

Structural alphabet

Structural alphabet: from a local point of view to a global description of protein 3D structures.

Alexandre G. de Brevern ^{*#}, Cristina Benros [#] & Serge Hazout

Equipe de Bioinformatique Génomique et Moléculaire (EBGM),
INSERM E03-46, Université Denis DIDEROT-Paris 7, case 7113,
2, place Jussieu, 75251 Paris, France

*** Corresponding author:**

mailing address: Dr. de Brevern A.G., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM E 03-46, Université Denis DIDEROT-Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France

E-mail : debrevern@ebgm.jussieu.fr

Tel: (33) 1 44 27 77 31

Fax: (33) 1 43 26 38 30

Running title: structural alphabet

key words: secondary structure, local folds, Bayesian prediction approach, structure-sequence relationship, protein structural classes, *ab initio*.

[#] both authors have contributed equally to this work.

Abstract

The study of protein structures' local conformations has a long history principally based on the analysis of the classical repetitive structures (i.e. α -helix and β -sheet), and also on the characterization of some particular structures in the coil state (e.g. turns). The secondary structures are interesting for describing the global protein fold but miss all the orientations of the connecting regions and so neglect many particularities of the coil state.

In order to take these structural features into account, we have identified a local structural alphabet composed of 16 folding patterns of five consecutive residues, called Protein Blocks (PBs). Conversely to the secondary structures, the PBs are able to approximate every part of the protein structures. These PBs have been used both to describe precisely the 3D protein backbones with an average *rmsd* of 0.42 Å, and to perform a local structure prediction with a rate of correct prediction of 48.7%.

In this chapter, we present the interest of the Protein Blocks by comparing the secondary structure assignment with the assignment in terms of PBs. We highlight the discrepancies between different secondary structure assignment methods and show some interesting correspondence between particular local folds and the Protein Blocks. Then, we use the Protein Block prediction to classify proteins into the classical structural classes, namely all α , all β and mixed. The prediction rate of these different classes is good, i.e. 71.5%, with no confusion between all α and all β classes. Finally, we present a new approach named TopKAPi that stands for "Triangular Kohonen Map for Analyzing Proteins". It enables to classify and analyze proteins

Structural alphabet

according to their Protein Block frequencies using for this purpose a novel unsupervised clustering method: a triangular self-organizing Kohonen map. This method enables to determine new relationships between local structures and amino acid distributions. This new methodology could be of great interest in proteomics and sequence alignment.

Introduction

The first protein structure obtained by X-ray diffraction [1] had marked the beginning of the description and analysis of protein structures. Two ways have since been followed with the theoretical methods and the descriptive methods. In the first years, due to the limited number of available structures, the first kind of approaches was used for proposing potential local structures based on physico-chemical properties (e.g. the γ -helix [2]). Their presence and interest in experimentally determined structures were confirmed or not only in a second step [3]. In this chapter, we focus on the second kind of approaches based on the description and characterization of particular local fold structures observed in experimentally determined protein structures. Figure 1 summarizes different levels of description of the protein folds that we are going to follow in this introduction: (i) the 3-states secondary structures, (ii) the secondary structures with more distinct states, (iii) the structural alphabets and (iv) the description of the complete protein topology.

(i) *the 3-states secondary structures.* One of the major events in the protein history is the series of seven consecutive papers of Pauling and Corey in 1951. They described an impressive number of potential local folds including the α -helix and the β -sheet [2, 4]. The average characteristics of these local structures are described in Table 1. The α -helix (or 3.6_{13} helix) is characterized by intramolecular hydrogen bonds between amino acid residues i and $i + 4$ [5]. Its extremities show specific physicochemical stabilizations [6]. The β -sheet is defined by hydrogen bonds between neighboring parallel (or antiparallel) chains [4]. As for the α -helix, its edges have been widely analyzed [7 - 9]. Many studies have also analyzed the packing of these two repetitive structures [10 - 13] and the sequence - structure relationship between sequence and structure [14 -

17].

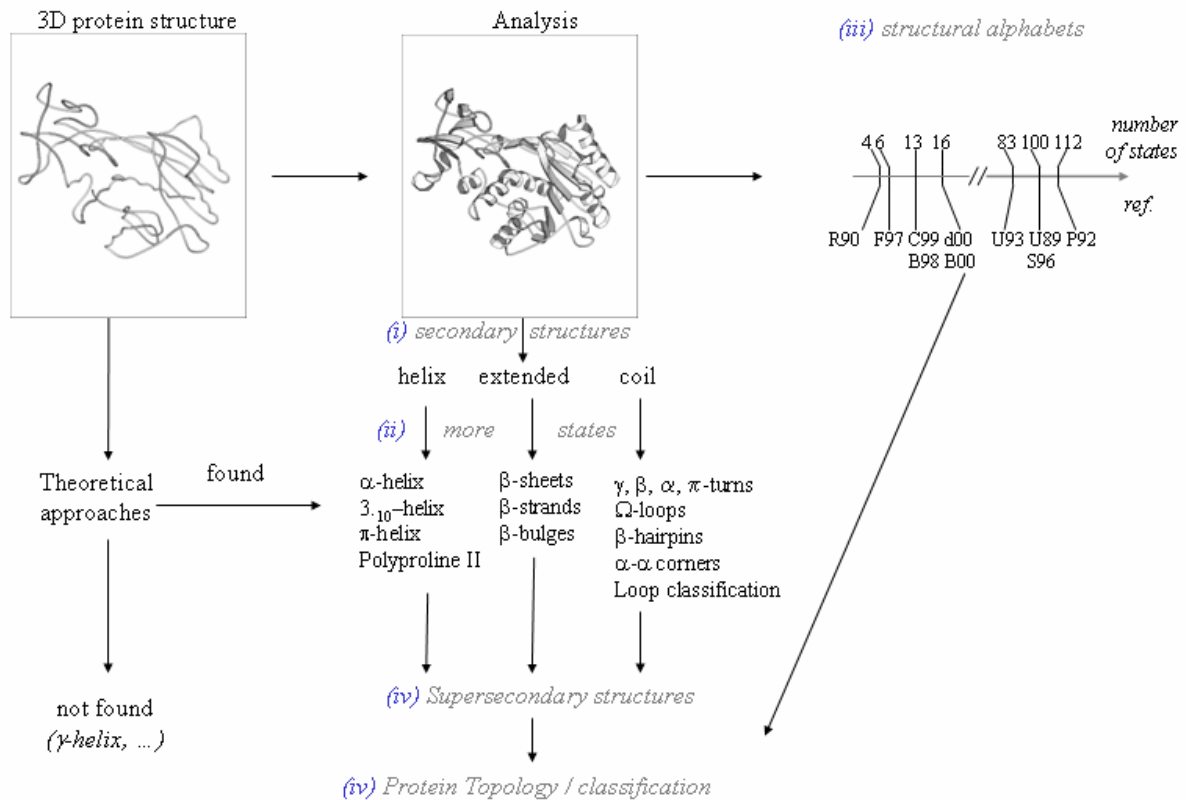


Figure 1. *Protein structure analysis.* From the 3D atom coordinates of a protein structure different analyses are possible. Theoretical approaches consist in predicting potential interesting local structures based on physicochemical criteria. Descriptive approaches are based on the analysis of known local structures at different levels of organization: (i) *the secondary structures* defined by three states (helicoidal, extended and non-helicoidal /non-extended). (ii) a more detailed description (see Text). Another way to describe the protein structures is by (iii) *the use of structural alphabets* that are characterized by different states (references : R90 [125], F97 [126], C99 [129], B98 [127], d00 [128], B00 [161], U93 [120], U89 [118], S96 [121], P92 [119]). The local folds combinations create local topologies referred to as (iv) *super-secondary structures* which describe the *complete protein topology*.

The high propensity of these helicoidal and extended local structures in experimentally determined structures has since achieved one kind of dogma, the ‘secondary structures’ composed of the α -helix, the β -strand and a state corresponding to everything else, the coil. The structures are often limited to this simple description.

Structural alphabet

	ϕ (°)	ψ (°)	ω (°)	nrp	tpr (Å)	pitch (Å)	iHb	atHloop	b. rad. (Å)
α - helix (3.6 ₁₃)	-64 <i>-57.8</i>	-41 <i>-47.0</i>	180	3.6	1.5	5.5	i - i+4	13	2.3
3.10 -helix (3.0 ₁₀)	-71 <i>-74</i>	-18 <i>-4</i>	180	3.0	2.0	6.0	i - i+3	10	1.9
π - helix (4.4 ₆)	-57	-70	180	4.4	1.15	5.0	i - i+5	6	2.8
polyproline II	-75	145	180	3.0	3.12	9.0			
Antiparallel β	-139	135	-178	2.0	3.4	6.8			
Parallel β	-119	113	180	2.0	3.2	6.9			

Table 1. Characterization of classical local folds with their dihedral angles (ϕ , ψ and ω , in *italics* are given the theoretical values), the number of residues per turns (nrp), the translation per residues (tpr), the pitch, the intramolecular hydrogen-bond between (CO, NH) (iHb), the atoms found in the H-bonded loop (atHloop) and the backbone radius.

(ii) *the secondary structures with more distinct states.* Different studies have examined and extended these definitions both in creating new states and refining the assignment criteria. They have improved our knowledge of the repetitive helicoidal and extended structures and have highlighted the interest of the coil state too many times badly described as ‘random’ coil.

The new local folds pointed out exhibit interesting energetic and / or geometrical properties. Thus, less common helices, like 3₁₀ – helices are characterized by intramolecular hydrogen bonds between amino acid residues i and $i + 3$ [18 - 20] and π -helices (i.e. 4.4₆-helices) with hydrogen bonds between amino acid residues i and $i + 5$ [21 - 25]. These two types of helicoidal structures are often encountered at the extremities of longer α -helices and seem to play an important role in the stabilization of longer helicoidal structures [26]. The π -bulges constitute a particular kind of discontinuity in helicoidal structures. Like the π -helices, they are not frequent but seem directly associated to protein functions [23, 27]. The Polyproline II helices correspond to a specific local fold initially found in fibrous proteins [28, 29]. They contribute to the creation of coiled coil supersecondary structures characteristic of these proteins. They are also found in globular proteins and are not composed only of Proline [30 - 33]. Among the predicted helicoidal

Structural alphabet

local folds never observed in proteins we can quote the γ -helix (or 5.11₁₇-helix) [2, 3], the 2.2₇-helix and the 4.3₁₄-helix [18]. As for the Polyproline I, it is only found in apolar solvents.

In the same way, accurate analyses have been carried out for the β -sheet category. An interesting point is that since the description of the β -strands, several analyses have shown that a strand can be found independently of a β -sheet and named the E-strand [34]. Moreover, orthogonal $\beta\beta$ motifs, i.e. consecutive strands, have been identified, forming a 'L' structure with an angle of 90° [13, 35]. Globally, the irregularities within the β -strands have been classified into 4 distinct classes of β -bulges [36, 37] and can be related to the of proteins' function and stability [38].

The regions between the repetitive helicoidal and extended structures have been intensively studied too. Thus, Venkatachalam, using a theoretical approach close to Ramachandran method [39], determined small local folds characterized by the reversing of the polypeptide chain maintained by a hydrogen bond between two close residues, i.e. the turns [40]. After this description, a classification was done and has greatly evolved. The tight turns are characterized by precise dihedral angle values and short distance between their ends [41]. The two most studied turns are the γ - (3 residues) and the β -turns (4 residues). The γ - turns are composed of two categories, *classic* and *inverse* [42 - 45]. The β -turns have a more complex history. At the beginning, the four main categories were the types I, I', II and II' [46, 47]. The extension of the β -turn classification created new categories: the turns III, III', V and V', the turn VI characterized by a Proline, the turn VII associated with a kink and the turn IV corresponding to all the non classified turns [48]. The first analyses of turns in protein structures used this classification [49 - 52]. However, at the beginning of the 80's, different turns have been excluded, the turns III and III' which were too close to the 3.10 helix, the turns V, V' and VII which were too rare and their

definitions inaccurate [53]. Wilmot and Thornton defined the turn VIII which is associated with an important number of observations [54]. It is the first turn not directly associated with a stabilizing bond between its ends. The definitions used by Thornton's group [37, 55] are considered as the standard. Nevertheless, some analyses have been done using the excluded turns V, V' and VII [56, 57]. Shorter turns (e.g. 2 residues δ -turns) [58] and longer ones (e.g. 5 residues α -turns [59, 60] and 6 residues π -turns [61, 62]) have been less studied. The different classes of turns can be overlapping, e.g. two β -turns can have 3 positions in common. The turns can also be multiple at the same position, e.g. a β -turn can encompass a γ -turn [63 - 65]. The turns account for some 25% of the structures.

Other interesting local structures, less frequent than the turns have also been identified in the coil state. For instance, the Ω -loops constitute a particular category characterized by a small distance at their extremities and an important number of contacts in their structure [66, 67]. They correspond to compact globular loops mainly located at the surface of the proteins [68]. They may be directly associated with the protein functions [69, 70] and folding [70].

However, even if the coil state is better characterized, some local folds still remained unassigned. Hence, another approach is developed and consists in classifying the protein fragments between α -helices and / or β -sheets. Different kind of classifications has been carried out. The first type consists in analyzing only specific successions from one state to another. For instance, the study of the connections between two successive β -strands has been studied and has resulted in the classification of the β -hairpins [71 - 74]. Interestingly, the short length hairpins are often characterized by a specific turn [13] and the longer ones by a β -bulge in one of the strands [75]. Sometimes stabilization by disulfide bonds can be observed [76]. The β -hairpins are well studied in molecular dynamics [77, 78]. The same approach was performed for the short loops

Structural alphabet

connecting two α -helices and resulted in the characterization of the α - α corners which are similar to the 'L' structure of orthogonal $\beta\beta$ [79]. Other studies have focused on one precise loop category like α -helix-turn- β -strand [80], or on particular combinations of β -strands like the Ψ -loop [55, 81]. The second type of classification consists in more systematic analyses of the short and medium loops. Many studies have been carried out for short loops connecting α -helices and β -sheets [82 - 84]. Others have also been done systematically for the short [85, 86] and medium loops [87, 88], whatever the flanking regions.

All these methods can only be used for short length loops [89] or for combination of small loops [90] since longer loops are less frequent and considered as too variables. Nevertheless, the different loop classifications have shown their interest in local structure prediction to construct loops in non-complete structures [91 - 95]. Databases of loops useful for molecular modeling have been created [88, 96 - 98].

Even if the repetitive secondary structures have been intensively analyzed [17, 99], the characterization of the α -helices and the β -strands has led to different assignment methods based upon energetic, geometrical and/or angular criteria, which do not always agree particularly at the edges. The first software has been developed by Levitt and Greer and used only the C_α positions as these atoms are the most precisely defined by X-ray crystallography [100]. Table 2 summarizes the different methods analyzed in this study in this research with the number and type of states they focus on. DSSP [101] is the most popular method. Moreover, it is the basis of the secondary structure assignment given by the Protein DataBank [102, 103]. It is based on the hydrogen bonding patterns.

Structural alphabet

methods	helicoidal state	strand state	coil	states
DSSP	α -helix ('H')	β -strand ('b')	turn ('T')	8
	3_{10} -helix ('G')	β -sheet ('E')	bend ('N')	
	π -helix ('I')		coil	
STRIDE	α -helix ('H')	β -strand ('b')	turns ('T')	7
	3_{10} -helix ('G')	β -sheet ('E')	coil	
	π -helix ('I')			
PSEA	α -helix	β -strand	coil	3
DEFINE	α -helix	β -strand	coil	3
PCURVE	α -helix	β -strand	coil	3
XTLSSTR	α -helix	β -strand	h-bonded turn ('T')	7
	3_{10} -helix		unh-bonded turn ('N')	
			polyproline II ('P')	
SECSTR	α -helix	β -strand	coil	5
	3_{10} -helix			
	π -helix			
HELANAL	α -helix [5]	/	/	5
EXTENDED-BETA	/	β -sheet [5]	/	6
PROMOTIF	α -helix	β -strand	γ -turn [2]	25
		β -strand	β -turn [10]	
		β -bulge [10]	β -hairpins	
SUMMARY	α -helix [5]	β -sheet [6]	γ -turn [2]	
	3_{10} -helix	β -strand	β -turn [10]	
	π -helix	β -bulge [10]	β -hairpins	
			polyproline II	
TOTAL	7	17	14	38

Table 2. Secondary structure assignment methods with the number of states for the helicoidal states, the extended states and the non-repetitive states. In brackets are given the number of states corresponding to one specific category and in parenthesis is given the one letter code corresponding to the state.

