Hydrophobic compaction, curvature of space and deciphering protein sequences*

Jean-Paul Mornon, Systèmes Moléculaires & Biologie Structurale
Laboratoire de Minéralogie-Cristallographie de Paris, Universités Paris 6 & Paris 7, Paris, France

In the July 2000 issue of the *Bulletin de la SFP* [1], Yves-Henri Sanejouand and Georges Trinquier presented an overview of the surprising ability of simple topological models (cubic nets) to explain the fundamental three-dimensional (3D) folding properties of the polymers that we call proteins, those essential components of life. Here, through a number of different examples, I will illustrate another equally surprising aspect of another apparent "simplicity", that relating to hydrophobic compaction, which governs the folding of these macromolecules and which directly serves as a useful tool to decipher genes, thereby opening new prospects in this "post-genomic" era.

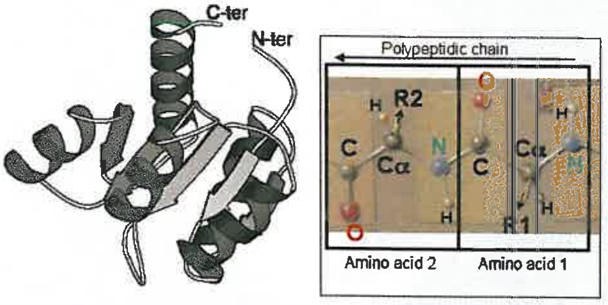
A protein is a linear, unbranched polymer consisting of anywhere from a few dozen to a few thousand links. Nature has limited the chemical diversity of proteins, with the occasional exception, to twenty different types (the twenty common amino acids). All amino acids share the same backbone, differing in terms of their side chains (Fig. 1). Seven amino acids have an aliphatic or aromatic side chain, making them strongly hydrophobic: V (valine), I (isoleucine), L (leucine), F (phenylalanine), M (methionine), Y (tyrosine) and W (tryptophane). Six have a strongly hydrophilic side chain: D (aspartic acid), E (glutamic acid), N (asparagine), Q (glutamine), K (lysine) and R (arginine), while the other seven have intermediate properties: A (alanine), C (cysteine), T (threonine), G (glycine), P (proline), S (serine), and H (histidine). This distribution of hydrophobicity/hydrophilicity offers a clever range of blocks with which to build macromolecules exhibiting remarkable physicochemical properties.

Under normal conditions, any fairly long polypeptide (from a few dozen to a few hundred amino acids) folds spontaneously in the presence of water into globular domains with a stable three-dimensional architecture; some can also fold specifically (often in helical form) within lipid membranes. It is the dichotomy between hydrophobicity/hydrophilicity that acts as the driving force for these processes (e.g. [2]), as it does for many other physicochemical situations in the world around us.

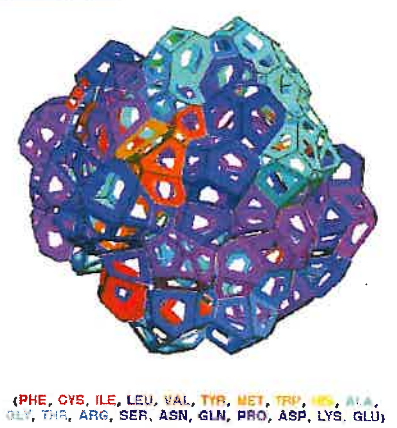
The succession of different types of amino acids along the polymer, which is specific to each protein, is called the primary structure, or sequence. This information is sufficient for the polypeptide chain to adopt a stable and unique three-dimensional structure in a suitable medium (mainly water), with the occasional exception (Fig. 1). Yet not all the positions of a given sequence have the same influence on the cooperative process of folding. As a result, ancestral proteins have undergone considerable modification during evolution, amino acid after amino acid, without altering their resulting three-dimensional structure nor affecting their associated biological function(s). Thus, it is not infrequent that proteins with only a very low number of chemically conserved homologous positions along the polypeptide chain (cf. Fig. 4A), i.e. a very low level of sequence identity (10 % for example), are in fact close cousins within a same structural or functional family.

