Theor Chem Acc (1999) 101:2–8
DOI 10.1007/s002149800m91

Theoretical
Chemistry Accounts

Regular article

"Topohydrophobic positions" as key markers of globular protein folds*

Anne Poupon, Jean-Paul Mornon

Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS UMRC7590, Universités P6 et P7, T16, Case 115, 4 place Jussieu, F-75232 Paris Cedex 05, France

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 16 November 1998

Abstract. The positions of a given fold always occupied by strong hydrophobic amino acids (V, I, L, F, M, Y, W), which we call "topohydrophobic positions", were detected and their properties demonstrated within 153 non-redundant families of homologous domains, through 3D structural alignments. Sets of divergent sequences possessing at least four to five members appear to be as informative as larger sets, provided that their mean pairwise sequence identity is low. Amino acids in topohydrophobic positions exhibit several interesting features: they are much more buried than their equivalents in non-topohydrophobic positions, their side chains are far less dispersed; and they often constitute a lattice of close contacts in the inner core of globular domains. In most cases, each regular secondary structure possesses one to three topohydrophobic positions, which cluster in the domain core. Moreover, using sensitive alignment processes such as hydrophobic cluster analysis (HCA), it is possible to identify topohydrophobic positions from only a small set of divergent sequences. Amino acids in topohydrophobic positions, which can be identified directly from sequences, constitute key markers of protein folds, define long-range structural constraints, which, together with secondary structure predictions, limit the number of possible conformations for a given fold.

Key words: Hydrophobic core – Solvent accessibility – Hydrophobicity – Folding – Modelling

1 Introduction

Many proteins are able to fold in physiologic conditions without the help of chaperon proteins [1]. Thus, in principle, it should be possible to predict the structure of a protein knowing only its sequence. However, our understanding of the driving forces of protein folding is still insufficient for this task, although there is ample evidence that hydrophobic amino acids play a key role in protein folding [2–8]. Hydrophobicity is one of the best conserved characteristics (of both buried and exposed amino acids) during evolution [9-12], but, surprisingly, buried hydrophobic amino acids are more often mutated than non-hydrophobic ones [13].

By comparing pairs of sequences for homologous domains of known 3D structure, two major populations of strong hydrophobic amino acids can be distinguished: those which share the same position in the two structures (and consequently in the two sequences), whatever their chemical nature, and those which are replaced in the other structure by non-strong hydrophobic amino acids. Calculation of the mean solvent accessibilities of these two populations showed that conserved amino acids are more buried than non-conserved ones.

That unpublished study has been extended to the analysis of families of proteins of known structure, within a non-redundant bank of 150 folds constituted for this purpose. Each family was structurally aligned and the properties of the amino acids in positions where only hydrophobic amino acids were found, which we call "topohydrophobic positions", were studied [14]. The results show that these amino acids must play a special role in folding and stability.

The properties of topohydrophobic positions are demonstrated here through structural alignments, using known 3D structures. However, even more interesting is the possibility of identifying these positions from sequence only, using sensitive sequence comparison methods such as bidimensional hydrophobic cluster analysis (HCA) [15–17], although with a lower accuracy than with structural alignments.

2 Methods

Protein databanks were searched using the BLAST network server at the NCBI (National Center for Biotechnology Information) with

*Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambercy, France

Correspondence to: A. Poupon
e-mail: poupon@lmcp.jussieu.fr

default parametrization [18] in order to harvest all known 3D structures belonging to a same structural family. Each BLAST search output was analyzed as a single multiple alignment of the significant pairwise alignments, through processing by MULBLAST [19]. Editing of multiple alignments and profile database screenings were performed with the programs LINEUP and PROFILESEARCH from the software package GCG 7.0 (Genetic Computer Group Inc., Wis., USA). The SWISSPROT Database (release 32) was used for profile screenings. The statistical validation of the similarities came from the Poisson law probabilities calculated by the program BLAST and from the PROFILESEARCH Z score, the TOPITS search [29] and the homemade program TZscore [15]. 3D visualization was performed on a UNIX workstation using the program XmMol [20], and on a Silicon-Graphics using the program INSIGHT II, release 2.3.0 (Biosym Technologies, San Diego, Calif., USA). The program COMPOSER was used to accomplish preliminary 3D superimposition, preliminary 3D alignments and root mean square distance (RMSD) calculations [21]. Alignments were manually checked and refined using an iterative process.

