Local Protein Structures

Bernard Offmann, Manoj Tyagi, Alexandre de Brevern

Preprint for Current Bioinformatics 2007

Local Protein Structures

Offmann B.[1], Tyagi M.[1+] & de Brevern A.G.[2*]

[1] Laboratoire de Biochimie et Génétique Moléculaire, Université de La Réunion, 15, avenue René Cassin, BP7151, 97715 Saint Denis Messag Cedex 09, La Réunion, France
[2] Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726, Université Paris Diderot, 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 UMR-S 726, Université Paris Diderot, 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

key words: secondary structure, protein folds, structure-sequence relationship, structural alphabet, protein blocks, molecular modeling, *ab initio*.

+ Present adress : Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, MD 20894.

Abstract

Protein structures are classically described as composed of two regular states, the α-helices and the β-strands and one non-regular and variable state, the coil. Nonetheless, this simple definition of secondary structures hides numerous limitations. In fact, the rules for secondary structure assignment are complex. Thus, numerous assignment methods based on different criteria have emerged leading to heterogeneous and diverging results. In the same way, 3 states may over-simplify the description of protein structure; 50% of all residues, *i.e.*, the coil, are not genuinely described even when it encompass precise local protein structures. Description of local protein structures have hence focused on the elaboration of complete sets of small prototypes or "structural alphabets", able to analyze local protein structures and to approximate every part of the protein backbone. They have also been used to predict the protein backbone conformation and in *ab initio* / *de novo* methods. In this paper, we review different approaches towards the description of local structures, mainly through their description in terms of secondary structures and in terms of structural alphabets. We provide some insights into their potential applications.

Introduction

Protein folds are often described as a succession of secondary structures. Their repetitive parts (α-helices and β-strands) have been intensively analyzed since their initial description by Pauling and Corey [1]. Nonetheless, this description of the 3D structures in terms of secondary structures is not simple and different major drawbacks must be carefully addressed. Indeed, the rules for secondary structure assignments are not trivial, and so numerous assignment methods based on different criteria have emerged. The greatest discrepancies are found mainly at the caps of the repetitive structures. These small differences

can result in different lengths for the repetitive structures, depending on the algorithm used. In addition, a classification of the backbone conformation limited to 3 states (the classical repetitive secondary structures and coils) does not precisely describe the protein structures, because it fails to describe the relative orientation of connecting regions. Besides, the coil state covering almost 50% of all residues corresponds to a large set of distinct local protein structures.

Thus, to circumvent these difficulties, other approaches were developed. They led to a new view of 3D protein structures which are now thought to be composed of a combination of small local structures or fragments, also called prototypes. A given complete set of prototypes defines a "structural alphabet." Different groups described these local protein structures according to different criteria. Structural alphabets have been used to approximate and analyze local protein structures and to predict backbone conformation.

This paper is divided in two parts. First, we focus on the detailed analysis of known secondary structures with respect to the different secondary structure assignment methods. Second, we present the complete panorama of known structural alphabets, *i.e.*, libraries of protein local structures used in *ab initio* methods.

Thus, we present in the *Secondary structure section* the classical and less represented repetitive structures, the irregularities within these structures, the different kinds of turns, the Polyproline II and the loops. We focus on the problematic issue of secondary structure assignment and conclude on the step from secondary structure to 3D. The *Structural alphabets section* shows firstly the structural libraries dedicated to the structure approximation, secondly the different developed prediction methods based on structural alphabet description and finally some applications.

Secondary structures

Introduction. The description of protein structures in terms of secondary structures is widely used for analysis or prediction purposes (see the example of Human Liver Glycogen Phosphorylase A [2] taken from the Protein DataBank [3] in Figure (**1**)). The secondary structures are composed of well-known α-helix [4] and β-sheet [5]. Secondary structure assignment is directly implemented in all 3D structure visualization softwares (e.g. Rasmol [6], molmol [7] or VMD [8]) which helps for the analysis of the protein scaffold. They are also used as the basis of classification of protein structures like in SCOP [9] and CATH [10] databases. Since precise modeling of the structure of a protein remains a challenge, the prediction of secondary structures is an important research area [11] and has been included in many sophisticated prediction methods, like threading [12] or *de novo* approaches [13, 14].

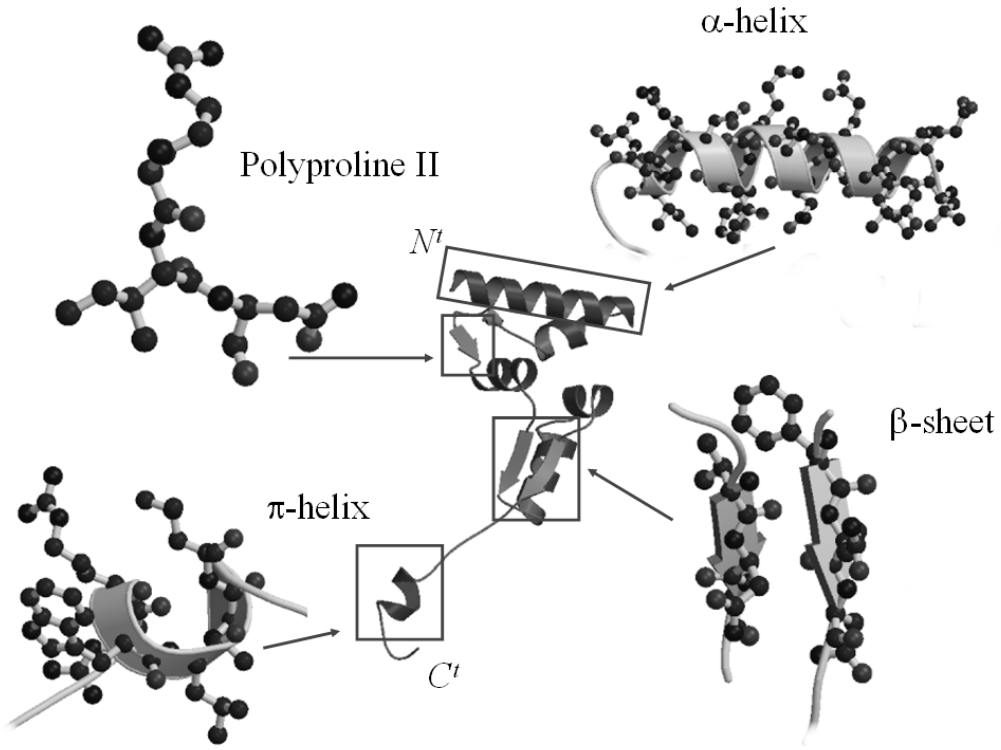

Figure 1. Example of analysis of a protein structure fragment (Human Liver Glycogen Phosphorylase A [2], Protein DataBank [3] code 1EXV, residues from 400 to 500) described by secondary structures. We can observe a long α-helix (residue 400 – 417), a Polyproline II (438 -440), π-helix (489-494) and β-sheet (452-454 and 479-481).

Classical secondary structures. Before the first protein structure was solved [15], Pauling and Corey have proposed many stable local protein structures [1, 4, 5], including two major local folds: (i) the α-helix (or 3.6_{13} helix) characterized by intramolecular hydrogen bonds between amino acid residues *i* and *i* +4 and (ii) the β-sheet composed of extended chains with hydrogen bonds between adjacent chains. They roughly represent 1/3rd and 1/5th of the residues found in proteins. Long and short α-helices do not have the same amino acid composition according to Richardson and Richardson [16] and Pal and co-workers [17]. α-helices extremities have specific amino acid propensities [18-22] and specific physicochemical stabilizations [23]. For instance, C-capping motifs of α-helices are often stabilized by hydrophobic interactions between helical residues and residues outside the repetitive structures [24], e.g. the Pro C-capping motif [25]. For instance, helix 9, the major structural element in the C-terminal region of class Alpha glutathione transferases (GSTs), forms part of the active site of these enzymes where its dynamic properties modulate both catalytic and binding functions. The importance of the conserved aspartic acid N-capping motif for helix 9 was identified by Dirr and co-workers using sequence alignments of the C-terminal regions of class GSTs and *in silico* approaches [26, 27]. Indeed, the replacement of N-cap residue Asp-209 destabilizes the complete region.

A β-sheet is formed by the association of several β-strands *via* hydrogen bonds between residues from two distinct strands [5]. Thus, a fundamental difference between the two main regular secondary structures, α-helices and β-sheets, is the non-local nature of hydrogen bonds: partners can be far from each other in the sequence space. Depending on the strand orientation, a β-sheet can be parallel, anti-parallel or mixed, resulting in different hydrogen-bonded patterns [28]. This kind of planar arrangement introduces a periodicity in the side-chain orientation: side-chains point alternatively toward one side and the other side of the sheet. As for the α-helix, the sequence specificity of β-strands has been widely studied [29],

as well as the terminal residues of strands [30]. Nonetheless, the experimental [31-33], or statistical works on pair correlation [34, 35] have not given simple conclusion to analyses the specificity of pair interaction between neighboring residues of adjacent β-strands. The β-sheet assembly is more complex than simple pair complementarities [28, 36].

A consequence of this difference is the more complex aspect of β-sheet formation and our weakest understanding of the underlying mechanism [37]. Synthetic peptide combinatorial libraries arose as a source of new lead compounds [38]. Combinatorial libraries of genes provided new proteins or protein domains [39] and peptide libraries built on α-helical scaffolds appeared as a useful strategy for the identification of new antimicrobial and catalytic synthetic α-helical peptides [40]. On the other hand, the construction of peptides libraries that fold as β-sheet structures is more recent [41]. This is mainly due to the scarcity of data and incomplete understanding of the factors determining formation of such secondary structure motifs [36, 42]. For the last three decades, more than a thousand secondary structure prediction methods have been elaborated from the early statistical approach [43-46] to complex Artificial Neural Networks and hidden Markov Model [47-51].

Other repetitive structures. 3_{10}- and π-helices are less frequent helical states representing coarsely 4% and 0.02% of the residues in proteins. The 3_{10}-helix is characterized by intramolecular hydrogen bonds between amino acid residues *i* and *i*+3 [52-54]. Majority of 3_{10}-helices are short, containing three (one-turn) or four residues but two-turn and longer 3_{10}-helices have been reported [53]. They are commonly found at termini of α-helices [55, 56] and act as connectors between two α-helices [54, 56] and their sequence content is different from α-helix [57]. An analysis of sequence and structural features of 3_{10}-helix adjoining α-helix and β-strand has recently been done. It shows that composites of 3_{10}-helices and β-strands are much more conserved among members in families of homologous structures than

those between two types of helices; often, the 3_{10}-helix constitutes the loops in β-hairpin or β-β-corner motifs [58].

In the π-helix (i.e., 4.4_{16}-helices) hydrogen bonds are formed between amino acid residues *i* and *i* +5. This helix conformation is less stable due to steric constraints [59-62]. Fodje *et. al.* [63] showed that π-helix would occur more frequently in protein structures that was previously described and would be conserved within functionally related proteins. Weaver found in 8 out of 10 confirmed crystal structures which contained π-helices, that its unique conformation was directly linked to the formation or stabilization of a specific binding site within the protein [64]. A dynamic relationship would exist between the different kinds of helices as shown for instance between α– and π–helices [65]. 3_{10}-helices, and to a lesser extent π-helices, have been proposed to be intermediates in the folding/unfolding of α-helices [66-68].

Since the description of β-strands, several analyses have shown that a strand can be found independently of a β-sheet, *i.e*. the isolated E-strand [69]. These isolated E-strands are clearly distinct from classical β-strands involved in β-sheets: (i) they exhibit particular sequence specificity, as for example an over-representation of Proline residues and (ii) they display high solvent exposure in the structures. Hence, the isolated strands are related to loops but with an extended geometry. Due to the low occurrences of these different states, the number of prediction methods dedicated to them is very limited. We could note the method SSPRO8 that performs reasonably well [49].

Irregularities in repetitive structures. The π-bulges form a particular kind of discontinuity in helical structures. Like the π-helices [64], they are not frequently observed but they seem to be directly associated to protein function [70]. For instance, Erb2 protein transmembrane domain has been shown, using molecular dynamics approaches [71-73], to

display a transition state from an α-helix to a π-bulge motif, and this has further been confirmed by experimental approaches [74]. The π-bulge is also named α-aneurism as this structural motif was revealed in an insertion mutant of staphylococcal nuclease [75]. Since then, other cases have been found, *e.g. Plasmodium falciparum* 1-Cys peroxiredoxin [76] or μ-opioid receptor [77].

In the same way, most of the observed α-helices are distorted due to presence of proline residues [78, 79], solvent induced distortions [80] or peptide bond distortions [81]. Due to these local modifications, the three dimensional path of a α-helix often becomes non-linear [82]. Barlow and Thornton found in their set only 15% of linear helices, 17% were kinked and 58% curved [56]. These conclusions were confirmed by Kumar and Bansal with an enlarged dataset [83] and very recently by Martin and co-workers [84]. Bansal and co-workers have analyzed [85] and developed specific a software to classify the helices and showed again that important proportion are in fact curved or composed of two (or more) distinct helices [86].

A β-bulge is defined as a region between two consecutive β-type hydrogen bonds, which includes two or more residues on one strand opposite a single residue on the other strand [87, 88]. Found primarily in anti-parallel β-sheets, β-bulges are common, on average twice per protein [89]. These irregularities were first classified by Richardson and co-workers into two types [87] and later in five classes by Thornton [89].

The extra residue(s) on the bulged strand not only disrupts the normal alternation of side chain direction, but also impacts the directionality of β- strands and accentuates the typical right-handed twist of β-sheets. For these reasons, β-bulges are often well conserved in proteins. Their role is not clear; they may facilitate insertions or deletions in β-strands or position crucial residues by accentuating the local twist of the strands [90, 91], as it has been shown with insertions and deletions in a β-bulge region of *Escherichia coli* dihydrofolate reductase [91] and ubiquitin [92]. As they are more exposed than other β-strands residues,

they play an important role in protein–protein interaction and in protein function [93, 94] and have been suggested to be associated with some pathologies, like the aggregation of proteins into a fibrillar structure in the case of several neurodegenerative disorders [95]. However, the underlying molecular basis for the formation of β- bulges in proteins remains poorly understood.

Turns. Regions connecting repetitive helical and extended structures, known as loops, have been extensively studied for the last decades. However, their classification is difficult to achieve namely for loop regions composed of more than 8 residues [96-99] where more precise descriptions are needed to encompass their whole diversity.

Alongside the helices and the strands, turns are perhaps one of the most interesting local fold. By definition, turns are small elements of secondary structure. They are constituted of *n* consecutive residues (denoted *i* to *i*+*n*) with a distance between Cα(s) of residues *i* and *i*+*n* that has to be smaller than 7 Å (or 7.5 Å depending on the authors). The tight turns are composed of γ-turns (*n* = 3), β-turns (*n* = 4), α-turns (*n* = 5) and π-turns (*n* = 6). The restrictive distance of 7 Å imply a particular geometry to the backbone which can therefore turn back on itself or more generally change of direction. As they orient α-helices and β-strands, they play a major role for the final protein topology. In order to not mix up with α-helices (which can be obviously considered as a succession of turns), the central residues of turns have to not be helical, e.g. residues *i*+1 and *i*+2 for the β-turns. Often, hydrogen bonds between the N-H of residue *i* and the C=O of residue *i*+*n*-1 stabilize the turn structure. Turns are classified into types according to the values of dihedral angles ϕ and ψ of the central residues. For the β- and α- turns, a deviation of ± 30° from these canonical values is allowed on 3 of these angles while the fourth can deviate of ± 45° [100].

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* (see Table 1) [101-105]. The β-turns as defined by Venkatachalam are characterized by a hydrogen bond between N-H and C=O of residues *i* and *i*+3 and types I, II, III, and their corresponding mirror images I', II' and III' were characterized [106]. These results have been confirmed with a limited set of proteins [107, 108]. Lewis enlarged this definition to several new categories: the β-turns V and V', the β-turn VI which is characterized by the presence of a Proline, the β-turn VII which is associated with a kink and the β-turn IV corresponding to all the non classified β-turns [109]. The very first documented analyses of turns in protein structures used this classification scheme [110-114]. However different turns have been excluded since then. The β-turns III and III' are too close to the 3_{10}-helix, the turns V, V' and VII are too rare and their definitions are inaccurate [100]. On the other hand, type VI were divided into 2 sub-types, that is, VIa and VIb. Lastly, Venkatachalam also noticed that some distorted type I β-turns have their ϕ_{i+2} in the β-strand region (instead of α). Later, Wilmot and Thornton precisely defined type VIII [115] which is basically based on Richardson's type Ib. Finally, Hutchinson and Thornton [116] divided type VIa in 2 sub-types VIa1 and VIa2. The definitions used by Thornton's group [89, 117] are nowadays considered as the standard (see Table 1). They are widely analyzed in molecular dynamics [118] and prediction methods have been developed [119-126]. Motifs and conformational analysis of amino acid residues adjoining β-turns in proteins have also been extensively described [127].

So, γ- and β-turns are the most important secondary structures following the α-helix and β-sheet. β-turns correspond roughly to 25 to 30% of the residues [128]. An interesting point is that they are often observed as repeated tandems leading sometimes to long series of γβ, βγ, ββ or γγ turns [129]. It is also noteworthy that γ and β turns are found associated to the same residues [130, 131].

γ-turn[a]	ϕ_{i+1}	ψ_{i+1}
Classic	75.00	-64.00
Inverse	-79.00	69.00

β-turn[b]	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
I	-60.00	-30.00	-90.00	0.00
I'	60.00	30.00	90.00	0.00
II	-60.00	120.00	80.00	0.00
II'	60.00	-120.00	-80.00	0.00
III	*obsolete*			
III'	*obsolete*			
IV[c]	----	----	----	----
V	*obsolete*			
VIa1[d]	-60.00	120.00	-90.00	0.00
VIa2[d]	-120.00	-120.00	-60.00	0.00
VIb[d]	-135.00	135.00	-75.00	160.00
VII	*obsolete*			
VIII	-60.00	-30.00	-120.00	120.00

α-turn[b]	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}	ϕ_{i+3}	ψ_{i+3}
I RS	-60.00	-29.00	-72.00	-29.00	-96.00	-20.00
I LS	48.00	42.00	67.00	33.00	70.00	32.00
II RS	-59.00	129.00	88.00	-16.00	-91.00	-32.00
II LS	53.00	-137.00	-95.00	81.00	57.00	38.00
I RU	59.00	-157.00	-67.00	-29.00	-68.00	-39.00
I LU	-61.00	158.00	64.00	37.00	62.00	39.00
II RU	54.00	39.00	67.00	-5.00	-125.00	-34.00
II LU	-65.00	-20.00	-90.00	16.00	86.00	37.00
I C	-103.00	143.00	-85.00	2.00	-54.00	-39.00

Table 1. Values of dihedral angles of γ-turns [105], β-turns [117] and α-turns [134].

[a] Allowed angles variations: +/- 40 °.
[b] Allowed angles variations: +/- 30 ° for the angles with at most one angle allows to deviate by +/- 45°.
[c] Turns which do not fit any of the above criteria are classified as type IV.
[d] Types VIa1, VIa2 and VIb are characterized by a cis-proline (*i*+2).

Shorter turns (e.g. 2 residues δ-turns) [132] and longer ones (e.g. 5 residues α-turns

[133-135] and 6 residues π-turns [136]) have been less studied. Only the α-turns has been the object of a classification scheme (see Table 1) [134]. α-turns have a functional role in molecular recognition and protein folding. For instance, residues in the α-turn in protein human lysozyme participate in a cluster of hydrogen bonds, and are located in the active site cleft, suggesting the possibility of a functional role [137]. Some are also involved in metal ion coordination [138, 139]. Moreover, α-turns are also relevant structural domains in small peptides, particularly in cyclopeptides containing 7–9 residues in their sequence [140-142]. Recently, a very elegant classification of α-turns has been proposed and the analysis of sequence – structure correspondence has highlighted the potential implication of α-turns in helix folding [143].

Polyproline II. The Polyproline II (PII) helices correspond to a specific local fold first discovered in fibrous proteins [144-146]. They contribute to the creation of coiled coil supersecondary structures characteristic of these fibrous proteins but are also found in numerous globular proteins. Because of their characteristic backbone angles and trans isomers peptide bonds, PII helix is a left-handed helical structure with an overall shape resembling a triangular prism. It is extended, with a helical pitch of 9.3 Å / turn, 3 residues per turn. This α-helical conformation is characterized by canonical values of ϕ around -75 ° and ψ around +145°, *i.e.* characteristic dihedral angle values of β-strands. There has recently been an increase of interest in PII conformations [147-151], especially in the field of molecular dynamics [148, 152-154]. Even if they are called polyproline, they are not only composed of Proline successions and some PII helices have no Proline at all [155-159] like short stretches of poly-glutamines [160]. Adzhubei and Sternberg [155] found 96 PII helices in a databank of 80 proteins. This was thought to be unexpectedly common. They found that these PII helices were highly solvent-exposed and tended to have high crystallographic temperature factors. PII

are not stabilized by salt bridges [161]. It was suggested that PII helices are often stabilized by main-chain-water hydrogen bonds (in the absence of main-chain-main-chain H-bonds), and tend to have a regular pattern of hydrogen bonds with water [162]. They are, however, still much less solvent-accessible than experimentally studied peptides. Stapley and Creamer [157] additionally suggested that local side-chain to main-chain hydrogen bonds are important in stabilizing PII helices. Cubellis and co-workers recently highlighted that PII helices are stabilized by non-local interactions [150]. They do not display strong sequence propensities in contrast with other extended conformations, such as β-strands [163]. The non-local stabilization of hydrogen-bond donors and acceptors does, however, result in PPII conformations being well suited for participating in protein-protein interactions. They are suspected to be implicated in amyloid formation [164, 165] and nucleic acid binding [166]. As recently highlighted, actual molecular dynamics parameters seem to underestimate the polyproline II and so diminish its frequencies [167].

Loops. Even after classification of protein backbone using classical three-states described above, many residues are still associated to the coil states (*i.e.*, nearly half of the residues). Several studies have hence focused on distinct conformation subsets of loops linking specific secondary structures. There are of 4 distinct loop classes (α-α, α-β, β-α and β-β) [168] and most of the studies focused on loops of less than 9 residues.

The β-hairpins correspond to loops connecting two adjacent antiparallel strands. They have been widely analyzed since they are widespread in globular proteins. Different classes have been identified resulting in the definition of structural families [169-172]. Interestingly, the short length hairpins are often characterized by a specific turn, *i.e.*, a quick return of the protein backbone [173], like a β-turn [174]. Sometimes, stabilization by disulfide bonds are observed [175]. Characteristic sequence patterns have also been highlighted, *e.g.* in

erythropoietin receptor agonist peptides [176], and used to aid loop homology modelling [177]. The β-hairpins have been well studied in molecular dynamics [178, 179]. Other types of motifs connecting two β-strands have also been analyzed like the β-β corners [173]. Orthogonal ββ motifs, i.e. consecutive strands forming an 'L' structure with an angle of 90°, have been identified [173, 180]. These motifs are often associated with particular types of loops making the connection.

α-α turn motifs, and corners, in proteins have also been described in detail [181, 182]. A recent study showed that two predominant linking backbone conformations are observed for a given short link length and some linking backbone conformations correlate strongly with distinctive inter-helical geometric parameters [183]. Wintjens and co-workers [184] presented an automatic classification procedure of protein short fragments and described ten α–α turn families that tend to exhibit some conserved sequence features. As for α-β and β-α loops, preferred conformations have also been found [96, 185, 186].

Other interesting local structures, less frequent than the turns have also been described in the coil state. For instance, the Ω-loops constitute a particular category characterized by a small distance between their extremities and an important number of contacts in their structure [187-189]. They correspond to compact globular loops mainly located at the surface of the proteins [190]. They may be directly associated with protein function [191-195] and folding [192]. An important number of studies focused on the cytochrome c and the use of compatible Ω-loops to replace existing local 3D conformations [191, 196-198]. These studies have to be viewed as complementary to investigations on closed loops, Tight End Fragment (TEF) and MIR (Most Interacting Residues) which define loop fragments that are able in three-dimensional (3D) space to nearly close their ends [199-201]. These fragments are not only composed of residues associated to coil residues but also with regular secondary structures [202].

Loops within protein families have been extensively analyzed with respect to their role in the stabilization of proteins [104]. Tramontano and Lesk used *rmsd* (*root mean square deviation*) criteria to describe structural determinants of the conformations of medium-sized loops in proteins and focus on immunoglobulins [203, 204]. This early lead already highlighted the difficulty to analyze long length loops. Complete classifications have been attempted only for short and medium size loops due to the low occurrences of longer loops and to their larger variability. For instance, Rice and co-workers showed the example of a helix – turn – strand motif found in α-β proteins that was well characterized in short loops, but not in longer loops due to the absence of local constraints [205].

Ring and co-workers [206] are the only one to propose a classification scheme for loops based on their linearity and planarity defining three categories, named linear strap loops, the non-linear and planar omega loops (not to be confused with the Ω-loops), and the non-linear and non-planar zeta loops. Their databank was composed of 432 loops. Interestingly, they proposed for the longest loops to categorize them in a forth category defined as any combination of the first three ones. They used their analysis to propose a prediction approach based on genetic algorithms named Bloop [207].

Sun and Jiang have used a non – redundant databank of 240 proteins with a resolution of less than 2.5 Å, focusing on loops of length from one to five [208]. The classification is based on a clustering of the phi-psi space into zones and 34 classes of supersecondary motifs occurring at least five times have been identified, most of them were commonly occurring supersecondary structure motifs.

In Sloop, elaborated by Donate and co-workers [209, 210], loops were classified according to their length (from one to eight residues), the type of bounding secondary structures and the conformation of the main chain. The clustering was performed thanks to a hierarchical clustering based on *rmsd* distance between the loops. Thus, 161 well populated

conformational classes were determined, and further grouped into families. For each conformational class, amino acid sequence preferences were identified. Residues located in highly conserved positions were shown to be mainly involved in the stabilization of the loop conformation or to be associated with specific functions, new classes included a 2:4 type IV hairpin, a helix-capping loop, and a loop that mediates dinucleotide-binding [210]. Their databank comprises 2,024 loops taken from 223 proteins (resolution < 2.7 Å). An approach for loop prediction was further proposed based on the identification of preferred loop conformational classes in the databank [211]. For every query, the procedure consisted in identifying among the 161 conformational classes, those that were compatible in terms of sequence preferences and disposition of bounding secondary structures. Further prediction was performed with a new evaluation dataset that comprises 1,785 loops extracted from 138 new proteins that share less than 35% of identity sequence with the initial set of proteins. Updates of this databank of supersecondary fragments were then performed, with a considerable increase in the number of conformational classes amounting 560 well populated categories with loops up to 20 residues in length [212, 213].

