Nucleic Acids Research, 2001, Vol. 29, No. 1 323–326

BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations

Anne Bahr, Julie D. Thompson, J.-C. Thierry and Olivier Poch*

Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP), BP 163, 67404 Illkirch Cedex, France

Received August 22, 2000; Accepted October 4, 2000

ABSTRACT

BAliBASE is specifically designed to serve as an evaluation resource to address all the problems encountered when aligning complete sequences. The database contains high quality, manually constructed multiple sequence alignments together with detailed annotations. The alignments are all based on three-dimensional structural superpositions, with the exception of the transmembrane sequences. The first release provided sets of reference alignments dealing with the problems of high variability, unequal repartition and large N/C-terminal extensions and internal insertions. Here we describe version 2.0 of the database, which incorporates three new reference sets of alignments containing structural repeats, transmembrane sequences and circular permutations to evaluate the accuracy of detection/prediction and alignment of these complex sequences. BAliBASE can be viewed at the web site http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/index.html or can be downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE2/.

INTRODUCTION

The alignment of protein sequences is an essential tool in molecular biology and genome analysis. A good alignment is crucial to reliable phylogenetic studies and in the identification of characteristic motifs and conserved residues in protein families. However, the quality of an alignment is strongly dependent on the domain organization and the residue composition of the sequences to be aligned. In previous studies, we addressed the problems of aligning sequences with low similarity, and sequences containing large internal insertions or N/C-terminal extensions (1,2). The remaining problems encountered by most alignment programs concern mainly protein families containing either repeats, low complexity regions such as transmembrane helices, inverted domains or circular permutations. Obviously, proteins containing repeated domains/motifs or circular permutations pose particular problems for global alignment programs. In the case of transmembrane proteins, alignment difficulties are related to the important bias in residue composition in the transmembrane regions and the frequent insertion of large domains inbetween. Effectively, these proteins contain a large proportion of fairly conserved aliphatic amino acids distributed in multiple transmembrane helices and the optimal alignment is often uncertain.

A number of programs have been developed to detect repeats in protein sequences, either by comparison with an existing database of known repeats (3,4), or *ab initio* (5,6). Much more work has been done on transmembrane prediction (7–12). The problem is in some ways similar to repeat detection since transmembrane regions may be considered as 'repeats with low sequence complexity'. Although this bias in residue composition allows easier detection of transmembrane helices, the alignment of such sequences is more problematic. Concerning the detection of circular permutations, very few algorithms have been developed so far (13). However, few of these methods consider the global alignment of such proteins. In order to construct a multiple alignment, three steps are necessary: (i) detection of the repeat, transmembrane helix or permutated domains, (ii) local alignment of these elements and (iii) integration of the local alignment into a global family alignment. Comprehensive evaluation of the programs specifically dedicated to repeats, transmembrane or circular permutation requires a large number of accurate reference alignments that can be used as test cases. For this reason, we have developed BAliBASE (2.0), a new version of a manually refined multiple sequence alignment database.

CONTENTS OF BALIBASE 2.0

The first version of BAliBASE (1) was dedicated to the evaluation of multiple alignment programs and was divided into five hierarchical reference sets of (i) equidistant sequences with various levels of conservation, (ii) families aligned with a highly divergent 'orphan' sequence, (iii) subgroups with <25% residue identity between groups, (iv) sequences with N/C-terminal extensions and (v) internal insertions. For release 2.0 of BAliBASE, these alignments have been manually verified and corrected by superposition of all known three-dimensional structures, using the lsqman program (14).

