*Protein Structure Prediction,* Edited by Huzefa Rangwala and George Karypis.
ISBN 0-471-XXXXX-X 

# Chapter 5

# Local Structure Alphabets

***Agnel Praveen Joseph, Aurélie Bornot & Alexandre G. de Brevern***

INSERM UMR-S 665 ; Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB) ; Université Paris Diderot - Paris 7 ; Institut National de la Transfusion Sanguine (INTS) ; 6, rue Alexandre Cabanel, 75739 Paris cedex 15 - FRANCE

## Abstract

Protein structures are classically described in terms of secondary structures, *i.e.*, two regular states, the α-helices and the β-strands and one default state, the coil. Even if the regular secondary structures have relevant physical meaning, the definition of secondary structures has some important (and often forgotten) limitations: the rules for secondary structure assignments are (i) not simple, (ii) not unique and (iii) 50% of all residues, which occur in the coil, are not described. Hence, different research groups have described local protein structures with the aim of analyzing them and to approximate every part of the protein backbone. These libraries of local structures consist of sets of small prototypes named "structural alphabets". They have also been used to predict the protein backbone conformation. In this chapter, we first present the secondary structures, *i.e.*, the most classical approach to describe protein structures, followed by the different structural alphabets designed till date. We focus on the different prediction schemes developed with these structural alphabets.

# Introduction

Proteins play a crucial key role in most of the cellular processes. They act as enzymes, transcription factors, mediators in cell signalling, transporters, storage molecules or have structural, regulatory or protective roles. Many diseases are associated with abnormality in protein functions. At this day, proteins are also the most important drug targets. The protein three-dimensional (3D) structure is directly dependent on its biological function. So a good understanding of 3D structure often gives sufficient hints to understand the protein functions and this forms the basis of structure based drug design [1]. Only about a percent of the total number of sequenced proteins have experimentally determined structures [2] and a considerable number of these proteins are without known functions [3]. Considering the fact that the amino acid sequence of a protein determines its 3D structure, one often tries to extract the structural information embedded in the sequence.

Even before the first protein structure was solved, Pauling and Corey proposed two major repetitive structures than could occur within protein structures: the $\alpha$-helix and the $\beta$-sheet [4,5]. Since then, these repetitive structures are not only being used to analyze the protein structures, but also to predict them. Nonetheless, this description has some limitations that have lead to the definition of a more complex concept of structural alphabets. Here, we will present the secondary structures and the different structural alphabets designed at this day.

## Repeating structural elements in proteins

A number of repeating structural elements have been observed in the known protein structures. Representing proteins in terms of secondary structures like helices and strands, is known to be useful for visualisation, prediction, classification and analysis of protein structures [6-8]. Several methods for assigning secondary structures and other repeating elements (discussed in the following paragraphs), have been developed. Methods like DSSP [9] or STRIDE [10] use the information on the hydrogen bonding patterns to characterize these secondary structures. PROSS [11] and SEGNO [12] uses torsion angle information for assignments while others [13], use the inter Cα distances either alone or along with the information on the hydrogen bonding pattern and dihedral angles, for assigning secondary structures.

***Classical secondary structures.*** The classical way of describing protein structures is in terms of alpha helices and beta sheets, the two major repetitive local structures in proteins [14]. These repeating units are characterized by the pattern of hydrogen bonds formed by the protein backbone. α-helices involve hydrogen bonds between $i^{th}$ and $i+4^{th}$ residues while β-sheets are composed of extended strands with hydrogen bonds formed between adjacent strands. β-sheets help to bring together parts of protein that are far apart in the sequence, while helices involve consecutive residues in a sequence. The planar arrangement of beta strands gives rise to steric constraints that cause consecutive side-chains to point in opposite sides of the plane.

Analysis of sequence-structure relationships has shown over- and under-representations of certain amino acids. Richardson and Richardson and Pal and et al. have

made a detailed analysis and shown that short and long helices have different amino acid compositions [15,16]. The sequence specificities of beta strands have also been studied [17] as of their ends [18]. Experimental and statistical works on analysis of specificity of pairs of interacting residues in neighbouring strands have given limited results and failed to present and pertinent laws for their associations. The recent studies mainly focus on the crucial question of protein aggregation [19]. Analysis of helix signals in proteins highlighted the hydrophobic capping, a hydrophobic interaction that straddles the helix terminus and is always found to be associated with hydrogen-bonded capping [18,20,21].

During the seventies, predictions of regular secondary structures have been carried out using statistical approaches [22]. The introduction of Artificial Neural Networks coupled with evolutionary information has led to an impressive increase in the prediction rate, *e.g*. PHD methodology [23]. The secondary structure prediction rate has reached a maximum limit that is slightly better than 80%. The two most widely used programs are PSI-PRED [24] and SSPRO [25,26]. No new significant improvements have been seen during the last few years. It is considered that the secondary structure prediction is no more a research area that can be really improved.

***Other Helical and Extended conformations.*** Several other repeating structural elements are also observed (see Figure 1 for some examples). Apart from $\alpha$-helices, other helical states like $3_{10}$ and $\pi$ are also found, covering around 4% and 0.02% of residues respectively. $3_{10}$-helices are characterized by inter-residue hydrogen bonds between $i^{th}$ and $i$+$3^{th}$ residues. Majority of $3_{10}$-helices involve only one turn [27]. They are usually found at the termini of a $\alpha$-helices, often link two alpha helices like what is observed in

hairpins and corner motifs [28]. π-helices involve inter-residue hydrogen bonds between $i^{th}$ and $i+5^{th}$ residues. Dynamic transitions between alpha and $3_{10}$- and α- and π-helices have been proposed to occur during the folding and unfolding process [29]. As shown with Figure 1, these different helices are short and thus difficult to assign precisely. For instance, the $3_{10}$ helix shown is the only one assigned by DSSP and PROSS, each ones assigned the other $3_{10}$ helices of this cytochrome as coil or turn (see [13,18,30]for more details).

Isolated extended structures that are not part of a β-sheet are also found in proteins and they are generally exposed to solvent [31]. SSPRO8 has the potentiality to predict them. However due to low occurrences, the prediction of π-helices or isolated extended structures becomes difficult.

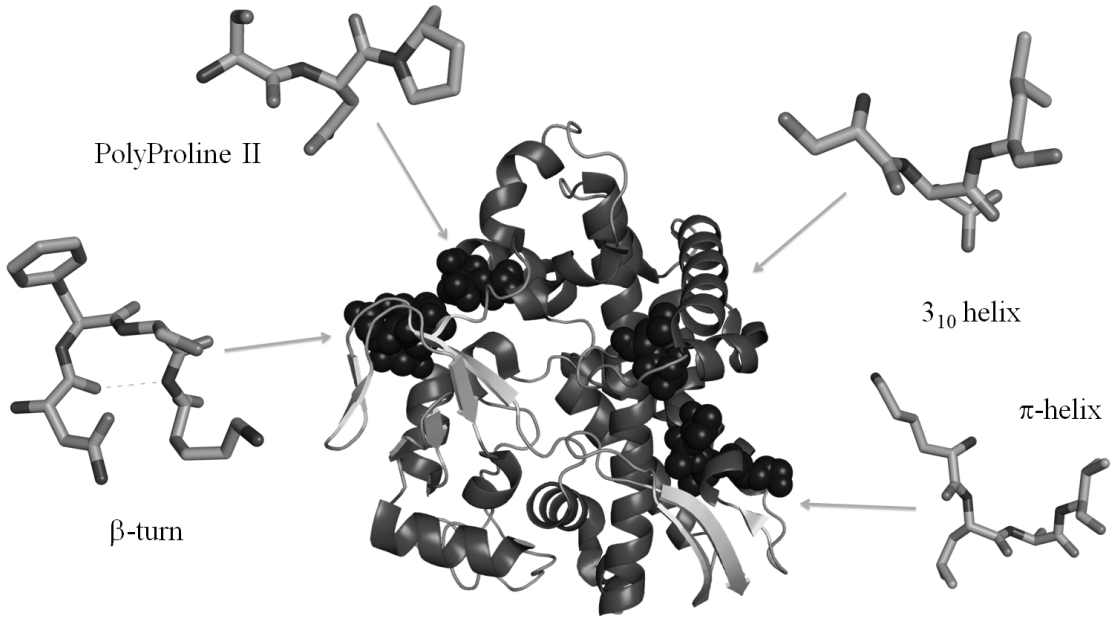


**Figure 1.** *Some less common 'secondary structures'.* A cytochrome P450 (PDB code 1IO7 [32]) has been assigned using DSSP and PROSS. One $3_{10}$ helix has been assigned by both approaches (positions 148-150) while PROSS is the only one to have assigned one π-helix (positions 120-123). PROSS has also assigned one Polyproline II (positions 71-73). Different b-turns have also been located; the one represented encompassed the amino acid from positions 35 to 38. Visualisation has been done using PyMol [33].

***Turns.*** The first description and analysis of turns has been made by Venkatachalam [34]. The turns correspond to a short return of the protein backbone. It is the third most studied secondary structure. A turn with *n* residues has a distance less than 7 Å between the Cα carbons of residues *i* and *i*+*n*. Also, the central residues are not helical and at least one residue must not be extended. There are four different types of tight turns: γ-turns (3 residues), β-turns (4), α-turns (5) and π-turns (6 residues). Each of these is further classified into different types based on the φ / ϕ dihedral angles. γ- and β–turns are the most widely studied types of turns. About 25 to 30 % of residues correspond to β-turns. Till date, seven types of β-turns have been characterized [35]. As seen on Figure 1, the β-turn can be easily confounded within a helical structure, *e.g*. α-helix. Moreover, they are often multiple, *i.e*. successive β-turns overlap.

The first secondary structure prediction method was also dedicated to predict the β-turns [22]. However, due to the difficulty associated with its prediction, the secondary structure prediction had been rapidly limited to the prediction of α-helix, β-sheet and coil. Nowadays, the prediction of β-turns is done mainly after a prediction of three-states secondary structures, as in PSI-PRED [24] or the method based on statistical approaches [36] or advanced classifiers like Support Vector Machines [37]. Prediction accuracy of turns is nowadays quite acceptable; however the prediction of some rarely seen turns remains low [36]. Very recently Klebe's group had done a new learning of the 'turns' to define a novel classification of open and hydrogen-bond turns [38,39]. They also developed a prediction method.