The result is that, while the number of natural sequences in the biosphere is undoubtedly very large (perhaps more than 10^{12}), the number of distinct 3D folds found in nature is much smaller. Recent estimates [3] put the number at approximately 10^3 and, interestingly, about half of these (≅ 600) have already been precisely characterized at the atomic level through X-ray crystallography or NMR studies over the past four decades (the first x-ray structure of a protein dates back to 1958). Thus, a protein domain encoded on a gene currently has a roughly one in two chance of being closely related in structure (and often in function) to a previously characterized protein fold. This favorable situation can only get better in the future, considering the ever growing number of known protein folds and the even faster rise in the number of whole genomes being sequenced.

The major obstacle to uncovering the underlying correlations in this expanding bulk of data is to recognize the relatedness between evolutionarily divergent proteins sharing low sequence identity, solely on the basis of the available data, i.e. the sequences themselves. We shall see below how topological considerations alone, which directly reflect the physicochemical principles governing polypeptide stability, can meet this challenge by going one step beyond the classical methods of statistical and lexical sequence analysis, which rapidly lose their sensitivity as sequences



▲ Fig. 1: Simplified "ribbon" illustration of an experimentally determined protein structure, depicting the regular secondary structures [α helices (helices) and β strands (arrows)]. The positions of the two ends, N-ter (for NH2-terminal) and C-ter (for COOH-terminal) of the polypeptide chain (128 amino acids – PDB 3CHY) are indicated and, in this case, are close to one another, as frequently observed. The insert shows a peptide segment (from [15]) composed of two links, two amino acids whose backbone (main chain) are linked through a planar peptide bond (CO-NH). The side chains (R1, R2) are carried by the alpha carbons (Cα) of the main chain and vary according to amino acid (20 different side chains; e.g. the side chain of methionine (hydrophobic) has the structure –CH2-CH2-S-CH3). The NH and CO groups of the backbone form hydrogen bonds with the CO and NH groups, respectively, of distant amino acids in the sequence to form regular secondary structures such as a helices and β sheets (assemblies of β strands).


{PHE, CYS, ILE, LEU, VAL, TYR, MET, TRP, HIS, ALA,
GLY, THR, ARG, SER, ASN, GLN, PRO, ASP, LYS, GLU}

▲ Fig. 2: Voronoi tesselation of the same protein Che Y (PDB: 3CHY shown in Fig. 1). The colors of the cells correspond to the hydrophilic/hydrophobic properties of the amino acids. Black and white lines respectively depict faces specific to a single cell and those common to several cells. Only the surface of the protein is shown (source: B. Angelov and J.F. Sadoc).

become more divergent. But first, let us take a look at this topological environment.

Voronoi tesselation of common protein globular domains

Consider the geometrical barycenters of each amino acid and, starting from these few dozen to few hundred points, let us build contiguous polyhedral cells for which any point inside their volume is closer to their associated barycenter than to any other barycenter. These are Voronoi cells, characterized in particular by their mean number of faces and edges per face (Fig. 2). In a representative sample of proteins, these values are in almost perfect agreement with what would be expected for a compact packing of spheres [4]: the protein domains behave like condensed matter, at least in this regard. The advantage of this cell-building method is that it gives "absolute" access to the concept of neighbors, i.e. among others and without arbitrary numerical cutoffs, to concepts as important as, for instance, the mean burying of such and such a category of amino acids, by direct counting of the Voronoi cells at the core and at the surface. This concept, widely employed in condensed matter physics, opens new perspectives into the world of proteins.

Proteins as flowers

Compilation of the geometrical distances between peptide links along a protein sequence reveals that segments separated by some 20 to 30 amino acids are sometimes in close proximity to one another [5]. This heretofore unnoticed property is also revealed in the physics of certain nonbiological polymers [6] (such loops are formed of segments of chains which are about 4 persistence lengths). Thus, like the edge of a daisy petal which loops around before returning to the center of the flower, proteins can be thought of as a "disorderly bouquet" occupying the volume of the folded protein globe (Fig. 3). The recent demonstration of these "closed loops" sheds new light on the origin of proteins, suggesting that they arose from the concatenation of primordial peptide fragments having the right length to self-stabilize by pseudo-closure [7].