BAliBASE 2.0 includes three new alignment references sets (references 6–8) specifically dedicated to repeats, circular permutations and transmembrane proteins. The protein families included in the new reference sets were selected from the Pfam database (15) or from the literature and new family

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 00; Fax: +33 3 88 65 32 01; Email: poch@igbmc.u-strasbg.fr

Table 1. New reference sets in BAliBASE version 2.0

A					
	Alignment name	No. of sequences	Total no. of repeats	Max. repeats per sequence	Mean repeat length
Reference 6, repeats	SH3	111	174	5	55
	ZF	29	260	36	25
	APO	46	419	14	22
	SUSHI	45	385	37	55
	MYB	141	284	3	50
	ANK	30	214	24	30
	KRINGLE	35	187	38	100
	DEAD	22	31	2	230
	TRK	33	45	2	84
	FAA	37	39	2	200
	LRR	24	339	41	24
	ION_TM	49	153	4	250
Total		602	2316		
B					
	AlignmentName	No. of sequences	No. of transmembrane helices per sequence		
Reference 7, transmembrane	ION	52	6		
	ACR	43	12		
	DTD	56	12		
	PTGA	51	12		
	NaT	59	14		
	PHOTO	33	7		
	MSL	13	2		
	7TM	128	7		
Total		422			
C					
	Alignment Name	No. of sequences			
Reference 8, circular permutations	PTGA	20			
	LECTIN	31			
	GSH	18			
	SH3/SH2	44			
	CELLULASE	5			
Total		118			

members were detected by searching the sequence databases with BLAST (16). The alignments were confirmed by superposition of the three-dimensional structures where possible. As very few three-dimensional structures of transmembrane proteins have been resolved, a Gribskov profile (17) was constructed of the known family members and new sequences were included in the alignment based on the score obtained in the profile search. BAliBASE 2.0 now consists of 167 reference alignments, with more than 2100 sequences. The three new reference sets (references 6–8) contain 26 protein families with 12 distinct repeat types, eight transmembrane families and five families with inverted domains (Table 1), representing more than 1100 sequences. As in references 1–5, core blocks are defined that only include the regions of the sequences that can be reliably aligned.

BAliBASE 2.0 can be viewed on the WWW at http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2 or can be downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE2/. For each new reference set, the alignments are provided in both RSF and MSF formats with an associated annotation file containing a description of the alignment, including the number and position of repeats/transmembrane helices. The unaligned sequences are also available in FastA format.

Reference 6

The difficulties encountered when detecting or aligning proteins containing repeats are strongly related to the residue similarity of the repeated regions. Therefore, for each of the 12 reference families, a multiple alignment has been constructed by fragmenting the sequences in order to align all the repeated

regions. The repeats were then classified into a number of subtypes according to their residue similarity. The number of repeats and the presence of additional domains also affect the ability of a given program to construct an accurate alignment. To address these questions, subsets of each reference alignment are proposed, selected according to repeat subtype and presence of additional domains (Table 2). These 83 subsets present a number of aligned sequences verifying specific criteria:

- C1(a) The same number of repeats of a unique subtype
- C1(b) A variable number of repeats of a unique subtype
- C2(a) The same number of repeats with different subtypes in the same order
- C2(b) The same number of repeats with different subtypes, but in a different order
- C2(c) A variable number of different repeat subtypes
- C3 The presence of an additional non-repeated conserved domain
- C4 The presence of several different repeat types

By combining these alignment subsets, a flexible benchmark test system can be designed to evaluate and compare the various algorithms currently available for the detection and alignment of repeats in protein sequences.

Table 2. Reference 6, subsets of sequences containing repeats showing the number of alignments in each subset

	C1		C2			C3	C4	Total
	A	B	A	B	C			
SH3	1	2	1	2	3	1	1	11
ZF	1	1	0	2	2	0	1	7
APO	0	0	3	3	2	0	0	8
SUSHI	1	1	1	6	3	2	1	15
MYB	0	0	2	1	2	0	0	5
ANK	0	0	0	4	2	2	1	7
KRINGLE	0	0	2	2	2	2	0	8
DEAD	1	1	0	0	1	1	0	4
TRK	1	1	0	1	1	0	0	4
FAA	0	0	0	0	3	0	0	3
LRR	0	0	3	2	2	1	1	9
Total	5	6	12	23	23	9	5	83