Final superimpositions, according to the structural alignments, RMSD calculations, solvent accessibilities (using the algorithm of Lee and Richards [22], multimers were completed before calculation, incomplete residues were excluded) and generation of aleatory sub-families were performed by the program CHAP (Poupon, 1997, unpublished).

3 Results and discussion

A large non-redundant bank of fold families containing as much divergence as possible was needed for this study. The existing banks contain mainly families sharing around 30% sequence identity, and a few families with no detectable sequence identity. To harvest families, a subset of the Protein Data Bank (PDB, 1/1/97 release [23]), containing only sequences sharing less than 30% identity with any other one, was initially used to constitute the fold families. Each sequence of the subset was used as a probe for screening of the sequence banks with the program BLAST [18]. Each sequence suspected to share 3D similarities with the probe was checked using the HCA method [15-17], then used as a probe for a new screening of the sequence banks. Finally, only the sequences with known 3D structure were retained in the family.

In each putative fold family, the largest subset of proteins sharing less than 55% identity with any other member of the subset was determined using Hobohm and Sander's bank [24, 25]. This ensures that the proteins with the best resolution are chosen preferentially.

The proteins of each family were structurally aligned using the program COMPOSER [21]. All alignments were then manually checked through an iterative process. For each step, the alignment was used for superimposition. C_α pairs that are too distant (>3Å) were discarded, leading to a new alignment used for superimposition. C_α pairs which became close enough in this process are reintroduced in the alignment for the next step. In this manner, it was possible to obtain RMSD lower than 3 Å for 97% of the families.

The bank contains 648 proteins, divided into 153 families comprising at least two members; 292 proteins of the original 30% identity subset remained single (Table 1).

It is essential to emphasize that the alignments used hereafter for all calculations derive from structural superimpositions and not at all from sequence comparisons. Sequence alignments were used only when generating the sequence family for each protein of the original 30% identity subset of the PDB.

All the families of other structural databases such as CATH [26], SCOP [27] or Homstrad [28] are represented in our bank, but a significant number of families in our bank are not represented in the other databases, mainly because the homology between their members cannot be automatically detected.

Table 1. Size of the families

Number of members	2	3	4	5	>5
Number of families	61	33	19	13	27

Table 2. Tests of significance of the difference between mean solvent accessibility in topohydrophobic and non topohydrophobic positions. In each family, and for each concerned amino acid, the difference between mean solvent accessibility in topohydrophobic and non topohydrophobic positions was tested for a first kind error α of 0.05 and 0.1. Part A summarizes the results obtained with group I amino acids and strict topohydrophobic versus non topohydrophobic positions. Part B summarizes the results obtained with group I and II amino acids and extended topohydrophobic versus non topohydrophobic positions.

A.

Number of proteins	Total number of tests	Significant tests for $\alpha = 0.05$ (%)	Significant tests for $\alpha = 0.1$ (%)	Non significant tests for $\alpha = 0.05$ (%)	Impossible tests (%)
2	427	6.8	14.5	50.0	43.2
3	245	31.0	37.1	35.1	33.9
4	126	33.1	40.2	34.6	32.3
5	91	50.0	53.0	20.6	28.3
>5	182	36.8	40.1	24.7	38.4

B.

Number of proteins	Total number of tests	Significant tests for $\alpha = 0.05$ (%)	Significant tests for $\alpha = 0.1$ (%)	Non significant tests for $\alpha = 0.05$ (%)	Impossible tests (%)
2	854	6.3	10.4	40.0	53.7
3	490	15.6	19.6	31.2	64.5
4	252	19.4	21.8	25.0	55.9
5	182	34.1	36.3	31.9	34.1

Strong hydrophobic amino acids considered in the present study are those used in HCA [15], i.e., V, I, L, F, M, Y and W. They have been shown to mainly constitute the internal sides of regular α and β secondary structures, by computation of their observed preferential participation in these elements. Indeed, sorting amino acids by decreasing order of the sum of their respective propensities to be α, β or coil, allows three obvious groups to be distinguished : group I, for which the propensity for the coil is lower than the propensities for alpha helices and beta strands (V, I, L, F, M, Y and W); group II, with a coil propensity similar to those for alpha and beta conformations (A, R, C, Q, T, E and K); and group III, preferentially associated with loops in the increasing order H, S, N, D, P, and G.

