CMLS, Cell. Mol. Life Sci. 58 (2001) 492–498
1420-682X/01/030492-07 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Research Article

Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding?

M. Lamarine [a], J.-P. Mornon [a], I.N. Berezovsky [b] and J. Chomilier [a, *]

[a] Equipe Systèmes Moléculaires et Biologie Structurale, LMCP, Universités Paris 6 and Paris 7, CNRS UMR 7590, Case 115, 75252 Paris cedex 05 (France), Fax +33 1 44 27 37 85, e-mail: Jacques.Chomilier@lmpc.jussieu.fr
[b] Department of Stuctural Biology, The Weizmann Institute of Science, Rehovot, 76100 (Israel)

Received 1 February 2001; accepted 7 February 2001

Abstract. Using a set of 372 proteins representative of a variety of 56 distinct globular folds, a statistical correlation was observed between two recently revealed features of protein structures: tightened end fragments or 'closed loops', i.e. sequence fragments that are able in three-dimensional (3D) space to nearly close their ends (a current parameter of polymer physics), and 'topohydrophobic positions', i.e. positions always occupied in 3D space by strong hydrophobic amino acids for all members of a fold family. Indeed, in sequence space, the distribution of preferred lengths for tightened end fragments and that for topohydrophobic separation match. In addition to this statistically significant similarity, the extremities of these 'closed loops' may be preferentially occupied by topohydrophobic positions, as observed on a random sample of various folds. This observation may be of special interest for sequence comparison of distantly related proteins. It is also important for the ab initio prediction of protein folds, considering the remarkable topological properties of topohydrophobic positions and their paramount importance within folding nuclei. Consequently, topohydrophobic positions locking the 'closed loops' belong to the deep cores of protein domains and might have a key role in the folding process.

Key words. Protein folding; protein fragment; closed loops; topohydrophobic position.

The structure of globular domains of proteins is mainly characterised by the dichotomy between a hydrophilic surface (typically two-thirds of the amino acids of a domain) and an internal hydrophobic core (one-third). Within this internal core, two species of positions occupied by hydrophobic amino acids are distinguishable: on the one hand, positions always occupied by strong hydrophobic amino acids for all the members of a family sharing the same fold [1–2], recently called 'topohydrophobic positions' [3–6], and on the other hand, positions filled either by hydrophilic, neutral or hydrophobic amino acids (non-topohydrophobic positions). Precise knowledge of topohydrophobic positions from sequences alone is possible by means of multiple alignment of representative sequences of a common fold.

Topohydrophobic positions exhibit remarkable properties: e.g. they are significantly more buried than non-topohydrophobic ones; they constitute a set of mutually interacting side chains, evidence of short-distance-interacting residues and they are the main components of folding nuclei [5–8]. Topohydrophobic positions represent between one-third and one-half of the total number of hydrophobic amino acids for a domain and constitute a smaller inner entity of the hydrophobic core (e.g. T-folded proteins [9] and the uteroglobin fold [10]). In the protein set used in the present paper, the mean number of

* Corresponding author.

topohydrophobic positions in helices, strands and coils is 2.25, 1.67 and 0.54 positions, respectively [4]. It should be noted that in our data set, 80% of the topohydrophobic residues occur in regular secondary structures.

A recent study using a polymer-statistical approach has demonstrated the universal existence of 'closed loops', consisting of chain fragments with both ends in close contact in the three-dimensional (3D) space [11], typically less than 10 Å between alpha carbons (C_αs). These fragments show a strong preference to be built by 22–32 amino acids and to gather their nearly closed extremities in the internal core of proteins rather like petals in a daisy flower (see fig. 1B). To avoid any ambiguity with classical 'loops', we prefer to name these nearly standard size closed loops of length centred on 27 amino acids, tightened end fragments or TEFs. Therefore, a globular domain can be split into TEFs, i.e. 'closed loops' [11], the sizes of which often include between 22 and 32 amino acids.

