to be published in *Protein Science* (2009)

Analysis of loop boundaries using different local structure assignment methods.

Manoj Tyagi[1,2,3,+], Aurélie Bornot[2,4,+], Bernard Offmann[1,5,6] and Alexandre G. de Brevern[2,4,§]

[1] Laboratoire de Biochimie et Génétique Moléculaire, Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France.
[2] INSERM UMR-S 726, DSIMB, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), Université Paris Diderot – Paris 7, case 7113, 2, place Jussieu, 75251 PARIS Cedex 05 France.
[3] Current Address: Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, MD 20894.
[4] INSERM UMR-S 665, DSIMB, Université Paris Diderot – Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
[5] INSERM UMR-S 665, DSIMB, Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France.
[6] PEACCEL, 37, rue de Vienne, 97430 Le Tampon, France.

Short title : loop boundary behaviors.

[§] Corresponding author: Alexandre G. de Brevern, INSERM UMR-S 665, DSIMB, Université Paris Diderot – Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

Email addresses: tyagim@ncbi.nlm.nih.gov , aurelie.bornot@univ-paris-diderot.fr , bernard.offmann@univ-reunion.fr , alexandre.debrevern@univ-paris-diderot.fr

[+] The first two authors contributed equally to this article.

Abstract

Loops connect regular secondary structures. In many instances, they are known to play important biological roles. Analysis and prediction of loop conformations depend directly on the definition of repetitive structures. Nonetheless, the secondary structure assignment methods (SSAMs) often lead to divergent assignments.

In the present study, we analyzed, both structure and sequence point of views, how the divergence between different SSAMs affect boundary definitions of loops connecting regular secondary structures.

The analysis of SSAMs underlines that no clear consensus between the different SSAMs can be easily found. Since these latter greatly influence the loop boundary definitions, important variations are indeed observed, *i.e*. capping positions are shifted between different SSAMs. On the other hand, our results show that the sequence information in these capping regions are more stable than expected, and, classical and equivalent sequence patterns were found for most of the SSAMs.

This is, to our knowledge, the most exhaustive survey in this field as (i) various databank have been used leading to similar results without implication of protein redundancy and (ii) the first time various SSAMs have been used. This work hence gives new insights into the difficult question of assignment of repetitive structures and addresses the issue of loop boundaries definition. Though SSAMs give very different local structure assignments; capping sequence patterns remain efficiently stable.

Key-words: protein structures; biochemistry; amino acids; secondary structures; propensities.

Introduction

The knowledge of the three-dimensional (3D) structures of proteins contributes to understand their biological functions. Protein 3D structures are often described as a succession of repetitive secondary structures (mainly α-helices and β-sheets (Pauling and Corey 1951; Pauling et al. 1951)). This mono-dimensional description helps to simplify coarsely this 3D information. It can also be used to describe more complex local 3D motifs, *e.g.* the Greek key (Hutchinson and Thornton 1993), or even complete 3D structures in 2D views, *e.g.* HERA (Hutchinson and Thornton 1990) or TOPS (Michalopoulos et al. 2004).

Numerous approaches exist to assign secondary structure and rely on various descriptors (see Table 1).

A first class of methods is based solely on H-bond patterns. In this category, DSSP (Kabsch and Sander 1983) remains the most popular Secondary Structure Assignment Methods (SSAMs). It identifies the secondary structures by particular hydrogen bond patterns detected from the protein geometry and an electrostatic model. DSSP is the basis of the assignment done by the Protein DataBank (PDB) (Bernstein et al. 1977; Berman et al. 2000). A recent version of DSSP called DSSPcont was proposed by Rost (Andersen et al. 2002). SECSTR is also an evolution of DSSP method dedicated to improved π-helices detection (Fodje and Al-Karadaghi 2002).

A second class of SSAMs add dihedral angle properties to H-bond patterns. In this category, STRIDE, developed in 1995, is the second widely used SSAM (Frishman and Argos 1995). PROMOTIF derives also from the DSSP approach, namely the software SSTRUC (Smith 1989), but focus on the characterization of γ- and β-turns, β-hairpins and β-bulges (Hutchinson and Thornton 1996).

The third class of secondary structure assignment methods relies on distances between residues inside protein structures. Additionally, this criterion has also been extended by taking into account angles. The DEFINE method (Richards and Kundrot 1988), like the Levitt's and Greer's method (Levitt and Greer 1977), 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 new assignment method of assignation using the inter-C_α distances and dihedral angles criteria (Martin et al. 2005). PSEA assigns the repetitive secondary structures from the sole C_α position using distance and angles criteria (Labesse et al. 1997). XTLSSTR uses all the backbone atoms to compute two angles and three distances (King and Johnson 1999).

Fourth, some SSAMs are defined solely on angles. PROSS is based only on the computation of Φ and Ψ dihedral angles. The Ramachandran map is divided into mesh of 30° or 60° and the secondary structures are assigned in regards to their successions of encoded mesh (Srinivasan and Rose 1999). SEGNO uses also the Φ and Ψ dihedral angles coupled with other angles to assign the secondary structures (Cubellis et al. 2005).

Fifth, VoTap (Voronoï Tessellation Assignment Procedure) is a geometrical tool that associates with each amino acid a Voronoï polyhedron (Dupuis et al. 2005), the faces of which define contacts between residues (Dupuis et al. 2004). 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 (Taylor et al. 2005).

A sixth category of SSAM relies on geometrical definitions and Cα coordinates. PCURVE is based on the helical parameters of each peptide unit, generates a global

peptide axis and makes use of an extended least-squares minimization procedure to yield the optimal helical description (Sklenar et al. 1989). PALSSE delineates secondary structure elements from protein Cα coordinates, and specifically addresses the requirements of vector-based protein similarity searches (Majumdar et al. 2005); this approach leads to surprising assignment where a residue can be associated to a α-helix and also to a β-strand. Very recently, PROSIGN proposed a different approach based solely on Cα coordinates (Hosseini et al. 2008). Hosseini and co-workers introduce four certain relations between Cα three-dimensional coordinates of consecutive residues, their method gives interesting information about helix geometry.

Finally, some SSAMs like Beta Spider could be considered more as hybrid or consensus methods. For instance, Beta Spider focuses only on β-sheet (the α-helix assignment is performed by DSSP) by considering all the stabilizing forces involved in the β-sheet phenomenon (Parisien and Major 2005).

As a consequence, these different assignment methods have generated specific weaknesses. For example, DSSP can generate very long helices that can be classified as linear, curved or kinked (Kumar and Bansal 1996; 1998; Bansal et al. 2000). This was one of the motivations of KAKSI methodology to define linear helices instead of long kinked helices (Martin et al. 2005). Moreover, the disagreement between the different SSAMs is not negligible, leading to only 80% of agreement between two distinct methods (Woodcock et al. 1992; Colloc'h et al. 1993; Fourrier et al. 2004; Martin et al. 2005). Consensus methods have been proposed using (i) DEFINE, P-CURVE and DSSP (Colloc'h et al. 1993) and (ii) more recently, P-SEA, KAKSI, SECSTR and STRIDE (Zhang et al. 2007), to diminish such features.

The coil state is in fact composed of really distinct local folds (Richardson

1981; Fitzkee et al. 2005a; Fitzkee et al. 2005b; Offmann et al. 2007), such as turns (Rose and Seltzer 1977; Rose et al. 1985; Hutchinson and Thornton 1994; 1996; Fuchs and Alix 2005; Bornot and de Brevern 2006; Street et al. 2007). Several studies have attempted to analyze conformation of loops linking specific secondary structures forming distinct subsets (Edwards et al. 1987; Thornton et al. 1988; Ring et al. 1992; Wintjens et al. 1996; Boutonnet et al. 1998; Wintjens et al. 1998; Efimov 2008). They are biologically essential regions (Espadaler et al. 2006), *e.g*. loops of protein kinases (Rekha and Srinivasan 2003; Fernandez-Fuentes et al. 2004). They are also used to analyze protein homology (Srinivasan et al. 1996; Panchenko and Madej 2004; 2005; Panchenko et al. 2005; Madej et al. 2007; Wolf et al. 2007), *e.g*. for structure-based phylogenetic study (Jiang and Blouin 2007). Due to their flexible nature they raise crucial questions in protein docking approaches (Huang et al. 2007; Nabuurs et al. 2007; Wong and Jacobson 2007), to predict protein loop conformations (Lessel and Schomburg 1999; Miyazaki et al. 2002; Wohlfahrt et al. 2002; Rohl et al. 2004; Boomsma and Hamelryck 2005; Monnigmann and Floudas 2005; Fernandez-Fuentes and Fiser 2006; Fernandez-Fuentes et al. 2006a; Fernandez-Fuentes et al. 2006b; Zhu et al. 2006; Kanagasabai et al. 2007; Olson et al. 2007; Prasad et al. 2007; Soto et al. 2007), to enhance protein thermostability (Reetz et al. 2006), to design proteins (Hu et al. 2007) or to obtain protein structures (Rapp et al. 2007). According to the repetitive secondary structures of their extremities, connecting loops are of 4 distinct classes (α-α, α-β, β-α and β-β) (Thornton et al. 1988; Efimov 1991b; a; Rufino et al. 1997). The research on loops has always been limited by the number of available loops in protein structures from the Protein DataBank (PDB (Bernstein et al. 1977; Berman et al. 2000)), so most of the works focus on loops of less than 9 residues (Wojcik et al. 1999; Michalsky et al. 2003).