For each alignment, positions occupied only by strong hydrophobic amino acids were determined and are called strict topohydrophobic positions. Partial topohydrophobic positions were defined as positions where the percentage of strong hydrophobic amino acids is greater than or equal to 75%, and strictly lower than 100%, the other amino acids (mainly A, C, T) in the position belonging to group II, defined above. The expression "extended" topohydrophobic position is used for positions which are strict or partial topohydrophobic positions.

3.1 Solvent accessibility of topohydrophobic positions

Figure 1A shows that, for each hydrophobic amino acid, the mean ratio between solvent accessibilities in non-topohydrophobic versus topohydrophobic positions (strict or extended) is high. The ratio is largely greater

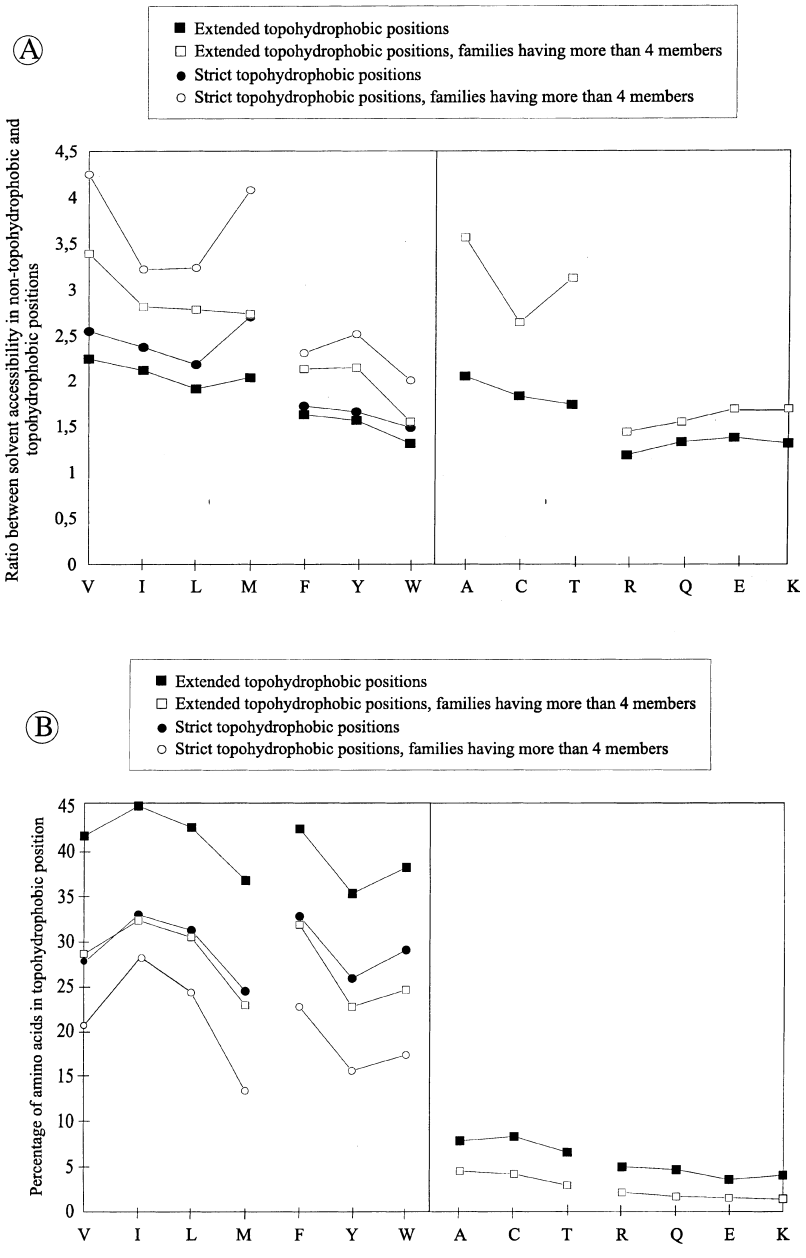


Fig. 1. Ratio between mean solvent accessibilities in topohydrophobic and non-topohydrophobic positions **A** For each amino acid, the ratio between mean solvent accessibilities when involved in topohydrophobic or non-topohydrophobic positions is computed, in the case of strict topohydrophobic positions (● all families; ○ families having more than four members) for group I amino acids, and in the case of extended topohydrophobic positions (■ all families; □ families having more than four members) for group I and II amino acids. **B** Percentage of each amino acid involved in strict (● all families; ○ families having more than four members) or extended (■ all families; □ families having more than four members) topohydrophobic positions. The number of amino acids in one category, which ranges from 164 to 9127, allows representative score calculations