STRIDE [104] uses the same criteria with parameters slightly different and the computation of backbone dihedral angles. SECSTR focuses on the correct assignment of 3_{10} – and π -helices [24]. Recently, DSSPcont tries to optimize the parameters of DSSP by taking into account

Structural alphabet

multiple NMR models assignment and tries to compensate at best the fluctuations of the assignment between the different model observations [105 - 107]. However, the results are not really improved. DEFINE [108] like Levitt and Greer method, uses only the C_α positions. It computes inter- C_α distance matrices and compares the results to ideal repetitive secondary structures. PCURVE [109] is based on the helicoidal parameters of each peptide unit and generates a global peptide axis. PSEA [110] assigns the repetitive secondary structures from the sole C_α position using distance and angles criteria. XTLSSTR uses all the backbone atoms to compute two angles and three distances [111]. It is especially dedicated to the spectroscopists. PROMOTIF uses an implementation similar to DSSP but focuses on the characterization of γ - and β -turns, β -hairpins and β -bulges [55].

The assignment methods may generate particular problems. Hence, DSSP may assign very long helices which do not correspond to reality [112]. Bansal and co-workers have analyzed and classified the helices and showed that important part of them are in fact curved or composed of distinct helices [113]. In the same way, Woodcock and co-workers [114] noted that these methods do not assign the same state to certain residues, especially those located at the beginnings and ends of repetitive structures, *i. e.* the secondary structure assignments differ according to the chosen method. This observation has led to the development of a consensus approach [115] which represents an average measure of DSSP, DEFINE and PCURVE. This study has shown that less than 2/3 of the residues are associated to the same state by these three algorithms. The use of one or another method does not reflect the same type of reality. For instance, the α -helix defined by DSSP, with its eight states grouped in only three states, does not correspond only to the α -helix (3.16₁₃ helix), but incorporates the 3.10 helix and the π -helix (4.4₆-helix). In the same way, β -sheets (DSSP 'E' state) correspond to β -strands implicated in parallel

Structural alphabet

or anti-parallel characteristic patterns but not β -strands without hydrogen bond partner (DSSP 'B' state). These features may induce difficulties in analyzing the protein structures or dynamic trajectories. So, it is important to note that the repetitive structures definitions only reflect a given classification.

(iii) *the structural alphabet*. Various teams have tried to proceed without using classic secondary structure descriptions. Instead, they categorize the 3D structures without any *a priori*. Thus, every local fold is associated to one specific small prototype. The complete set of prototypes defines "a structural alphabet" [116, 117]. Numerous structural alphabets have been defined and differed by the description parameters of the protein backbone (C_α coordinates, C_α distances, α or dihedral angles) and by the method used for defining them (hierarchical clustering, empirical function, Kohonen Maps, neural network or Hidden Markov Model) [116]. Each structural alphabet or fragment library is defined as a series of N prototypes of l residues length. N is highly variable (between 4 and 123), l only varies between 4 and 7.

Two main types of research must be distinguished. The first one consists in describing an important number of prototypes to reconstruct precisely a protein structure. The second one aims at predicting the 3D structures from the sole knowledge of the sequence and so is limited to few prototypes.

Hence, the earliest works used hundred of prototypes ($N = 83$ to 120) to reconstruct protein structures [118 - 121]. Levitt's group [122, 123] and Micheletti and co-workers [124] tried to optimize the construction of such libraries from geometrical point of view. This structural description allows new insight into protein 3D structures and reveals peculiar sequence specificity [116].

Structural alphabet

However, to perform a prediction from the sequence, the number of prototypes, N , must be smaller, i.e. a correct prediction implies the selection of a more limited number of local conformations as shown by Rooman's and Fetrow's works [125, 126]. Indeed to capture most of the local folds, it is advisable to have a balance between a number of states sufficient for approximating correctly the local folds and limited for ensuring a correct prediction level. An alphabet composed of $N = 10$ to 20 states corresponds to this goal [127, 128]. These methods have proved their efficiency both in the description and the prediction of small loops [129, 130] or long fragments [131 - 137]. Byströff and Baker's I-Sites must be pointed out as one of the most interesting structural alphabet. It has been used with a high efficiency for improving *new fold* methods [138, 139].

The different alphabets have in common to describe more precisely the repetitive structures (helical and extended) and their edges, and to focus on a better description of the coil state.

In this work, we use the structural alphabet we have defined in a previous study. It is composed of 16 mean protein fragments of 5 residues length called Protein Blocks. These PBs have been used both to describe the 3D protein backbones and to perform a local structure prediction [128, 133]. A comparison between different structural alphabets has shown its informativity [140].

(iv) *the local structures describe the protein topology.* The succession of secondary structures defines the supersecondary structures, i.e. $\alpha\beta$, $\beta\alpha$, $\beta\beta$ and $\alpha\alpha$. Their combinations generate some particular motifs like the greek key ($\beta\beta\beta\beta$) or the Rosmann fold ($\beta\alpha\beta\alpha\beta$) [141]. They can be used to define the complete topology of the proteins like in the TOPS family

Structural alphabet

database [142]. Thus, they are used to classify proteins into different structural families like that of SCOP [143] or CATH [144], even if a recent study has shown the difficulties to find a good consensus between all these classifications [145]. Most of these classifications give few major families and then a important number of sub-families. Nevertheless, these descriptions have proved their efficiency to find distant structural homologues [146] or to work with genomic data [147]. They are the classic benchmark of fold recognition [140].

Here, we examine the relationship between the structural alphabet and the 3D structure topology.

The results of our study are divided into three consecutive parts. (i) As proteins are classically described by their content in secondary structures, we looked at the correspondence between the different secondary structure states and our Protein Blocks. We highlighted the differences between many secondary structure assignment methods. (ii) We have previously described a Bayesian approach to predict the Protein Blocks from the sequence [128]. It has been improved and gives now a prediction rate of 48.7%. In the second part of this chapter, we analyzed the results of the prediction and their use to protein classification according to their structural classes. These classes are defined as all- α , all- β and mixed ($\alpha+\beta$ and α/β) following the definitions of Michie and co-workers [148]. (iii) Finally, we described a new method called TopKAPi, for Triangular Kohonen map for Analyzing Proteins. Firstly, it allows classifying and analyzing protein structures based only on the Protein Blocks frequencies. Then, it permits to analyze the amino acid distributions associated with this new classification. We show new insights into the sequence and structure relationships.

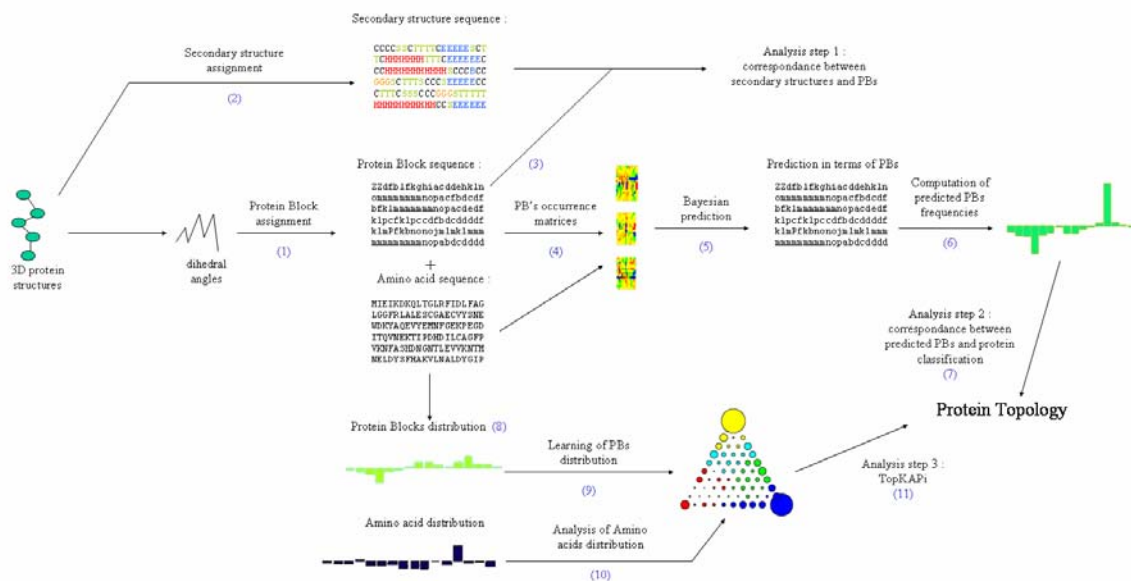
Materials and Methods:

Figure 2. *Main principles of this study.* The first step consisted to extract from a non redundant databank the dihedral angles of the protein structures. (1) These last allowed to encode the 3D structures in terms of Protein Blocks. (2) Using different assignment algorithms the 3D structures were encoded also in terms of secondary structures. (3) We first analyzed the agreement between secondary structures and PBs. (4) Then, using the amino acid sequences, we computed amino acid occurrence matrices associated to each PB. (5) We used this information to perform Bayesian prediction and from the prediction in terms of PBs, (6) we computed the predicted frequencies of PBs per protein. (7) We analyzed the correspondence between the predicted frequencies of PBs and the class of the proteins (all α , all β or mixed). (8) In parallel, from the true PBs and the amino acids of the proteins, we computed the frequencies of PBs and amino acids per protein. (9) We learnt these frequencies of PBs per protein using an adapted triangular Self-Organizing Maps, named TopKaPi. Then, we analyzed (10) the amino acid distributions in TopKaPi and (11) the different clusters of PBs.

Figure 2 gives the main steps of the research carried out in this chapter and described here.

Data sets: Different sets of proteins were used in this work. The four first ones have already been used in a recent work [133]: *PAPIA* from PDB-REPRDB database [149], *PDBselect*

Structural alphabet

databank [150, 151], *culled-Pdb* (now *PSICES*) [152], *SCOP - ASTRAL* [143, 153]. We have preferentially used the *PAPIA* set which is composed of 717 protein chains and 180,854 residues. The set contained proteins with no more than 30% pairwise sequence identity. The selected chains had X-ray crystallographic resolutions less than 2.0 Å and an R-factor less than 0.2. Each selected structure had *RMSD* value more than 10 Å with every representative chain. A new updated data set (noted *PAPIA03*) has been composed from PDB-REPRDB database [149] with the same criteria as *PAPIA* and is composed of 1,407 protein chains and 293,507 residues. We have verified that the amino acid compositions were not significantly modified between the two protein sets. Each chain was carefully examined with geometric criteria to avoid bias from zones with missing density.

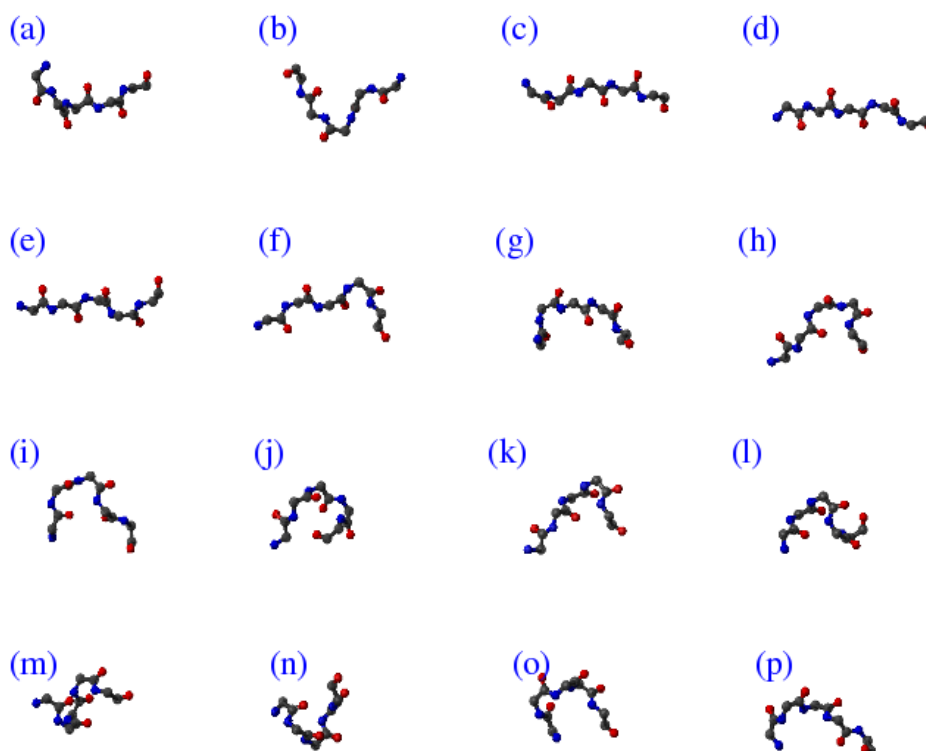


Figure 3. Backbone representation of the 16 Protein Blocks with MOLMOL software [162]. PB *a* to PB *p* are displayed from left to right and from top to bottom.

Protein Blocks (PBs): The structural alphabet we defined in a previous study [128] is composed of 16 local prototypes called “Protein Blocks” (PBs). They are overlapping fragments of 5 residues in length, encoded as sequence windows of 8 consecutive (ψ , ϕ) pairs. They were obtained by an unsupervised classifier similar to Kohonen Maps [154, 155] and Hidden Markov Models [156]. Figure 3 gives a representation of the 16 Protein Blocks. The PBs m and d correspond to the prototypes for the central α -helix and the central β -strand, respectively. PBs a through c primarily represents β -strand N-caps and e and f , C-caps. PBs g through j are specific to coils, PBs k and l to α -helix N-caps, and n through p to their C-caps. This structural alphabet allows a reasonable approximation of the protein 3D-structures with a *RMSD* now evaluated at 0.42 Å. This value has been assessed again in this study with the new databanks.

Protein coding: Protein structures are encoded as sequences of ϕ - ψ dihedral angles, so that a protein of M amino acids long is defined by a signal of $2(M-1)$ dihedral angular values. Each fragment of M residues ($M=5$) centered at the α -carbon $C\alpha_n$ is represented by a vector of 8 dihedral angles (Ψ_{n-2} , Φ_{n-1} , Ψ_{n-1} , Φ_n , Ψ_n , Φ_{n+1} , Ψ_{n+1} , and Φ_{n+2}). The fragment is compared to each PB with the *RMSDa* measure [121], i.e. Euclidean distance using angle values. The lowest *RMSDa* value for the $2(M-1)$ angles determines the assignment of the PB (Fig2, arrow 1).

Secondary structure assignments: They have been done with ten distinct softwares. The seven first ones correspond to DSSP [101] (CMBI version 2000), DEFINE [108] (version 2.0), PCURVE [109] (version 3.1), STRIDE [104], PSEA [110] (version 2.0), XTLSSTR [111], SECSTR [24] (Table 2). Default parameters were used for each softwares. Three more programs

Structural alphabet

have been used: HELANAL [113] to analyze the α -helices, EXTENDED-BETA, which corresponds to an alphabet developed in Kevin Karplus' laboratory to study more precisely the β -strands [140] and PROMOTIF [55]. DSSP was used to define the α -helices analyzed by HELANAL. Hence, we encoded the 3D protein structures in terms of secondary structures using these different algorithms (Fig. 2, arrow 2), and analyzed the correspondence between the secondary structure and the PB assignments (Fig. 2, arrow 3).

Z-score : Amino acid occurrence matrices were computed for each PB and normalized into Z-scores as follows : $Z\text{-score} = (n_{obs}(i,x) - n_{th}(i,x)) / \sqrt{n_{th}(i,x)}$, with $n_{obs}(i,x)$ the occurrence number of observing amino acid i in PB x , and $n_{th}(i,x)$ the occurrence number expected. $n_{th}(i,x) = N_x \cdot f_i$, where N_x and f_i denote the occurrence number of PB x and the frequency of amino acid i in the entire databank respectively (Fig. 2, arrow 4). Positive Z-scores, more than a user-fixed threshold ε (respectively negative, less than $-\varepsilon$) correspond to overrepresented amino acids (respectively underrepresented).