Geetha and Munson [214, 215] used a set of 330 proteins sharing less than 45% of sequence identity and a resolution better than 2.25 Å. The clustering algorithm proceeded with the use of two criteria: C_α distance within the loop fragments and dihedral angles of the protein backbone. They analyzed 3,313 loops of length two to eight, highlighting for instance the orthogonal architecture of the α-class proteins. They described new clusters and new relationship between sequences and structures.

Wloop is an interesting approach developed since 1996 by Chomilier and his group [216] that proposes taxonomy of the loops. Wloop proceeds by clustering loops of three to eight residues in length. Loops of the same length were placed in a common reference frame and classified within families of similar three-dimensional structures. The dataset used was

composed of 243 proteins sharing less than 50% of sequence identity. Contrarily to most of the known loop classification procedures, the clustering methodology does not rely on the nature of the neighbouring secondary structures. In total, 1,586 loops were grouped into 183 clusters. Sequence and conformational signatures were then deduced. The loop taxonomy differentiates clusters, relying on the mean distance between the first and last alpha carbon and the distance to the centre of gravity of the cluster. The database was then extended to 13,563 loops extracted from 1,411 protein structures sharing less than 50% sequence identity [97]. Using this new classification scheme, a prediction was performed using a new evaluation dataset of 47 and 48 entries sharing respectively a redundancy inferior and superior to 95% with the PDB. The Wloop web service has recently been upgraded to facilitate the newly implemented prediction scheme [217].

Wintjens and co-workers used a two-step methodology to define their loop clusters [96, 184]. The first step consisted in clustering the loop fragments according to zones within Ramachandran maps. In second instance, the loops within each class were superimposed to evaluate the quality of the clusters. A cluster was split if *rmsd* values exceeded a fixed threshold. From a dataset of 141 proteins sharing less than 20% sequence identity, they analyzed 15 αβ and 15 βα kinds of loops [96]. Previously, they had characterized 10 αα categories of loops. They focused on the most occurring clusters. This databank was used by Boutonnet and co-workers to characterize ααβ and ββα supersecondary structures [98].

ArchDB is from Oliva and co-workers [218]. They analyzed 3005 loops coming from a non-redundant databank of 283 proteins sharing less than 25% of sequence identity and classified them into five major types according to their flanking secondary structures: α-α, β-β links, β-β hairpins, α-β and β-α. The clustering algorithm, based on both the loop main-chain dihedral angles and the geometry of the bracing secondary structures, generated 56 classes that were further subdivided into 121 sub-classes. Consensus sequences were then

derived. The clustering procedure was then improved and fully automated resulting in ArchDB database [219]. In addition, updates enabled the inclusion of clusters for many long loops. ArchDB was to provide functional information. So, they have used this approach to classify the loops obtained from a set of 141 protein structures classified as kinases. A total of 1813 loops were classified into 133 subclasses (9 ββ links, 15 ββ hairpins, 31 αα, 46 αβ and 32 βα). Functional information and specific features relating subclasses and function were included in the classification. Functional loops were classified into structural motifs e.g. the P-loop shared by different folds. Hence a common mechanism for catalysis and substrate binding was sustained for most kinases [220]. ArchDB has also been used in prediction process with excellent results [221], the dataset used was based on SCOP 40 of the 1.61 SCOP release [9]. A recent application of such an approach has found more than 500 new putative function-related motifs not reported in PROSITE [222].

Li *et al.* [223] developed a database of loops extracted from a set of homologous proteins taken from FSSP database [224] where the structures had a resolution better than 2.5 Å. In their study, loops were grouped into families when they had well-superimposed bounding secondary structures. They used a hierarchical average linkage cluster analysis, which resulted in 84 loop families of 2 to 13 residues long. Subfamilies were generated and sequence features were characterized. This work enabled them to observe the diversity of loops on specific protein frameworks.

The "Loops In Proteins" (LIP) database that was developed by Michalsky and co-workers [225] is based on a non-redundant protein databank (sharing less than 20% of sequence identity) of excellent resolution (less than 1.8 Å). It included all protein segments ranging from 1 to 15 residues in length contained in the Protein Data Bank, which amounts to about 10^8. This database was used for loop prediction in the framework of homology modelling. The prediction strategy consisted in efficiently selecting loop candidates from the

database and in ranking them. The main-chain atoms of the top-scoring loop candidates were chosen as templates. Accurate prediction results were obtained, particularly for long loops.

Name	web address	database	prediction	last update
Sloop [211]	http://www-cryst.bioc.cam.ac.uk/sloop/	yes	no	2002
Archdb [218]	http://gurion.imim.es/cgi-bin/archdb/loops.pl	yes	no	2004
WLoop [97]	http://bioserv.rpbs.jussieu.fr/cgi-bin/WLoop http://psb11.snv.jussieu.fr/wloop/	no	yes	2006 *(obsolete version)*
LIP [225]	http://www.drug-redesign.de/LIP/	no	no	*Test set* (2003)

Table 2. Web services about protein loops.

Table 2 summarizes the available web services. Nonetheless, it must be noted that the major difficulty remains the definition of the regular secondary structure elements since the assignment of their boundaries directly defines the loops (see assignment methods section below). Similarly, in most of these studies, not all the protein loops were taken into account, most of the time some of the low occurring loops were withdrawn.

The assignment methods. Often the secondary structure assignment methods (SSAMs) are considered not as a specific problem, the visualization tools doing "naturally" the assignment. However as noted by Arthur Lesk in his book [226], "*what is unfortunate is that people use these secondary structure assignments unquestioningly; perhaps the greatest damage the programs do is to create an impression (for which Levitt, Greer, et al., [i.e., authors of SSAMs] cannot be blamed) that there is* ***A RIGHT ANSWER****. Provided that the danger is recognized, such programs can be useful*". Indeed different SSAMs exist. The difference between prediction and assignment of true structures is known for a long time [227]. However, the difference between secondary structure assignments is less frequently highlighted [228]. As noted earlier by Colloc'h and Cohen [30] and Woodcock and co-workers [229], a serious issue raised by the variety of methods for secondary structure

assignment is that they often yield diverging results. Different methodologies also differ in the level of detail they offer (*i.e.*, the number of secondary structures they distinguish). Here, in the following paragraphs, we describe some of the existing assignment methods. Table 3 presents a summary of available SSAMs with different number of states that they can assign.

The first developed software was proposed by Levitt and Greer and used only the C_α positions as these atoms are the best precisely defined by X-ray crystallography [230]. In this paper, the authors described another assignment criteria based on torsion angle α and hydrogen bonds. They compared their assignments with the available assignment of 33 α-helices and 25 β-sheets. They highlighted the difficulty to assign precisely the short β-sheets with only the use of C_α, so they combined inter C_α distances assignment with H-bond assignment. Nowadays, the most widely used approaches are based on the identification of hydrogen bond patterns (DSSP [231], DSSPcont [232], SECSTR [63] and STRIDE [233]).

To date, DSSP (Dictionary of Secondary Structure Protein) [231] is the most popular method. In this methodology, secondary structure segments are identified by particular hydrogen bond patterns detected from the protein geometry and an electrostatic model. After computing all the H-bonds, the algorithm first assigns helical states (with a minimum length of four residues, three for the 3_{10} and five for the π-helix) and then the β-sheets (with a minimum length of one residue). DSSP assigns eight different types of secondary structure including the aperiodic coil. DSSP is the basis of the assignment done by the Protein DataBank [3, 234]. Most of the *prediction* methods use the secondary structure *assignment* performed by DSSP to derive their parameters.

Methods	Year	Helical state	Extended state	Coil
DSSP [231] (DSSPcont [232])	1983 (2001)	α-helix 3_{10}-helix π-helix	β-strand β-bridge	turn bend coil
DEFINE [82]	1988	α-helix	β-strand	coil
PCURVE [247]	1989	α-helix	β-strand	coil
Consensus [253]	1992	α-helix	β-strand	coil
STRIDE [233]	1995	α-helix 3_{10}-helix π-helix	β-strand β-bridge	turns coil
PSEA [241]	1997	α-helix	β-strand	coil
XTLSSTR [246]	1999	α-helix 3_{10}-helix	β-strand	h-bonded turn un h-bonded turn polyproline II coil
PROSS [242]	1999	α-helix	β-strand	coil polyproline II
STICK [258]	2001	α-helix	β-strand	coil
SECSTR [63]	2002	α-helix 3_{10}-helix π-helix	β-strand	coil
VoTap [248]	2004	α-helix	β-strand	coil
t-number [250]	2005	α-helix	β-strand	coil
KAKSI [84]	2005	α-helix	β-strand	coil
Beta-Spider [251]	2005	α-helix (DSSP)	β-strand	coil
SEGNO [150]	2005	α-helix 3_{10}-helix π-helix	β-strand	coil polyproline II
PALSSE [252]	2005	α-helix	β-strand	coil
HELANAL [86]	2000	α-helix (5)	/	/
EXTENDED-BETA [259]	2002	/	β-sheet (5) β-strand	/
PROMOTIF [117]	1996	α-helix	β-strand β-bulge (10)	γ-turn (2) β-turn (10) β-hairpins

Table 3. Different available SSAMs with the states they can assign.

A recent version of DSSP called DSSPcont (Continuous DSSP) was proposed by Rost

[232]. It is based on the principle that any discrete assignment is incomplete, because the continuum of thermal fluctuations cannot be simply described. Hence, a continuous assignment of secondary structure that replaces 'static' by 'dynamic' states is used similarly to NMR studies which have emphasised the importance of structural changes over multiple length and time scales. Protein structure determination by NMR spectroscopy finds many models, the ensemble that is consistent with experimental constraints. The variations between these models result partially from experimental inconsistencies and incomplete data sets, but they are also believed to result partially from intrinsic fluctuations. Thus, DSSPcont assignments are obtained as weighted averages over ten DSSP assignments with different hydrogen bond thresholds. The continuous DSSP assignments calculated from a single set of coordinates may reflect the structural variations due to thermal fluctuations. The goal is to compensate at best the fluctuations of the assignment between the different models [232, 235-237].

SECSTR is an evolution of DSSP method. As crystallographers do not find correctly existing π-helices [238, 239], Fodje and Al-Karadaghi developed improved π-helices detection parameters. In particular, the hierarchy of detection was modified in order to focus on the correct assignment of 3_{10} and π-helices [63]. This method logically assigns more π-helices than others (10 times more than DSSP).

STRIDE [233] is also a widely-used method. It was done because DSSP often assigns too short helices. The principle of STRIDE is identical to DSSP but also takes into account dihedral angles. Hydrogen bonds are detected with an empirical energy function. The different parameters have been optimized regarding the definition of helices and strands in PDB files. The number of distinct states in this method is seven (including aperiodic), the bend defined by DSSP is here associated to classical turns. Its assignment is really close to the one done by DSSP (95% of identity). The differences are due to confusion between α-helices

and coil (1%) and to confusion between β-strand and coil (4%) [99].

PROMOTIF derives also from DSSP approach (an unpublished software called SSTRUC [240]) but focus on the characterization of γ- and β-turns, β-hairpins and β-bulges [117]. It uses the detection of repetitive structure reading the remarks of PDB files and when none is available, it supplies it using an assignment done by SSTRUC.

The periodicity of α–helices and β–strands generates some regularity in the backbone topology. Hence, some assignment methods do not use the detection of hydrogen bounds, but other characteristics of repetitive secondary structures.

DEFINE method [82], like Levitt's and Greer's method, uses only the C_α positions. It computes inter-C_α distance matrix and compares it with matrices produced by ideal repetitive secondary structures.

KAKSI is a novel assignment method that uses inter-C_α distances and dihedral angles as criteria [84]. Its principle is hierarchical (see Figure (**2**)). Firstly, the helices are assigned if the inter-C_α distances and / or dihedral angles are in a defined range. Secondly, the β-sheets are assigned if both inter-C_α distances and dihedral angles are in a defined range. The range of allowed inter-C_α distances and dihedral angle patterns have been optimized using the helices and sheets found in PDB. KAKSI have some specific features, such as a procedure for kink detection in α-helices resulting in the assignment of several distinct short helices instead of a long curved one. It is also less affected by the quality of the protein structure resolution. Indeed, higher are the resolution values, lower are the secondary structure contents.

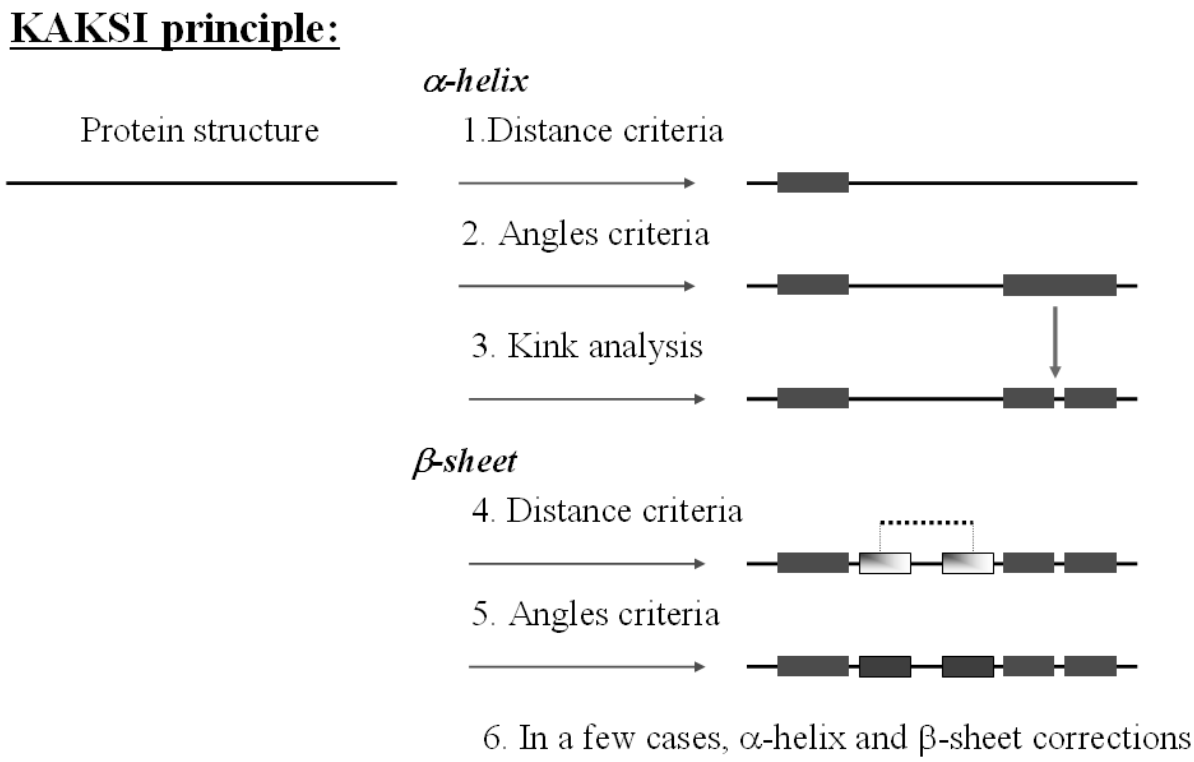

Figure 2. Principle of KAKSI assignment process. Firstly, α-helices are assigned (1) using distance criteria and / or (2) angles criteria. (3) Kinks are then detected. Secondly, the β-sheets are detected using sliding windows, if both (4) distance criteria and (5) angles criteria are within the selected ranges, the two β-strands are assigned. (6) If a α-helix and a β-strand are continuous, the α-helix is shortening.

PSEA [241] assigns secondary structures solely from C_α position using distance and angles criteria. This approach is also not much sensitive to the quality of the structures as the C_α are always the best resolved atoms. It is particularly sensitive with respect to the assignment of small β-strands.

PROSS [242] is based only on the computation of ϕ and ψ dihedral angles. The Ramachandran map is divided into meshes of 30° or 60° and the secondary structures (helices, sheet, polyproline II) are assigned according to their successions of encoded mesh. This approach has been widely used to analyze the folding of polyproline II [148, 159], the continuity between C-terminal end of α-helix and N-terminal end of β-strand [243] or the compilation of the coil library, *i.e.* a convenient repository of all remaining structure after

these two regular secondary structure elements [244].

A new method called SEGNO [245] uses also the ϕ and ψ dihedral angles coupled with other angles to assign the secondary structures. This method has been used to analyze the Polypoline II helix [150].

XTLSSTR uses all the backbone atoms to compute two angles and three distances [246]. It is especially dedicated to spectroscopy and focus on amide-amide interactions.

PCURVE methodology [247] is based on the helical parameters of each peptide unit and generates a global peptide axis. The global shapes of secondary structures are then taken into account. This approach makes use of an extended least-squares minimization procedure to yield the optimal helical description where structural irregularities are distributed between changes in the orientation of the successive peptide groups and curvature of the overall helical axis.

A recent method uses Voronoï tessellation around C_α positions to compute a contact map [248]. It is called VoTap (Voronoï Tessellation Assignment Procedure) and is based on the Voro3D software [249]. This geometrical tool associates with each amino acid a Voronoi polyhedron, the faces of which define contacts between residues. It permits the distinction between strong and normal contacts. This new definition yields new contact matrices, which are analyzed and used to assign secondary structures. This assignment is performed in two stages. The first one uses contacts between residues along the primary structure and is mainly dedicated to local assignment, *e.g*. helices. The second step focuses on the strand assignment and uses contacts between distant residues.

In the same way, Vaisman and co-workers have developed a simple, five-element descriptor, derived from the Delaunay tessellation of a protein structure in a single point per residue representation, which can be assigned to each residue in the protein [250]. The descriptor characterizes main-chain topology and connectivity in the neighborhood of the

residue and does not explicitly depend on putative hydrogen bonds or any geometric parameter, including bond length, angles, and areas.

Beta Spider is the name of an European car of the 70's but also the name of a SSAM [251]. It focuses only on β-sheet (the α-helix assignment is performed by DSSP) and for this purpose by considering all the stabilizing forces involved in the β-sheet phenomenon. Thus, not only the C=O...H-N hydrogen bonds are considered but also the C=O...C=O electrostatic dipoles and bifurcated H-bonds C=O...H-Cα. Beta-Spider also uses some geometrical factors, to make sure that the side-chains of the beta-sheet partners are pointing in the same direction.

Grishin and co-workers have recently developed a new method called PALSSE (Predictive Assignment of Linear Secondary Structure Elements) [252]. It delineates secondary structure elements from protein C_α coordinates, and specifically addresses the requirements of vector-based protein similarity searches. This program identifies two types of secondary structures: α-helix and β-strand, typically those that can be well approximated by vectors. In opposition to all the other SSAMs, this approach leads to surprising assignment for where a residue can be associated to a α-helix and also to a β-strand. It assigns about 80% of the protein chain to regular secondary structure. The authors declared that their method is robust to coordinate errors and can be used to define secondary structures elements even in poorly refined and low-resolution structures. This method is not dedicated to the analysis of protein structure but more to potentially perform a prediction.

As a consequence, these different assignment methods have generated particular problems. For example, DSSP can generate very long helices which do not correspond to the reality [56]. It is the main reason why Bansal and co-workers have analyzed and classified the helices as linear, curved or kinked [86]. In the same way, Woodcock and co-workers [229] noted that these methods do not assign the same state to certain residues, especially those located at the beginnings and ends of repetitive structures. This observation has led to the

development of a consensus approach [253] 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. That was one of the motivations of KAKSI methodology, *i.e.* to define linear helices instead of long kinked helices (see L-mandelate dehydrogenase [254] in Figure (**3**)).

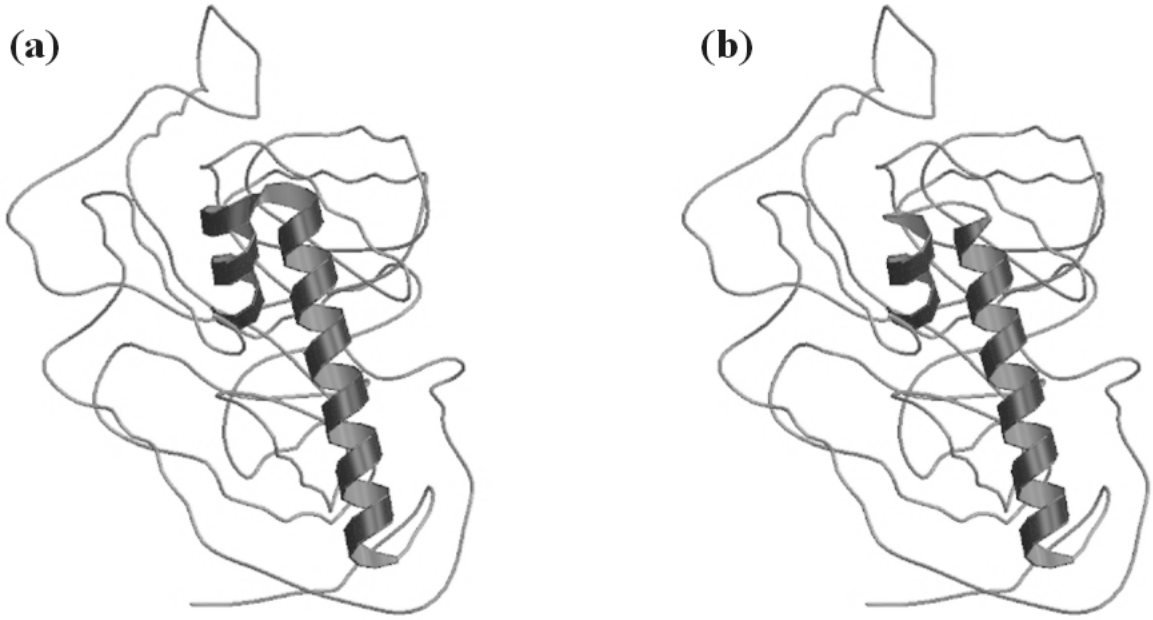

Figure 3. (a) Assignment of an α-helix of the L-mandelate dehydrogenase [254] (PDB code: 1P4C) by DSSP. (b) This helix is split in two by KAKSI.

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 α-helix (3.16_{13} helix), but incorporates the 3_{10} helix and the π-helix (4.4_6-helix) too. In the same way, β-sheets (DSSP 'E' state) correspond to β-strands implicated in parallel or anti-parallel characteristic patterns but not to isolated E-strands. This can induce difficulties in analyzing the protein structures or dynamic features.

A recent study has compared five assignment software (DSSP, STRIDE, DEFINE, PCURVE and PSEA) [99]. It used an agreement rate, denoted as C_3, which is the proportion

of residues associated with the same state between two assignment methods. The results of this work clearly highlight three points: (i) DEFINE yielded results very different from the other methods, as shown by its C_3 values, close to 62%; (ii) DSSP and STRIDE produced nearly identical assignments, with C_3 equal to 95%; (iii) all the other comparisons gave a mean C_3 of 80%, with 6–7% confusion between α-helices and coils and 12–13% between β-strands and coils. In addition, DEFINE was the only method where confusion between α–helices and β-strands was observed. These results show the difficulties for defining an appropriate length for α-helices, β-strands and coils and locating their ends. Another recent study has also shown the consequences of this differences on β-turn assignment [255].

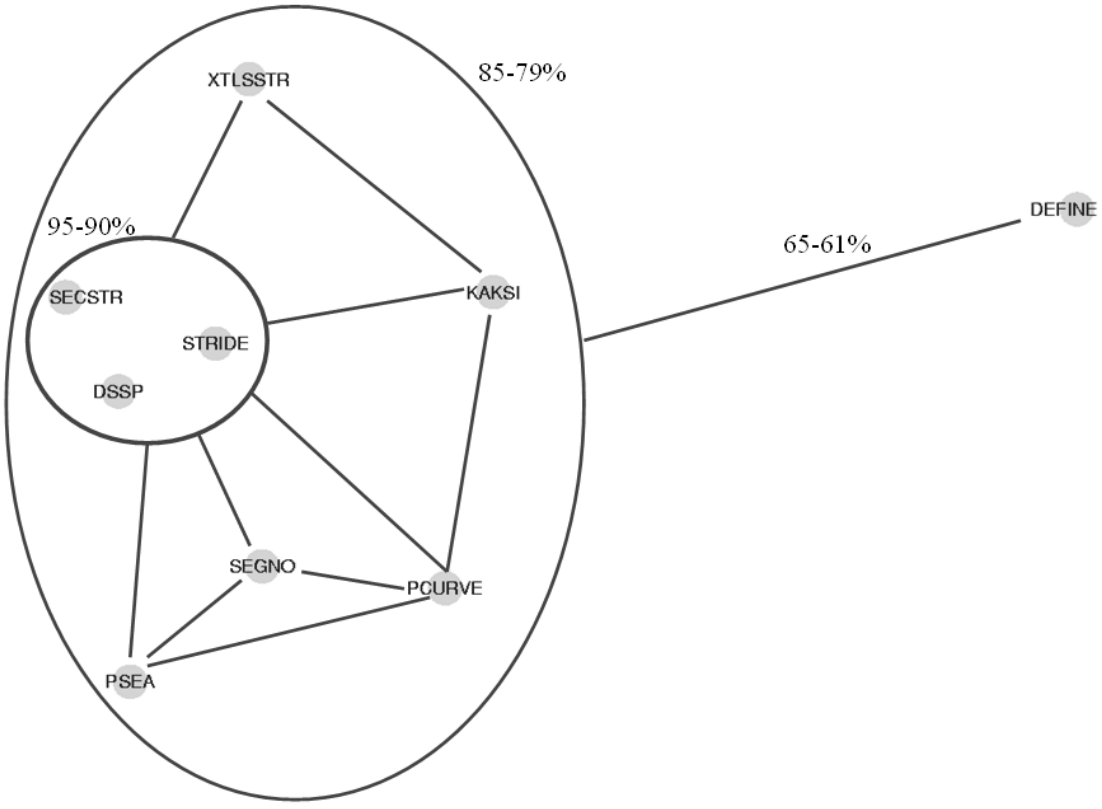

Figure 4. 2D projections of the distance between different SSAMs (adapted from [84, 99, 256] and unpublished data).