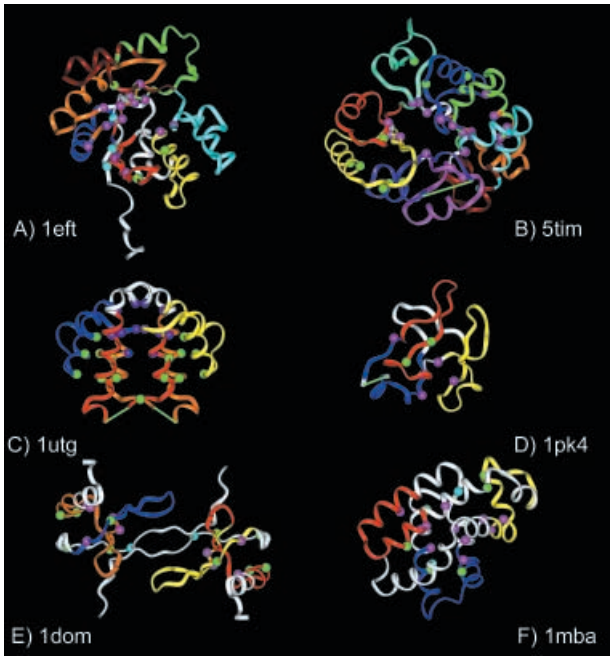

Figure 1. Ribbon representation of six distinct folds: elongation factor-Tu, PDB code 1eft (*A*), triosephosphate isomerase, PDB code 5tim (*B*), uteroglobin, PDB code 1utg (*C*), chemoattractant cytokin, PDB code 1pk4 (*D*), Plasminogen Kringle 4, PDB code 1dom (*E*), Myoglobin, PDB code 1mba (*F*). Different colours show the tightened end fragments (TEFs) [11]. The chains outside TEFs are left white. Purple spheres indicate the location of topohydrophobic positions within ±3 amino acids from the extremities of TEFs. Topohydrophobic positions inside TEFs but outside this range are coloured green and those far from a TEF extremity (more than 8 amino acids) in blue. A green line represents a virtual connection between N and C termini which may allow additional TEFs to be built through both ends of a chain, e.g. the purple one of the Tim-barrel (*B*), red and orange ones in the dimeric uteroglobin (*C*) and the blue one in the hydrolase (*D*). Overlapping parts of TEFs are indicated by alternate colours on the ribbons.

We show in this paper that in sequence space, the statistical distribution of lengths for TEFs and that of topohydrophobic separation match. They both appear to be often in physical concordance: TEFs preferentially begin, or end, at a topohydrophobic position, or very close to it. Generally, TEFs are built by two regular secondary structures and consequently are large parts of or full classical super-secondary structures. This correspondence may constitute a step in the field of ab initio protein structure prediction [12], as it may help to identify pairs of interacting topohydrophobic positions within protein cores.

Materials and methods

The protein set analysed in this study comprised 372 structures from the Protein Data Base (PDB) [13], grouped into 56 families of more than three members and representative of the different classes of fold, the sequences of which have been aligned using 3D structures and excluding too closely related proteins (pairwise sequence identity between pairs is below 55%) [3]. When at least three-quarters of the residues occupying a given position in the alignment are clearly hydrophobic (VILFMWY), this position was considered as topohydrophobic. All residues were assigned to one of three groups: group 1 comprised hydrophobic residues, group 2 comprised ACTPGSH and group 3 comprised all other residues. If a position was occupied by less than one-quarter of residues from group 2, it was still considered topohydrophobic, while members of group 3 were excluded from these positions [6]. Typically, one-third to one-half of the amino acids of group 1 occupy topohydrophobic positions, depending upon the protein family [5]. From this data set, we computed the distances between all pairs of C_α. The data were then entered in large data files and processed with SAS statistical software (SAS Institute, Cary, N.C.). In such an analysis, histograms can then be produced for any kind of residue pair, and cut-off distances can be imposed.