Analyses have shown that capping regions of repetitive structures have specific amino acid compositions. George Rose analysis of helix signals in proteins highlighted the hydrophobic capping (Presta and Rose 1988), an hydrophobic interaction that straddles the helix terminus is always associated with hydrogen-bonded capping. From a global survey of protein structures, they identified seven distinct capping motifs, three at the helix N-terminus and four at the C-terminus (Aurora and Rose 1998). Recently, Kruus and coworkers have studied helix-cap sequence motifs. Their study is based on a very innovative approach. Indeed, they firstly assigned the helix of well-determined protein structures. Then, they searched for the sequence motifs corresponding at best to the capping regions. This search is based on Gibbs sampling method. They showed an important number of frameshifts of ±1 amino acid residue (Kruus et al. 2005). To date, no similar properties have been reported directly on β-strands.

In this paper, we focus on the analysis of loop boundaries *i.e.* capping regions of repetitive structures. We analyzed the disagreement between SSAMs for the definition of these capping regions and evaluated if the structural disagreement is associated with clear frameshift at the sequence level.

Results

Protein databanks

The constitution of the protein dataset is always crucial for protein structure analysis and prediction. In the case of loop predictions, another major problem is the right choice of the sequence similarity cut-off used to construct training datasets. Indeed, a 30% sequence identity non-redundant dataset corresponds to 10 - 20% sequence identity in coil regions. Thus, we have used different cut-off criteria ranging

from 20 to 90% and constructed 10 different datasets (see Supplementary material 1) to sample different sequence identity rates and analyze the influence of sequence identity on capping regions. Crystallographic structures in these datasets were selected at two resolution levels: 3 datasets were filtered for high resolution quality (resolution better than 1.6 Å) and 7 were filtered for good resolution quality (resolution better than 2.5 Å). The datasets have been extracted from PISCES database (Wang and Dunbrack 2003; 2005).

Table 2 summarizes, for each of the 10 datasets in our study, the secondary structure assignment done by different secondary structure assignment methods (SSAMs). The classical differences observed between (SSAMs) are found again (Fourrier et al. 2004), *i.e.* α-helices frequency ranges mainly between 28 and 34% and β-strand between 18 and 24%. Some SSAMs have particular behaviors like KAKSI (Martin et al. 2005) that is associated to a high β-strand frequency (~28%) or DEFINE (Richards and Kundrot 1988) with a low α-helix frequency (~24%). Nonetheless, for each SSAM, both mean frequency of secondary structures and length of repetitive structures remain surprisingly highly comparable for all the datasets; neither number of residues, nor sequence identity rate, nor resolution quality had an effect on the secondary structure features. In the following, the presented results will concern DB0 except when noted.

Figure 1 shows an example of *Hhai Methyltransferase* (Sheikhnejad et al. 1999) assigned by different SSAMs, it highlights visually how the differences can be important (see also Supplementary material 2). In the same way, the computation of C_3, *i.e.* the agreement rates between SSAMs (see Methods section), gives also similar results to previous works (Fourrier et al. 2004; de Brevern 2005; Martin et al. 2005) (see Figure 2). Briefly, SSAMs based on hydrogen bond assignments (DSSP,

STRIDE and SECSTR) produced nearly identical assignments, with C_3 more than to 90%. Otherwise, a mean C_3 of 80% was observed, with SEGNO displaying a closer C_3 value to hydrogen bond assignments than the others. DEFINE remains very different from the other methods with C_3 values close to 60%. Comparison of all theses SSAMs clearly highlights the intricacy of obtaining a simple consensus between all the methods.

Analyses of the structural agreement between the capping regions of repetitive secondary structures

These results highlight the difficulties to define an appropriate length for α-helices, β-strands and coils and locating their extremities (Presta and Rose 1988; Doig and Baldwin 1995; Aurora and Rose 1998; Mandel-Gutfreund et al. 2001; Mandel-Gutfreund and Gregoret 2002; Bang et al. 2006; Rose 2006). Inaccuracies in defining the repetitive structures have direct repercussions on the definition of loops. Thus, we have analyzed the positions of capping positions of repetitive structures as assigned by DSSP and systematically looked for their counterparts in assignments performed by another SSAM (only long repetitive structures of more than 6 residues have been used). Figure 3 shows some examples of this systematic comparison (see Supplementary material 3 for all the examples). Each figure compares a SSAM with DSSP. On the x-axis are given the positions of the N- and C-caps of α-helices (top) and β-strands (bottom) obtained by each method with respect to reference DSSP assignments (labeled "N-cap" or "C-cap" on this x-axis). On the y-axis are given the corresponding observed frequencies. For instance, C-cap position of α-helix assigned by DSSP corresponds to 43% of C_1, 42% of C_{cap} and 4% of C_1' positions assigned by STRIDE (see Figure 3, pattern 2). Five characteristic patterns could be identified:

- *pattern 1*, the capping position of the SSAM is the same than DSSP (in red),
- *pattern 2*, same capping position as DSSP and an adjacent positions are found,
- *pattern 3*, No preferred capping positions could be identified, they are distributed over the whole window range,
- *pattern 4*, it is another position that is considered preferably as the capping residue by the other SSAM,
- *pattern 5*, due to the definition of repetitive structures, the capping position is not within the range -4 to +4 around the capping position of DSSP.

Using the above categorization scheme, we can conveniently classify assignment methods based on how their capping positions differ from DSSP (see Supplementary material 4). It can also be used to show how well the four different capping regions are resolved. Hence, α-helix N cap displays four patterns 1, whereas β-sheet N cap displays only two patterns 1, but also two patterns 3 and two patterns 5, *i.e.*, the capping regions of β-sheet are more variably described than those of α-helix for which the correspondence between SSAMs is quite easily found. For the C caps, it goes to a higher level of complexity. Thus, α-helix C cap has only one pattern 2, two patterns 3 and three patterns 4, while the β-sheet C cap is characterized by four patterns 4, *i.e.* the correspondence between SSAMs are quite complex. Surprisingly, even the SSAM related to DSSP are not strictly equivalent to it, *e.g.* β-sheet N cap of STRIDE and SECSTR are shifted by (-1) residue. These results highlight greatly the difficulties to assign the β-strand extremities, while α-helix is in comparison more "conserved". Previous works done using other SSAMs as standard gave similar results.

Amino acid distributions in capping regions

Table 3 shows the over – and under – representation of amino acid of the different SSAMs in terms of Z-scores (de Brevern et al. 2000). Thus, at each position of each SSAM is given the important amino acids. KLd (Kullback and Leibler 1951) values were also computed to locate the most informative positions (see Supplementary material 5). For the following paragraphs, we use a notation $(x / y)^{pz}$ that corresponds to the amino acid (x) over - and (y) under – represented at the position z. N capping regions of α-helices (Table 3a) show a strong pattern (PSTND / IVLMAFYQERK)^{p1} (PE / GN)^{p2} (AQDE / IGN)^{p3} where p1 corresponds to position N_1' for DSSP, STRIDE, SECSTR, PSEA and SEGNO and N_{cap} for XTLSSTR and KAKSI. This position p1 is associated to a high KLd value.

At the opposite, KLd values of C capping regions of α-helices are weaker; multiple positions are in the same range of values. Repeated patterns (LAERK / IVPG) are found before p1, then (GN / IV)^{p1}, (PG / -)^{p2} and (PK / -)^{p3} where p1 corresponds to position C_1' for DSSP, STRIDE, SECSTR, XTLSSTR, PSEA and SEGNO and C_{cap} for DEFINE and KAKSI. It is noteworthy that the different positions, even if they are related, cannot be interchanged. The pattern of over-represented amino acids ([LAERK], [LAERK], [LAERK], [GN], [PG], [PK]) can correspond for instance, to the sequence (L A L N P K). The succession LAL of C_2, C_1, C_{cap} cannot be shifted as they are mainly under-represented at positions C_1', C_2' and C_3'.

N capping regions of β-strand (Table 3b) are more informative than C capping regions, they are characterized by a strong succession of patterns (PGND / IVL), followed by a pattern (IVFYT / APND)^{p1} followed by compatible patterns (IVFY / AQPGNDERK); this latter corresponding to the β-strand; position p1 correspond to

N_{cap} for DSSP, STRIDE, SECSTR and KAKSI, and to N_1 for PSEA, DEFINE and SEGNO.

