Theor Chem Acc (2001) 106:113–120
DOI 10.1007/s002140100255

Theoretical
Chemistry Accounts

Regular article

Deciphering globular protein sequence–structure relationships: from observation to prediction

A. Poupon[1], **J.-P. Mornon**[2]

[1] Laboratoire de Structure des Protéines, DIEP, CEA/Saclay, 91191 Gif-sur-Yvette, France
e-mail: apoupon@cea.fr
[2] Equipe Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS 7590, Universités Paris 6 et 7, Case 115,
4 place Jussieu, 75252 Paris Cédex 05, France

Received: 7 October 2000 / Accepted: 5 January 2001 / Published online: 3 April 2001
© Springer-Verlag 2001

Abstract. Careful comparison of proteins sharing a same fold but only low or no sequence identity should allow a better understanding of the coding of three-dimensional structures by amino acid sequences. It has already been shown that positions of a given fold occupied mainly by hydrophobic residues in the different proteins of a structural family share very specific physical properties and participate in stabilization of the protein domain. They probably also play a crucial role in the very first steps of folding [Poupon A, Mornon J.-P (1999) *FEBS Lett.* 452: 283–289; Mirny LA, Shaknovich EI (1999) *J. Mol. Biol.* 291: 177–196]. To further understand the sequence–structure relationship, we studied the correlation between allowed mutations at a given three-dimensional position and some of its physical properties. The different amino acids were divided in three groups (hydrophobic, nonpolar or weakly polar and polar or charged), and a correlation was established between the occupation rate of each group at a given position in the fold and the burying, the side-chain dispersion, the interposition distances and the ability to form a network of directly interacting residues. The results are then applied to predict some solvent accessibility. We show that this property can be accurately predicted for about 70% of the residues, providing precious information concerning the corresponding three-dimensional structures. The results are used to predict other structural features, as secondary structures, compactness or long-range interactions between residues remote in sequence. This information will allow the number of possible structures for a given sequence to be reduced considerably, simplifying the ab initio modelling problem to a level where it might be solved by computing methods.

Key words: Hydrophobicity – Topohydrophobicity – Solvent accessibility – Ab initio prediction

Correspondence to: A. Poupon

Contribution to the Symposium Proceedings of Computational Biophysics 2000

1 Introduction

As the number of known protein sequence keeps growing, it is quite clear now that we will never be able to study experimentally each of them. Homology modelling is an alternative solution for about half of them, those for which a homologous sequence with known three-dimensional structure exists. Threading methods can also be used to find a compatible fold for some sequences. However, for many proteins, these methods cannot be applied, and an ab initio prediction method has to emerge. Recent developments in this field look promising, and new methods have been published, such as ROSETTA [1] (see Refs [2, 3, 4] for reviews). These methods allow the prediction of the overall fold for small proteins (one or two secondary structures), within 4 Å of the real structure. They also give good results for 80–100 residue proteins, but in most cases there are no results on larger proteins. Even when the overall fold can be predicted, the models built are too far from the real structure, on the atomic level, to give information on the function of the protein. To reach that goal, structural parameters for individual residues, such as secondary structures or interresidue contacts and distances, have to be better predicted, as they considerably improve the quality of the structure prediction [2].

Methods for secondary structure prediction give Q3 values around 74% (meaning that 74% of the predicted residues are assigned the correct secondary structure, which can be one of three states: β-strand, α-helix or coil) [2]. Methods for interresidue contact and distance prediction give quite disappointing results, with only 8–15% of correctly predicted contacts [5, 6].

We have already shown that structure comparisons allows significant differences for different physical properties for conserved hydrophobic residues to be demonstrated. For this purpose, proteins with known three-dimensional structures where superimposed, and the multiple alignment of the corresponding sequences was deduced from the superimposition. The study of the properties of topologically conserved hydrophobic residues (residues at positions occupied by hydrophobic residues in almost all proteins sharing a same fold, which we call " topohydrophobic " residues) shows significant differences with unconserved hydrophobic residues: they are far more buried, they are found in the inner part of the hydrophobic core, they have very low side-chain dispersion, and probably the most important, they do form a network of directly interacting residues [7, 8]. Moreover, comparison of these topohydrophobic residues with residues that have be demonstrated to be essential for the folding process has shown a very good correlation between these two properties [9]. This demonstrates that topohydrophobic residues are essential in every aspect of protein structure: folding, stability and determinism.