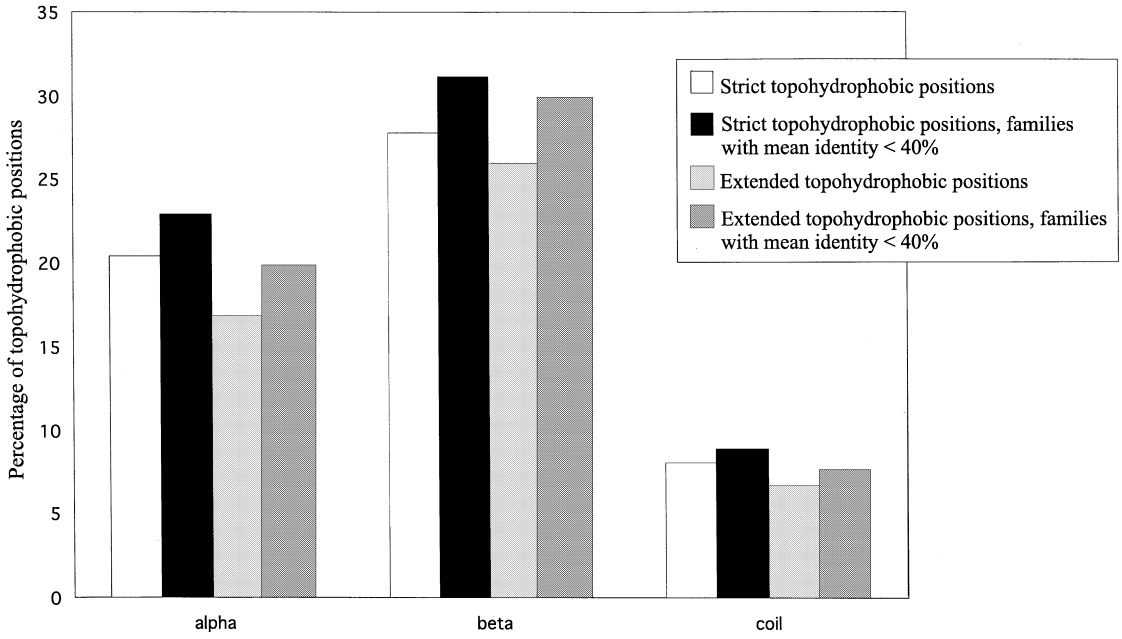
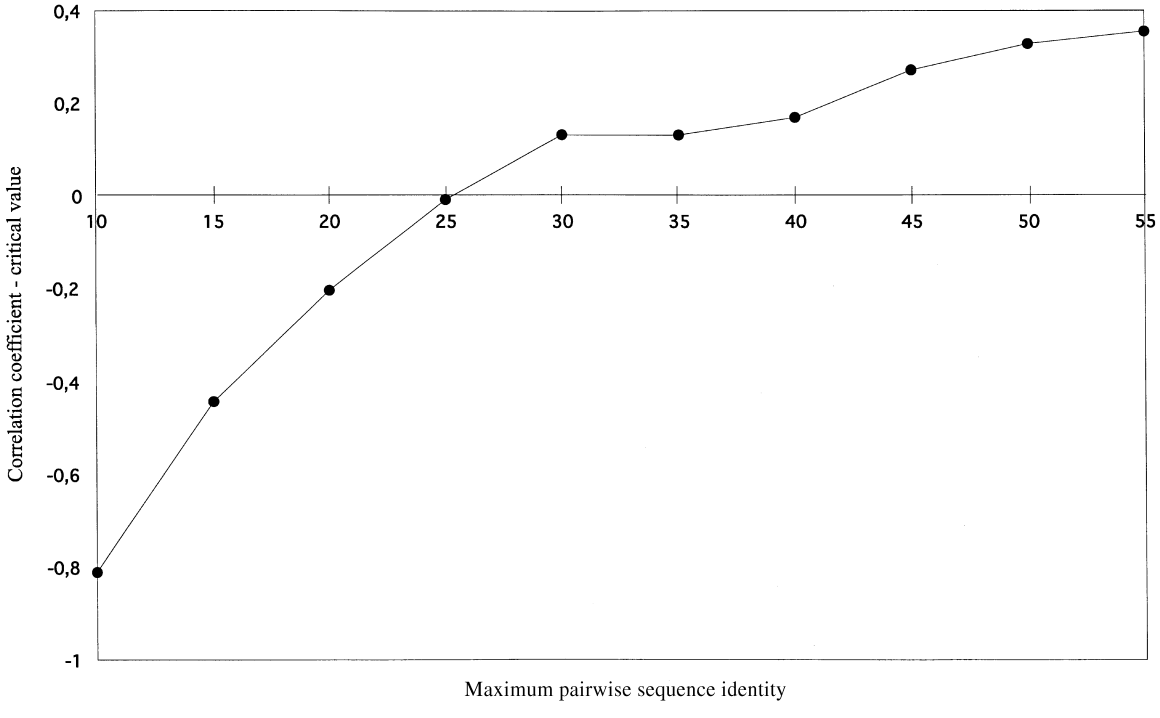