Prediction of PBs by a Bayesian probabilistic approach: The goal is to predict the optimal PB for each position along a sequence of length L (Fig. 2, arrow 5). To this end, we used a Bayesian probabilistic approach similar to that proposed in a previous work [128]. We focused on the conditional probability of observing the PB_k given an amino acid chain X , (a_1, a_2, \dots, a_p) , noted $P(PB_k / X)$. Bayes' theorem accomplishes the inversion between the sequence X and the structure PB_k . This leads to:

$$P(X | PB_k) = P(a_1 | PB_k) \times P(a_2 | PB_k) \times \dots \times P(a_p | PB_k)$$

Structural alphabet

A window of length p ($p=15$ here) is slide along the sequence and centered on a position s . To define the optimal Protein Block, PB^* for a given amino acid fragment X at a site in a protein, we used the prediction score R_k :

$$R_k = P(X | PB_k) / P(X) = P(PB_k | X) / P(PB_k)$$

The ratio R_k measures the information provided by the knowledge of the amino acid chain X in the prediction of the Protein Block PB_k . This criterion is equivalent to a ratio of likelihood's. The optimal structural block PB_k among the 16 possible blocks is defined as $PB^* = \text{argmax}\{R_k\}$. Then PB^* is assigned to the central residue of the chain X . The final prediction rate, noted Q_{16} , is the ratio between the number of PBs correctly predicted and all the PBs of the protein.

To assess the prediction, the databank was divided into two sets. The first one was used to define the PBs sequence-structure relationship and hence to compute : $P(PB_k / X)$. The second set was used to perform the prediction.

To improve the prediction rate, we used the concept of the sequence families which lies on the fact that a local fold can be associated to different clusters of sequences. The initial prediction rate was of 34.2% and was improved to 40.7% [128]. Now, with a new approach (*manuscript in preparation*), the prediction rate reaches 48.7%.

Protein classes assignment: To define for a protein its class, we have used the definition of Michie and co-workers: an all- α protein is characterized by a frequency of α -helix of more than 60% and of β -strand less than 15%, and, an all- β protein by a frequency of α -helix of less than

Structural alphabet

15% and of β -strand of more than 35% [148].

Analysis of the protein structural classes from the prediction in terms of PBs: the predicted PBs frequencies per proteins were computed from the results of the Bayesian prediction. They were analyzed using a Principal Component Analysis (PCA) [157] in regards to their structural classes [148] (all α , all β or mixed) (Fig. 2, arrows 6 and 7).

Prediction of the structural classes from the prediction in terms of PBs: For the 3 structural classes, mean values of each predicted PB frequencies were computed. The prediction step is a comparison between the each target protein predicted PB frequencies and the mean values for the 3 structural classes using an Euclidean distance. The smallest distance defines the assignment to the predicted class.

TopKAPi. In parallel, we computed the true PB frequencies per protein (Fig. 2, arrow 8) and learnt them using a Self-Organizing Map, named TopKaPi for Triangular Kohonen map for Analyzing Proteins (Fig. 2, arrow 9). Kohonen Map or SOM (Self – Organizing Map) is an efficient way to classify data [155]. The analysis of the results is highly facilitated by the nearness of related clusters. The main specificity of our SOM, is to be a triangle. It is composed of $w = G \times (G-1)/2$ neurons (triangle side of G neurons). A neuron w^f is similar to the vector v (dimension 16). The learning is iterative and consists in 5 consecutive steps:

- (i) random choice of an observation vector v .
- (ii) v is compared to every w neurons using an Euclidean distance.
- (iii) the winning neuron w^* , the closest to v , is identified, i.e. the Euclidean distance is

Structural alphabet

minimal.

- (iv) each neuron w^t of the SOM is modified :

$$w^t + 1 \leftarrow w^t + (v - w^t) \alpha(n) \pi(n)$$

with $\alpha(n)$ the learning factor and $\pi(n)$ the neighbourhood factor $\alpha(n)$ is defined as $\alpha_0 / (1 + (n / N))$, with $\alpha_0 = 5 / 1000$, n the number of observation vectors learnt and N the total number of observation vectors. $\pi(n)$ controls the diffusion process and is defined by $\exp(-2(r-r^*)^2 / \rho(n)^2)$, r is the coordinates of the neurons w^t , r^* the coordinates of the winning neuron w^* , $\rho(n) = \rho_0 / (1 + (n / N))$ with $\rho_0 = 2.4$.

- (v) the process is reiterated from (i) to (iv) with another vector.
- (vi) To learn all the observation vectors of the databanks, the whole databank is used C times ($C = 50$).

Then, we analyzed the correspondence between the different clusters obtained and the relationship between local structure in terms of PB distribution and frequencies of amino acids (Fig. 2, arrows 10 and 11).

	stride	psea	pcurve	define	xtlsstr	secstr
dssp	95.28	80.40	77.56	61.81	80.36	93.53
stride		81.44	77.99	62.07	80.50	91.40
psea			83.26	64.66	75.90	80.11
define				64.92	60.11	61.55
pcurve					74.56	77.38
xtlsstr						79.47

Table 3. Agreement rate between the different states defined by seven secondary structure assignment methods.

Part I: Secondary structures and PBs

Correspondence between the different secondary structure assignment methods. The secondary structure definitions are often considered as fixed and the assignment unique [4, 5]. However, as we have noted in the introduction, the reality is less simple. Table 3 gives the agreement ratios between the different *Secondary Structure Assignment Methods* (SSAMs). These values are computed as the proportion of identical assignment between two SSAMs. To compute these ratios, we have carried out for the SSAMs with more than 3 states a reduction of their N states to a classical 3-state alphabet (helix, extended and coil). We have done the classical associations for the helicoidal states (i.e. α -helix, 3_{10} -helix and π -helix), the extended states (i.e. extended strand and β -sheet) and the coil state (the other states: turn, bend, polyproline II and coil), even if these associations are not always pertinent. Table 3 shows that two SSAMs can strongly disagree.

A first cluster of SSAMs can be distinguished with DSSP [101], STRIDE [104] and SECSTR [24], which have agreement rates within the range [91.4%; 95.3%]. These 3 SSAMs will be noted DSS (*DSSP – STRIDE – SECSTR*). They have in common a similar assignment criterion, i.e. the hydrogen bonds computation. Interestingly, between these three methods, the extended structures are not the ones that have the most disruptive assignment. When a divergence in the assignment occurs, in 80% of the case it is between the helicoidal state and the coil state. This fact is particularly clear for SECSTR which was designed to better assign the less frequent helicoidal states (3_{10} and π -helices).

When different assignment criteria are compared (e.g. distances or angles), the agreement rates are within the range [75.9%; 83.2%] for DSS, PSEA [110], PCURVE [109] and XTLSSTR [111]. DEFINE [108] is clearly an outlier SSAM. It creates long successions of identical states,

Structural alphabet

and its helix frequency is only equal to 27%. This value is largely inferior to all the other SSAMs since the helix frequency is always greater than 31%. Moreover, it is the only one to create high assignment confusion between helix and strand. This awkward confusion is within the range [2% - 5%] between DEFINE and all the other methods. For the others, this confusion α / β is always less than 0.05%. Thus, we show that the definition of new rules and methods since the beginning of the 90's has not changed the heterogeneity of the secondary structure assignments and that the remarks of Woodcock and co-workers [114] about the difficulties of comparing different assignment methods still remain true.

Example of a protein coding using different secondary structure assignment methods.

Figure 4 gives the example of the secondary structure assignments obtained for the Hhai Methyltransferase protein (PDB code: 10MH) using the different SSAMs. As Table 3, this figure highlights the difficulties of comparing these methods.

Firstly, the secondary structures of the Hhai Methyltransferase are assigned with a 3-state alphabet by DSSP, STRIDE, PSEA, DEFINE and PCURVE, and the difficulties of finding a consensus (cons.) between all these methods appear clearly since they are very few stars corresponding to a perfect match between the 5 SSAMs. The main disagreements are observed for the α -helices and β -strands edges and length. With the consensus method (noted C93) described by Colloc'h, Etchebest and co-workers [115] which involves the three oldest methods among these five ones, i.e. DSSP, DEFINE and PCURVE, we observe a rate of complete non agreement (i.e. one method assigns a α -helix, the second one a β -strand and the third one a coil) of 1%.

Structural alphabet

Secondly, the secondary structure assignments are represented for the methods that give more than three states and the results are even more difficult to analyze. For instance, at the N-terminus of the protein (box 1), when XTLSSTR assigns two positions as PolyProline II followed by a strand and a series of turns, DSSP shows a small bend followed by a turn, STRIDE a longer series of turns, and SECSTR assigns a 3_{10} helix instead of the turns. The box 2 is another interesting example, because it reflects classical confusing problems. The 3-state descriptions give a region mainly in coil or with a short helix (except for DEFINE). The consensus C93 also gives a short helix. SECSTR and DSSP assign a 3_{10} helix which stays coherent with C93. Nevertheless, XTLSSTR and STRIDE assign those positions as turns. This classical confusion between 3_{10} helix and turns is the main reason of the exclusion of type III β -turn from the β -turn classification. In the following positions, we observe the same feature but inversed. XTLSSTR assigns a 3_{10} helix when STRIDE and DSSP give a turn.

The interest of the Protein Blocks appears through this example. When most of the SSAMs agree, the assigned PBs are coherent with the regular secondary structure states. For instance, the core of the α -helices are described by PB *m* and the core of the β -strands by PBs *c* and *d*. In addition, the PBs give a more detailed description in the confused positions. For example, in box 2, the Polyproline II helix assigned by XTLSSTR has its N-cap characterized by the PBs *fkllpc*, a well characterized series of 5 PBs identified in a previous work as a Structural Word (SW). A SW is a series of PBs which is found with an important occurrence in the databank. The SWs identified have shown a particularly high structural stability [133]. This SW *fkllpc* is characterized by a strong kink which induces a significant change in the backbone orientation. Its last dihedral angles are mainly associated to β -strand values. Thus, it is coherent with a transition between a tight turn and a Polyproline II helix, this last having dihedral angle values in the β -strand upper

Structural alphabet

left region of the Ramachandran Map.

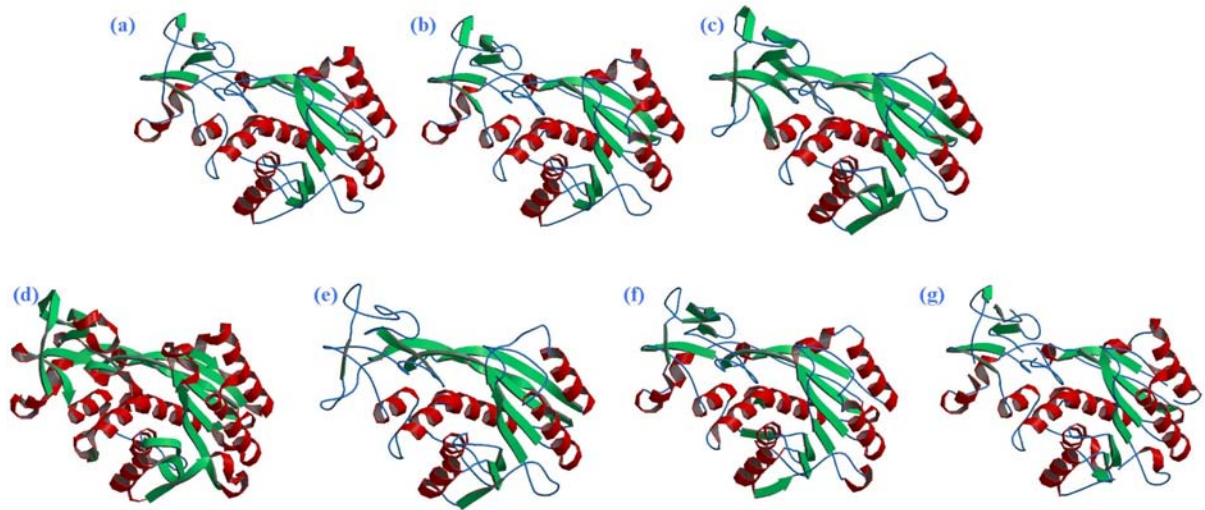


Figure 5. Example of secondary structure assignments for the protein 10MH with (a) DSSP, (b) STRIDE, (c) PSEA, (d) DEFINE, (e) PCURVE, (f) XTLSSTR and (g) SECSTR. All the methods have been reduced to three states with the helicoidal states in red ribbons, the extended state in green arrows and the coil in blue line.

Figure 5 shows the global 3D structure of the Hhai Methyltransferase according to the seven assignment methods. This picture highlights again the heterogeneity of the secondary structure assignments. For instance, we can note the helices in the upper right of each picture. With DSSP and STRIDE, two helices are found creating an α - α corner, i.e. two helices which are orthogonal [79]. With PSEA, only one helix remains, the shorter one is not considered. DEFINE assigns all in helicoidal state in a surprising way, i.e. even the residues which are in the kink (deviation of 90°). XTLSSTR gives the same result as PSEA, but shorten more the remaining helix. SECSTR, as already noted, does a treatment very similar to DSSP and STRIDE with slight differences at the extremities. With the exception of some particularly well characterized structures like the Schellman box, the precise determination of repetitive structure capping limits is highly difficult [16].

Structural alphabet

(a)

PB	helicoïdal state							coil state						
	DS.	ST.	PS.	DE.	PC.	XT.	SE.	DS.	ST.	PS.	DE.	PC.	XT.	SE.
<i>a</i>	0.13	0.12	0.08	15.96	0.07	2.38	0.72	79.80	73.55	66.77	57.79	67.50	72.92	82.95
<i>b</i>	0.29	0.20	0.19	11.32	0.01	0.26	1.18	85.55	84.73	96.32	54.01	80.22	80.00	98.43
<i>c</i>	0.02	0.04	0.83	15.28	0.03	0.10	0.20	55.42	54.26	49.20	50.15	52.51	58.47	57.50
<i>d</i>	0.00	0.02	0.02	9.13	0.01	0.03	0.00	27.69	26.48	19.76	40.30	19.99	40.90	32.29
<i>e</i>	0.14	0.08	0.07	9.91	0.00	0.04	0.30	47.70	46.17	43.78	47.75	55.77	65.51	50.05
<i>f</i>	0.02	0.05	0.11	11.40	0.04	18.49	0.33	71.54	69.28	69.94	51.55	78.14	64.95	74.28
<i>g</i>	13.10	11.93	10.63	24.70	1.69	11.67	21.35	79.34	79.81	84.74	58.81	93.44	82.95	72.95
<i>h</i>	3.53	2.45	0.30	11.97	0.02	0.22	7.06	76.59	76.81	75.75	59.92	88.39	89.32	76.39
<i>i</i>	3.70	2.36	0.60	11.05	0.03	0.82	10.37	90.20	91.01	96.79	64.48	81.65	91.21	89.17
<i>j</i>	10.63	8.60	0.99	12.68	2.02	8.79	12.92	79.05	79.65	75.00	57.90	93.50	85.86	80.44
<i>k</i>	47.69	48.91	33.07	22.96	23.47	53.06	48.21	51.89	50.69	66.15	56.86	73.86	46.70	51.74
<i>l</i>	59.72	59.91	41.05	29.34	42.60	61.59	61.14	39.62	39.28	58.67	55.20	56.10	38.17	38.46
<i>m</i>	90.09	91.83	85.96	51.08	86.72	91.70	92.63	9.74	7.98	14.01	40.41	13.20	8.26	7.25
<i>n</i>	67.99	72.36	62.36	44.09	56.78	71.99	70.64	31.56	27.13	37.54	46.41	43.00	27.84	28.88
<i>o</i>	29.30	49.73	39.83	37.07	7.84	43.15	28.97	70.20	49.85	59.83	52.19	91.68	56.66	70.72
<i>p</i>	16.05	15.93	9.48	32.51	1.90	17.27	22.27	82.56	82.30	86.81	54.93	90.97	73.37	76.55

Table 4 (a). helicoïdal and coil state frequencies for each of the 16 PBs with DSSP (DS.), STRIDE (ST.), PSEA (PS), DEFINE (DE), XTLSSTR (XT) and SECSTR(SE).

