BIOINFORMATICS

Vol. 19 no. 4 2003, pages 506–512
DOI: 10.1093/bioinformatics/btg016

Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures

Mounir Errami, Christophe Geourjon and Gilbert Deléage*

Pôle de BioInformatique Lyonnais, Institut de Biologie et de Chimie des Protéines, Centre National de la Recherche Scientifique, UMR 5086, 69367 Lyon CEDEX 07, France

Received on March 14, 2002; revised on August 26, 2002; accepted on October 8, 2002

ABSTRACT

Motivation: Multiple sequence alignments are essential tools for establishing the homology relations between proteins. Essential amino acids for the function and/or the structure are generally conserved, thus providing key arguments to help in protein characterization. However for distant proteins, it is more difficult to establish, in a reliable way, the homology relations that may exist between them. In this article, we show that secondary structure prediction is a valuable way to validate protein families at low identity rate.

Results: We show that the analysis of the secondary structures compatibility is a reliable way to discard non-related proteins in low identity multiple alignment.

Availability: This validation is possible through our NPS@ server (http://npsa-pbil.ibcp.fr).

Contact: g.deleage@ibcp.fr

INTRODUCTION

Sequencing genome projects have generated a massive surge of data and a dramatic growth of publicly available DNA and protein sequences. The remain work consists to analyze these genomes, to locate the genes and to assign a biological function and possibly a structure to each protein resulting from their traduction. The proteins can be gathered in families and subfamilies, characterized by typical folds, sites, functions. An essential basis upon which this classification is established is the comparison of protein sequences in the form of multiple alignments, helping to establish predictions about biological functions and/or phylogenetic relations between proteins. These multiple alignments, offer through residues conservation analysis, a rapid way to characterize a protein. Homology is easy to establish when sequences are similar (sharing an identity $> 30\%$). This does not imply that nonsimilar proteins are not related. The difficulty is to validate protein families when the similarity is low. Various approaches exist, but they primarily use alignment of two protein sequences. An approach consists to exploit the homology transitivity and to use one or more relay proteins to establish the relations between proteins at low identity rate (Teichmann *et al.*, 2000). Another solution has been the improvment of similarity search algorithms to make them more sensitive, like PSI-BLAST (Altschul *et al.*, 1997). One more recent way consists in using the information brought by the predicted secondary structures to validate the structural homology that can bind two proteins, even at low identity rate (Geourjon *et al.*, 2001). Indeed, secondary structure predictions are known to succefully help in fold recognition, and various methods based on this approach exist (Jones *et al.*, 1999; Rost, 1995). Furthermore, since CASP3 (Critical Assesment of Technics for Protein Structure Prediciton round 3; see *Proteins* suppl. 3, 1999), all succefull methods in the field of fold recognition make use of secondary structures predictions, showing that secondary structure is a valuable way to establish structural relationship between proteins. Multiple alignments analysis, as for it, can be made by the analysis of positional conservation (Pei and Grishin, 2001), or by measuring the statistical sgnificance calculated for multiple alignments (Hertz and Stormo, 1999). Another way consists in the use of a scoring function like norMD (Thompson *et al.*, 2001). But we must admit that analysis and validation tools are missing, leaving the user to cope with manual analysis. In order to provide a solution to this problem, we present here a novel way to validate protein families within multiple alignments at low identity rate (10 to 30%). Our method consists in analyzing the agreement between predicted secondary structures of the aligned sequences. We show that it is then possible to validate structural families within multiple alignments at low identity rate, by discarding the nonrelated sequences.

*To whom correspondance should be adressed.

MATERIALS AND METHODS

Overview

Is SOV parameter a valuable tool to validate protein families? To answear this question, SOV is calculated using reference alignments. It is compared to SOV calculated for control alignments. For each alignment, we proceed in three steps. Firstly, control alignments are obtained from a reference alignment in which a sequence is made non related to the others by random shuffling. Secondly, secondary structures are predicted for all the aligned sequences. Thirdly SOV parameters are calculated for reference and control alignments and they are compared through the calculation of a corrected difference ΔSOV.