Fig. 2. Proportions of topohydrophobic positions in secondary structure elements. For each type of secondary structure element, the percentage of positions which are strict or extended topohydrophobic is calculated, for all families, and for families with a mean pairwise sequence identity lower than 40%

than 2 for the aliphatic V, I, L and M amino acids, and often above 1.5 for the aromatic amino acids (F, Y, and W). Consequently, hydrophobic amino acids in topohydrophobic positions are clearly more buried than the same amino acids in non-topohydrophobic positions. When computed on the whole bank, the differences between these values are significant for all amino acids, and indicate that topohydrophobic positions mainly occupy the inner core of globular domains (Table 2A). Figure 1 also shows that, in extended topohydrophobic positions, A, C, and T are naturally distinguished from the remaining amino acids of group II.

Fig. 3. Topohydrophobic positions and sequence identity. For each family, the percentage of amino acids involved in strict topohydrophobic positions is computed. The correlation coefficient between this percentage and the mean pairwise sequence identity in the family is then computed for all families having a mean pairwise identity lower than 10, 15, ..., 55%. The deviation between the correlation coefficient and the critical value for a first kind error $\alpha = 0.05$ is plotted as a function of the maximum mean pairwise sequence identity



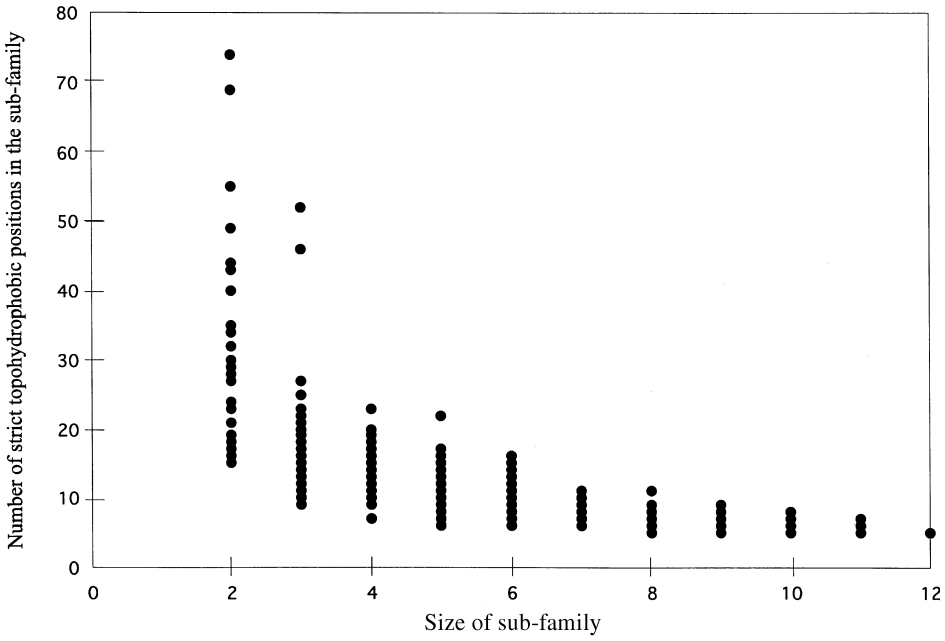


Fig. 4. Number of topohydrophobic positions in sub-families of the virus coat protein family. For each generated sub-family of the virus coat protein family, the number of strict topohydrophobic positions is plotted as a function of the sub-family size. For 2-member and 11-member sub-families, all the possible sub-families were generated; for the other sizes, 50 sub-families of each size were randomly chosen