***Polyproline II.*** Polyproline II helices (PII) are left handed helical structures that help in the formation of coiled-coils in fibrous proteins [40]. The left handedness is characterized by specific dihedral angles and trans-isomers of peptide bonds. The φ / ϕ dihedral angles (approximately -75° and 145° respectively) fall in the region that is characteristic of β-strand. These helices are often solvent exposed and also associated with high temperature factors [41]. Non-local interactions suggest a prominent role for PII helices in protein-protein and protein-ligand interactions [42,43]. It must be noticed that PII can exist without any Proline. For instance, the only Polyproline observed within this cytochrome P450 contains only one Proline (see Figure 1). So it has been noted that designation PII is a bit misleading, since the conformation is not just associated with Pro but can be adopted by all amino acids. In a recent and fine study, about one-third of the residues in the center of PII tripeptides are Pro; the rest include all types of amino acids. The authors proposed that the common name could be changed to a more general "polypeptide-II" conformation [44]. Only PROSS [11], XTLSSTR [45] and SEGNO [12] are capable of PII assignment, it is not the case for instance of DSSP [9], STRIDE [10], P-SEA [46], VOTAP [47] or PROSIGN [48]. To the best of our knowledge our knowledge, only one group had recently developed prediction methods of PII [49].

***Loops.*** Even after performing helical, strand and turn assignments, about 50% of the residues are left out and are associated to the coil state. Thus different classification approaches have been developed to analyze the regions connecting repetitive structures. β-hairpins are the most studied type of specific loops, thanks to their high frequency of occurrence. They connect two adjacent anti-parallel beta strands. They are grouped into

different classes based on their length and conformation. Other types of loops joining beta strands like the β-β corners and orthogonal β-β motifs have also been studied. Characteristic sequence patterns are often observed in the strand-loop-strand motifs and some dedicated prediction strategies based on neural networks, have been developed [50-52]. Prediction rates of β-hairpins go up to 80%, leading to an overall prediction rate of 65% for the four states [52]. α-α turns and corners have also been studied extensively [53]. Complete loop regions have also been analysed. Most of these studies are focussed on loops of length less than nine residues leading to some classifications [54]. ArchDB is an online method available to find potential compatible loops [55].

## Beyond secondary structures

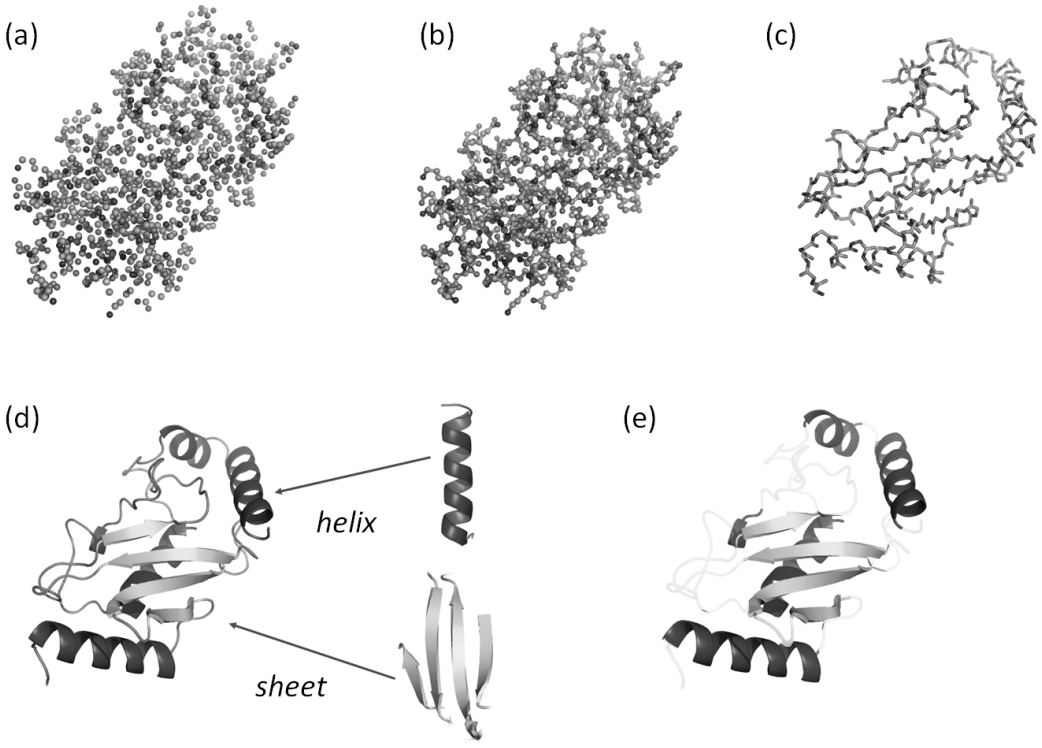


**Figure 2.** *The different descriptions of a protein structure.* (a) The atoms are presented as in the PDB file. (b) Links are done between the atoms. (c) Only the backbone is shown. (d) The secondary structures are assigned. (e) Only the regular structures are really assigned.

Secondary structure assignments are widely used to analyze protein structures. However, it often gives a wrong representation of real protein structures. Figure 2 shows the idea behind the secondary structure assignment. From the atomic co-ordinates in the PDB file, (cf. Figure 2a) covalent bonds can be assigned to link the atoms (cf. Figure 2b) or only the protein backbone can be considered (cf. Figure 2c). The secondary structure assignment as shown in Figure 2d is the classical way to see it, but as shown in Figure 2e, about half of the residues are not assigned any secondary structure.

Moreover, it could give a wrong impression that helices and/or strands are ideal. Though helices and strands are geometrically defined as stable structural elements, local irregularities are often seen. The majority of α-helices is not linear but curved (58%) and even kinked (17%) [13,56]. Contiguous stretches of intra-helical residues exhibiting non-helical geometry have also been well-defined; they are named π-bulges [57]. They are not frequently observed but are implicated in the protein function.

Like α-helices, β-strands are also found to have local stretches with non extended conformation, called β-bulges [58,59]. An elaborate classification of β-bulges has been made by Thornton's group [60]. They are observed quite frequently.

Secondary structure assignment is often considered as a resolved problem and assignment made by DSSP is considered as THE true and the only possible secondary structure assignment. However, it is not the case and the huge number of different assignment methods proved it [10,11,13,18,45-48,61-65]. The most important factor is the choice of descriptors and the parameters used, e.g. distances, angles. Even with similar descriptors, the assignments could be different as shown by [62]. It has a strong

impact of the assignment by itself. Protein flexibility also plays an important role. Comparison of different secondary structure assignment methods has shown some surprising results: difference in assignments could be seen in about one in five residues [66,67]. These different problems had led to the idea that some other descriptions of local protein structures can be useful.

## Local structure libraries

The absence of secondary structure assignment for an important proportion of the residues has lead some scientific teams to develop local protein structure libraries (i) that are able to approximate all (or almost all) of the local protein structures and (ii) that do not take into account the description of classical secondary structures. These libraries brought about the categorization of 3D structures without any *a priori* knowledge of small prototypes that are specific for local folds found in proteins. The complete set of local structure prototypes defines a structural alphabet [68]. A structural alphabet, being able to approximate the local structures in proteins, helps to represent the structural information in one dimension as a sequence. Such a representation also presents methods that are effective and computationally cheap for the comparison and analysis of protein structures (see Table 1 [69]).

***Building Blocks.*** Unger et al. were the first to develop a structural alphabet using a clustering approach based on Cα root mean square deviation (RMSD) [70]. They had chosen hexapeptides as the smallest units that can represent unique local structural information. Using a clustering method called “of annexation” and an RMSD threshold

of 1 Å for clustering. They were able to select about 100 representatives (which they called as "building blocks"). They were able to cover 76% of hexapeptide fragments in the dataset, with an RMSD less than 1.0 Å. [71]. They then carried out a first detailed study of those building blocks associated with extended strands.

***Hierarchical clustering.*** Rooman and Wodak extended their work on protein secondary structure prediction to the description of local protein structures [72]. For this purpose, they performed a hierarchical clustering based on Cα root mean square deviation (RMSD). They were mainly interested on prototypes of different lengths and they tested fragments of lengths ranging from four to seven residues long [73]. They selected four different prototypes for each length. This limited number was chosen based on their final purpose: perform a prediction of these local protein structures from the sequence. Using a simple statistical approach, they obtained a correct prediction rate ranging from 41 to 47% [74].

***Cα distances and dihedral angles.*** Pretrelski and et al. have developed a structural alphabet to support their experimental studies on trypsin-like proteins [75]. For this purpose, they used a combination of linear Cα distances and the Cα dihedrals to generate a set of local structural prototypes. The scoring function designed is a complex combination of Cα distances and the tangent of the dihedral angles. They could find 113 prototypes that are of five residues in length [76]. Their approach was only based on structural approximation.

***Self-Organizing Maps.*** Schuchhardt and et al. designed a complex Self-Organizing Map [77,78] to generate local structure prototypes. Their learning approach was based on protein fragments that are nine residues in length encoded as series of $\varphi$ / $\phi$ angles, *i.e.*, 16 dihedral angles. They could characterize 100 structural prototypes [79]. Interestingly, they could also identify amino acid preferences associated with some structural prototypes that can be considered as part of protein loops.

***Auto-associative Neural Network.*** Fetrow and et al. generated a set of local protein structures using a learning method more complex than the earlier ones [80]. They used an auto-associative neural network (autoANN). This specific neural network has input and output layers with similar dimensions. The hidden layer thus does a compaction of the information. They used this hidden layer to characterize seven residue long fragments encoded as distances, bond and dihedral angles. They generated six structural prototypes and also performed an analysis on the amino acid composition of each prototype, underlining some specificities related to repetitive structures.