Reference Alignments

For this work, benchmark alignments are needed, with objective criteria to assess the quality of an alignment. Structural alignments can provide reference alignments. Indeed, this kind of alignment is more reliable when the identity level between sequences is low, since it is obtained after three-dimensional structures superimposition, ensuring an optimal alignment of amino acids sequences so that the structure, and possibly the function are preserved. Two principal sources of structural alignments were used: SSSD (Friedberg *et al.*, 2000) and BAliBASE (Bahr *et al.*, 2001).

SSSD. The SSSD database is obtained starting from DAPS database (Distant Aligned Protein Sequences, Rice and Eisenberg, 1998; http://siren.bio.indiana.edu/daps). SSSD contains 126 pairs of aligned structures sharing on average 12% of sequence identity (8 to 13%), with variable gap rates (0 to 60%). These alignments include proteins of more than 30 residues, with determined structure of, at least, 3.5 Å resolution. The similarity between the sequences for each of the 126 pairs is below the detection threshold of Smith and Waterman dynamic programming algorithm.

BALIBASE

BAliBASE (version 1.0) is a database of multiple structural alignments, containing five groups of alignments also called references. Reference 1 alignments consist of equidistant sequences of similar length. For each alignment, the identity level between any two sequences is within a specified range, resulting in three sets (Table 1). Reference 2 alignments contain families composed of closely related sequences (sharing at least 25% of identity) and orphan sequences representing distant members of the family, sharing at the most 20% identity with any other sequence within the alignment. Reference 3 alignments contain up to four families per alignment. The identity rate between two sequences from different families never exceeds 25%. References 4 and 5 contain alignments with large N-terminal extensions and large C-terminal insertions. All these references were created with the aim of covering the majority of biological cases and difficulties, which can be encountered by multiple alignments programs. Thus, these references offer benchmark alignments to assess the quality of such programs (Karplus and Hu, 2001; Thompson *et al.*, 1999).

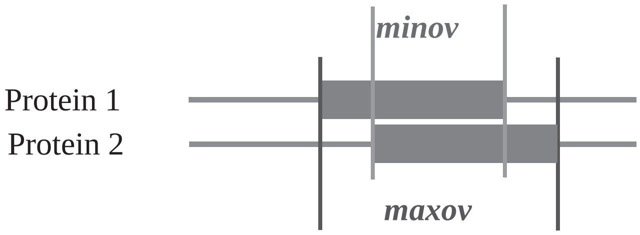

Fig. 1. Shematic representation of minov and maxov as used in equation (1).

We used these two sources of structural alignments since they are complementary. Indeed, SSSD alignments present a relatively constant identity rate with variable gap rates, allowing to know gap rate influence on the discriminate capacity of secondary structure compatibility parameter SOV (Structural OverLap; Rost *et al.*, 1994a). BALiBASE alignments, as for them, will help in studying the possible correlation between the identity level and SOV parameter within a multiple alignment.

Secondary structure compatibility

For each aligned pair, the agreement between secondary structure was estimated by calculating SOV parameter as most recently defined (Zemla *et al.*, 1999) and adapted to the comparison of two different proteins (Geourjon *et al.*, 2001):

$$\mathrm{Sov} = 100 \times \left[\frac{1}{N} \sum_{i\in[H,E,C]} \sum_{S(i)} \frac{\mathrm{minov(Sq,St)} + \delta\mathrm{(Sq,St)}}{\mathrm{maxov(Sq,St)}} \times \mathrm{len(Sq)} \right] \quad (1)$$

in which N is the alignment length minus the number of gaps; len is the sequence length; H, E and C are the Helix, Extended, and Coil states, minov is the length of actual secondary structures overlap of the query Sq and the target St; maxov is the maximal length of overlapping secondary structures Sq and St (Fig. 1) and δ is defined as

$$\delta\mathrm{(Sq,St)} = \min\{(\mathrm{maxov(Sq,St)} - \mathrm{minov(Sq,St)}); \mathrm{minov(Sq,St)}; \mathrm{len(Sq)}/2; \mathrm{len(St)}/2\}.$$

Whereas it is recognized that identity must be at least about 25% for the selection of a structural template in a

Table 1. BAliBASE version 1.0 status. Alignments number in each Reference. (from Thompson *et al.*, 1999). For each alignment, an average gap rate is calculated for each sequence. The average is obtained using all possible pairs between this sequence and the other ones of the alignment. If the average gap rate exceeds 30%, the sequence is removed from the study (SOV values obtained with all pairs implying this sequence are not considered)