Two-level micelles

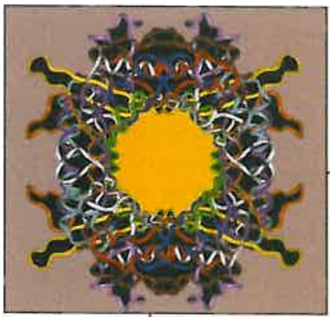
In a typical protein globular domain, about one-third of the amino acid side chains are strongly hydrophobic while the other two-thirds are polar or neutral. The first type clusters within the protein core, insulated from the solvent (water) by a layer of neutral or polar amino acids. However, not all the positions in the sequence (and the resulting 3D structure) occupied by hydrophobic amino acids have the same properties. Two categories may be distinguished: positions at which, for all sequences coding for a same fold, strong hydrophobicity is always conserved ("topohydrophobic positions") [8], and those at which this hydrophobicity is not conserved between sequences ("non-topohydrophobic positions"). The first account for about 1/3 of all positions in a sequence occupied by hydrophobic amino acids (or about 1/9 of the sequence) and display remarkable properties. They are buried deep within the core and therein constitute a network of mutually interacting amino acids, very largely coinciding with the amino acids which, very early during the folding process, are experimentally found to be in contact ("folding nuclei" [9]).

Moreover, it was recently shown that the ends of the closed loops described above are frequently occupied by these topohydrophobic positions [10]. One advantage, and not the least, is that these topohydrophobic positions can be predicted from a limited number of sufficiently divergent sequences coding for a same fold.

Sequences, or the game of chance and time

A protein sequence is only very rarely stable. Over the course of evolution and during long periods of time, certain amino acids might be substituted by others, while others are added and still others deleted. This game of musical chairs (with a variable number of chairs!) lasts for as long as the modifications do not cause unfavorable, or even lethal, effects to the organism, i.e. for as long as they remain compatible with conservation of the fold and its resultant biological function. Yet these geological periods are much too short, by far, to allow this game to explore more than just a tiny fraction of the possibilities. Nevertheless, the result is that a sequence coding for a protein with a given biological function may become unrecognizable: its length may have varied considerably (by addition or deletion of amino acids in any number or position) as well as the identity of the amino acids that make up the polypeptide chain (Fig. 4).

Therefore, comparing sequences and ordering them on the basis of homology into families within genomes is no simple matter. Below a certain level of conserved homologous positions, corresponding to about one quarter of the amino acids, it is very difficult, if not impossible, to succeed by using common automated lexical methods that statistically compare sequences like strings of independent letters. This methodological obstacle is of major importance because it is encountered in a great many genomes, substantially hindering their analysis and use.



▲ Fig. 3: Reduction of the urate oxidase tetramer into closed loops, some of which from this viewpoint are clearly visible at the periphery (in yellow, orange and pink) (source: M. Lamarine). These closed loops should not be confused with the usual loops found in proteins, which are normally shorter (segments of the polypeptide chain joining two successive regular secondary structures – thin segments in Fig. 1).

We shall see below that this obstacle can now be often overcome, but for the moment the price to pay is a loss of automaticity and, consequently, an inability to process massive amounts of genomic data.

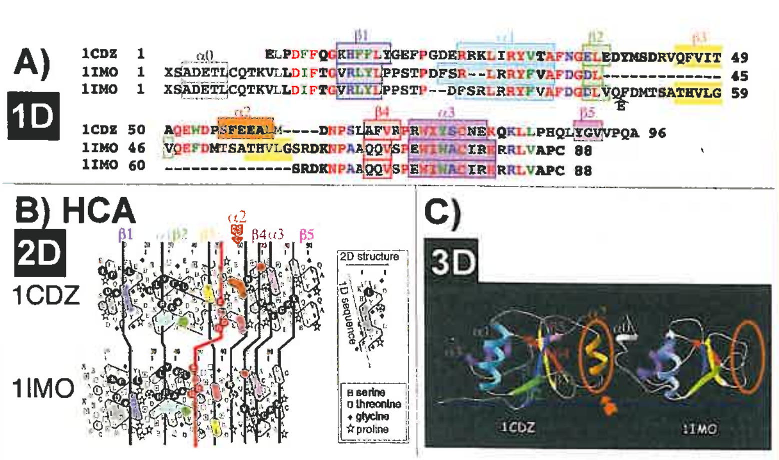
Sequences are more than just strings of letters, or how giving them curvature can help decipher them

