PB alignment can extract protein of similar fold from a databank within given a class. A jackknife approach (as done in previous analysis, cf. *infra*) was applied to calculate statistics for identifying true FOLD of a protein as defined by SCOP at various levels. Figure 1 shows the distribution of true hits at different rank positions. It is noteworthy that 81.3% of true hits are from first rank while 89.3% true hits are within top 10 ranking alignments.

Further efficiency of mining similar folds within each seven major classes was studied and results are reported in table 2. For each class, hit rate was calculated at three different levels, top10, top5 and top1 where first 10, 5 and first ranking alignments were considered respectively. The ability of our method to extract same SCOP FOLD within top10 level vary from a hit rate of 86.1% for alpha class to 93.6% for alpha/beta class. Similarly at top1 level, the hit rate varies from 70% (small protein class) to 88.4 % (alpha/beta class). Consistent good level of hit rates across various classes to mine similar fold using PB alignment method gives support to basic ability of the method and quality of the substitution matrix.

These results hence illustrates that the use of PB substitution matrix with simple DP algorithm along with naïve scoring function is efficient to extract proteins sharing structural similarities from large dataset.

PBE-ALIGNm provides this facility for mining structural similarities from a databank using a reduced representation of protein structures. User can upload a protein structure in PDB format and can decide against which databank the structure is to be queried. PBE gives option to select local or global alignment algorithm, setting up parameters like minimum length of proteins against which query should be aligned. Option is also given to decide if you want to align against whole databank or with some specific SCOP CLASS proteins. Typically, the runtime for a query is less than one minute.

INTERFACE TO PBE DATABASE

PBE server provides another feature for protein structure analysis using structural alphabets. We have created two databases of protein structures and are grouped under the PBE-SAdb facility. First is a database of 9392 protein domains extracted from SCOP95 that were translated into PB sequences. Second is a database of all possible pairwise PB sequence alignments within each SCOP class.

In both instances, an interface gives option for querying at superfamily or family level by entering appropriate SCOP code. List of all family and superfamily codes and their description present in our database is available in our help section. In addition PB sequences or alignments can also be accessed by specifying a PDB id of a protein. Because PDB was filtered for 95% sequence identity cut-off, the list of the available PDB

structures can also be checked in help section. Outputs can be easily downloaded with PB sequences or PB alignments in Fasta format.

Facility to query and download PB sequences or PB alignments at family or superfamily level is expected to be of great help in studying protein structure conservation. This can also aid studies on variations in homologous proteins in terms of structural alphabets, which might provide better insight into sequence to structure relationship. Analysis of PB alignments to study conservation or variability of local structures is expected to provide better understanding of relationship between structure and function of homologous proteins.

TECHNICAL ASPECTS

Pairwise PB alignments for each given class were calculated using 32 processor IBM AIX52 machine. Database for PB sequences and alignments are maintained using MySQL server. Web server front end and back end processing are handled using HTML, CGI and PERL scripting along with JAVA (PBs encoding of protein) and C (dynamic programming) programs. Job requests in PBE-ALIGNm are queued and provided a randomly generated job-id that guarantees the inaccessibility of jobs to other users of the servers. PBE Server is maintained on a Linux-based single processor machine and is accessible at http://bioinformatics.univ-reunion.fr/PBE/.

DISCUSSION AND PERSPECTIVES

Decrease in the complexity of protein space from three dimensions to one dimension with the combination of sequence analysis methods to study protein structure has opened a simple and exciting way of looking at protein structure space. Initial results

of mining similar fold structures in databases and finding local structural similarities between proteins has been a promising start though far from exploiting full potential of such methodology. Low confusion rate across various SCOP classes and high efficiency rate to mine similar fold protein from large database based on naïve scoring scheme indicates that PB alignment method is efficient enough to discriminate between different topologies despite lack of topological information,. This success can be attributed to both, efficiency of PBs to represent local structural properties in more refined form compared to simple SSE representation and quality of substitution matrix. Sequence of PBs between regular SSEs and their alignment or misalignment might be playing important role in discriminating true from false.

Pairwise comparison of proteins using PBE-ALIGNc performed decently when compared to standard methods like DALI and this was further been validated on a large-scale basis here (7195x7195 pairwise alignments). Structure alignment using PBs has also shown its efficiency to locate subtle similarities at local level and to very efficiently mine for local structural similarities from large structural databases. Still some fine-tuning e.g. gap penalty optimization, is required to obtain better results at the level of residue-residue alignment. Though, PB alignment is expected to be very advantageous in cases of distantly related proteins where residue-residue alignment is difficult to obtain..

Further, selection of more robust DP algorithm, calculation of statistical significance of alignment, confidence index for an alignment are few of the areas where we have to look into. Future work will also include analysis of the method to extract similar proteins at level of family and super family. Usage of class specific PB substitution matrix to mine similar folds will be of active interest.

Finally, because prediction of protein backbone in terms of PB sequences is possible from amino acid sequence (14), this work opens up interesting perspectives for large scale structural annotation of genomic sequences.

ACKNOWLEDGEMENT

We would thank all the colleagues both within the group and outside who gave useful feedback during the testing phase of this web server. We would like to thank informatics department for their support during the access of super computer facility. This work has been supported by a PhD grant to MT from Conseil Régional de La Réunion. NS is a senior fellow of the Wellcome Trust, London and is a visiting professor to the Reunion University. We thank anonymous reviewers for fruitful comments.

REFERENCES

1. Gibrat, J.F., Madej, T. and Bryant, S.H. (1996) Surprising similarities in structure comparison. *Curr Opin Struct Biol*, **6,** 377-385.

2. Singh, A.P. and Brutlag, D.L. (1997) Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proc Int Conf Intell Syst Mol Biol,* **5,** 284-293.