	Alignment Number				
Reference 1	<100 residues	200 < 300 residues	>500 residues	Average gap rate (%)	Removed sequences
ld < 25% (set1)	7	8	8	11.27 ± 5.60	0
20 < ld < 40% (set2)	10	9	10	11.44 ± 6.79	0
ld > 35% (set3)	10	10	8	12.09 ± 6.89	0
Reference 2	9	8	7	10.69 ± 5.03	0
Reference 3	5	3	5	17.25 ± 5.72	0
	Extensions (ref. 4)	Insertions (ref. 5)			
Alignment number	12	12			
Average gap rate	22.12 ± 4.03	16.00 ± 5.92			
Removed sequences	73	16			

molecular modeling process, the use of SOV parameter within PROCSS method (PROtein Compatibility from Secondary Structure; Geourjon *et al.*, 2001) allows to decrease this threshold by 10% with the contribution of the secondary structures information. At low identity level (below 25%), sequence similarity alone fails to find distantly related proteins. The SOV brings an additional dimension, thus decomposing the information contained in a pair of aligned sequences, and gives a reliable way to assess the homology between two proteins when they share 10 to 30% identity. By applying a SOV threshold of 60%, it is possible to validate the homology between two proteins at a success rate of 95% (Geourjon *et al.*, 2001). The SOV parameter is a particularly interesting tool, insofar it makes a clear improvement of homology molecular modeling processes by increasing the number of potentially usable structural templates. For this reason, SOV parameter is used in automatic molecular modeling processes available through Internet, like Geno3D (Combet *et al.*, 2002) integrated within the NPS@ server (Network Protein Sequence Analysis; Combet *et al.*, 2000, http://npsa-pbil.ibcp.fr/).

Secondary structures prediction

SOV calculation requires secondary structures. With this goal, three predictive methods were used: SOPMA (Geourjon and Deléage, 1995), DSC (King *et al.*, 1997) and PHD (Rost *et al.*, 1994b).

The consensus prediction is calculated using these three methods: the most frequently observed state is kept. With this consensus the prediction accuracy obtained is a little bit more reliable than any given method alone as shown in Table 2.

Table 2. Accuracy levels of secondary structure predictions. Q3 is the prediction accuracy considering three secondary structure states (Helix, Extended or Sheet, Coil)

	Q3%			
Prediction method	Coil	Helix	Sheet	Average
SOPMA	75.5	75.3	62.1	72.5
DSC	78.0	64.5	56.2	68.5
PHD	74.9	74.3	64.8	72.5
SOPMA-DSC-PHD[a]	80.1	72.9	59.4	72.8

[a] Consensus pradiction method from all three methods as calculated in NPS@ (Combet *et al.*, 2000)

Control alignments and SOV parameters. Control alignments are obtained by the random shuffling of one sequence in the alignment taking care about preserving gap and identity rates between the two sequences. For each SSSD aligned pair, two control alignments groups are obtained. In the first control alignments group, the first sequence is kept unchanged, and the second is modified (Fig. 2). This procedure is performed three times resulting in a first group composed of three alignments. The second group is obtained in the same way, but here the second sequence of each alignment being kept unchanged while the first one is modified. The secondary structure of randomised sequences is then predicted. The SOV Parameter is then calculated for the six control pairs and for the actual SSSD alignment. This process is applied to all SSSD alignments.

The average control SOV ($\text{SOV}_{\text{control}}$) is calculated with control alignments and is compared to the average actual SOV ($\text{SOV}_{\text{actual}}$) resulting from SSSD alignments. This comparison is made by the calculation of the corrected