***I-sites.*** Based on a library of short sequence patterns having high correlation with the 3D structure, Bystroff and Baker developed an efficient method for predicting local protein structures [81]. They identified frequently occurring sequence motifs by automatic clustering and characterized their corresponding local structures. They further developed an iterative method to optimize the correspondence between sequence and structure. Sequence based clusters were generated with the HSSP protein families [82] and the most frequent local structure in each cluster was chosen as the structural

*paradigm*. An iterative process similar to the *k-means* approach was then employed, by re-estimating the *paradigms* obtained from clusters formed from the dataset. The clustering on the structure space was done using criteria of Cα distance and dihedral angle measure. A library of 82 sequence clusters that are 3 to 19 residues long, were obtained finally. The local structural paradigms corresponding to these clusters were then structurally aligned to get 13 different sequence-structure motifs, which they called "I-sites". The library of I-sites presented new sequence-structure relationships.

In combination with the secondary structure prediction method based on profile based neural networks, PHD, the sequence-structure relationships in the I-sites were used to develop a local structure prediction method leading to a prediction rate of ~ 50%. The prediction method performed well in the CASP2 trials and the prediction for α-spectrin SH3 domain had good correlation with NMR results [83].

They further generated a set of hidden Markov model based profiles called HMMSTR for the sequences in the I-sites library. This HMM was built using overlapping I-sites using an updated dataset [84].

***Hidden Markov Model.*** The first work done by Pr. Serge Hazout (also see *Protein Blocks* section) was on short protein fragments of 4 residues. Described as series of Cα distance, these fragments were learnt by a classical Hidden Markov Model [85]. 13 structural prototypes were obtained from the model and some of them showed specific amino acid preferences. A work dedicated for the prediction of short loops was carried out [86]. A specific work focuses on the reconstruction of protein backbone from Cα traces [87]. Another one was based on the specific learning of fragments from outer

membrane proteins [88], it has lead to propose 20 structural prototypes that show some amino acid specificities. These structural models were used to discriminate CASP models.

***Oligons.*** Michetelli and et al. used an iterative procedure to generate local structure prototypes based on RMSD [89]. At the first stage the fragments were clustered based on the RMSD distribution. The representatives chosen from each cluster, named "oligons", were clustered again and this process was repeated. The optimization process is similar to the classical Monte-Carlo approach. This method helps to generate prototypes with hierarchical weights associated with them, *i.e.*, the first set of oligons is more significant than those that follow. The main aim behind this approach was to generate an increasing number of local structural prototypes. They had tested this approach on fragments of lengths varying from three to ten residues. Highly satisfying results were obtained on structure reconstruction trials using oligons. The importance of the fragment length is highlighted, showing that, for longer fragments, a large number of prototypes are required for a similar 3D approximation. No specific study of amino acid specificities associated with these local protein structures was done.

***Centroids.*** Using a hyper-cosine clustering method, Hunter and Subramaniam [90] clustered 7 residue fragments. RMSD was used as the distance measurement. They chose a threshold to define the optimum number of clusters, which they called, the centroids. Despite a detailed analysis of parameters used to select the threshold, the fragment distribution among the 28 clusters finally chosen, is highly uneven. To develop

a prediction method based on the set of centroids generated, they used a Bayesian predictor that gives the probability of each centroid to occur at a position in the sequence. This prediction is highly related to the prediction used for the Protein Blocks (see *Protein Blocks* section) [91]. An overall prediction accuracy of 40% was obtained. However, this correct prediction rate gives a wrong impression, as it is in fact highly biased. Indeed, 11 of the 28 centroids are not predicted at all, which diminish greatly the interest of the approach [92]. Moreover, some major divergences can be noted between the two papers describing the approach.

***k-means.*** Sander and et al. have developed a novel approach based on the use of Cα distance matrix comparison [93] using a 'complex' *k*-means. They defined 27 prototypes of eight residues comparable to those developed by Hunter and Subramaniam [92]. They also incorporated protein family information by using profiles instead of simple sequences. They have tested numerous prediction methods: C.5 classifier, Support Vector Machines and random forest. All these approaches have led to an unbiased prediction unlike the predictions made using Hunter and Subramaniam approach [92].

***Kappa-alpha map***. Tung and et al. have defined a structural alphabet dedicated to mine the Protein DataBank [94]. The main principle used in this approach is a measure based on Cα distance and a nearest-neighbor clustering (NNC) algorithm. A set of 23 local prototypes were selected and used to identify similar protein structural domains and corresponding SCOP superfamilies [95,96]. The search methodology is based on the direct use of BLAST algorithm; similar to the work done earlier with Protein Blocks (see

*Protein Blocks* section), *i.e.* PBE [69]. Analysis of sequence – structure relationship was not done.

***SOMs and k-means.*** Recently, Ku and Hu [97] used the idea developed by Schuchhardt *et al.* [79] and that was used for Protein Blocks design [91], namely defining the protein in terms of φ / ϕ dihedrals. Like Protein Blocks, they used five residue long fragments to define the prototypes. The first step is a classical learning using a Self-Organizing Map [77,78]. After many simulations with different number of neurons, they selected a large map and analyzed it using U-matrix visualization. From these data, they clustered the results using *k*-means approach. Then, a substitution matrix was computed and it is optimized to detect SCOP class similarity. A FASTA methodology is used to compute the similarity score. Analysis of sequence – structure relationship was not done.

***Protein Folding Shape Code.*** Recently Yang described a novel approach based on the description of protein local structures as a vector of angle and distances. He had only used Cα distances and obtained 27 prototypes of length 5 [98].

## Protein Blocks

***Design of Protein Blocks.*** Following an earlier work, Pr. Serge Hazout developed a novel structural alphabet, with two specific goals: (i) to obtain a good local structure approximation and (i) to predict local structures from sequence. Fragments that are five residues in length were coded in terms of the φ / ϕ dihedral angles. A Root Mean Square

Deviation on Angle (RMSDA) score was used to quantify the structural difference among the fragments. This idea was already used by Schuchhardt and et al. [79]. Using an unsupervised cluster analyser related to self organised Kohonen maps [77,78], a three step training process was carried out: (i) the learning of structural difference of fragments has been performed only using the minimal RMSDA as criterion to associated a fragment to a cluster, (ii) the transition probability (probability of transition from one fragment to another in a sequence) was also added to select the cluster associated to the protein fragment, (iii) this last constraint was removed. The optimal number of prototypes was obtained by considering both the structural approximation and the prediction rate. A set of 16 prototypes called "protein blocks (PBs)", represented as average dihedral vectors, were obtained at the end of this process [91]. Figure 3a shows the 16 PBs. Figure 4 gives an example of PB assignment.

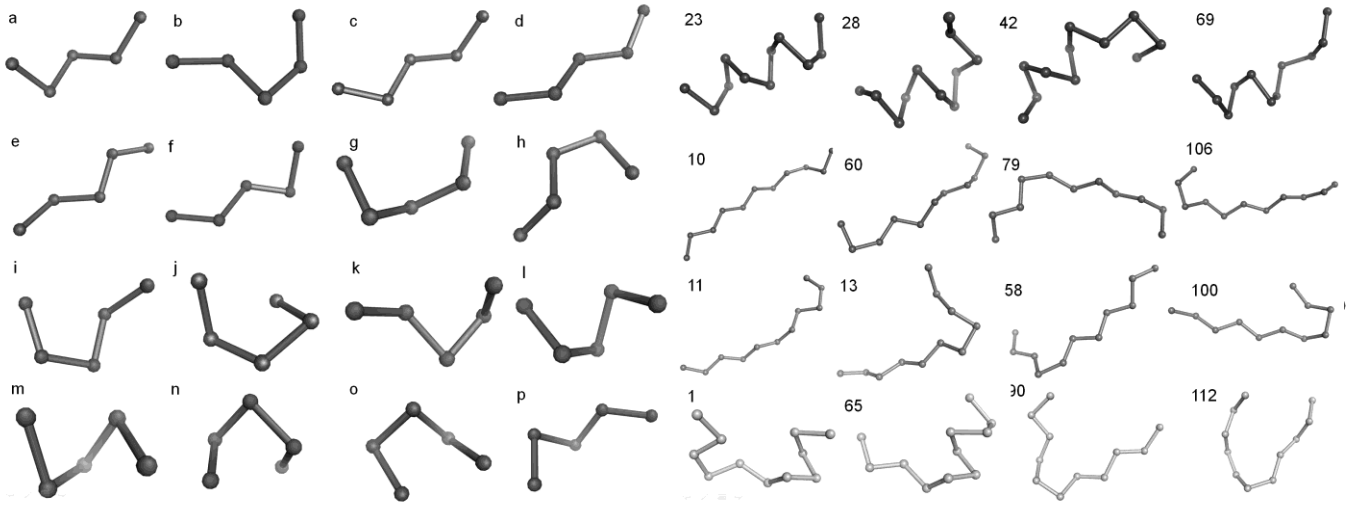


**Figure 3.** *Protein Blocks and Local Structure Prototypes.* (left) are shown the 16 PBs (5 residues in length)), (right) some examples of the 120 LSPs (11 residues in length). LSPs 23, 28, 42 and 69 belongs to the *helical* LSP, LSPs 10, 60, 79 and 106 to *extended* LSP, LSPs 11, 13, 58 and 100 to *extended edges* LSPs , and, LSPs 1, 65, 90 and 112 to *connection* LSP.

***Analysis of PBs.*** The relationship between PBs and secondary structures was analysed. PB *m* corresponds to the central part of helices while PB *d* corresponds to strands. Some PBs are associated with the N-and C-caps of helices and strands representing subtle variations in the termini. Some PBs also represent conserved features in the coils. Specific or highly preferential transitions are observed between consecutive PBs in a sequence. The three major transitions observed correspond to about 76% of the possible transitions. The distribution of PBs, transition probabilities and structural definitions has been evaluated and cross-checked using different datasets of proteins. These features were found to be highly consistent among the different datasets [99]. Table 2 shows the correspondence of all the 16 PBs and the different secondary structure elements. It has been computed with a non-redundant databank with 25% of sequence identity and a resolution better than 2.5 Å. Protein list has been taken from PISCES web server [100] and the secondary structure assignment has been done with DSSP [9]. Table 2a shows the frequencies of classical secondary structures for each PB, while Table 2b shows the opposite. It highlights that a-helix and other helical structures are associated only to PBs *k* to *o*, while turns are found spread over all the PBs. It underlines also the non-equivalence of turns and coils that have specificities.