C capping regions of β-strand are less informative, but are also clearly cut into two successive patterns, the first is the one characteristic of β-strand (IVLFYN / AQPGNDERK) followed by (GND / IVLAF)^{p1}. The final position of p1 is harder to define than previously, but correspond most of the time to C_1' that is also the less informative position in terms of KLd. Analysis of the position informativity with KLd values, emphases the results seen on Table 3b. Positions C_2, C_1, and C_2' have a strong amino acid distribution associated with high KLd values, whereas the boundary region, *i.e.* C_{cap} and C_1', have fewer amino acids over and under-represented and low KLd values.

Finally, every amino acid distributions of DSSP capping regions with the other SSAMs have been compared (see Supplementary material 6). N capping α-helix regions of DSSP is strictly equivalent to SECSTR, STRIDE, PSEA and SEGNO. A light difference at N_2' position (associated to a low informative position) is found between DSSP and DEFINE and a clear frameshift from N_{cap} of DSSP to N_1 for XTLSSTR and KAKSI.

For the C capping regions of α-helix, the situation is more complex, the only strict equivalent amino acid matrices is find between DSSP and SECSTR. A limited divergence is found at position C_1 for PSEA and at C_2' for XTLSSTR. Surprisingly, STRIDE has only three strict corresponding positions with DSSP, but it remains highly comparable as C_2 and C_1 positions have very close amino acid distributions as C_2' and C_3'. Concerning KAKSI, we observe a shift of (+1) for the positions ranging from C_2 to Cc_{ap}. For SEGNO, only the central positions are equivalent to DSSP. C_2 and C_1 positions of SEGNO correspond to C_1 and C_{cap} positions of DSSP, but all

these amino acid distributions are very close. Only position $C_{3'}$ of SEGNO is particular due to an over- representation of Glycine not found in any other SSAM and thus more related to $C_{2'}$ position of DSSP than $C_{3'}$ position.

Contrary to the α-helix, the β-strand capping regions show few strong amino acid distribution divergences as the α-helix. Thus, we find that SECSTR, STRIDE and DEFINE are equivalent to DSSP N capping region of β-strand. For the others, only the clear cut between [$N_{3'}$ - $N_{1'}$] and [N_{cap} - N_2] positions of DSSP are found. For instance, [$N_{3'}$ - $N_{1'}$] of XTLSSTR correspond to $N_{2'}$ position of DSSP.

For the C capping regions of β-strand, SECSTR, STRIDE and KAKSI are equivalent to DSSP. For XTLSSTR and SEGNO only their $C_1^{'}$ positions is not equivalent to $C_1^{'}$ of DSSP. PSEA adds to this, a shift of positions C_1 and C_{cap}; it is mainly due to lower informativity at these positions.

Discussion

Analysis of different SSAMs based on diverse structural protein databanks gave results that are in line with previous studies including our own (Colloc'h et al. 1993; Fourrier et al. 2004; de Brevern et al. 2005a; Martin et al. 2005; Zhang et al. 2007). Indeed, each SSAM - based on different criteria- gives a different assignment. Thus no simple consensus of secondary structure assignments could be done. Repetition of over- and under-represented amino acids are found as expected within the regular secondary structures, i.e. positions N_{cap}, N_1, N_2 and positions C_2, C_1, C_{cap} (de Brevern et al. 2000). Analysis of position of N and C cap of DSSP in regards to capping positions given by other SSAMs lead to a similar view. Even the SSAM closely related to DSSP could have systematically a very different N or C cap position.

Amino acid distributions surprisingly do not reflect this fact: A structural

frameshift does not imply a "sequence" frameshift. α-helix capping regions possess a true amino acid patterns (see Table 3), the classical over – and under – representations of amino acids are found again. For the N cap α-helix, we observe a clear frameshift of (+1) for KAKSI & XTLSSTR assignment method and for the C cap α-helix, we observe a clear frameshift of (-1) for KAKSI. Thus, the sequence informativity characterizing 'the' α-helix capping regions is found for all the SSAMs with some slight sliding. Only DEFINE assignment does not correspond. However, its KLd values are 20 to 50 times less informative than other SSAMs. For the β-strands capping regions as classically noted, a simple differentiation exists between the central regions mainly composed of aliphatic hydrophobic residues and "outside" regions with polar and "breakers". This very simple rule is found for all the SSAMs.

The capping regions are the most important differences between SSAMs, but they do not create different amino acid patterns, only minor shift, *e.g.* DSSP and KAKSI helices. These results are in agreement with the results of Kruus and co-workers (Kruus et al. 2005) that elegantly analyze the question of capping regions of α-helices. They have shown that strong patterns are found in these regions, but on the structure, even if does not correspond perfectly, they shift often in a very close vicinity. We observe the same kind of results, but in our case, the average created by the use of one occurrence matrix each time gives a global view of the amino acid patterns.

We have also analyzed the repetitive structures assigned by our structural alphabet (Offmann et al. 2007), namely the Protein Blocks (de Brevern et al. 2000; de Brevern et al. 2001; de Brevern et al. 2002; de Brevern and Hazout 2003; de Brevern et al. 2004; de Brevern 2005; de Brevern et al. 2005a; de Brevern et al. 2005b; Benros et al. 2006; de Brevern et al. 2007; Etchebest et al. 2007; Benros et al. 2009; Bornot et

al. 2009; Faure et al. 2009). Their results are a bit different from the SSAMs, *e.g.* N_{cap} and C_{cap} have always lower KLd values than other positions. Contrary to the SSAMs, they approximate even the non-repetitive states, *i.e.* loops, so they can be used to predict them from the knowledge of sequence.

Secondary structure assignment is too often considered as a finished research field with only one golden standard DSSP. As noted by Arthur M. Lesk (Lesk 2005), "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 [authors of SSAMs] cannot be blamed) that there is **A RIGHT ANSWER**. Provided that the danger is recognized, such programs can be useful". SSAMs lead to different assignments, and, to different analysis of protein structures.

Robson and Garnier have written: "In looking at a model of a protein, it is often easy to recognize helix and to a lesser extent sheet strands, but it is not easy to say whether the residues at the ends of these features be included in them or not (Robson and Garnier 1986.)." Indeed, the discrepancies are often found at the extremities of repetitive structures and loop boundaries are essential in loop conformation prediction (Lessel and Schomburg 1999). Nonetheless, we have shown here that systematically differences do not appear in terms of sequence. This result reinforce the results of Kruus and co-workers (Kruus et al. 2005). This study is also related to the elegant research done by Zhang and co-workers (Zhang et al. 2007). They have proposed to assess secondary structure assignment using recognized pairwise sequence-alignment benchmarks. They have so highlighted the interest of two assignment methods and also underline the repetitive structure extremities. Here, we went further and quantified the discrepancies in terms of amino acid propensities in a very systematic

way using various SSAMs. We showed that, though SSAMs give different local structure assignments, capping sequence patterns remain in fact surprisingly stable. In someway, it emphasised the idea of Grishin with PALSSE, that focus on the sequence property as on the structure properties to assign the repetitive structure (Majumdar et al. 2005).

Moreover, the definition of assignment of secondary structure has a direct impact on the quality of the prediction. Cuff and Barton have used three different SSAMs (DSSP, STRIDE and DEFINE) and combined their *assignments* to improve secondary structure *prediction* rate (using assignment done by DSSP as reference) (Cuff and Barton 1999). Recently, Zhang and co-workers showed that the consensus of STRIDE, KAKSI, SECSTR, and P-SEA improves assignments over the best single method in each benchmark by an additional 1% (Zhang et al. 2008). Our analysis underlines that the amino acid contents of capping regions is encompassed by numerous various SSAMs. Thus, the amino acid contents of capping regions could help to define more precisely the assignments by helping to find a consensus between divergent assignment methods. Thus, this new consensus SSAM encompassing different SSAMs and amino acid behaviors would help the prediction.

In the same way, Dovidchenko and co-workers showed that loop boundary prediction methods relying on sequence specificities seem to be more efficient that methods based on physical properties of amino-acids (Dovidchenko et al. 2008). Actually, the PSIPRED prediction method (based on assignment performed by DSSP) achieved 73 % correct prediction rates from the single sequence that is between 7 and 9% better than physics based methods. Thus, protein sequence conservation is critical for predicting loop boundaries. Our contribution is substantial in the sense that equivalent sequence patterns were found for most of the SSAMs.

Thus prediction from these patterns could provide a unified decision of loops boundaries. Furthermore, this pattern stability, despite of assignment shifts, enlightens an interesting property of protein sequences that allow some fuzziness at loop boundaries. This phenomenon might physically support the conformational adaptations of proteins for function or for stability in variable cell environments.

Methods

Data sets

The 10 sets of proteins are based on the PISCES database (Wang and Dunbrack 2003; 2005) and represents between 162,830 and 1,572,412 residues. They are available at http://www.dsimb.inserm.fr/~debrevern/DOWN/DB/new. The sets are defined as containing no more than *x*% pairwise sequence identity with *x* ranging from 20 to 90%. The selected chains have X-ray crystallographic resolutions less than 1.6 Å with an R-factor less than 0.25 or less than 2.5 Å with an R-factor less than 1.0. Each chain was carefully examined with geometric criteria to avoid bias from zones with missing density. Table 2 presents all the details of these databanks.