Figure (**4**) [84, 99, 256] shows a projection of comparison studies between SSAMs. A small compact cluster is found; it encompasses all the "DSSP-like" hydrogen bonds related

method, i.e. DSSP, STRIDE and SECSTR. Spreading around them are found the other methods, i.e. based on different criteria, they have average disagreement rates around 20%. DEFINE always remains distinct from all these methods because it over-assigns regular secondary structure and, with respect to this, is closer to PALSSE than the other SSAMs. It is important to note that the repetitive structures definitions only reflect a given classification and can disagree on structure description, especially on the segment extremities and on the presence of very short segments.

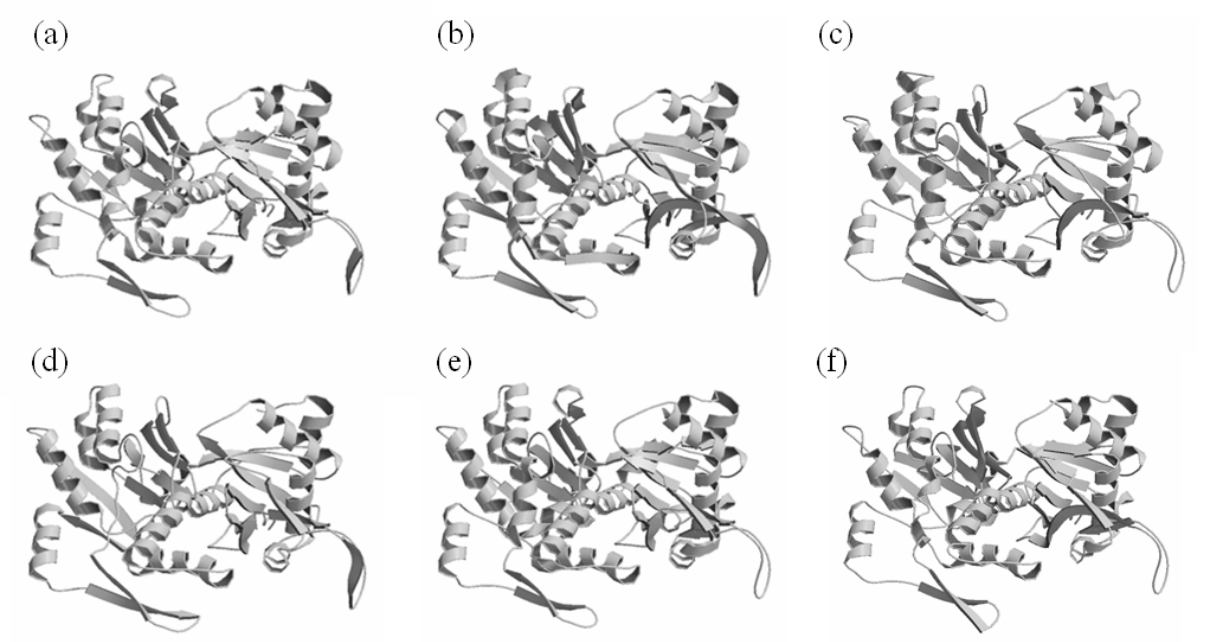

Figure 5: an example of multiple SSAMs for the beginning of bovine pancreatic deoxyribonuclease I protein (PDB code: 1ATN) [257].

Figures (**5**) and (**6**) show an example of multiple SSAMs on bovine pancreatic deoxyribonuclease I protein [257]. Figure (**5**) clearly highlights the variability of the assignments even for helices and for sheets. Figure (**6**) gives a visual representation of discrepancies in the various assignments.

Figure 6: an example of multiple SSAMs of bovine pancreatic deoxyribonuclease I protein. (a) DSSP, (b) DEFINE, (c) P-SEA, (d) P-CURVE, (e) SECSTR, (f) BETA-SPIDER.

From SSEs to 3Ds. It is obvious from the above sections that the organization of three-dimensional proteins structures can be represented as an assembly of different secondary structure elements arranged in a particular topology [100, 258] which characterizes a unique and particular fold [259]. Several distinct secondary structure combinations, generally between 2 and 4, form particular supersecondary motifs that can be found in many different folds. Many of them have been well characterized such as the simple β-hairpins [177] or more complex associations like triple-strand beta-sheets [186] and Greek key [100, 260]. Unfortunately, many folds contain very few or no super-secondary structure, *e.g*. the knottins [261], or contain secondary structure arrangements that are not very frequent in known protein structures.

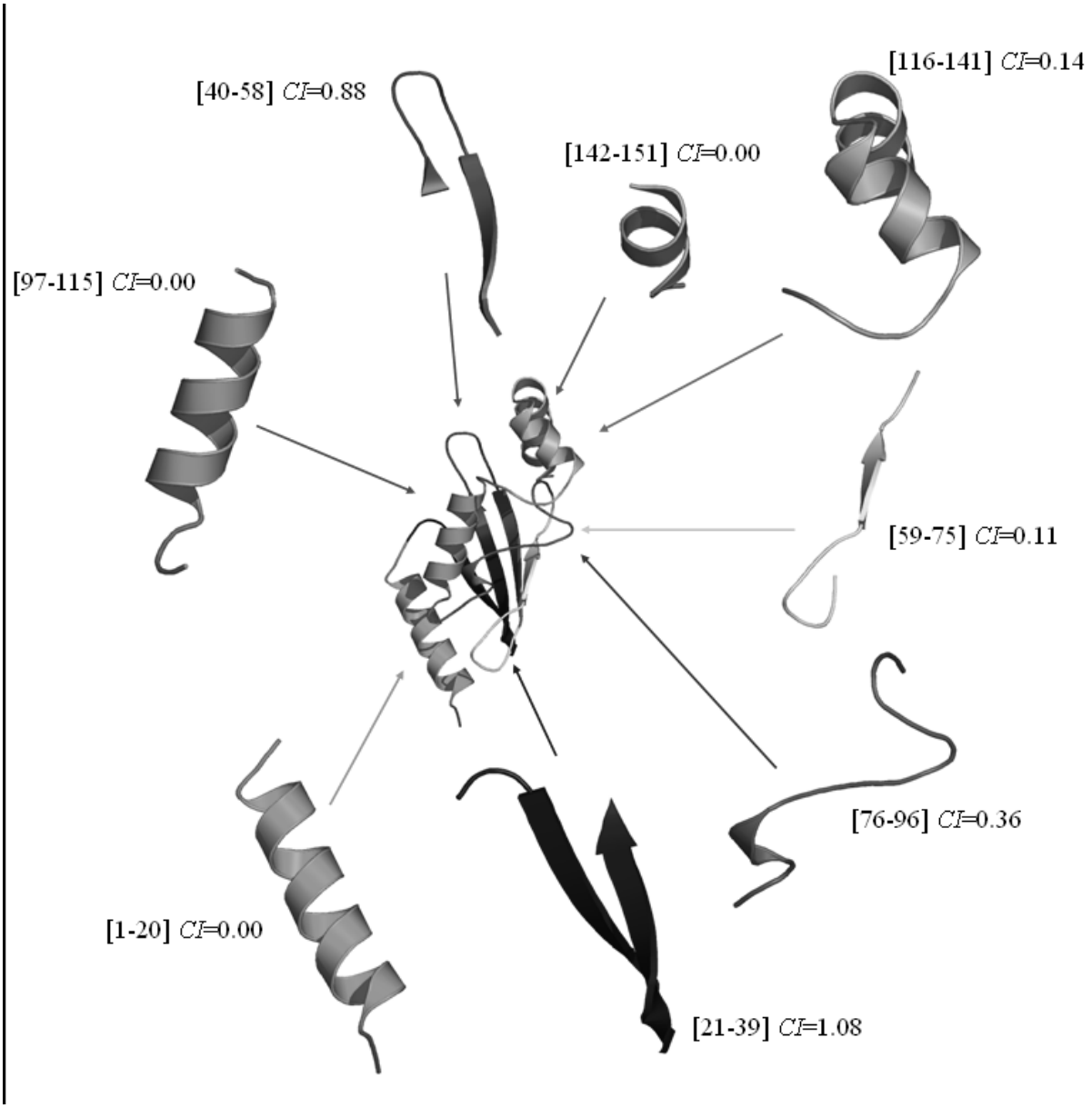

Figure 7. The different Protein Units composing the 2aak protein. Are noted the positions of the PUs in brackets with their corresponding Compaction Index (*CI*), an index measuring the number of internal contacts [267, 268].

Since the 80s, many authors have proposed different methods to hierarchically split protein structures into small compact units in the aim of describing the different levels of protein structure organization [262-264]. The rules used by these methods are quite different. To identify compact units, Lesk & Rose described the protein fragments as inertial ellipsoid and selected the most compact ones using a progressive growing approach [262]. Method proposed by Sowdhamini & Blundell to identify protein domain and supersecondary element

was based on C_α-C_α distances between secondary structures [263]. The algorithm developed by Tsaï & Nussinov used a complex scoring function, based on compactness, hydrophobicity and isolatedness, that measures stability of a candidate building block [264]. Another description at an intermediate level of organization, between secondary structures elements and domains, called Protein Units (PUs) has recently been proposed [265, 266]. A PU is a compact sub-region of the 3D structure corresponding to one sequence fragment. The basic principle is that each PU must have a high number of intra-PU contacts, and, a low number of inter-PU contacts (see Figure (**7**)).

Thus, organization of protein structures can be considered in a hierarchical manner: secondary structures are the smallest elements, protein units are intermediate elements leading to the structural domains.

Conclusion. Secondary structures are really a powerful tool to analyze the protein structures. The number of secondary structure prediction methods incredibly amounts as much as one thousand [11], beginning from simple statistical methods [43] to complex Artificial Neural Network combined with homology information like PSI-PRED [47] or SSPRO [49], that reaches 80% correct prediction. Similarly, 3-states secondary structures have been used in threading / fold recognition approach [12] and *de novo* approach [267]. Nonetheless, the assignment rules are not trivial. It is not due to the difficulty to accept a common definition with fixed values, but more to classical problem of classification where rules must be applied to delimitate the frontier of one class, such as the α-helix, and also the intrinsic flexibility of protein structure [268, 269]. This question is crucial and the scientific community seems to appreciate it more and more as the number of (and different) point of views grows. For instance, 4 different new SSAMs were proposed in 2005 only (KAKSI, PALSSE, Delaunay tessellation and Beta-Spider). A very recent elegant example of such

interest has been shown by Raveh and co-workers [270]. In a fully unsupervised manner, and without assuming any explicit prior knowledge, they were able to rediscover the existence of α-helices, parallel and anti-parallel β-sheets and loops, as well as various non-conventional hybrid structures.

Libraries of local protein structures

Introduction. As we have seen in the previous paragraphs, the secondary structures have directed our conception of the protein structure analysis [271]. Nonetheless, the secondary structures focus on two kinds of regular local structures that compose only a part of the protein backbone. The remaining residues are only assigned if they can be associated with some particular structures such as the β-turns. In fact, the secondary structure assignment is highly hierarchic. The absence of assignment for an important proportion of the residues has lead some scientific teams to develop local protein structure libraries (i) able to approximate all (or almost all) the local protein structures and (ii) that do not take into account the classical secondary structures. To start with is the precursor lead of Unger and co-workers [272] whose work has led to numerous applications, from the reconstruction of protein structures [273] to the prediction of short loops [99]. These libraries brought about the categorization of 3D structures without any *a priori* into small prototypes that are specific to local folds found in proteins. The complete set of *local structure prototypes* defines a *structural alphabet* [274-276]. The term "structural alphabet" was first introduced by Ring and co-workers to define a more precise description of the loops using 3 categories [206]. Numerous structural alphabets and names have been defined: *Buildings blocks (BBs)* for Unger and co-workers [272], *Short Structural Motifs (SSM)* for Unger and Sussman [277], *Substructures* for Prestrelski and co-workers [278], *Local Structural Motifs (LSMs)* for Schuchhardt and co-workers [279], *Recurrent Local Structural Motifs (RLSMs)* for Rooman and co-workers [280], *Structural Buildings Blocks (SBBs)* for Fetrow and co-workers [281], *Local Structures (LSs)* for Bystroff and Baker [267], *Short Structural Building Blocks (SSBs or SSBBs)* for Camproux and co-workers [282, 283], *oligons* for Micheletti and co-workers [284] and *Protein Blocks* (PBs) for de Brevern and co-workers [285]. They differ in the parameters used to describe the protein backbone like C_α coordinates, C_α distances, α or dihedral angles and in the methods used to

define them such as *k-means* [286], empirical function, Kohonen Maps [287, 288], artificial neural network [289] or Hidden Markov Model [290]. Each structural alphabet or fragment library is defined as a series of *N* prototypes of *l* residues length. *N* is highly variable, *l* only varies between 4 and 8. In the following paragraphs, we will present the most important works in this area. They are summarized in Table 4.

History. The increasing number of protein sequences and structures has supported the concept of protein evolution through the divergence of the sequences and conservation of protein structures and, in some cases, convergence of protein sequences to a common structure. The number of related protein sequences has led to the generalization of homology modeling with softwares like Modeller [291-293] or Composer [294].

In 1986, Jones and Thirup reconstructed a retinol-binding protein (RBP) using fragments of the main chain from three unrelated proteins leading to a model with a Cα *rmsd* of 1.0 Å from the known structure [295]. It was the first usage of short 3D structural motifs. This work has led to the use of known substructures for completing / refining low resolution X ray structures and suggested potential use in homology modelling for insertions. In 1989, Claessens [296] followed similar approach to rebuild the protein backbone with recurrent motifs derived from 66 high resolution structures. The constructed model was built using overlapping fragments of variable length. The final model was also less than 1.0 Å Cα *rmsd* deviation compared to crystal structures. Levitt [297] suggested construction of full atom models including side chains by pulling fragments from the PDB based on both sequence and structure consideration. This strategy is particularly efficient as most of the local protein structures are present in the PDB [298], even for the coils.

Team	Year	Name of library	Number of proteins	Number of residues	Learning method	Distance used	Prototypes number	Prototypes length
Unger *et al*	1989	Building Blocks	4 \ 82	426 \ 12 973	*k*-means	*rmsd* on Cα	103	6
Rooman *et al*	1990	Recurrent local structural motifs	75	12 978	Hierarchical clustering	*rmsd* on Cα	4	4, 5, 6 and 7
Prestrelski *et al*	1992	Substructures	14	2 347	Function	Linear distance and α angle	113	8
Zhang *et al*	1993	Structural Building Blocks	74	13 114	AutoANN	Cα distances, dihedral and valence angles	6	7
Schuchhardt *et al*	1996	Local structural motifs	136	24 239	Kohonen map	Dihedral angles	100	9
Fetrow *et al*	1997	Structural Building Blocks	116	23 335	AutoANN	Cα distances, dihedral and valence angles	6	7
Bystroff and Baker	1998	Local Structures	471	NA	*k*-means	Sequence profiles and *rmsd / dma*	13 from 82 (updated to 16 in 2000)	Structure : 3 to 15 Sequence : 8
Camproux *et al*	1999	Short Structural Building Blocks	100	19 137	HMM	Cα distances	12	4
Micheletti *et al*	2000	Oligons	75	11 086	Iterative clustering by removing the biggest clusters	*rmsd* on Cα	28, 202, 932 & 2 561	4, 5, 6 and 7
de Brevern *et al*	2000	Protein Blocks	342	87 996	Unsupervised classifier (~SOM + transitions)	Dihedral angles	16	5
Kolodony *et al*	2002	-	145 \ 200	NA (~5 000 to 9 000)	k-means simulated annealing clustering	*rmsd* on Cα	4 to 14, 10 to 225,40 to 300, 50 to 250	4, 5, 6 and 7
Hunter and Subramaniam	2003	centroids	790	156 643	Hypercosine clustering	Hypercosine Cα	28 to 16 336 (28 for prediction)	7
Camproux *et al*	2004	SBBs	250 x 2	NA	HMM	Cα distances	27	4
De Brevern, Etchebest *et al*	2005	Protein Blocks	1 407	293 507	*New evaluation*	Dihedral angles	16	5
Benros *et al*	2006	*LSP*	675 & 1 401	139 503 & 251 497	Hybrid Protein Model	PBs and *rmsd* on Cα	120	11
Sander *et al.*	2006	Structural representatives	1 999	295 411	Leader algorithm and *k*-means	Cα distance matrices	28	7
Tung *et al*.	2007	Kappa-alpha	1 348	225 523	Nearest-neighbor clustering	κ and α angles	23	5

Table 4. Synopsis of the different available local protein structure libraries or structural alphabets. The characteristics of the datasets used, the learning methods and distance criteria used are featured as well as total number of prototypes and prototypes lengths.

In fragment based studies two main approaches exists. The first consists in the description of an important number of prototypes to reconstruct precisely a protein structure. The second aims at predicting the 3D structures directly from the sequence. This last approach is only feasible when the number of prototypes is limited.

Prototyping local structures. The first approach in fragment analysis deals with approximating known protein folds after construction of a library of structural prototypes. Here the number of prototypes, *N*, is often important, and the higher *N* is, the finer is the local structure description. The different approaches towards the generation of local structure libraries and their characteristics are described hereafter.

Building Blocks. The aim of the leading work of Unger and co-workers (1989) was to obtain an important number of prototypes, called Building Blocks (BBs) able to rebuild protein structures approximated by these BBs [272]. Their method consists in calculating the average standard deviation (Cα *rmsd*) between two structures. After a preliminary calculation carried out on 4 proteins (426 residues on the whole), Unger and collaborators decided to focus on hexamers which was considered as the smallest prototype length necessary for differentiating protein fragments. They conceived a method so-called "of annexation" to create fragment clusters based on a fixed threshold of 1 Å. The process was performed in iterative refinement steps. At the end, they selected 103 BBs. They showed that 76% of the protein fragments were close to one of the BBs with a *rmsd* less than 1 Å, 92% with less than 1.25 Å while 65% were close only to one BB (with less than 1 Å) and 5% were close to more than 2 blocks (with less than 1 Å). By preserving only fragments having less than 1 Å of

difference with reality, 99% of their database was covered. This study gave the true picture of how powerful and meaningful these building blocks can be. Not only they represented regular secondary structure elements but also the complex motifs which connected helices and strands and the random coil regions.

This precursor work highlighted the difficulty to use *rmsd* as a simple measure to define local structure prototypes. For instance, using 4 other proteins, the method yielded 144 building blocks. By combining 8 proteins, to the number of prototypes jumped to 170 BBs [272]. BBs have also been used coupled with dihedral angles [299].

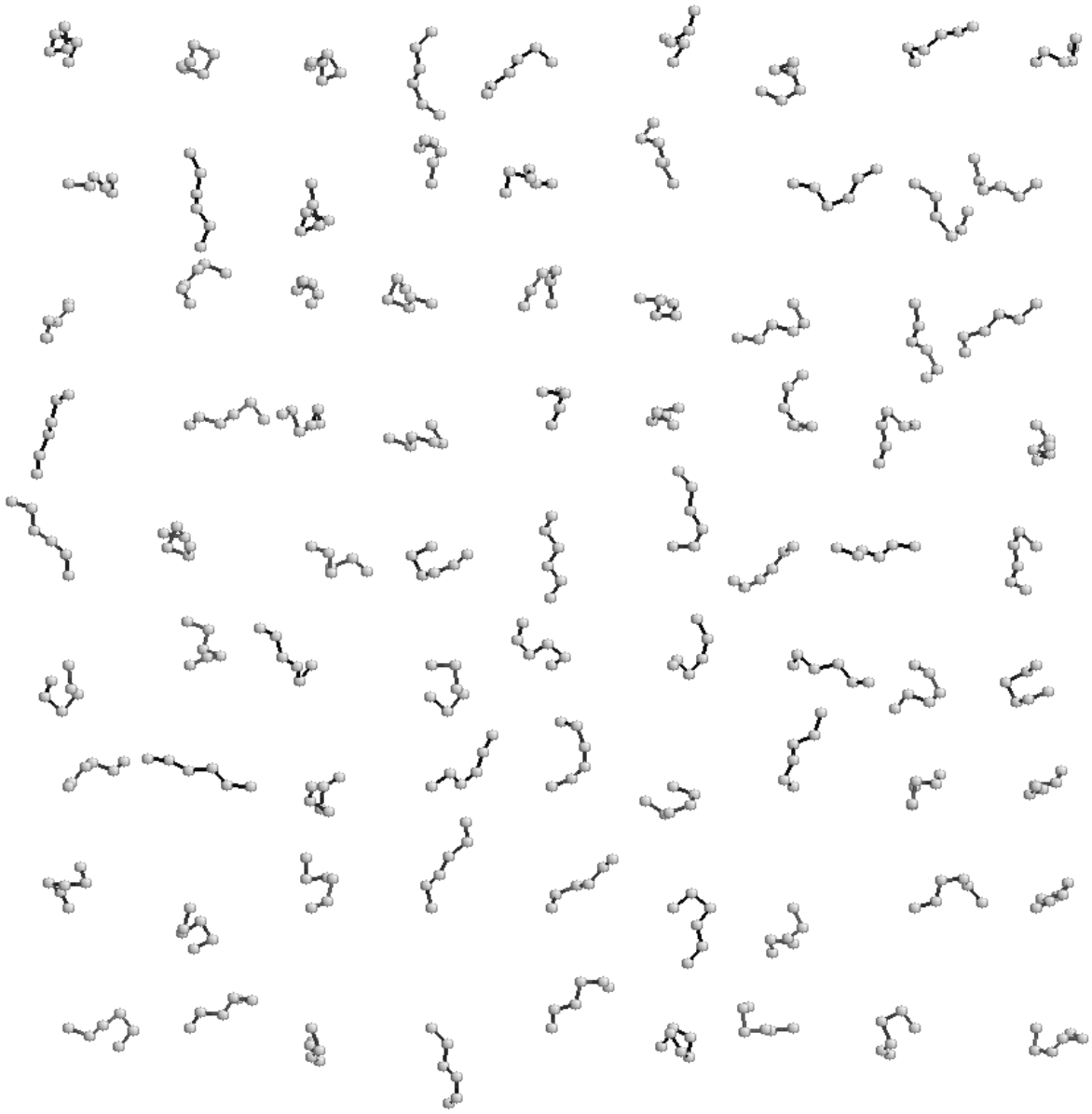

Figure 8. Representation of the 81 most frequent Building Blocks of Unger and co-workers [276]. They are presented from left to right and up to down according to their occurrence.

In 1993, Unger and Sussman proposed to keep only 81 BBs [277] which corresponded to fragments that are observed more than 35 times in their database, *i.e*. a minimal frequency of 0,3 %. They also precisely studied the blocks that corresponded to extended strand as defined by DSSP [231]. The sequence specificity for structural alphabets was also analyzed in terms of a matrix giving the occurrence of amino acid at each position for every alphabet. These BBs are available at Ron Unger's web site: http://faculty.biu.ac.il/~unger/unger.html.

Substructures. Prestrelski and collaborators created a library of local structure prototypes without any *a priori* on the type of secondary structures. They wish to use them to find structural homologies, in a similar way to that of Jones and Thirup [295], but by designing a fixed library of prototypes not specifically conceived for a single protein [278]. The method consists in generating small numbers of distinct local structure prototypes. The selected criterion for similarity is less traditional than in the preceding work. They worked primarily with a description of 4 successive Cα of the polypeptide backbone. The linear distance (DL) at a residue *i* is the sum of the distances $C\alpha_i$- $C\alpha_{i+j}$ (with *i* fixed and *j* ranging from 2 to 4). This criterion makes it possible to differentiate the repetitive structures. It was used to observe insertions-deletions in crystallographic structures. However, it has some limitations, *e.g.* it is not possible to differentiate a left propeller from a right propeller. To thwart this type of problem, they associated the angle α. To differentiate two protein fragments, a performance index dependent on two parameters C_1 and C_2 were defined using tangents of difference on angle α. The maximum value of the difference of the angles α was limited to 85°. Moreover, to work on fragments of size higher than 5, a sliding window of length 4 was used. They fixed other values to compare the fragments as they focussed on local structure prototypes of length 8. When the values of the linear distances are close, the angle α could be very different. The complementary use of the two criteria thus seems discriminating.

The calibration of the two coefficients C_1 and C_2 was done using ten prototypes of propellers, layers and periodic structures. They considered two structures as equivalent when rmsd was lower than 1 Å. The method used for groupings is not detailed, but results which recapitulate the 30 most frequent blocks are given yielding a set of 113 distinct blocks. As an illustration, the second most populated block clustered only 17 fragments. This methodology is not very conventional, but was well designed with regard to the limited number of known folds at that time [278]. It has been used to propose architecture for a serine protease [300].

Structural Buildings Blocks. Fetrow and co-workers' obtained and analyzed a limited number of local protein structures that are pertinent for protein local structure prediction [301, 302]. The methodology they have used is more complex than the previous ones. Their approach used an auto-associative Artificial Neural Network (autoANN), a particular kind of ANN with an output layer with identical dimension to the input layer. AutoANN does not perform a classical prediction, but tries to restore the information that it learned. In fact, its main interest is the hidden layer that performs a compression of the information (see Figure (**9**)). An autoANN can be considered as equivalent to a classical Sammon Map [303].