(b)

PB	strand state									
	DS.	ST.	PS.	DE.	PC.	XT.	SE.			
<i>a</i>	20.07	26.32	33.15	26.25	32.43	24.69	16.33			
<i>b</i>	14.16	15.07	3.49	34.67	19.76	19.73	0.40			
<i>c</i>	44.55	45.69	49.97	34.57	47.45	41.45	42.29			
<i>d</i>	72.31	73.50	80.22	50.57	80.00	59.08	67.71			
<i>e</i>	52.16	53.76	56.14	42.35	44.23	34.46	49.64			
<i>f</i>	28.44	30.68	29.95	37.04	21.82	16.55	25.40			
<i>g</i>	7.56	8.25	4.63	16.49	4.87	5.37	5.71			
<i>h</i>	19.88	20.74	23.95	28.12	11.59	10.46	16.55			
<i>i</i>	6.09	6.61	2.61	24.46	18.31	7.97	0.46			
<i>j</i>	10.33	11.76	24.01	29.42	4.48	5.36	6.64			
<i>k</i>	0.43	0.40	0.78	20.18	2.67	0.24	0.04			
<i>l</i>	0.66	0.81	0.28	15.47	1.30	0.23	0.41			
<i>m</i>	0.18	0.20	0.02	8.51	0.08	0.05	0.13			
<i>n</i>	0.46	0.51	0.10	9.50	0.23	0.17	0.48			
<i>o</i>	0.49	0.42	0.34	10.73	0.48	0.20	0.31			
<i>p</i>	1.39	1.75	3.71	12.56	7.13	9.34	1.17			

Table 4 (b). extended state frequencies for each of the 16 PBs with DSSP (DS.), STRIDE (ST.), PSEA (PS), DEFINE (DE), XTLSSTR (XT) and SECSTR(SE).

Structural alphabet

Moreover, some algorithms are highly sensitive to the quality of the protein structures, i.e. resolution and temperature factors. For instance, a limited change in resolution or temperature factors can modify the DSSP secondary structure assignments.

The Protein Blocks and the classical secondary structure 3-state description. Table 4 (parts a and b) gives the complete distribution of the 3-state secondary structures for the seven studied methods in the Protein Blocks. As seen in the last paragraphs, DEFINE has a distinct behaviour in regards to the other methods, so we will not take it into account in the following sections.

For all the methods, we observe that the PB *m* is associated to the α -helix with a mean frequency of 90%. The 10% left correspond to coil. The α -helix is also described by the PBs *n* (67%) and *l* (54%). The PB *d* is associated at 72% to the β -strand and at 28% to the coil. These values underline one more time the β -strand definition problem. The main interest of our structural alphabet is a better description of the coil state by PBs *i* (90%), *b* (87%), *j* (82%), *p* (82%), *h* (80%), *f* (71%), *o* (66%), *k* (56%), *c* (54%) and *e* (51%).

Tables 4a and 4b enable to further analyze the differences between the secondary structure assignment methods. For seven PBs (i.e. PBs *a*, *b*, *g*, *i*, *j*, *o* and *p*), we observe some significant differences in their secondary structure assignments. For instance, the PB *o* has an α -helix frequency of 29.3 % and a coil frequency of 70.2% with DSSP whereas these frequencies are equal to 49.7% and 49.9%, respectively with STRIDE, although these two methods are really close. The PBs *g*, *i* and *p* also present high variations in their α -helix and coil frequencies according to the different methods. The PB *a* shows high differences with β -strand and coil frequencies of 20.0 % and 79.8% with DSSP versus 26.3% and 73.6% with STRIDE. Furthermore, its β -strand frequency is only equal to 16.3% with SECSTR. A low value of

Structural alphabet

SECSTR compared to DSSP and STRIDE is also observed for PBs b and i : the β -strand frequencies of PB i and b are equal to 6.1% and 14.1% for DSSP, to 6.6% and 15.1% for STRIDE and only 0.5% and 0.4% for SECSTR. This last point is intriguing as SECSTR was specifically dedicated to perform a better assignment of the helicoidal states than DSSP and STRIDE while giving a similar assignment for the extended state.

We compare all the SSAMs according to their 3-state secondary structure frequencies in the different PBs. For the helicoidal state, we observe a hierarchy $\text{XTLSSTR} > \text{DSS} > \text{PCURVE} > \text{PSEA}$ in the frequencies associated to PB m . For the extended state, it is the inverse for the frequencies characterizing the PB d , with $\text{XTLSSTR} < \text{DSS} < \text{PCURVE} < \text{PSEA}$. Finally, for the coil state, we can roughly note the hierarchy $\text{PSEA} > \text{PCURVE} > \text{DSS} > \text{XTLSSTR}$.

The Protein Blocks and the secondary structure N-state description. Table 5 focuses on three SSAMs that describe the secondary structures with more than three states, i.e. DSSP, STRIDE and SECSTR, and shows the correspondence with the PBs. The helicoidal state is characterized by α -helices, 3_{10} -helices and π -helices. For the α -helix state, the frequencies in the different PBs are close except as previously noted for the PB o which has an α -helix frequency equal to 22.2% for DSSP, 41.4% for STRIDE and only 14.5% for SECSTR. All the α -helix frequencies are lower for SECSTR. However, it is the opposite for the 3_{10} -helices and in a lesser extent for the π -helices, since for the PBs from g to p the 3_{10} -helix frequencies are 2 to 10% greater than DSSP and STRIDE. This last fact is consistent with the main purpose of SECSTR. The PB l has a 3_{10} -helix frequency of 19.2%. The PB g and p are also especially well furnished with 3_{10} -helix frequencies of about 17%. The π -helix frequencies are low, but far superior with SECSTR.

Structural alphabet

(a)

PB	α - helix			3_{10} - helix			π - helix		
	DSSP	STRIDE	SECSTR	DSSP	STRIDE	SECSTR	DSSP	STRIDE	SECSTR
<i>a</i>	0.07	0.06	0.09	0.05	0.04	0.60	0.01	0.02	0.03
<i>b</i>	0.15	0.14	0.19	0.12	0.06	0.95	0.02	0.00	0.04
<i>c</i>	0.01	0.04	0.03	0.01	0.00	0.17	0.00	0.00	0.00
<i>d</i>	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>e</i>	0.03	0.02	0.03	0.09	0.04	0.24	0.02	0.02	0.03
<i>f</i>	0.02	0.04	0.07	0.00	0.01	0.26	0.00	0.00	0.00
<i>g</i>	4.55	4.56	4.02	8.52	7.37	17.26	0.03	0.00	0.07
<i>h</i>	0.43	0.04	0.48	3.08	2.39	6.56	0.02	0.02	0.02
<i>i</i>	0.57	0.02	0.55	3.11	2.32	9.80	0.02	0.02	0.02
<i>j</i>	5.24	4.43	4.75	5.34	4.17	8.12	0.05	0.00	0.05
<i>k</i>	34.68	35.51	32.37	12.99	13.39	15.79	0.02	0.01	0.05
<i>l</i>	43.42	43.47	41.76	16.25	16.40	19.20	0.05	0.04	0.18
<i>m</i>	85.64	87.29	81.95	4.39	4.49	9.83	0.06	0.05	0.85
<i>n</i>	60.40	62.20	57.57	7.57	10.14	12.54	0.02	0.02	0.53
<i>o</i>	22.19	41.36	14.49	7.10	8.36	14.24	0.01	0.01	0.24
<i>p</i>	4.56	6.26	4.62	11.49	9.67	17.56	0.00	0.00	0.09

Table 5 (a). The three helicoidal states frequencies for each of the 16 PBs defined by DSSP, STRIDE and SECSTR.

Structural alphabet

(b)

PB	turn		bend		coil			iso. β -bridge		extended strand		
	DSSP	STRIDE	DSSP	STRIDE	DSSP	STRIDE	SECSTR	DSSP	STRIDE	DSSP	STRIDE	SECSTR
<i>a</i>	3.28	32.53	16.58	59.94	41.02	82.95	3.01	3.70	17.06	22.62	16.33	
<i>b</i>	11.79	42.14	48.24	25.52	42.59	98.43	0.14	0.15	14.02	14.92	0.40	
<i>c</i>	0.47	21.23	13.34	41.61	33.03	57.50	3.74	3.38	40.81	42.31	42.29	
<i>d</i>	0.11	3.60	5.36	22.22	22.88	32.29	1.79	1.73	70.52	71.77	67.71	
<i>e</i>	1.01	30.56	7.97	38.72	15.61	50.05	2.86	3.00	49.30	50.76	49.64	
<i>f</i>	0.52	30.28	7.41	63.61	39.00	74.28	3.67	3.74	24.77	26.94	25.40	
<i>g</i>	15.49	60.36	35.81	28.04	19.45	72.95	3.52	3.76	4.04	4.49	5.71	
<i>h</i>	48.47	64.91	13.87	14.25	11.90	76.39	2.15	1.94	17.73	18.80	16.55	
<i>i</i>	61.79	79.17	21.13	7.28	11.84	89.17	0.25	0.23	5.84	6.38	0.46	
<i>j</i>	26.78	42.09	31.47	20.80	37.56	80.44	1.83	2.70	8.50	9.06	6.64	
<i>k</i>	34.91	45.51	10.27	6.71	5.18	51.74	0.04	0.01	0.39	0.39	0.04	
<i>l</i>	24.36	34.90	9.02	6.24	4.38	38.46	0.25	0.30	0.41	0.51	0.41	
<i>m</i>	5.36	6.23	1.70	2.68	1.75	7.25	0.10	0.10	0.08	0.10	0.13	
<i>n</i>	22.77	23.80	4.68	4.11	3.33	28.88	0.18	0.17	0.28	0.34	0.48	
<i>o</i>	56.90	41.00	8.81	4.49	8.85	70.72	0.39	0.23	0.10	0.19	0.31	
<i>p</i>	45.00	38.71	17.46	20.10	43.59	76.55	0.25	0.36	1.14	1.39	1.17	

Table 5 (b). Turns, bends, coil, isolated β -bridge (iso. β -bridge) and extended strand states frequencies for each of the 16 PBs defined by DSSP, STRIDE and SECSTR.

PB	α - helix					extended strand						
	short	linear	curved	kinked	una.	b	p	q	a	z	m	e
<i>a</i>	0.04	0.00	0.00	0.03	0.00	3.05	0.38	1.96	1.63	11.44	0.47	0.92
<i>b</i>	0.08	0.00	0.04	0.02	0.00	0.14	0.02	0.18	0.02	0.23	0.01	13.48
<i>c</i>	0.01	0.00	0.00	0.00	0.00	3.71	3.93	7.65	6.14	20.75	1.79	0.53
<i>d</i>	0.00	0.00	0.00	0.00	0.00	1.77	8.03	9.97	17.81	29.39	4.77	0.67
<i>e</i>	0.02	0.00	0.00	0.02	0.00	2.95	6.05	6.66	8.77	25.13	2.15	0.54
<i>f</i>	0.01	0.00	0.00	0.01	0.00	3.61	1.15	4.98	1.50	16.24	0.49	0.33
<i>g</i>	1.85	0.14	1.03	1.30	0.00	3.59	0.00	0.95	0.04	2.57	0.11	0.46
<i>h</i>	0.18	0.00	0.00	0.25	0.00	2.15	2.37	3.41	1.27	8.86	0.29	1.38
<i>i</i>	0.21	0.00	0.00	0.30	0.00	0.24	0.00	0.16	0.00	0.16	0.00	5.33
<i>j</i>	1.48	0.59	2.42	0.69	0.05	1.88	0.00	1.93	0.56	4.16	0.05	1.83
<i>k</i>	13.52	2.99	15.31	2.68	0.12	0.04	0.00	0.01	0.00	0.02	0.00	0.35
<i>l</i>	16.76	3.75	19.08	3.52	0.18	0.23	0.00	0.07	0.01	0.17	0.01	0.15
<i>m</i>	17.52	8.47	48.63	10.34	0.58	0.10	0.00	0.02	0.00	0.05	0.00	0.00
<i>n</i>	18.61	5.74	29.08	6.31	0.29	0.19	0.00	0.15	0.00	0.13	0.00	0.00
<i>o</i>	6.92	1.93	10.63	2.43	0.11	0.36	0.00	0.00	0.01	0.03	0.01	0.03
<i>p</i>	2.06	0.34	1.38	0.71	0.01	0.26	0.00	0.01	0.02	0.94	0.00	0.18

Table 6. Analysis of the repetitive structures defined by DSSP in terms of Protein Blocks. The α -helices are divided into 5 categories as short helices, linear helices, curved helices, kinked helices and unassigned helices (una.) using HELANAL. The extended strands are described as residue in isolated β -bridge (b) or as extended strand in : parallel in sheet (p), parallel edge (q), antiparallel sheet (a), antiparallel edge (z), parallel and antiparallel mixed (m) and strand, alone (e).

Structural alphabet

The coil state defined in Table 4 is decomposed in four categories, namely the coil ('C'), the turn ('T'), the bend ('N') and the isolated β -bridge ('B'). As observed in Table 4, the non-repetitive structures present equivalent frequencies in the different PBs according to the three SSAMs. However, Table 5 shows that between the different types of classification, the results are clearly distinct. The turns of DSSP (states 'T' and 'N') are not equivalent to the turns of STRIDE ('T'). Their average frequencies per PBs differ by more than 10%. The turns of DSSP are more frequent in PBs *o* (+25%), *p* (+25%), *b* (+19%) and *j* (+17%) and the turns of STRIDE ('T') are more frequent in PBs *f* (+21%), *e* (+21%), *a* (+11%) and *g* (+7%). This point is particularly important as the turns are commonly used to describe more precisely the protein structures. For the isolated β -bridge ('B'), the frequencies are really similar between DSSP and STRIDE, the difference between the two assignment methods is always less than 0.9%. For the extended strand, the results are the same as previously found. These results highlight the complexity of describing only particular regions. The differences in the number of analyzed local folds can bias the analysis of the results.

The Protein Blocks and the precise description of the repetitive secondary structures. Table 6 summarizes the distribution of the α -helices and extended strands using more detailed descriptions.

The helices of the Protein DataBank [102, 103] are known not to be ideal helices according to the thermodynamical properties. The use of the SSAMs often creates helices that are too long. The longest helices in the Protein DataBank contain about 60 residues. Barlow and Thornton [112] have shown that 3/4 of the helices are not linear, i.e. they are curved or kinked. HELANAL [113] allows to redefine helices into 5 categories: short (less than 9 residues), linear, curved,

Structural alphabet

kinked or unassigned. We have used DSSP definitions of the helices to compute the assignment. As expected, the most frequent PB associated with the different categories is PB *m* with 48.6% associated to curved helices, 17.5% to short helices, 10.3% to kinked helices and only 8.5% to linear helices. PB *m* represents 70.1% of the PBs associated to short helices, 82.6% to linear helices, 84.2% to curved helices and 84.4% to kinked helices. We observe that the short helices are described by several other PBs including PB *n* (18.6%), *l* (16.8%) and *k* (13.5%), with PB *n* frequency greater than that of PB *m*. For the other types of helix, PB *m* remains the most important although for the curved helices for instance many other PBs are involved in their description.

In the same way than for HELANAL, the extended beta alphabet used the DSSP outputs to define different new labels : isolated β -bridge ('b'), extended strand in parallel sheet ('p') or parallel edge ('q'), antiparallel sheet ('a'), antiparallel edge ('z'), parallel and antiparallel mixed ('m') and strand only ('e'). As expected, the PBs *d*, *c* and also *e* are the most frequent ones. In addition, two interesting facts must be pointed out. The first one is the isolated β -bridge ('b') which is characterized by the PB *g* with a frequency of 3.6% even though it is not a PB particularly associated to extended structures. The second one is the distribution of the extended strand alone ('e') which is mainly associated to the PB *b*, a PB associated with long loops and Ccap of β -strand, and to PB *i*, which is more associated to the coil state than to the extended state.

These results show that even for the repetitive not-so ideal structures, the Protein Blocks constitute an interesting analyzing tool.