A natural polypeptide chain folds upon itself to give the compacted polymer its biological property or properties (Fig. 1). In principle, to "mimic" this compaction, one must obviously and necessarily curve the polypeptide chain which, in the symbolic representation of its sequence, i.e. a string of letters, is considered to be fully extended. This string might be thought of as a straight line (1D). In 2D space, the ideal, uniform curvature of this line is the circle (in which the two ends of the chain – canonically called N and C – touch). Transposition into ordinary 3D space turns this isotropic curve, by continuity, into a helix. Starting from these intuitive principles, it can be shown that with a suitable pitch, the sequence so curved into a uniform helix gives direct information about the local hydrophobic compactions associated with the typical regular secondary structures of polypeptide chains (a helices and β strands) (Fig. 1). Hydrophobic Cluster Analysis (HCA) is based on such an approach [11]. Fig. 4B illustrates in two-dimensional form the result of helical curvature of a sequence, supporting the analysis and the HCA sequence comparison.

This strategy thus makes use of this curvature of space of sequences and also takes into account a suitable separator of information [the "connectivity distance", equal to the number of amino acids, for a helix of given pitch, separating amino acid i from its farthest near neighbor j (e.g. 4 for the common α helix of proteins)]. For the usual sequence, this connectivity distance is simply equal to 1 (amino acid i and its neighbors $i \pm 1$). Using these ingredients, HCA reduces the sequence into protein "words" that make physicochemical sense: they form the internal hydrophobic faces of the regular secondary structures of the polypeptide chain and appear as hydrophobic clusters on the helical representation of the sequence (Fig. 4B). This works in much the same way as blanks and spaces give meaning to a text by separating it into words and sentences: a given string of letters does not have the same meaning in the presence and absence of these punctuation marks. Geometrically speaking, the internal hydrophobic surfaces of these secondary structures are like strips of velcro which, by sticking together, form the hydrophobic core of the domain and constitute the signature of its fold.

It is these "words", these cores of molecular compaction, that allow one to compare and classify highly divergent sequences, because during evolution these clusters have been much more resistant to the shuffling of amino acids than the amino acids considered in isolation. Currently, this can still only be done by dint of a largely multifactorial, context-dependent human expertise comprising both structural and functional data, much like a police investigation. However, more powerful automated methods may one day become feasible, opening a vast field of study.

Proteins are compact, and the first HCA curvature of space of the sequence, which exploits the intrinsic properties of helical compaction (e.g. the Coxeter helix as a continuous assembly of tetrahedra attached by their faces [12]) gives information about local compaction nuclei, which serve as building blocks of global hydrophobic compaction. This makes it tempting to explore the potentials of a higher dimension, i.e. a second curvature in topological space to four dimensions, with the aim of revealing, even in a roundabout way, the global compaction, i.e. the 3D fold itself, based solely on sequence information (the so-called *ab initio* fold, an approach developed in collaboration with Jean François Sadoc (Laboratoire de Physique des Solides ,Orsay)) and which, with the occasional exception, is still outside of our reach, though substantial progress has been made through the use of structural fragment banks (e.g. Rosetta [13]).