When computed in each family for strict or partial topohydrophobic positions, these differences are significant in half the families containing three or more members, where the test is possible (Table 2B). For families containing two members, conservations of hydrophobic amino acids are mainly due to sequence relatedness and not to structural necessities. There are only very few cases in which the mean accessibility of an amino acid in a topohydrophobic position is higher than that of the same amino acid in a non-topohydrophobic position, and in this latter case, the difference is never significant. The main difference between the results obtained with strict or partial topohydrophobic positions is the proportion of impossible tests (one of the categories is empty) which is higher in the case of partial positions.

Approximately 15–30% of the strong hydrophobic amino acids are involved in strict topohydrophobic positions (Fig. 1B), and about 25–35% in extended ones, while only 8% of A, C or T, and 4% of remaining group II amino acids are involved in extended topohydrophic positions. Interestingly, it can be seen in Fig. 1 that V, I, L for aliphatic amino acids, and F for aromatic ones, constitute the hard nucleus of hydrophobic amino acids, as previously reported [15, 17, 19].

3.2 Distribution of topohydrophobic positions

In the studied set of protein families, strict topohydrophobic positions respresent 12.4% of all positions, a proportion which reaches 14.9% when partial positions are included (10.7 and 13.4% for families with mean pairwise sequence identity lower than 40%).

Table 3 and Fig. 2 show that, as expected, topohydrophobic positions mainly occupy regular secondary structures (usually one or two strict topohydrophobic positions per β-strand, two or three per α-helix).

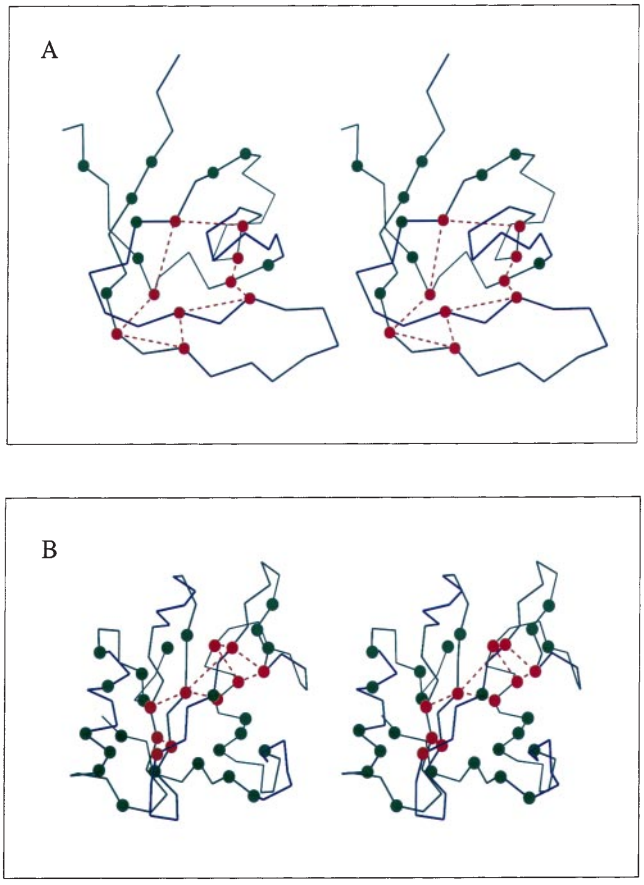


Fig. 5. Topohydrophobic lattice. In an SH3 domain (**A**) and a Tyrosine phosphatase (**B**), C_α of amino acids involved in topohydrophobic positions are shown in *red*, and C_α of hydrophobic amino acids involved in non-topohydrophobic positions in *green*. The topohydrophobic lattice is indicated with *red dashed lines*; each line associates amino acids whose gravity centers are closer than 6 Å