Results

Figure 1 illustrates, on six distinct folds, TEFs calculated according to Berezovsky et al. [11] as well as the distribution of topohydrophobic positions deduced from multiple alignment of the corresponding protein families [3]. The histograms of C_α-C_α distances were derived from the data set of 372 proteins, for pairs of topohydrophobic or non-topohydrophobic positions (fig. 2). Basic statistics for amino acid pairs separated by a given sequence distances, are given in figure 3. Direct comparison of topohydrophobic positions with the extremities of TEFs

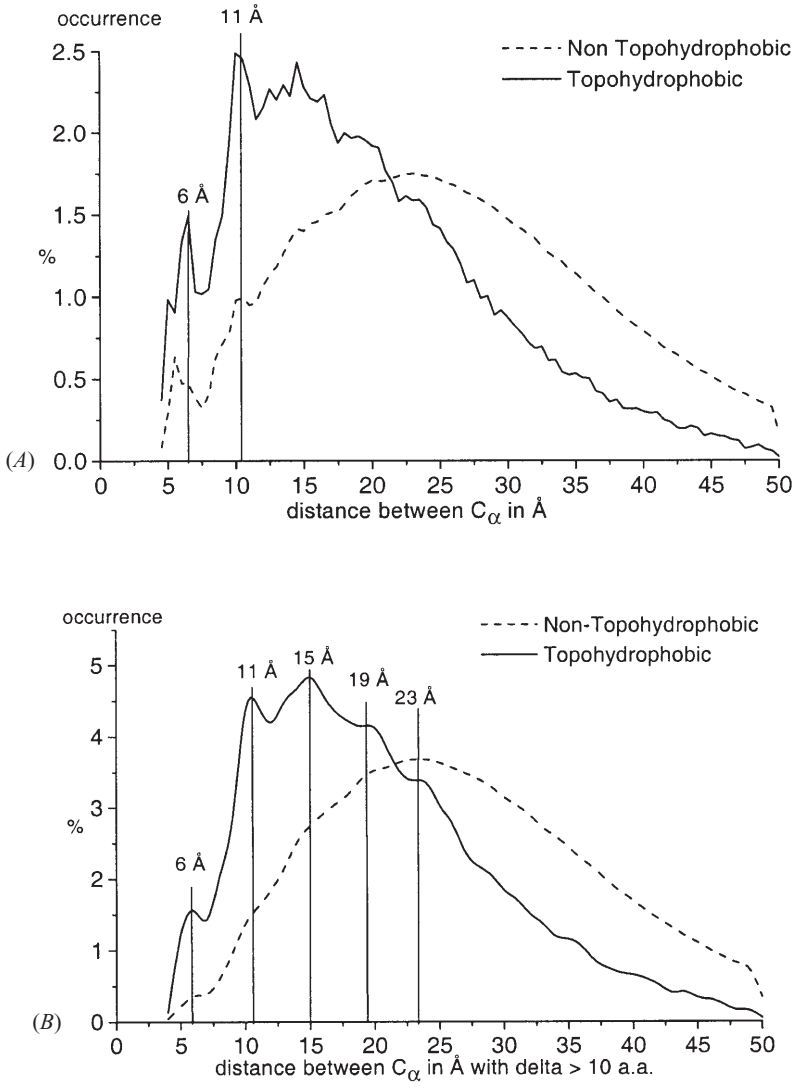

Figure 2. (*A*) Histogram, sampled every 0.5 Å, of the distances between residue pairs in the entire data set without any restriction on sequence separation between the two members of a pair, for topohydrophobic pairs (solid line) and non-topohydrophobic pairs (broken line). (*B*) Similar histogram, sampled every 1 Å in order to have meaningful statistics, of the distances between residue pairs in the entire data set with a sequence separation (delta) larger than ten amino acids to avoid the effects of local contacts, for topohydrophobic pairs (solid line) and for non-topohydrophobic pairs (broken line).

(fig. 4) was performed on a sample set of 15 distinct folds (PDB codes: 1pk4, 1utg, 1mba, 4blm, 1uox, 1eft, 1fiv, 1frd, 3c2c, 5tim, 1dom, 2aaa, 2fcr, 1gky, 1amy).