Secondary structure assignments

They have been done with five distinct software: DSSP (Kabsch and Sander 1983) (CMBI version 2000), STRIDE (Frishman and Argos 1995), SECSTR (Fodje and Al-Karadaghi 2002) (version 0.2.3-1), XTLSSTR (King and Johnson 1999), PSEA [ref] (version 2.0), DEFINE (Richards and Kundrot 1988) (version 2.0), KAKSI (Martin et al. 2005) (version 1.0.1) and SEGNO (Cubellis et al. 2005) (version 3.1). PBs (de Brevern 2005) have been assigned using in-house software (available at http://www.dsimb.inserm.fr/DOWN/LECT/), it follows similar rules to

assignment done by PBE web server (http://bioinformatics.univ-reunion.fr/PBE/) (Tyagi et al. 2006). DSSP, STRIDE, SECSTR, XTLSSTR and SEGNO give more than three states, so we have reduced them: the α-helix contains α, 3_{10} and π - helices, the β-strand contains only the β-sheet and the coil everything else (β-bridges, turns, bends, polyproline II and coil). Default parameters are used for each software. The first residue of a repetitive structures is noted N_{cap} and the following N_n (n =1 to 3 in this study), while the previous residues are noted N'_n (n = 1 is so the closest residue to N_{cap} position). In the same way, the last residue of repetitive structure is noted C_{cap} and the following C'_n, while the previous residues are noted C_n. The N_n and C_n residues are so inside the repetitive structures, N'_n and C'_n residues belongs to coil regions.

Agreement rate

To compare two distinct secondary structure assignment methods, we used an agreement rate which is the proportion of residues associated with the same state (α-helix, β-strand and coil). It is noted C_3 (Fourrier et al. 2004).

To compare capping regions of repetitive secondary structures, we have taken as standard the capping regions of repetitive secondary structures defined by DSSP. Then, we simply search the positions corresponding to N and C cap defined by DSSP with other assignments. In the same way, we have compared the amino acid distribution of capping regions of repetitive secondary structures defined by DSSP with the amino acid distribution of capping regions of repetitive secondary structures defined by other SSAMs.

Z-score

The amino acid occurrences for each secondary structure have been normalized into a Z-score:

$$Z(\ _{,j}) = \frac{n_{i,j}^{obs} - n_{i,j}^{th}}{\sqrt{n_{i,j}^{th}}}$$

with $n_{i,j}^{obs}$ the observed occurrence number of amino acid *i* in position *j* for a given secondary structure and n_{ij}^{th} the expected number. The product of the occurrences in position *j* with the frequency of amino acid *i* in the entire databank equals $n_{i,j}^{th}$. Positive Z-scores (respectively negative) correspond to overrepresented amino acids (respectively underrepresented); threshold values of 4.42 and 1.96 were chosen (probability less than 10^{-5} and 5.10^{-2} respectively).

Asymmetric Kullback-Leibler measure

The Kullback-Leibler measure or relative entropy (Kullback and Leibler 1951), denoted by *KLd,* evaluates the contrast between two amino acid distributions, *i.e.* the amino acid distribution observed in a given position *j* and the reference amino acid distribution in the protein set (DB). The relative entropy $KLd(j|S_x)$ in the site *j* for the secondary structure S_x is expressed as :

$$KLd(j|S_x) = \sum_{i=1}^{i=20} P(aa_j = i|S_x) \ln\left(\frac{P(aa_j = i|S_x)}{P(aa_j = i|S\_)}\right)$$

where $P(aa_j = i|S_x)$ is the probability of observing the amino acid *i* in position *j* (*j* = -*w*, ...,0, ..., +*w*) of the sequence window (15 residue long, *w*=7) given a secondary structure S_x, and, $P(aa_j = i|DB)$ the probability of observing the same amino acid in

the databank (named DB). Thus, it allows one to detect the "informative" positions in terms of amino acids for a given secondary structure (de Brevern et al. 2000).

Acknowledgements

This work was supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, Université de Saint-Denis de la Réunion and the French Institute for Health and Medical Care (INSERM). MT had a grant from the Conseil Régional de La Réunion and European Union. AB has a grant from Ministère de la Recherche.

Figures

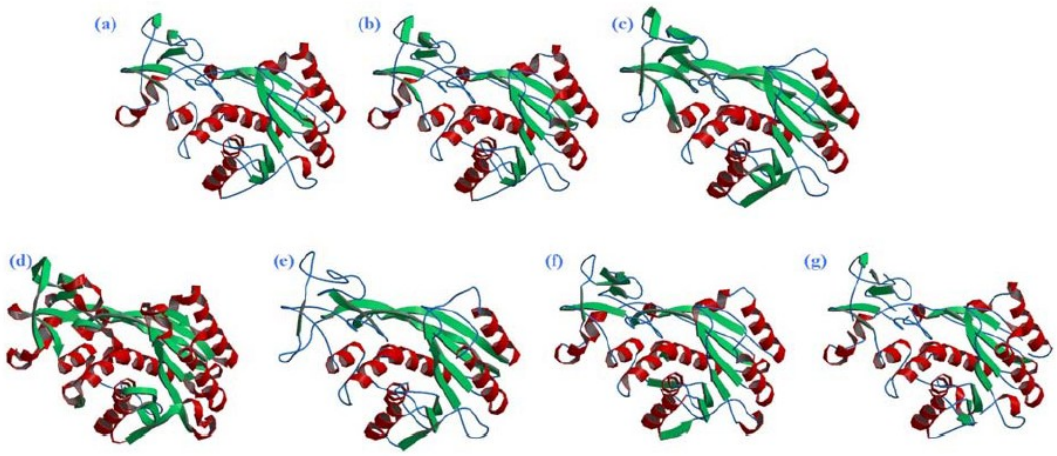

Figure 1 - SSAMs of Hhai Methyltransferase.
Example of secondary structure assignments for the *Hhai Methyltransferase* (PDB code :10MH (Sheikhnejad et al. 1999)) with (a) DSSP, (b) STRIDE, (c) PSEA, (d) DEFINE, (e) PCURVE, (f) XTLSSTR and (g) SECSTR. All the methods have been reduced to three states with the helical states in red ribbons, the extended state in green arrows and the coil in blue line.

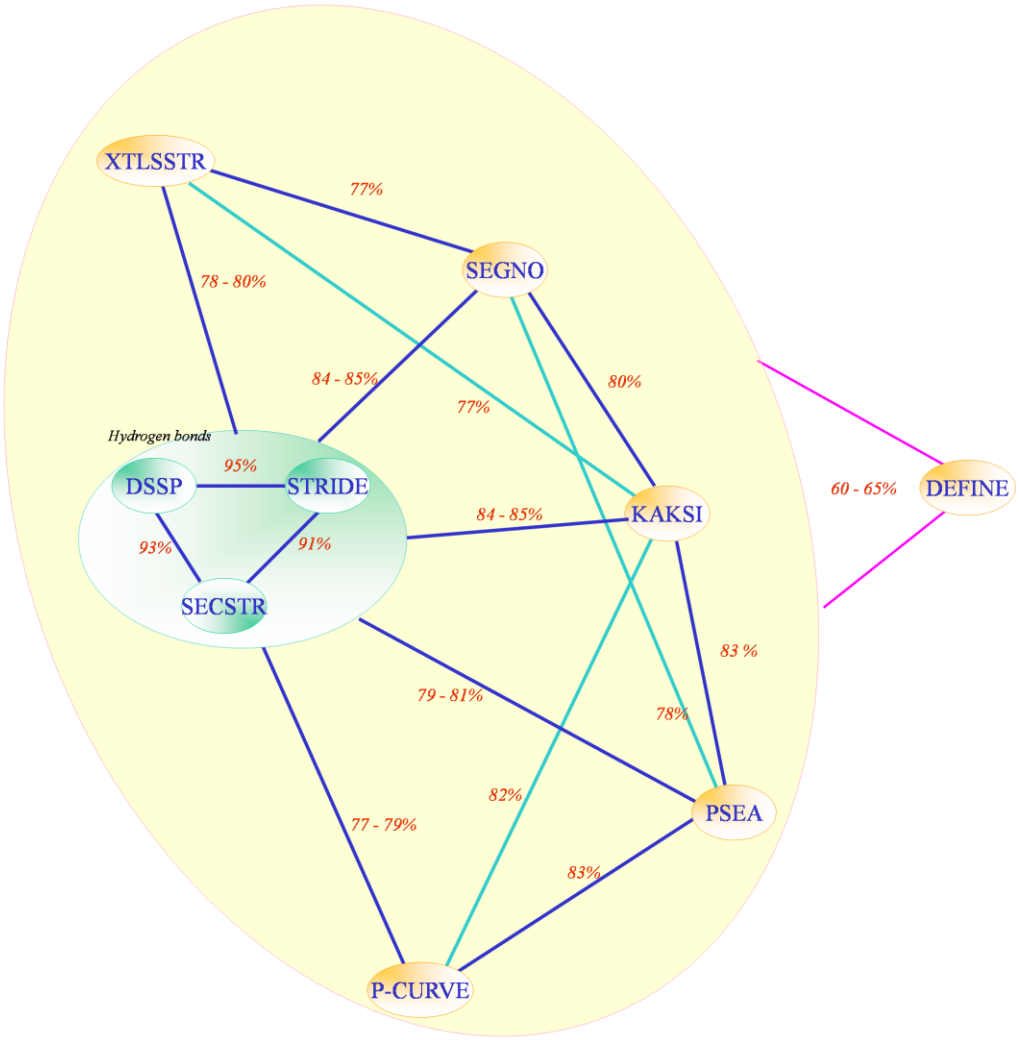
XTLSSTR
77%
SEGNO
78 - 80%
84 - 85%
80%
77%
Hydrogen bonds
95%
DSSP
STRIDE
60 - 65%
DEFINE
84 - 85%
KAKSI
93%
91%
SECSTR
83 %
79 - 81%
78%
82%
PSEA
77 - 79%
83%
P-CURVE

Figure 2 – C_3 values for different SSAMs (DB0 dataset).

