Analysis of loop boundaries using different local structure assignment methods.
Manoj Tyagi, Aurélie Bornot, Bernard Offmann, Alexandre de Brevern

To cite this version:

HAL Id: inserm-00392504
https://inserm.hal.science/inserm-00392504
Submitted on 7 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Analysis of loop boundaries using different local structure assignment methods.

Manoj Tyagi⁴, Aurélie Bornot⁵, Bernard Offmann⁵, ⁶ and Alexandre G. de Brevern⁴, ⁶

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 $\alpha$-helices and $\beta$-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 $\alpha$-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 $\gamma$- and $\beta$-turns, $\beta$-hairpins and $\beta$-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 Cα positions. It computes inter-Cα distance matrix and compares it with matrices produced by ideal repetitive secondary structures. KAKSI is a new assignment method of assignation using the inter-Cα distances and dihedral angles criteria (Martin et al. 2005). PSEA assigns the repetitive secondary structures from the sole Cα 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 an 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 ($\alpha$-$\alpha$, $\alpha$-$\beta$, $\beta$-$\alpha$ and $\beta$-$\beta$) (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 \( \alpha \)-helices, \( \beta \)-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 \( \alpha \)-helices (top) and \( \beta \)-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 \( \alpha \)-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, $\alpha$-helix N cap displays four patterns 1, whereas $\beta$-sheet N cap displays only two patterns 1, but also two patterns 3 and two patterns 5, *i.e.*, the capping regions of $\beta$-sheet are more variably described than those of $\alpha$-helix for which the correspondence between SSAMs is quite easily found. For the C caps, it goes to a higher level of complexity. Thus, $\alpha$-helix C cap has only one pattern 2, two patterns 3 and three patterns 4, while the $\beta$-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.* $\beta$-sheet N cap of STRIDE and SECSTR are shifted by (-1) residue. These results highlight greatly the difficulties to assign the $\beta$-strand extremities, while $\alpha$-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 \((x / y)^p_z\) that corresponds to the amino acid \(x\) over - and \(y\) under – represented at the position \(z\). N capping regions of \(\alpha\)-helices (Table 3a) show a strong pattern (PSTND / IVLMAFYQERK)\(^{p_1}\) (PE / GN)\(^{p_2}\) (AQDE / IGN)\(^{p_3}\) where \(p_1\) corresponds to position \(N_1\) for DSSP, STRIDE, SECSTR, PSEA and SEGNO and \(N_{\text{cap}}\) for XTLSSTR and KAKSI. This position \(p_1\) is associated to a high KLD value.

At the opposite, KLD values of C capping regions of \(\alpha\)-helices are weaker; multiple positions are in the same range of values. Repeated patterns (LAERK / IVPG) are found before \(p_1\), then (GN / IV)\(^{p_1}\), (PG / -)\(^{p_2}\) and (PK / -)\(^{p_3}\) where \(p_1\) corresponds to position \(C_1\) for DSSP, STRIDE, SECSTR, XTLSSTR, PSEA and SEGNO and \(C_{\text{cap}}\) 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 (\([\text{LAERK}], [\text{LAERK}], [\text{LAERK}], [\text{GN}], [\text{PG}], [\text{PK}]\)) can correspond for instance, to the sequence (L A L N P K). The succession LAL of \(C_2\), \(C_1\), \(C_{\text{cap}}\) cannot be shifted as they are mainly under-represented at positions \(C_1\)', \(C_2\)' and \(C_3\)'.

N capping regions of \(\beta\)-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)\(^{p_1}\) followed by compatible patterns (IVFY / AQPGNDERK); this latter corresponding to the \(\beta\)-strand; position \(p_1\) 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 (IVLFY\(N\) / AQPGN\(DR\)K) followed by (G\(N\)D / 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 K\(Ld\). Analysis of the position informativity with K\(Ld\) 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 K\(Ld\) values, whereas the boundary region, i.e. C_{cap} and C_1, have fewer amino acids over and under-represented and low K\(Ld\) 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 C_{cap}. 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 C₃' of SEGNO is particular due to an over-representation of Glycine not found in any other SSAM and thus more related to C₂' position of DSSP than C₃' 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 [N₃' - N₁'] and [N₃cap - N₂] positions of DSSP are found. For instance, [N₃' - N₁'] of XTLSSTR correspond to N₂ 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 C₁' positions is not equivalent to C₁' of DSSP. PSEA adds to this, a shift of positions C₁ and C₃cap; 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 N₃cap, N₁, N₂ and positions C₂, C₁, C₃cap (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
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](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 (Frisman 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/](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:

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

with $n_{ij}^{obs}$ 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 $n_{ij}^{th}$. 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:

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

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

**Pattern 1:** N-cap of $\alpha$-helix (P-SEA)  
**Pattern 2:** C-cap of $\alpha$-helix (STRIDE)  
**Pattern 3:** N-cap of $\beta$-sheet (XTLSSTR)  
**Pattern 4:** C-cap of $\beta$-sheet (SECSTR)

**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

<table>
<thead>
<tr>
<th>Methods</th>
<th>year</th>
<th>Assignment based on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greer &amp; Levitt</td>
<td>1977</td>
<td>Distance</td>
</tr>
<tr>
<td>DSSP</td>
<td>1983</td>
<td>H-bond</td>
</tr>
<tr>
<td>DEFINE</td>
<td>1988</td>
<td>Distance</td>
</tr>
<tr>
<td>PCURVE</td>
<td>1989</td>
<td>Axis</td>
</tr>
<tr>
<td>SSTRUC</td>
<td>1989</td>
<td>H-bond</td>
</tr>
<tr>
<td>CONCENSUS</td>
<td>1993</td>
<td>Mean (DSSP, DEFINE and PCURVE)</td>
</tr>
<tr>
<td>STRIDE</td>
<td>1995</td>
<td>H-bond / dihedral</td>
</tr>
<tr>
<td>PROMOTIF</td>
<td>1996</td>
<td>H-bond / dihedral</td>
</tr>
<tr>
<td>PSEA</td>
<td>1997</td>
<td>Distance / angle</td>
</tr>
<tr>
<td>PROSS</td>
<td>1999</td>
<td>Dihedral</td>
</tr>
<tr>
<td>XTLSSTR</td>
<td>1999</td>
<td>Distance / angle</td>
</tr>
<tr>
<td>DSSPcont</td>
<td>2002</td>
<td>H-bond</td>
</tr>
<tr>
<td>SECSTR</td>
<td>2002</td>
<td>H-bond</td>
</tr>
<tr>
<td>VORO3D</td>
<td>2004</td>
<td>Voronoï</td>
</tr>
<tr>
<td>KAKSI</td>
<td>2005</td>
<td>Distance / dihedral</td>
</tr>
<tr>
<td>SEGNO</td>
<td>2005</td>
<td>angle / multiple</td>
</tr>
<tr>
<td>Beta-Spider</td>
<td>2005</td>
<td>$\beta$-sheet + DSSP for $\alpha$-helix</td>
</tr>
<tr>
<td>PALSSE</td>
<td>2005</td>
<td>$C_\alpha$ (vector similarity)</td>
</tr>
<tr>
<td>Delaunay tessellation</td>
<td>2005</td>
<td>Delaunay</td>
</tr>
<tr>
<td>SKSP</td>
<td>2007</td>
<td>Mean (STRIDE, DSSP, SECSTR, KAKSI, P-SEA, and SEGNO)</td>
</tr>
<tr>
<td>PROSIGN:</td>
<td>2008</td>
<td>$C_\alpha$ deviation values</td>
</tr>
</tbody>
</table>

Table 1 – Secondary structure assignment methods.
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, 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.

### (a)

<table>
<thead>
<tr>
<th>Protein Databanks</th>
<th>DB 0</th>
<th>DB 1</th>
<th>DB 2</th>
<th>DB 3</th>
<th>DB 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
</tr>
<tr>
<td>DSSP</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>STRIDE</td>
<td>2.77</td>
<td>5.51</td>
<td>22.38</td>
<td>5.45</td>
<td>21.78</td>
</tr>
<tr>
<td>SECDRT</td>
<td>40.05</td>
<td>44.37</td>
<td>44.97</td>
<td>44.93</td>
<td>44.25</td>
</tr>
<tr>
<td>PSE</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>SECDRT</td>
<td>2.77</td>
<td>5.51</td>
<td>22.38</td>
<td>5.45</td>
<td>21.78</td>
</tr>
<tr>
<td>KANS</td>
<td>40.05</td>
<td>44.37</td>
<td>44.97</td>
<td>44.93</td>
<td>44.25</td>
</tr>
<tr>
<td>SECON</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
</tr>
<tr>
<td>DSSP</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>STRIDE</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>SECDRT</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>PSE</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>SECON</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>Rfactor</td>
<td>0.25</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### (b)