PB	helix			turns			PII	PII Ccap	coil	extended
	α	3_{10}	3_{10} Ccap	h.-bonded	unh.-bond	PII				
<i>a</i>	1.64	0.73	0.01	3.34	2.07	22.15	0.42	44.94	24.69	
<i>b</i>	0.14	0.01	0.11	9.31	7.40	0.91	0.03	62.35	19.73	
<i>c</i>	0.06	0.02	0.02	0.66	0.75	23.13	5.36	28.57	41.45	
<i>d</i>	0.02	0.01	0.00	0.24	0.37	17.09	4.80	18.40	59.08	
<i>e</i>	0.04	0.00	0.00	12.14	3.19	19.96	5.64	24.58	34.46	
<i>f</i>	12.77	5.72	0.00	12.19	4.04	0.56	24.96	23.20	16.55	
<i>g</i>	4.17	0.11	7.39	23.54	7.81	14.70	0.35	36.55	5.37	
<i>h</i>	0.20	0.02	0.00	35.09	6.92	1.69	23.51	22.11	10.46	
<i>i</i>	0.73	0.09	0.00	39.17	5.87	2.10	0.16	43.91	7.97	
<i>j</i>	6.13	2.66	0.00	34.63	8.02	0.05	0.31	42.85	5.36	
<i>k</i>	36.93	16.13	0.00	31.42	9.09	0.02	0.07	6.10	0.24	
<i>l</i>	43.18	18.40	0.01	24.89	7.68	0.13	0.13	5.34	0.23	
<i>m</i>	75.73	13.11	2.86	4.49	1.60	0.02	0.01	2.14	0.05	
<i>n</i>	52.45	14.44	5.10	14.10	4.24	0.10	0.21	9.19	0.17	
<i>o</i>	30.05	1.88	11.22	28.11	5.21	0.09	0.03	23.22	0.20	
<i>p</i>	4.48	0.33	12.46	11.39	2.67	0.21	0.00	59.10	9.34	

Table 7. Correspondence between the states defined by XTLSSTR and the 16 Protein Blocks with the α -helix, the 3_{10} helix and its Ccap, the turns defined by the presence (h.-bonded) or absence (unh.-bond) of an hydrogen stabilizing bond, the polyproline II (PII) and its Ccap, coil and extended strand.

The Protein Blocks and XTLSSTR 9-state description. Table 7 summarizes the correspondence between the 16 Protein Blocks and the 9 states defined by XTLSSTR. This method has some interesting particularities like the assignment of turns not with the classical dihedral angle criteria but defined as hydrogen- or non hydrogen bonding-turns, and of polyproline II helices. Moreover, it identifies the Ccaps of the 3_{10} -helices and of the polyproline II helices. This SSAM gives different results in regards to the precedent methods. The PB *f* is now associated with a non negligible proportion of α -helix (12.7%) and of 3_{10} -helix (5.7%). This last fact is related to the PB *g* 3_{10} -Ccap frequency (7.4%) since the main transition of PB *f* is PB *g*. This value is coherent with the DSS frequency of 3_{10} -helix associated to PB *g* (DSSP, STRIDE and SECSTR frequencies are equal to 8.5%, 7.4% and 17.3%, respectively; cf. Table 5a). The PBs *o* and *p* are associated to the 3_{10} -Ccap (frequencies of 11.2% and 12.5%, respectively). In addition, we observe that globally more hydrogen bond turns are found than unhydrogen bond turns. Several PBs are involved in their description with no particular specificity related to the hydrogen bond stabilization. As for the polyproline II helices, they appear more frequent in globular proteins than expected [23]. Their dihedral angle distribution is often confused with β -strands in the upper left of Ramachandran Map and so is confused with the β -sheet assignment. This feature is observed again in these results where PBs *a*, *c*, *d*, *e* and *g* have polyproline II frequencies equal to 22.1%, 23.1%, 17.1%, 20.0% and 14.7%, respectively. Some PBs are specific to the C-cap of the polyproline II helix, e.g. PBs *f* and *h*. These observations are in agreement with the main transitions between successive PBs since PB *e* often goes to PB *f* and PB *g* to PB *h*.

Thus, the goal of this detailed analysis was to emphasize the fact that the “classic”

Structural alphabet

secondary structures can be described with different criteria which results in ambiguous assignments. Moreover, we have highlighted the interest of using more than three states for better describing protein structures through a structural alphabet. The 16 PBs enable to analyze specifically every part of the protein structures.

In the next section of this chapter, we propose a prediction scheme of the protein structural classes from the PB prediction.

Part II: PBs and protein structural classes

Goal. The question tackled in this section is the potentiality of classifying one protein into its true protein class from the sole knowledge of its prediction in terms of Protein Blocks. The process used is in three steps : (i) Protein Blocks are predicted from the sequence, (ii) the relative frequencies of the 16 PBs are computed and the three mean frequencies vectors, called prototypes, representing the three protein classes are computed from the learning protein set (iii) the comparisons between the three mean prototypes representing the protein classes and the relative frequencies of the 16 PBs are done for target proteins to predict their classes.

Having a good idea of the protein classes can be an efficient way for refining the prediction research. For instance, it can enable to direct the prediction of a protein, i.e. if a protein is all- α , the information derived from all- β proteins would not be used for this protein. The interest of this study is not to use the *true* PBs, but the *predicted* PBs. The prediction rate is, as previously said, equal to 40.7% [128]. Hence, the difficulty here is to predict accurately protein classes from this partial information.

Protein classes. As noted by Thornton and co-workers, the secondary structures form particular motifs that define the global protein topology, e.g. the TIM barrel fold [141]. This information is used to classify the protein structures. Different algorithms have been developed and the classification is done either automatically like in the CATH database [144] or mainly manually like in the SCOP database [143]. Different classes are defined and give information about the relationships between the proteins. Interestingly, the different methods give the same types of hierarchical relationships with few superfamilies and many subfamilies. On the basis of their secondary structures, the folds are grouped into four main classes: all- α (essentially α -helices), all- β (essentially β -strands), $\alpha + \beta$ (α -helices and β -strands are largely segregated) and α / β (α -helices and β -strands are largely interspersed). The assignment of a structure to a particular class is in some cases a difficult task. In fact, even with an automatic classification, a manual inspection is needed. From the sole knowledge of the amino acid sequence, the task is even more complicated when no homologous sequence is found.

Bayesian prediction of the Protein Blocks: In a previous study [128], we have tackled the Protein Blocks prediction from the amino acid sequence. To this end, we extracted the amino acid preferences for each local pattern and used this information in a Bayesian process to predict the structural motifs able to be adopted by a given protein chain. With this strategy, for each amino acid sequence, the potential series of Protein Blocks is predicted [128, 133]. To evaluate the prediction, we computed a Q_{16} ratio which corresponds to the number of well predicted PBs. This value is similar to the Q_3 of secondary structures with more states to predict, i.e. $N=16$

possibilities against $N=3$ for the secondary structures.

PB	PB frequency (%)	Bayesian prediction	Sequence Families	Sequence Families
		<i>simple</i>	(<i>Proteins</i> , 2000)	<i>New approach</i>
<i>a</i>	3.9	59.2	53.5	57.4
<i>b</i>	4.4	12.7	27.0	23.3
<i>c</i>	8.1	26.4	32.9	35.8
<i>d</i>	18.8	28.3	34.8	47.3
<i>e</i>	2.4	40.1	35.9	38.2
<i>f</i>	6.7	29.7	36.2	33.0
<i>g</i>	1.1	30.3	35.1	29.8
<i>h</i>	2.4	42.6	42.7	40.9
<i>i</i>	1.9	37.7	41.0	37.5
<i>j</i>	0.8	49.1	47.2	48.5
<i>k</i>	5.5	38.5	35.2	34.9
<i>l</i>	5.5	37.5	32.1	36.7
<i>m</i>	30.2	39.7	50.8	68.3
<i>n</i>	2.0	51.2	44.7	51.7
<i>o</i>	2.8	49.2	45.8	47.9
<i>p</i>	3.5	30.5	33.9	31.1
Q_{16}		35.4	40.7	48.7

Table 8. Bayesian prediction with Q_{16} value for the 16 Protein Blocks with their corresponding frequencies and prediction rate for the (*simple*) Bayesian prediction and the improved prediction using the sequence families with the original results (*Proteins*, 2000) [ADB00] and the new one (*new approach*).

With the new databank used in this study, we have an initial Q_{16} ratio equal to 35.4%, which is very similar to the result of our previous work, i.e. 34.4% [128]. Table 8 (col 3) gives the prediction rates for each of the 16 PBs. The high differences of the PB frequencies in the databank need to be taken into account (Table 8, col 2) since some of the PBs are overrepresented like PBs *m* and *d* (30.2% and 18.8% respectively). Consequently, we made sure that the Q_{16} ratio was not biased by over-predictions of PBs *m* and *d*.

However, associating one PB with one class of sequences is a restrictive point of view. A

Structural alphabet

same fold pattern (or PB) may be associated with different types of sequences (1 Protein Block \rightarrow n sequences). Thus, we have defined a new process to split the set of fragment sequences associated with one PB into different clusters. These clusters allow a better description of the sequence specificities associated with each PB. They are called sequence families (see [128] for more details). In our previous work, this process allowed to increase the Q_{16} ratio from 34.4% to 40.7% (see Table 8, col. 4). Since, we have improved the splitting process (*unpublished results*) which gives a better Q_{16} ratio now equal to 48.7% (see Table 8, col. 5). A clear improvement was observed for the PBs m and d (+28.6 and +19.0%, respectively), but also for the other PBs (with a mean improvement rate of +2.1%).

From this improved prediction, we have analyzed the association between the predicted PB frequencies and the protein structural classes. To carry out this analysis, we have focused on three protein classes, i.e. all- α , all- β and mixed. In this last class, we have merged the $\alpha + \beta$ and α / β proteins as our approach does not take account of the secondary structures sequentiality. The categorization of the proteins has been done using the criteria of Michie and co-workers [148] and DSSP secondary structure assignment.

Analysis of the prediction informativity. In a first step, we have analyzed the informativity of the prediction, i.e. the information contained in the relative predicted frequencies of Protein Blocks. For this purpose, we have performed a Principal Component Analysis (PCA) on the relative frequencies of the predicted PBs obtained for each protein. Such a descriptive approach allows quantifying this informativity by coding all the data in their original dimension, i.e. without information loss. The first component explains 92% of the information; the others represent less than 1.0%. This shows that the information of the predicted Protein Block

Structural alphabet

frequencies has an important determinism.

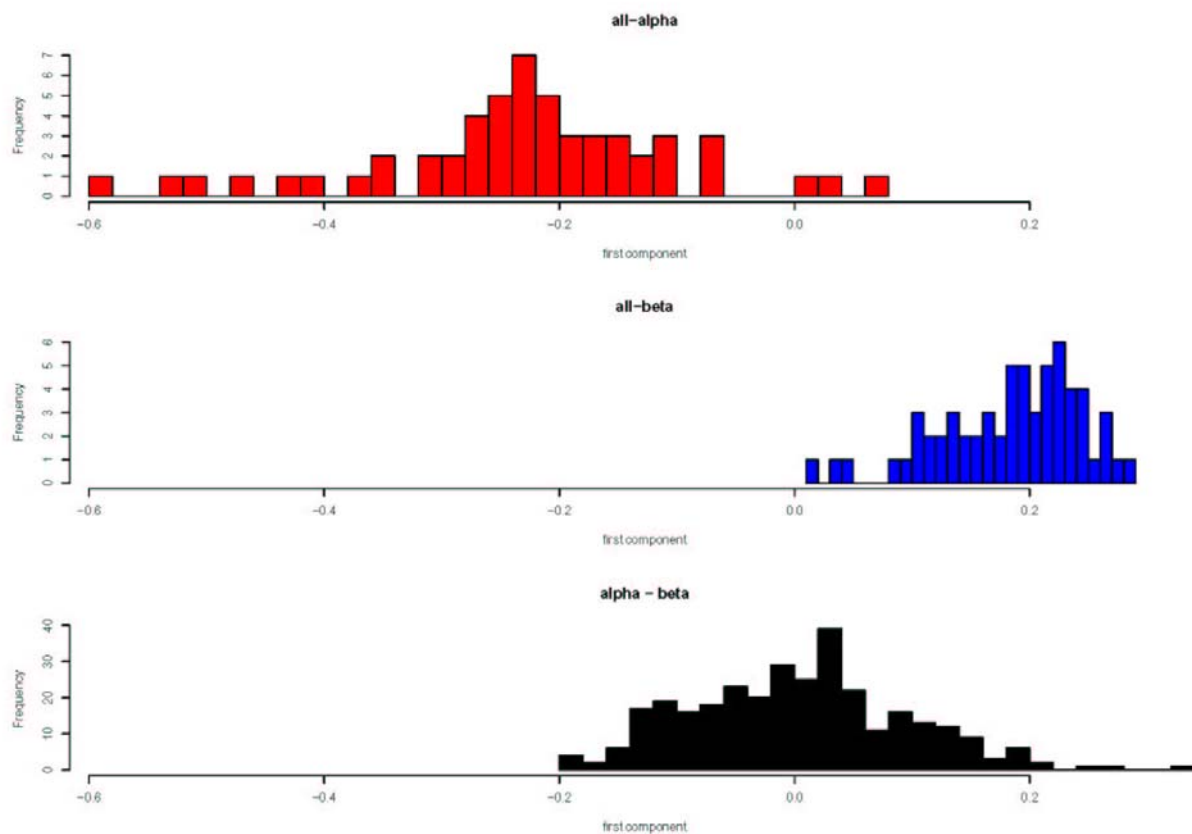


Figure 6. Values of the first component of the Principal Component Analysis carried out on the predicted frequencies of Protein Blocks, for the three studied protein classes all- α (blue), all- β (red) and mixed (black).

Figure 6 shows for each structural class the distribution of the first component values. We observe that the all- α proteins are the only one to have a first component value that can be less than -0.2. At the opposite, none of the all- β proteins is found at a value less than 0.0. The mixed group is between the two first ones and overlaps more the all- β proteins than the all- α proteins. Figure 7 shows the projection of the two first components. The all- α (red crosses) and the all- β

Structural alphabet

proteins (blue crosses) seem to be distinguishable using the PCA and form two distinct clusters. The first component is the most discriminative. However, the all- β proteins used more the second component than the all- α proteins. This analysis shows that we are able to discriminate the three structural classes according to the protein PB predicted frequencies.

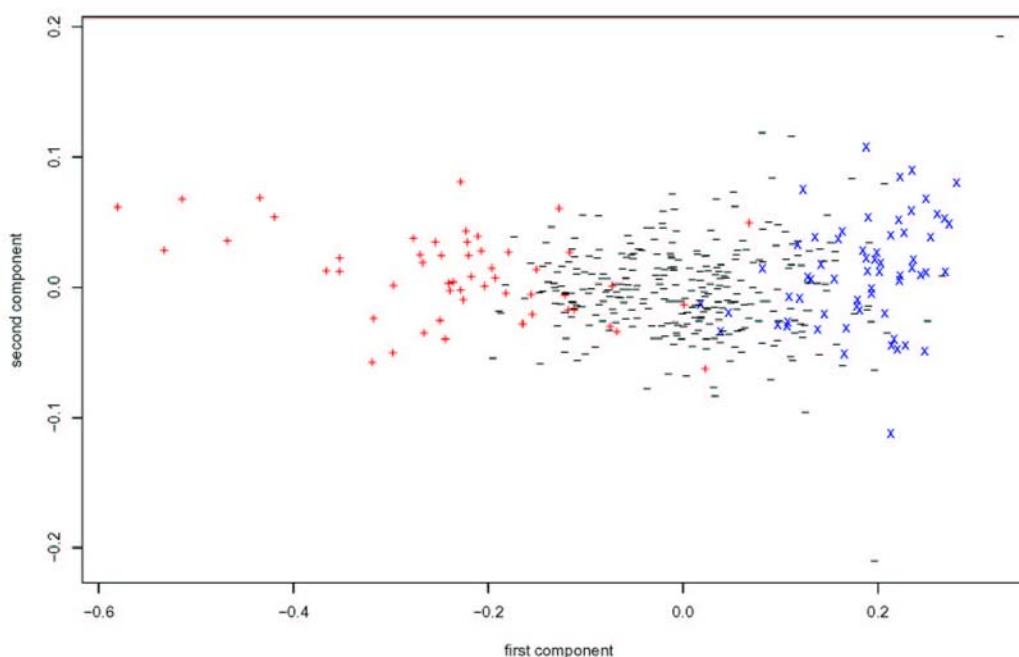


Figure 7. Description of the two first components (see Figure 6) for the three studied protein classes, all- α (blue '+'), all- β (red 'x') and mixed (black '-').