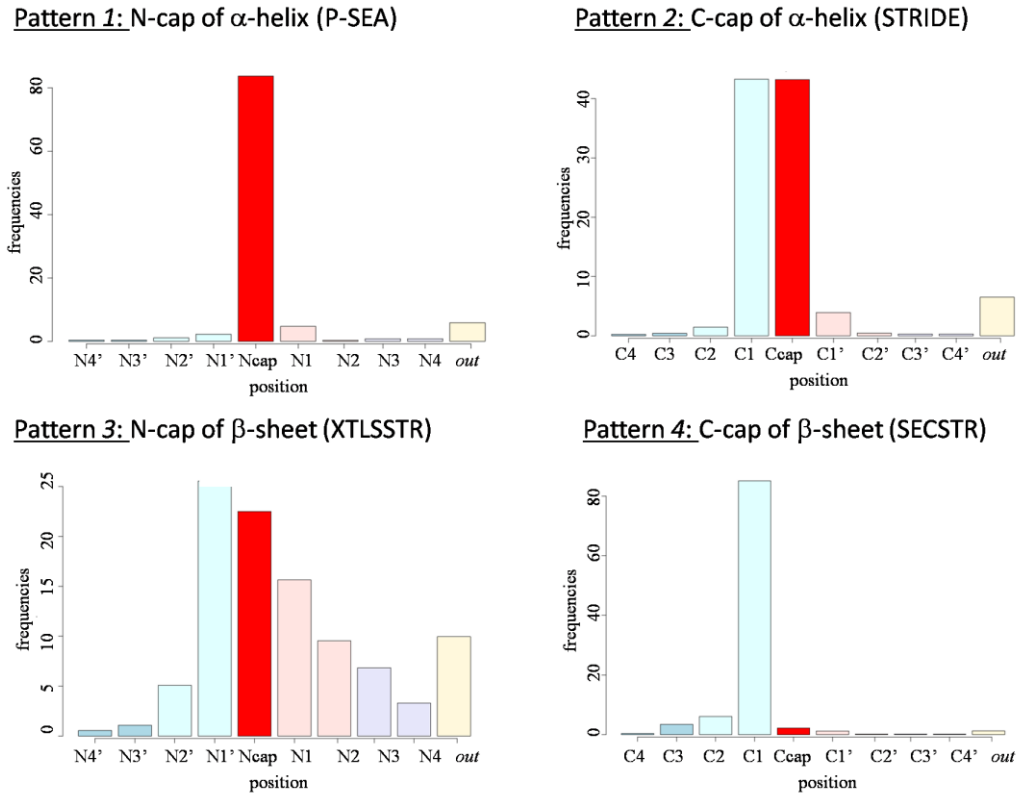

Figure 3 – Examples of discrepancies between N or C cap positions assigned by DSSP with other SSAMs. Examples of the four kinds of differences are shown. (x-axis): the position of the capping region, (y-axis) frequencies of N or C cap central positions of SSAMs according to DSSP. Central positions are in red colour.

Tables

Methods	year	Assignment based on
Greer & Levitt	1977	Distance
DSSP	1983	H-bond
DEFINE	1988	Distance
PCURVE	1989	Axis
SSTRUC	1989	H-bond
CONCENSUS	1993	Mean (DSSP, DEFINE and PCURVE)
STRIDE	1995	H-bond / dihedral
PROMOTIF	1996	H-bond / dihedral
PSEA	1997	Distance / angle
PROSS	1999	Dihedral
XTLSSTR	1999	Distance / angle
DSSPcont	2002	H-bond
SECSTR	2002	H-bond
VORO3D	2004	Voronoï
KAKSI	2005	Distance / dihedral
SEGNO	2005	angle / multiple
Beta-Spider	2005	β-sheet + DSSP for α-helix
PALSSE	2005	Cα (vector similarity)
Delaunay tessellation	2005	Delaunay
SKSP	2007	Mean (STRIDE, DSSP, SECSTR, KAKSI, P-SEA, and SEGNO)
PROSIGN:	2008	Cα deviation values

Table 1 – Secondary structure assignment methods.

(a)

		DB 0		DB 1		DB 2		DB 3		DB 4	
		freq	lg	freq	lg	freq	lg	freq	lg	freq	lg
DSSP	α	33.17	10.66	34.51	11.21	34.46	11.14	34.07	11.09	33.70	11.02
	β	21.52	5.30	21.60	5.44	21.64	5.42	21.85	5.41	21.86	5.39
	coil	45.3		43.88		43.91		44.08		44.44	
STRIDE	α	30.78	11.12	34.15	11.76	34.07	11.69	33.74	11.63	33.47	11.56
	β	19.7	5.34	20.89	5.47	21.10	5.45	21.38	5.44	21.39	5.42
	coil	49.51		44.96		44.83		44.88		45.14	
SECSTR	α	31.38	10.93	32.72	11.56	32.62	11.48	32.25	11.43	31.88	11.36
	β	20.32	4.98	20.22	5.11	20.29	5.10	20.48	5.09	20.57	5.07
	coil	48.28		47.06		47.10		47.27		47.56	
XTLSSTR	α	32.13	10.64	32.83	11.18	32.62	11.10	32.23	11.04	31.87	10.98
	β	19.57	4.91	19.05	5.02	19.14	5.01	19.34	5.00	19.38	4.99
	coil	48.3		48.12		48.24		48.44		48.75	
PSEA	α	34.04	10.78	35.56	11.30	35.48	11.23	35.09	11.17	34.68	11.11
	β	24.01	5.16	24.49	5.27	24.48	5.26	24.72	5.25	24.84	5.24
	coil	41.95		39.94		40.04		40.18		40.48	
DEFINE	α	28.35	10.95	25.60	11.42	26.25	11.36	26.38	11.30	26.12	11.24
	β	25.89	5.39	22.39	5.47	23.12	5.47	23.48	5.46	23.48	5.45
	coil	45.76		52.01		50.63		50.14		50.40	
KAKSI	α	29.66	11.12	27.36	11.57	28.25	11.51	28.45	11.45	28.83	11.40
	β	28.91	5.53	25.87	5.59	26.69	5.59	27.12	5.58	27.84	5.58
	coil	41.43		46.78		45.06		44.43		43.34	
SEGNO	α	30.17	10.99	31.64	11.43	31.71	11.37	31.32	11.31	30.92	11.27
	β	21.26	5.58	21.26	5.65	21.36	5.65	21.50	5.63	21.52	5.63
	coil	48.58		47.10		46.93		47.17		47.56	
PBs	α	31.39	10.65	33.02	11.11	32.84	11.05	32.45	10.99	32.05	10.94
	β	18.25	5.39	18.64	5.46	18.64	5.45	18.77	5.44	18.85	5.44
	coil	50.35		48.35		48.51		48.79		49.10	
Nb res		162 830		565 364		712 075		870 094		1 132 639	
nb chains		887		2 722		3 325		3 983		5 081	
pc		20		20		25		30		40	
res		1.6		2.5		2.5		2.5		2.5	
Rfactor		0.25		1.00		1.00		1.00		1.00	

(b)