Histogram of C_α-C_α distances

In figure 2 A, the histogram of distances between all pairs of topohydrophobic residues (104,505 pairs) in the entire data set (solid line) is compared to the corresponding histogram for non topohydrophobic residues (7,866,263 pairs) (broken line). Mixed pairs (topohydrophobic with non topohydrophobic) were not taken into account. To have the same scale for both curves, the number of pairs at a given distance is expressed as a percentage of the corresponding total number of pairs. Figure 2 A shows that the distribution maximum occurs around pair distances of 15 Å for topohydrophobic residues, and is shifted towards 25 Å for non-topohydrophobic amino acids. This shift is expected, since topohydrophobic positions form the inner core of globular protein domains [4–5]. There are two secondary maxima in both curves centred around 6 and 11 Å, i.e. at short distances, corresponding to the first and second neighbour shells in globular proteins. Interestingly, globular proteins were recently shown to be as compact as a random packing of hard spheres encountered in the case of condensed matter [14, 15]. In the case of compact hexagonal or cubic structures, the 12 first neighbours are at a distance of 6.5 Å, and the 24 third neighbours are at 11.2 Å. There are only 6 second neighbours

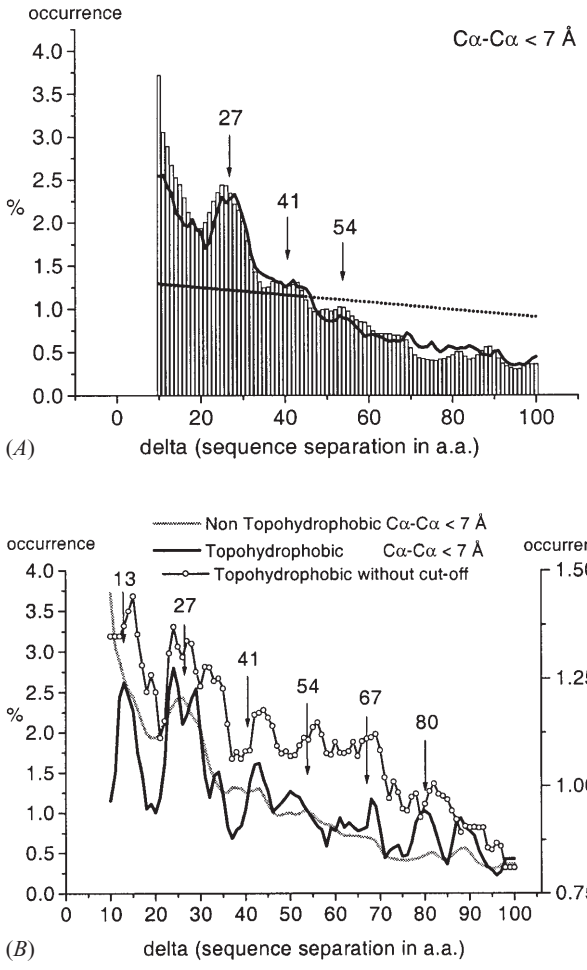

Figure 3. (*A*) Histogram of C_α carbon distances less than 7 Å from the entire protein set of this study as a function of sequence separation (delta), given for all residues and compared to the same data derived from a previously published data set [11] (thick solid line). The histogram is expressed as a percentage of the total number of pairs observed in the data set. Mean peak positions are indicated by arrows and their delta values are given. The dotted line corresponds to the distribution of pairs without any distance cut-off. (*B*) Histogram of C_α-C_α distances less than 7 Å for topohydrophobic (solid line) and non-topohydrophobic (grey line) amino acids, using the data base of this study (left scale). The topohydrophobic plot without distance cut-off is shown by circles (right scale).