***Structural alignment.*** Based on PBs, a new structure comparison method (PB-ALIGN) useful for mining protein structural databases has been developed. Using the structural homologues in PALI database [101] encoded in terms of PBs, a dedicated PB substitution matrix was computed [69]. Using this matrix with a classical alignment approach, it is possible to find structural homologues [102], similar to what is done in the

case of amino acid sequences. A recent benchmark has proved that this method is most efficient for mining the PDB to find structural homologues [103].

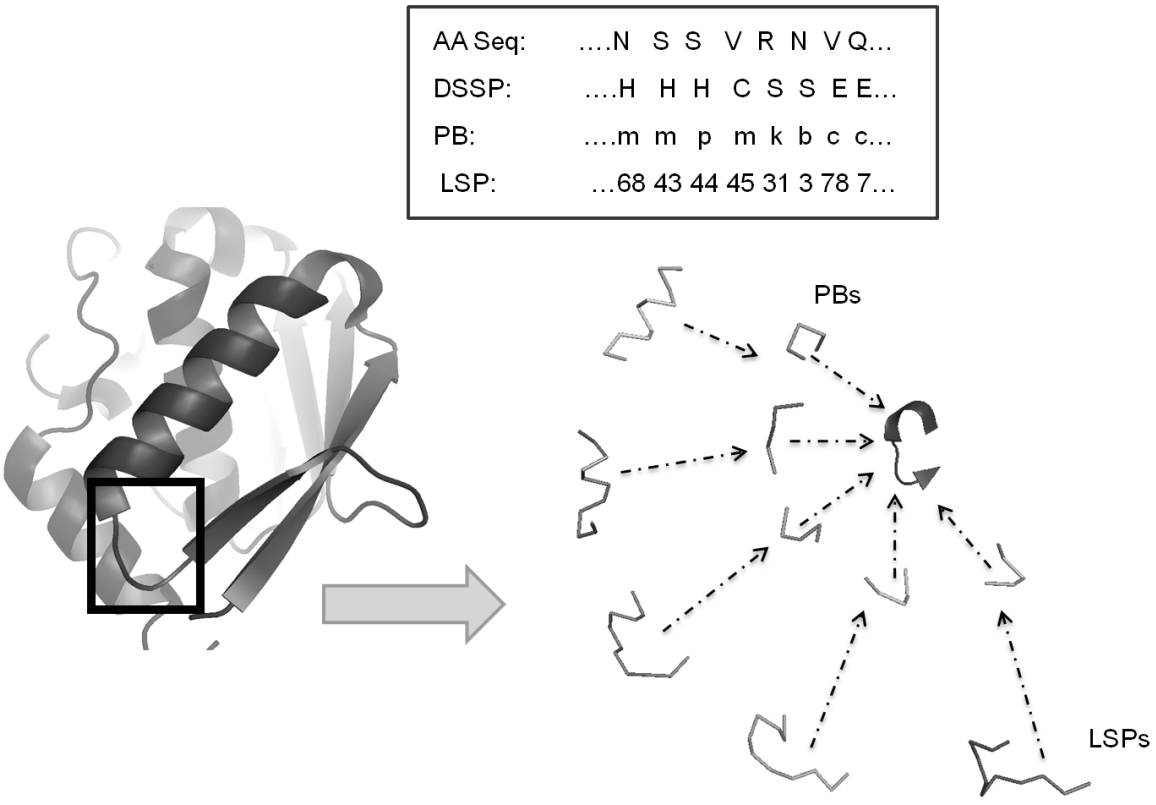


**Figure 4.** *Example of assignment.* The zinc endoprotease (PDB code 1c7k [104]) has been encoded in terms of secondary structures with DSSP (shown in 3D on the left), but also in terms of Protein Blocks and Local Structure Prototypes. The short protein fragment in the black box is detailed with the PB and the LSP sequence. The corresponding prototypes are shown also.

***Longer fragments.*** An analysis of preferential transitions of PBs of various lengths, suggested that the series of 5 PBs (or 9 residues) present interesting structural features [105]. The distribution and consistency of structural features associated with fragments representing set of 5 PBs were checked on different datasets and significant variation was not observed. Based on the extent to which a set of such fragments can

cover a protein chain, an optimal set of 72 fragments called as "Structural Words (SW)" were selected. They represented 92% of the databank residues, nearly all the repetitive structures and 80% of the "coil". Most of these SWs were found to overlap; some had even four PBs in common. These structural words represent local structure transitions and irregularities. Quality of structural approximation was assessed, showing that a structural alphabet is meaningful even for longer fragments.

Following this idea, a novel approach was developed: the Hybrid Protein Model (HPM [106]). This specific clustering allows associating longer protein fragments to create structural prototypes with high transition between them [107-111]. From a dataset of proteins coded in the form of PB sequences, fragment sequences of PBs of varying lengths were derived. Similarity between the fragments is decided based on the propensities of PBs to occur at each position in the fragment. In this process, for a given fragment length, a hybrid protein of an optimal length that can represent the sets of preferential transitions of local structures in continuity, is generated. The length of the hybrid protein and the propensities of PBs to occur at a position varied during learning. Redundant sets of PB transitions (similar propensities at the same positions). The results of a HPM approach on a dataset of fragments of length 10 residues, could be effectively used for fine description of protein structures and the data was used efficiently for the identifying local structural similarities between two cytochromes P450 [107]. An hybrid protein of length 233 based of 13 residue long fragments, gave a better description of various local structural features [108]. Recent development has given a new hybrid protein that has been used for prediction purpose [109,112].

***Structure Prediction Using PBs.*** A Bayesian probabilistic approach was utilized for the prediction of PBs from amino acid sequence. For learning the amino acid propensities associated with each PB, the set of proteins chains used in training were then encoded in terms of PBs, using the minimal RMSDA criterion. Sequence windows of length 15 residues were considered for calculating the propensities associated with each PB. For every PB, the probability of occurrence of an amino acid at each position in the sequence window was calculated and an occurrence matrix was generated for each of the sixteen PBs. Bayes theorem was used to predict the structure of new sequences. A prediction rate of 34.4% was achieved [91,113].

One of the limitation of this approach is to average the sequence information associated with a PB as only one amino acid occurrence matrix corresponds to one PB. Thus, using a clustering approach related to SOM [77], amino acid occurrence matrices was split for some PBs, increasing their sequence specificities. Bayesian prediction was carried out to achieve an improved prediction rate of 40.7 % [91,113]. In the process of generating sequence families, including a simulated annealing approach that maximizes the prediction rate, helped to improve the overall prediction to 48.7% [113,114]. No biased or unbalanced improvements were detected among the PBs, with this approach. Combining the secondary structure information with the Bayesian prediction did not result in significant improvement of the prediction rate. A java based program named LocPred (see Figures 5 and 6) is available to perform these predictions [113].

A Bayesian prediction approach (without optimization of sequence-structure relationships) similar to what was used for PB prediction was also carried out for the SWs. A 4% improvement in prediction rate could be achieved [105]. Preferable transitions were also observed between SWs occurring in a sequence and certain series of SWs were found to be highly frequent. Use of this information with an approach called "pinning strategy", helped to improve the prediction rate significantly [115]. Principle of pinning strategy is quite simple: (i) a classical Bayesian prediction is done with SWs, (ii) the positions with a high prediction confidence index are selected as "seeds", (iii) at a seed position *i*, a SW (5 PBs) is predicted; so, a selection is also done at position *i*-1 (and *i*+1 respectively), one SW that overlaps this SW is selected through the most probable SW. It is an iterative process, from *i*-1 (and *i*+1 respectively), the prediction is extended through *i*-n (and *i*+n' respectively); it stops when a probability threshold is reached.

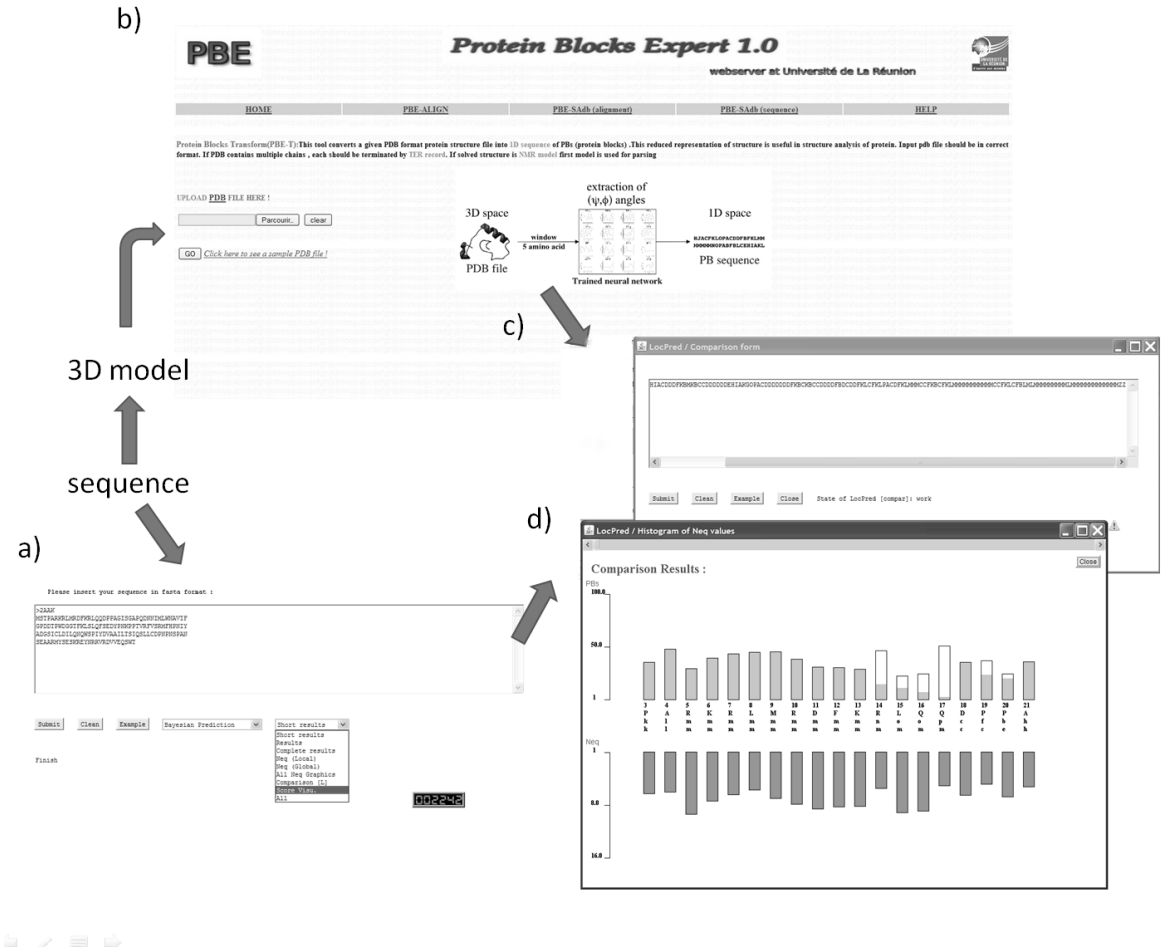


**Figure 5.** *LocPred use with a structural model.* It is possible to confront PB predictions with 3D structural model obtained by another approach. (a) The *fasta* sequence is given and the prediction options are selected. (b) The structural model is encoded in terms of PBs with PBE website (http://bioinformatics.univ-reunion.fr/PBE/). (c) The PB sequence corresponding to the structural model is put into the *Comparison form*. (d) The compatibility between prediction and structural model is given graphically.