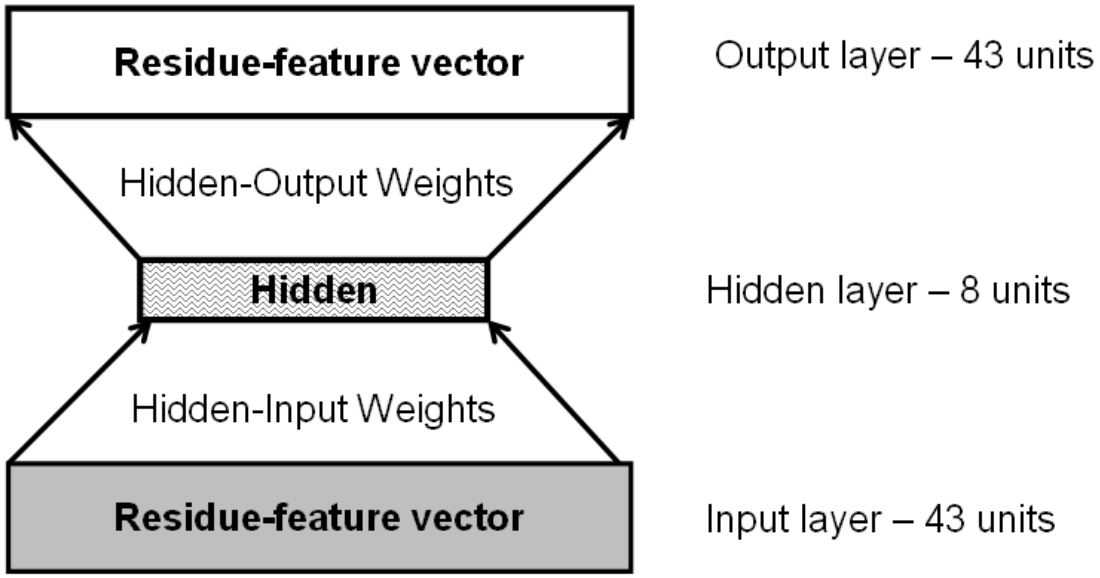

Figure 9. Principle of the autoANN used by Fetrow's group [281, 302]. Each local protein structure is encoded as a 43 unit vector (see text), the hidden layer (a 8 unit vector) performed the compression step as the output layer is again a 43 unit vector.

Here, the local protein structures are seven residues long and are encoded as distances, bond and dihedral angles. One of the difficulties is the coding of these different parameters to ensure an unbiased learning. They performed a very elegant normalization of the data as the

distances are normalized following the bimodal distribution of Cα distances on two bits, and the angles using cosine and sine values. Each local protein structures are then coded as a 43 unit vector. The databank is learned and in a second step, the coding of each local protein structures is performed using only the hidden layer (a 8 unit vector). This new encoding is used in a classical *k-means* clustering approach to determine the mean local structure prototypes. The clustering tool generated six local protein structures called Structural Building Blocks (SBBs) which correspond to the regular α-helix and β-strand, and to their respective N- and C-caps. The different databanks gave highly similar results. Nonetheless, a limitation of this approach – probably due to the limited size of the non-redundant protein databanks and the length of the local protein structure - was the absence of SBBs related to non-repetitive structures. The coding of some protein structures using these SBBs is available at http://www.cs.albany.edu/compbio/.

The study of amino acid frequencies in the SBBs showed clear preference for amino acids at specific positions and were consistent with known amino acid preferences in the case of helical regions. Importance of these structures was highlighted in loop modelling. Rooman *et al* [304] have already emphasized on the importance of the recurrent local structure motifs over secondary structure classification and on the relationship between structure and amino acid sequence. A complementary work was performed by Fetrow and Berg who showed that the use of local protein structure can reveal different distributions of rotamers classes [305].

Local Structural Motifs. In 1996, Schuchhardt and co-workers [279] used Self Organizing Map (SOM) developed by Kohonen [287, 288] to perform unsupervised classification of local structure prototypes. Contrarily to the previous studies, they did not use 3D coordinates as the direct information, but the dihedral angular values (ϕ and φ). The obtained map was of size 10 x 10 defining 100 local structure prototypes. Each motif was

constituted by a set of 16 angular values (ϕ and φ, *i.e.* 9 residues) representing most common structures found. Because of the learning method (SOM), the network was expectedly well spread out in two regions with one belonging to helices and the other to strands. The neighbouring regions gave variation on these regular structures and the in-between cells represented the various transition structures like helix-strand, helix-loop, helix-loop-helix, strand-turn-strand and other coil regions. It was suggested that this information can be used to distinguish between two structures and can give more insight to sequence-structure relationship. As an example, a neuron with over-representation of Glycine in its central position was shown.

Even though the length of the prototype is important and use of dihedral angles distance (called *rmsda* for *root mean square deviation on angular values*) discussed, the final choice of an important number of prototypes allows a fine discretization of the protein space. The only simple improvement that could have been done easily with this approach would have been to use a non planar map. A toroïd map does not have limitations during the diffusion step. A comparison between these two approaches has shown that the standard deviation of each neurons decrease [306].

Oligons. Micheletti *et al.* [284] did similar studies to construct library of recurrent oligomers in proteins which they called *oligons*. Micheletti's team used an iterative approach and computed every *rmsd* between every fragments of their databank. Their approach (i) creates clusters of local folds based on the distribution of *rmsd*, (ii) searches for the most populated cluster, (iii) selects it as an *oligon*, (iv) eliminates the fragments associated with it from the analysis, and then (v) returns backwards to step (i). Hence, it creates a hierarchy in the cluster definition: the first is more important than those that follow. The interest of their approach is to propose an increasing number of local structure prototypes, coming from the

most classical repetitive to the lowest occurring local structures. Moreover, they tested different lengths from 3 to 10 residues. They concluded that for any meaningful classification of fragments, the lengths should be between four and six residues. The coordinates of local protein structures of different libraries are available at http://www.sissa.it/~michelet/prot/repset/index.html.

They also performed an approximation of the 3D structures by reconstructing the protein structures from the *oligons* with very correct results on a dataset of 10 proteins. The optimization was performed using a classical Monte Carlo approach. The importance of the length of *oligons* was clearly highlighted, *i.e.* to ensure a similar 3D approximation for a longer length, the number of local structure prototypes must be significantly higher. The results of the sequence – structure relationship analysis is more difficult to evaluate as the number of occurrences in the dataset is low.

Librairies. Michael Levitt's work is one of the most established and precise work in this field. Moreover, he had proposed an interesting index called "complexity index" that gives an average number of states per residue and allows comparing different sets of local structure prototypes with different lengths. Park and Levitt [307] and Kolodony *et al.* [308] have constructed protein structure accurately using these small libraries of protein fragments. The method presented by Kolodony and coworkers is based on a *k-means* simulated annealing clustering approach. The different libraries were designed for the best fit and for reconstruction of protein structures. A large set of different libraries (from 4 to 300 structural prototypes, k) was designed with four different prototype lengths, r (from 4 to 7 residues). It is noteworthy that fragments never overlapped but the starting positions were sampled randomly from the dataset. The learning method consists in (i) selecting k protein fragments as k initial prototypes, (ii) associating each protein fragment with its closest prototype, with the criterion

of *rmsd*, (iii) modifying the *k* prototypes according to their associated fragments, and (iv) repeating the process until convergence in a Monte Carlo fashion. Fragments too distant from any prototype were considered outliers and eliminated. This approach gave excellent structural approximation with clear improvement over previous experiments [284, 307].

Another study pointed out that a small database-derived library of short fragments can adequately represent all protein structures, and uses this library to generate sets of protein decoys [309]. They constructed self-avoiding and compact protein decoys by repeatedly assembling pieces from their library of local protein structure. The pieces used for the assembly of the chains were chosen at random, with a limited bias based on the secondary structure content of the protein used. Despite the extreme simplicity of this method, the sets of decoys were of excellent quality [309]. The coordinates of different libraries of local protein structures are available at http://csb.stanford.edu/rachel/fragments/ .

It must also be noticed that M. Levitt proposed an index called complexity index useful for comparing different libraries [307]. It corresponds to an average number of states per residue.

(Short) Small Building Blocks. In 1999, Camproux, Hazout and *co-workers* used Hidden Markov Model (HMM [290]) to identify recurrent short structural 3D building blocks (SSBs) [282]. The major interest of HMM is to take into account the local dependencies between the different local protein structure, *i.e.* the transitions between the established states. The final model was able to give the most probable path connecting various SSBs and so could be introduced in the reconstruction of the protein backbone. Each SSB is four residue long defined by a vector of four distances: three distances are between the non-consecutive C_α atoms and the fourth distance is the projection of last C_α on the plane formed by first three C_α atoms. In the first study, a final number of 12 SBBs was selected. Analysis of patterns of

SSBs between regular secondary structures was performed. It was found that series of fragments between secondary structures were more dependent on following structure than the preceding one. It was also found that the combination of SSBs was specific to the kind of regular structures it connects [283].

Recently the number of SSBs was re-evaluated with regard to Bayesian Information Criterion (BIC [310]). This index estimates if the growth of the number of states for a given HMM is informative or not. The analysis was performed with two independent non-redundant datasets and increased the number of prototypes to 27 SBBs, which all displayed strong connection logic. The quality of the structural approximation was assessed also with respect to protein structure reconstruction. An interesting point was the finding of two very close α-helical cores with slight structural differences (*rmsd* < 0.15 Å) but with very different transitions with other states. The reconstruction process of the protein structure was improved using Go-based energy function and greedy algorithm [273, 311]. This approach was improved with force field such OPEP [312] or EEF1 [313]. Recently, an analysis of this structural alphabet had shown that it is suitable for deciphering some local sequence – structure relationship [314]. One application of this structural alphabet has been the web server SCit that allows analysing the protein side chain conformation [315]. It is available at http://bioserv.rpbs.jussieu.fr/cgi-bin/SCit. Another recent one is SABBAC [316], an on-line service devoted to protein backbone reconstruction from Cα trace. It is based on the assembly of fragments taken from a library encoded into SBBs. The assembly of the fragments is achieved by a greedy algorithm, using an energy-based scoring. It can be accessed at http://bioserv.rpbs.jussieu.fr/SABBAC.html.

Prototyping and prediction of local structures. The second approach in fragment analysis deals with predicting protein fold from sequence using fragment libraries, giving more insight

into sequence to structure relationship. To perform prediction from sequence, the number of prototypes, *N*, must be small enough as correct prediction requires limited number of local conformations as shown by Rooman & Wodak 's and Fetrow's works [281, 304].

To capture most of the local folds, it is necessary to optimize the number of states. It should be sufficiently large to approximate correctly the local folds and limited enough to ensure correct prediction levels from sequence alone. An alphabet composed of $N = 10$ to 20 states is particularly suited for this goal [267, 285]. Bystroff and Baker's I-Sites is one of the most interesting local protein structure libraries. It has been used with a high efficiency for improving *de novo* methods [317, 318].

Rooman and Wodak. Following their leading works on the local aspect of secondary structure, on the critical size of protein databanks to reach correct secondary structure prediction [319] and on the characterization of turn specific amino acid signatureq [320], Rooman and Wodak described local protein structures that can be considered as stable structural units that fold independently of the rest of the structure [280, 321]. The training method is based on a hierarchical classification, which uses as criterion the *rmsd* between protein fragments. First, the fragments are compared two by two by calculating *rmsd* on C_α of the protein backbone. Then, a hierarchical clustering is carried out. Lengths going from 4 to 7 Cα were tested and, for each length, 4 distinct groups were selected. As expected, a template corresponding to the α-helix and another to the β-strand was identified for each length. Since these structures have high correlations between sequence and structure, they proposed a prediction from sequence alone [304]. The prediction approach is related to the classical statistical prediction of secondary structures [320]. However, they also proceeded to a filtering of the data to ensure a stronger[322] relationship between the local protein structures and their associated sequences. With this approach, the authors found, starting from the

sequence, a prediction rate ranging between 41% and 47%. This rate must be compared with the rates for secondary structures prediction that was prevailing at that time and which was about 60% per 3 states [323, 324]. This prediction approach has been used to predict local backbone conformation based on zones of Ramachandran map [325, 326].

I-sites. Bystroff and Baker developed an innovative method for local protein structure prediction based on library of short sequence patterns having strong correlation with 3D structure. Their method is based on previous observations done by Han and Baker who identified recurring local sequence motifs using automatic clustering [327] and extended their results to the characterization of corresponding local protein structures [328-330]. Following these results, Bystroff and Baker developed an iterative method that would optimize the correspondence between protein sequence content and local protein structures, leading to a high correlation of the sequence – structure relationship [267]. Based on HSSP families [331], sequence based clusters were created and the most frequently occurring structure in each cluster was chosen as the structural "paradigm". Then through an iterative process, related to a *k-means* approach, a dynamic modification of the clusters was performed as (i) local structure protein with structure different from the paradigm were removed, (ii) sequence patterns were recalculated from the remaining members, (iii) new members were identified and (iv) the paradigm was assessed again. Different criteria were also used to ensure a high correlation between sequence content and 3D local approximation for each cluster [267].

At the end, a library comprising 82 sequence patterns (corresponding to HSSP profiles) of 3 to 19 residues long was obtained. These 82 profiles were structurally aligned and grouped into 13 different sequence-structure motifs. The generic term for these clusters is "I-sites". The library not only contained previously defined sequence-structure correlation but presented few new relationships like diverging type-II β-turn, a frayed helix, a proline-

terminated helix, and serine-containing β-hairpin. It must be noted this approach does not use all the protein fragments of the databank, *i.e*. not all the protein fragments are encoded.

Following this characterization, a prediction method was elaborated. This prediction combined a secondary structure prediction done by PHD [332, 333] and an adequacy score of the 82 clusters. It must be noticed that I-sites length in the prediction step ranges not from 3 to 19 residues length but are only of 3 and 9 residues length. The secondary structure prediction has a very limited weight. The prediction accuracy was based on the use of (ϕ,ψ) dihedral angles of the I-site paradigm and compared with the true dihedral angles. A good prediction corresponds to a protein local structure with no predicted dihedral angles with more than 60° from the reality. Bystroff and Baker have proposed protein local structures ranging from 3 to 17 residue length, but in their prediction protocol, they used fixed length of 8 residues. The results of their approach lead to excellent results in regards to the number of their I-sites [267].

This process was tested with correct results during the second Critical Assessment of Structure Prediction experiment [334] and a very good correlation with NMR characterization of α-spectrin SH3 [335].

The same group also built a hidden Markov model called HMMSTR for protein sequences based on I-sites library [336]. This HMM was built using overlapping I-sites with an updated databank and 3 new I-sites were identified. Using the previous prediction method, the prediction rates surprisingly increased, probably due to the computation of accuracy rates [337].

In HMMSTR extended I-sites library, (i) new sequence-structure correlations such as α-α corner, Type-I' β-hairpin and Glycine-rich α-helix N-cap were identified, (ii) adjacencies of different sequence-structure motifs *i.e*. non-random transitions between various sequence motifs were described and (iii) overlapping motifs are presented in more condensed form,

resulting in better description with fewer parameters. HMM model is useful for *ab initio* structure prediction and homology modelling specially loop building problem.

Rosetta algorithm, *de novo* protein structure prediction method [14, 338] from Baker's lab has been one of the most successful prediction methods in the CASP experiments [318, 339]. It had predicted accurate models that have correct global topology, correct architecture of secondary structure elements and functional residues often clustered in active site region [14, 340]. Rosetta method has been extended to other protein modelling problems like *de novo* protein design [341, 342], protein-protein docking [343], protein modelling based on limited experimental data [344, 345] and loop modelling or modelling structurally variable regions in homologous proteins [346]. It must be noticed that I-SITES, HMMSTR and ROSETTA have combined in a fully automated *ab-initio* protein structure prediction that gives excellent results for instance for secondary structure prediction [347]. HMMSTR / Rosetta server is available at: http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php.

Protein Blocks. Another set of structural alphabet was identified by de Brevern, Etchebest and Hazout [285]. The main purpose of this work was to construct a structural alphabet composed of local protein structures both able to approximate and to predict the local structure from the sequence. It differs from the previous studies as it has been constructed in two steps: (i) approximation and (ii) prediction. Different sets of local protein structures have been defined and the choice of the selected set has been directed both by a correct structural approximation and by an acceptable prediction rate. It diverges from I-sites approach as (i) all the local protein fragments are used and (ii) no profiles were used.

This structural alphabet is composed of 16 local structure prototypes of five consecutive C_α, called Protein Blocks (PBs), representing local structural features of proteins (see Figure (**10**)). Protein blocks were identified using unsupervised cluster analyzer, taking into account

the sequential dependence of the blocks, *i.e.* related to Kohonen maps [287, 288] and hidden Markov models [290]. They are overlapping fragments, 5 residues in length, encoded as sequence windows of 8 consecutive dihedral angles (ψ, ϕ). The distance used in the training approach was the root mean square deviation on angular values namely *rmsda*, *i.e.* a simple Euclidean distance on angular values [279]. Each of the PBs is denoted by letters *a* to *p*. The relationship of PBs with secondary structure was also studied and can be characterized by their secondary structure composition. For example, PB *m* forms the central part of helix and PB *d* is ideal for β-strand. Each represented respectively 30.0 and 18.9% of the protein fragments, the others ranged between 8.7 and 1.0%. PBs from *a* to *c* and *d* to *f* are mainly concerned with N and C caps of β-strand respectively. Similarly PBs *k, l* and *n, o, p* form N and C caps of helix respectively. The remaining PBs labelled from *g* to *j* are mainly concerned with coils. PBs are not only capable of representing regular secondary structures but also the subtle variations present in the beginning and end of regular structures along with local features of coil regions. This set was selected because it ensured a good structural approximation and a correct prediction rate.

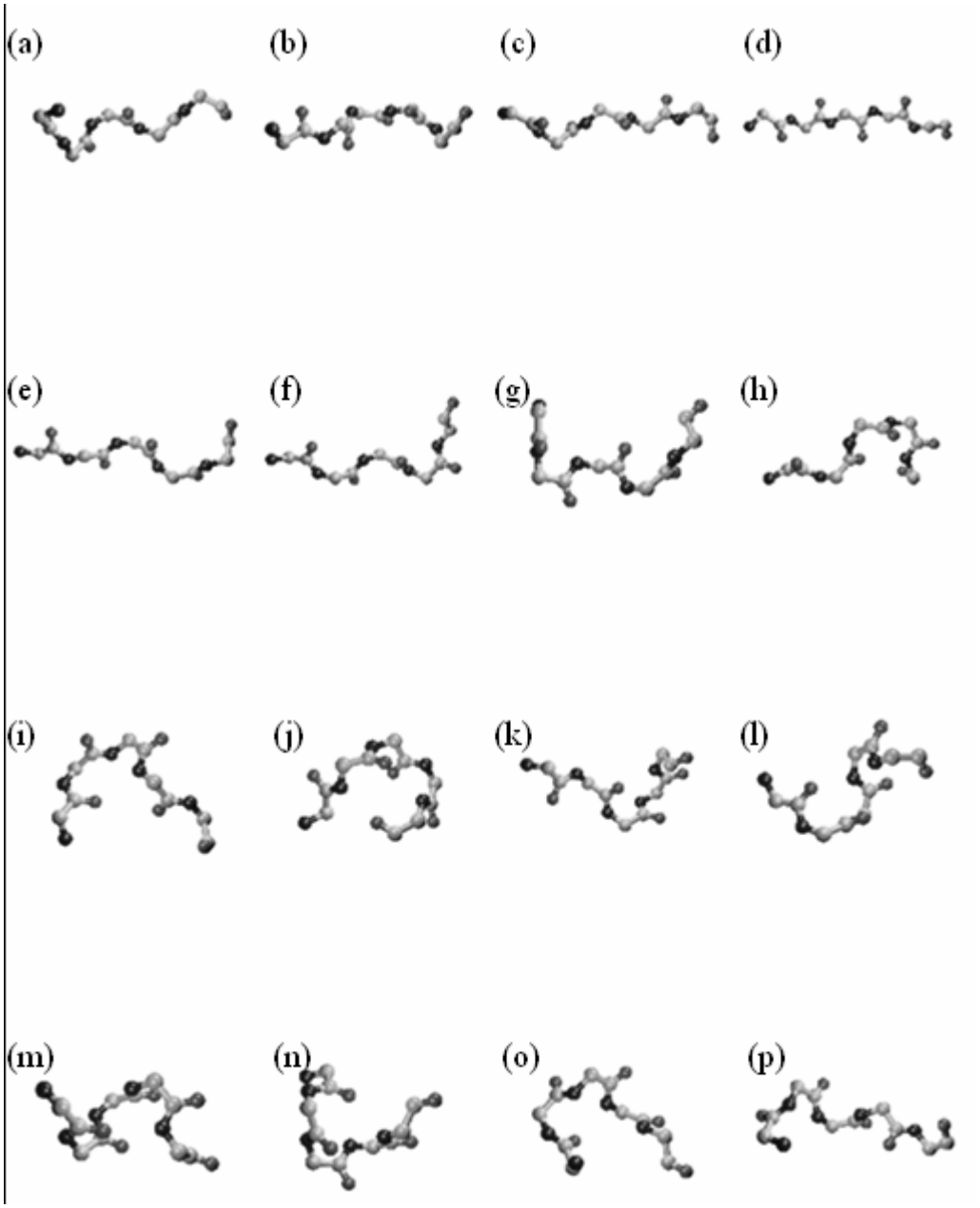

Figure 10. From left to right and top to bottom the 16 Protein Blocks (labelled from *a* to *p*) are shown. For each PB, the N-cap is on the left and the C-cap is on the right.

Analysis of PBs showed also that transition from one PB to another one is highly constrained due to protein topology. The three main observed transitions amounts to a mean value of 76%.

The critical assessment of Protein Blocks in terms of the stability of their distribution frequencies, in terms of their main transitions and in terms of their geometrical features has been extensively reported [348]. It was hence verified that PB definitions remained valid after

the size of the databank was more than tripled (from 86,628 [285] to 293,507 residues). It was also highlighted that the distribution of PB frequencies remained equivalent in all the non-redundant databank. The transitions between all the PBs remained also highly constant.

Besides, it was shown that the use of *rmsda* distance allows a good discrimination. The comparisons of the *rmsd* and *rmsda* values for PBs show that *rmsda* is more sensitive to small structural variations than *rmsd*. The *rmsda* is more discriminative to split up the N- or C-caps of repetitive structures from the core of the repetitive structures. For instance, PBs *m* and *n* have discriminative *rmsda* values for a low non-discriminative *rmsd* values.

PBs have also been compared with the classical secondary structures. As shown above (see *comparison of secondary structure assignment section*), the comparison with classical secondary structure assignments is not trivial as the assignment methods differ and the correspondence with PBs is not direct. For instance, the PB *d* has the geometrical feature of a β-strand, but is assigned by PSEA [241] and by STRIDE [233] to coil state with a rate equal to 19.6% and to 29.0% respectively [348]. Specific correspondences have also been highlighted like the one between some PBs and Polyproline II core and C-cap [256].

The extension of Protein Blocks concept to overlapping sequences of five PBs corresponding to 9 Cα long has lead to characterization of Structural Words (SWs) [349]. Combination of most of the SWs in a protein network encompasses more than 90% of protein residues of a non-redundant protein structural database. Interestingly, more than 80% of the coils are included in the network. This SWs concept provides a good structural approximation for fragments of nine C_α long. Regular structures are represented by most frequently occurring SWs *e.g. mmmmm* or *ddddd* related to α-helix and β-sheet respectively. SWs also represent variation of these regular structures at N and C caps *e.g. lmmmm, klmmm, cdddd, ddddf* etc. Most of the SWs are overlapping either on one or both ends, last four PBs of a

given SW may be identical to the first four PBs of another and vice versa. For example, *mnopa* overlap with *nopac, nopab* and *nopaf*. Other examples of overlapping SWs involving PB *d* related to β strands are *ccddd* with *bccdd, cdddd, cdddf,* and *cddde* indicating the flexibility of these local structures. SWs containing repetitive PB *m* like *mmmpm, mlmmm,* and *mklmm* with one or two PB change correspond to irregular, curved or kinked α-helices.

The prediction performed with the PBs from the sequence information alone is based on Bayes' rule. Figure (**11**) presents the principle. (a) All the protein fragments associated to a given PB are used to compute (b) an occurrence matrix that represents the frequencies of the amino acids. These frequencies are normalized in accordance to the frequencies of the amino acids in the databank. (c) Then, to predict the PBs for a new sequence of unknown structure, the probability of the corresponding amino acid is evaluated (d) using the Bayes' theorem. (e) The predicted PB is associated with the best prediction score. The prediction rate Q_{16} so reached 34.4% [285].

A Bayesian probabilistic approach has also been performed with the SWs and had lead to an improvement of 4% of the prediction rate but with no optimization of the sequence-structure relationship [349]. A new approach called pinning strategy has recently increased this prediction rate [350].

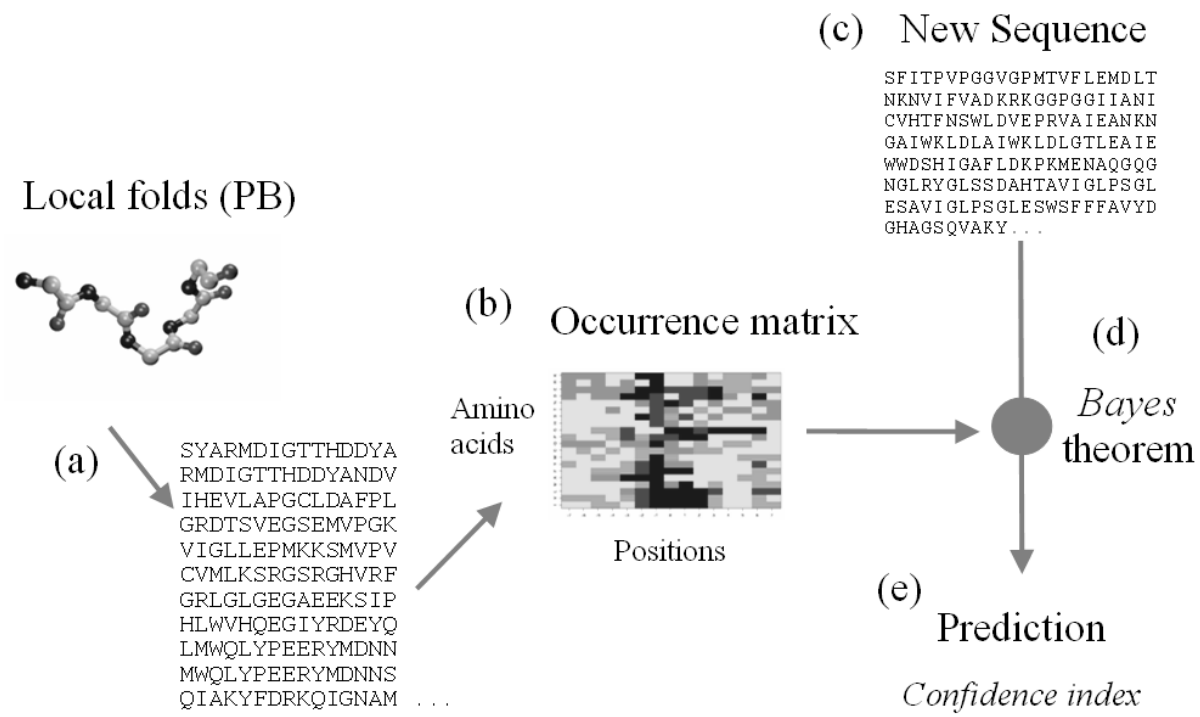

Figure 11. *Prediction principle*. (a) All the protein fragments associated to a given PB are selected. (b) From this information, an amino acid occurrence matrix is computed. (c) A new sequence is presented and (d) the occurrence matrix associated to the PB is used to compute a prediction score based on Bayes' rule. (e) The best score is conserved as the predicted PB.