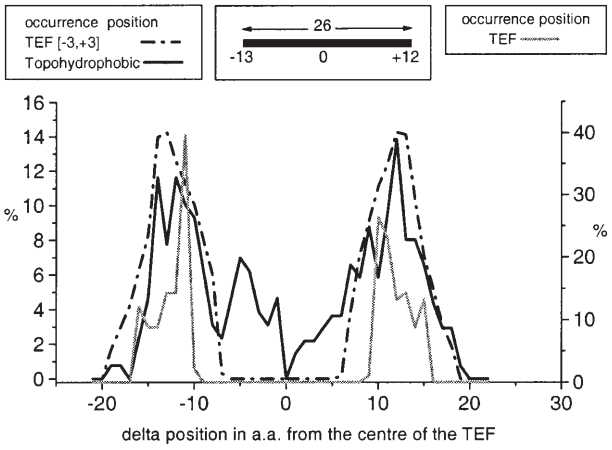

Figure 4. Histogram of the occurrences of topohydrophobic positions on a set of 15 proteins of various common folds (continuous line) compared to the positions of TEF ends (grey line). Numbering origin is chosen at the centre of the TEF (e.g. a TEF of 26 residues ranges from −13 to +12). Otherwise, to permit some uncertainty at end positions covering the bell-shaped peak at delta 27 of figure 3A, TEFs were allowed to shift by 3 residues in either direction (dashed-dotted line).

at 9.2 Å, not visible in figure 2A because they are situated in the shoulder of the next peak (11 Å).

To further investigate these first peaks, we restricted the pair distance distribution to residues separated by more than 10 amino acids along the sequence (fig. 2B.) In this case, there are 55,413 topohydrophobic pairs and 3,786,239 non-topohydrophobic pairs, corresponding to a slightly larger proportion of topohydrophobic pairs (1.5%) than in the case of figure 2A. These data therefore indicate short-range interactions in the 3D space corresponding to long-range distances in sequence, i.e. more or less to the core of the protein, as is typical of the topohydrophobic distribution. The peaks at 6 and 11 Å are still present for the topohydrophobic positions but practically vanish for the other distribution. For longer distances, 'topohydrophobic' peaks are regularly observed at 15, 19 and 23 Å. These distances are more difficult to interpret, first because, from the fourth neighbour shell, distances diverge between cubic and hexagonal compact packing, and second, because these distributions of neighbours start to become fairly continuous as distances increase.

Closed-loops size distribution compared to that of topohydrophobic separation

For the remainder of this study, we therefore decided to focus on shorter pair distances, limiting them, unless otherwise specified, to 7 Å between their respective C_αs, i.e. to the first peak of figure 2B centred on 6 Å, as in the case in figure 1B by Berezovsky et al. [11]. Thus figure 3A presents the histogram of all pairs (belonging to either topohydrophobic or non-topohydrophobic positions) at distances smaller than 7 Å as a function of sequence separation, which we will henceforth call 'delta'. Proteins with fewer than 120 amino acids were removed for a sufficiently long separation range in the plots. The data in figure 3A are expressed as a proportion of the total area under the curve, which corresponds to 279,359 pairs of residues from all amino acid groups. This histogram with a cut-off at 7 Å shows a series of regularly distant peaks (near 13 or 14 amino acids between each other), whose mean delta values are indicated by arrows. It is very similar to the equivalent histogram built on another 3D data base (thick solid line) by Berezovsky et al. [11], and therefore indicative that those peaks are not base depen-