These previous studies brought two questions. Would similar properties be found if studying other types of conservation (conservation of the hydrophilicity for example)? What is the predicting power of such results? To answer these questions, we classified the different amino acids in three groups: hydrophobics (H, VILFMYW), hydrophilics (C, REKNDQ) and nonpolar or weakly polar (N, ACTHPGS). Cysteines were not included in the hydrophobic group because, on average, these residues are not very hydrophobic. Cystine residues are very buried, but were not included for two main reasons: they are not really hydrophobic, meaning that they do not have many hydrophobic interactions with other hydrophobic residues; and the main goal of this work is to establish structural parameters that could be used for sequence-based predictions, and the distinction between cysteine and cystine cannot be accurately made from sequence only. The type of a given position was then defined as a function of the proportions of residues from each group found at this position in the different proteins of the same family. For each type of position, we studied different properties: burying, side-chain dispersion, mean distance between two positions of the same type, direct interactions between two positions of the same type. We show that each type of position has different properties and that the properties demonstrated for topohydrophobic positions are correlated not only with conservation, but also with hydrophobicity.

To investigate the possibility to use these results, obtained from the study of known structures, we used the values found to predict solvent accessibility. Using this method, solvent accessibility can be predicted for more than 70% of the residues (depending of the chosen threshold), knowing only a multiple alignment of divergent protein sequences from the same family. We also describe the prediction of solvent accessibility ranges for each type of position, for which the accuracy is over 77%.

2 Material and methods

2.1 Fold families and multiple alignments

Clustering of proteins with known three-dimensional structures into families of proteins sharing a same fold was done primarily by sequence homology using PSI-BLAST [10], known homologies (from bibliographic references or from the CATH databank [11]) were also taken into account (for details see Refs. [7, 8]). The families obtained were then explored visually to group families with similar folds. Redundancy was eliminated by keeping, in each family, only sequences sharing less than 50% pairwise sequence identity. For property computations, only families having six or more members were used [7, 8]. Preliminary structure superimpositions were done using COMPOSER [12] and were refined manually through an iterative process [7, 8]. Multiple sequence alignment were deduced from structure superimposition.

Only positions represented for every protein of the family were considered for computations. The positions corresponding to active-site residues were also excluded.

2.2 Properties computations

1. Position types. The type of a given position is defined by the proportions of residues from the three groups (H: VILFMYW; C: REKNDQ; N: ACTHPGS) occupying this position in the different proteins of the family. The proportions are rounded to the closest tenth. For example, in a family of 24 members, a position where 14 group H, seven group N and three group C residues are found will be classified by the type $t = (0.6H; 0.3N; 0.1C)$, which will further be noted (0.6, 0.3, 0.1). The number of residues occupying each position type is given Table 1.
2. Residue solvent accessibilities for each protein were taken from DSSP files [13]. For each type, *t*, of position, the mean solvent accessibility is the sum of the solvent accessibility of all the amino acids occupying a position of type *t* divided by the number of such residues. It is to be noted that the value is not averaged in each family, but rather over the whole bank.
3. Mean distance between two positions of the same type. The distance between two positions is defined, for positions *i* and *j* (both of type *t*), as the distance between the average centres of gravity of the side chains for both positions. These distances are normalized by the mean distance between any two positions in the family. These normalized distances are then averaged over the whole bank. Here again, it is to be noted that the distance values are not averaged within a family but over the whole bank.
4. Distance between one position and its two closest neighbours of the same type. For one position *a* of type *t*, the distances to its two closest neighbours *b* and *c,* also belonging to type *t,* is the distance between the average centres of gravity of the side chains in positions *a* and *b* and *a* and *c*, respectively. The mean value between these two distances is computed, then averaged over the whole family for each position type.
5. Dispersion at a given position is computed as the mean distance between the centres of gravity of the different side chains (after superimposition) occupying this position in the different

Table 1. Number of residues for each type of position. The position type is defined by the proportion of hydrophobic (H, residues VILFMYW) and nonpolar or weakly polar (N, residues ACTHPGS) residues present in the different members of the family. Consequently, the proportion of charged residues is 1–(H + N). For the computation of the solvent accessibility, the mean distance and the network distance, individual amino acids are used; the number of positions of each type are used only for dispersion calculations

N\H	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	420	47	48	78	23	51	67	84	145	127	1553
0.1	270	95	80	167	101	93	162	89	167	886	
0.2	191	46	199	126	80	91	80	168	682		
0.3	535	304	180	287	217	250	287	433			
0.4	329	188	240	159	123	181	377				
0.5	272	105	248	186	118	317					
0.6	274	240	43	129	203						
0.7	189	171	119	468							
0.8	226	114	152								
0.9	184	357									
1	912										

proteins of the family. For each type of position, this value is averaged over the whole bank.

2.3 Solvent accessibility predictions

Multiple alignments used to test solvent accessibility predictions were taken from the PFAM bank [14, 15]. For each position of the multiple alignment, the proportions of residues from H, N and C groups were computed, then the solvent accessibility value observed for the type of position was assigned.