A detailed analysis of PB distribution in short loop regions (6 to 10 residues) has been done [30]. The description in terms of PBs helped to understand the ambiguity associated with the assignment of the boundaries of regular secondary structures based on different assignment methods. Specific sequence-structure relationships in the short loops could be derived. A Bayesian prediction carried out based on this information gave an accuracy rate of 41.2% for the short loops and 36% for the loops in general. A recent

study has shown that a specific learning of the different kinds of short loops improved greatly the prediction [116].

LocPred is useful to predict the protein structures in terms of PBs, but also to analyze the sequence – structure relationship of the protein of interest. The simplest output of LocPred is a list with the raw prediction values with their confidence indexes and the different probabilities. Graphical outputs give visual representations of the probabilities associated to each predicted PB, it helps to have an idea of the local tendencies, and the confidence index associated to each position, *i.e*. lower the confidence index is, better it is. This option could be so helpful even if the user does not want to use PBs, it quantifies the sequence – structure relationship of this protein.

Figure 5 gives another possibility given by LocPred, *i.e.* the comparison of a structural model and PB prediction. A prediction is performed as given in Figure 5a. Many different approaches, softwares and web services allow the obtaining of structural model. Thanks to PBE web server (see Figure 5b), it is simple to translate a protein structure in terms of PBs. Then, in LocPred, it is possible to compare the assigned PBs of the structural model with the PB predictions (see Figure 5c and 6d). Figure 5d shows an example of such comparison. For each amino acid position is given the amino acid, the position in the sequence and the two PBs, i.e. the assigned and the predicted one. The histogram corresponds to the prediction of the best predicted PBs. When the predicted and assigned PBs are the same, the histogram bar is plain, otherwise the colour is smaller as in the second part of the example (positions 14 to 17). It helps to localize critical structural regions of the structural model.

***Prediction with the Hybrid Protein Model.*** In order to extend the analyses of long structural fragments, the HPM was used to construct a new library of local structures. 120 structural clusters were proposed to describe fragments of 11-residue long [109]. For each class, a mean representative prototype, named Local Structure Prototype (LSP, see Figure 3b), was chosen according to Cα RMSD criteria. These 120 LSPs enabled a satisfying average approximation of 1.6 Å for all local structures observed in known proteins. The consequences of long-range interactions are taken into account thanks to the high length of fragments. Moreover, the major advantage of this library is its capacity to capture the continuity between the identified recurrent local structures. The overlapping properties of LSPs were used to identify very frequent transitions between them and characterize their involvement in longer super secondary structures [112]. Figure 4 gives an example of LSP assignment.

For each one of the 120 structural classes, high sequence-relationships were observed and led the development of an original prediction method from single sequence and based on logistic regressions. The main purpose of local structure prediction methods is to reduce the combinatory of structural possibilities for a sequence. Thus, it is worth noting that this method proposed a short list of the best structural candidates among the 120 LSPs of the library. Moreover, to identify directly regions easier or difficult to predict, each prediction is associated to a confidence index. With a geometrical assessment, a prediction rate of 51.2 % was reached. This result was already very satisfying given the high length of fragments and the high number of classes [109].

Recently, an improved prediction method relying on Support Vector Machines (SVM) and evolutionary information was proposed. A global prediction rate of 63.1 %

was achieved and corresponded to an improved prediction of 85 % of proteins. A confidence index was also defined for directly assessing the relevance of the prediction at each sequence site. This method was shown to be among the most efficient cutting-edge local structure prediction strategies [112]. Taking advantage of the high length of fragments, the relationships between their structural flexibility and their predictability are now under study.

***Solving a biological problem – DARC.*** Local structure prediction based on PBs was used along with threading, *ab initio* and secondary structure prediction methods to determine the fold of the Duffy Antigen/Receptor for Chemokines (DARC) [117]. DARC occurs on the surface of erythrocytes and serves as a receptor for various chemokines. It was also identified as the erythrocyte receptor for *Plasmodium vivax* and *Plasmodium knowlesi* parasites. In the absence of well-defined homologues of known structure, modelling of transmembrane proteins remains a difficult task. PB predictions from the regions of low information content were highly relevant for the analysis of the models generated by energy minimization and molecular dynamics refinements. This example was a very good example of interest that helped to analyze the results of simulated annealing based prediction with a finer description. We have recently described the use of such approaches for the DARC [118] to define pertinent structural models [119]. Figure 6 describes the protocol used, which is based on (i) biochemical data, some residues must be accessible, (ii) transmembrane predictions and (iii) Protein Blocks approach.

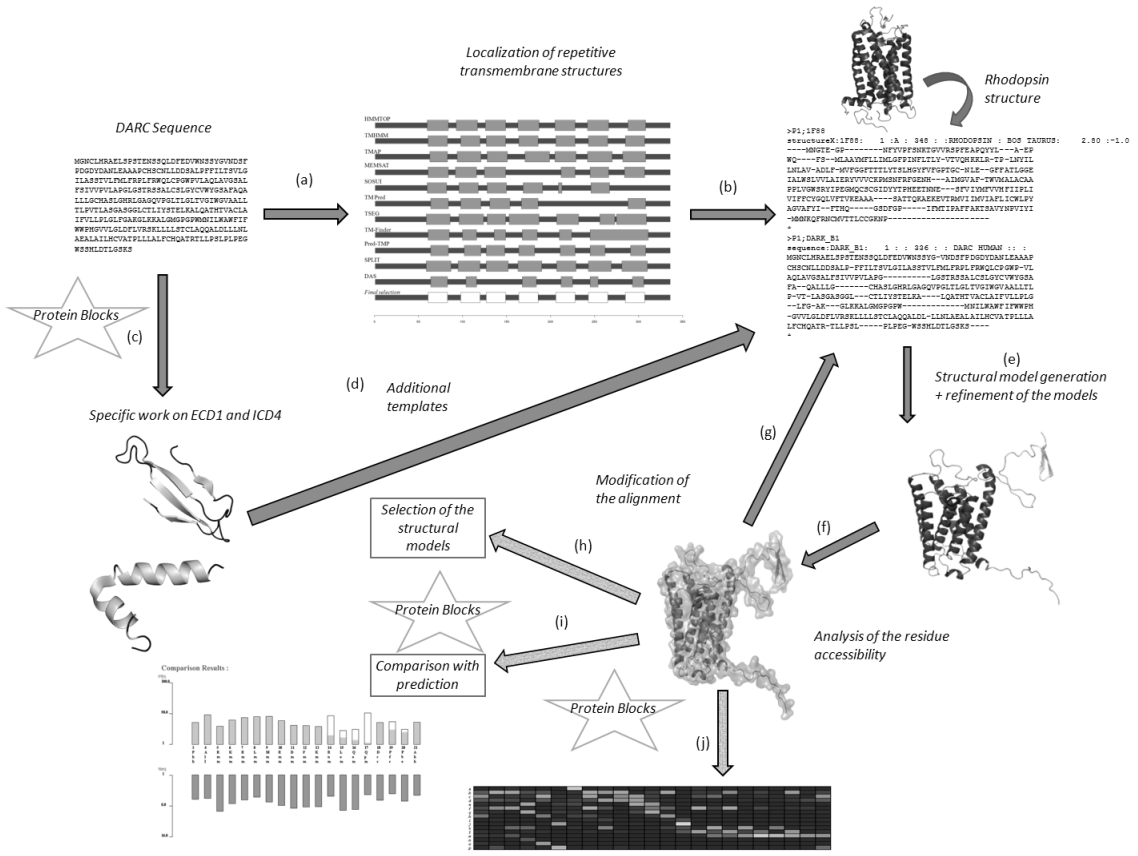


**Figure 6.** *Building structural models of DARC***.** (a) Prediction of transmembrane helices. (b) Alignment of helical regions with corresponding regions of rhodopsin structure. (c) Potential structural templates for the extremities are done thanks to Protein Blocks (d) Addition of these results to the complete alignment for comparative modeling. (e) Structural model generation and refinement of these models. (f) Accessibility computation of amino acids and known to be exposed. (g) In regards to the results, the alignment is modified. (h) At last, some models are selected. (i) As seen on Figure 5, comparison between PB prediction and the PB assignment can help to locate arduous regions. (j) PBs can also be used to analyze protein molecular dynamics as in [117].