dent. Figure 3 B shows the same data separately for topohydrophobic and non-topohydrophobic amino acids. In particular, it shows a large peak centred at a delta of 27 residues, which corresponds to the 22–32 nearly standard loop size previously discussed by Berezowsky et al. (fig. 3 A) [11] and that we now call TEFs. The highest maximum also occurs with our data set for a delta centred around 26–27. Topohydrophobic amino acids are clearly more concerned than non-topohydrophobic ones by preferred regular occurrences along delta at a mean separation of 13–14 amino acids. In figure 3 B there are 7,159 topohydrophobic pairs with a cut-off at 7 Å and 55,413 without. The restricted curve is therefore not as statistically significant as without the cut-off, although the six maxima of figure 3 B are present in both curves, suggesting that the cut-off has no effect on the distribution of topohydrophobic peaks. Clearly, the preferred separation (in amino acids) between topohydrophobic positions (fig. 3 B) matches the three first preferred differences of 'closed loops' end positions shown in figure 3 A, i.e. a mean separation between maxima of approximately 13–14 amino acids. This result may be important regarding the remarkable properties of topohydrophobic positions as key markers of folds and folding [3–7]. The non-uniform character of the distribution of topohydrophobic amino acids constituting the core of proteins, i.e. residues at short distances one from another (cut-off curve), is conserved in the distribution of topohydrophobic residues along the sequence (no cut-off curve). An important result of this is the occurrence of a large peak in the range 22–32 when the allowed pair distances are limited to small values. It reflects the appearance of the first neighbour layer among topohydrophobic residues, which statistically favours this particular range, even if it also detects residues distant along the sequence.

TEFs and regular secondary structure elements

We next investigated the number of secondary structure elements (SSEs) included in TEFs (i.e. the part of the closed loops which are included in the range 22–32) and observed that the most frequent situation occurs for pairs of residues separated by one to two SSEs. If the sequence separation between the two residues of each pair is restricted to the range 22–32, thus corresponding to the large peak of figure 3 A, then most of these pairs are separated by two SSEs. Restriction of this histogram to the pairs of residues separated by 22–32 residues without any distance cut-off leads to the same position of the maximum (data not shown).

Direct comparison between TEF ends and topohydrophobic positions

It should be noted that the closed loops in figure 3 could have various amino acid sizes, but only the preferred-size (22–32) TEFs are used to map globular domains.

In the majority of cases, TEFs approximately start or end at a position occupied by one or several topohydrophobic residues (e.g. fig. 1) as illustrated for a sample of 15 proteins of various folds (fig. 4). Amino acids included in TEFs are numbered relative to the TEF central position, if the TEF contains an odd number of residues, or to the upstream one for an even number. In the ordinate axis the histogram of topohydrophobic positions is represented (continuous line) as a percentage of the topohydrophobic residues involved in TEFs in the data set (topohydrophobic residues in regions where TEFs cannot be defined are excluded from fig. 4). For comparison the histogram of TEF ends (grey line) is also shown. To accept a certain tolerance (that observed on the main '27' peak in fig. 3 A), TEFs are allowed to slide by 3 residues in each direction, their length remaining, nevertheless, constant (dash-dotted line). For example a TEF of 26 amino acids will participate in seven positions: [–16, +9], [–15, +10], [–14, +11], [–13, +12], [–12, +13], –11, +14], [–10, +15]. This is in agreement with Berezovsky and Trifonov's study that admitted a 'window' of 7 residues which more or less corresponds to 90 % of the peak centred at position 27 in figure 3 A (unpublished data). Negative and positive values correspond to the Nter and Cter side of the TEFs, respectively. As the major contribution to TEFs is the one between 22–32 amino acids, the larger peaks occur for this length, i.e. at positions centred around [–13, +12]. These maxima match fairly well the maxima of the topohydrophobic positions. This is the first statistical evidence of a common distribution of TEF ends and topohydrophobic positions. Although our data set is limited, this is an optimistic indication of a statistical correspondence between the ends of TEFs or 'closed loops' and the location of many topohydrophobic residues. Indeed, among the 15 folds of the test set, 13 exhibit low probabilities that TEF extremities (± 3) randomly match topohydrophobic positions, taking into account all these positions, as calculated by the hypergeometrical law; between 10^{-3} and 1.2×10^{-1} (mean 2.4×10^{-2}). The two extra cases concern 1 gky and 4 blm. The probability of a random match between TEF ends [–3, +3] and topohydrophobic positions for 1gky falls from 0.4 to 5×10^{-3} if the 125–147 TEF is extended to 120–166, with no variation in the C_α-C_α distance. For 4 blm, the quasi-random value of 0.8 is due to a special fold which includes, among others, an internal helix where many positions are topohydrophobic but which are unable to be included in a TEF.