3 Results and discussion

3.1 Properties of the different types of positions

Different properties were computed on positions of the multiple alignments represented for every protein of the same family. As a consequence, most loops are excluded and β sheets are overrepresented. Catalytic residues were also excluded because their properties are correlated to their activity and not to structural needs. For each position type, the properties studied are

1. The solvent accessibility, for which two different parameters are computed: the mean value and the median value (Fig. 1A, B).
2. The average distance between two positions of the same type (Fig. 1C).
3. The mean distance between one position and its two closest neighbours of the same type (Fig. 1D).
4. The side-chain dispersion (Fig. 1E).

The number of residues for each type of position is also shown (Fig. 1F).

The different position types are not equally represented in the families studied (Table 1). This is not surprising, as the three groups of residues have different physicochemical properties; however, some interesting features have to be noted. There is a great gap between completely hydrophobic positions (1553 residues) and (0.9, 0.1, 0) positions (886 residues), and this is even greater for (0.9, 0, 0.1) positions. This is interesting as nonpolar or weakly polar residues are the most common ones (especially alanine). It is usually considered that nonpolar or weakly polar residues can replace hydrophobic ones without damaging the structure. The frequencies we found seem to indicate, on the contrary, that some positions, representing about 10% of the sequence, have to be maintained strictly hydrophobic. This is also true for conserved nonpolar or weakly polar positions: they concern 912 residues, whereas (0.1, 0.9, 0) and (0, 0.9, 0.1) positions are much lower (357 and 184 residues, respectively).

Some of the variations appearing in the number or residues are artifacts. For example, positions with 0.3 of any of the three groups are overrepresented, whereas positions with 0.1 are underrepresented. This is due to the fact that some of the fold families have fewer than ten members, thus having no occurrences for some position types. For example, a family with six members gives values only for 0, 0.2 (1 out of 6), 0.3, 0.5, 0.7, 0.8 and 1. The solution to this problem is not evident: one can take only families with at least ten members, but this reduces the database and thus the precision of the data obtained; another way is to keep more structures in each family, but this introduces more redundancy, biasing the results; the last solution is to reduce the number of position types (with steps of 0.2 instead of 0.1), but the consequence would be to average very different values. For example, the median solvent accessibility value for strictly hydrophobic positions (1, 0, 0) is 1 Å^2, whereas it is 13 Å^2 for (0.9, 0, 0.1) positions. Simlarily, the median solvent accessibility increases sharply as the proportion of charged residues rises.

Computation of solvent accessibility for the different position types shows that only few of them are systematically buried: (0, 1, 0), positions with 90% or more hydrophobic residues, and positions with 60%–80% hydrophobic residues, the remaining ones being nonpolar or weakly polar (Fig. 1A). However, the distributions are very asymmetric and cannot be fitted with Gaussian distributions (Fig. 2B). To take into account this dissymmetry, together with the mean accessibility, the median value was also computed (Fig. 1B). The median value is the value separating the population in two halves. For symmetric distributions, the median value is very close to the mean value, but is very different for asymmetric distributions. As illustrated in Fig. 2B, for (1, 0, 0) and (0, 1, 0) positions, the median accessibility value (2 and 1 Å^2, respectively) is more indicative than the mean value (11 and 15 Å^2, respectively) of the nature of the distribution. Computation of solvent accessibility for conserved hydrophobic positions (0, 0, 1) shows two categories: positions where the charge is conserved (RKQN or DE), which are buried (here it should be remembered that all active-site residues have been excluded from computations); in contrast, positions where the charge is not conserved are exposed (Fig. 2A).

The computation of the mean distance between two positions of the same type (Fig. 1C) clearly shows that positions which occupy the inner part of the protein are fairly well conserved hydrophobicity, and very few polar or charged residues are allowed at these positions. These positions also have low side-chain dispersion (Fig. 1E). (0, 0, 1) positions for which the charge is conserved are also found in the inner part of the protein (data not shown) and have low side-chain dispersion. These results are consistent with the fact that in the protein core, where the hydrophobicity (or the charge) is well conserved for each position, the interactions are also well conserved. This implies that the axis of the side chain must be conserved, even when the chemical nature of the residue is not. In contrast, positions that can be occupied either by hydrophilic or by hydrophobic residues have very high dispersion and a large mean distance between two positions of the same type. This shows that, at these positions, when occupied by a hydrophobic residue, the side chain points towards the centre of the protein, and when occupied by an hydrophilic residue, the side chain points towards the solvent. This also explains why, for these positions, the solvent accessibility distributions are very wide.