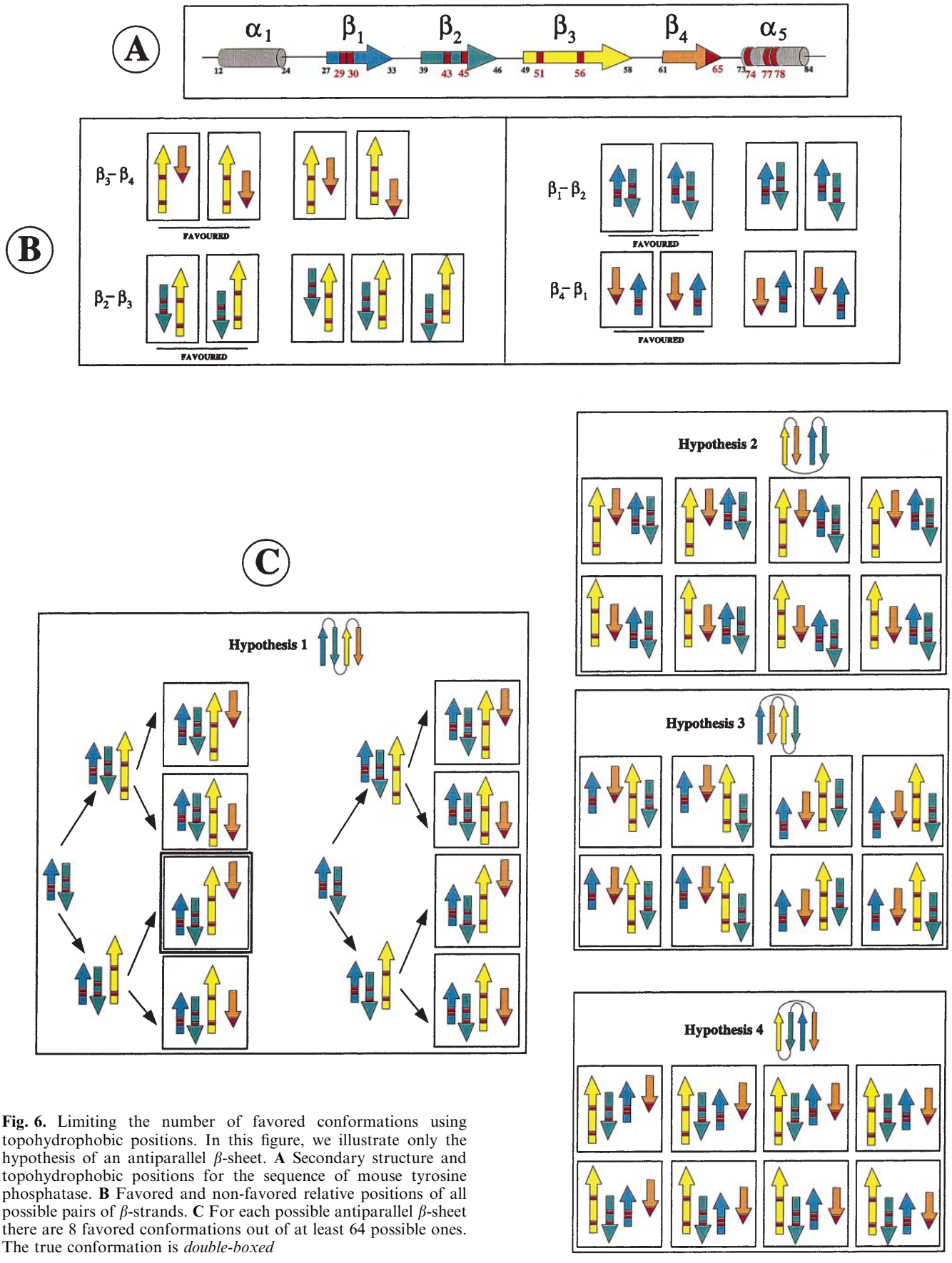


Fig. 6. Limiting the number of favored conformations using topohydrophobic positions. In this figure, we illustrate only the hypothesis of an antiparallel β-sheet. **A** Secondary structure and topohydrophobic positions for the sequence of mouse tyrosine phosphatase. **B** Favored and non-favored relative positions of all possible pairs of β-strands. **C** For each possible antiparallel β-sheet there are 8 favored conformations out of at least 64 possible ones. The true conformation is *double-boxed*

Table 3. Distribution of topohydrophobic positions in secondary structure elements. For the three main types of secondary structure elements (α helix, β strand and coil), the mean number of strict and partial topohydrophobic positions (lanes 3 and 4) in one element was determined. The number of representatives of each element in the bank is indicated in lane 2 (an element is counted only once per family).