Prediction of the structural classes: In this part, we have identified the PBs principally involved in the different classes and we have used the predicted PB frequencies for prediction purpose. Figure 8 shows the relative frequencies of the predicted Protein Blocks for each structural class. These frequencies are normalized in regards to the frequencies of the predicted Protein Blocks in the databank and centred on 0, *i.e.* 0 represents the background frequency of the PB. The α - β proteins have average frequencies close to the background random Protein Block

Structural alphabet

frequencies. The all - β proteins have over - representations of the Protein Blocks from PB a to PB j and under - representations for the others. The all - α proteins have under - representations of the Protein Blocks from PB a to PB j and an important over - representation of PB m . These results are highly coherent with the definition of the different classes.

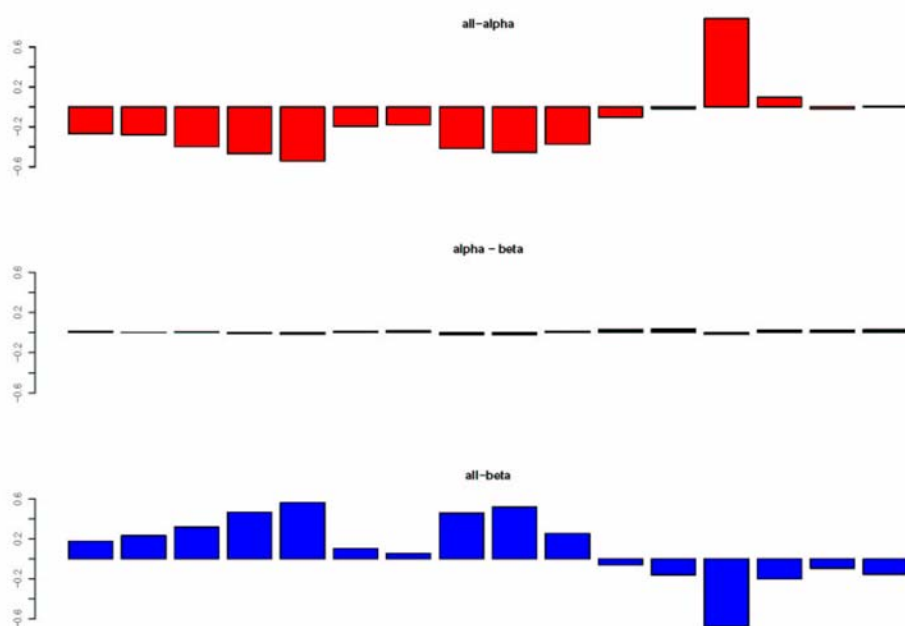


Figure 8. The three average prototypes of the relative predicted frequencies (RF) of the 16 Protein Blocks for the three studied protein classes all- α (blue), α - β (black) and all- β (red). In abscise are the Protein Blocks from PB a to PB p , in ordinate are given $(RF - 1.0)$ to highlight the frequency differences in regards to the random distribution.

		predicted		
		all - α	α - β	all - β
true	all - α	10.2	2.32	0.00
	α - β	12.7	50.6	9.7
	all - β	0.00	3.7	10.7

Table 9. Correspondence between the true classification of the proteins and the predicted class using the results of the Bayesian prediction.

Structural alphabet

From these 3 mean representations of the protein classes, we have developed a simple prediction strategy of the protein class from the PB predicted frequencies. Each target protein's predicted PB frequencies vector is compared with the three vectors of the protein classes using an Euclidean distance. The minimal distance gives the predicted class, i.e. the minimal difference between the protein and the prototype. Table 9 summarizes the results of the prediction by giving the confusion matrix between true and predicted protein classes. The prediction rate is equal to 71.5%. The prediction rates for the two extreme classes, i.e. all - α and all - β are high, i.e. 84% and 74%, respectively. The rate of correct prediction is lower for the mixed class, i.e. 69%. These results show that our simple prediction scheme is efficient. No confusion between all - α and all- β proteins is observed.

This work shows that from the prediction step, we can tackle the protein class with a good accuracy. It is interesting since the prediction is actually done indiscriminately for all types of globular proteins. By focusing on one particular class, we can supervised more efficiently the prediction. Hence, it can be useful to learn separately (all - α and α - β) proteins and (all - β and α - β) proteins.

Part III: Triangular Kohonen map for Analyzing Proteins (TopKAPi)

In the continuity of our previous work, we have attempted to assess the contribution of our structural alphabet for discriminating protein structures. A first step consisted in encoding the protein structures of our non-redundant databank in terms of Protein Blocks, and, in translating this information into PB frequencies, then into Z-scores. This work is related to the representation of relationships into a set of related proteins [159, 160].

Structural alphabet

A Sammon Map [158] was computed with this data and defined in a 2D space. Interestingly, it showed a triangular pattern (figure not shown). It must be noted that the Sammon Map is only an approximated projection of the protein set. To carry out an unsupervised clustering of the proteins, the approach SOM (Self – Organizing Map) is a good choice [154, 155]. We computed a specific SOM that takes into account the triangular projection, “a triangular SOM”. This particular SOM is called TopKAPi, for *Triangular Kohonen map for Analyzing Proteins*. After having performed different trials with TopKAPi, we have selected a triangular network with a side of $G = 11$, i.e. a total of 66 neurons.

The interest of such an approach is to control the training to avoid a possible redundancy in the clustering. In fact, we have defined the initial PBs Z-score distributions in the network. From the previous analysis, we have selected the three proteins forming the largest triangle in the space, i.e. the sum of the distances between the three selected proteins is maximal. After locating these normalized PBs distributions at the vertices of the triangular network, we have defined the PBs distribution in each neuron by linear interpolation. The PBs distribution neurons evolve thanks to the concept of information diffusion around the winner neuron.

This choice states on a protein location evenness over the network allowing a number of proteins per neuron adequate for estimating the Z-scores. After training, we observed a median number equals to 13 proteins per neuron. The two maximums are equal to 95 and 89.

Learning step. Figure 9 shows the distribution of the proteins in the network neurons. The size of the circles is proportional to the numbers. Every neuron is characterized by an average distribution normalized in Z-scores. The 66 PB distributions are shared into five classes (labelled from G1 to G5; see the colors of Figure 9) by a hierarchical clustering. The average distributions associated with the clusters are displayed. Two subsets of proteins are easily pointed out: the

Structural alphabet

structures “all- α helix” (characterized by a high positive Z -score for PB m) and the structures “all- β sheet” (characterized by a high positive Z -score for PB c, d, e, h and i). The other clusters are intermediate groups since the Z -scores distributions are reduced in magnitude.

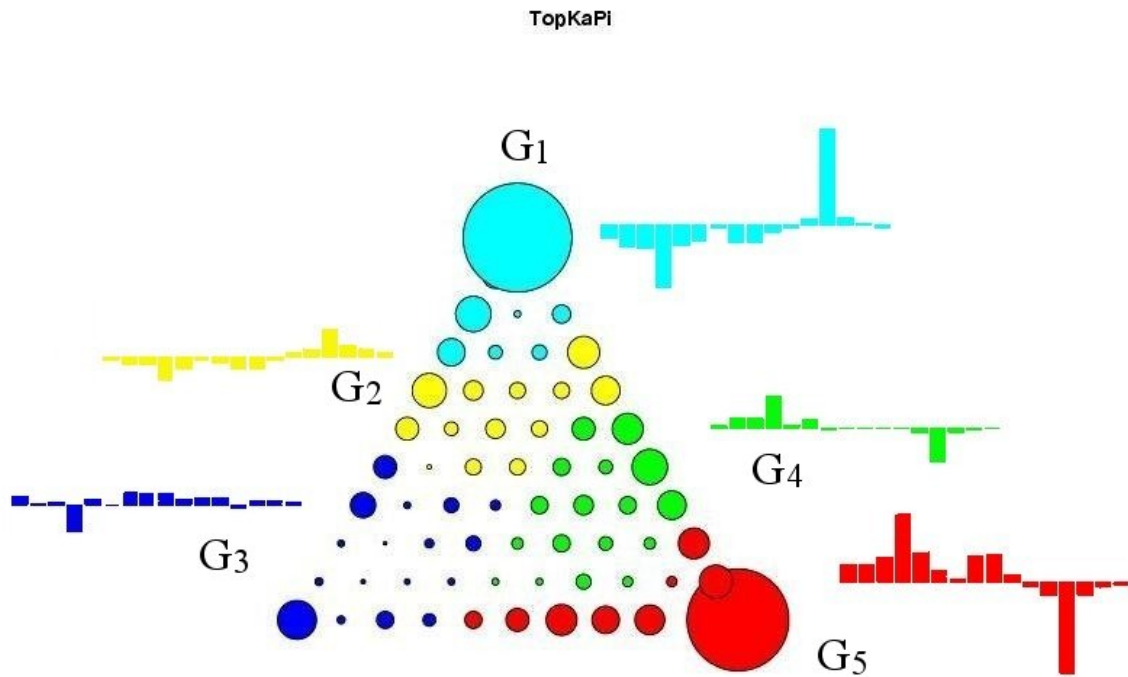


Figure 9. Final Triangular Kohonen map for Analyzing Proteins (TopKaPi) with the number of associated data for each neuron. From a hierarchical clustering, 5 clusters have been identified (*blue, red, green, yellow and cyan*) and the mean distribution of Protein Block Z -scores are indicated.

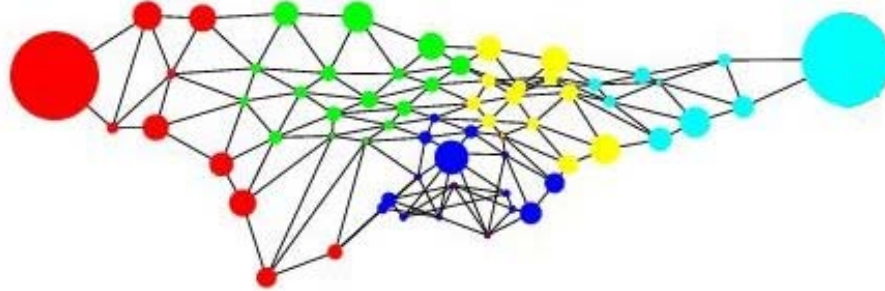


Figure 10. Sammon map of the TopKAPi neurons (cf. Figure 9).

Figure 10 gives the Sammon map [158] obtained using inter-neuron distance matrix. We observe a coherent repartition of the five clusters more linear than expected. Nevertheless, it shows the complete antagonism between all- α proteins and all- β proteins. It highlights too the interest of the triangular representation of TopKAPi which allows the emergence of the cluster colored in blue characterized by an underrepresentation of PBs d and m .

Protein Blocks distributions. Figure 11 gives the overrepresentation (in pink and red) and the underrepresentation (in yellow and blue) observed in every neuron (the grey corresponds to values close to 0). Before interpreting these figures, we have carried out a hierarchical clustering of these Z-score distributions per block. Figure 12 gives the dendrogram showing the PB associations in the protein description. Two specific protein blocks are clearly distinguishable, as expected PBs m and d which correspond to α -helix and β -strand cores, respectively.

Structural alphabet

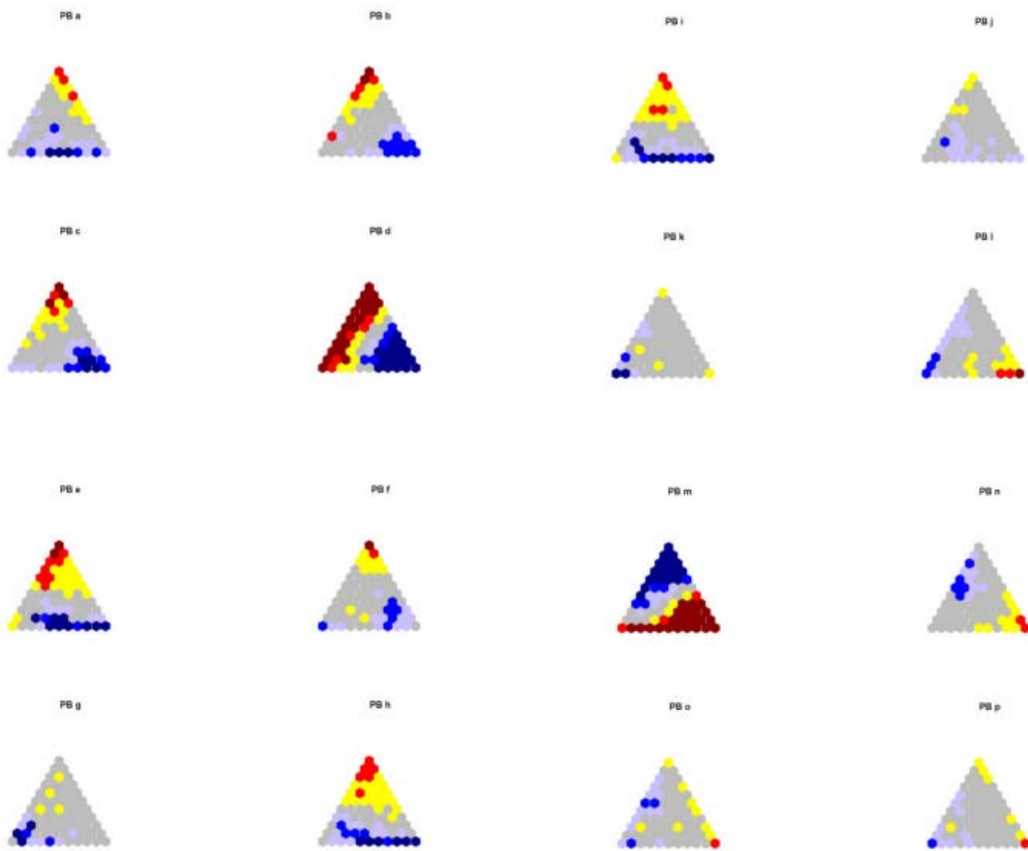


Figure 11. Distribution of the 16 Protein Blocks (PB *a* to PB *p*) in every neuron of TopKAPi (brown : Z-score < -4.4, red : Z-score within the range [-4.4; -1.96], grey : Z-score within the range [-1.96; +1.96], cyan : Z-score within the range [1.96; 4.4], red : Z-score > 4.4).

Another class of PBs is defined: $\{k, o, p, l, n\}$ associated to the N- and C-caps of the α -helix. The last class is heterogeneous; it groups the N- and C-caps of the β -strand and the coils. The PB subset $\{e, h, i\}$ is more frequently associated to β - β transition. The PB subset $\{g, a, j\}$ is more specific to coil or flexible regions. The last subset $\{f, b, c\}$ is more ambiguous containing N- and C-caps of β -strand and coils.

Structural alphabet

From the protein classification into five clusters (G_1 - G_5), previously defined, and this PB classification, we point out:

(i) class G_1 : an overrepresentation of PB m and an underrepresentation of a large subset of PBs $\{a, b, c, d, e, f, h, i\}$. The graph of PB m shows large contrast in the Z -scores. This group is specific to all- α proteins.

(ii) class G_2 : an overrepresentation of PBs located in helices and their extremities i.e. $\{l, m, n, o\}$ and an underrepresentation of PBs located in β -strands. The Z -scores are less high than in G_1 .

(iii) class G_3 : an overrepresentation of certain PBs $\{a, e, g, h, i, j, k, l, p\}$ related to protein showing little regular secondary structures, i.e. mainly composed of coils.

(iv) class G_4 : an overrepresentation of PBs b, c, d and f associated to protein showing high frequency of β -strands.

(v) class G_5 : an overrepresentation of PBs a to f, h to i and an underrepresentation of PBs l, m and n . This group is specific to “all β ” proteins associated to high Z -scores for PBs c and d .

Amino acid distribution. Figure 12 shows the Z -score distributions per amino acid. We observe a higher heterogeneity in the locations of over- and under-representations. This can be explained by the fact that the training is only based on the structure, i.e. the PB distribution.

In each protein cluster, we highlight some amino acid propensities to be located in certain protein classes such as:

(i) class G_1 : overrepresentations of L, Q, E, K with underrepresentation of C, P and G. The presence of these amino acid frequencies is a specificity of all- α proteins.