		DB 5		DB 6		DB 7		DB 8		DB 9	
		freq	lg	freq	lg	freq	lg	freq	lg	freq	lg
DSSP	α	32.18	10.69	33.60	11.10	33.37	10.99	33.17	10.98	31.56	10.70
	β	21.77	5.31	22.03	5.45	21.76	5.37	21.90	5.37	22.18	5.31
	coil	46.05		44.37		44.87		44.93		46.25	
STRIDE	α	29.96	11.15	33.60	11.66	33.25	11.54	33.09	11.53	29.60	11.18
	β	19.88	5.34	21.57	5.47	21.37	5.40	21.53	5.40	20.34	5.34
	coil	50.16		44.83		45.38		45.38		50.06	
SECSTR	α	30.41	10.96	31.90	11.46	31.58	11.34	31.40	11.34	29.83	10.98
	β	20.73	4.99	20.67	5.13	20.53	5.06	20.67	5.06	21.15	4.99
	coil	48.86		47.43		47.89		47.93		49.02	
XTLSSTR	α	31.13	10.65	31.95	11.08	31.63	10.96	31.45	10.96	30.61	10.68
	β	19.83	4.92	19.48	5.04	19.32	4.97	19.44	4.97	20.21	4.93
	coil	49.05		48.57		49.05		49.10		49.18	
PSEA	α	32.96	10.80	34.47	11.22	34.30	11.10	34.11	11.09	32.41	10.83
	β	24.37	5.17	25.00	5.28	24.80	5.22	24.97	5.23	24.86	5.18
	coil	42.67		40.52		40.90		40.93		42.73	
DEFINE	α	28.02	10.95	26.70	11.34	26.52	11.22	26.41	11.22	26.91	10.97
	β	26.10	5.39	24.29	5.49	23.91	5.44	24.01	5.44	26.12	5.40
	coil	45.89		49.01		49.57		49.58		46.97	
KAKSI	α	29.45	11.14	29.14	11.49	28.84	11.38	28.66	11.38	27.98	11.14
	β	30.00	5.56	28.27	5.62	28.29	5.58	28.23	5.58	29.16	5.56
	coil	40.55		42.58		42.88		43.11		42.86	
SEGNO	α	29.41	11.00	31.34	11.36	30.61	11.24	30.43	11.24	28.24	11.00
	β	21.28	5.61	22.06	5.68	21.49	5.64	21.66	5.64	21.55	5.62
	coil	49.31		46.60		47.91		47.92		50.21	
PBs	α	30.62	10.65	32.04	11.02	31.71	10.91	31.54	10.91	30.08	10.65
	β	18.59	5.41	18.88	5.48	18.78	5.45	18.90	5.45	18.87	5.42
	coil	50.79		49.09		49.51		49.56		51.05	
Nb res		276 586		415 360		1 513 629		1 572 412		312 219	
nb chains		1 425		5 847		6 823		7 141		1 630	
pc		50		50		70		80		90	
res		1.6		2.5		2.5		2.5		1.6	
Rfactor		0.25		1.00		1.00		1.00		0.25	

Table 2 - 10 Protein databanks.

This table summarizes all the 10 protein databanks (noted from DB 0 to DB 9) used in this study. Each databank is analyzed using different SSAMs, are given the frequencies of secondary structure (*freq*) and average length of repetitive structures (*lg*), with the total number of amino acids (*NB res*), the number of protein chains (*nb chains*), the maximum percentage of sequence identity (pc), the resolution (*res*) and *Rfactor*.

Ncap alpha		N3'	N2'	N1'	Ncap	N1	N2
DSSP	(+)	G	M	PSTND	WPE	AQDE	QDE
STRIDE	(+)	PG	M	PGSTND	P	ADE	QDE
SECSTR	(+)	PG	MP	STND	PE	ADE	AQDE
XTLSSTR	(+)	G	P	STND	PSTND	APE	QDE
PSEA	(+)	G	MG	GSTND	PE	AQDE	AQDE
DEFINE	(+)		D	PSD	AE	E	AE
KAKSI	(+)	P	G	P	PSTND	APE	ADE
SEGNO	(+)	G	MP	GSTND	WPE	AQDE	AQDE
PBs	(+)	PSTND	PD	DE	QDE	ILAF	LAQERK
DSSP	(-)			IVLMAFYQERK	GN	IVLFG	PGN
STRIDE	(-)			IVLAFYWQERK	GN	IVLFG	PGN
SECSTR	(-)			IVLMAFYERK	GN	IVLFG	PGN
XTLSSTR	(-)			VAEK	IVLAF	VGTN	IPGN
PSEA	(-)			IVLAFQERK	GN	IVLG	PGN
DEFINE	(-)			IV	N		PG
KAKSI	(-)				IVLMAFK	GN	IVG
SEGNO	(-)			IVLMAFYQERK	GTN	IVLG	PGN
PBs	(-)	IVLAFQERK	IVL	IVLC	IP	PGTD	PGS

Ccap alpha		C2	C1	Ccap	C1'	C2'	C3'
DSSP	(+)	LAERK	LAERK	LAERK	GN	PG	PK
STRIDE	(+)	LAERK	AQERK	LAN	GN	P	PDK
SECSTR	(+)	LMA	AERK	LAQERK	LAGN	PGN	K
XTLSSTR	(+)	LAEK	AERK	LAE	GN	PK	K
PSEA	(+)	ILAERK	AQERK	LAHNRK	GN	PG	PD
DEFINE	(+)			G	P	P	V
KAKSI	(+)	LAQERK	LARK	GN	PG	PGD	P
SEGNO	(+)	ILA	LAERK	LAQERK	GHN	PHN	PGD
PBs	(+)	LA	LMAC	AQERK	QERK	LTN	PGN
DSSP	(-)	VPGT	PGSTD	IVPGD	IVWPTE	IVLMAFYE	V
STRIDE	(-)	PGTN	IVPGT	IVPGD	IVLAPTE		VL
SECSTR	(-)	VPGT	PGST	VPGD	IVPTD	IVLMFYWTE	
XTLSSTR	(-)	PG	PG	VPG	IVPT		V
PSEA	(-)	PGSTD	IVFPGT	IVPGD	IVLAE		A
DEFINE	(-)						
KAKSI	(-)	VPGT	VPGD	IVP		IVL	
SEGNO	(-)	PGSTND	VPGTD	IVFPGD	IVPT	IVF	IVLFYT
PBs	(-)	PGST	VPGTD	VGT	IVG	IVG	IVLMAFYWTE

Table 3 (a) – Amino acid over- and under –representation at capping regions.

The over (+) (respectively under (-)) - representation have been selected using a Z-score more than 4.4 (respectively less than -4.4). The first part of the table presents the N and C capping regions of α-helix. Results have been obtained with DB0.

Ncap beta		N3'	N2'	N1'	Ncap	N1	N2
DSSP	(+)	PGND	PGND	PGND	IVFYT	IVLFY	IVFY
STRIDE	(+)	PGSND	PGND	PGND	IVFYT	IVLFY	IVFY
SECSTR	(+)	PGN	PGND	PGND	IVFYWT	IVLFY	IVLFY
XTLSSTR	(+)	PG	PGNK	PGN	G	IVFYT	IVLFY
PSEA	(+)	GNK	PN	GND	VG	IVYPT	IVFY
DEFINE	(+)	G	G	G		IVP	V
KAKSI	(+)	PGNDK	PGND	GN	VPT	IVFY	IVLFY
SEGNO	(+)	GNK	PGN	GND	PG	IVFYT	IVLFYW
PBs	(+)	GN	GDK	VP	IVFYP	IVFY	IVFYPT
DSSP	(-)	IVLA	IVLMFWT	IVLAFE	APND	AQPGSNDEK	AQPGNDERK
STRIDE	(-)	IVLAF	IVLMFWT	IVLAE	APGND	APGSNDEK	QPGNDERK
SECSTR	(-)	LAF	VLMAFYW	IVLAFE	APNDE	AQPGSNDEK	AQPGNDEK
XTLSSTR	(-)	IL	ILAY	E	A	APGNDE	AQPGNDEK
PSEA	(-)	IVL		IVLAFYC	LPD	AGNDE	AQPGSNDEK
DEFINE	(-)					A	
KAKSI	(-)	IVLY	IVLMAFYW	LAF	E	APGNDE	AQPGSNDEK
SEGNO	(-)	IVL	LYW	IVLAFYW	LND	APGNDE	AQPGSNDEK
PBs	(-)	ILMAYE	LAP	LD	AGSNDE	AGSNDEK	AQGDERK

Ccap beta		C2	C1	Ccap	C1'	C2'	C3'
DSSP	(+)	IVLFYW	IVFYWC	IVFYD	GSND	PGSND	GSND
STRIDE	(+)	IVFYW	IVFYWC	IVYD	GND	PGSND	GSND
SECSTR	(+)	IVLFYW	IVFYWCT	IVFY	GND	PGSND	GSND
XTLSSTR	(+)	IVF	IVFYWT	IVFYTD	PGND	PGSD	PGSND
PSEA	(+)	IVLFY	IVFYCT	PSTD	PND	GSND	GSD
DEFINE	(+)		P	PSD	PD	GD	G
KAKSI	(+)	IVLFY	IVFYW	ND	PGND	PGSND	GSND
SEGNO	(+)	IVLFYW	IVFYWC	IVCTD	PGND	GSND	GSND
PBs	(+)	IVF	IVFY	IVFY	PSTND	P	GSND
DSSP	(-)	AQPGSNDERK	APGSNDEK	AGE	IFQER	IVLMAFY	IVLAF
STRIDE	(-)	AQPGSNDERK	APGNDEK	AGE	VFYQER	IVLMAFY	IVLAF
SECSTR	(-)	AQPSNDERK	APGNDE	AQGEK	AFYQEK	IVLMAFY	IVLAF
XTLSSTR	(-)	APNDE	AQPGNDEK	APGE	IVAQR	IVLF	IVLMAF
PSEA	(-)	AQPGNDEK	AQGNDEK	LAERK	IVLAFY	IVLAF	IVLAF
DEFINE	(-)			I	L	IV	
KAKSI	(-)	AQPGSNDEK	APGNDE	AEK	IVLF	IVLMAFY	IVLAF
SEGNO	(-)	AQPGSNDEK	APGSNDE	APGEK	IVLMAFYQR	IVLMFY	IVLAF
PBs	(-)	GNDERK	AGSNDE	AQGNDE	LAQERK	G	IVLMAFYP

Table 3(b) - Amino acid over- and under –representation at capping regions.

