Journal of Bioinformatics and Computational Biology
Vol. 1, No. 3 (2003) 505–520
© Imperial College Press

Imperial College Press
www.icpress.co.uk

CONSERVATION OF AMINO ACIDS INTO MULTIPLE ALIGNMENTS INVOLVED IN PAIRWISE INTERACTIONS IN THREE-DIMENSIONAL PROTEIN STRUCTURES

MOUNIR ERRAMI*, CHRISTOPHE GEOURJON† and GILBERT DELÉAGE‡

Pôle Bioinformatique Lyonnais-Institut de Biologie et Chimie des Protéines,
Laboratoire de Bioinformatique et RMN structurales,
7, passage du Vercors, 69367 Lyon cedex, France
**m.errami@ibcp.fr*
†*c.geourjon@ibcp.fr*
‡*g.deleage@ibcp.fr*

Received 13 December 2002
Revised 9 March 2003
Accepted 21 March 2003

We present an original strategy, that involves a bioinformatic software structure, in order to perform an exhaustive and objective statistical analysis of three-dimensional structures of proteins. We establish the relationship between multiple sequences alignments and various structural features of proteins. We show that amino acids implied in disulfide bonds, salt bridges and hydrophobic interactions are particularly conserved. Effects of identity, global similarity within alignments, and accessibility of interactions have been studied. Furthermore, we point out that the more variable the sequences within a multiple alignment, the more informative the multiple alignment. The results support multiple alignments usefulness for predictions of structural features.

Keywords: Conservation; three-dimensionnal interactions; multiple alignment; similarity; accessibility.

1. Introduction

There are two main ways to investigate the residue importance for a protein structure or function. The first is experimental, and consists in examining the effect of mutations. The second is computational, by comparing sequences of proteins in a family and studying the distribution of residues. Conserved positions are suspected to play important roles regarding to protein structure or function. Previous studies established some basic principles like conserved hydrophobic residues in protein core,[1–3] conserved physico-chemical properties regarding functional sites[1,4] or conserved polar residues at protein surfaces.[5] Various studies focused on interaction conservation in protein families, but they considered only one interaction type such as hydrophobic interaction, electrostatic interaction such as salt bridges[6,7] or disulfide bonds.[8] Furthermore, they were not exhaustive, since the conservation was studied on few proteins or families: Musafia *et al.* (1995) used 94 proteins,[9] Schueler and Margalit (1995) used eight protein families.[10] However, in these studies, the effect of various parameters has been studied like accessibility or secondary structure. We present an original automatic strategy, that allowed an exhaustive analysis of the conservation of disulfide bonds, salt bridges, hydrophobic interactions in automatic computed multiple alignments. The aim of this work is to investigate the relationships between the conservation of residues in multiple alignment of a protein family (obtained from sequence similarity search) and their involvement into pairwise interactions (derived from three-dimensional structures). This study was led with a permanent purpose of exhaustiveness and objectivity. Our results indicate that structural roles of residues correlate with their preferential conservation in multiple alignments, supporting the usefulness of multiple alignments in predicting structural features.

2. Materials and Methods

2.1. *Protein structures and interaction detection*

A total of 1567 protein structures of the PDB[11] have been used. The identity between any two sequences is less than 25 percent, ensuring a sequence non-redondancy protein set. An interaction database has been created using DSSP[12] (*Dictionary of Secondary Structures of Proteins*). We have modified the DSSP program (called DSSPm) so as it becomes able to detect and list, the position and the accessibility of residues implied in disulfide bonds, salt bridges and hydrophobic interactions from a protein structure (PDB file). A chemical type has been defined for each residue, and important functional atoms have been used for interaction searching (Table 1).

Disulfide is detected if two atoms "SG" are distant of less than 3 Å. A salt bridge is detected if functional atoms of two opposed charged residues are distant of less than 3 Å. An hydrophobic interaction is detected if functional atoms of two

Table 1. Residues chemical types and atoms used for interaction detection.

Residue		Functional Atoms	Chemical Type
Gly	G		GLY
Arg	R	NH1, NH2	BASE
Asp	D	OD1, OD2	ACID
Cys	C	SG	CYS
Glu	E	OE1, OE2	ACID
His	H	ND1, CD2, CE1, NE2	BASE
Ile	I	CD1	HYDROPHOBIC
Leu	L	CD1, CD2	HYDROPHOBIC
Lys	K	NZ	BASE
Met	M	CE	HYDROPHOBIC
Phe	F	CD1, CD2, CE1, CE2, CZ	HYDROPHOBIC
Trp	W	CD1, CD2, CE2, CE3, CZ2, CZ3, NE1, CH2	HYDROPHOBIC
Val	V	CG1, CG2	HYDROPHOBIC