Structural alphabet

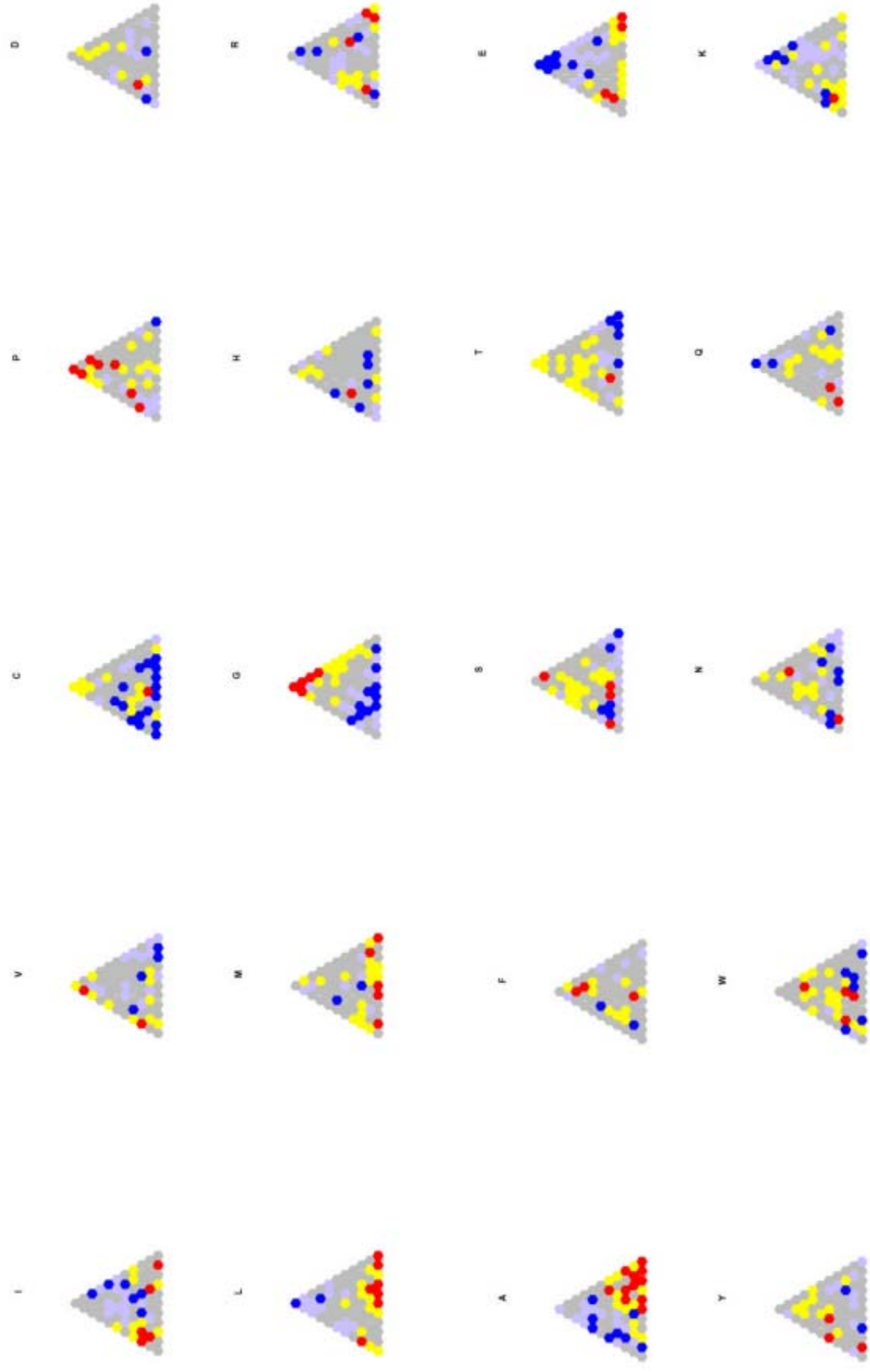


Figure 12. Distributions of the 20 types of amino acids (I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K) in every neuron of TopKAPi (brown : Z-score < 4.4, red : Z-score < -4.4, grey : Z-score within the range [-1.96; +1.96], cyan : Z-score within the range [1.96; 4.4], red : Z-score > 4.4).

Structural alphabet

(ii) class G_2 : presence of I, A, E and K. Few differences are observed between clusters G_1 and G_2 . It seems that the amino acids specific to α -helices, i.e. A and L, are not found in the same protein types.

(iii) class G_3 : this group containing low proportion of regular secondary structures shows high affinities for amino acids such as A, C, G and S.

(iv) class G_4 : this intermediate group close to the “all- β ” does not show significant relationship with the amino acids. We observe a large heterogeneity of the Z-score values.

(v) class G_5 : The graph reveals high contrast on positive Z-scores for certain amino acids, e.g. C, V, G, S and T.

This study has allowed the definition of five protein classes characterized by PB distributions and associated with under or overrepresentations of certain amino acids. This classification is not only based on the proportions of the PBs located in α -helix or β -strand, but also on the proportions of PBs located in this extremities of the regular secondary structures or in the coils. A further work associating the PB distributions and the amino acid distributions in the training should improve the definition of the protein classes.

Conclusion

Through this chapter, our aim was primarily to highlight the interest of a structural alphabet composed of local structural prototypes, i.e. the 16 Protein Blocks, to describe every part of protein 3D structures. Also, we accurately compared the 3-state secondary structure assignment with the assignment in terms of PBs. The analysis of the correspondence between the different

Structural alphabet

secondary structure assignment methods showed high discrepancies: 20% of the residues are assigned to different states. The agreement ratio between two Secondary Structure Assignment Methods (SSAMs) is highly dependent of the metrics used and the difficulties of comparing different assignment methods have been pointed out several times.

The Protein Blocks encompass most of the features of the secondary structure description with 3 or more states. They describe more precisely the repetitive structures (helical and extended), their edges and the coil state which is composed of really distinct local folds. We highlighted interesting correspondences between particular local folds and the PBs.

Use Protein Blocks prediction to classify proteins into the classical structural classes, namely all- α , all- β and mixed gave a good prediction rate, i.e. 71.5%, with no confusion between all α and all β classes. Predicting protein structural classes can be interesting for then directing the prediction in terms of PBs from sequence and for understanding the protein folding of particular protein classes.

Finally, our novel clustering approach, TopKAPi enables to classify and analyze proteins according to their Protein Block frequencies. Moreover, we have characterized some propensities of amino acids to be located in the five protein clusters previously defined.

In conclusion, the structural alphabet is a tool for analyzing local protein structures, so allowing one to work with a chain of characters rather than with carbon α coordinates. Moreover, a structural alphabet facilitates the definition of 'Structural Words', PB series of high frequencies, specific to a folding pattern. Hence, it constitutes a new way for characterizing local folds. Also, it represents an intermediate step for protein modelling in terms of the 16 PBs. The present study reveals the interest of a structural alphabet for defining the protein classes. Other multiple perspectives appear with this way of representing 3D protein structures and particularly in

Structural alphabet

molecular modelling.

Acknowledgments

We would like to thank Estelle Calvez and Maxime Huvet for previous works on TopKAPi methods, Laurent Fourrier and Aurélie Urbain for different local fold analyses, Joelle Hochez for the informatic support, Romain Gautier for a previous databank, Catherine Etchebest, Patrick Fuchs and Anne-Claude Camproux for fruitful discussions and Rachel Karchin for the Extended Beta alphabet script.

This work was supported by grants from the Ministère de la Recherche and from "Action Bioinformatique inter EPST" number 4B005F 2001-2002 and 2003-2004. AdB was supported by a grant from the Fondation de la Recherche Médicale. CB has a grant from the Ministère de la Recherche. SH is Professor at the University Paris 7 - Denis-Diderot, Paris. AdB is a researcher at the french Institute for Health and Medical Research (INSERM).

References

- [1] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Philips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666.
- [2] Pauling, L. and Corey, R.B. (1951) Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains *Proc Natl Acad Sci, USA* **37**, 235-240.
- [3] Eisenberg, D. (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci, USA* **100**, 11207-11210.

Structural alphabet

- [4] Pauling, L. and Corey, R.B. (1951) The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proc Natl Acad Sci, USA* **37**, 251-256.
- [5] Pauling, L. and Corey, R.B. (1950) Two Hydrogen-Bonded Spiral Configurations of the Polypeptide Chain. *J. Am. Chem. Soc* **72**, 53.
- [6] Ho, B.K., Thomas, A. and Brasseur, R. (2003) Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Prot Sci.* **12**, 2508-2522.
- [7] Colloc'h, N. and Cohen, F.E. (1991) Beta-breakers: an aperiodic secondary structure. *J Mol Biol.* **221**, 603-613.
- [8] Regan, L. (1994) Protein structure. Born to be beta. *Curr Biol.* **4**, 656-658.
- [9] Richardson, J.S. and Richardson, D.C. (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci, USA* **99**, 2754-2759.
- [10] Richardson, J.S. (1976) Handedness of Crossover Connections in Beta Sheets. *Proc Natl Acad Sci, USA* **73**, 2619-2623.
- [11] Chotia, C., Levitt, M. and Richardson, D. (1977) Structure of Proteins: Packing of Alpha-Helices and Pleated Sheets. *Proc Natl Acad Sci, USA* **74**, 4130-4134.
- [12] Chotia, C., Levitt, M. and Richardson, D. (1981) Helix to helix packing in proteins. *J Mol Biol* **145**, 215-250.
- [13] Efimov, A.V. (1991) Structure of coiled beta-beta-hairpins and beta-beta-corners. *FEBS Lett.* **284**, 288-292.
- [14] Aurora, R., Srinivasan, R. and Rose, G.D. (1994) Rules for alpha-helix termination by glycine. *Science* **264**, 1126-1130.
- [15] Hutchinson, E and Sessions, RB and Thornton, JM and Woolfson, DN (1998) Determinants of strand register in antiparallel beta-sheets of proteins, *Protein Sci*, **7**, 2287-2300.
- [16] Aurora, R. and Rose, G.D. (1998) Helix Capping. *Protein Sci.* **7**, 21-38.

Structural alphabet

- [17] Liu, W.M. and Chou, K.C. (1998) Singular points of protein beta-sheets. *Protein Sci.* **7**, 2324-2330.
- [18] Donohue, J. (1953) Hydrogen bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* **39**, 470– 478.
- [19] Pal, L. and Basu, G. (1999) Novel protein structural motifs containing two-turn and longer 3(10)-helices. *Protein Eng.* **12**, 811-814.
- [20] Pal, L., Basu, G. and Chakrabarti, P. (2002) Variants of 3₁₀-helices in proteins. *Proteins* **48**, 571-579.
- [21] Low, B.W. and Baybutt, R.B. (1952) The π -helix -A hydrogen bonded configuration of the polypeptide chain. *J Am Chem Soc* **74**, 5806.
- [22] Low, B.W. and Greenville-Wells, H.J. (1953) Generalized mathematical relationships for polypeptide chain helices. The coordinates of the π -helix. *Proc Natl Acad Sci USA* **39**, 785–801.
- [23] Weaver, T.M. (2000) The pi-helix translates structure into function. *Protein Sci.* **9**, 201-206.
- [24] Fodje, M.N. and Al-Karadaghi, S. (2002). Occurrence, conformational features and amino acid propensities for the π -helix. *Protein Eng.* **15**, 353-358.
- [25] Sudha, R., Kohtani, M., Breaux, G.A. and Jarrold, M.F. (2004) π -Helix Preference in Unsolvated Peptides. *J. Am. Chem. Soc.* **126**, 2777-2784.
- [26] Armen, R., Alonso, D.O. and Daggett, V. (2003) The role of alpha-, 3(10)-, and pi-helix in helix-->coil transitions. *Protein Sci.* **12**, 1145-1157.
- [27] Cartailier, J.P. and Luecke, H. (2004) Structural and functional characterization of pi bulges and other short intrahelical deformations. *Structure* **12**, 133-144.
- [28] Pauling, L. and Corey, R.B. (1951) The Structure of Fibrous Proteins of the Collagen-

Structural alphabet

- Gelatin Group. *Proc Natl Acad Sci, USA*. **37**, 272-281.
- [29] Cowan, P.M. and McGavin, S. (1955) Structure of poly-L-proline. *Nature* **176**, 501-503.
- [30] Adzhubei, A.A. and Sternberg M.J. (1993) Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol*. **229**, 472-493.
- [31] Stapley, B.J. and Creamer, T.P. (1999) A survey of left-handed polyproline II helices. *Protein Sci*. **8**, 587-595.
- [32] Creamer, T.P. (1998) Left-handed Polyproline II helix formation is (very) locally driven. *Proteins* **33**, 218-226.
- [33] Creamer T.P. and Campbell, M.N. (2002) Determinants of the polyproline II helix from modeling studies. *Adv Protein Chem* **62**, 263-282.
- [34] Eswar, N., Ramakrishnan C. and Srinivasan, N (2003) Stranded in isolation: structural role of isolated extended strands in proteins. *Protein Eng*. **16**, 331-339.
- [35] Sowdhamini, R., Srinivasan, N., Ramakrishnan, C. and Balaram P. (1992) Orthogonal beta beta motifs in proteins. *J Mol Biol* **223**, 845-851.
- [36] Richardson, J.S., Getzoff, E.D. and Richardson, D.C. (1978) The beta bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci USA* **75**, 2574-2578.
- [37] Chan, A.W., Hutchinson, E.G., Harris, D. and Thornton, J.M. (1993). Identification, classification, and analysis of β -bulges in proteins. *Protein Sci*. **2**, 1574-1590.
- [38] Axe, D.D., Foster, N.W. and Fersht, A.R. (1999) An irregular beta-bulge common to a group of bacterial RNases is an important determinant of stability and function in barnase. *J Mol Biol* **286**, 1471-1485.
- [39] Ramachandran, G. N. and Sasisekharan, V. (1968) Conformation of polypeptides and proteins *Advan. Protein Chem*. **23**, 283.

Structural alphabet

- [40] Venkatachalam, C.M. (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **10**, 1425-1436.
- [41] Richardson, J.S., Getzoff, E.D. and Richardson, D.C. (1978). The β bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci, USA* **75**, 2574-2578.
- [42] Némethy, G. and Printz, M.P. (1972) The gamma turn, a possible folded conformation of the polypeptide chain. Comparison with the beta turn. *Macromolecules* **5**, 755-758.
- [43] Matthews, B.W. (1972) the gamma-turn. Evidence for a new folded conformation in Proteins. *Macromolecules* **5**, 818-819.
- [44] Milner-White, E.J. (1988) Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with α -helices and β -bulge loops. *J Mol Biol.* **199**, 503-511.
- [45] Rose, G.D., Gierasch, L.M. and Smith, J.A. (1985) Turns in peptides and proteins. *Adv. Prot. Chem.* **37**, 1-109.
- [46] Lewis, P.N., Momany, F.A. and Scheraga, H.A. (1971) Folding of polypeptide chains in proteins: a proposed mechanism for folding. *Proc Natl Acad Sci USA.* **68**, 2293-2297.
- [47] Kuntz, I.D. (1972) Protein folding. *J Am Chem Soc.* **94**, 4009-4012.
- [48] Lewis, P.N., Momany, F.A. and Scheraga, H.A. (1973) Chain reversals in proteins. *Bioch Biophys Acta* **303**, 211-229.
- [49] Chou, P.Y. and Fasman, G.D. (1977) Beta-turns in proteins. *J Mol Biol* **115**, 135-175.
- [50] Rose, G.D. and Seltzer, J.P. (1977) A New algorithm for finding the peptide chain turns in a globular protein. *J Mol Biol* **113**, 153-164,
- [51] Rose, G.D. and Wetlaufer, D.B. (1977) The number of turns in globular proteins, *Nature* **5622**, 769-770.