Discussion

This study confirmed our previous findings concerning the folding of proteins by fragments, with a clear advantage for fragment lengths in the range of 22–32 amino

acids. Here, we show that topohydrophobic residues, which actually constitute the deep core of proteins, are not regularly distributed along the sequence, but rather display some sequence periodicity compatible with the 22–32 length of TEFs. This is also in good agreement with the fact that TEFs have mainly hydrophobic ends that are located at the centre of globular proteins (data not published). This argues in favour of cooperative folding of proteins [16], with topohydrophobic positions acting as anchors for protein folding [17]. This can be related to the numerous studies on the formation of β hairpins, either by lattice simulation [18], molecular dynamics [19] or Monte-Carlo simulation [20], demonstrating that an important stage of β sheet nucleation concerns the contacts between residues in hydrophobic clusters. Our data suggest that topohydrophobic residues, and particularly the first neighbour shells that they constitute in the protein core, play a key role in the formation of these hydrophobic clusters.

Furthermore, figure 1 suggests that extra TEFs do exist, linking the N- and C-terminal parts of protein domains. For example, the dimeric uteroglobin protein (PDB code 1utg) shows a unique TEF of 30 amino acids between L13 and K42 (K42 is a very conserved and important marker of this fold [10]). The structure reveals a putative second one of 35 amino acids between Q50 and L14 through an artificial link between the C and N termini, which are often very close in the 3 D space within globular domains (here 6 Å between their C_αs). The first TEF (L13–K42) is built by helices 2 and 3 (yellow and blue), the 'extended' one bringing together helices 4 and 1 (red and orange). Such a situation is expected to occur in many other proteins [21] (e.g. 5tim and 1pk4 in fig. 1).

In figure 1, topohydrophobic positions which are located at or near the extremities of TEFs are coloured purple and constitute, in the sample of 15 folds considered here (fig. 4), nearly half the total number of topohydrophobic positions. Therefore, as for Russian dolls, we can consider three 'concentric' populations of hydrophobic amino acids: the classical hydrophobic core (roughly one-third of a globular domain), the topohydrophobic positions (around one-third to one-half of the hydrophobic core domain, depending on each domain) and, finally, the topohydrophobic residues at the extremities of the TEFs as discussed above (near 50% of the preceding population).

The mean separation between peaks observed in figure 3 (13.5 amino acids) is likely the result of the mean length values for α helices, β strand, and coil regions (classical loops) observed for protein globular domains. Indeed, for an observed rough repartition of amino acids, about half are found in loops, and one-quarter in both helices and strands. As a mean TEF is built by two SSEs and their connecting loop, the mean length of a TEF is expected to be of the order of the sum of the mean length of its constituting elements: 12 amino acids for an α helix, 7 amino acids for a β strand and 7 amino acids for a loop in the present data. In the majority of cases, TEFs are composed of two helices or two strands connected by a loop, which implies that a TEF length is included between 21 and 31 amino acids in very good agreement with the peak in the range 22–32 (fig. 3A).

Finally, as topohydrophobic positions can be directly identified from a sufficient set of divergent sequences of a protein family (typically five or six members exhibiting mean pairwise sequence identity between 10 and 30%) [3, 4, 9, 10], particularly through the hydrophobic cluster analysis approach [22], the 22–32-amino acid preferred length of TEFs may then be used to roughly sort topohydrophobic positions and identify some amino acids of the inner circle of the considered fold.

These additional constraints derived from an accessible punctuation of sequences [23] may be a new efficient tool to further characterise distant members of a family (post-genomic analysis), as well as constituting a valuable determinant for the ab initio prediction of protein folds, based on sequence data and distance constraint methods [24–25].

Acknowledgements. The CNRS, through the Genome Project, is acknowledged for its financial support. I.N.B. is a postdoctoral fellow of the Feinberg Graduate School at the Weizmann Institute of Science. Discussions and comments by E.N. Trifonov are appreciated.