Reference 7

In the case of transmembrane proteins, the problems are similar to those defined previously for repeats, i.e. detection, local and global alignment. Reference 7, presented in Table 1B, consists of eight families of transmembrane proteins containing approximately 400 aligned protein sequences. A global multiple alignment of each family is available, in which the known/predicted transmembrane helices are identified. Unfortunately, due to the variability observed in the automatic prediction of transmembrane helices and the lack of known three-dimensional structures, the exact locations of the transmembrane helices cannot be reliably determined. We have therefore defined two types of 'blocks' in reference 7. The core blocks contain those regions in which all the sequences can be reliably aligned, as before. In addition, we define 'fuzzy' blocks in which predicted transmembrane helices overlap, but where the optimal multiple alignment may vary by several residues. In addition, the sequences in each family alignment are divided into a number of sub-groups using the Secator program (N.Wicker, unpublished results). Within each sub-group, the sequences are highly related and may constitute tests for the accuracy and reliability of the prediction for transmembrane prediction programs.

Reference 8

Reference 8 concerns two different, but related, problems. Inversions in proteins can result from various phenomena, such as the insertion of a complete domain at different sites in a protein, or the transfer of part of the C-terminal of the protein to its N-terminal, thus causing a discontinuity of the terminal domain. Examples of both of these cases are presented in Table 1C. Reference 8 consists of five protein families in which the sequential ordering of the domains is not preserved, corresponding to 118 sequences. For each family in this reference, we propose an independent alignment of each permuted domain.

CONCLUSIONS AND PERSPECTIVES

With the addition of the three new reference sets, BAliBASE (2.0) now provides specific test cases representing most of the problems encountered in sequence detection and alignment. A comparison of the programs currently available (work in progress) for the detection and/or alignment of repeats, transmembrane or inverted regions will enable us to identify the strong and weak points of each. The results of the study should allow us to further develop the multiple alignment program DbClustal (18). The modular design of this program facilitates the incorporation of new modules, allowing the integration of information from different sources into the multiple alignment process. The resulting program should enhance the automatic and reliable multiple sequence alignment of these highly modular proteins.

ACKNOWLEDGEMENTS

We would like to thank D. Moras for his continuous support during this work, and F. Plewniak and R. Ripp for helpful discussions. The work was supported by institute funds from Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg and the Fond de Recherche Hoechst Marion Roussel.

REFERENCES

1. Thompson,J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
2. Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
3. Andrade,M.A., Ponting,C.P., Gibson,T.J. and Bork,P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.*, **298**, 521–537.
4. Bornberg-Bauer,E., Rivals,E. and Vingron,M. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res.*, **26**, 2740–2746.

5. Heringa,J. and Argos,P. A method to recognize distant repeats in protein sequences. (1993) *Proteins*, **17**, 391–441.
6. Pellegrini,M., Marcotte,E.M. and Yeates,T.O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins*, **35**, 440–446.
7. Cserzo,M., Wallin,E., Simon,I., von Heijne,G. and Elofsson,A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.*, **10**, 673–676.
8. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
9. Kihara,D., Shimizu,T. and Kanehisa,M. (1998) Prediction of membrane proteins based on classification of transmembrane segments. *Protein Eng.*, **11**, 961–970.
10. Gromiha,M.M. (1999) A simple method for predicting transmembrane alpha helices with better accuracy. *Protein Eng.*, **12**, 557–561.
11. Pasquier,C., Promponas,V.J., Palaios,G.A., Hamodrakas,J.S. and Hamodrakas,S.J. (1999) A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. *Protein Eng.*, **12**, 381–385.
12. Lio,P. and Vannucci,M. (2000) Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, **16**, 376–382.
13. Uliel,S., Fliess,A., Amir,A. and Unger,R. (1999) A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, **15**, 930–936.
14. Kleywegt,G.J. and Jones,T.A. (1995) Where freedom is given, liberties are taken. *Structure*, **3**, 535–540.
15. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
18. Thompson,J.D., Plewniak,F., Thierry,J.-C. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.* **15**, 2919–2926.