Computation of the mean distance between one position and its two closest neighbours of the same type

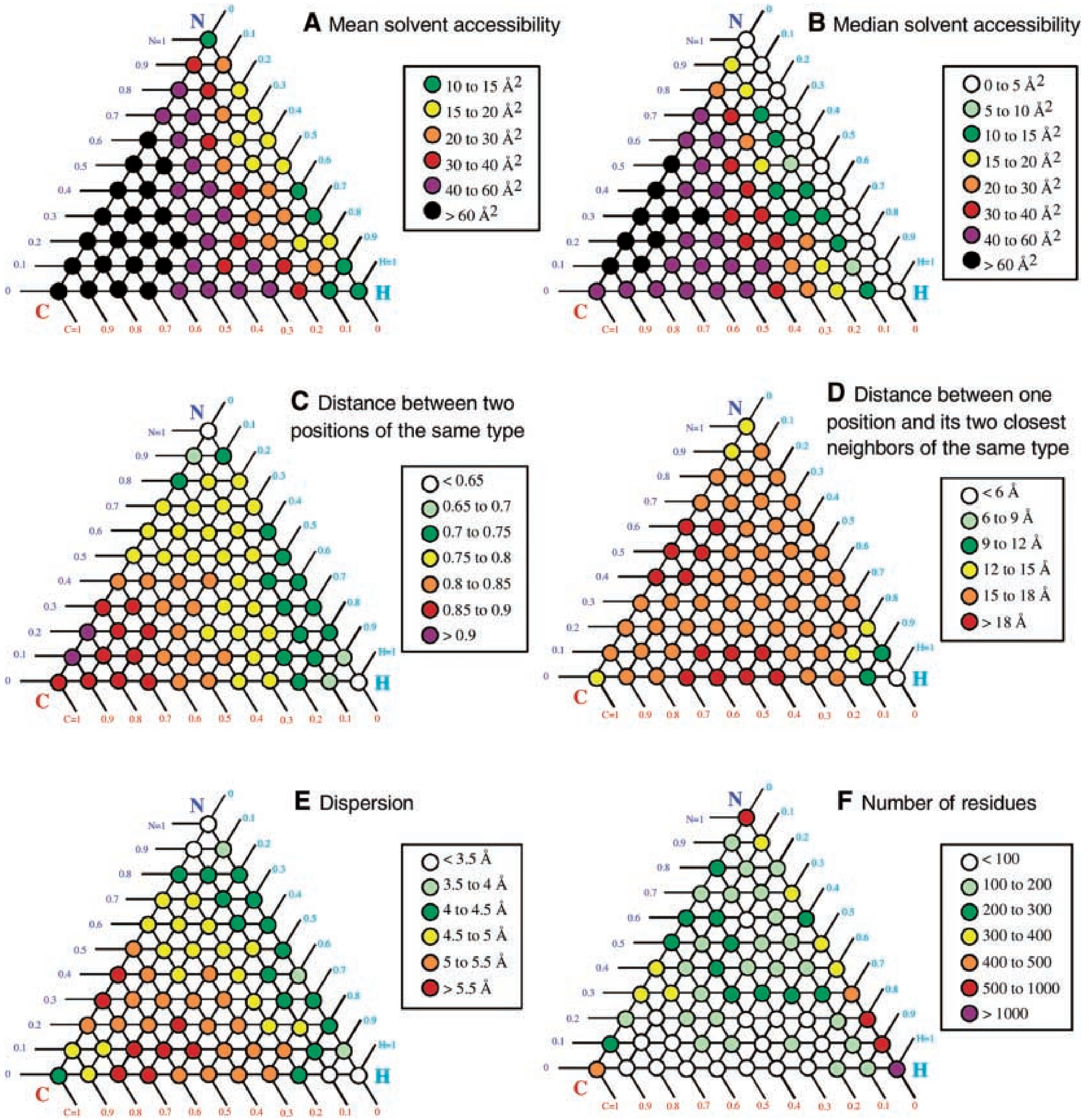

Fig. 1A–F. Properties computed for the different position types (see Sect. 2). **A:** Mean distance between two positions of the same type; **B:** mean distance between one position and its two closest neighbours of the same type; **C:** mean side-chain dispersion; **D:** number of residues occupying each type; **E:** mean solvent accessibility; **F:** median solvent accessibility

(Fig. 1D) shows that, for most position types, there is no direct interaction between positions of the same type: most positions give values from 15 to 20 Å. Only conserved positions give lower values. For conserved hydrophobic positions, this value is very low (5.8 Å), showing that each position makes direct interactions with its two closest conserved hydrophobic neighbours. We have already shown that these positions form a network of directly interacting residues in the inner core of the protein [7, 8]. The values obtained for conserved polar or charged positions and conserved nonpolar or weakly polar positions are higher, and that there is often interaction with one of the two closest neighbours, but not with both of them.