Secondary structure	Total number of considered secondary structures	Mean number of strict topohydrophobic positions	Mean number of extended topohydrophobic positions
α helix	574	2.25	2.52
β strand	1005	1.67	1.87
Coil	1660	0.54	0.59

Figure 3 shows that, above the sequence "twilight zone" (25-30%), the detection of topohydrophobic positions is perturbed by sequence relatedness. Consequently, the divergence area, accessible through HCA below this zone, would be of great interest.

3.3 Influence of family size

To check the influence of family size on the detection of topohydrophobic positions, large families possessing more than ten members were used to generate numerous random sub-families of lower size. Figure 4 illustrates a typical result for a 12 member family (virus coat proteins). The number of topohydrophobic positions detected decreases sharply as the size of the sub-family increases, and rapidly converges toward the final number (here, 5 topohydrophobic positions), provided that the mean pairwise sequence identity in the sub-families is similar to that of the complete family (data not shown).

In many cases, five divergent members are sufficient to ensure a virtually complete detection of all topohydrophobic positions. This property will become especially useful in the automatic detection of topohydrophobic positions from limited sets of divergent sequences, using sensitive alignment procedures such as HCA [15, 16].

3.4 Topohydrophobic positions as key markers of folds

Compilation of the whole set of families shows that the side chains of amino acids in topohydrophobic positions are far less dispersed than those of the same amino acids in non topohydrophobic positions [14]. Moreover, they constitute a lattice of positions in contact with each other (most of the time situated in the inner part of the hydrophobic core, data not shown) as illustrated in Fig. 5.

Consequently, the identification of topohydrophobic positions from sequences only is a useful structural constraint to considerably limit the number of favored conformations, as schematically illustrated in Fig. 6. Indeed, we directly detect, with topohydrophobic positions, long-range key markers of folds, distributed all along family sequence alignments.

Prediction of the nature and approximate limits (±2 aa) of secondary structure elements, which is becoming increasingly possible with the use of divergent sequence families (data not shown), combined with the detection of topohydrophobic positions, would therefore represent a promising new contribution to the ab-initio prediction of protein globular domain 3D structures.

Acknowledgement. The authors thank the CNRS Program Genome for financial support of these studies.

References

1. Anfinsen CB (1973) Science 181: 223–230
2. Nozaki Y, Tanford C (1971) J Biol Chem 246: 2211–2217
3. Tanford C (1978) Science 200: 1012
4. Wolfenden R, Andersson L, Cullis PM, Southgate CCB (1981) Biochemistry 20: 849–855
5. Wolfenden R (1983) Science 222: 1087–1093
6. Chothia C (1976) J Mol Biol 105: 1–14
7. Chothia C (1984) Annu Rev Biochem 53: 537–572
8. Janin J (1979) Nature (Lond) 277: 491–492
9. Lawrence C, Auger I, Manella C (1987) Proteins 2: 153–161
10. Ladunga I, Smith RF (1997) Prot Eng 10: 187–196
11. Koshi JM, Goldstein RA (1997) Proteins 27: 336–344
12. Rost B, Sander C (1994) Proteins 20: 216–226
13. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Prot Sci 1: 216–226
14. Poupon A, Mornon J-P (1998) Proteins (in press)
15. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon J-P (1997) Cell Mol Life Sci 53: 621–645
16. Gaboriaud C, Bissery V, Benchetrit T, Mornon JP (1987) FEBS Lett 224: 149–155
17. Woodcock S, Mornon J-P, Henrissat B (1992) Prot Eng 5: 629–635
18. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) J Mol Biol 215: 403–410
19. Labesse G (1996) CABIOS 12: 463–467
20. Tuffery P (1995) J Mol Graphics 13: 67–72
21. Sutcliffe MJ, Haneef I, Carney D, Blundell T (1987) Prot Eng 1: 377–384
22. Lee BK, Richards FM (1971) J Mol Biol 55: 379–400
23. Bernstein FC, Koetzle TF, Williams G (1977) J Mol Biol 112: 535–542
24. Hobohm U, Scharf M, Schneider R, Sander C (1992) Prot Sci 1: 409–417
25. Hobohm U, Sander C (1994) Prot Sci 3: 522–529
26. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5: 1093–1108
27. Overington JP, Zhu Z-Y, Sali A, Johnson MS, Sowdhamini R, Louie GV, Blundell TL (1993) Biochem Soc Trans 21: 597–604
28. Gernstein M, Levitt M (1998) Prot Sci 7: 445–456
29. Rost B (1995) ISmb 3: 314–321