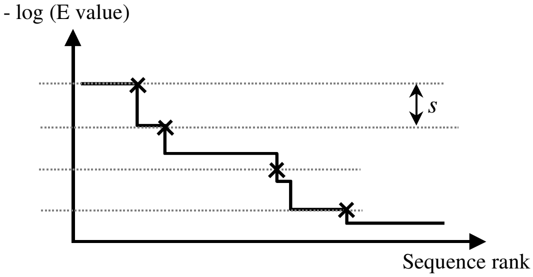

Fig. 1. Schematic distribution of sequences found by BLAST. s represents the stage calculated by Eq. (1), the crosses are the position of selected sequences in order to obtain a mixed sample.

distinct hydrophobic residues are distant of less than 3.3 Å.

2.2. *Similarity search and multiple alignments*

For each protein, a similarity search is performed using BLASTP[13] in SWISSPROT + TrEMBL.[14] In order to get a set of the maximum number of representative and related sequences, a procedure for sampling sequences in the similarity result file has been developed. This sampling function has to be efficient on most protein families. The customized selection is achieved using different criteria:

- If several sequences are found with a zero E-value, only the first one is kept so as to minimize uninformative redundancy;
- The selection procedure proceeds in stages. Each BLAST result file is decomposed in different stages, according to the E-value range covered by sequences and the number of found sequences. For each stage, one sequence is selected. The stages (s) are defined using E-values (Eq. (1)).

$$s = \frac{\log(\mathrm{E}_{(\mathrm{n})}) - \log(\mathrm{E}_{(\mathrm{l})})}{\mathrm{n} - 1} \tag{1}$$

With n: the number of sequences satisfying $0 < \text{E-value} < 1\mathrm{e} - 6$; $\mathrm{E}_{(1)}$: the first sequence with a non-zero E-value.

During the stage selection, the first non-zero E-value sequence is selected. The second sequence is selected when it has an E-value so that Eq. (2) is verified. The E-value of the second sequence is then used for the selection of the third sequence. This procedure is iteratively used until the E value exceeds 1E-6.

$$\log(\mathrm{E}_{\mathrm{i+1}}) - \log(\mathrm{E}_{\mathrm{i}}) \geq s \tag{2}$$

The aim of this procedure is to avoid the overrepresentation of identical or very close sequences (Fig. 1): this procedure allows to obtain a sample of related sequences, in which structural features like weak interactions are conserved. In the same time, this selection procedure introduces a minimal variability between

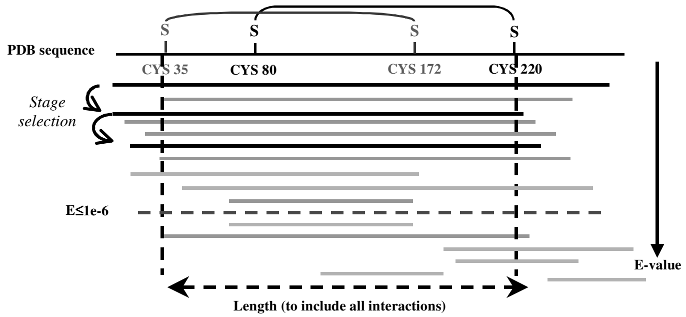

Fig. 2. Example of selection procedure of sequences found by BLAST. The selected sequences are represented by straight black lines, the discarded one are grey. In this example, the length is defined to include all dissulfide bonds.

sequences (in the form of mutations) providing key material to observe and demonstrate that these mutations tend to preserve weak interactions.

• The selected sequences (Fig. 2) are truncated on the basis of the two most distant residues involved in a interaction, so as to comprise all the interactions found in the PDB structure. These precautions ensure the calculation of a multiple alignment, containing the interactions, and a weak number of gaps.

All the sequences that verify these criteria are selected and constitute a database of related sequences. The sequences are then aligned using CLUSTALW[15] (version 1.8) with default parameters. The process is applied to each PDB protein.

2.3. *Interaction conservation and control pairs definition*

Firstly, it is necessary to identify the interactions (listed using DSSPm) within the multiple alignments. The positions numbers have to be corrected for the presence of gaps. Secondly, the conservation of an interaction (**f**) is calculated as:

$$\mathbf{f}_{\mathrm{A(i)B(j)}} = \mathrm{n}_{\mathrm{A(i)B(j)}}/\mathrm{N} \tag{3}$$

with $\mathrm{n}_{\mathrm{a(i)b(j)}}$, the number of sequences in which a residue of type A is present at the position i and a residue of type B is present at the position j and N, the number of sequences in the multiple alignment. In Eq. (3), the type of the residue is considered: Arg10-Asp40 is equivalent to Lys10-Glu40 as BASE10-ACID40. Furthermore, for salt bridges, a permutation is equivalent to a conservation: Asp10-Arg40 is equivalent to Arg10-Asp40. So as to have an idea of the statistical meaning of the calculated values, controls are needed. Concerning salt bridges, the charged residues that are obvioulsy not implied in a salt bridge (d = 15 Å) have been used. Each control pair is formed pairing two opposite-charged residues. The same rules are applied to hydrophobic interactions. Concerning disulfide bonds, reduced cysteins are randomly paired.