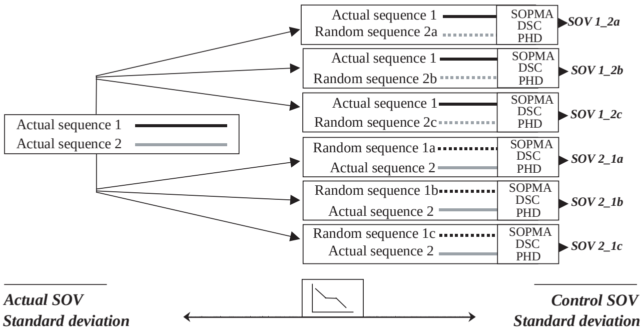

Fig. 2. Calculation of actual SOV parameter, control alignments and control SOV parameter. Random sequences (dotted lines) are obtained by random shuffling of sequences, taking care of preserving gap and identity rates between sequences. For each alignment pair, each sequence is randomly shuffled three times. The process is applied to all SSSD alignments. SOV parameters are calculated on aligned secondary structures as defined in equation (1). Secondary structures predictions are a consensus prediction calculated using three prediction methods (see Materials and Methods): SOPMA (Geourjon and Deléage, 1995), DSC (King *et al.*, 1997) and PHD (Rost *et al.*, 1994b).

difference:

$$\Delta\text{SOV} = \text{SOV}_{\text{actual}} - (\text{SOV}_{\text{control}} + \sigma_{\text{SOVactual}} + \sigma_{\text{SOVcontrol}}) \quad (2)$$

with $\text{SOV}_{\text{actual}}$: the average SOV on SSSD alignments; $\sigma_{\text{SOVactual}}$: standard deviation on $\text{SOV}_{\text{actual}}$; $\text{SOV}_{\text{control}}$: the average SOV on control alignments; $\sigma_{\text{SOVcontrol}}$: standard deviation on $\text{SOV}_{\text{control}}$. This process is applied to all BAliBASE multiple alignments. So, for each multiple alignment, the actual SOV parameter is calculated for every pair in the alignment. It is compared to the control SOV parameter obtained with all the possible control pairs (in which the amino acids positions of one sequence had been randomly changed). Thus for a multiple alignment of n sequences, a total of $n(n-1)$ possible control pairs are obtained. These pairs are recomputed three times, leading to a total of $3n(n-1)$ control pairs, on which the average control SOV is calculated.

RESULTS

On the one hand, the effect of the gap rate on SOV discrimination has been studied in SSSD pairwise alignments. On the other hand, the effect of the identity rate on the capability to detect related proteins in multiple alignments has been studied by using BAliBASE as a set of reference multiple alignments.

SOV comparison

SSSD. In order to determine gap effect, the corrected difference ΔSOV is represented as a function of gap rate in aligned pairs (Fig. 3). The observed SOV is always greater than the control SOV regardless the gap rate (Fig. 3A). The plot of the corrected difference between actual and control SOV's shows a significant difference (Fig. 3B). SOV parameter is able to set appart, in a identity range from 8 to 13%, the related sequences pairs from from pairs containing a randomised sequence. This distinction can be observed up to about 30% gap rate. Beyond this threshold, ΔSOV is not sufficient any more to permit this reliable discrimination. We can note the weaker the gap rate, the greater the ΔSOV and the easier the discrimination. This comes from the fact that secondary structure cannot be predicted in gap regions. So the higher the gap rate, the lower the $\text{SOV}_{\text{actual}}$. Thus the difference ΔSOV becomes too low to be reliably used.