<table>
<thead>
<tr>
<th>Protein Databanks</th>
<th>DB 5</th>
<th>DB 6</th>
<th>DB 7</th>
<th>DB 8</th>
<th>DB 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
</tr>
<tr>
<td>DSSP</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>STRIDE</td>
<td>2.77</td>
<td>5.51</td>
<td>22.38</td>
<td>5.45</td>
<td>21.78</td>
</tr>
<tr>
<td>SECDRT</td>
<td>40.05</td>
<td>44.37</td>
<td>44.97</td>
<td>44.93</td>
<td>44.25</td>
</tr>
<tr>
<td>PSE</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>SECON</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>KANS</td>
<td>40.05</td>
<td>44.37</td>
<td>44.97</td>
<td>44.93</td>
<td>44.25</td>
</tr>
<tr>
<td>SECON</td>
<td>3.14</td>
<td>10.68</td>
<td>33.60</td>
<td>11.15</td>
<td>33.37</td>
</tr>
<tr>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
<td>freq</td>
<td>lg</td>
</tr>
<tr>
<td>DSSP</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>STRIDE</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>SECDRT</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>PSE</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>SECON</td>
<td>1.49</td>
<td>3.51</td>
<td>33.51</td>
<td>3.51</td>
<td>33.51</td>
</tr>
<tr>
<td>Rfactor</td>
<td>0.25</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
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.
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.

<table>
<thead>
<tr>
<th>Ncap beta</th>
<th>N3'</th>
<th>N2'</th>
<th>N1'</th>
<th>Ncap</th>
<th>N1</th>
<th>N2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSP</td>
<td>PGND</td>
<td>PGND</td>
<td>PGND</td>
<td>IYFT</td>
<td>IFLY</td>
<td>IFLY</td>
</tr>
<tr>
<td>STRIDE</td>
<td>PGND</td>
<td>PGND</td>
<td>PGND</td>
<td>IYFT</td>
<td>IFLY</td>
<td>IFLY</td>
</tr>
<tr>
<td>SECSTR</td>
<td>PGN</td>
<td>PGND</td>
<td>PGND</td>
<td>IYFT</td>
<td>IFLY</td>
<td>IFLY</td>
</tr>
<tr>
<td>XTLSSR</td>
<td>PGK</td>
<td>PGK</td>
<td>PGK</td>
<td>G</td>
<td>IYFT</td>
<td>IFLY</td>
</tr>
<tr>
<td>PSEA</td>
<td>GNK</td>
<td>PN</td>
<td>GN</td>
<td>IVY</td>
<td>IMLY</td>
<td>IFLY</td>
</tr>
<tr>
<td>DEFINE</td>
<td>G</td>
<td>G</td>
<td>V</td>
<td>IFLY</td>
<td>IFLY</td>
<td>IFLY</td>
</tr>
<tr>
<td>KASSO</td>
<td>PGND</td>
<td>PGN</td>
<td>GN</td>
<td>IVF</td>
<td>VIPY</td>
<td>VIPY</td>
</tr>
<tr>
<td>PBs</td>
<td>GN</td>
<td>GCK</td>
<td>VP</td>
<td>IMLY</td>
<td>IFLY</td>
<td>IFLY</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ccap beta</th>
<th>C2</th>
<th>C1</th>
<th>Ccap</th>
<th>C1'</th>
<th>C2'</th>
<th>C3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSP</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>STRIDE</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>SECSTR</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>XTLSSR</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>PSEA</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>DEFINE</td>
<td>P</td>
<td>P</td>
<td>G</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>KASSO</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
<tr>
<td>PBs</td>
<td>IFLY</td>
<td>IFLY</td>
<td>GND</td>
<td>IFLY</td>
<td>GND</td>
<td></td>
</tr>
</tbody>
</table>

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


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