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

	Q3 Percentage			
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*	80.1	72.9	59.4	72.8

*Consensus prediction method from all three methods as as calculated in NPS@[20]

2.4. *Alignment quality assessment: secondary structure compatibility*

In order to check the quality of the automatically computed multiple alignments, we have used secondary structures information. Indeed, it has been established that SOV is a valuable way to assess the homology relationships between sequences in multiple alignments.[16] The secondary structure of each protein has been predicted using three different methods: SOPMA,[17] DSC[18] and PHD.[19] Then a consensus prediction has been calculated.

In the consensus prediction the most frequently predicted state is kept. Then the prediction accuracy obtained is a little bit more reliable than any given method alone as shown in Table 2.

The agreement between the aligned secondary structures has been measured with SOV parameter as most recently defined[21] and adapted to the comparison of two different proteins[22]:

$$\mathrm{SOV} = 100 \times \left[\frac{1}{N} \sum_{i \in [H,E,C]} \sum_{S(i)} \frac{\mathrm{minov}(Sq, St) + \delta(Sq, St)}{\mathrm{maxov}(Sq, St)} \times \mathrm{len}(Sq) \right] \quad (4)$$

in which N is the alignment length minus the number of gaps; len is the sequence length; H, E et C are 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 and δ is defined as:

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

2.5. *Identity and global conservation, interaction accessibility*

The identity level is calculated counting the number of strictly conserved positions divided by length of the multiple alignment.

Global conservation is calculated using AL2CO.[23] AL2CO calculates a conservation index for each position as a function of the frequency of each residue.

In order to favour the structural similarity between residues, the matrix HSDM[24] (*Homologous Structural Derived Matrix*) was used in the command line:

$$\text{AL2CO} \ -\text{i inFile} \ -\text{o outFile} \ -\text{c } 2 \ -\text{s hsdm}\,.$$

We checked the ability of AL2CO to give relevant conservation indexes onto BAliBASE[25] alignments, that have been manually refined. The indexes values were compared to the ones obtained with the BAliBASE alignments recomputed using CLUSTALW. AL2CO gives better conservation indexes for BAliBASE alignments (data not shown), showing its ability to give relevant global conservation indexes.

Accessibility is calculated for each residue using DSSP[12] (geodesic sphere integration algorithm).

DSSPm, Extractblast and Extractfasta programs, which have been developed for this study, are available at http://pbil.ibcp.fr/. A graphical interface in Tcl/Tk named BioRead has been written to integrate the functions of Exctractblast, and Extractfasta (equivalent of Extractblast for FASTA[26] and SSEARCH[27] programs).

3. Results

3.1. *Strategy*

In order to determine how important residues for structures are conserved in multiple sequence alignments, an automated strategy had to be developed. For each protein of the Protein Data Bank, we proceed in three major steps: first, the sequence is used to automatically find, select and align related sequences in order to compute a representative multiple alignment. Second, the structural interactions within this protein structure are listed in a automatic generated database. Finally, the conservation of these interactions are calculated in the multiple alignment. This process is repeated for all the proteins of the PDB, allowing an exhaustive analysis, since it is fully automatic (Fig. 3).

In order to validate this strategy, we applied it on the disulfide bonds conservation. Indeed, since disulfide bonds play a major role in protein structure establishment, evolution processes had to maintain these bonds in order to guarantee the structure integrity. Multiple alignments reflect some of the mutational changes that had occured during evolution. To be biologically relevant, our strategy must point out an important conservation of disulfide bonds comparatively to controls (pairs of reduced cysteins). The conservation of disulfide bonds in mutiple alignment has been calculated for actual SS bonds and for control. The results are given for alignments containing less or more than ten sequences (Fig. 4). In all cases, disulfide bonds conservation is higher for actual SS bonds than for control.

Disulfide bonds conservation is 94.5 percent, clearly more than controls conservation of 57.6 percent. The conservation varies with the size of the alignments. The difference between disulfide bonds and controls conservations increases with the size of alignments since it is of 21.26 in alignment of less than ten sequences and reaches 55.55 in alignments countaining more than ten sequences.