1 Mirny L. A., Shakhnovich E. I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol. **291**: 177–196
2 Ptitsyn O. B., Ting K. L. (1999) Non-functional conserved residues in globins and their possible role as a folding nucleus. J. Mol. Biol. **291**: 671–682
3. Poupon A., Mornon J. P. (1998) Populations of hydrophobic amino acids within protein globular domains; identification of conserved 'topohydrophobic' positions. Proteins **33**: 329–342
4 Poupon A., Mornon J. P. (1999) 'Topohydrophobic positions' as key markers of globular protein folds. Theor. Chem. Acc. **101**: 2–8
5 Poupon A., Mornon J. P. (1999) Predicting the protein folding nucleus from sequences. FEBS Lett. **452**: 283–289
6 Poupon A., Mornon J. P. (in press) Deciphering globular proteins sequence/structure relationships: from observation to prediction. Theor. Chem. Acc.
7 Kragelund B. B., Osmark P., Neergaard T. B., Schiod J., Kristiansen K., Knudsen J., et al. (1999) The formation of a native like structure containing eight conserved hydrophobic residues is rate limiting in two states protein folding of ACBP. Nat. Struct. Biol. **6**: 594–601
8 Lorch M., Mason J. M., Clarke A. R., Parker M. J. (1999) Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. Biochemistry **38**: 1377–1385
9 Colloc'h N., Poupon A., Mornon J. P. (2000) Sequence and structural features of the T-fold, an original tunnelling building unit. Proteins **39**: 142–154
10 Callebaut I., Poupon A., Bally R., Demaret J. P., Housset D., Delettré J. et al. (2000) The uteroglobin fold. Ann. N. Y. Acad. Sci. **923**: 90–112

11 Berezovsky I. N., Grosberg A. Y., Trifonov E. N. (2000) Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett. **466**: 283–286
12 Dunbrack R. L., Jr, Gerloff D. L., Bower M., Chen X., Lichtarge O., Cohen F. E. (1996) Meeting review: the Second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, 13–16 December. Fold Des **2**: R27–42
13 Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al. (2000). The Protein Data Bank. Nucleic Acids Res. **28**: 235–242
14 Zimmer R., Wöhler M., Thiele R. (1998) New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics **14**: 295–308
15 Soyer A., Chomilier J., Mornon J.-P., Jullien R., Sadoc J.-F. (2000) Voronoï tesselation reveals that folded proteins belong to condensed matter. Phys. Rev. Lett **85**: 3532–3535
16 Hao M. H., Scheraga H. A. (1998) Molecular mechanisms for cooperative folding of proteins. J. Mol. Biol. **277**: 973–983.
17 Karplus M., Weaver D. L. (1996) Protein folding dynamics: the diffusion collision model and experimental data. Prot. Sci. **3**: 650–668
18 Klimov D. K., Thirumalai D. (1998) Lattice models for proteins reveal multiple folding nuclei for nucleation collapse mechanism. J. Mol. Biol. **282**: 471–492.
19 Pande V. S., Roskhsar D. S. (1999) Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. Proc. Natl. Acad. Sci. USA **96**: 9062–9067
20 Dinner AR, Lazaridis T, Karplus M. (1999) Understanding β-hairpin formation. Proc. Natl. Acad. Sci. USA **96**: 9068–9073
21 Martinez J. C., Viguera A. R., Berisio R., Wilmanns M., Mateo P. L., Filimonov V. V. et al. (1999) Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. Biochemistry **38**: 549–559
22 Callebaut I., Labesse G., Durand P., Poupon A., Canard L., Chomilier J. et al. (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell. Mol. Life Sci. **53**: 621–645
23 Brocchieri L, Karlin S. (1995) How are close residues of protein structures distributed in primary sequence? Proc. Natl. Acad. Sci. USA **92**: 12136–12140
24 Havel T. F. (1991) An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Prog. Biophys. Mol. Biol. **56**: 43–78
25 Havel T. F., Snow M. E. (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. J. Mol. Biol. **217**: 1–7