The over (+) (respectively under (-)) - representation have been selected using a Z-score more than 4.4 (respectively less than -4.4). The second part of the table presents the N and C capping regions of β-sheet. Results have been obtained with DB0.

Supplementary materials

Supplementary material 1 – *The non-redundant protein structure databanks*

Supplementary material 2 – *Multiple alignments of SSAMs.*
Example of multiple secondary structure assignments for the N-terminal extremity of the Methyltransferase protein with DSSP3 and STRID3 (DSSP and STRIDE reduced to 3 states), PSEA, DEFINE, PCURVE, a consensus method (cons. with a star when the 5 methods agree), the consensus defined by Colloc'h and co-workers, XTLSSTR, SECSTR, DSSP, STRIDE, HELANAL and the extended BETA alphabet (BETAEX).

Supplementary material 3 – *Discrepancies between N or C cap positions assigned by DSSP with other SSAMs (see Figure 3 for details)*.
From up to bottom are presented the discrepancies of STRIDE, SECSTR, XTLSSTR, PSEA, DEFINE, PCURVE, KASKI, SEGNO in regards to DSSP. From left to right are given N_{cap} and C_{cap} of α-helix and, then N_{cap} and C_{cap} of β-sheet.

Supplementary material 4 – *Grading of capping regions correspondence with DSSP assignment.*

Supplementary material 5 – *analysis of the most important positions in the capping regions.*

Supplementary material 6 – *Correspondence between amino acid distributions of capping regions*

References

Andersen, C.A., Palmer, A.G., Brunak, S., and Rost, B. 2002. Continuum secondary structure captures protein flexibility. *Structure (Camb)* **10:** 175-184.

Aurora, R., and Rose, G.D. 1998. Helix capping. *Protein Sci* **7:** 21-38.

Bang, D., Gribenko, A.V., Tereshko, V., Kossiakoff, A.A., Kent, S.B., and Makhatadze, G.I. 2006. Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis. *Nat Chem Biol* **2:** 139-143.

Bansal, M., Kumar, S., and Velavan, R. 2000. HELANAL: a program to characterize helix geometry in proteins. *J Biomol Struct Dyn* **17:** 811-819.

Benros, C., de Brevern, A.G., Etchebest, C., and Hazout, S. 2006. Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins* **62:** 865-880.

Benros, C., de Brevern, A.G., and Hazout, S. 2009. Analyzing the sequence-structure relationship of a library of local structural prototypes. *J Theor Biol.* **256:** 215-226.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112:** 535-542.

Boomsma, W., and Hamelryck, T. 2005. Full cyclic coordinate descent: solving the protein loop closure problem in Calpha space. *BMC Bioinformatics* **6:** 159.

Bornot, A., and de Brevern, A.G. 2006. Protein beta-turn assignments. *Bioinformation* **1:** 153-155.

Bornot, A., Etchebest, C., and de Brevern, A.G. 2009. A new prediction strategy for long local protein structures using an original description. *Proteins*.

Boutonnet, N.S., Kajava, A.V., and Rooman, M.J. 1998. Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins* **30:** 193-212.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng* **6:** 377-382.

Cubellis, M.V., Cailliez, F., and Lovell, S.C. 2005. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics* **6 Suppl 4:** S8.

Cuff, J.A., and Barton, G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34:** 508-519.

de Brevern, A.G. 2005. New assessment of a structural alphabet. *In Silico Biol* **5:** 283-289.

de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S., and Etchebest, C. 2004. Local backbone structure prediction of proteins. *In Silico Biol* **4:** 381-386.

de Brevern, A.G., Benros, C., and Hazout, S. 2005a. Structural Alphabet: From a Local Point of View to a Global Description of Protein 3D Structures. In *Bioinformatics: New Research* (ed. P.V. Yan), pp. 127-169. Nova Publishers.

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

de Brevern, A.G., Etchebest, C., Benros, C., and Hazout, S. 2007. "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence. *J Biosci* **32:** 51-70.

de Brevern, A.G., Etchebest, C., and Hazout, S. 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41:** 271-287.

de Brevern, A.G., and Hazout, S. 2003. 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics* **19:** 345-353.

de Brevern, A.G., Valadie, H., Hazout, S., and Etchebest, C. 2002. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci* **11:** 2871-2886.

de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C., and Etchebest, C. 2005b. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta* **1724:** 288-306.

Doig, A.J., and Baldwin, R.L. 1995. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci* **4:** 1325-1336.

Dovidchenko, N.V., Bogatyreva, N.S., and Galzitskaya, O.V. 2008. Prediction of loop regions in protein sequence. *J Bioinform Comput Biol* **6:** 1035-1047.

Dupuis, F., Sadoc, J.F., Jullien, R., Angelov, B., and Mornon, J.P. 2005. Voro3D: 3D Voronoi tessellations applied to protein structures. *Bioinformatics* **21:** 1715-1716.

Dupuis, F., Sadoc, J.F., and Mornon, J.P. 2004. Protein secondary structure assignment through Voronoi tessellation. *Proteins* **55:** 519-528.

Edwards, M.S., Sternberg, J.E., and Thornton, J.M. 1987. Structural and sequence patterns in the loops of beta alpha beta units. *Protein Eng* **1:** 173-181.

Efimov, A.V. 1991a. Structure of alpha-alpha-hairpins with short connections. *Protein Eng* **4:** 245-250.

Efimov, A.V. 1991b. Structure of coiled beta-beta-hairpins and beta-beta-corners. *FEBS Lett* **284:** 288-292.

Efimov, A.V. 2008. Structural Trees for Proteins Containing phi-Motifs. *Biochemistry (Mosc)* **73:** 23-28.

Espadaler, J., Querol, E., Aviles, F.X., and Oliva, B. 2006. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics* **22:** 2237-2243.

Etchebest, C., Benros, C., Bornot, A., Camproux, A.C., and de Brevern, A.G. 2007. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J* **36:** 1059-1069.

Faure, G., Bornot, A., and de Brevern, A.G. 2009. Analysis of protein contacts into Protein Units. *Biochimie*.

Fernandez-Fuentes, N., and Fiser, A. 2006. Saturating representation of loop conformational fragments in structure databanks. *BMC Struct Biol* **6:** 15.

Fernandez-Fuentes, N., Hermoso, A., Espadaler, J., Querol, E., Aviles, F.X., and Oliva, B. 2004. Classification of common functional loops of kinase super-families. *Proteins* **56:** 539-555.

Fernandez-Fuentes, N., Oliva, B., and Fiser, A. 2006a. A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res* **34:** 2085-2097.

Fernandez-Fuentes, N., Zhai, J., and Fiser, A. 2006b. ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res* **34:** W173-176.

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

Fitzkee, N.C., Fleming, P.J., and Rose, G.D. 2005b. The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* **58:** 852-854.

Fodje, M.N., and Al-Karadaghi, S. 2002. Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng* **15:** 353-358.

Fourrier, L., Benros, C., and de Brevern, A.G. 2004. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* **5:** 58.

Frishman, D., and Argos, P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23:** 566-579.

Fuchs, P.F., and Alix, A.J. 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins* **59:** 828-839.

Hosseini, S., Sadeghi, M., Pezeshk, H., Eslahchi, C., and Habibi, M. 2008. PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. *Comput Biol Chem.* **32:** 406-411.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. 2007. High-resolution design of a protein loop. *Proc Natl Acad Sci U S A* **104:** 17668-17673.

Huang, Z., Wong, C.F., and Wheeler, R.A. 2007. Flexible protein-flexible ligand docking with disrupted velocity simulated annealing. *Proteins*.

Hutchinson, E.G., and Thornton, J.M. 1990. HERA--a program to draw schematic diagrams of protein secondary structures. *Proteins* **8:** 203-212.

Hutchinson, E.G., and Thornton, J.M. 1993. The Greek key motif: extraction, classification and analysis. *Protein Eng* **6:** 233-245.

Hutchinson, E.G., and Thornton, J.M. 1994. A revised set of potentials for beta-turn formation in proteins. *Protein Sci* **3:** 2207-2216.

Hutchinson, E.G., and Thornton, J.M. 1996. PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* **5:** 212-220.