Different sequence clusters may be associated with the same fold. In a previous study, we developed the concept of "*n* sequences for one fold", with *n* the number of sequence clusters associated with a given PB. For each protein block or *PB*, the corresponding set of sequences is divided into *n* groups. Each is represented by one amino acid occurrence matrix (as seen in Figure (**11**)b). This approach is based on an unsupervised clustering close to SOM [287, 288] and is called Sequence Families. With this approach, the Q_{16} rate improved to 40.7% [285].

This approach has further been improved recently using an approach related to simulated annealing; it had lead to a Q_{16} rate equaled to 48.7% [351]. Moreover, this improvement was distributed to all the PBs without decreasing the less frequent PBs. A study about the combination of secondary structure prediction with simple Bayesian prediction or Sequence

Families had shown no particular improvement due to the limited correspondence between the three states of secondary structures and the 16 PBs [351].

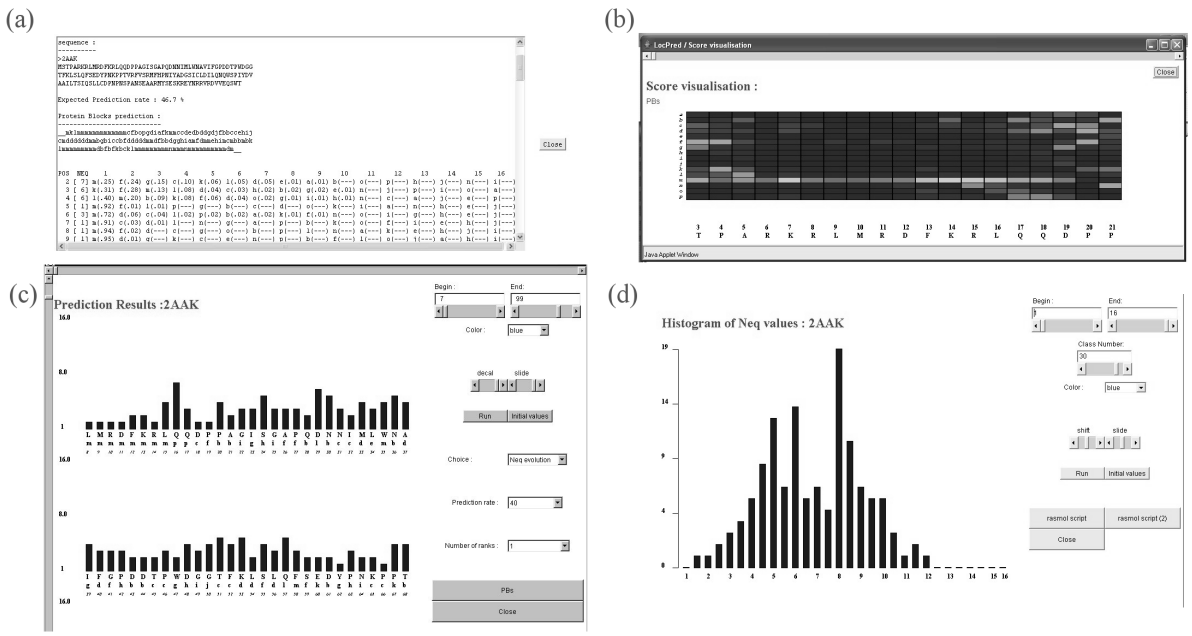

Figure 12. A screenshot montage of *LocPred* [352] output. (a) Simple text presentation of the results with corresponding probabilities. (b) Graphical representation of the prediction results. (c) Confidence index at each residue. (d) Confidence index for the whole protein.

These different predictions are available to the scientific community through a Java applet, *LocPred* [352]. From the sole information of the sequence (see Figure (**12**)a), simple prediction or prediction with Sequence Families can be performed. The results are given with corresponding probabilities both in text and graphically (see Figure (**12**)b). Moreover, the use of Bayes' rule allows computing a confidence index of the prediction. This confidence index can be local, i.e. for each residue position (see Figure (**12**)c), or represented for the whole protein (see Figure (**12**)d). *LocPred* is available at http://www.ebgm.jussieu.fr/~debrevern/LOCPRED/index.html and through the RPBS web server [217] http://bioserv.rpbs.jussieu.fr/LocPred/index.html.

Another approach for predicting local structure in terms of PBs was recently developed

by our team (Offmann *et al*, in preparation). Extraction of large number of pentapeptides from protein structure space has been used to build a database of PB annotated pentapeptides. Using this pentapeptide database, a novel knowledge-based PB prediction method was proposed. The scheme that was used for the prediction is summarized in Figure (**13**) and was assessed on a representative subset of 100 families from SCOP where each homologous member was tested using a jack-knife approach.

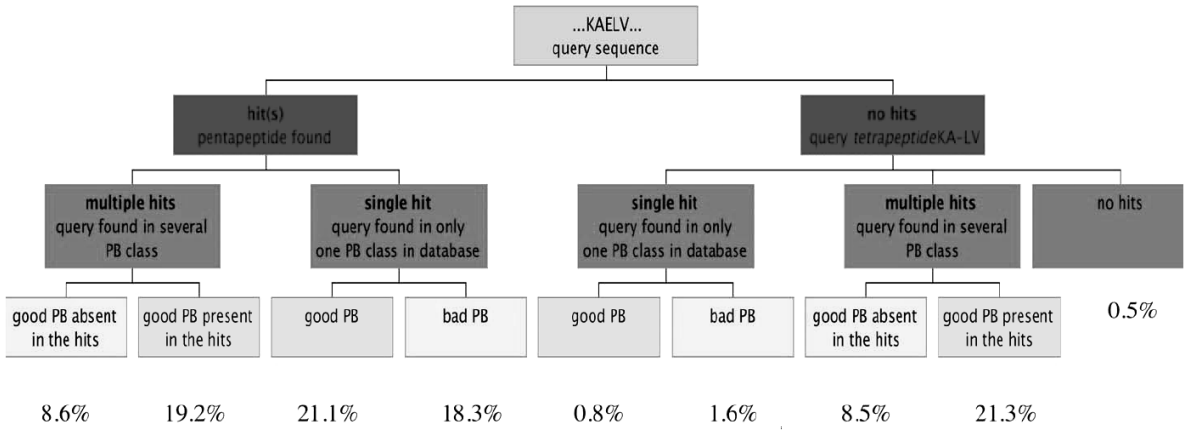

Figure 13. Scheme for predicting local backbone structure in terms of Protein Blocks from amino acid sequence by querying a database of pentapeptides that was extracted from known structures (Offmann *et al*, in preparation). Using a sliding window of 5 residues, sequences of pentapeptides are extracted and queried against the database. When a pentapeptide is present in the database ("hits"), two situations can be distinguished : (i) either it is mapping to a single PB ("single hit") or it is mapping to several PBs ("multiple hits"). The occurrence of the true PB in the list of hits is counted. When a pentapeptide is absent in the database ("no hits") the corresponding tetrapeptide (by considering residues 1, 2, 4 and 5) is searched. Three situations can arise; two corresponding to the "hits" and "multiple hits" situations described above and the third to the absence of both the pentapeptide and tetrapeptide in the database ("no hits"). The occurrence of the true PB in the first two situations is counted.

Preliminary prediction results have shown the viability of this approach. First, 67% of the query peptides matched (hit) to an entry in the database and in more than 60% of such cases, the true PB was indeed present. Second, by relaxing the identity of the residue in the middle position and by checking the availability of the corresponding tetrapeptide in the database helped considerably to increase the overall success rates. On average, a global 62% success rate was achieved. This knowledge-based prediction scheme is available at

http://bioinformatics.univ-reunion.fr/PBE/pb_prediction/.

Hunter and Subramaniam. The approach developed by Hunter and Subramaniam is more classical and encompass long fragments of 7 residue long [353]. The local protein structures were compared using a hypercosine clustering method, *i.e*. a faster method than *rmsd* superimposition. Then, the authors chose a threshold to select the final number of clusters called centroids. They performed an in-depth analysis of the parameters used to select the centroids, but the analysis of the obtained local protein structures is weaker. Indeed, the clustering approach seems to create big unbalanced clusters. For instance, with a threshold that leads to find 28 centroids, 13 centroids represent 99.9% of the protein fragments and 4 more than 75%.

In a second step, they performed a prediction method highly similar to de Brevern *et al*. using Bayes' rule with a slight modification to take into account of the repetitive structures [354]. The Bayesian prediction rate reaches 40%. However, this value is strongly biased towards the most frequent centroids at the expense of the others. Eight of the 28 local protein structures are predicted at a rate above 20% but only four above 50%. In addition, eleven centroids are not predicted at all. The approach developed by Hunter and Subramaniam hence can be improved to find an equilibrium between a better description of the protein structures and a prediction better distributed. Moreover, a surprising change in the approximation done by the 28 centroids appeared without any explanation between the initial report on the structural approximation (1.71 Å) [353] and the one on the prediction (1.23 Å) [354].

Local structure clustering. Sander and co-workers have recently described a novel approach to create local structure prototypes [355]. They defined 27 prototypes of 8 residues long comparable to Hunter and Subramaniam [353]. The clustering was based on Cα distance

matrix comparison. They also introduced a new approach to local protein structure prediction. In contrast to Baker's approach [267], they took into account structural information while partitioning sequence space. As sequence diversity is much higher than structural variation, it was expected that unsupervised learning in sequence space would be harder than unsupervised learning in structure space. In contrast to Hunter and Subramaniam's approach [353], they also incorporated protein family information by using profiles instead of sequences. They have tested numerous prediction approach using C.5 classifier [356], Support Vector Machines [357] and random forest [358]. All these approaches have led to a prediction not biased as it was the case for Hunter and Subramaniam.

Multiple alphabets. Karchin and co-workers have done an important work by defining numerous potential structural alphabets based on geometrical properties such as the α angle [275, 359]. They have also adapted other approaches such as a subdivision of STRIDE or DSSP secondary structure assignment methods. They have so compared the features of 9 structural alphabets using information content. Their results show clearly that some descriptions have a very limited interest and other seems highly informative. Their results clearly showed that PB alphabet was highly informative with the best predictive ability [359]. Recently, they also highlighted that the best predictive ability was not always associated to the best final prediction depending on the prediction method, *i.e*. some prediction methods are not well suited for some specific alphabets [360]. For instance, PBs used as a secondary hidden Markov model track reverse-sequence null model gives scoring results in significantly more false positives than geometric null scoring [360].

Other developments. Structural alphabets have been used for various purposes. For instance, structural alphabets defined by Camproux *et al*. [282, 361] and by de Brevern *et al*.

[285, 352] have proved their efficiency both in the description and the prediction of small loops [99, 283, 362, 363] or long fragments [349, 364-368]. In the following, we will detail some interesting developments based on the application of these local protein structures.

SA-Search. SA-Search is a web tool developed by Guyon and co-workers (see Figure (**14**)) that can be used to mine for proteins with similar fold and extract structural similarities [369]. It is based on the structural alphabet developed by Camproux (see *Short Structural Building Blocks* section) [361]. Using such a representation, classical methods developed for amino acid sequences can be employed. In SA-Search fast 3D similarity searches such as the extraction of exact words using a suffix tree approach can be achieved and the search for fuzzy words can be viewed as a simple 1D sequence alignment problem. At this time, SA-Search gives results often with shorter alignment length than DALI [370], but one strong point of SA-Search is that it allows the fast mining of protein structures, a typical run being on the order of a few seconds. SA-Search is available at the following url at the RPBS web server [217]: http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search. A more sophisticated approach, based also on the structural alphabet developed by Camproux, was used recently to mine SCOP, namely MinSet: http://mathbio.nimr.mrc.ac.uk/~jkleinj/MinSet [371].

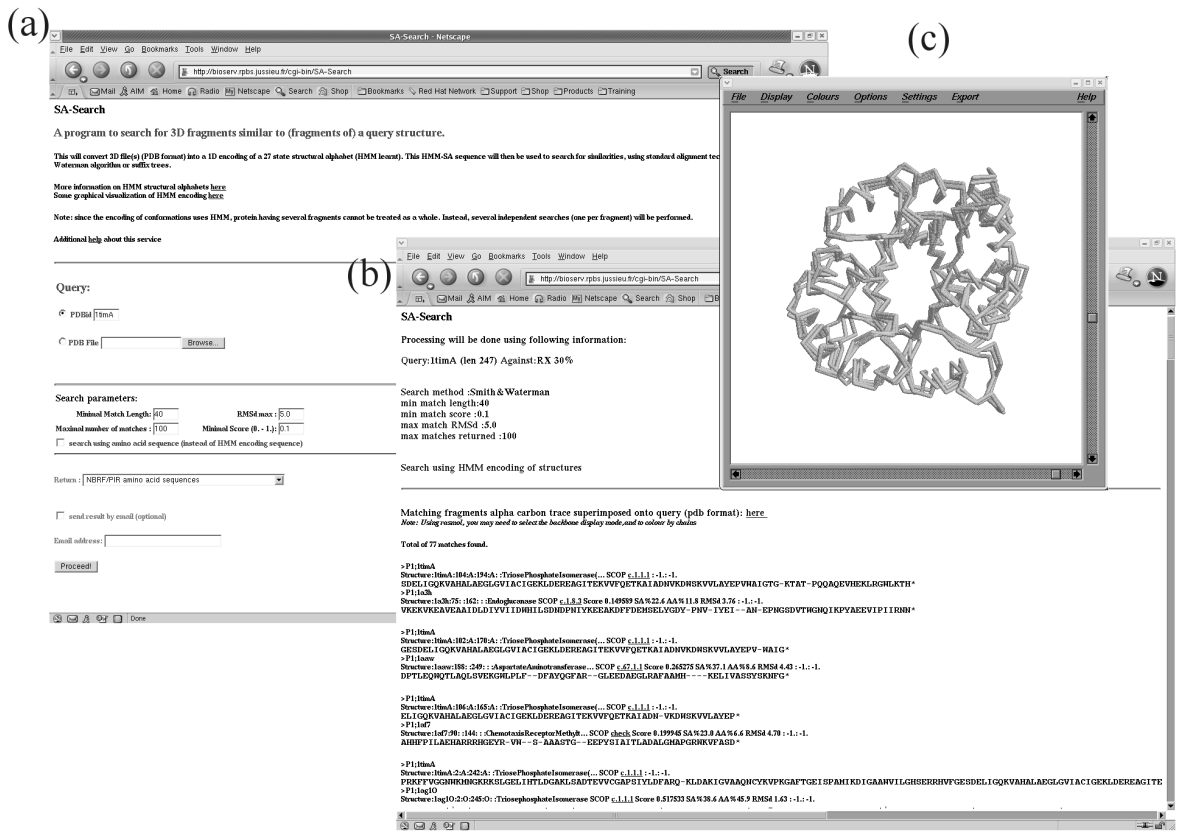

Figure 14. *SA-Search*. (a) Query page. (b) Results of the extraction of structural similarities and (c) rasmol representation of superimposed structures.

PBE server. Very recently, a substitution matrix in terms of Protein Blocks was developed based on the analysis of local structure variations in the alignments of homologous proteins from the large PALI database [372]. The number of substitutions between any two PBs was counted based on the alignment corresponding only to structurally conserved regions identified in PALI [373-377]. This caution was exercised, as the alignment of residues in the structurally variable regions is meaningless in the rigid body alignments. The raw frequencies were normalized and were then expressed as log-odds scores (see Table 5). It was shown that most of the off-diagonal elements are negative suggesting that most of the local conformations in the homologous protein structures are conserved while only few off-diagonal elements were with positive substitution scores *i.e.* with favourable substitutions. This example nicely illustrates how a structural alphabet was used to derive local structure

variations that are more or less frequently observed in aligned structures.

Protein Blocks	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
a	2.73															
b	-0.37	2.92														
c	0.27	-0.49	2.09													
d	-0.70	-0.86	-0.26	1.48												
e	-1.86	-0.76	-1.39	-0.78	3.46											
f	-0.83	-1.81	-0.71	-0.96	0.42	2.54										
g	0.15	-0.88	0.01	-1.60	1.19	-0.54	3.89									
h	-1.42	-0.10	-1.95	-1.65	0.35	-0.68	-0.83	3.43								
i	0.16	-0.02	-1.37	-1.52	-1.42	-1.38	-0.35	-1.13	3.81							
j	-1.32	0.13	-1.25	-1.28	-0.96	-0.59	-0.59	0.96	1.35	4.31						
k	-2.07	-0.36	-2.78	-3.12	-0.73	-0.40	-1.62	0.15	-0.47	-0.18	2.81					
l	-0.90	-0.26	-2.53	-2.02	-2.08	-0.69	-0.93	-0.70	-0.51	-0.33	-0.16	2.60				
m	-2.92	-3.51	-3.24	-5.88	-5.29	-2.72	-1.51	-3.48	-3.67	-2.46	-1.62	-1.23	0.97			
n	-1.73	-1.17	-2.03	-3.56	-0.93	-2.37	0.84	-1.43	-1.30	-0.71	-0.96	-0.64	-1.36	3.95		
o	-0.85	-0.88	-0.99	-3.14	-2.80	-1.78	-0.22	0.61	-1.20	-0.70	-2.08	-0.29	-1.85	-0.07	3.67	
p	-0.60	0.04	-0.30	-2.53	-2.52	-2.24	0.33	-2.10	1.05	0.40	-1.69	-1.54	-1.64	0.01	-0.07	3.23

Table 5. Normalized substitution frequencies [372] expressed as log-odds scores between any two protein blocks as determined by structure-based pairwise alignments of homologous proteins of known three-dimensional structure from PALI database [375].

The generated PB substitution matrix was validated by benchmarking how well it can be used to identify structurally equivalent regions in closely or distantly related proteins using a dynamic programming approach. The alignment results obtained are very comparable to well established structure comparison methods like DALI and STAMP [372].

An extension of the application of this PB substitution matrix in the direction of protein structure comparison and mining protein structure databanks for similar folds has been attempted [378]. Here, simplified 1D representation of protein structure in terms of PBs was analyzed just like amino acid sequence analysis to find structural local similarity, dissimilarity and relationship among protein structures. This has been implemented in the SCOP based Protein Block Expert (PBE) Server (see Figure (**15**)). Two protein structures can be compared

using simple dynamic programming algorithm by aligning their PB sequences using the PB substitution matrix (PB-Align). Similarly, PBE server uses the same PB alignment method to extract similar fold or related protein structures from a given databank with an average of 81.3% of first ranked hits belonging to the same SCOP FOLD as the benchmarked query. This server is available at (http://bioinformatics.univ-reunion.fr/PBE/).

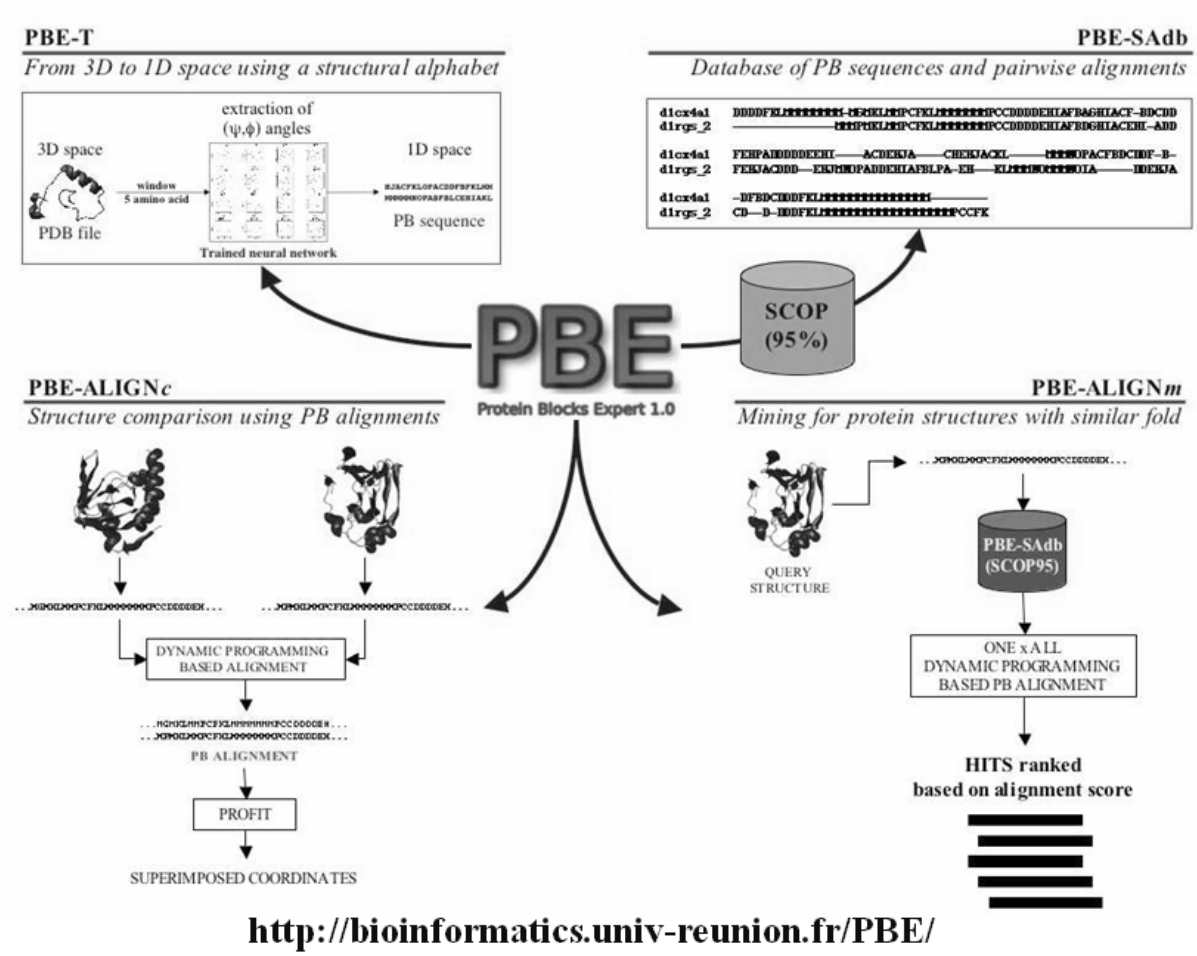

Figure 15. Architecture of PBE server. The main facilities available in PBE are (i) the encoding of protein structures into 1D sequences of PBs (PBE-T), (ii) the comparison between two structures by aligning their PB sequences (PBE-ALIGNc), the mining of structural databases for similar folds (PBE-ALIGNm) and a database of PB sequences and aligned PB sequences from homologous families (PBE-SAdb).

It is further demonstrated by a rigorous comparative analysis (Table 6) that performance of PB-Align for comparing structures [379] and for mining similar folds was at least

equivalent if not better to well-established structure comparison methods including CE, DALI, DEJAVU, FATCAT, VAST and YAKUSA (Tyagi *et al*., *submitted*).

Program	Mainly α (19)	Mainly β (19)	Mixed αβ (15)	Few SSEs (8)	Total (%)
PB-ALIGN	18[+]	17[*]	14	8	96.6
YAKUSA	17	19	14	8	95
CE	17	19	13	8	93
DALI	14	19	14	8	90
MATRAS	11	19	14	8	85
VAST	12	17	15	7	84
TOP	14	18	12	7	84
DEJAVU	14	19	9	4	75
TOPSCAN	15	12	9	7	70
TOPS	2	15	14	7	62
PRIDE	14	14	7	3	62
LOCK	0	14	11	8	54
SSM	5	13	10	5	54

Table 6. Comparison of PB-ALIGN with existing structure comparison methods. Comparison of PB-ALIGN with 12 structure comparison methods was based on results from Carpentier *et al.* [379]. Are indicated the total number of successful queries for each method in each main SCOP class. The numbers along with the header gives total number of queries belonging to each class. All the hits are counted based on first 10 ranking alignments compared to 100 hits taken by Carpentier *et al*. [379].
[+] One query has no target in our database. [*] For mainly b class, query protein 1vmo has no target in our database and query 1ciy misses target in top ten ranks.

3D-BLAST. A similar approach to the one developed in SA-Search and PB-Align and which uses a 23 states structural alphabet to describe the backbone has been developed very

recently [380]. This method named 3D-Blast, uses BLAST as a search method using a structural alphabet substitution matrix to find the longest common substructures with high-scoring segment pairs. Interestingly, this method uses an E-value as measure of statistical significance of an alignment and generates results with performance comparable to known methods.

Its general principle has recently been detailed [381]. The proteins were cut into structural fragments of 5 residues length encoded as κ and α angles. The authors discretized the (κ, α) plot into 648 representative segments and clustered them into 23 clusters grouping similar fragments. One the cluster was dedicated to the fragments not associated to a given cluster. The obtained clusters were consistent with secondary structure content defined by DSSP program [231]. The dedicated substitution matrix was obtained using 674 superimposed pairs of proteins taken from SCOP. Classical BLAST approach [382, 383] was used to perform the mining of structural database. Using a greedy algorithm [308], the authors proved the validity of their structural alphabet by reconstructing 39 protein structures with efficiency.

Short loops. The concept of local protein structures has been applied to predict short loop conformations. A first attempt has used the structural alphabet composed of 12 local protein structures defined by Camproux and co-workers [282]. It followed an analysis of the dependence between the SBBs [283] and focussed on the prediction of "exact" succession of local protein structures, *i.e.* exact words [362]. The quality of the results was highly dependant of the number of occurrence of these words.