3. Lu, G. (2000) TOP: a new method for protein structure comparisons and similarity searches. *J Appl Crystallogr*, **33,** 176–183.

4. Jung, J. and Lee, B. (2000) Protein structure alignment using environmental profiles. *Protein Eng*, **13,** 535-543.

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

6. Jones, T.A. and Thirup, S. (1986) Using known substructures in protein model building and crystallography. *Embo J,* **5,** 819-822.

7. Unger, R., Harel, D., Wherland, S. and Sussman, J.L. (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins,* **5,** 355-373.

8. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, **226,** 507-533.

9. Unger, R. and Sussman, J.L. (1993) The importance of short structural motifs in protein structure analysis. *J Comput Aided Mol Des*, **7,** 457-472.

10. Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, **281,** 565-577.

11. Park, B.H. and Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, **249,** 493-507.

12. Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. (2002) Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, **323,** 297-307.

13. Fourrier, L., Benros, C. and de Brevern, A.G. (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics*, **5,** 58.

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

15. de Brevern, A.G. (2005) New assessment of a structural alphabet. *In Silico Biology*, **5,** 26.

16. Needleman, S.B. and 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.

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

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

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

20. Gowri, V.S., Pandit, S.B., Karthik, P.S., Srinivasan, N. and 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.

21. Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D. and Wrede, P. (1996) Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng*, **9,** 833-842.

22. de Brevern, A.G., Valadie, H., Hazout, S. and Etchebest, C. (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci*, **11,** 2871-2886.

23. Etchebest, C., Benros, C., Hazout, S. and de Brevern, A.G. (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins*, **59,** 810-827.

24. de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C. and Etchebest, C. (2005) A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta*, **1724,** 288-306.

25. Russell, R.B. and Barton, G.J. (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol*, **244,** 332-350.

26. McLachlan, A.D. (1982) Rapid Comparison of Protein Structres. *Acta Cryst*, **A38,** 871-873.

27. Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res*, **28,** 257-259.

28. Brenner, S.E., Koehl, P. and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, **28,** 254-256.

FIGURES

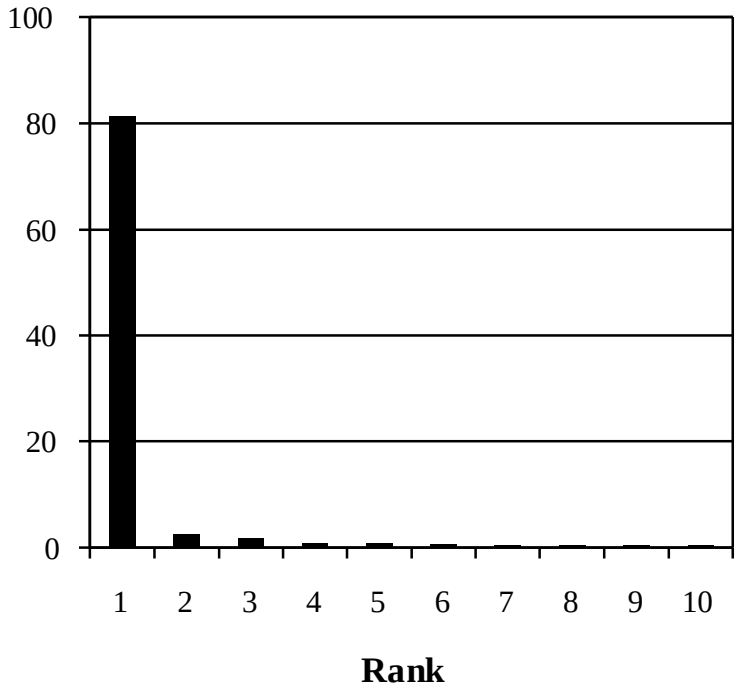

Figure 1. Mining SCOP for similar structures using PB alignment. Distribution of number of hits in top 10 ranking alignments. If a given query and extracted alignment have same FOLD, a hit is counted at that position.

TABLES

Table 1. Mining SCOP for similar structures using PB alignment. Confusion matrix between true (vertical) and predicted (horizontal) SCOP classes.

True class vs hit class	ALPHA	BETA	ALPHABETA	APLUSB	MULTIDOM	MEMBRANE	SMALL	Total
ALPHA	245 (88.1%)	1	12	9	1	5	5	278
BETA	2	404 (94.4%)	5	10	0	1	6	428
ALPHABETA	3	5	255 (89.5%)	18	3	0	1	285
APLUSB	16	23	27	240 (76.2%)	0	1	8	315
MULTIDOM	0	0	5	2	11 (61.1%)	0	0	18
MEMBRANE	10	5	0	1	1	12 (41.3%)	0	29
SMALL	2	15	0	8	0	0	122 (84.7%)	144
								1500

Table 2. Mining SCOP for similar structures using PB alignment. Hit rates (in percentage) for identifying similar FOLD within each SCOP classes. Are given rates that take into account top 10, 5 and 1 ranking alignments. Exact numbers for each case is given within brackets.

SCOP class	Top10 (%)	Top5 (%)	Top1 (%)
Alpha (1312)	86.1 (1130)	82.6 (1087)	75.0 (985)
Beta (2076)	92.9 (1930)	91.4 (1897)	87.2 (1811)
AlphaBeta (1386)	93.6 (1298)	92.0 (1275)	88.4 (1226)
AplusB (1500)	88.3 (1325)	86.3 (1294)	81.3 (1219)
Small (700)	87.7 (614)	84.3 (590)	70.3 (492)
Membrane (139)	91.4 (127)	89.2 (124)	81.3 (113)
MultiDomain (82)	85.4 (70)	84.1 (69)	81.7 (67)