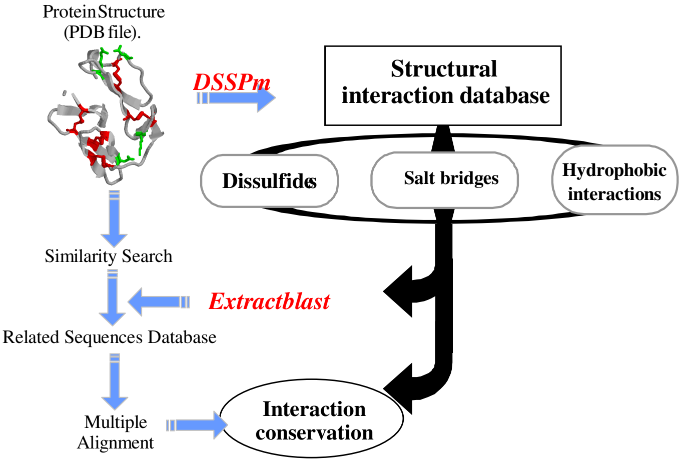

Fig. 3. Analysis strategy.

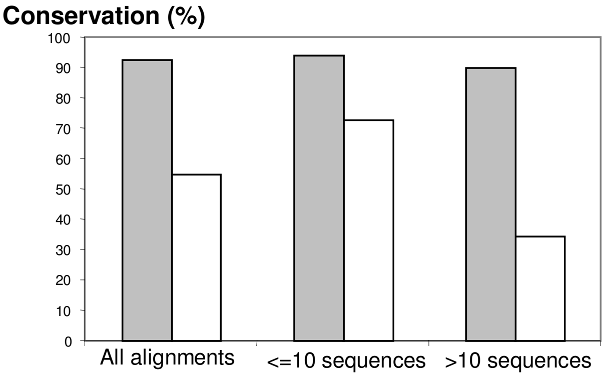

Fig. 4. Disulfide bonds conservation (grey) and reduced cystein pairs conservation (white).

We also found that the simultaneous presence of oxidized and reduced cyteins is rarely observed. Only 34 proteins of the 597 counting at least one disulfide bonds, also count at least two reduced cysteins that is to say 5.7 percent.

As our strategy points out the clear conservation of disulfide bonds and confirms our thoughts. This strategy can be regarded as relevant and the developed bioinformatic structure as functional. Then salt bridges and hydrophobic interactions conservation have been studied using the same process.

3.2. *Alignment quality*

In order to assess the quality of the automatic computed alignments used in this study, the secondary structure compatibility between aligned sequences has been checked using SOV parameter. It has been previously demonstrated that a minimal SOV of 60 percent was sufficient to establish that two aligned proteins are related.[22] Furthermore it has also been shown that SOV parameter could be made profitable to assess the relevance of a multiple alignment, even at low identity rate (10 to 30 percent).[16] The average SOV calculated for all the alignments is 89 percent ± 9 percent suggesting that the automatic computed alignments are biologically relevant. Furthermore, one aim of the selection procedure was to minimize the gap rate, since sequences were selected with a length that depends on the position of the interactions. This goal has been reached since the average gap rate within the alignments is of 5 percent ± 4 percent.

3.3. *Conservation of interactions*

3.3.1. *Global analysis*

As cysteins, charged and hydrophobic residues play an important role in three-dimensional structures. Thus, they should be more conserved than control pairs. However, these residues play various roles in proteins: solubilisation physiological solutions, allosteric regulation, interaction with other molecules, enzymatic catalysis etc. Consequently, conservation of these residues cannot be attributed to their only structural role. The aim of this study is to determine how the structural role influences their conservation within multiple alignments.

Results show a clear preferential conservation of amino acids when they are in an interaction (electrostatic or hydrophobic). Amino acids within salt bridges (Table 3) and hydrophobic interactions (Table 4) are more conserved than controls. Furthermore, this preferential conservation is more marked in alignments containing more than ten sequences: the differences between interactions and controls conservation rates are greater for these alignments.

With regard to salt bridges, permutations are more observed when residues are implied in an interaction, since the permutation rate is 12.41 for salt bridges versus 8.07 for controls. The difference is higher when alignments count more than ten sequences (8.72). This difference, while real, is not sufficient to be reliably applied in a predictive purpose.

The results are comparable to the ones obtained with disulfide bonds: the tendencies are the same. However, the difference between interactions and controls

Table 3. Salt bridges and control pairs conservation and permutation rates.

Percentage	All Alignments (570)		≤ Sequences (346)		> 10 Sequences (224)	
(Number)	Cons.	Perm.	Cons.	Perm.	Cons.	Perm.
Salt bridges	63.86	12.41	71.20	4.62	55.05	26.16
	(3075)		(1992)		(1083)	
Controls	53.87	8.07	62.00	3.03	42.76	17.44
	(14,382,706)		(9,393,334)		(4,989,372)	

Table 4. Hydrophobic interactions and control pairs conservation rates.