It is obvious from these results that conserved hydrophobic, hydrophilic and nonpolar or weakly polar positions have very different properties. As already shown in previous publications, topohydrophobic positions (conserved hydrophobic positions) are very buried; they occupy what could be called the “inner

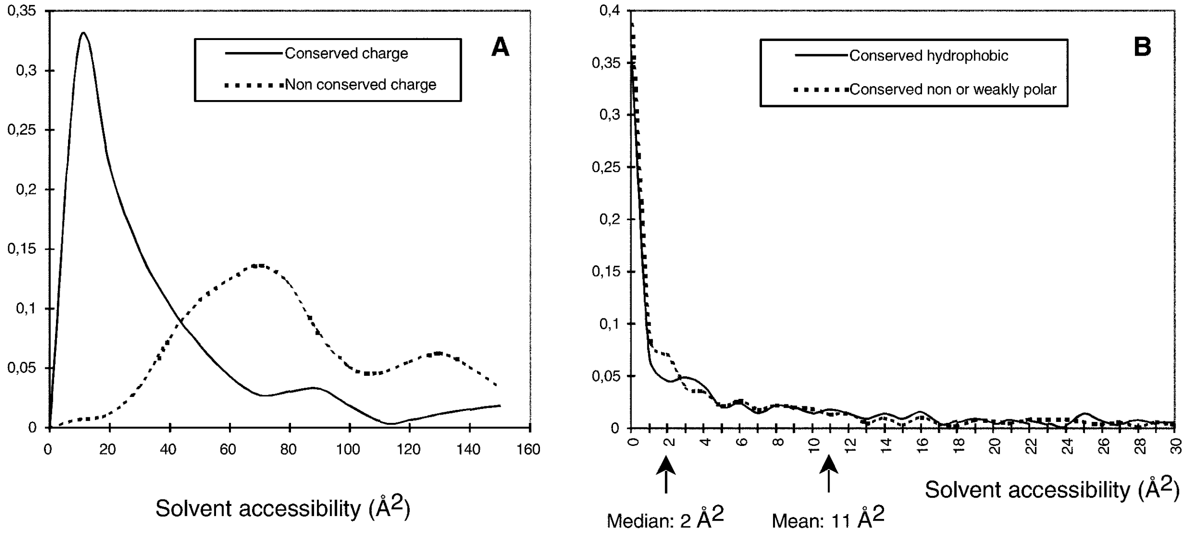

Fig. 2. A: Distribution of solvent accessibility for conserved hydrophilic positions with conserved charge (204 positions) or nonconserved charge (294 positions). **B:** Distribution of solvent accessibility for conserved hydrophobic and conserved nonpolar or weakly polar positions

core", the size of which is greatly reduced compared to that of the usual hydrophobic core; they are close to each other and they form a network of directly interacting residues in the core of the protein; the dispersion is very low, meaning that the orientation of the side chain is conserved from one protein to another within the same family and for the same position. Nonpolar or weakly polar conserved positions are also very buried, are found in the inner core of the protein and exhibit low dispersion, but they do not form a network. Mixed hydrophobic/nonpolar or weakly polar positions have properties similar to conserved hydrophobic and conserved nonpolar or weakly polar positions: they are very burred, they are found in the protein core, they have low dispersion, but they are not in direct interaction with each other. Mixed positions containing hydrophilic residues have very different properties: they are exposed and are found mostly at the surface of the protein, the side chains occupying such positions are very dispersed, and they do not make contact with each other.

3.2 Solvent accessibility distributions and prediction

As already shown for conserved hydrophilic positions, the distribution of solvent accessibility for a given position type is very asymmetric, especially for buried residues (Fig. 2B). Consequently, the mean value is not very relevant for the accessibility of the position type in a prediction perspective. The median value (the value dividing the population in two halves) is much more indicative (Fig. 3): for (1, 0, 0) positions the mean solvent accessibility is 11 Å^2 [15 Å^2 for (0, 1, 0) positions], whereas the median value is 1 Å^2 [3 Å^2 for (0, 1, 0) positions].

To investigate the possibility of predicting structural parameters, from sequences only, using the values obtained, we predicted the solvent accessibility for 90 proteins of known structure, from the multiple alignment in Pfam [15] to which they belong. These proteins were selected according to the following criteria: they should not be present in the set of structures used to compute the different values; the length of the multiple alignment should be more than 60 residues and has to contain more than ten sequences.

For each protein, a residue is assigned the median value corresponding to the position type it occupies in the multiple alignment. The accuracy of these predictions was then investigated for different thresholds (Fig. 4). The solvent accessibility is correctly predicted for more than 70% of the residues for any threshold value. Residue solvent accessibility is predicted if the corresponding position is occupied in all the sequences, which represent 97.6% of the residues present in Pfam alignments and 84% of the residues of the complete domains (because protein N- and C-terminal parts are most often omitted in the alignments).

The Q2 values obtained do not vary much with the threshold; however, the number of residues predicted in one category that truly belong to this category does depend on the threshold (Fig. 5A), as does the number of residues that belong to one category and are predicted in that category. As expected, the proportion of residues predicted to be buried that are truly buried increases as the threshold increases, and the proportion of residues predicted to be exposed that are truly exposed decreases. However, for every threshold tested, the numbers of residues predicted in one category and the number of residues which truly belong to that category are very close (Fig. 5B).