***Comparison of predictions.*** As most of the structural alphabets are not available for use for the scientific community, it is very difficult to make a comparison. Comparison of prediction is not trivial, but can be done, even if they are based on unrelated methodologies. Yang and Wang developed a database of sequence profiles of nine residue fragments, the members of each profile having similar backbone conformational state and similar sequences. These profiles are generated in a two step

process. In the first step, seed sequence profiles were generated based on φ / ϕ dihedral states defined by [120] and also on the sequence similarity calculated based on structure specific amino acid substitution matrices, [121]. The preliminary profiles, in the form of PSSMs were then used to search for more fragments with identical backbone conformation and a good sequence-profile match score. A Bayesian prediction pseudo count method was used to represent the amino acid occurrence propensities in the preliminary PSSMs. For the prediction purpose, the set of sequence profiles with a good sequence profile matching score and having at least 60% consistency with the secondary structure prediction by PSIPRED, were chosen. For each of the selected profiles, a consensus score giving an indication of the extent of backbone conformational similarity with others in the set is calculated. The one with the highest consensus score is chosen as the predicted candidate. The percent of correct predictions on a dataset were comparable to those obtained with HMMSTR. However, based on RMSD between the true and the predicted structure, this method is reported to perform better than HMMSTR. The prediction accuracy was later improved with the use of SVMs and Neural Networks [122]. Prediction made using HPM with linear regression [109,112] was comparable to these approaches, and the results are better with our new approaches that use SVMs with evolutionary information.

More recently, another method for predicting PBs from sequence has been developed. Li and et al. propose an innovative combination of PB prediction, taking into account the information on secondary structure and solvent accessibilities [123]. Prediction rates were improved, and, interestingly their approach was found useful for fragment threading, pseudo sequence design, and local structure predictions.

Zimmermann and Hansmann developed a method in the recent times, named *Locustra*, for predicting local structures encoded in terms of PBs from sequence [124]. The prediction was carried out using SVMs with a radial basis function kernel. For the prediction of each class of PB, a two layer classification scheme was used. In the first step, the samples belonging to one class was considered as the positive set while those belonging to another class were considered as the negative class, *i.e.*, a pairwise coupling classifier. 120 classifiers were required. The input sequence data was enriched using the information derived from the homologues and a profile of amino acid propensities was obtained. The sequence window of 15 residues indicated a feature vector of size 315. To estimate the class probability, a cross-validation based method was used. The probabilities at each sequence position, obtained from the 120 pairwise coupling classifiers were used as features for the second layer. Here, a one-per-class classifier was used, where the samples belonging to one class is considered as the positive set while those belonging to all the other classes were included in the negative set. The PB having the highest number of votes in the output of the second layer was chosen as the predicted PB. The major secondary structures like helices or strands were chosen in cases of multiple predictions. The prediction accuracy reaches 61%. It was also noted that the PBs that are mispredicted were often structurally related to the true PB and these mispredictions often correspond to exposed regions of the structure.

Prediction of PBs is very simple as only a sequence in Fasta format is needed. Protein Blocks are the only structural alphabets with web-service for prediction and moreover, three different approaches are available.

# Conclusions and Perspectives

In this paper, we have presented different facets of the protein structures at a local level, underlining some limitations of using secondary structures for describing protein structures. Global protein structures can be described by a limited set of recurring local structures [125] and in this context, the use of structural alphabets is obvious. As it is not easy to build relevant structural models directly with structural prototypes, I-sites have been added to a prediction method, namely Rosetta [126].

Recently, Dong and et al. developed a set of structural alphabets with the aim of finding an optimal structural alphabet sequence from which an accurate model of the protein can be regenerated [127]. Using the standard *k-means* algorithm they clustered fragments that are seven residues in length, based on the Cα RMSD. The set of alphabets generated were used to reconstruct the structure of the protein such that the global RMSD is minimal. For doing so, they adopted a combination of greedy and dynamic programming algorithms. Sets of structural alphabets of sizes 4 to 100 prototypes were evaluated for both local and global structure approximations and finally a set of 28 letters were chosen. When compared with the global approximation based on PBs, this set of alphabets is reported to give slightly better results. Thus, the future of local protein structures is promising in the area of building relevant structural models.

Till this day, nearly all the structural alphabets are only used within the research groups that have developed them (see Table 1). Hence, Protein Block structural alphabet is an exception. Protein Blocks is one of the most widely used structural alphabet. Indeed, it is easy to use PBs for various applications. Protein Blocks have been used both to describe the 3D protein backbones [99] and to perform a local structure prediction

[91,113,114,116]. The efficiency of PBs have also been proven in the description and the prediction of long fragments [67,105,107-111,115,128], to compare protein structures [69,102,103], to build globular [127] and transmembrane protein structures [117], to define a reduced amino acid alphabet dedicated to mutation design [129], to design peptides [130] or to define binding site signatures [131]. The features of this alphabet have been compared with those of 8 other structural alphabets showing clearly that the PB alphabet is highly informative, with the best predictive ability of those tested [132].

Future of structural alphabets is also coupled with the taking into account more biophysical feature. One of our main axes of research is so the link between local protein structure prediction and the protein flexibility [133]. For this purpose, we have studied protein dynamics from two different points of view, *i.e.*, X-ray experiments and molecular dynamics (MD) simulations. Prediction results are quite good in comparison to available methodologies.

## Acknowledgments

This work was supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, Université de Saint-Denis de la Réunion, National Institute for Blood Transfusion (INTS) and the Institute for Health and Medical Care (INSERM). APJ has a grant from CEFIPRA number 3903-E and AB has a grant from the Ministère de la Recherche.

| Research Team | Number of proteins in dataset | Fragment Length | Distance measure | Learning Method | prototype number | prediction |
|---|---|---|---|---|---|---|
| Unger *et al.* | 4/82 | 6 | Cα RMSD | K-means | 103 | N |
| Rooman et al | 75 | 4,5,6, 7 | Cα RMSD | Hierarchical Clustering | 4 | Y |
| Prestrelski *et al.* | 14 | 8 | Linear Cα disance and α torsion angle | Function of Cα disance and torsion angle | 113 | N |
| Schuchhardt *et al.* | 136 | 9 | Dihedral angles | Kohonen map | 100 | N |
| Fetrow *et al.* | 116 | 7 | C α distance, dihedral and bond angles | Auto-ANN | 6 | N |
| Bystroff and Baker | 471 | 3-19 | Sequence profiles, RMSD, MDA | k-means | 13(later updated to 16) | Y |
| Camproux *et al.* | 100 | 4 | Cα distance | HMM | 12 | N |
| Micheletti *et al.* | 75 | 4,5,6,7 | Cα RMSD | Iterative clustering (Monte-carlo like) | 28,202,932,2561 | N |
| de Brevern *et al.* | 342 | 5 | Dihedral angles | Unsupervised classifier (SOM with transition probabilities) | 16 | Y |
| Kolodony *et al.* | 145/200 | 4,5,6,7 | Cα RMSD | Simulated annealing based on k-means | 4-14,10-225,40-300,50-250 | N |
| Hunter and Subramaniam | 790 | 7 | Hypercosine Cα | Hypercosine clustering | 28-16336 | Y |
| Camproux *et al.* | 250 * 2 | 4 | C α distance | HMM | 27 | N |
| Etchebest *et al.* | 1407 | 5 | Dihedral angles | Unsupervised classifier | 16 *(New evaluation)* | Y |
| Benros *et al.* | 675 & 1401 | 11 | Cα RMSD, PB based | Hybrid Protein Model | 120 | Y |
| Sander *et al.* | 1999 | 7 | Cα distance | Leader algorithm and k-means | 28 | Y |
| Tung *et al.* | 1348 | 5 | κ and α angle | Nearest Neighbor Clustering | 23 | N |
| Ku and Hu | 18 | 5 | Dihedral angle | SOM & k-means | 18 | N |
| Bornot *et al.* | 675 & 1401 | 11 | Cα RMSD, PB based | Hybrid Protein Model | 120 | Y |
| Yang | 268 | 5 | Cα distances and angles | Shape object clustering | 27 | N |

**Table 1.** *The different sets of structural alphabet.*

(a)

| Protein Blocks | secondary structures | | | | | | |
|---|---|---|---|---|---|---|---|
| | α-helix | $3_{10}$ helix | π-helix | turn | coil | β-strand | freq PB |
| *a* | *0.14* | *0.13* | *0.00* | **19.35** | **62.64** | **17.74** | 3.92 |
| *b* | *0.13* | *0.10* | *0.00* | **58.38** | **25.84** | **15.54** | 4.16 |
| *c* | *0.00* | *0.01* | *0.00* | **13.51** | **43.83** | **42.65** | 7.93 |
| *d* | *0.00* | *0.00* | *0.00* | 5.42 | **21.77** | **72.81** | 18.28 |
| *e* | *0.05* | *0.18* | *0.00* | 9.15 | **38.51** | **52.11** | 2.36 |
| *f* | *0.01* | *0.01* | *0.00* | 7.67 | **66.36** | **25.96** | 6.52 |
| *g* | *4.56* | 7.83 | *0.00* | **52.76** | **29.67** | 5.18 | 1.10 |
| *h* | *0.27* | *2.54* | *0.00* | **62.35** | **16.66** | **18.17** | 2.30 |
| *i* | *0.24* | *2.08* | *0.00* | **84.33** | 7.63 | 5.72 | 1.79 |
| *j* | *4.55* | 5.34 | *0.00* | **59.58** | **21.35** | 9.18 | 0.79 |
| *k* | **35.21** | **13.69** | *0.02* | **43.98** | 6.34 | *0.76* | 5.41 |
| *l* | **44.90** | **17.24** | *0.02* | **31.14** | 6.13 | *0.57* | 5.38 |
| *m* | **86.37** | *4.51* | *0.07* | 6.42 | *2.51* | *0.12* | 31.50 |
| *n* | **64.02** | 7.41 | *0.14* | **24.26** | *3.49* | *0.69* | 2.17 |
| *o* | **23.08** | 6.45 | *0.02* | **66.30** | *3.87* | *0.28* | 2.86 |
| *p* | *4.05* | **12.37** | *0.00* | **62.87** | **18.91** | *1.81* | 3.53 |