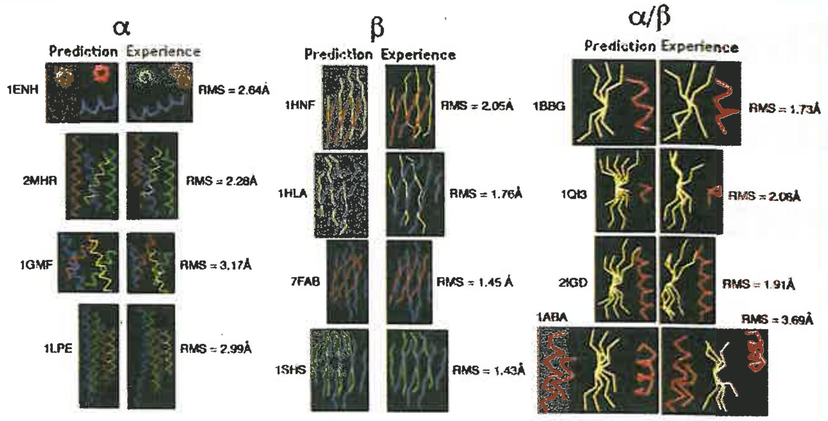
```
1CDZ  1   ELPDFFQGKHFFLYGEFPGDERRKLIRYVTAFNGELEDYMSDRVQFVIT 49
1IMO  1   XSADETLCQTKVLLDIFTGVRLYLPPSTPDFSR--LRRYFVAFDGDL-------------- 45
1IMO  1   XSADETLCQTKVLLDIFTGVRLYLPPSTP--DFSRLRRYFVAFDGDLVQFDMTSATHVLG 59
                                                            E

1CDZ 50   AQEWDPSFEEALM----DNPSLAFVRFRVXYSCNEKQKLLPHQLYGVVPQA 96
1IMO 46   VQEFDMTSATHVLGSRDKNPAAQQVSPEWIWACIRHRRLVAPC 88
1IMO 60   --------------SRDKNPAAQQVSPEWIWACIRHRRLVAPC 88
```

▲ Fig. 4A: Alignment of two protein sequences sharing a same fold and a common ancestor (belonging to the BRCT family, a family of domains found in several proteins associated with DNA repair and cell cycle control, among which the breast cancer- associated BRCA1 protein [16]). The sequences have considerably evolved and share only less than 30 % amino acid identity (red) in homologous folding positions, whereas conservative substitutions of the principal physicochemical properties (green for hydrophobic amino acids, blue for the others) and non-conservative substitutions have occurred at other positions in the alignment, in addition to insertions or deletions of amino acids (hyphens). Secondary structures, as observed experimentally (panel C), are reported in coloured boxes. The alignment shown on the two first lines, based solely on lexical comparison of the two sequences, can be refined by HCA (see panel B) and the characteristic profiles of this family of sequences [16] (first and third lines). The result is a displacement of the insertion. Strands β3 are not aligned correctly by the lexical procedure between the first and second lanes, as well as the begining of helix α1. The cluster of identical (Q, E, D) or similar (W or F) amino acids, depicted in white on a red background on panel B, has wrongly attracted the automatic lexical procedure of alignment (red line of correspondence on the HCA plots – panel B), shown on the first and second lines of this panel.

▲ Fig. 4B: Comparison of the HCA plots of these same two sequences, revealing the conservation of the internal faces of the secondary folding structures (hydrophobic clusters shaded, according to the corresponding structures). The orange arrow shows a large insertion comprising a typical hydrophobic cluster with regular secondary structure. The HCA hydrophobic clusters depicted in this figure represent a broken and transposed 2D image of the local nuclei of hydrophobic compaction within the 3D core of the protein.

▲ Fig. 4C: Comparison of the three-dimensional structures of these two proteins, illustrating the overall conservation of folding and the presence in one of the sequences of a large insertion in a loop, including a "decorative" regular secondary structure (orange arrow).



▶ **Fig. 5:** Representative sample of the RUSSIA procedure to predict the folding of small protein domains, based on a topological search for optimal compaction of hydrophobic side chains. Only regular secondary structures (helices and strands) are shown (long chains). Loops (in the usual meaning of the term) are not depicted. The mean quadratic distances in Å (RMS) between the positions of homologous alpha carbons in actual structures (experience) and the predicted ones are indicated.

RUSSIA

RUSSIA (Rigid Unconnected Secondary Structure Iterative Assembling) [Znamenskiy D., Le Tuan K., Chomilier J. and Mornon J.P., submitted for publication] is a topological procedure to predict the folding of small to intermediate size globular domains based solely on seeking the maximal compaction of hydrophobic amino acids. The starting data consist of the sequence, the nature and the limits (approximate) of secondary structures, a helices and b strands, predicted by multiple sequence alignments. Helices are modelled as rigid cylinders and assemblies of β strands (β sheets) by helical surfaces with a predefined pitch. The simplicity of this description allows an exhaustive study and an efficient sorting of all possible conformations, as long as the global hydrophobic barycentric description of the domain under study can be considered small (ideally a single point) (Fig. 5).