Conservation Percentage (Number)	All Alignments (762)	≤ 10 Sequences (353)	> 10 Sequences (409)
Hydrophobic interactions	76.37 (2248)	80.13 (1209)	75.12 (1039)
Controls	66.66 (17,994,378)	73.68 (10,829,637)	64.32 (7,164,741)

conservations is lower, pointing out the difficulty of predicting salt bridges from the only sequences.

3.3.2. *Multiple alignment identity effect*

With the aim to measure the identity effect on interaction conservation, three alignments groups were made. The first, composed of alignments showing less than five percent identity, the second of alignments showing between 5 and 50 percent identity, and the third in which alignments present more than 50 percent identity (Fig. 5). First, the greater the identity rate within the alignment, the more conserved the interactions and the controls. Second, the difference between controls and interactions conservation rates is all the more high since identity rate is weak: at five percent identity this difference reaches 12.61 for salt bridges and 8.4 for hydrophobic interactions, when at > 50 percent identity this difference downs to 8.0 for salt bridges and 1.3 for hydrophobic interactions. The preferential conservation of interactions is then better when the sequences of alignments are more variable.

3.3.3. *Global similarity effect*

The identity is not a precise measure of the global conservation in a multiple alignment, since identity complies with the "all or nothing" law. No difference is made

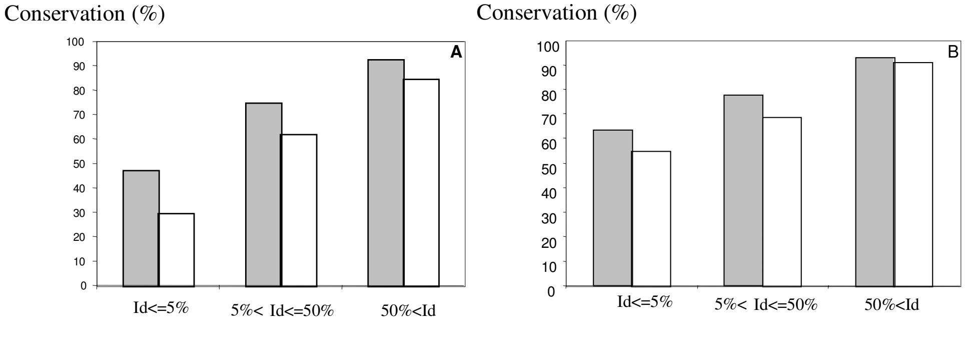

Fig. 5. Identity effect on salt bridges (A), hydrophobic interactions (B) and respective controls (white) conservation rates.

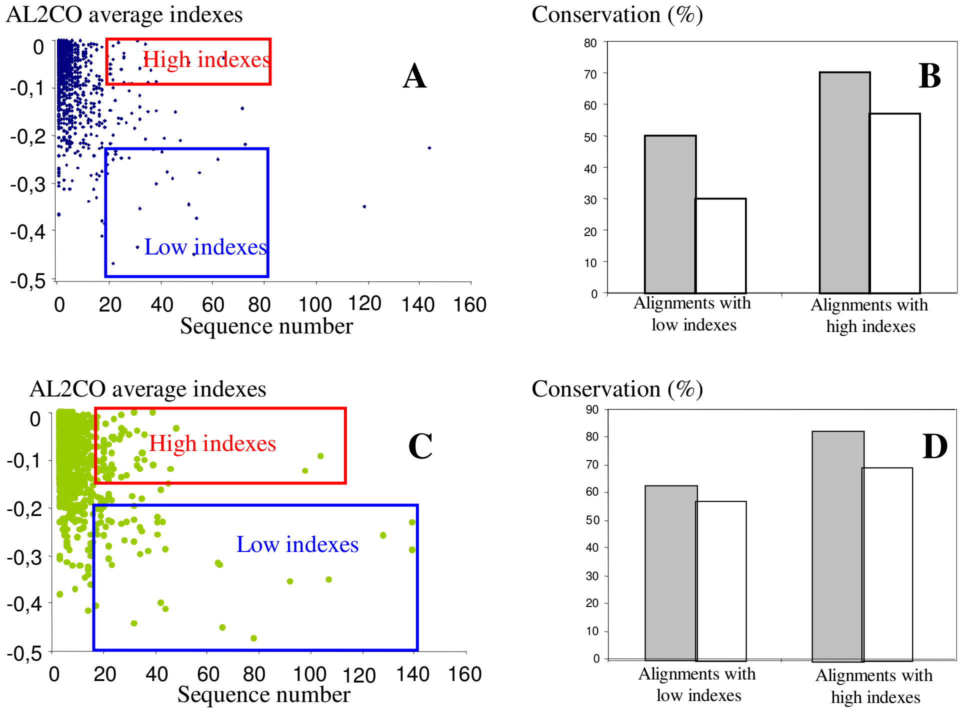

Fig. 6. (A) AL2CO average indexes for the alignments used in the analysis of salt bridges conservation. (B) AL2CO average indexes for the alignments used in the analysis of hydrophobic interaction conservation. One plot is obtained per alignment. The indexes are represented as a function of the number of the sequences. (C) Global similarity effect on salt bridges and controls (white) conservation rates. (D) Global similarity effect on hydrophobic interactions and controls (white) conservation rates.