(b)

| Protein Blocks | secondary structures | | | | | |
|---|---|---|---|---|---|---|
| | α-helix | $3_{10}$ helix | π-helix | turn | coil | β-strand |
| *a* | *0.02* | *0.12* | *0.00* | 3.67 | **12.57** | *3.19* |
| *b* | *0.02* | *0.11* | *0.00* | 11.74 | 5.49 | *2.96* |
| *c* | *0.00* | *0.02* | *0.00* | 5.18 | **17.78** | **15.52** |
| *d* | *0.00* | *0.02* | *0.00* | *4.79* | **20.34** | **61.04** |
| *e* | *0.00* | *0.11* | *0.00* | *1.04* | *4.64* | 5.63 |
| *f* | *0.00* | *0.02* | *0.00* | *2.42* | **22.13** | 7.77 |
| *g* | *0.15* | *2.10* | *0.00* | *2.81* | *1.67* | *0.26* |
| *h* | *0.02* | *1.42* | *0.00* | 6.93 | *1.96* | *1.92* |
| *i* | *0.01* | *0.91* | *0.00* | 7.30 | *0.70* | *0.47* |
| *j* | *0.11* | *1.03* | *0.00* | *2.28* | *0.86* | *0.33* |
| *k* | 5.63 | **18.01** | *4.44* | **11.50** | *1.75* | *0.19* |
| *l* | 7.14 | **22.58** | *4.44* | 8.10 | *1.69* | *0.14* |
| *m* | **80.42** | **34.55** | **77.78** | 9.78 | *4.05* | *0.18* |
| *n* | *4.11* | *3.91* | 11.11 | *2.55* | *0.39* | *0.07* |
| *o* | *1.95* | *4.49* | *2.22* | 9.17 | *0.57* | *0.04* |
| *p* | *0.42* | **10.62** | *0.00* | **10.73** | *3.41* | *0.29* |
| freq. S2 | 33.83 | 4.11 | 0.03 | 20.68 | 19.56 | 21.80 |

**Table 2.** *S2 ←→ PBs. (a) (a) Is given the relative frequencies of PBs for each secondary structures. (b) The relative frequencies of secondary structures in each PB. In bold are the frequencies more than 10, in italics frequency less than 5%.*

# References

[1] Doppelt, O., Moriaud, F., Delfaud, F. and de Brevern, A.G. (2009). Analysis of HSP90 related folds with MED-SuMo classification approach. Drug Design, Development and Therapy 9, 3.

[2] Slabinski, L., Jaroszewski, L., Rodrigues, A.P., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007). The challenge of protein structure determination--lessons from structural genomics. Protein Sci 16, 2472-82.

[3] Doppelt, O., Moriaud, F., Bornot, A. and de Brevern, A.G. (2007). Functional annotation strategy for protein structures. Bioinformation 1, 357-9.

[4] Pauling, L. and Corey, R.B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci U S A 37, 251-6.

[5] Pauling, L., Corey, R.B. and Branson, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37, 205-11.

[6] Sayle, R.A. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem Sci 20, 374.

[7] Perez-Iratxeta, C. and Andrade-Navarro, M.A. (2008). K2D2: estimation of protein secondary structure from circular dichroism spectra. BMC Struct Biol 8, 25.

[8] 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, 536-40.

[9] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-637.

[10] Frishman, D. and Argos, P. (1995). Knowledge-based protein secondary structure assignment. Proteins 23, 566-79.

[11] Srinivasan, R. and Rose, G.D. (1999). A physical basis for protein secondary structure. Proc Natl Acad Sci U S A 96, 14258-63.

[12] Cubellis, M.V., Cailliez, F. and Lovell, S.C. (2005). Secondary structure assignment that accurately reflects physical and evolutionary characteristics. BMC Bioinformatics 6 Suppl 4, S8.

[13] Martin, J., Letellier, G., Marin, A., Taly, J.-F., de Brevern, A.G. and Gibrat, J.-F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Structural Biology 5, 17.

[14] Eisenberg, D. (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A 100, 11207-10.

[15] Richardson, J.S. and Richardson, D.C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. Science 240, 1648-52.

[16] Pal, L., Chakrabarti, P. and Basu, G. (2003). Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. J Mol Biol 326, 273-91.

[17] Regan, L. (1994). Protein structure. Born to be beta. Curr Biol 4, 656-8.

[18] Tyagi, M., Bornot, A., Offmann, B. and de Brevern, A.G. (2009). Analysis of loop boundaries using different local structure assignment methods. Protein Sci, in press.
[19] Khare, S.D. and Dokholyan, N.V. (2007). Molecular mechanisms of polypeptide aggregation in human diseases. Curr Protein Pept Sci 8, 573-9.
[20] Aurora, R. and Rose, G.D. (1998). Helix capping. Protein Sci 7, 21-38.
[21] Kruus, E., Thumfort, P., Tang, C. and Wingreen, N.S. (2005). Gibbs sampling and helix-cap motifs. Nucleic Acids Res 33, 5343-53.
[22] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120, 97-120.
[23] Rost, B. and Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A 90, 7558-62.
[24] Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
[25] Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21, 1719-20.
[26] Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47, 228-35.
[27] Pal, L., Basu, G. and Chakrabarti, P. (2002). Variants of 3(10)-helices in proteins. Proteins 48, 571-9.
[28] Pal, L., Dasgupta, B. and Chakrabarti, P. (2005). 3(10)-Helix adjoining alpha-helix and beta-strand: sequence and structural features and their conservation. Biopolymers 78, 147-62.
[29] Lee, K.H., Benson, D.R. and Kuczera, K. (2000). Transitions from alpha to pi helix observed in molecular dynamics simulations of synthetic peptides. Biochemistry 39, 13737-47.
[30] Fourrier, L., Benros, C. and de Brevern, A.G. (2004). Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 5, 58.
[31] Eswar, N., Ramakrishnan, C. and Srinivasan, N. (2003). Stranded in isolation: structural role of isolated extended strands in proteins. Protein Eng 16, 331-9.
[32] Park, S.Y., Yamane, K., Adachi, S., Shiro, Y., Weiss, K.E., Maves, S.A. and Sligar, S.G. (2002). Thermophilic cytochrome P450 (CYP119) from Sulfolobus solfataricus: high resolution structure and functional properties. J Inorg Biochem 91, 491-501.
[33] DeLano, W.L.T. (2002). The PyMOL Molecular Graphics System DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org
[34] Venkatachalam, C.M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6, 1425-36.
[35] Hutchinson, E.G. and Thornton, J.M. (1994). A revised set of potentials for beta-turn formation in proteins. Protein Sci 3, 2207-16.

[36] Fuchs, P.F. and Alix, A.J. (2005). High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59, 828-39.
[37] Zheng, C. and Kurgan, L. (2008). Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics 9, 430.
[38] Koch, O. and Klebe, G. (2009). Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. Proteins 74, 353-67.
[39] Meissner, M., Koch, O., Klebe, G. and Schneider, G. (2009). Prediction of turn types in protein structure by machine-learning classifiers. Proteins 74, 344-52.
[40] Makowska, J., Rodziewicz-Motowidlo, S., Baginska, K., Vila, J.A., Liwo, A., Chmurzynski, L. and Scheraga, H.A. (2006). Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. Proc Natl Acad Sci U S A
[41] Stapley, B.J. and Creamer, T.P. (1999). A survey of left-handed polyproline II helices. Protein Sci 8, 587-95.
[42] Eker, F., Griebenow, K. and Schweitzer-Stenner, R. (2004). Abeta(1-28) fragment of the amyloid peptide predominantly adopts a polyproline II conformation in an acidic solution. Biochemistry 43, 6893-8.
[43] Hicks, J.M. and Hsu, V.L. (2004). The extended left-handed helix: a simple nucleic acid-binding motif. Proteins 55, 330-8.
[44] Hollingsworth, S.A., Berkholz, D.S. and Karplus, P.A. (2009). On the occurrence of linear groups in proteins. Protein Sci 18, 1321-5.
[45] King, S.M. and Johnson, W.C. (1999). Assigning secondary structure from protein coordinate data. Proteins 35, 313-20.
[46] Labesse, G., Colloc'h, N., Pothier, J. and Mornon, J.P. (1997). P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci 13, 291-5.
[47] Dupuis, F., Sadoc, J.F. and Mornon, J.P. (2004). Protein secondary structure assignment through Voronoi tessellation. Proteins 55, 519-28.
[48] Hosseini, S., Sadeghi, M., Pezeshk, H., Eslahchi, C. and Habibi, M. (2008). PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. Comput Biol Chem. 32, 406-11.
[49] Vlasov, P.K., Vlasova, A.V., Tumanyan, V.G. and Esipova, N.G. (2005). A tetrapeptide-based method for polyproline II-type secondary structure prediction. Proteins 61, 763-8.
[50] Kuhn, M., Meiler, J. and Baker, D. (2004). Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins 54, 282-8.
[51] Kumar, M., Bhasin, M., Natt, N.K. and Raghava, G.P. (2005). BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33, W154-9.
[52] Hu, X.Z. and Li, Q.Z. (2008). Prediction of the beta-hairpins in proteins using support vector machine. Protein J 27, 115-22.
[53] Efimov, A.V. (1996). A structural tree for alpha-helical proteins containing alpha-alpha-corners and its application to protein classification. FEBS Lett 391, 167-70.

[54] Wojcik, J., Mornon, J.P. and Chomilier, J. (1999). New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 289, 1469-90.

[55] Fernandez-Fuentes, N., Querol, E., Aviles, F.X., Sternberg, M.J. and Oliva, B. (2005). Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. Proteins 60, 746-57.

[56] Bansal, M., Kumar, S. and Velavan, R. (2000). HELANAL: a program to characterize helix geometry in proteins. J Biomol Struct Dyn 17, 811-9.

[57] Cartailler, J.P. and Luecke, H. (2004). Structural and functional characterization of pi bulges and other short intrahelical deformations. Structure (Camb) 12, 133-44.

[58] Milner-White, E.J. (1987). Beta-bulges within loops as recurring features of protein structure. Biochim Biophys Acta 911, 261-5.

[59] 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 U S A 75, 2574-8.

[60] Chan, A.W., Hutchinson, E.G., Harris, D. and Thornton, J.M. (1993). Identification, classification, and analysis of beta-bulges in proteins. Protein Sci 2, 1574-90.

[61] Andersen, C.A., Palmer, A.G., Brunak, S. and Rost, B. (2002). Continuum secondary structure captures protein flexibility. Structure (Camb) 10, 175-84.

[62] Fodje, M.N. and Al-Karadaghi, S. (2002). Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng 15, 353-8.