Structural alphabet

- [52] Zimmermann, S.S. and Scheraga, H.A. (1977) Influence of local interactions on protein structure. I. Conformational energy studies of N-acetyl-N'-methylamides of Pro-X and X-Pro dipeptides. *Biopolymers* **16**, 811-843.
- [53] Richardson, J.S. (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**, 167-339.
- [54] Wilmot, C.M. and Thornton, J.M. (1988) Analysis and prediction of the different types of β -turn in proteins. *J Mol Biol* **5**, 221-232.
- [55] Hutchinson, E.G. and Thornton, J.M. (1996) PROMOTIF - A program to identify structural motifs in proteins. *Protein Sci.* **5**, 212-220.
- [56] Ball, S.G. and Hughes, AS (1993) beta-turn topography. *Tetrahedron* **49**, 3467-3478.
- [57] Ashish, X., Grover, A. and Kishore, R. (2000) Characterization of a novel type VII beta-turn conformation for a bio-active tetrapeptide rigin. *Eur J Biochem* **267**,1455-1463.
- [58] Chou, K.C. (2000) Prediction of tight turns and their types in proteins. *Anal. Biochem.* **286**, 1-16.
- [59] Pavone, V., Gaeta, G., Lombardi, A., Nastri, F., Maglio, O., Isernia, C. and Saviano, M. (1996) Discovering protein secondary structures: classification and description of isolated α -turns. *Biopolymers* **38**, 705-721.
- [60] Kaur, H. and Raghava, G.P.S. (2004) Prediction of α -turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins* **55**, 83-90.
- [61] Milner-White, E.J., Ross, B.M., Ismail, R., Belhadj-Mostefa, K. and Poet R. (1988) One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. *J Mol Biol.* **204**, 777-782.
- [62] Rajashankar, K.R., and Ramakumar, S. (1996). π -turns in proteins and peptides:

Structural alphabet

- classification, conformation, occurrence, hydration and sequence. *Protein Sci.* **5**, 932-946.
- [63] Guruprasad, K. and Rajkumar, S. (2000) beta- and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *J Biosci* **25**, 143-156.
- [64] Guruprasad, K., Prasad, M.S. and Kumar, G.R. (2000) Analysis of gammabeta, betagamma, gammagamma, betabeta multiple turns in proteins, *J Pept Res* **56**, 250-263.
- [65] Guruprasad, K., Rao, M.J., Adindla, S., and Guruprasad L. (2003) Combinations of turns in proteins. *J Pept Res.* **62**, 167-174.
- [66] Leszczynski, J.F. and Rose, G.D. (1986) Loops in globular proteins: a novel category of secondary structure. *Science* **234**, 849-855.
- [67] Fetrow, J.S. (1995) Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J.* **9**, 708-717.
- [68] Pal, M. and Dasgupta, S. (2003) The nature of the turn in omega loops of proteins, *Proteins* **51**, 606-616.
- [69] Fetrow, J.S., Cardillo, T.S. and Sherman, F. (1989) Deletions and replacements of omega loops in yeast Iso-1-cytochrome c. *Proteins* **6**, 372-381.
- [70]. Krishna, M.M.G., Lin, Y., Rumbley, J.N. and Englander, S.W. (2003) Cooperative Omega loops in Cytochrome c: Role in folding and function. *J Mol Biol* **331**, 29-36.
- [71] Sibanda, B.L. and Thornton, J.M. (1985) Beta-hairpin families in globular proteins. *Nature* **316**, 170-174.
- [72] Milner-White, E.J. and Poet, R. (1986) Four classes of beta-hairpins in proteins. *Biochem J.* **240**, 289-292.
- [73] Sibanda, B.L., Blundell, T.L. and Thornton, J.M. (1989) Conformation of beta-hairpins in

Structural alphabet

- protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol.* **206**, 759-777.
- [74] Sibanda, B.L. & Thornton, J.M. (1991). Conformation of β hairpins in protein structures: classification and diversity in homologous structures. *Methods Enzymol.* **202**, 59-82.
- [75] Sibanda, B.L. and Thornton, J.M. (1993) Accommodating sequence changes in beta-hairpins in proteins. *J Mol Biol* **229**, 428-447.
- [76] Gunasekaran, K., Ramakrishan, C. and Balaram, P. (1997) Beta-hairpins in proteins revisited: lessons for de novo design. *Protein Eng.* **10**, 1131-1141.
- [77] Blandl, T., Cochran, A.G. and Skelton N.J. (2003) Turn stability in beta-hairpin peptides: Investigation of peptides containing 3:5 type I G1 bulge turns. *Protein Sci.* **12**, 237-247.
- [78] Kim, J., Brych, S.R., Lee, J., Logan, T.M., Blaber, M. (2003) Identification of a key structural element for protein folding within beta-hairpin turns. *J Mol Biol* **328**, 951-961.
- [79] Efimov A.V. (1984) A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. *FEBS Lett* **166**, 33-38.
- [80] Rice, P.A., Goldman, A. and Steitz, T.A. (1990) A helix-turn-strand structural motif common in α - β proteins. *Proteins* **8**, 334-340
- [81] Tang, J., James, M.N., Hsu, I.N., Jenkins, J.A. and Blundell, T.L. (1978) Structure evidence for gene duplication in the evolution of the acid protease. *Nature* **271**, 618-622.
- [82] Wintjens, R.T., Rooman, M.J. and Wodak, S.J. (1996) Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.* **255**, 235-253.
- [83] Wintjens, R.T., Rooman, M.J. and Wodak, SJ (1998) Typical interaction patterns in alphabeta and betaalpha turn motifs. *J Mol Biol* **11**, 505-522.
- [84] Boutonnet, N.S., Kajava, A.V. and Rooman, M.J. (1998) Structural classification of

Structural alphabet

- alphabetabeta and betabetaalpha supersecondary structure. *Proteins* **30**, 193-212.
- [85] Kwasigroch, J.-M., Chomilier, J. and Mornon, J.-P. (1996) A global taxonomy of loops in globular proteins. *J Mol Biol.* **259**, 855-872.
- [86] Geetha, V. and Munson, P.J. (1997) Linkers of secondary structures in proteins. *Protein Sci.* **6**, 2538-2547.
- [87] Tramontano, A., Chothia, C. and Lesk, A.M. (1989) Structural determinants of the conformations of medium-sized loops in proteins. *Proteins.* **6**, 382-394.
- [88] Donate, L.E., Rufino, S.D., Canard, L.H.J. and Blundell T.L. (1996) Conformational analysis and clustering of short and medium size loops connecting secondary structures: a database for modeling and prediction. *Protein Sci.* **5**, 2600-2616.
- [89] Li, W., Liu, Z. and Lai, L. (1999). Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers.* **49**, 481-495.
- [90] Ring, C.S., Kneller, D.G., Langridge, R. and Cohen, F.E. (1992). Taxonomy and conformational analysis of loops in proteins. *J Mol Biol.* **224**, 685-699.
- [91] Fichteler, T., Dengler, U. and Schomburg, D. (1995), Prediction of protein three-dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. *J Mol Biol* **267**, 114-131.
- [92] Rufino, S.D., Donate, L.E., Canard, L.H.J. and Blundell, T.L. (1997) Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J Mol Biol.* **267**, 352-367.
- [93] Vlijmen, H.W.T and Karplus, M. (1997) PDB-based protein loop prediction: parameters for selection and methods of optimization. *J Mol Biol.* **267**, 975-1001.
- [94] Wojcik, J., Mornon, J.-P. and Chomilier, J. (1999). New efficient statistical sequence-

Structural alphabet

- dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol.* **289**,1469-1490.
- [95] Fiser, A., Do, R.K.G. and Sali, A (2000) Modeling of loops in protein structures. *Prot Sci.* **9**, 1753-1773.
- [96] Fidelis, K., Stern, P.S., Bacon, D. and Moulton, J. (1994) Comparison of systematic search and database methods for constructing segments of protein structure. *Prot Eng.* **7**, 953-960.
- [97] Oliva, B., Bates, P.A., Querol, E., Aviles, F.X. and Sternberg, M.J.E. (1997) An automated classification of the structure of protein loops. *J Mol Biol.* **266**, 814-830.
- [98] Michalsky, E., Goede, A. and Preissner, R. (2003) Loops In Proteins (LIP)--a comprehensive loop database for homology modeling. *Protein Eng.* **16**, 979-985.
- [99] Bansal, M., Kumar, S. and Velavan, R. (2000). HELANAL - A program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn* **17**, 811-819.
- [100] Levitt, M. and Greer, G. (1977) Automatic identification of secondary structure in globular proteins. *J Mol Biol.* **114**, 181-293.
- [101] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
- [102] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol.* **112**, 535-540.
- [103] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.
- [104] Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins.* **23**:566-579.
- [105] Andersen, C.A.F, Palmer, A.G.H., Brunak, S. and Rost, B. (2002) Continuum secondary

Structural alphabet

- structure captures protein flexibility. *Structure* 10, 175-154.
- [106] Andersen, C.A.F. and Rost, B. (2002) Secondary structure assignment in Structural Bioinformatics, Bourne (ed), pp. 341-364.
- [107] Carter, P., Andersen, C.A.F. and Rost, B. (2003) DSSPcont: continuous secondary structure assignments for proteins. *Nucleic Acids Res.* **31**, 3293-3295.
- [108] Richards, F.M. and Kundrot, C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure, *Proteins*. **3**, 71-84.
- [109] Sklenar, H., Etchebest, C. and Lavery, R. (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **6**, 46-60.
- [110] Labesse, G., Colloc'h, N., Pothier, J. and Mornon, J.-P. (1997) PSEA: a new efficient assignment of secondary structure from C α trace of proteins. *Comput Appl Biosci.* **13**, 291-295.
- [111] King, S.M. and Johnson, WC. (1999). Assigning secondary structure from protein coordinate data. *Proteins*. **35**, 313-320.
- [112] Barlow, D.J. and Thornton, J.M. (1988) Helix geometry in proteins. *J Mol Biol.* **201**, 601-619.
- [113] Bansal, M., Kumar, S. and Velavan, R. (2000) HELANAL: a program to characterize helix geometry in proteins. *J Bio Struct Dyn.* **17**, 811-820.
- [114] Woodcock, S., Mornon, J.-P. and Henrissat, B. (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Prot Eng.* **5**, 629-635.
- [115] Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. and Mornon, J.-P. (1993)

Structural alphabet

- Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* **6**, 377-382.
- [116] de Brevern, A.G., Camproux, A.C., Hazout, S., Etchebest, C. and Tuffery, P. (2001) Beyond the secondary structures : the structural alphabets, in Recent Adv In Prot Eng., Sangadai SG (ed). Research signpost, Trivandrum,India, pp. 319-331.
- [117] Karchin R. (2003). Evaluating local structure alphabets for protein structure prediction. Ph.D. Computer Science.
- [118] Unger, R., Harel, D., Wherland, S. and Sussman, J.L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins.* **5**, 355-373.
- [119] Prestelski, S.J., Williams Jr., A.L. and Liebman, M.N. (1992) Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results. *Proteins* **14**, 430-439.
- [120] Unger, R. and Sussman J.L. (1993). The importance of short structural motifs in protein structure analysis. *J Comput Aid Mol Des* **7**, 457-472.
- [121] Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D. and Wrede, P. (1996). Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng.* **9**, 833-842.
- [122] Park B.H. and Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J Mol Biol.* **249**, 493-507.
- [123] Kolodny R., Koehl, P., Guibas, L., Levitt, M. (2002) Small libraries of protein fragments model native protein structures accurately. *J Mol Biol.* **323**, 297-307.
- [124] Micheletti, C., Seno, F. and Maritan, A. (2000), Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* **40**, 662-674.

Structural alphabet

- [125] Rooman, M.J., Rodriguez, J., and Wodak, S.J. (1990) Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol.* **213**, 327-336.
- [126] Fetrow, J.S., Palumbo, M.J., and Berg, G. (1997) Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins.* **27**, 249-271.
- [127] Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motif. *J Mol Biol.* **281**, 565-577.
- [128] de Brevern, A.G., Etchebest, C. and Hazout, S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins.* **41**, 271-287.
- [129] Camproux, A.C., Tuffery, P., Buffat, L., Andre, C., Boisvieux, J.F. and Hazout, S. (1999) Using short structural building blocks defined by a Hidden Markov Model for analysing patterns between regular secondary structures. *Theor. Chem. Acc.* **101**, 33-40.
- [130] Camproux, A.C., de Brevern, A.G., Hazout, S. and Tuffery, P. (2001). Exploring the use of a structural alphabet for a structural prediction of protein loops. *Theor Chem Acc.* **106**(1/2), 28-35.
- [131] de Brevern, A.G. and Hazout, S (2001). Compacting local protein folds by a "Hybrid Protein Model". *Theor Chem Acc.* **106**(1/2): 36-47.
- [132] de Brevern A.G. and Hazout S. (2000), Hybrid Protein Model (HPM) : une nouvelle approche pour caractériser les relations séquence-structure dans les protéines. *Premières Journées Ouvertes de Biologie, Informatique et Mathématiques. Recueil des Actes*, pp.105-112.
- [133] de Brevern, A.G., Valadié, H., Hazout, S. & Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. *Protein Sci.* **11**, 2871-2886.

Structural alphabet

- [134] de Brevern A.G. and Hazout S. (2001) Compactage d'une base de données protéiques recodées dans un alphabet structural. *Secondes Journées Ouvertes de Biologie, Informatique et Mathématiques. Recueil des Actes*, pp.85-92.
- [135] de Brevern, A.G. and Hazout, S. (2003). Improvement of "Hybrid Protein Model" to define an optimal repertory of contiguous 3D protein structure fragments. *Bioinformatics*. **19**, 345-353.
- [136] Benros, C., Hazout S. and de Brevern, A.G. (2002) "Hybrid Protein Model": a new clustering approach for 3D local structures. *Proceedings of Workshop on Bioinformatics ISMIS*. Chapter 5, pp. 1-6.
- [137] Benros, C., de Brevern A.G. and Hazout S. (2003) Hybrid Protein Model (HPM): A method for building a library of overlapping local structural prototypes. sensitivity study and improvements of the training. *IEEE Int Work. NNSP 2003*, 53-70.
- [138] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E.M., and Baker, D. (2001) Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins* **37**(S5):119-126.
- [139] Bonneau, R., and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct.* **30**, 173-189.
- [140] Karchin, R., Cline, M., Mandel-Gutfreund, Y. and Karplus, K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*. **51**, 504-514.
- [141] Thornton, J.M., Sibanda, B.L., Edwards, M.S. and Barlow, D.J. (1988) Analysis, design and modification of loop regions in proteins. *Bioessays* **8**, 63-69.
- [142] Westhead, D.R., Slidel, T.W.F., Flores, T. P. J. and Thornton J. M. (1999) Protein structural topology: automated analysis, diagrammatic representation and database

Structural alphabet

- searching. *Protein Sci* **8**, 897-904.
- [143] Murzin A. G., Brenner S. E., Hubbard T. and Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* **247**, 536-540.
- [144] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108.
- [145] Day, R., Beck, D.A., Armen, R.S. and Daggett, V. (2004) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* **12**, 2150-2160.
- [146] Bray, J.E., Todd, A.E., Pearl, F.M.G., Thornton, J.M. and Orengo, C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues *Prot Eng.* **13**, 153-165.
- [147] Pearl, F.M.G, Lee, D., Bray, J.E, Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28**, 277-282.
- [148] Michie, A.D., Orengo, C.A. and Thornton, JM. (1996) Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol.* **262**, 168-185.
- [149] Noguchi, T., Matsuda, H. and Akiyama, Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.* **29**, 219-220.
- [150] Hobohm, U., Scharf, F., Schneider R. and Sander, C. (1992) Selection of a representative set of structures from the Brookhaven Protein Databank. *Prot Sci.* **1**, 409-417.
- [151] Hobohm, U. and Sander, C. (1994) Enlarged representative set of protein structures. *Prot*

Structural alphabet

Sci. **3**, 522-524.

- [152] Wang, G. and Dunbrack, R.L., Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.
- [153] Brenner, S.E., Koehl, P. and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucl Acids Res.* **28**, 254-256.
- [154] Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69.
- [155] Kohonen, T. (2001) Self-Organizing Maps, 3rd edition. Springer-Verlag, Berlin, Germany.
- [156] Rabiner, L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE.* **77**, 257-285.
- [157] Mardia, K.V., Kent, T. and Bibby, J.M. (1979) Multivariate Analysis, Academic Press.
- [158] Sammon Jr. (1969) A non-linear mapping for data structure analysis. *IEEE Transactions on Computers C* **18**, 401-409.
- [159] Agrafiotis, D.K. (1997) A new method for analyzing protein sequence relationships based on Sammon maps. *Prot Sci.* **6**, 287-293.
- [160] Andrade, M.A., Casari, G., Sander, C. and Valencia, A. (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol Cyber.* **74**, 441-450.
- [161] Bystroff, C., Thorsson, V. & Baker, D. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol.* **301**:173-190.
- [162] Koradi, R., Billeter, M. and Wuthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph.* **14**, 29-32.