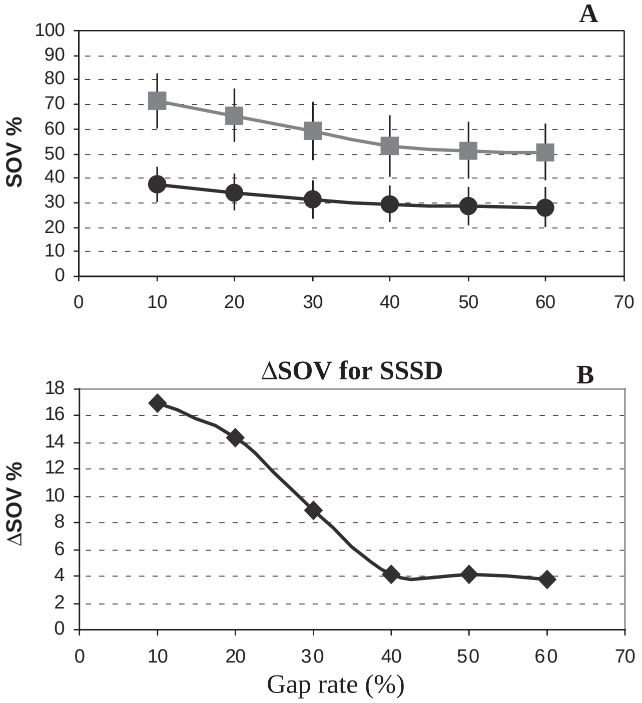

Fig. 3. Calculated SOV for SSSD. (a) $\text{SOV}_{\text{actual}}$ (squares) and $\text{SOV}_{\text{control}}$ (circles) are calculated as defined equation (1) on SSSD aligments and control alignments obtained as described Figure 2. ΔSOV is the corrected difference between $\text{SOV}_{\text{actual}}$ and $\text{SOV}_{\text{control}}$ calculated as described in equation (2).

BAliBASE. The SOV parameter variation for SSSD aligned pairs results in the definition of a reasonable 30% gap threshold. This threshold is applied to BAliBASE. For each alignment, and for each sequence, an average gap rate is calculated beetween this sequence and the other sequences of the alignment. For all alignments of references 1 to 3, the gap rates for all sequences do not exceed this threshold, thus allowing to include all of them

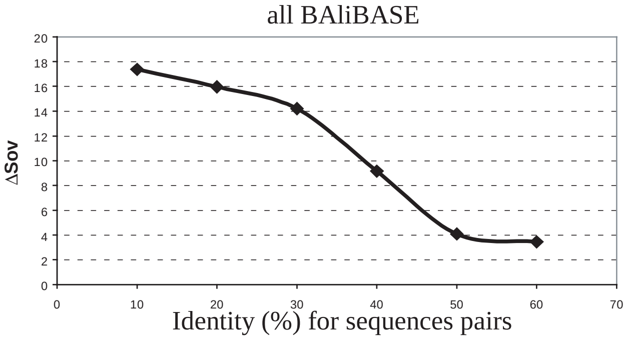

Fig. 4. ΔSOV for all the BAliBASE Références. ΔSOV is calculated as defined in equation (2).

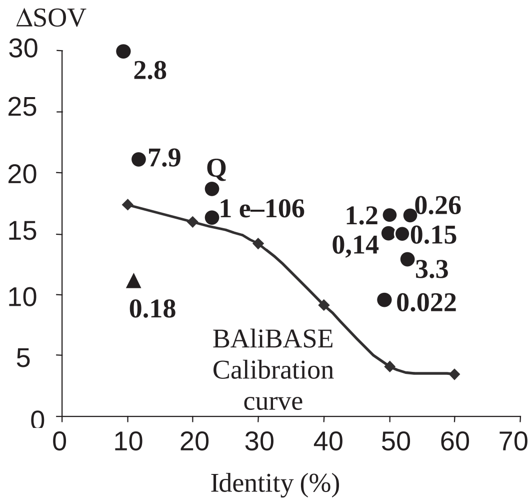

Fig. 5. ΔSOV for selected sequences in the multiple alignment, with their PSI-BLAST E-value (Q = Query). ▲ = sequence predicted as non related with ΔSOV and found with PSI-BLAST; ● = sequence predicted as related with ΔSOV.

in the study. The extensions in reference 4 alignments, lead us to remove 73 sequences from the study (reference 4 counts 108 sequences in 12 alignments). For these sequences, the average gap rate exceeds 30%. The same problem was encountered with reference 5 (which counts 100 sequences in 12 alignments). 16 sequences were not considered in SOV calculation. BAliBASE, by providing representative alignments of various biological cases, in form of multiple alignments with variable identity and gap rates (Table 1), enabled us to study SOV parameter discrimination capabilites as a function of identity rates in multiple alignments. The results show that, when the identity is above 40%, ΔSOV tends to decrease quickly and becomes weak. This shows that when identity is above 40%, it becomes difficult to distinguish between a related protein sequence and a randomised sequence in a multiple alignment (Fig. 4). Reversely, below this threshold, SOV discrimination capacity is all the stronger since identity rate is weak: ΔSOV is 17% at 10% sequence identity to decrease to 9% when the identity reaches 40% (Fig. 4). This general tendency observed for all BAliBASE alignments also appears when the references are considered in an individual way (data not schown). These results obtained with BAliBASE (Fig. 4) can serve as a ΔSOV calibration curve. This curve is identity dependant. With a given identity for a protein in a alignment, this protein must show a ΔSOV superior or equal to BAliBASE threshold (at that identity rate) to be predicted as related to other sequences in the alignment. An example is given in next section to illustrate the utility and reliability of SOV predictions.