Recently, local protein structures in terms of Protein Blocks has been predicted from amino acid sequence [285, 351, 352]. Similarly the backbone of short loops was predicted in terms of PBs [99]. The prediction was performed using a classical Bayesian approach, but

contrarily to the previously reported studies, the prediction was performed specifically for these loop regions. As expected, the knowledge of the protein zones induced a significant improvement of the prediction rate of the PBs from amino acid sequence. New sequence – structure relationships were highlighted. Nonetheless, the size of the databank (a non – redundant databank with a too low identity sequence rate) remained an important limitation.

HMMSTR developments. HMMSTR has also proven to be useful in many other instances. It has successfully been used to predict protein three-dimensional local structures, secondary structures, to identify protein-coding ORFs, or to design a sequence to fit a structure. Recently, a two-dimensional approach has been developed with HMMSTR-CM [384]. The latter predicts the likelihood of pairwise inter-residue contacts. The resulting contact maps can be projected into three-dimension using methods based on distance geometry such as those used to solve NMR structures. Interestingly HMMSTR-CM contains a set of rules that describe how secondary structure elements can arrange themselves in 3D.

Similarly, the remote homologue detection method called SVM-HMMSTR has been developed that overcomes the reliance on detectable sequence similarity by transforming the sequences into strings of hidden Markov states that represent local folding patterns. It uses a Support Vector Machine (SVM) that combines an order-independent feature which captures the amino acid and local structure composition and an order-dependent feature which captures the sequential ordering of the local structures, and its performance overcomes numerous equivalent approaches [385].

Another program, SCALI, which is based on HMMSTR, has recently been developed to find all possible ways for arranging secondary structure units in space, and to model them as HMMs. To do this task, a computationally feasible method was developed to perform alignments of protein structures to find conserved packing arrangements, even if they are non-

sequentially ordered in space. The results of SCALI are better than current methods that cannot find non-sequential similarities in proteins [386].

Very recently, Huang and Bystroff extended local structure predictions from HMMSTR to improve the quality of the most difficult pairwise alignments, those with less than 25% sequence identity using a profile-profile alignment method [387]. A new model, called HMMSUM (HHMSTR-based Substitution Matrices) was developed where a set of matrices, one for each 281 local structure contexts defined by HMMSTR, were summed from a training set of multiple sequence alignments in the same way PAM and BLOSUM matrices were summed. When HMMSUM model is used in an alignment, target and template HMMSTR descriptors are predicted before an alignment matrix is calculated using the model. It is argued that the improved accuracy of HMMSUM matrices over other equivalent single matrices methods like SDM is due to the fact that, since different local structures may have significant different amino acid preferences, these can only be captured using multiple substitution matrices like in HMMSUM.

Building blocks folding model. Somewhat different but also interesting is the "building blocks folding model" proposed by Nussinov's group. Unlike alphabet approaches based mainly on 3D similarities between fragments, Nussinov and her collaborators focused on the elementary folding units that lead through a hierarchical process to the folded state [388, 389]. These folding units are obtained from a progressive and hierarchical dissection based mainly on native 3D interactions in the folded state. The process concludes with fragments of variable length (at least 13 residues) and may be used to engineer new naturally occurring folds with low homology to existing proteins [390]. Such elementary folding units have been used in a homology prediction strategy. A target sequence is compared with and aligned to the building block sequences in the database. A graph approach then assigns the building block

automatically to the query sequence. Before this method can be used for *ab initio* structure prediction purposes, further exploration needs to establish position-specific sequence-structure relations from these elementary folding units. This methodology, while it may constitute an alternative to structural alphabets, is clearly based on a distinct approach. Any direct comparison between these folding units and the fragments defined by a structural alphabet is less than straightforward.

Other related works. In the same way, Lee and co-workers have presented preliminary results highly related to the work of Camproux [283, 361] using a two-stage strategy to cluster local protein structures using BIC criteria and Expected Maximization [391]. Their results seemed promising but lacked biological analyses [392]. It has been used to classify protein 3D folds in regards to a simple categorization within SCOP superfamily [393]. Yi and co-workers also tried to create a library of local protein structures with a methodology highly similar to Schuchhardt *et al.* and de Brevern *et al.* [279, 285] using dihedral angles (ϕ, ψ) [394]. Nonetheless, these approaches had led to the proposition to a too high number of clusters (1858 for only 3636 fragments), showing that some parameters in the learning approach needed to be improved. Zheng and Liu have proposed a set of local protein structures and an associated substitution matrix [395, 396]. To compute this matrix, they have used protein families from FSSP database and derived a substitution matrix in the same way as BLOSUM [397]. It is noteworthy that their set of protein local structures have been built in the absence of any comprehensive analysis on the structural alphabet and their learning method is questionable, *i.e.* a mixture model of pseudo-bond angles. Same remarks can be made about a recent prediction method based on Artificial Neural Net works – derived from HYPLOSP approaches [398, 399] and dedicated to the prediction of PBs [400]. The results are higher than the pervious ones based only on one sequence, but no analysis are done to assess the

absence of bias in favor to the most frequent PBs.

On the other hand, we must note the excellent PhD thesis of Tang who proposes novel approaches to obtain local protein structure useful in prediction process [401, 402]. Hence, he proposes a method to obtain structural alphabets based on a variation of *k-means* algorithm. In the same way, Wang and co-workers have developed local structure-based sequence profile database called LPBSP1 and LPBSP2 [403-405]. Local structure-based sequence profiles are equivalent to sequence-structure motifs as they both represent the consensus of a group of segments sharing similar compositions and structures.

Other applications

Assessment of structural diversity of pentapeptides. Taking advantage of protein blocks which offer a standardized and refined description of protein local conformations, the question of the structural diversity of pentapeptides, namely of identical pentapeptides, was asked to provide new insights into local sequence-structure relationship (Offmann *et al.*, *in preparation*). Results obtained generally confirmed the known paradigm that identical stretches of amino acids can adopt different conformations as it was established previously [406-408]. More than 73% of identical pentapeptides are surveyed to adopt different local conformations. Amino acid content seems to critically define the inherent capability of identical pentapeptides to adopt different local conformations. For example, higher propensities in glycine content in these pentapeptides confirmed its well ubiquitous distribution in the phi-psi Ramachandran's map. It is suggested that this largely distributed property of identical pentapeptides to adopt different local structures can be viewed as to be evolutionarily advantageous. Points of structural flexibility of constituent pentapeptides inside a polypeptide chain may be viewed as naturally necessary during its lifetime with regard to its folding and function.

Evidence for strong local sequence to structure relationship at the level of pentapeptides in protein structures has nevertheless been brought in this work (Offmann *et al*., *in preparation*). Examples of identical pentapeptides that have conserved local structures in otherwise different environments have been identified. It is shown that they are not widely distributed; only about 25% of multicopy pentapeptides indeed map to a single PB. Most of them, surprisingly are found in helical-like structures (map to PB *m*) and to some extent to strand-like conformations. Reasons for recruitment of structurally "stable" pentapeptides has yet to be established, but their presence is suggested to be a landmark in protein structures and may represent structural "hotspots" in proteins.

Hybrid Protein Model. The reliability of the Protein Blocks for characterizing long fragments enabled the development a novel unsupervised clustering method called Hybrid Protein Model (HPM). This method, which can capture long-range features of a succession of PBs, compresses a structural protein databank into a limited set of clusters. The HPM training principle is similar to that of Kohonen's SOM [287, 288]. Its originality is that it can learn long protein fragments previously encoded into series of PBs. HPM compacts a databank of such fragments into one "Hybrid Protein" (HP), by stacking them on the basis of the similarity of their PB series. Through this process, it builds a library of clusters that group structurally similar fragments; each cluster is represented by a mean local structure prototype. Unlike standard clustering methods, HPM generates a library of overlapping prototypes. Its principal advantage is that it takes into account the dependence between successive local structures along the proteins by maintaining their continuity.

Two main characteristics affect the features of the final library built by HPM: the length of the protein fragments and the number of clusters in the library. A first HPM of 100 clusters, grouping a series of 10-PB fragments, was used for fine description of protein 3D structures

and efficiently identified local structure similarities between two cytochromes P450 [364]. Subsequent examination of a new learning approach led to an HPM of 233 clusters that grouped a series of 13-PB fragments [368]. Recent work has focused on improving detection of similarities between long fragments [366, 367].

Recently, the library of local structure prototypes constructed by the HPM has been used to develop a prediction strategy, aiming at optimizing exploitation of the sequence-structure relations in this library. This was achieved by setting up a system of experts, each defined by logistic regression and best able to discriminate from sequence a given local structure prototype relative to the others. The experts then computed probabilities for each prototype for a target sequence window, and the top scorers become structural candidates [409]. Recent use of Support Vector Machines coupled with evolutionary data has greatly improved the prediction rates (Bornot *et al*., *in preparation*).

HPM was also used to analyze the sequence – structure relationship of globular proteins [410] and an adaptation of the algorithm was done for genomic data [411, 412].

Spectacular example of application of local structures for modelling DARC. The Duffy Antigen/Receptor for Chemokine (DARC) is an erythrocyte receptor for malaria parasites (*Plasmodium vivax* and *Plasmodium knowlesi*) and for chemokines. In contrast to other chemokine receptors, DARC is a promiscuous receptor that binds chemokines of both CC and CXC classes. The extracellular domains of DARC are characterized by a long N-terminal chain essential for the interaction both with the malaria erythrocyte-binding proteins and the chemokines. Structural models of the DARC have been elaborated and analyzed [413]. The building of the 3D models was based on a comparative modelling process. Protein Blocks were used in complement of classical transmembrane prediction, threading, *ab initio* and secondary structure approaches. The chosen structural models correlated at best available

experimental data (see Figure (**16**)).

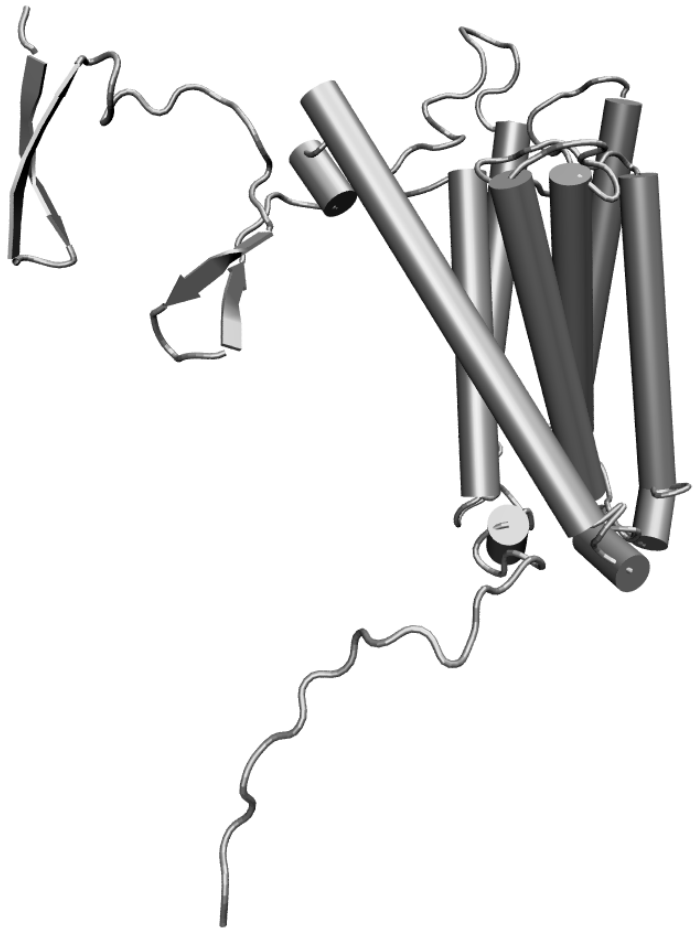

Figure 16. Visualisation of DARC protein using VMD software [8]. The N terminus region is presented at the top left of the figure.

The analysis of the flexibility of the extracellular domains is performed with simulated annealing. The second and fourth extracellular loops are strongly constrained. Protein Blocks were used (i) for the analysis of the sequence – structure relationship using *LocPred* software [352] and (ii) also to analyze the simulated annealing results from Gromacs [414, 415]. In both case, the number of local protein structures have given a better analysis of the results than the classical secondary structures.

Discovering Structural Motifs using a Structural Alphabet. To test a PBs-based strategy for discovering metal binding site structural motifs, Dudev and Lim scan a database of zinc binding sites [416]. They searched for proteins containing a characteristic sequence motif [417]. All of these proteins were found to possess a PB structural motif *f*(2)*o*(13-15)*f*(2)*m* showing that structural-alphabet based approach for discovering new structural motifs seems promising. The study was extended to Mg^{2+}-binding sites which are more difficult protein families. The resulting structural patterns were more complex and sometimes fuzzier but showed the interest of such an approach, e.g. similar Mg^{2+} binding site structures were found in otherwise unrelated protein sequences [418].

Comparison between approaches towards local structures. One of the major difficulties of the local protein structures – and contrarily to secondary structure assignment – is the lack of comparisons between different approaches. It is mainly due to three factors: (i) the use of very different criteria and / or methodology for each approach, (ii) the absence of software able to perform an easy encoding of the protein structures and (iii) the use of local protein structures of different length. Most of the time, the analysis focus only on the distribution of secondary structures in each local protein structure. We can note an old comparison that was done a few years ago between the Protein Blocks and (a) Rooman and Wodak's local protein structure [280], (b) Fetrow's local protein structure [281] and (c) Camproux's first structural alphabet [337]. The comparison was mainly based on protein encoded by two structural libraries. It has highlighted the difficulties of comparison (see Table 7), but also the interest of different approaches for the analysis of repetitive and non-repetitive local structures.

For instance, between PBs and SBBs, PB *d* that represents the central core of a β-strand corresponds to SBBs β2 (55.8%), γ2 (16.4%) and β1 (14.7%), and PB *h*, a local protein structure associated to the coil corresponds to SBBs γβ (23.3%), γ2 (22.9%) and β2 (17.6%).

SBBs β are related to β-strand, while γ are coil states, thus γβ is a N-cap of a β-strand. Recently, another comparison has been performed between the Protein Blocks and libraries elaborated by Levitt and *oligons* elaborated by Micheletti [351]. Due to the absence of protein structure assignment, the comparison has been directly done between the local protein structures; this comparison nevertheless gave some insight towards the understanding of the specificity of each approach.

		Protein Blocks															
		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
SBBs	η		5.7						**11.4**	1.6	2.2	**11.9**	9.2	**23.1**	**15.4**	**11.7**	2.7
	α												**10.6**	**86.1**			
	τ	7.1	6.4	**10.2**	**16.0**		2.2	5.0		**12.7**			5.2	7.3	1.2	8.2	**16.4**
	ζ	2.0		1.5	1.3	9.4	**34.9**	1.9	5.1		2.1	**27.2**	2.7	9.0			1.1
	ι	7.1	**10.6**	**15.8**	**40.4**	6.5	3.7	2.0	1.9	1.5	1.1	1.0	1.0	1.2			4.9
	β	7.2	7.1	**18.3**	**53.2**	2.4	6.1		1.6								1.9

Table 7. Comparison of LSPs assignment with PBs assignment using the protein databank used by Fetrow and co-workers [281]. The frequencies f_{ij} correspond to the percentage of LSP *i* found associated to PB *j*.

Conclusion. Local protein structure libraries are an important and relevant area of research. Nonetheless, it is often minimized or underestimated for a major reason. It is due to the importance of regular secondary structures, i.e. α-helix and β-sheet. They represent half of the residues and possess critical physico-chemical properties essential for the protein fold. It allows a very interesting simplification of protein structures and so has been widely used.

Local protein structure libraries or structural alphabets complexify greatly the information from a simple 3-states (two regular and one undefined) to more than a dozen local protein structures at least, *i.e*. more difficult to apprehend.

This review attempted to show that research area in this field is very active and that, up to now, yielded very interesting results. Several points should be stressed on. One of the most important one has been highlighted by Karchin and co-workers' excellent investigations [275]: in final, it is difficult to compare between the different local protein structure libraries.

In the same way, classical predictions were assessed by indexes that may not be suited for a structural alphabet such as the Matthews' correlation coefficient [419], an index elaborated for only two states. So, other indexes must be used to ensure a correct and unbiased prediction. For instance, the prediction done by Hunter and Subramaniam is highly biased as 11 out of 28 local protein structures are not populated nor used. Baldi and co-workers present a very good review on different indexes that can help towards this aim [420]. Serge Hazout has also proposed an index based on Shannon entropy that can help such comparison [285].

This difficulty is widely observed even with a sophisticated approach such as SSPRO [49], a complex HMM that used evolutionary profiles. It is one of the best secondary structure prediction method (Q_3 ~80%), but its prediction rate drops to 62% when the number of predicted state climb to 8, i.e. DSSP assignment. Moreover, two states cannot be predicted. For the ASs, few have been used in prediction approaches and they emphasises the same problem: "higher the number of states rise, harder the prediction". For instance, the prediction done by Hunter and Subramaniam equals 40% but is highly biased as 11 out of 28 local protein structures are not populated nor used [354].

We observed a similar evolution to the secondary structure prediction going from simple statistical approach based on one sequence to sophisticated methods. Efficient algorithms for

predictions using structural libraries have been developed. I-sites library is such an example coupling complex prediction method with evolutionary information [267]. Similar approach has been developped by Sander and co-workers that used SVMs [355] and the coupling of SVMs with evolutionary information has increased the prediction of local structure prototypes developed by Benros and co-workers (Bornot *et al*., *submitted*).

As we have seen in this review, the vision of the protein structures is not as basic and trivial as often admitted. First, the secondary structures are more complex than "simple" hydrogen bonds, they are not rigid bodies and so it is sometimes difficult to precisely delimit them. Second, they are not the only possible solution to analyze and predict the protein structures and with respect to this, the use of structural alphabets seems a very promising research avenue.

Acknowledgments

This paper is dedicated to the memory of Pr. Serge Hazout. This work was supported in parts by grants from the Ministère de la Recherche, Université Paris Diderot, French Institute for Health and Medical Research (INSERM) and Conseil Régional de la Réunion.

References

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

[2] Rath VL, Ammirati M, Danley DE, Ekstrom JL, Gibbs EM, Hynes TR, Mathiowetz AM, McPherson RK, Olson TV, Treadway JL,Hoover DJ. Human liver glycogen phosphorylase inhibitors bind at a new allosteric site. *Chem Biol* **2000**; 7: 677-82.

[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN,Bourne PE. The Protein Data Bank. *Nucleic Acids Res* **2000**; 28: 235-42.

[4] Pauling L, Corey RB,Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **1951**; 37: 205-11.

[5] Pauling L,Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* **1951**; 37: 251-6.

[6] Sayle RA,Milner-White EJ. RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **1995**; 20: 374.

[7] Koradi R, Billeter M,Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **1996**; 14: 51-5, 29-32.

[8] Humphrey W, Dalke A,Schulten K. VMD: visual molecular dynamics. *J Mol Graph* **1996**; 14: 33-8, 27-8.

[9] Murzin AG, Brenner SE, Hubbard T,Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **1995**; 247: 536-40.

[10] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB,Thornton JM. CATH--a hierarchic classification of protein domain structures. *Structure* **1997**; 5: 1093-108.

[11] Rost B. Review: protein secondary structure prediction continues to rise. *J Struct Biol* **2001**; 134: 204-18.

[12] Marin A, Pothier J, Zimmermann K,Gibrat JF. FROST: a filter-based fold recognition method. *Proteins* **2002**; 49: 493-509.

[13] Simons KT, Bonneau R, Ruczinski I,Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **1999**; Suppl 3: 171-6.

[14] Rohl CA, Strauss CE, Misura KM,Baker D. Protein structure prediction using Rosetta. *Methods Enzymol* **2004**; 383: 66-93.

[15] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H,Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958**; 181: 662-6.

[16] Richardson JS,Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. *Science* **1988**; 240: 1648-52.

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

[18] Presta LG,Rose GD. Helix signals in proteins. *Science* **1988**; 240: 1632-41.

[19] Aurora R, Srinivasan R,Rose GD. Rules for alpha-helix termination by glycine. *Science* **1994**; 264: 1126-30.

[20] Aurora R,Rose GD. Helix capping. *Protein Sci* **1998**; 7: 21-38.

[21] Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry* **1978**; 17: 4277-85.

[22] Imai K,Mitaku S. Mechanisms of secondary structure breakers in soluble proteins. *Biophysics* **2005**; 1: 55-65.

[23] Ho BK, Thomas A,Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* **2003**; 12: 2508-22.

[24] Ermolenko DN, Thomas ST, Aurora R, Gronenborn AM,Makhatadze GI. Hydrophobic interactions at the Ccap position of the C-capping motif of alpha-helices. *J Mol Biol* **2002**; 322: 123-35.

[25] Prieto J,Serrano L. C-capping and helix stability: the Pro C-capping motif. *J Mol Biol* **1997**; 274: 276-88.

[26] Dirr HW, Little T, Kuhnert DC,Sayed Y. A conserved N-capping motif contributes significantly to the stabilization and dynamics of the C-terminal region of class Alpha glutathione S-transferases. *J Biol Chem* **2005**; 280: 19480-7.

[27] Kuhnert DC, Sayed Y, Mosebi S, Sayed M, Sewell T,Dirr HW. Tertiary interactions stabilise the C-terminal region of human glutathione transferase A1-1: a crystallographic and calorimetric study. *J Mol Biol* **2005**; 349: 825-38.

[28] Penel S, Morrison RG, Dobson PD, Mortishire-Smith RJ,Doig AJ. Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. *Protein Eng* **2003**; 16: 957-61.

[29] Regan L. Protein structure. Born to be beta. *Curr Biol* **1994**; 4: 656-8.

[30] Colloc'h N,Cohen FE. Beta-breakers: an aperiodic secondary structure. *J Mol Biol* **1991**; 221: 603-13.

[31] Zaremba SM,Gregoret LM. Context-dependence of amino acid residue pairing in antiparallel beta-sheets. *J*

Mol Biol **1999**; 291: 463-79.
[32] Merkel JS,Regan L. Aromatic rescue of glycine in beta sheets. *Fold Des* **1998**; 3: 449-55.
[33] Merkel JS, Sturtevant JM,Regan L. Sidechain interactions in parallel beta sheets: the energetics of cross-strand pairings. *Structure Fold Des* **1999**; 7: 1333-43.
[34] Wouters MA,Curmi PM. An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins* **1995**; 22: 119-31.
[35] Hutchinson EG, Sessions RB, Thornton JM,Woolfson DN. Determinants of strand register in antiparallel beta-sheets of proteins. *Protein Sci* **1998**; 7: 2287-300.
[36] Mandel-Gutfreund Y, Zaremba SM,Gregoret LM. Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J Mol Biol* **2001**; 305: 1145-59.
[37] Lacroix E, Kortemme T, Lopez de la Paz M,Serrano L. The design of linear peptides that fold as monomeric beta-sheet structures. *Curr Opin Struct Biol* **1999**; 9: 487-93.
[38] Dooley CT, Chung NN, Wilkes BC, Schiller PW, Bidlack JM, Pasternak GW,Houghten RA. An all D-amino acid opioid peptide with central analgesic activity from a combinatorial library. *Science* **1994**; 266: 2019-22.
[39] West MW, Wang W, Patterson J, Mancias JD, Beasley JR,Hecht MH. De novo amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci U S A* **1999**; 96: 11211-6.
[40] Perez-Paya E, Houghten RA,Blondelle SE. Functionalized protein-like structures from conformationally defined synthetic combinatorial libraries. *J Biol Chem* **1996**; 271: 4120-6.
[41] Pastor MT,Perez-Paya E. Combinatorial chemistry of beta-hairpins. *Mol Divers* **2003**; 6: 149-55.
[42] Ho BK,Curmi PM. Twist and shear in beta-sheets and beta-ribbons. *J Mol Biol* **2002**; 317: 291-308.
[43] Garnier J, Osguthorpe DJ,Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **1978**; 120: 97-120.
[44] Robson B,Suzuki E. Conformational properties of amino acid residues in globular proteins. *J Mol Biol* **1976**; 107: 327-56.
[45] Chou PY,Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**; 13: 211-22.
[46] Chou PY,Fasman GD. Prediction of protein conformation. *Biochemistry* **1974**; 13: 222-45.
[47] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999**; 292: 195-202.
[48] Pollastri G,McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* **2005**; 21: 1719-20.
[49] Pollastri G, Przybylski D, Rost B,Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **2002**; 47: 228-35.
[50] Dor O,Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins* **2006**; 66: 838-45.
[51] Martin J, Gibrat JF,Rodolphe F. Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct Biol* **2006**; 6: 25.
[52] Donohue J. Hydrogen bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* **1953**; 39: 470– 78.
[53] Pal L,Basu G. Novel protein structural motifs containing two-turn and longer 3(10)-helices. *Protein Eng* **1999**; 12: 811-4.
[54] Pal L, Basu G,Chakrabarti P. Variants of 3(10)-helices in proteins. *Proteins* **2002**; 48: 571-9.
[55] Baker EN,Hubbard RE. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol* **1984**; 44: 97-179.
[56] Barlow DJ,Thornton JM. Helix geometry in proteins. *J Mol Biol* **1988**; 201: 601-19.
[57] Karpen ME, de Haseth PL,Neet KE. Differences in the amino acid distributions of 3(10)-helices and alpha-helices. *Protein Sci* **1992**; 1: 1333-42.
[58] Pal L, Dasgupta B,Chakrabarti P. 3(10)-Helix adjoining alpha-helix and beta-strand: sequence and structural features and their conservation. *Biopolymers* **2005**; 78: 147-62.
[59] Low BW,Baybutt RB. The π-helix -A hydrogen bonded configuration of the polypeptide chain. *J Am Chem Soc* **1952**; 74: 5806.
[60] Low BW,Greenville-Wells HJ. Generalized mathematical relationships for polypeptide chain helices. The coordinates of the π-helix. *Proc Natl Acad Sci USA* **1953**; 39: 785–801.
[61] Ramachandran GN,Sasisekharan V. Conformation of polypeptides and proteins. *Advan. Protein Chem* **1968**; 23: 283-438.
[62] Rohl CA,Doig AJ. Models for the 3(10)-helix/coil, pi-helix/coil, and alpha-helix/3(10)-helix/coil transitions in isolated peptides. *Protein Sci* **1996**; 5: 1687-96.
[63] Fodje MN,Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the pi-