between a position conserved at 90 percent or one conserved at 10 percent. AL2CO program provided us a solution to this problem. AL2CO was used to calculate a global conservation index for each multiple alignment. Two groups of alignments have been made (Figs. 6(A) and 6(C)): the first in which calculated indexes are high showing a high global conservation, the second with alignments presenting low indexes and then a weak global conservation. The groups were constituted so that the number of alignments and the size of these alignments are comparable.

The results (Figs. 6(B) and 6(D)) confirm the observation made with the identity. Indeed, preferential conservation of salt bridges and hydrophobic interactions is more marked when the global similarity is weak.

3.3.4. *Accessibility effect*

Charged residues present at molecule surfaces help in protein solubilization. Substitution studies did not reveal any significant influence of exposed salt bridges on

Table 5. Salt bridges (SB), Hydrophobic interactions (Hyd. int.) and control pairs (controls) conservation rates. Accessibility threshold is applied to each residue of interactions and control pairs. Numbers in parenthesis are the strength used for the study.

SB	Accessibility ≤ 10 Å^2		Global Analysis	
Conservation (%) (number)	Salt Bridges	Controls	Salt Bridges	Controls
Alignments (723)	79.40 (277)	59.04 (233)	64.37 (3075)	54.28 (8198)
≤ 10 sequences (449)	82.50 (184)	71.24 (144)	71.20 (1992)	62.00 (5519)
> 10 sequences (274)	72.59 (93)	41.57 (89)	55.05 (1083)	42.76 (2679)
Hyd int	Accessibility ≤ 10 Å^2		Accessibility > 30 Å^2	
Conservation (%) (number)	Hydrophobic Interactions	Controls	Hydrophobic Interactions	Controls
Alignments (1130)	78.85 (1434)	75.27 (1612)	64.60 (135)	54.28 (1240)
≤ 10 sequences (632)	82.58 (770)	80.66 (1097)	71.05 (73)	65.00 (509)
> 10 sequences (498)	77.60 (664)	73.47 (856)	62.46 (62)	50.95 (731)

structure stability. But, buried salt bridges are favourable over isolated charges, the latter tend to destabilize protein structures. Concerning hydrophobic residues, they tend to decrease their solvent exposure and are generally located within the protein core. Accessibility is then an important parameter to consider when dealing with salt bridges and hydrophobic interactions conservation. With the aim to study the conservation of more buried interactions, an arbitrary maximal threshold of 10 Å^2 has been fixed for both residues within a buried interaction.

Conservation results (Table 5) clearly show that the preferential conservation of salt bridges is more important for buried ones, and more particularly for alignments containing more than ten sequences, since the difference between true positive and control reaches 31.02, versus 12.29 for the global analysis (independently of interaction accessibility). The accessibility effect on hydrophobic interactions is opposed: the difference between interactions and controls conservation rates is all the greater since residues are exposed. Thus, the preferential conservation of hydrophobic interactions is more important when they are exposed. As exposure of hydrophobic residues is not energetically favourable, exposed interactions must be of biological importance so that they are maintained at protein surfaces, explaining their conservation in multiple alignments.

4. Discussion

In this paper, we have described an original automatic strategy, that allowed an exhaustive analysis of the conservation of disulfide bonds, salt bridges, hydrophobic interactions in automatic computed multiple alignments. A crucial point is to check the relevance of these alignments. The calculated indexes with AL2CO on these alignments, are comparable with values obtained on BAliBASE[25] (Bahr *et al.*, 2001) structural alignments (data not shown) suggesting that the alignments of this study are correct. Furthermore, the quality of the alignments has been checked by measuring the agreement between aligned secondary structures (SOV) and by measuring the gap rate. Our results clearly show that our alignments can be regarded as relevant. This is not suprising, since the maximal E-value threshold of selected sequences (found by BLAST) has been fixed as classically to 1e-6. This rather low value has been chosen so as to avoid the selection of foreign sequences.