[63] 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.

[64] Majumdar, I., Krishna, S.S. and Grishin, N.V. (2005). PALSSE: A program to delineate linear secondary structural elements from protein structures. BMC Bioinformatics 6, 202.

[65] Parisien, M. and Major, F. (2005). A New Catalog of Protein Beta-Sheets. Proteins, in press.

[66] Martin, J., Gibrat, J.F. and Rodolphe, F. (2006). Analysis of an optimal hidden Markov model for secondary structure prediction. BMC Struct Biol 6, 25.

[67] de Brevern, A.G., Benros, C. and Hazout, S. (2005) Structural Alphabet: From a Local Point of View to a Global Description of Protein 3D Structures. In Bioinformatics: New Research (Yan, P.V., ed.^eds), pp. 128-187. Nova Publishers

[68] Offmann, B., Tyagi, M. and de Brevern, A.G. (2007). Local Protein Structures. Current Bioinformatics 3, 165-202.

[69] Tyagi, M., Gowri, V.S., Srinivasan, N., de Brevern, A.G. and Offmann, B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 65, 32-9.

[70] 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, 355-73.

[71] Unger, R., Harel, D., Wherland, S. and Sussman, J.L. (1990). Analysis of dihedral angles distribution: The doublets distribution determines polypeptides conformations. Biopolymers 30, 499-508.
[72] Rooman, M.J. and Wodak, S.J. (1988). Identification of predictive sequence motifs limited by protein structure data base size. Nature 335, 45-9.
[73] Rooman, M.J., Rodriguez, J. and Wodak, S.J. (1990). Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 213, 327-36.
[74] Rooman, M.J., Rodriguez, J. and Wodak, S.J. (1990). Relations between protein sequence and structure and their significance. J Mol Biol 213, 337-50.
[75] Prestrelski, S.J., Byler, D.M. and Liebman, M.N. (1992). Generation of a substructure library for the description and classification of protein secondary structure. II. Application to spectra-structure correlations in Fourier transform infrared spectroscopy. Proteins 14, 440-50.
[76] Prestrelski, S.J., Williams, A.L., Jr. 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, 430-9.
[77] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biol. Cybern 43, 59-69.
[78] Kohonen, T. (2001) Self-Organizing Maps (3rd edition), Springer
[79] 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, 833-42.
[80] Fetrow, J.S., Horner, S.R., Oehrl, W., Schaak, D.L., Boose, T.L. and Burton, R.E. (1997). Analysis of the structure and stability of omega loop A replacements in yeast iso-1-cytochrome c. Protein Sci 6, 197-210.
[81] Bystroff, C. and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281, 565-77.
[82] Schneider, R., de Daruvar, A. and Sander, C. (1997). The HSSP database of protein structure-sequence alignments. Nucleic Acids Res 25, 226-30.
[83] Bystroff, C. and Baker, D. (1997). Blind predictions of local protein structure in CASP2 targets using the I-sites library. Proteins Suppl 1, 167-71.
[84] Bystroff, C., Thorsson, V. and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301, 173-90.
[85] Camproux, A.C., Tuffery, P., Chevrolat, J.P., Boisvieux, J.F. and Hazout, S. (1999). Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 12, 1063-73.
[86] Camproux, A.C., de Brevern, A.G., Hazout, S. and Tufféry, P. (2001). Exploring the use of a structural alphabet for structural prediction of protein loops. Theo Chem Acc 106, 28-35.
[87] Maupetit, J., Gautier, R. and Tuffery, P. (2006). SABBAC: online Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace. Nucleic Acids Res 34, W147-51.
[88] Martin, J., de Brevern, A.G. and Camproux, A.C. (2008). In silico local structure approach: a case study on outer membrane proteins. Proteins 71, 92-109.

[89] 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, 662-74.
[90] Hunter, C.G. and Subramaniam, S. (2003). Protein fragment clustering and canonical local shapes. Proteins 50, 580-8.
[91] de Brevern, A.G., Etchebest, C. and Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41, 271-87.
[92] Hunter, C.G. and Subramaniam, S. (2003). Protein local structure prediction from sequence. Proteins 50, 572-9.
[93] Sander, O., Sommer, I. and Lengauer, T. (2006). Local protein structure prediction using discriminative models. BMC Bioinformatics 7, 14.
[94] Yang, J.M. and Tung, C.H. (2006). Protein structure database search and evolutionary classification. Nucleic Acids Res 34, 3646-59.
[95] Tung, C.H., Huang, J.W. and Yang, J.M. (2007). Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search. Genome Biol 8, R31.
[96] Tung, C.H. and Yang, J.M. (2007). fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res 35, W438-43.
[97] Ku, S.Y. and Hu, Y.J. (2008). Protein structure search and local structure characterization. BMC Bioinformatics 9, 349.
[98] Yang, J. (2008). Comprehensive description of protein structures using protein folding shape code. Proteins 71, 1497-518.
[99] de Brevern, A.G. (2005). New assessment of Protein Blocks. In Silico Biology 5, 283-289.
[100] Wang, G. and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics 19, 1589-91.
[101] Gowri, V.S., Pandit, S.B., Karthik, P.S., Srinivasan, N. and Balaji, S. (2003). Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res 31, 486-8.
[102] Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G. and Offmann, B. (2006). Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 34, W119-23.
[103] Tyagi, M., de Brevern, A.G., Srinivasan, N. and Offmann, B. (2008). Protein structure mining using a structural alphabet. Proteins 71, 920-37.
[104] Kurisu, G., Kai, Y. and Harada, S. (2000). Structure of the zinc-binding site in the crystal structure of a zinc endoprotease from Streptomyces caespitosus at 1 A resolution. J Inorg Biochem 82, 225-8.
[105] de Brevern, A.G., Valadie, H., Hazout, S. and Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 11, 2871-86.
[106] de Brevern, A.G. and Hazout, S. (2000). Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties. IEEE - Computer Society S1, 49-54.

[107] de Brevern, A.G. and Hazout, S. (2001). Compacting local protein folds with a "hybrid protein model". Theo Chem Acc 106, 36-47.
[108] de Brevern, A.G. and Hazout, S. (2003). 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19, 345-53.
[109] Benros, C., de Brevern, A.G., Etchebest, C. and Hazout, S. (2006). Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62, 865-80.
[110] Benros, C., Hazout, S. and de Brevern, A.G. (2002) Extension of a local backbone description using a structural alphabet. "Hybrid Protein Model": a new clustering approach for 3D local structures. In International Workshop on Bioinformatics ISMIS ed.^eds), pp. 36-45, Lyon, France.
[111] 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. In IEEE Workshop on Neural Networks for Signal Processing ed.^eds), pp. 53-72
[112] Bornot, A., Etchebest, C. and de Brevern, A.G. (2009). A new prediction strategy for long local protein structures using an original description. Proteins, in press.
[113] de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S. and Etchebest, C. (2004). Local backbone structure prediction of proteins. In Silico Biol 4, 381-6.
[114] Etchebest, C., Benros, C., Hazout, S. and de Brevern, A.G. (2005). A structural alphabet for local protein structures: Improved prediction methods. Proteins, 810-827.
[115] de Brevern, A.G., Etchebest, C., Benros, C. and Hazout, S. (2007). "Pinning strategy": a novel approach for predicting the backbone structure in terms of Protein Blocks from sequence. J Biosciences 32, 51-72.
[116] Tyagi, M., Bornot, A., Offmann, B. and de Brevern, A.G. (2009). Protein short loop prediction in terms of a structural alphabet. Computational Biology and Chemistry, in press.
[117] de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C. and Etchebest, C. (2005). A structural model of a seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC). Biochim Biophys Acta 1724, 288-306.
[118] de Brevern, A.G., Autin, L., Colin, Y., Bertrand, O. and Etchebest, C. (2009). In silico studies on DARC. Infect Disord Drug Targets 9, 289-303.
[119] de Brevern, A.G. (2009). New opportunities to fight against infectious diseases and to identify pertinent drug targets with novel methodologies. Infect Disord Drug Targets 9, 246-7.
[120] Oliva, B., Bates, P.A., Querol, E., Aviles, F.X. and Sternberg, M.J. (1997). An automated classification of the structure of protein loops. J Mol Biol 266, 814-30.
[121] Yang, A.S. and Wang, L.Y. (2002). Local structure-based sequence profile database for local and global protein structure predictions. Bioinformatics 18, 1650-7.
[122] Yang, A.S. and Wang, L.Y. (2003). Local structure prediction with local structure-based sequence profiles. Bioinformatics 19, 1267-74.

[123] Li, Q., Zhou, C. and Liu, H. (2008). Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. Proteins

[124] Zimmermann, O. and Hansmann, U.H. (2008). LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. J Chem Inf Model 48, 1903-8.

[125] Fitzkee, N.C., Fleming, P.J., Gong, H., Panasik, N., Jr., Street, T.O. and Rose, G.D. (2005). Are proteins made from a limited parts list? Trends Biochem Sci 30, 73-80.

[126] Bonneau, R., Strauss, C.E. and Baker, D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 43, 1-11.

[127] Dong, Q.W., Wang, X.L. and Lin, L. (2007). Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med 37, 1610-6.

[128] de Brevern, A.G., Camproux, A.-C., Hazout, S., Etchebest, C. and Tuffery, P. (2001) Protein structural alphabets: beyond the secondary structure description. In Recent Research Developments in Protein Engineering (Sangadai, S., ed.^eds), pp. 319-331. Research Signpost, Trivandrum

[129] Etchebest, C., Benros, C., Bornot, A., Camproux, A.C. and de Brevern, A.G. (2007). A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36, 1059-69.

[130] Thomas, A., Deshayes, S., Decaffmeyer, M., Van Eyck, M.H., Charloteaux, B. and Brasseur, R. (2006). Prediction of peptide structure: how far are we? Proteins 65, 889-97.

[131] Dudev, M. and Lim, C. (2007). Discovering structural motifs using a structural alphabet: Application to magnesium-binding sites. BMC Bioinformatics 8, 212.

[132] 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, 504-14.

[133] Bornot, A., Offmann, B. and de Brevern, A.G. (2007). How flexible protein structures are? New questions on the protein structure plasticity. BIOFORUM Europe, 11, 24-25.