helix. *Protein Eng* **2002**; 15: 353-8.
[64] Weaver TM. The pi-helix translates structure into function. *Protein Sci* **2000**; 9: 201-6.
[65] Lee KH, Benson DR,Kuczera K. Transitions from alpha to pi helix observed in molecular dynamics simulations of synthetic peptides. *Biochemistry* **2000**; 39: 13737-47.
[66] Millhauser GL. Views of helical peptides: a proposal for the position of 3(10)-helix along the thermodynamic folding pathway. *Biochemistry* **1995**; 34: 3873-7.
[67] Millhauser GL, Stenland CJ, Bolin KA,van de Ven FJ. Local helix content in an alanine-rich peptide as determined by the complete set of 3JHN alpha coupling constants. *J Biomol NMR* **1996**; 7: 331-4.
[68] Armen R, Alonso DO,Daggett V. The role of alpha-, 3(10)-, and pi-helix in helix-->coil transitions. *Protein Sci* **2003**; 12: 1145-57.
[69] Eswar N, Ramakrishnan C,Srinivasan N. Stranded in isolation: structural role of isolated extended strands in proteins. *Protein Eng* **2003**; 16: 331-9.
[70] Cartailler JP,Luecke H. Structural and functional characterization of pi bulges and other short intrahelical deformations. *Structure (Camb)* **2004**; 12: 133-44.
[71] Duneau JP, Genest D,Genest M. Detailed description of an alpha helix-->pi bulge transition detected by molecular dynamics simulations of the p185c-erbB2 V659G transmembrane domain. *J Biomol Struct Dyn* **1996**; 13: 753-69.
[72] Aller P, Voiry L, Garnier N,Genest M. Molecular dynamics (MD) investigations of preformed structures of the transmembrane domain of the oncogenic Neu receptor dimer in a DMPC bilayer. *Biopolymers* **2005**; 77: 184-97.
[73] Sajot N,Genest M. Structure prediction of the dimeric neu/ErbB-2 transmembrane domain from multi-nanosecond molecular dynamics simulations. *Eur Biophys J* **2000**; 28: 648-62.
[74] Goetz M, Carlotti C, Bontems F,Dufourc EJ. Evidence for an alpha-helix --> pi-bulge helicity modulation for the neu/erbB-2 membrane-spanning segment. A 1H NMR and circular dichroism study. *Biochemistry* **2001**; 40: 6534-40.
[75] Keefe LJ, Sondek J, Shortle D,Lattman EE. The alpha aneurism: a structural motif revealed in an insertion mutant of staphylococcal nuclease. *Proc Natl Acad Sci U S A* **1993**; 90: 3275-9.
[76] Sarma GN, Nickel C, Rahlfs S, Fischer M, Becker K,Karplus PA. Crystal structure of a novel Plasmodium falciparum 1-Cys peroxiredoxin. *J Mol Biol* **2005**; 346: 1021-34.
[77] Fowler CB, Pogozheva ID, LeVine H, 3rd,Mosberg HI. Refinement of a homology model of the mu-opioid receptor using distance constraints from intrinsic and engineered zinc-binding sites. *Biochemistry* **2004**; 43: 8700-10.
[78] Love WE, Klock PA, Lattman EE, Padlan EA, Ward KB, Jr.,Hendrickson WA. The structures of lamprey and bloodworm hemoglobins in relation to their evolution and function. *Cold Spring Harb Symp Quant Biol* **1972**; 36: 349-57.
[79] Piela L, Nemethy G,Scheraga HA. Proline-induced constraints in alpha-helices. *Biopolymers* **1987**; 26: 1587-600.
[80] Blundell T, Barlow D, Borkakoti N,Thornton J. Solvent-induced distortions and the curvature of alpha-helices. *Nature* **1983**; 306: 281-3.
[81] Chakrabarti P, Bernard M,Rees DC. Peptide-bond distortions and the curvature of alpha-helices. *Biopolymers* **1986**; 25: 1087-93.
[82] Richards FM,Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* **1988**; 3: 71-84.
[83] Kumar S,Bansal M. Structural and sequence characteristics of long alpha helices in globular proteins. *Biophys J* **1996**; 71: 1574-86.
[84] Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG,Gibrat J-F. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. **submitted**.
[85] Kumar S,Bansal M. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J* **1998**; 75: 1935-44.
[86] Bansal M, Kumar S,Velavan R. HELANAL: a program to characterize helix geometry in proteins. *J Biomol Struct Dyn* **2000**; 17: 811-9.
[87] Richardson JS, Getzoff ED,Richardson DC. The beta bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci U S A* **1978**; 75: 2574-8.
[88] Milner-White EJ. Beta-bulges within loops as recurring features of protein structure. *Biochim Biophys Acta* **1987**; 911: 261-5.
[89] Chan AW, Hutchinson EG, Harris D,Thornton JM. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci* **1993**; 2: 1574-90.
[90] Sondek J,Shortle D. Accommodation of single amino acid insertions by the native state of staphylococcal nuclease. *Proteins* **1990**; 7: 299-305.
[91] Sondek J,Shortle D. Structural and energetic differences between insertions and substitutions in

staphylococcal nuclease. *Proteins* **1992**; 13: 132-40.
[92] Chen PY, Gopalacushina BG, Yang CC, Chan SI,Evans PA. The role of a beta-bulge in the folding of the beta-hairpin structure in ubiquitin. *Protein Sci* **2001**; 10: 2063-74.
[93] Axe DD, Foster NW,Fersht AR. An irregular beta-bulge common to a group of bacterial RNases is an important determinant of stability and function in barnase. *J Mol Biol* **1999**; 286: 1471-85.
[94] Richardson JS,Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A* **2002**; 99: 2754-9.
[95] Siepen JA, Radford SE,Westhead DR. Beta edge strands in protein structure prediction and aggregation. *Protein Sci* **2003**; 12: 2348-59.
[96] Wintjens R, Wodak SJ,Rooman M. Typical interaction patterns in alphabeta and betaalpha turn motifs. *Protein Eng* **1998**; 11: 505-22.
[97] Wojcik J, Mornon JP,Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol* **1999**; 289: 1469-90.
[98] Boutonnet NS, Kajava AV,Rooman MJ. Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins* **1998**; 30: 193-212.
[99] Fourrier L, Benros C,de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* **2004**; 5: 58.
[100] Richardson JS. The anatomy and taxonomy of protein structure. *Adv Protein Chem* **1981**; 34: 167-339.
[101] Némethy G,Printz MP. The gamma turn, a possible folded conformation of the polypeptide chain. Comparison with the beta turn. *Macromolecules* **1972**; 5: 755-58.
[102] Matthews BW. the gamma-turn. Evidence for a new folded conformation in Proteins. *Macromolecules* **1972**; 5: 818-19.
[103] Rose GD, Gierasch LM,Smith JA. Turns in peptides and proteins. *Adv Protein Chem* **1985**; 37: 1-109.
[104] Milner-White EJ. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops. *J Mol Biol* **1988**; 199: 503-11.
[105] Milner-White EJ. Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *J Mol Biol* **1990**; 216: 386-97.
[106] Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **1968**; 6: 1425-36.
[107] Lewis PN, Momany FA,Scheraga HA. Folding of polypeptide chains in proteins: a proposed mechanism for folding. *Proc Natl Acad Sci U S A* **1971**; 68: 2293-7.
[108] Kuntz ID. Protein folding. *J Am Chem Soc* **1972**; 94: 4009-12.
[109] Lewis PN, Momany FA,Scheraga HA. Chain reversals in proteins. *Biochim Biophys Acta* **1973**; 303: 211-29.
[110] Chou PY,Fasman GD. Beta-turns in proteins. *J Mol Biol* **1977**; 115: 135-75.
[111] Rose GD,Seltzer JP. A new algorithm for finding the peptide chain turns in a globular protein. *J Mol Biol* **1977**; 113: 153-64.
[112] Rose GD,Wetlaufer DB. The number of turns in globular proteins. *Nature* **1977**; 268: 769-70.
[113] Zimmerman SS,Scheraga HA. Influence of local interactions on protein structure. I. Conformational energy studies of N-acetyl-N'-methylamides of Pro-X and X-Pro dipeptides. *Biopolymers* **1977**; 16: 811-43.
[114] Wilmot CM,Thornton JM. Analysis and prediction of the different types of beta-turn in proteins. *J Mol Biol* **1988**; 203: 221-32.
[115] Wilmot CM,Thornton JM. Beta-turns and their distortions: a proposed new nomenclature. *Protein Eng* **1990**; 3: 479-93.
[116] Hutchinson EG,Thornton JM. A revised set of potentials for beta-turn formation in proteins. *Protein Sci* **1994**; 3: 2207-16.
[117] Hutchinson EG,Thornton JM. PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* **1996**; 5: 212-20.
[118] Mattos C, Petsko GA,Karplus M. Analysis of two-residue turns in proteins. *J Mol Biol* **1994**; 238: 733-47.
[119] Kaur H,Raghava GP. An evaluation of beta-turn prediction methods. *Bioinformatics* **2002**; 18: 1508-14.
[120] Kaur H,Raghava GP. BetaTPred: prediction of beta-TURNS in a protein using statistical algorithms. *Bioinformatics* **2002**; 18: 498-9.
[121] Kaur H,Raghava GP. BTEVAL: a server for evaluation of beta-turn prediction methods. *J Bioinform Comput Biol* **2003**; 1: 495-504.
[122] Kaur H,Raghava GP. Prediction of beta-turns in proteins from multiple alignment using neural network. *Protein Sci* **2003**; 12: 627-34.
[123] Kaur H,Raghava GP. A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* **2004**; 20: 2751-8.

[124] Fuchs PF,Alix AJ. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins* **2005**; 59: 828-39.
[125] Zhang Q, Yoon S,Welsh WJ. Improved method for predicting beta-turn using support vector machine. *Bioinformatics* **2005**; 21: 2370-4.
[126] Pham TH, Satou K,Ho TB. Support vector machines for prediction and analysis of beta and gamma-turns in proteins. *J Bioinform Comput Biol* **2005**; 3: 343-58.
[127] Ramakrishnan C, Srinivasan N,Nataraj DV. Motifs and conformational analysis of amino acid residues adjoining beta-turns in proteins. *Int J Pept Protein Res* **1996**; 48: 420-8.
[128] Guruprasad K,Rajkumar S. Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *J Biosci* **2000**; 25: 143-56.
[129] Guruprasad K, Prasad MS,Kumar GR. Analysis of gammabeta, betagamma, gammagamma, betabeta continuous turns in proteins. *J Pept Res* **2001**; 57: 292-300.
[130] Guruprasad K, Prasad MS,Kumar GR. Analysis of gammabeta, betagamma, gammagamma, betabeta multiple turns in proteins. *J Pept Res* **2000**; 56: 250-63.
[131] Guruprasad K, Rao MJ, Adindla S,Guruprasad L. Combinations of turns in proteins. *J Pept Res* **2003**; 62: 167-74.
[132] Chou KC. Prediction of tight turns and their types in proteins. *Anal Biochem* **2000**; 286: 1-16.
[133] Nataraj D, Srinivasan N, Sowdhamini R,Ramakrishnan C. Alpha-turns in pro tein structures. *Curr. Sci.* **1995**; 69: 434-47.
[134] Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C,Saviano M. Discovering protein secondary structures: classification and description of isolated alpha-turns. *Biopolymers* **1996**; 38: 705-21.
[135] Ramakrishnan C,Nataraj DV. Energy minimization studies on alpha-turns. *J Pept Sci* **1998**; 4: 239-52.
[136] Rajashankar KR,Ramakumar S. Pi-turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence. *Protein Sci* **1996**; 5: 932-46.
[137] Artymiuk PJ, Blake CC,Sippel AE. Genes pieced together--exons delineate homologous structures of diverged lysozymes. *Nature* **1981**; 290: 287-8.
[138] Baker EN. Structure of azurin from Alcaligenes denitrificans refinement at 1.8 A resolution and comparison of the two crystallographically independent molecules. *J Mol Biol* **1988**; 203: 1071-95.
[139] Stout CD. Refinement of the 7 Fe ferredoxin from Azotobacter vinelandii at 1.9 A resolution. *J Mol Biol* **1989**; 205: 545-55.
[140] Zanotti G, Wieland T, Benedetti E, Di Blasio B, Pavone V,Pedone C. Structure-toxicity relationships in the amatoxin series. Synthesis of S-deoxy[gamma(R)-hydroxy-Ile3]-amaninamide, its crystal and molecular structure and inhibitory efficiency. *Int J Pept Protein Res* **1989**; 34: 222-8.
[141] Wang JH, Yan YW, Garrett TP, Liu JH, Rodgers DW, Garlick RL, Tarr GE, Husain Y, Reinherz EL,Harrison SC. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature* **1990**; 348: 411-8.
[142] Chou KC. Prediction and classification of alpha-turn types. *Biopolymers* **1997**; 42: 837-53.
[143] Dasgupta B, Pal L, Basu G,Chakrabarti P. Expanded turn conformations: characterization and sequence-structure correspondence in alpha-turns with implications in helix folding. *Proteins* **2004**; 55: 305-15.
[144] Pauling L,Corey RB. The structure of fibrous proteins of the collagen-gelatin group. *Proc Natl Acad Sci U S A* **1951**; 37: 272-81.
[145] Cowan PM, McGavin S,North AC. The polypeptide chain configuration of collagen. *Nature* **1955**; 176: 1062-4.
[146] Makowska J, Rodziewicz-Motowidlo S, Baginska K, Vila JA, Liwo A, Chmurzynski L,Scheraga HA. 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* **2006**.
[147] Vlasov PK, Kilosanidze GT, Ukrainskii DL, Kuz'min AV, Tumanian VG,Esipova NG. [Left-handed helix conformation of poly-L-proline II type in globular proteins. Statistics of incidence and a role of sequence]. *Biofizika* **2001**; 46: 573-6.
[148] Fleming PJ, Fitzkee NC, Mezei M, Srinivasan R,Rose GD. A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). *Protein Sci* **2005**; 14: 111-8.
[149] Kelly MA, Chellgren BW, Rucker AL, Troutman JM, Fried MG, Miller AF,Creamer TP. Host-guest study of left-handed polyproline II helix formation. *Biochemistry* **2001**; 40: 14376-83.
[150] Cubellis MV, Caillez F, Blundell TL,Lovell SC. Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. *Proteins* **2005**; 58: 880-92.
[151] Chen K, Liu Z, Zhou C, Shi Z,Kallenbach NR. Neighbor effect on PPII conformation in alanine peptides. *J Am Chem Soc* **2005**; 127: 10146-7.
[152] Pappu RV,Rose GD. A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci* **2002**; 11: 2437-55.

[153] Shi Z, Olson CA, Rose GD, Baldwin RL,Kallenbach NR. Polyproline II structure in a sequence of seven alanine residues. *Proc Natl Acad Sci U S A* **2002**; 99: 9190-5.
[154] Mezei M, Fleming PJ, Srinivasan R,Rose GD. Polyproline II helix is the preferred conformation for unfolded polyalanine in water. *Proteins* **2004**; 55: 502-7.
[155] Adzhubei AA,Sternberg MJ. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol* **1993**; 229: 472-93.
[156] Creamer TP. Left-handed polyproline II helix formation is (very) locally driven. *Proteins* **1998**; 33: 218-26.
[157] Stapley BJ,Creamer TP. A survey of left-handed polyproline II helices. *Protein Sci* **1999**; 8: 587-95.
[158] Creamer TP,Campbell MN. Determinants of the polyproline II helix from modeling studies. *Adv Protein Chem* **2002**; 62: 263-82.
[159] Chellgren BW,Creamer TP. Short sequences of non-proline residues can adopt the polyproline II helical conformation. *Biochemistry* **2004**; 43: 5864-9.
[160] Chellgren BW, Miller AF,Creamer TP. Evidence for polyproline II helical structure in short polyglutamine tracts. *J Mol Biol* **2006**; 361: 362-71.
[161] Whittington SJ,Creamer TP. Salt bridges do not stabilize polyproline II helices. *Biochemistry* **2003**; 42: 14690-5.
[162] Liu Z, Chen K, Ng A, Shi Z, Woody RW,Kallenbach NR. Solvent dependence of PII conformation in model alanine peptides. *J Am Chem Soc* **2004**; 126: 15141-50.
[163] Berisio R, Loguercio S, De Simone A, Zagari A,Vitagliano L. Polyproline helices in protein structures: A statistical survey. *Protein Pept Lett* **2006**; 13: 847-54.
[164] Blanch EW, Morozova-Roche LA, Cochran DA, Doig AJ, Hecht L,Barron LD. Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J Mol Biol* **2000**; 301: 553-63.
[165] Eker F, Griebenow K,Schweitzer-Stenner R. Abeta(1-28) fragment of the amyloid peptide predominantly adopts a polyproline II conformation in an acidic solution. *Biochemistry* **2004**; 43: 6893-8.
[166] Hicks JM,Hsu VL. The extended left-handed helix: a simple nucleic acid-binding motif. *Proteins* **2004**; 55: 330-8.
[167] Zagrovic B, Lipfert J, Sorin EJ, Millett IS, van Gunsteren WF, Doniach S,Pande VS. Unusual compactness of a polyproline type II structure. *Proc Natl Acad Sci U S A* **2005**; 102: 11698-703.
[168] Thornton JM, Sibanda BL, Edwards MS,Barlow DJ. Analysis, design and modification of loop regions in proteins. *Bioessays* **1988**; 8: 63-9.
[169] Sibanda BL,Thornton JM. Beta-hairpin families in globular proteins. *Nature* **1985**; 316: 170-4.
[170] Milner-White EJ,Poet R. Four classes of beta-hairpins in proteins. *Biochem J* **1986**; 240: 289-92.
[171] Sibanda BL, Blundell TL,Thornton JM. Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol* **1989**; 206: 759-77.
[172] Sibanda BL,Thornton JM. Conformation of beta hairpins in protein structures: classification and diversity in homologous structures. *Methods Enzymol* **1991**; 202: 59-82.
[173] Efimov AV. Structure of coiled beta-beta-hairpins and beta-beta-corners. *FEBS Lett* **1991**; 284: 288-92.
[174] Ramirez-Alvarado M, Blanco FJ, Niemann H,Serrano L. Role of beta-turn residues in beta-hairpin formation and stability in designed peptides. *J Mol Biol* **1997**; 273: 898-912.
[175] Gunasekaran K, Ramakrishnan C,Balaram P. Beta-hairpins in proteins revisited: lessons for de novo design. *Protein Eng* **1997**; 10: 1131-41.
[176] Skelton NJ, Russell S, de Sauvage F,Cochran AG. Amino acid determinants of beta-hairpin conformation in erythropoeitin receptor agonist peptides derived from a phage display library. *J Mol Biol* **2002**; 316: 1111-25.
[177] Sibanda BL,Thornton JM. Accommodating sequence changes in beta-hairpins in proteins. *J Mol Biol* **1993**; 229: 428-47.
[178] Blandl T, Cochran AG,Skelton NJ. Turn stability in beta-hairpin peptides: Investigation of peptides containing 3:5 type I G1 bulge turns. *Protein Sci* **2003**; 12: 237-47.
[179] Kim J, Brych SR, Lee J, Logan TM,Blaber M. Identification of a key structural element for protein folding within beta-hairpin turns. *J Mol Biol* **2003**; 328: 951-61.
[180] Sowdhamini R, Srinivasan N, Ramakrishnan C,Balaram P. Orthogonal beta beta motifs in proteins. *J Mol Biol* **1992**; 223: 845-51.
[181] Efimov AV. Structure of alpha-alpha-hairpins with short connections. *Protein Eng* **1991**; 4: 245-50.
[182] Efimov AV. Structural similarity between two-layer alpha/beta and beta-proteins. *J Mol Biol* **1995**; 245: 402-15.
[183] Engel DE,Degrado WF. alpha-alpha linking motifs and interhelical orientations. *Proteins* **2005**.
[184] Wintjens RT, Rooman MJ,Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in

proteins. *J Mol Biol* **1996**; 255: 235-53.
[185] Edwards MS, Sternberg JE,Thornton JM. Structural and sequence patterns in the loops of beta alpha beta units. *Protein Eng* **1987**; 1: 173-81.
[186] Efimov AV. Super-secondary structures involving triple-strand beta-sheets. *FEBS Lett* **1993**; 334: 253-6.
[187] Leszczynski JF,Rose GD. Loops in globular proteins: a novel category of secondary structure. *Science* **1986**; 234: 849-55.
[188] Fetrow JS, Horner SR, Oehrl W, Schaak DL, Boose TL,Burton RE. Analysis of the structure and stability of omega loop A replacements in yeast iso-1-cytochrome c. *Protein Sci* **1997**; 6: 197-210.
[189] Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. *Faseb J* **1995**; 9: 708-17.
[190] Pal M,Dasgupta S. The nature of the turn in omega loops of proteins. *Proteins* **2003**; 51: 591-606.
[191] Fetrow JS, Cardillo TS,Sherman F. Deletions and replacements of omega loops in yeast iso-1-cytochrome c. *Proteins* **1989**; 6: 372-81.
[192] Krishna MM, Lin Y, Rumbley JN,Englander SW. Cooperative omega loops in cytochrome c: role in folding and function. *J Mol Biol* **2003**; 331: 29-36.
[193] Bek E,Berry R. Prohormonal cleavage sites are associated with omega loops. *Biochemistry* **1990**; 29: 178-83.
[194] Kishore R, Samuel M, Khan MY, Hand J, Frenz DA,Newman SA. Interaction of the NH2-terminal domain of fibronectin with heparin. Role of the omega-loops of the type I modules. *J Biol Chem* **1997**; 272: 17078-85.
[195] Sanschagrin F, Theriault E, Sabbagh Y, Voyer N,Levesque RC. Combinatorial biochemistry and shuffling of TEM, SHV and Streptomyces albus omega loops in PSE-4 class A beta-lactamase. *J Antimicrob Chemother* **2000**; 45: 517-20.
[196] Murphy ME, Fetrow JS, Burton RE,Brayer GD. The structure and function of omega loop A replacements in cytochrome c. *Protein Sci* **1993**; 2: 1429-40.
[197] Mulligan-Pullyblank P, Spitzer JS, Gilden BM,Fetrow JS. Loop replacement and random mutagenesis of omega-loop D, residues 70-84, in iso-1-cytochrome c. *J Biol Chem* **1996**; 271: 8633-45.
[198] Sinibaldi F, Piro MC, Howes BD, Smulevich G, Ascoli F,Santucci R. Rupture of the hydrogen bond linking two Omega-loops induces the molten globule state at neutral pH in cytochrome c. *Biochemistry* **2003**; 42: 7604-10.
[199] Berezovsky IN, Grosberg AY,Trifonov EN. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett* **2000**; 466: 283-6.
[200] Lamarine M, Mornon JP, Berezovsky N,Chomilier J. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell Mol Life Sci* **2001**; 58: 492-8.
[201] Papandreou N, Berezovsky IN, Lopes A, Eliopoulos E,Chomilier J. Universal positions in globular proteins. *Eur J Biochem* **2004**; 271: 4762-8.
[202] Berezovsky IN,Trifonov EN. Loop fold nature of globular proteins. *Protein Eng* **2001**; 14: 403-7.
[203] Tramontano A, Chothia C,Lesk AM. Structural determinants of the conformations of medium-sized loops in proteins. *Proteins* **1989**; 6: 382-94.
[204] Tramontano A,Lesk AM. Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins* **1992**; 13: 231-45.
[205] Rice PA, Goldman A,Steitz TA. A helix-turn-strand structural motif common in alpha-beta proteins. *Proteins* **1990**; 8: 334-40.
[206] Ring CS, Kneller DG, Langridge R,Cohen FE. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol* **1992**; 224: 685-99.
[207] Ring CS,Cohen FE. Conformational sampling of loop structures using genetic algorithm. *Israel J Chem* **1994**; 34: 245-52.
[208] Sun Z,Jiang B. Patterns and conformations of commonly occurring supersecondary structures (basic motifs) in protein data bank. *J Protein Chem* **1996**; 15: 675-90.
[209] Rufino SD, Donate LE, Canard L,Blundell TL. Analysis, clustering and prediction of the conformation of short and medium size loops connecting regular secondary structures. *Pac Symp Biocomput* **1996**: 570-89.
[210] Donate LE, Rufino SD, Canard LH,Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci* **1996**; 5: 2600-16.
[211] Rufino SD, Donate LE, Canard LH,Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J Mol Biol* **1997**; 267: 352-67.
[212] Burke DF,Deane CM. Improved protein loop prediction from sequence alone. *Protein Eng* **2001**; 14: 473-