Biological example: comparison with PSI-BLAST

In order to demonstrate the merit of the method, a comparison with PSI-BLAST (Altschul *et al.*, 1997) and a SCOP (Murzin *et al.*, 1995) validation have been made. Q925W1 is a serine protease inhibitor of 346 residues. A PSI-BLAST (version 2.2.1) search is performed in TrEMBL (release 70). All the sequences found in the last run (#4) with E-Value above 0.01 are selected, if they share at least 150 residues with Q925W1 between position 30 and 200 (Table 3). The sequences are aligned, secondary structures are predicted, SOV and ΔSOV are calculated (Table 3). The results show that, at high E-values (0.01 to 10), no possible discrimination can be made between related and non related proteins, using PSI-BLAST E-values or identity rates. In this case, this is particularly true for protein Q9UZM4 found with an E-value of 0.18, whereas SOV predictions identify it clearly as a nonrelated one. Indeed, by applying a SOV threshold of 60%, it is possible to validate the homology which can exists between two proteins (Geourjon *et al.*, 2001). As Q9UZM4 is found with a 40% SOV, this protein is predicted as non related to Q925W1 family. To bring an additionnal argument, we can use ΔSOV value. For Q9UZM4, ΔSOV is 11 and average identity between this protein and the other in the alignment is 11% (Table 3). Therefore, BAliBASE study lead to the determination of a minimal ΔSOV threshold (at 11% identity) of 17 (Fig. 4). Below this ΔSOV threshold, the protein is predicted as non related to Q925W1 family. Q9UZM4 is the only protein in this alignment showing a ΔSOV below BAliBASE threshold (Tables 3 and Figure 5). Furthermore, if we make a comparison between Q9UZM4, Q29014 and Q9AU61, we can see that all these proteins present an average identity within the alignment about 10%. Q29014 and Q9AU61 show ΔSOV

Table 3. SOV and ΔSOV for the selected proteins from PSI-BLAST (version 2.2.1) search results in TrEMBL (release 70). Id (column 5) represents the average identity for a sequence with all the other sequences within the multiple alignment. The proteins were aligned with ClustalW (version 1.8), SOV parameter was calculated on NPS@ server. BAliBASE threshold used is determined using Figure 4. All the proteins, except Q9UZM4 (bold) are predicted as related to Q925W1 family since they have a ΔSOV higher than BAliBASE threshold. Q9UZM4, Q29014, Q9AU61 (bold) present an average identity about 10% in the multiple alignment. Q29014 and A9AU61 are predicted as related. Furthermore, it is important to note that Q29014 and A9AU61 are found by PSI-BLAST with E-values higher than Q9UZM4 one. SCOP results are obtain through web site: http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/index.html, release 1.59

Name	Psi-Blast Evalue	Psi-Blast Identify	Overlap	Id in the alignment	Sov	ΔSov	Minimal ΔSov (BALLBASE)	Prediction	SCOP classification
Q925W1		(Query)		23	60	19	15.5	Related	Lipocalin
Q9DBJ9	1E−106	95	349	23	58	16	15.5	Related	Lipocalin
Q40251	0.022	16	167	49	62	9	4.5	Related	Lipocalin
Q40693	0.14	14	196	50	70	15	4	Related	Lipocalin
Q39249	0.15	15	181	51	67	15	3.9	Related	Lipocalin
Q9UZM4	0.18	11	177	11	40	11	17.1	Non-Related	P-LOOP containing nucleotide triphosphate hydrolase
AAL83562	0.26	15	194	52	71	17	3.8	Related	Lipocalin
Q9SM43	1.2	14	191	50	71	17	4	Related	Lipocalin
Q29014	2.8	13	168	9	59	30	17.5	Related	Lipocalin
AAL67858	3.3	16	197	52	70	13	3.8	Related	Lipocalin
Q9AU61	7.9	14	173	12	59	21	17	Related	Lipocalin

values above ΔSOV threshold. These two proteins are clearly predicted as related to Q925W1 family, although their E-values are higher than Q9UZM4 one. These results demonstrate that SOV parameter performs an effective discrimination between related and non related proteins at low identity rate. In order to confirm SOV predictions, a SCOP search has been performed. SCOP is a powerfull structural classification tool, which uses a structural profil database (release 1.59). SCOP predictions are in accordance with SOV ones (Table 3).