As a consequence of the method used to predict solvent accessibility, the accuracy of the prediction depends on the position type and on the chosen threshold. As shown in Fig. 6, for a given threshold, the positions for which the median value is close to the threshold are not well predicted (the lowest value is 0.46) and the positions for which the median value is very different from the threshold are very well predicted

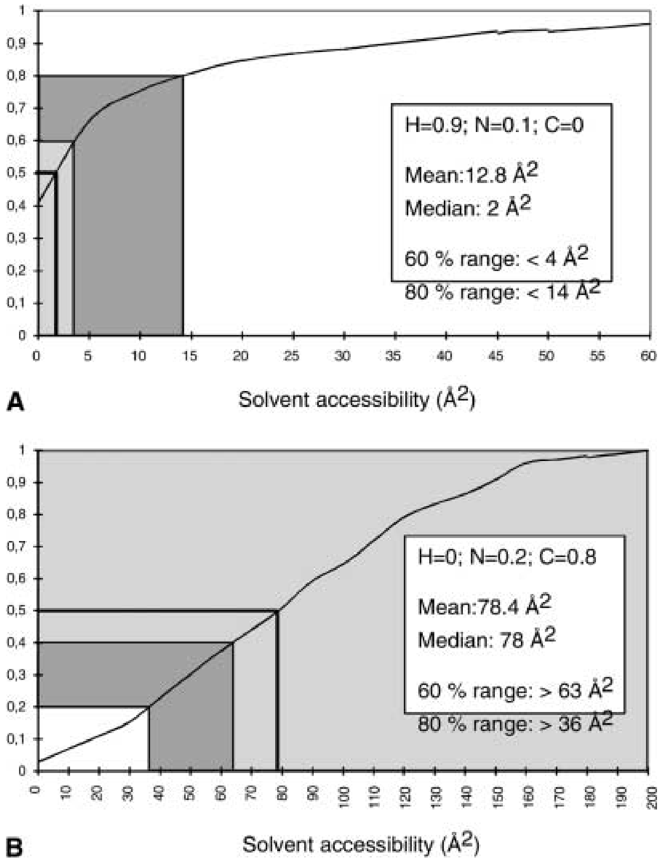

Fig. 3A, B. Solvent accessibility distributions for two different conserved position types. Accessibility range definitions. **A** For distributions having a median value under 30 Å^2, the 60% (or 80%) range is of the form [0, *x*] (or [0, *y*]), where *x* (or *y*) is the value for which 60% (or 80%) of the positions have accessibility lower than *x* (or *y*). **B** For distributions having a median value higher than 30 Å^2, the 60% (or 80%) range is of the form [*a*, +∞] (or [*b*, +∞]), where *a* (or*b*) is the value for which 60% (or 80%) of the positions have accessibility higher than *a* (or *b*)

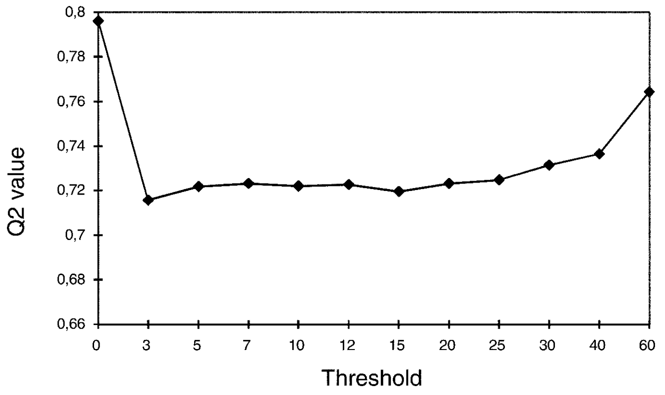

Fig. 4. Q2 values for solvent accessibility prediction as a function of the threshold. (The Q2 value is the number of correctly predicted residues divided by the number of predicted residues, times 100.)

(the highest value is 0.96). Thus, if, for a given threshold, only "well-predictable" positions are considered, the Q2 values are higher. For example, when the threshold is 15 Å^2, position types with median values between 5 and 15 Å^2 can be considered as difficult to predict. If the prediction is made only for the position types having median values outside this range, the Q2 value obtained is 75%, and 84% of the residues are predicted. Similarly, for a threshold of 25 Å^2, if position types with median values between 15 and 35 Å^2 are omitted, the Q2 value reaches 78% on 79% of the previously predicted residues.

Different approaches have previously been used for solvent accessibility predictions: neural networks [16, 17], Bayesian statistics [18] or logistic functions [19]. In particular, the later method shows that the neighbours of a residue can be used to predict its solvent accessibility. Such consideration is not taken into account in our method; thus, the combination of the two could give even better results.

3.3 Predicting solvent accessibility ranges

As already explained, the consequence of using a unique threshold is that some residues, whose accessibility is close to the threshold, are not well predicted. To overcome this problem, instead of predicting an accessibility relative to a threshold, we predict solvent accessibility ranges for each position type.