8.
[213] Burke DF, Deane CM,Blundell TL. Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics* **2000**; 16: 513-9.
[214] Geetha V,Munson P, Analysis of linkers of regular secondary structural regions in proteins, In, D. Marshak Eds, Techniques in Protein Chemistry. Academic Press, Orlando, Florida 1997 p. 667-77.
[215] Geetha V,Munson PJ. Linkers of secondary structures in proteins. *Protein Sci* **1997**; 6: 2538-47.
[216] Kwasigroch JM, Chomilier J,Mornon JP. A global taxonomy of loops in globular proteins. *J Mol Biol* **1996**; 259: 855-72.
[217] Alland C, Moreews F, Boens D, Carpentier M, Chiusa S, Lonquety M, Renault N, Wong Y, Cantalloube H, Chomilier J, Hochez J, Pothier J, Villoutreix BO, Zagury JF,Tuffery P. RPBS: a web resource for structural bioinformatics. *Nucleic Acids Res* **2005**; 33: W44-9.
[218] Oliva B, Bates PA, Querol E, Aviles FX,Sternberg MJ. An automated classification of the structure of protein loops. *J Mol Biol* **1997**; 266: 814-30.
[219] Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ,Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res* **2004**; 32: D185-8.
[220] Fernandez-Fuentes N, Hermoso A, Espadaler J, Querol E, Aviles FX,Oliva B. Classification of common functional loops of kinase super-families. *Proteins* **2004**; 56: 539-55.
[221] Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJ,Oliva B. Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. *Proteins* **2005**; 60: 746-57.
[222] Espadaler J, Querol E, Aviles FX,Oliva B. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics* **2006**; 22: 2237-43.
[223] Li W, Liu Z,Lai L. Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers* **1999**; 49: 481-95.
[224] Holm L,Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **1997**; 25: 231-4.
[225] Michalsky E, Goede A,Preissner R. Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. *Protein Eng* **2003**; 16: 979-85.
[226] Lesk AM, Introduction to Bioinformatics. 2005, Oxford: Oxford University Press.
[227] Schulz GE, Barry CD, Friedman J, Chou PY, Fasman GD, Finkelstein AV, Lim VI, Pititsyn OB, Kabat EA, Wu TT, Levitt M, Robson B,Nagano K. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature* **1974**; 250: 140-2.
[228] Robson B,Garnier J, Introduction to proteins and protein engineering. 1986, Amsterdam: Elsevier Press.
[229] Woodcock S, Mornon JP,Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng* **1992**; 5: 629-35.
[230] Levitt M,Greer J. Automatic identification of secondary structure in globular proteins. *J Mol Biol* **1977**; 114: 181-239.
[231] Kabsch W,Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**; 22: 2577-637.
[232] Andersen CA, Palmer AG, Brunak S,Rost B. Continuum secondary structure captures protein flexibility. *Structure (Camb)* **2002**; 10: 175-84.
[233] Frishman D,Argos P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**; 23: 566-79.
[234] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T,Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **1977**; 112: 535-42.
[235] Andersen CA,Rost B. Secondary structure assignment. *Methods Biochem Anal* **2003**; 44: 341-63.
[236] Andersen CAF,Rost B, Secondary structure assignement, In, P. Bourne Eds, Structural Bioinformatics. 2002 p. 341-64.
[237] Carter P, Andersen CA,Rost B. DSSPcont: Continuous secondary structure assignments for proteins. *Nucleic Acids Res* **2003**; 31: 3293-5.
[238] Al-Karadaghi S, Cedergren-Zeppezauer ES, Dauter Z,Wilson KS. Refined structure of Cu-substituted alcohol dehydrogenase at 2.1 A resolution. *Acta Crystallogr D Biol Crystallogr* **1995**; 51: 805-13.
[239] Al-Karadaghi S, Cedergren-Zeppezauer ES,Hovmoller S. Refined crystal structure of liver alcohol dehydrogenase-NADH complex at 1.8 A resolution. *Acta Crystallogr D Biol Crystallogr* **1994**; 50: 793-807.
[240] Smith D, *SSTRUC: A program to calculate a secondary structural summary.*, in *Department of Crystallography, Birkbeck College, University of London.* 1989.
[241] Labesse G, Colloc'h N, Pothier J,Mornon JP. P-SEA: a new efficient assignment of secondary structure

from C alpha trace of proteins. *Comput Appl Biosci* **1997**; 13: 291-5.
[242] Srinivasan R,Rose GD. A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* **1999**; 96: 14258-63.
[243] Fitzkee NC,Rose GD. Steric restrictions in protein folding: an alpha-helix cannot be followed by a contiguous beta-strand. *Protein Sci* **2004**; 13: 633-9.
[244] Fitzkee NC, Fleming PJ,Rose GD. The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* **2005**; 58: 852-4.
[245] Cubellis MV, Cailliez F,Lovell SC. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics* **2005**; 6 Suppl 4: S8.
[246] King SM,Johnson WC. Assigning secondary structure from protein coordinate data. *Proteins* **1999**; 35: 313-20.
[247] Sklenar H, Etchebest C,Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **1989**; 6: 46-60.
[248] Dupuis F, Sadoc JF,Mornon JP. Protein secondary structure assignment through Voronoi tessellation. *Proteins* **2004**; 55: 519-28.
[249] Dupuis F, Sadoc JF, Jullien R, Angelov B,Mornon JP. Voro3D: 3D Voronoi tessellations applied to protein structures. *Bioinformatics* **2005**; 21: 1715-6.
[250] Taylor T, Rivera M, Wilson G,Vaisman, II. New method for protein secondary structure assignment based on a simple topological descriptor. *Proteins* **2005**; 60: 513-24.
[251] Parisien M,Major F. A New Catalog of Protein Beta-Sheets. *Proteins* **2005**: in press.
[252] Majumdar I, Krishna SS,Grishin NV. PALSSE: A program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics* **2005**; 6: 202.
[253] Colloc'h N, Etchebest C, Thoreau E, Henrissat B,Mornon JP. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng* **1993**; 6: 377-82.
[254] Sukumar N, Dewanti AR, Mitra B,Mathews FS. High resolution structures of an oxidized and reduced flavoprotein. The water switch in a soluble form of (S)-mandelate dehydrogenase. *J Biol Chem* **2004**; 279: 3749-57.
[255] Bornot A,de Brevern AG. Protein beta-turn assignments. *Bioinformation* **2006**; 1: 153-5.
[256] de Brevern AG, Benros C,Hazout S, Structural Alphabet: From a Local Point of View to a Global Description of Protein 3D Structures, In, P.V. Yan Eds, Bioinformatics: New Research. Nova Publishers 2005 p. 128-87.
[257] Kabsch W, Mannherz HG, Suck D, Pai EF,Holmes KC. Atomic structure of the actin:DNase I complex. *Nature* **1990**; 347: 37-44.
[258] Michalopoulos I, Torrance GM, Gilbert DR,Westhead DR. TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res* **2004**; 32: D251-4.
[259] Chothia C,Finkelstein AV. The classification and origins of protein folding patterns. *Annu Rev Biochem* **1990**; 59: 1007-39.
[260] Hutchinson EG,Thornton JM. The Greek key motif: extraction, classification and analysis. *Protein Eng* **1993**; 6: 233-45.
[261] Gelly JC, Gracy J, Kaas Q, Le-Nguyen D, Heitz A,Chiche L. The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic Acids Res* **2004**; 32: D156-9.
[262] Lesk AM,Rose GD. Folding units in globular proteins. *Proc Natl Acad Sci U S A* **1981**; 78: 4304-8.
[263] Sowdhamini R,Blundell TL. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci* **1995**; 4: 506-20.
[264] Tsai CJ,Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci* **1997**; 6: 24-42.
[265] Gelly JC, de Brevern AG,Hazout S. 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. *Bioinformatics* **2006**; 22: 129-33.
[266] Gelly JC, Etchebest C, Hazout S,de Brevern AG. Protein Peeling 2: a web server to convert protein structures into series of protein units. *Nucleic Acids Res* **2006**; 34: W75-8.
[267] Bystroff C,Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* **1998**; 281: 565-77.
[268] Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE,Kern D. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **2005**; 438: 117-21.
[269] Teague SJ. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* **2003**; 2: 527-41.
[270] Raveh B, Rahat O, Basri R,Schreiber G. Rediscovering secondary structures as network motifs--an unsupervised learning approach. *Bioinformatics* **2007**; 23: e163-9.
[271] Benros C, Martin J, Tyagi M,de Brevern AG, Description of the local protein structure. I. Classical approaches, In, A.G. de Brevern Eds, Recent Research Developments in Protein Engineering. Research

Signpost, Trivandrum 2007 p. in press.
[272] Unger R, Harel D, Wherland S,Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **1989**; 5: 355-73.
[273] Tuffery P, Guyon F,Derreumaux P. Improved greedy algorithm for protein structure reconstruction. *J Comput Chem* **2005**; 26: 506-13.
[274] de Brevern AG, Camproux A-C, Hazout S, Etchebest C,Tuffery P, Protein structural alphabets: beyond the secondary structure description, In, S. Sangadai Eds, Recent Research Developments in Protein Engineering. Research Signpost, Trivandrum 2001 p. 319-31.
[275] Karchin R, *Evaluating local structure alphabets for protein structure prediction*. 2003. p. 301.
[276] Tyagi M, Benros C, Martin J,de Brevern AG, Description of the local protein structure. II. Novel approaches., In, A.G. de Brevern Eds, Recent Research Developments in Protein Engineering. Research Signpost, Trivandrum 2007 p. in press.
[277] Unger R,Sussman JL. The importance of short structural motifs in protein structure analysis. *J Comput Aided Mol Des* **1993**; 7: 457-72.
[278] Prestrelski SJ, Williams AL, Jr.,Liebman MN. Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results. *Proteins* **1992**; 14: 430-9.
[279] Schuchhardt J, Schneider G, Reichelt J, Schomburg D,Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng* **1996**; 9: 833-42.
[280] Rooman MJ, Rodriguez J,Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* **1990**; 213: 327-36.
[281] Fetrow JS, Palumbo MJ,Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* **1997**; 27: 249-71.
[282] Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF,Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng* **1999**; 12: 1063-73.
[283] Camproux AC, Tuffery P, Buffat L, Andre C, Boisvieux JF,Hazout S. Using short structural building blocks defined by a Hidden Markov Model for analysing patterns between regular secondary structures. *Theor. Chem. Acc* **1999**; 101(1-3): 33-40.
[284] Micheletti C, Seno F,Maritan A. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* **2000**; 40: 662-74.
[285] de Brevern AG, Etchebest C,Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **2000**; 41: 271-87.
[286] Hartigan JA,Wong MA. k-means. *Applied Statistics* **1979**; 28: 100-15.
[287] Kohonen T. Self-organized formation of topologically correct feature maps. *Biol. Cybern* **1982**; 43: 59-69.
[288] Kohonen T, Self-Organizing Maps (3rd edition). 2001: Springer. 501.
[289] Schneider G,Wrede P. Artificial neural networks for computer-based molecular design. *Prog Biophys Mol Biol* **1998**; 70: 175-222.
[290] Rabiner LR. A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE* **1989**; 77: 257-86.
[291] Sali A,Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **1993**; 234: 779-815.
[292] Fiser A, Do RK,Sali A. Modeling of loops in protein structures. *Protein Sci* **2000**; 9: 1753-73.
[293] Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F,Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **2000**; 29: 291-325.
[294] Sali A,Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* **1990**; 212: 403-28.
[295] Jones TA,Thirup S. Using known substructures in protein model building and crystallography. *Embo J* **1986**; 5: 819-22.
[296] Claessens M, Van Cutsem E, Lasters I,Wodak S. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng* **1989**; 2: 335-45.
[297] Levitt M. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* **1992**; 226: 507-33.
[298] Du P, Andrec M,Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng* **2003**; 16: 407-14.
[299] Unger R, Harel D, Wherland S,Sussman JL. Analysis of dihedral angles distribution: The doublets distribution determines polypeptides conformations. *Biopolymers* **1990**; 30: 499-508.
[300] Prestrelski SJ, Byler DM,Liebman MN. 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* **1992**; 14: 440-50.
[301] Zhang X, Fetrow JS, Rennie WA, Waltz DL,Berg G. Automatic derivation of substructures yields novel structural building blocks in globular proteins. *Proc Int Conf Intell Syst Mol Biol* **1993**; 1: 438-46.
[302] Zhang KY, Fetrow JS,Berg G, Design of an Auto-Associative Neural Network with Hidden Layer Activations that were used to Reclassify Local Protein Structure, In, Eds, Techniques in Protein Chemistry V. Academic Press, Inc 1994 p. 397-404.
[303] Sammon J, J. W. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* **1969**; 18: 401-09.
[304] Rooman MJ, Rodriguez J,Wodak SJ. Relations between protein sequence and structure and their significance. *J Mol Biol* **1990**; 213: 337-50.
[305] Fetrow JS,Berg G. *Using information theory to discover side chain rotamer classes: Analysis of the effects of local backbone structure*. in Pac Symp Biocomput. 1999.
[306] de Brevern AG, *Analyse et Prédiction de la structure locale des protéines*. 1998, Universite Paris 7 - Denis Diderot: Paris. p. 22.
[307] Park BH,Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol* **1995**; 249: 493-507.
[308] Kolodny R, Koehl P, Guibas L,Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* **2002**; 323: 297-307.
[309] Kolodny R,Levitt M. Protein decoy assembly using short fragments under geometric constraints. *Biopolymers* **2003**; 68: 278-85.
[310] Schwartz G. Estimating the dimension of a model. *Ann Stat* **1978**; 6: 461-64.
[311] Tuffery P,Derreumaux P. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Go potential and greedy algorithm. *Proteins* **2005**.
[312] Forcellino F,Derreumaux P. Computer simulations aimed at structure prediction of supersecondary motifs in proteins. *Proteins* **2001**; 45: 159-66.
[313] Lazaridis T,Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* **1999**; 288: 477-87.
[314] Camproux AC,Tuffery P. Hidden Markov Model-derived structural alphabet for proteins: The learning of protein local shapes captures sequence specificity. *Biochim Biophys Acta* **2005**; 1724: 394-403.
[315] Gautier R, Camproux AC,Tuffery P. SCit: web tools for protein side chain conformation analysis. *Nucleic Acids Res* **2004**; 32: W508-11.
[316] Maupetit J, Gautier R,Tuffery P. SABBAC: online Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace. *Nucleic Acids Res* **2006**; 34: W147-51.
[317] Bonneau R, Strauss CE,Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* **2001**; 43: 1-11.
[318] Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE,Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **2001**; Suppl 5: 119-26.
[319] Rooman MJ,Wodak SJ. Identification of predictive sequence motifs limited by protein structure data base size. *Nature* **1988**; 335: 45-9.
[320] Rooman MJ, Wodak SJ,Thornton JM. Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng* **1989**; 3: 23-7.
[321] Rooman MJ,Wodak SJ. Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *Proteins* **1991**; 9: 69-78.
[322] Rooman MJ,Wodak SJ. Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. *Biochemistry* **1992**; 31: 10239-49.
[323] Gibrat JF, Garnier J,Robson B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J Mol Biol* **1987**; 198: 425-43.
[324] Garnier J, Levin JM, Gibrat JF,Biou V. Secondary structure prediction and protein design. *Biochem Soc Symp* **1990**; 57: 11-24.
[325] Rooman MJ, Kocher JP,Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. *J Mol Biol* **1991**; 221: 961-79.
[326] Rooman MJ, Kocher JP,Wodak SJ. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry* **1992**; 31: 10226-38.
[327] Han KF,Baker D. Recurring local sequence motifs in proteins. *J Mol Biol* **1995**; 251: 176-87.
[328] Han KF,Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A* **1996**; 93: 5814-8.
[329] Bystroff C, Simons KT, Han KF,Baker D. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol* **1996**; 7: 417-21.
[330] Han KF, Bystroff C,Baker D. Three-dimensional structures and contexts associated with recurrent amino

acid sequence patterns. *Protein Sci* **1997**; 6: 1587-90.
[331] Schneider R, de Daruvar A,Sander C. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* **1997**; 25: 226-30.
[332] Rost B,Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A* **1993**; 90: 7558-62.
[333] Rost B,Sander C. Secondary structure prediction of all-helical proteins in two states. *Protein Eng* **1993**; 6: 831-6.
[334] Bystroff C,Baker D. Blind predictions of local protein structure in CASP2 targets using the I-sites library. *Proteins* **1997**; Suppl 1: 167-71.
[335] Yi Q, Bystroff C, Rajagopal P, Klevit RE,Baker D. Prediction and structural characterization of an independently folding substructure in the src SH3 domain. *J Mol Biol* **1998**; 283: 293-300.
[336] Bystroff C, Thorsson V,Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **2000**; 301: 173-90.
[337] de Brevern AG, *Nouvelles stratégies d'analyses et de prédiction des structures tridimensionnelles des protéines*, in *Biology (Analyses de Génomes et Modélisation Moléculaire)*. 2001, University Paris 7: Paris. p. 208.
[338] Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA,Baker D. Prediction of CASP-6 structures using automated Robetta protocols. *Proteins* **2005**.
[339] Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE,Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **2003**; 53 Suppl 6: 457-68.
[340] Chivian D, Robertson T, Bonneau R,Baker D. Ab initio methods. *Methods Biochem Anal* **2003**; 44: 547-57.
[341] Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL,Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**; 302: 1364-8.
[342] Kuhlman B, O'Neill JW, Kim DE, Zhang KY,Baker D. Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J Mol Biol* **2002**; 315: 471-7.
[343] Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA,Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* **2003**; 331: 281-99.
[344] Bowers PM, Strauss CE,Baker D. De novo protein structure determination using sparse NMR data. *J Biomol NMR* **2000**; 18: 311-8.
[345] Rohl CA,Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* **2002**; 124: 2723-9.
[346] Rohl CA, Strauss CE, Chivian D,Baker D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **2004**; 55: 656-77.
[347] Bystroff C,Shao Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* **2002**; 18 Suppl 1: S54-61.
[348] de Brevern AG. New assessment of Protein Blocks. *In Silico Biology* **2005**; 5: 283-89.
[349] de Brevern AG, Valadie H, Hazout S,Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci* **2002**; 11: 2871-86.
[350] de Brevern AG, Etchebest C, Benros C,Hazout S. "Pinning strategy": a novel approach for predicting the backbone structure in terms of Protein Blocks from sequence. *J Biosciences* **2007**; 32: 51-72.
[351] Etchebest C, Benros C, Hazout S,de Brevern AG. A structural alphabet for local protein structures: Improved prediction methods. *Proteins* **2005**: 810-27.
[352] de Brevern AG, Benros C, Gautier R, Valadie H, Hazout S,Etchebest C. Local backbone structure prediction of proteins. *In Silico Biol* **2004**; 4: 381-6.
[353] Hunter CG,Subramaniam S. Protein fragment clustering and canonical local shapes. *Proteins* **2003**; 50: 580-8.
[354] Hunter CG,Subramaniam S. Protein local structure prediction from sequence. *Proteins* **2003**; 50: 572-9.
[355] Sander O, Sommer I,Lengauer T. Local protein structure prediction using discriminative models. *BMC Bioinformatics* **2006**; 7: 14.
[356] Quinlan JR. Induction of decision trees. *Machine Learning* **1986**; 1: 81–106.
[357] Schölkopf B,Smola A, Learning with Kernels. 2002, Cambridge, MA: MIT Press.
[358] randomForest. p. [http://cran.r-project.org/src/contrib/Descriptions/randomForest.html].
[359] Karchin R, Cline M, Mandel-Gutfreund Y,Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **2003**; 51: 504-14.
[360] Karplus K, Karchin R, Shackelford G,Hughey R. Calibrating E-values for hidden Markov models with reverse-sequence null models. *Bioinformatics* **2005**.

[361] Camproux AC, Gautier R,Tuffery P. A hidden markov model derived structural alphabet for proteins. *J Mol Biol* **2004**; 339: 591-605.
[362] Camproux AC, Brevern AG, Hazout S,Tufféry P. Exploring the use of a structural alphabet for structural prediction of protein loops. *Theo Chem Acc* **2001**; 106: 28-35.
[363] Regad L, Martin J,Camproux AC. *Identification of non random motifs in loops using a structural alphabet*. in Computational Intelligence and Bioinformatics and Computational Biology. 2006. Toronto, Canada: IEEE.
[364] de Brevern AG,Hazout S. Compacting local protein folds with a "hybrid protein model". *Theo Chem Acc* **2001**; 106: 36-47.
[365] de Brevern AG,Hazout S. *Compactage d'une base de données protéiques recodées dans un alphabet structural*. in Secondes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM) pour la génomique. 2001. Toulouse.
[366] Benros C, Hazout S,de Brevern AG. *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. 2002. Lyon, France.
[367] Benros C, de Brevern AG,Hazout S. *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. 2003.
[368] de Brevern AG,Hazout S. 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics* **2003**; 19: 345-53.
[369] Guyon F, Camproux AC, Hochez J,Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res* **2004**; 32: W545-8.
[370] Holm L,Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* **1993**; 233: 123-38.
[371] Pandini A, Bonati L, Fraternali F,Kleinjung J. MinSet: a general approach to derive maximally representative database subsets by using fragment dictionaries and its application to the SCOP database. *Bioinformatics* **2007**; 23: 515-6.
[372] Tyagi M, Gowri VS, Srinivasan N, de Brevern AG,Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins* **2006**; 65: 32-9.
[373] Gowri VS, Pandit SB, Karthik PS, Srinivasan N,Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res* **2003**; 31: 486-8.
[374] Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R,Srinivasan N. SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res* **2002**; 30: 289-93.
[375] Balaji S,Srinivasan N. Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng* **2001**; 14: 219-26.
[376] Sujatha S, Balaji S,Srinivasan N. PALI: a database of alignments and phylogeny of homologous protein structures. *Bioinformatics* **2001**; 17: 375-6.
[377] Balaji S, Sujatha S, Kumar SS,Srinivasan N. PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res* **2001**; 29: 61-5.
[378] Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG,Offmann B. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res* **2006**; 34: W119-23.
[379] Carpentier M, Brouillet S,Pothier J. YAKUSA: a fast structural database scanning method. *Proteins* **2005**; 61: 137-51.
[380] Yang JM,Tung CH. Protein structure database search and evolutionary classification. *Nucleic Acids Res* **2006**; 34: 3646-59.
[381] Tung CH, Huang JW,Yang JM. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search. *Genome Biol* **2007**; 8: R31.
[382] Altschul SF, Gish W, Miller W, Myers EW,Lipman DJ. Basic local alignment search tool. *J Mol Biol* **1990**; 215: 403-10.
[383] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W,Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-402.
[384] Shao Y,Bystroff C. Predicting interresidue contacts using templates and pathways. *Proteins* **2003**; 53 Suppl 6: 497-502.
[385] Hou Y, Hsu W, Lee ML,Bystroff C. Remote homolog detection using local sequence-structure

correlations. *Proteins* **2004**; 57: 518-30.
[386] Yuan X,Bystroff C. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics* **2005**; 21: 1010-9.
[387] Huang YM,Bystroff C. Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. *Bioinformatics* **2005**.
[388] Tsai CJ, Maizel JV, Jr.,Nussinov R. Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci U S A* **2000**; 97: 12038-43.
[389] Tsai CJ,Nussinov R. The building block folding model and the kinetics of protein folding. *Protein Eng* **2001**; 14: 723-33.
[390] Tsai HH, Tsai CJ, Ma B,Nussinov R. In silico protein design by combinatorial assembly of protein building blocks. *Protein Sci* **2004**; 13: 2753-65.
[391] Lee I-Y, Soong T-T, Ho J-M,Hwang M-J. *Derivation and analysis of fragment librairies of protein structures*. in Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04). 2004.
[392] Soong T-T, Hwang M-J,Chen C-M. Discovery of recurrent structural motifs for approximating three-dimensional protein structures. *Journal of Chinese Chemical Society* **2004**; 51: 1107-14.
[393] Wang S-L, Chen C-M,Hwang M-J. *Classification of protein 3D folds by hidden Markov learning on sequences of structural alphabets*. in 3rd Asia Pacific Bioinformatic Conference. 2005. Singapore.
[394] Li Z, Brener NE, Iyengar SS, Seetharam G, Dua S, Ramakumar S, Manikandan K,Bahren J. *A robust grouping algorithm for clustering of similar protein folding units*. in Fourth virtual conference on Genomics and Bioinformatics. 2004.
[395] Zhen W-M,Liu X. A protein structural alphabet and its substitution matrix CLESUM. *eprint arXiv:q-bio/0412046* **2004**.
[396] Zhen W-M,Liu X. *A protein structural alphabet and its substitution matrix CLESUM*. in International Conference on Computational Science 2005. 2005. Emory University, Atlanta, USA.
[397] Henikoff S,Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **1992**; 89: 10915-9.
[398] Lin HN, Chang JM, Wu KP, Sung TY,Hsu WL. HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* **2005**; 21: 3227-33.
[399] Wu KP, Lin HN, Chang JM, Sung TY,Hsu WL. HYPROSP: a hybrid protein secondary structure prediction algorithm--a knowledge-based approach. *Nucleic Acids Res* **2004**; 32: 5059-65.
[400] Chen C-T, Lin H-N, Wu K-P, Sung T-Y,Hsu W-L. *A Knowledge-based Approach to Protein Local Structure Prediction*. in Asia Pacific Bioinformatics Conference. 2006. Taipei, Taiwan.
[401] Tang TCK, *Discovering protein sequence - structure motifs and two applications to structural prediction*, in *Computer science*. 2004, University of Waterloo: Waterloo, Ontario, Canada.
[402] Tang TCK, Xu J,Ming L. *Discovering protein sequence - structure motifs and two applications to structural prediction*. in Pacific Symposium on Biocomputing (PSB'05). 2005.
[403] Kuang R, Leslie CS,Yang AS. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **2004**; 20: 1612-21.
[404] Yang AS,Wang LY. Local structure prediction with local structure-based sequence profiles. *Bioinformatics* **2003**; 19: 1267-74.
[405] Yang AS,Wang LY. Local structure-based sequence profile database for local and global protein structure predictions. *Bioinformatics* **2002**; 18: 1650-7.
[406] Kabsch W,Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci U S A* **1984**; 81: 1075-8.
[407] Argos P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. *J Mol Biol* **1987**; 197: 331-48.
[408] Wilson IA, Haft DH, Getzoff ED, Tainer JA, Lerner RA,Brenner S. Identical short peptide sequences in unrelated proteins can have different conformations: a testing ground for theories of immune recognition. *Proc Natl Acad Sci U S A* **1985**; 82: 5255-9.
[409] Benros C, de Brevern AG, Etchebest C,Hazout S. Assessing a novel approach for predicting local 3D protein structures from sequence. **2005**: in revision.
[410] de Brevern AG,Hazout S. Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties. *IEEE - Computer Society* **2000**; S1: 49-54.
[411] de Brevern AG. Compartimentation chromosomique. *Biofutur* **2002**; 225: 20-22.
[412] de Brevern AG, Loirat F, Badel-Chagnon A, Andre C, Vincens P,Hazout S. Genome compartimentation by a hybrid chromosome model (HXM). Application to Saccharomyces cerevisae subtelomeres. *Comput Chem* **2002**; 26: 437-45.
[413] de Brevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C,Etchebest C. A structural model of a

seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta* **2005**; 1724: 288-306.
[414] Lindahl E, Hess B,van der Spoel D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* **2001**; 7: 306-17.
[415] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE,Berendsen HJ. GROMACS: fast, flexible, and free. *J Comput Chem* **2005**; 26: 1701-18.
[416] Dudev T, Lin YL, Dudev M,Lim C. First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. *J Am Chem Soc* **2003**; 125: 3168-80.
[417] Petkovich M, Brand NJ, Krust A,Chambon P. A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature* **1987**; 330: 444-50.
[418] Dudev M,Lim C. Discovering structural motifs using a structural alphabet: Application to magnesium-binding sites. *BMC Bioinformatics* **2007**; 8: 212.
[419] Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**; 405: 442-51.
[420] Baldi P, Brunak S, Chauvin Y, Andersen CA,Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**; 16: 412-24.