The success of this procedure has encouraged an approach based on both geometry and topology, which has the advantage of implicity taking into account a description and energy minimization that is still directly (formally) out of reach for such large molecular systems, considering the extreme complexity and lack of contrast in energy states.

Proteins never form knots

A glance at the many protein three-dimensional structures that have already been elucidated reveals that the two ends of the polypeptide chain are often in proximity to one another (Fig. 1); statistically this should be true even if the exercise is not straightforward. This observation might perhaps be related to another solidly established fact ([14]: except in rare circumstances, during the normal folding process polypeptide chains never form knots in the usual meaning of the term (a shoelace, for example – topologically speaking, only closed curves can contain irreducible knots). This might explain why, in contrast to the case of DNA for which topoisomerases exist, there appears to be no molecular machinery to prevent formation of potentially disastrous topological situations, such as knots, apart from the chaperone protein complexes which assist in polypeptide chain folding, perhaps in part by "holding" the two ends of the chain together. This question, which has not yet been widely addressed, once again underscores the importance of geometrical and topological considerations in the molecular machinery of the living world.

References

[1] Y. H. Sanejouand, G. Trinquier, Bulletin de la Société de Physique 125, 25 (2000).

[2] S. E. Bresler, D.L. Talmud, Doklady URSS XLIII 7, 310 (1944); S. E. Bresler, D.L. Talmud, Doklady URSS XLIII 8, 349 (1944)

[3] Z. X. Wang, Protein Eng. 11, 621 (1998); C. Zhang, C. DeLisi, J. Mol. Biol. 284, 1301 (1998); S. Govindarajan, R. Recabarren, R.A. Goldstein, Proteins 35, 408 (1999); Y. I. Wolf, N. V. Grishin, E.V. Koonin, J. Mol. Biol. 299, 897 (2000)

[4] A. Soyer, J. Chomilier, J.-P. Mornon, R. Jullien, J.-F. Sadoc, Phys. Rev. Lett. 85, 3532 (2000)

[5] I. N.Berezovsky, A. Y. Trifonov, FEBS Lett. 446, 283 (2000).

[6] C. R. Cantor, P.R. Schismel, Biophysical Chemistry (W.H. Freeman ed) 1980

[7] I. N.Berezovsky, A. Y. Trifonov, Mol. Biol. 35, 233 (2001)

[8] A. Poupon, J.-P. Mornon, Proteins 33, 329 (1998); A. Poupon, J.-P. Mornon, Theor. Chem. Acc. 101, 2 (1999)

[9] A. Poupon, J.-P. Mornon, FEBS Lett. 452, 283 (1999)

[10] M. Lamarine, J.-P. Mornon, I. N. Berezovsky, J. Chomilier, Cell. Mol. Life Sci. 58, 492 (2001)

[11] C. Gaboriaud, V. Bissery, T. Benchetrit, J.-P. Mornon, FEBS Lett. 224, 149 (1987); I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, Mornon J.-P., Cell. Mol. Life Sci. 53, 621 (1997)

[12] J.-F. Sadoc, Eur. Phys. J. E 5, 575 (2001)

[13] K.T. Simons, C. Strauss, D. Baker, J. Mol. Biol. 306, 1191 (2001)

[14] M. L. Mansfield, Nat. Struct. Biol. 1, 213 (1994); M. L. Mansfield, Nat. Struct. Biol. 4, 166 (1997); W. R.Taylor, Nature 406, 916 (2000)

[15] L. Pauling, R. Hayward, The architecture of molecules (W.H. Freeman & cie, Eds) 1964

[16] I. Callebaut, J.-P. Mornon, FEBS Lett. 400, 25 (1997)

* This paper was first published in french in the Bulletin de la Société Française de Physique (March 2002, 133, 4-7)

About the author

Jean-Paul Mornon is Research Director at CNRS. From crystallography, he has progressively moved towards sequence and protein folding analysis and has proposed in 1987 the Hydrophobic Cluster Analysis approach.