The ranges were computed from the structural alignments. We defined two ranges for each position type: the 60% range and the 80% range. For a given position, the 60% range (or the 80% range) is a region of solvent accessibility in which 60% (or 80%) of the residues occupying this type of position are found. If the median value for the given type is lower than 30 Å^2, the range starts at 0 Å^2 and so is more like a maximum value; if the median value is higher than 30 Å^2, the range is more like a minimum value and goes up to more than 200 Å^2 (Fig. 3, Tables 2, 3).

To test these ranges, we assigned these ranges to the residues of the previous 90 proteins. 59.4% of the residues have accessibility within the corresponding 60% range; 77.4% for the 80% range. Some of these ranges, especially for 80%, are very large. They correspond to positions which are not very well represented in the original database (Fig. 1D). These ranges could probably be narrowed by using more data for the computation; however, it should be noted that these large ranges also correspond to position types which are not very frequent in the alignments used to test the method, and not predicting them does not change the proportion of well-predicted residues.

4 Conclusion

The comparison of proteins sharing a same fold but no or low sequence identity allows the establishment of a correlation between the residues allowed at a given position of the fold and some structural parameters of this position. These results can be used to predict these structural parameters for proteins with unknown three-dimensional structure, using only a multiple alignment of divergent sequences. For example, solvent accessibil-

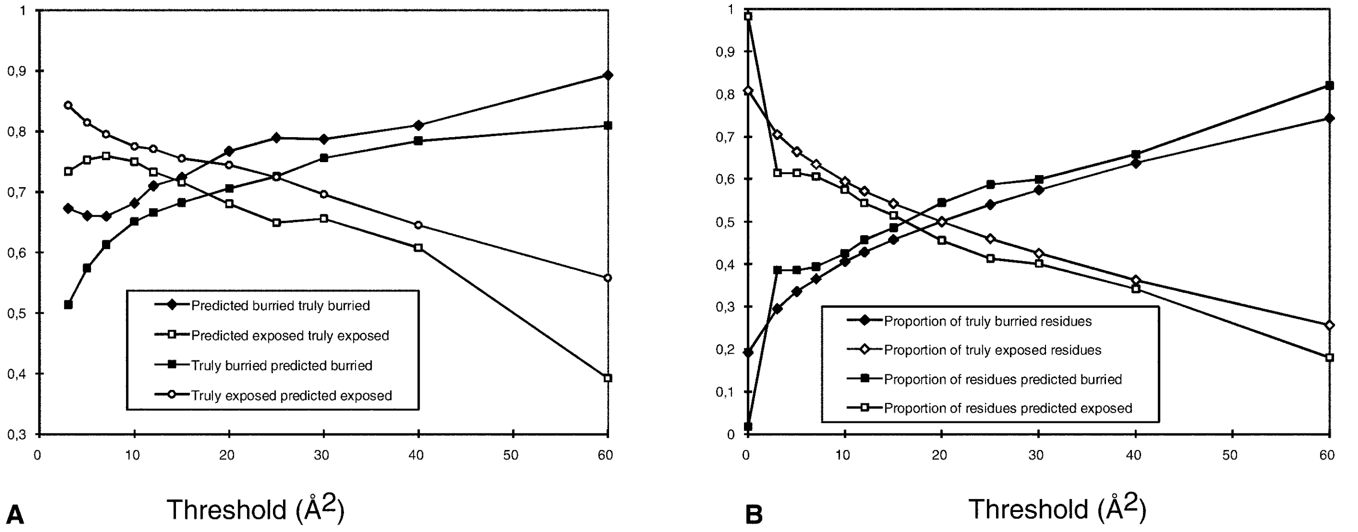

Fig. 5. **A** Proportions of residues predicted in one category which truly belong to that category and proportion of residues belonging to one category which are predicted in that category. **B** Proportions of residues which truly belong to each category and proportions of residues predicted in each category