These results show secondary structure information provides a way to detect a protein with no structural homology with the other sequences of the multiple alignment, even at low identity rates. Consequently, the SOV validation of structural families within multiple alignments at low identity has a real biological significance and can be regarded as reliable. This kind of validation is possible on the sequence analysis server NPS@ (http://npsa-pbil.ibcp.fr/), in the multiple alignment tools (secondary structure predictions and sov calculation).

DISCUSSION

It has been previously shown by Geourjon *et al.* (2001), that information brought by the secondary structures is a valuable way to identify structural homologous proteins even if their sequences are relatively divergent (10 to 30% identity). This discrimination between related sequences pairs and non-related ones, at low identity level, is possible by using the SOV parameter. It is a reliable tool primarily used in structural approaches, specially in molecular modeling processes, like low identity homology molecular modeling (Geourjon *et al.*, 2001) or threading techniques (Jones *et al.*, 1999).

Our study using SSSD database and its low identity aligned sequences (8 to 13% identity) confirms the SOV relialibilty, in its capability to assess proteins homology at low sequence identity rate. In addition, it leads to the definition of a 30% gap rate threshold, below which SOV distinction is reliable. SOV parameter is a particularly interesting tool to help in the comparison of two protein sequences. Nevertheless, it has never been made profitable in multiple sequences alignments. Here we propose a novel way to take advantage of the information brought by the secondary structures compatibility within low identity multiple alignments. Indeed, we have demonstrated SOV and ΔSOV abilities to detect unrelated sequences within BAliBASE multiple alignments. The lower the identity, the easier the detection. ΔSOV is able to perform a particularly effective detection of a non-related sequence, since it proves to be most reliable when it is most difficult: at low sequence identity level, and by only considering the sequences, it is hazardous to come to a conclusion about a biological relation between two sequences. Under these conditions, the secondary structures compatibility provides all its utility, and brings a considerable way to assess the homology relationships between sequences within low identity multiple alignments.

ACKNOWLEDGEMENTS

This work was supported by CNRS and the french 'Inter EPST bioinformatics program'. M.E. is a MESR doctoral fellowship.

REFERENCES

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.

Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.

Combet,C., Blanchet,C., Geourjon,C. and Deléage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.

Combet,C., Jambon,M., Deléage,G. and Geourjon,C. (2002) Geno3D: automatic comparative molecular modelling of protein. *Bioinformatics*, **18**, 213–214.

Friedberg,I., Kaplan,T. and Margalit,H. (2000) Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci.*, **9**, 2278–2284.

Geourjon,C. and Deléage,G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.*, **11**, 681–684.

Geourjon,C., Combet,C., Blanchet,C. and Deléage,G. (2001) Identification of related proteins with weak sequence identity using secondary structure information. *Protein Sci.*, **10**, 788–797.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Iwaasa,H., Takagi,T. and Shikama,K. (1989) Protozoan myoglobin from *Paramecium caudatum*. its unusual amino acid sequence. *J. Mol. Biol.*, **208**, 355–358.

Jones,D.T., Tress,M., Bryson,K. and Hadley,C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, **37**, 104–111.

Karplus,K. and Hu,B. (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**, 713–720.

King,R.D., Saqi,M., Sayle,R. and Sternberg,M.J. (1997) DSC: public domain protein secondary structure predication. *Comput. Appl. Biosci.*, **13**, 473–474.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

Proteins **37**; suppl. 3, 1999.

Rost,B., Sander,C. and Schneider,R. (1994a) PHD: an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.*, **10**, 53–60.

Rost,B., Sander,C. and Schneider,R. (1994b) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Rost,B. (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 314–321.

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

Teichmann,S.A., Chothia,C., Church,G.M. and Park,J. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics*, **16**, 117–124.

Thompson,J.D., Higgins,D.G. and 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. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Thompson,J.D., Plewniak,F., Ripp,R., Thierry,J.C. and Poch,O. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **314**, 937–951.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.