The strategy was validated onto the disulfide bonds pairing since it clearly points out the preferential conservation of cysteins within multiple alignments oxidized. Thus, multiple alignments are valuable tools to help in prediction of disulfide bonds within proteins.[28] The preferential conservation of oxidized cysteins is more marked in alignments counting more than ten sequences. Indeed, the more heterogeneous a family, the more important are the conserved residues within the family. In order to provide a representative sampling of the sequences of a given family, an extraction procedure by stages has been developed and implemented into a program called Extractblast. This program is a valuable tool to select sequences from BLAST files, in order to constitute a non redondant protein set, representative of a proteic family, allowing sufficient mutations to emphasize these important residues. Consequently, cysteins that are not essential to the structure may have disappeared but oxidized ones, which are particularly important are kept. These observations made with disulfide bonds also apply to salt bridges and hydrophobic interactions. Indeed, residues are more conserved when they are implied within an interaction, suggesting the importance of their structural role. Furthermore, this conservation is all the more important since the sequences within multiple alignments are variable. Nevertheless, this preferential conservation of residues involved in interactions, is less marked for salt bridges and hydrophobic interactions than for disulfide bonds. Several reasons can explain this. Firstly, salt bridges and hydrophobic interactions are weak interactions, it is easy for the protein to make up for the disappearing of weak interactions, by other weak interactions.[9] Russell and Barton (1994) has previously shown that weak interactions can be quite different (number, position and type) within proteins sharing the same structure.[29] Thus the conservation of a fold would not be due to the conservation of weak interactions but rather to the conservation of global physico-chemical properties. Secondly, the residues involved in weak interactions, like charged or hydrophobic residues are multivalent: enzymatic catalysis, interactions with other molecules, interactions with membranes, regulation targets. Consequently, it is not possible to ascribe the conservation of these residues to their only structural role.

An important parameter of the study is the accessibility of the interactions. Concerning electrostatic interactions, the more buried interactions show the greater preferential conservation (compared to controls). Due to the missing of water molecules in the interior of proteins, interactions are favourable over isolated charges. Accessibility is the more conclusive parameter, and has to be made profitable in any predictive purpose: the observed differences between interactions and controls conservation rates are the greatest when accessibility is considered. Accessibility influence on electrostatic interactions conservation has also been studied by others,[10,30] and the results confirm that the more conserved salt bridges are also the more buried. Furthermore, it is known that the consequences for the protein stability are minor when eliminating exposed salt bridges.[31,32] Concerning hydrophobic interactions, our results are consistent with previous studies[33,34] since accessibility has a visible effect on residue conservation. Unlike salt bridges, the difference between hydrophobic interactions and controls conservation rates is more important for the more exposed interactions. Nevertheless, buried interactions are very well conserved as previously suggested.[3]

An interesting point arose from this work: the more variables the sequences within an alignment, the more informative the alignment, as the observed difference between interactions and control conservation rates is all greater since global similarity is weak. Thus, it is interesting to compute alignments using related sequences, and by introducing the maximum variability. However, it is necessary to have a way at one's disposal to validate such alignments. To this goal, we have developed a strategy that helps in the detection of non related protein sequences, within low identity multiple alignments by using secondary structure predictions.[16]

Acknowledgments

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

References

1. D. Rennell, S. E. Bouvier, L. W. Hardy and A. R. Poteete, "Systematic mutation of bacteriophage T4 lysozyme," *J. Mol. Biol.* **222**, 67–88 (1991).
2. P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret and J. H. Miller, "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence," *J. Mol. Biol.* **240**, 421–433 (1994).
3. A. Poupon and J. P. Mornon, "Populations of hydrophobic amino acids within protein globular domains: identification of conserved 'topohydrophobic' positions," *Proteins* **33**, 329–342 (1998).
4. A. M. Lesk and C. Chothia, "How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins," *J. Mol. Biol.* **136**, 225–270 (1980).
5. Z. Hu, B. Ma, H. Wolfson and R. Nussinov, "Conservation of polar residues as hot spots at protein interfaces," *Proteins* **39**, 331–342 (2000).