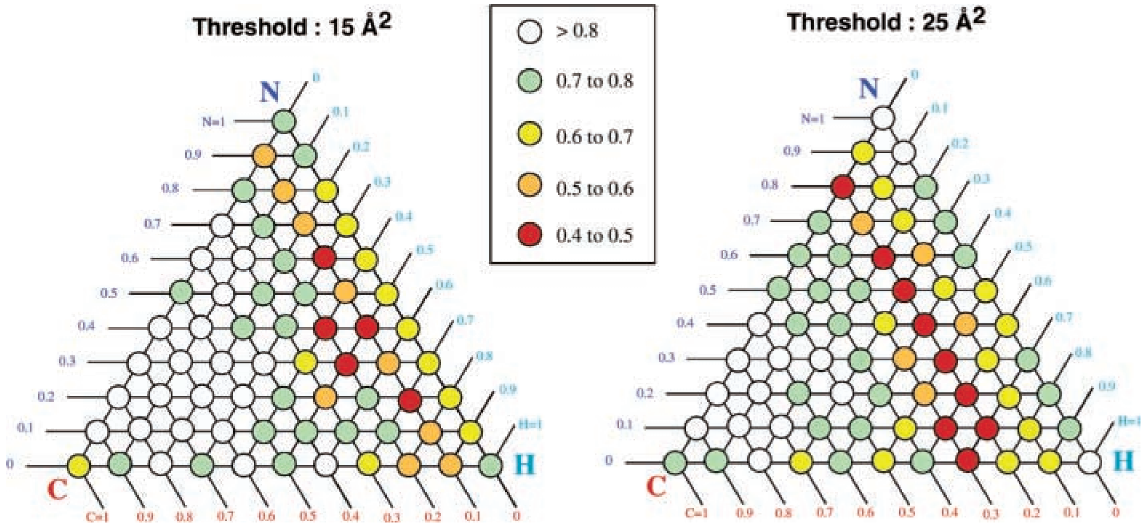

Fig. 6. Proportions of correct predictions for each position type and for two different thresholds: 15 and 25 Å^{2}

Table 2. 60% ranges for each position type

N\H	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	> 35[a]	> 36	> 18	> 34	> 25	< 54	< 47	< 28	< 36	< 22	< 4
0.1	> 57	> 55	> 40	> 29	> 41	> 28	< 54	< 32	< 18	< 4	
0.2	> 60	> 54	> 44	> 21	> 24	> 28	< 48	< 28	< 7		
0.3	> 57	> 47	> 47	> 24	< 60	< 30	< 22	< 6			
0.4	> 50	> 43	> 29	> 23	< 31	< 24	< 4				
0.5	> 50	> 39	> 28	< 40	< 18	< 4					
0.6	> 32	> 30	< 45	< 20	< 7						
0.7	> 37	< 66	< 25	< 4							
0.8	< 44	< 35	< 4								
0.9	< 34	< 3									
1	< 6										

[a] For hydrophilic positions with conserved charge, the 60% range is below 20

Table 3. 80% ranges for each position type

N\H	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	> 11[a]	> 8	> 1	> 13	> 2	< 70	< 64	< 42	< 57	< 37	< 17
0.1	> 22	> 19	> 10	> 13	> 4	> 8	< 78	< 58	< 36	< 17	
0.2	> 32	> 24	> 17	> 5	> 5	> 11	< 75	< 45	< 8		
0.3	> 25	> 21	> 14	> 7	< 74	< 54	< 46	< 27			
0.4	> 15	> 17	> 10	> 4	< 59	< 44	< 18				
0.5	> 15	> 17	> 6	< 56	< 46	< 14					
0.6	> 7	> 4	< 67	< 51	< 29						
0.7	> 14	< 73	< 48	< 22							
0.8	< 62	< 47	< 16								
0.9	< 46	< 13									
1	< 22										

[a] For hydrophilic positions with conserved charge, the 60% range is below 55

ity can be predicted with good accuracy: around 70%, depending on the threshold. Solvent accessibility ranges can be assigned to residues, with a rate of success of 59% or 77%, depending on the ranges chosen.

This method for solvent accessibility prediction does not consider the environment of each residue. As methods using this type of data also give very good results [19], a consensus method should give even better results.

References

1. Simons K, Bonneau R, Ruczinski I, Baker D (1999) Proteins Suppl 3: 171–176
2. Orengo C, Bray J, Hubbard T, Sillitoe I (1999) Proteins Suppl 3: 149–170.
3. Osguthorpe D (2000) Curr Opin Struct Biol 10: 146–152
4. Shortle D (1999) Curr Biol 9: R205–R209
5. Ortiz A, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Proteins Suppl 3: 177–185
6. Olmea O, Valencia A (1997) Folding Des 2: S25–S32
7. Poupon A, Mornon J-P (1998) Theor Chem Acc 101: 2–8
8. Poupon A, Mornon J-P (1998) Proteins 33: 329–342
9. Poupon A, Mornon J-P (1999) FEBS Lett 452: 283–289
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acid Res 25: 3389–3402
11. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5: 1093–1108
12. Sutcliffe MJ, Haneef I, Carney D, Blundell T (1987) Protein Eng 1: 377– 384
13. Kabsch W, Sander C (1983) Biopolymers 22: 2577–2637
14. Sonnhammer ELL, Eddy SR, Durbin R (1997) Proteins 28: 405–420
15. Bateman A, Birney E, Durbin R, Eddy S, Howe K, Sonnhammer E (2000) Nucl Acid Res 28: 263–266
16. Holbrook S, Muskal S, Kim S (1990) Protein Eng 3: 659–665
17. Rost B, Sander C (1994) Proteins 20: 216–226
18. Thompson M, Goldstein R (1996) Proteins 25: 38–47
19. Mucchielli-Giorgi M-H, Tuffery P, Hazout S (1999) Theor Chem Acc 101: 186–193