Jiang, H., and Blouin, C. 2007. Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics* **8:** 444.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kanagasabai, V., Arunachalam, J., Prasad, P.A., and Gautham, N. 2007. Exploring the conformational space of protein loops using a mean field technique with MOLS sampling. *Proteins* **67:** 908-921.

King, S.M., and Johnson, W.C. 1999. Assigning secondary structure from protein coordinate data. *Proteins* **35:** 313-320.

Kruus, E., Thumfort, P., Tang, C., and Wingreen, N.S. 2005. Gibbs sampling and helix-cap motifs. *Nucleic Acids Res* **33:** 5343-5353.

Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *Ann Math Stat* **22:** 79-86.
Kumar, S., and Bansal, M. 1996. Structural and sequence characteristics of long alpha helices in globular proteins. *Biophys J* **71:** 1574-1586.
Kumar, S., and Bansal, M. 1998. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J* **75:** 1935-1944.
Labesse, G., Colloc'h, N., Pothier, J., and Mornon, J.P. 1997. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput Appl Biosci* **13:** 291-295.
Lesk, A.M. 2005. *Introduction to Bioinformatics*. Oxford University Press, Oxford.
Lessel, U., and Schomburg, D. 1999. Importance of anchor group positioning in protein loop prediction. *Proteins* **37:** 56-64.
Levitt, M., and Greer, J. 1977. Automatic identification of secondary structure in globular proteins. *J Mol Biol* **114:** 181-239.
Madej, T., Panchenko, A.R., Chen, J., and Bryant, S.H. 2007. Protein homologous cores and loops: important clues to evolutionary relationships between structurally similar proteins. *BMC Struct Biol* **7:** 23.
Majumdar, I., Krishna, S.S., and Grishin, N.V. 2005. PALSSE: A program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics* **6:** 202.
Mandel-Gutfreund, Y., and Gregoret, L.M. 2002. On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J Mol Biol* **323:** 453-461.
Mandel-Gutfreund, Y., Zaremba, S.M., and Gregoret, L.M. 2001. Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J Mol Biol* **305:** 1145-1159.
Martin, J., Letellier, G., Marin, A., Taly, J.-F., de Brevern, A.G., and Gibrat, J.-F. 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Structural Biology* **5:** 17.
Michalopoulos, I., Torrance, G.M., Gilbert, D.R., and Westhead, D.R. 2004. TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res* **32:** D251-254.
Michalsky, E., Goede, A., and Preissner, R. 2003. Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. *Protein Eng* **16:** 979-985.
Miyazaki, S., Kuroda, Y., and Yokoyama, S. 2002. Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *J Struct Funct Genomics* **2:** 37-51.
Monnigmann, M., and Floudas, C.A. 2005. Protein loop structure prediction with flexible stem geometries. *Proteins* **61:** 748-762.
Nabuurs, S.B., Wagener, M., and de Vlieg, J. 2007. A Flexible Approach to Induced Fit Docking. *J Med Chem*.
Offmann, B., Tyagi, M., and de Brevern, A.G. 2007. Local Protein Structures. *Current Bioinformatics* **3:** 165-202.
Olson, M.A., Feig, M., and Brooks, C.L., 3rd. 2007. Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. *J Comput Chem*.
Panchenko, A.R., and Madej, T. 2004. Analysis of protein homology by assessing the (dis)similarity in protein loop regions. *Proteins* **57:** 539-547.

Panchenko, A.R., and Madej, T. 2005. Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol Biol* **5:** 10.

Panchenko, A.R., Wolf, Y.I., Panchenko, L.A., and Madej, T. 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins* **61:** 535-544.

Parisien, M., and Major, F. 2005. A new catalog of protein beta-sheets. *Proteins* **61:** 545-558.

Pauling, L., and Corey, R.B. 1951. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* **37:** 251-256.

Pauling, L., Corey, R.B., and Branson, H.R. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37:** 205-211.

Prasad, P.A., Kanagasabai, V., Arunachalam, J., and Gautham, N. 2007. Exploring conformational space using a mean field technique with MOLS sampling. *J Biosci* **32:** 909-920.

Presta, L.G., and Rose, G.D. 1988. Helix signals in proteins. *Science* **240:** 1632-1641.

Rapp, C.S., Strauss, T., Nederveen, A., and Fuentes, G. 2007. Prediction of protein loop geometries in solution. *Proteins* **69:** 69-74.

Reetz, M.T., Carballeira, J.D., and Vogel, A. 2006. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew Chem Int Ed Engl* **45:** 7745-7751.

Rekha, N., and Srinivasan, N. 2003. Structural basis of regulation and substrate specificity of protein kinase CK2 deduced from the modeling of protein-protein interactions. *BMC Struct Biol* **3:** 4.

Richards, F.M., and Kundrot, C.E. 1988. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* **3:** 71-84.

Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34:** 167-339.

Ring, C.S., Kneller, D.G., Langridge, R., and Cohen, F.E. 1992. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol* **224:** 685-699.

Robson, B., and Garnier, J. 1986. *Introduction to Proteins and Protein Engineering*. Elsevier Press; , Amsterdam:.

Rohl, C.A., Strauss, C.E., Chivian, D., and Baker, D. 2004. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55:** 656-677.

Rose, G.D. 2006. Lifting the lid on helix-capping. *Nat Chem Biol* **2:** 123-124.

Rose, G.D., Gierasch, L.M., and Smith, J.A. 1985. Turns in peptides and proteins. *Adv Protein Chem* **37:** 1-109.

Rose, G.D., and Seltzer, J.P. 1977. A new algorithm for finding the peptide chain turns in a globular protein. *J Mol Biol* **113:** 153-164.

Rufino, S.D., Donate, L.E., Canard, L.H., and Blundell, T.L. 1997. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J Mol Biol* **267:** 352-367.

Sheikhnejad, G., Brank, A., Christman, J.K., Goddard, A., Alvarez, E., Ford, H., Jr., Marquez, V.E., Marasco, C.J., Sufrin, J.R., O'Gara, M., et al. 1999. Mechanism of inhibition of DNA (cytosine C5)-methyltransferases by oligodeoxyribonucleotides containing 5,6-dihydro-5-azacytosine. *J Mol Biol* **285:** 2021-2034.

Sklenar, H., Etchebest, C., and Lavery, R. 1989. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **6:** 46-60.

Smith, D. 1989. SSTRUC: A program to calculate a secondary structural summary. . In *Department of Crystal-lography, Birkbeck College, University of London.*

Soto, C.S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. 2007. Loop modeling: sampling, filtering, and scoring. *Proteins*.

Srinivasan, N., Bax, B., Blundell, T.L., and Parker, P.J. 1996. Structural aspects of the functional modules in human protein kinase-C alpha deduced from comparative analyses. *Proteins* **26:** 217-235.

Srinivasan, R., and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* **96:** 14258-14263.

Street, T.O., Fitzkee, N.C., Perskie, L.L., and Rose, G.D. 2007. Physical-chemical determinants of turn conformations in globular proteins. *Protein Sci* **16:** 1720-1727.

Taylor, T., Rivera, M., Wilson, G., and Vaisman, II. 2005. New method for protein secondary structure assignment based on a simple topological descriptor. *Proteins* **60:** 513-524.

Thornton, J.M., Sibanda, B.L., Edwards, M.S., and Barlow, D.J. 1988. Analysis, design and modification of loop regions in proteins. *Bioessays* **8:** 63-69.

Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G., and Offmann, B. 2006. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res* **34:** W119-123.

Wang, G., and Dunbrack, R.L., Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* **19:** 1589-1591.

Wang, G., and Dunbrack, R.L., Jr. 2005. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33:** W94-98.

Wintjens, R., Wodak, S.J., and Rooman, M. 1998. Typical interaction patterns in alphabeta and betaalpha turn motifs. *Protein Eng* **11:** 505-522.

Wintjens, R.T., Rooman, M.J., and Wodak, S.J. 1996. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol* **255:** 235-253.

Wohlfahrt, G., Hangoc, V., and Schomburg, D. 2002. Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. *Proteins* **47:** 370-378.

Wojcik, J., Mornon, J.P., and Chomilier, J. 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol* **289:** 1469-1490.

Wolf, Y., Madej, T., Babenko, V., Shoemaker, B., and Panchenko, A.R. 2007. Long-term trends in evolution of indels in protein sequences. *BMC Evol Biol* **7:** 19.

Wong, S., and Jacobson, M.P. 2007. Conformational selection in silico: Loop latching motions and ligand binding in enzymes. *Proteins*.

Woodcock, S., Mornon, J.P., and Henrissat, B. 1992. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng* **5:** 629-635.

Zhang, W., Dunker, A.K., and Zhou, Y. 2007. Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* **71:** 61-67.

Zhang, W., Dunker, A.K., and Zhou, Y. 2008. Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* **71:** 61-67.

Zhu, K., Pincus, D.L., Zhao, S., and Friesner, R.A. 2006. Long loop prediction using the protein local optimization program. *Proteins* **65:** 438-452.