6. S. Kumar and R. Nussinov, "Fluctuations in ion pairs and their stabilities in proteins," *Proteins* **43**, 433–454 (2001).
7. S. Kumar and R. Nussinov, "Close-range electrostatic interactions in proteins," *Chembiochem.* **3**, 604–617 (2002).
8. P. Fariselli and R. Casadio, "Prediction of disulfide connectivity in proteins," *Bioinformatics* **17**, 957–964 (2001).
9. B. Musafia, V. Buchner and D. Arad, "Complex salt bridges in proteins: statistical analysis of structure and function," *J. Mol. Biol.* **254**, 761–770 (1995).
10. O. Schueler and H. Margalit, "Conservation of salt bridges in protein families," *J. Mol. Biol.* **248**, 125–135 (1995).
11. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.* **28**, 235–242 (2000).
12. W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers* **22**, pp. 2577 (1983).
13. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.* **25**, 3389–3402 (1997).
14. A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Res.* **28**, 45–48 (2000).
15. J. D. Thompson, D. G. Higgins and T. J. Gibson, "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 (1994).
16. M. Errami, C. Geourjon and G. Deléage, "Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures," *Bioinformatics* **19**, 506–512 (2003).
17. C. Geourjon and G. Deléage, "SOPMA: significant improvements in protein secondary structure prediction from multiple alignments," *Comput. Appl. Biosci.* **11**, 681–684 (1995).
18. R. D. King, M. Saqi, R. Sayle and M. J. Sternberg, "DSC: public domain protein secondary structure predication," *Comput. Appl. Biosci.* **13**, 473-474 (1997).
19. B. Rost, C. Sander and R. Schneider, "PHD — an automatic mail server for protein secondary structure prediction," *Comput. Appl. Biosci.* **10**, 53–60 (1994).
20. C. Combet, C. Blanchet, C. Geourjon and G. Deléage, "NPS@:network protein sequence analysis," *Trends Biochem. Sci.* **25**, 147–150 (2000).
21. A. Zemla, C. Venclovas, K. Fidelis and B. Rost, "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment," *Proteins* **34**, 220-223 (1999).
22. C. Geourjon, C. Combet, C. Blanchet and G. Deléage, "Identification of related proteins with weak sequence identity using secondary structure information," *Protein Sci.* **10**, 788–797 (2001).
23. J. Pei and N. V. Grishin, "AL2CO: calculation of positional conservation in a protein sequence alignment," *Bioinformatics* **17**, 700–712 (2001).
24. A. Prlic, F. S. Domingues and M. J. Sippl, "Structure-derived substitution matrices for alignment of distantly related sequences," *Protein Eng.* **13**, 545–550 (2000).
25. A. Bahr, J. D. Thompson, J. C. Thierry and O. Poch, "BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations," *Nucleic Acids Res.* **29**, 323–326 (2001).

26. W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).
27. T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.* **147**, 195–197 (1981).
28. A. Fiser and I. Simon, "Predicting the oxidation state of cysteines by multiple sequence alignment," *Bioinformatics* **16**, 251–256 (2000).
29. S. B. Russell and G. J. Barton, "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 (1994).
30. A. Fiser, I. Simon and G. J. Barton, "Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation," *FEBS Lett.* **397**, 225–229 (1996).
31. A. Horovitz, L. Serrano, B. Avron, M. Bycroft and A. R. Fersht, "Strength and cooperativity of contributions of surface salt bridges to protein stability," *J. Mol. Biol.* **216**, 1031–1044 (1990).
32. D. Sali, M. Bycroft and A. R. Fersht, "Surface electrostatic interactions contribute little of stability of barnase," *J. Mol. Biol.* **220**, 779–788 (1991).
33. I. Ladunga and R. F. Smith, "Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties," *Protein Eng.* **10**, 187–196 (1997).
34. C. Lawrence, I. Auger and C. Mannella, "Distribution of accessible surfaces of amino acids in globular proteins," *Proteins* **2**, 153–161 (1987).

Mounir Errami received his M.S. degree in functional and structural biology from Université Joseph Fourier, Grenoble, France, and his Ph.D. degree in Biochemistry/Bioinformatics from Université Claude Bernard, Lyon, France in July 1999 and November 2002 respectively. He is to start a postdoctoral training in Theoretical Molecular Biology and Bioinformatics Laboratory under direction of Pr. Valentin Ilyin at Northeastern University, Boston, USA.

Christophe Geourjon is permanent researcher at the French National Center for Research (CNRS). He is the author of several methods to predict the secondary structure of protein from their sequences and he was proposing one of the very first email server for protein prediction in 1994. C. Geourjon is now developing bioinformatics tools related with molecular modeling, the comparison of common 3D sites in protein structures and he is the author of Geno3D (http://geno3d-pbil.ibcp.fr), which is an automatic server for molecular modeling.

Gilbert Deléage is Professor of Structural Bioinformatics at the Claude Bernard University of Lyon. He is also one of the founder of the ⟨⟨Pôle BioInformatique Lyonnais⟩⟩. G. Deléage is at the head of a bioinformatic team at the Institute of Biology and Chemistry of Protein (UMR CNRS 5086). His team is specialized in protein structure prediction, structural bioinformatics, protein molecular modeling, sequence web services and databases. He is the author of ANTHEPROT software (http://antheprot-pbil.ibcp.fr first release in 1988) which is a integrated program dedicated to protein sequence and structure analysis.