Analysis of loop boundaries using different local structure assignment methods

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

[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]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

[4]INSERM UMR-S 665, DSIMB, Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France

[5]PEACCEL, 37, rue de Vienne, 97430 Le Tampon, France

Received 4 February 2009; Revised 30 April 2009; Accepted 2 June 2009
DOI: 10.1002/pro.198
Published online 15 July 2009 proteinscience.org

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 this 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. Because these latter greatly influence the loop boundary definitions, important variations are indeed observed, that is, 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. Although SSAMs give very different local structure assignments capping sequence patterns remain efficiently stable.

Keywords: protein structures; biochemistry; amino acids; secondary structures; propensities

---

Additional Supporting Information may be found in the online version of this article.

[†]Manoj Tyagi and Aurélie Bornot contributed equally to this work.

Manoj Tyagi's current address is Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, MD 20894.

Grant sponsor: The Ministère de la Recherche, Université Paris Diderot-Paris 7; Université de La Réunion; The French Institute for Health and Medical Research (INSERM); The Conseil Régional de La Réunion and European Union; Ministère de la Recherche.

*Correspondence to: 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. E-mail: alexandre.debrevern@univ-paris-diderot.fr

Table I. *Secondary Structure Assignment Methods*

Methods	Year	Assignment Based On
Greer and 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 tessalation
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 tessalation
SKSP	2007	Mean (STRIDE, DSSP, SECSTR, KAKSI, P-SEA, and SEGNO)
PROSIGN	2008	Cα deviation values

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[1,2]). This monodimensional description helps to simplify coarsely this 3D information. It can also be used to describe more complex local 3D motifs, for example, the Greek key,[3] or even complete 3D structures in 2D views, for example, HERA[4] or TOPS.[5]

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

A first class of methods is based solely on H-bond patterns. In this category, DSSP[6] 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).[7,8] A recent version of DSSP called DSSPcont was proposed by Rost.[9] SECSTR is also an evolution of DSSP method dedicated to improved π-helices detection.[10]

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.[11] PROMOTIF derives also from the DSSP approach, namely the software SSTRUC,[12] but focus on the characterization of γ- and β-turns, β-hairpins, and β-bulges.[13]

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,[14] like the Levitt's and Greer's method,[15] 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.[16] PSEA assigns the repetitive secondary structures from the sole C_α position using distance and angles criteria.[17] XTLSSTR uses all the backbone atoms to compute two angles and three distances.[18]

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.[19] SEGNO uses also the Φ and Ψ dihedral angles coupled with other angles to assign the secondary structures.[20]

Fifth, VoTap (Voronoï Tessellation Assignment Procedure) is a geometrical tool that associates with each amino acid a Voronoï polyhedron,[21] the faces of which define contacts between residues.[22] In the same way, Vaisman and coworkers 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.[23]

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.[24] PALSSE delineates secondary structure elements from protein Cα coordinates, and specifically addresses the requirements of vector-

based protein similarity searches[25]; 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.[26] Hosseini and coworkers 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.[27]

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.[28–30] This was one of the motivations of KAKSI methodology to define linear helices instead of long kinked helices.[16] Moreover, the disagreement between the different SSAMs is not negligible, leading to only 80% of agreement between two distinct methods.[16,31–33] Consensus methods have been proposed using (i) DEFINE, P-CURVE, and DSSP[32] and (ii) more recently, P-SEA, KAKSI, SECSTR, and STRIDE,[34] to diminish such features.

The coil state is in fact composed of really distinct local folds,[35–38] such as turns.[13,39–44] Several studies have attempted to analyze conformation of loops linking specific secondary structures forming distinct subsets.[45–51] They are biologically essential regions,[52] for example, loops of protein kinases.[53,54] They are also used to analyze protein homology,[55–60] for example, for structure-based phylogenetic study.[61] Because of their flexible nature they raise crucial questions in protein docking approaches,[62–64] to predict protein loop conformations,[65–78] to enhance protein thermostability,[79] to design proteins,[80] or to obtain protein structures.[81] According to the repetitive secondary structures of their extremities, connecting loops are of four distinct classes (α-α, α-β, β-α, and β-β).[46,82–84] The research on loops has always been limited by the number of available loops in protein structures from the Protein DataBank (PDB,[7,8]), so most of the works focus on loops of less than nine residues.[85,86]

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,[87] 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.[88] 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.[89] To date, no similar properties have been reported directly on β-strands.

In this article, we focus on the analysis of loop boundaries, that is, 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 nonredundant 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 Supporting Information 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: three datasets were filtered for high resolution quality (resolution better than 1.6 Å) and seven were filtered for good resolution quality (resolution better than 2.5 Å). The datasets have been extracted from PISCES database.[90,91]

Table II 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,[33] that is, α-helices frequency ranges mainly between 28 and 34% and β-strand between 18 and 24%. Some SSAMs have particular behaviors like KAKSI[16] that is associated to a high β-strand frequency (∼28%) or DEFINE[14] 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*[92] assigned by different SSAMs, it highlights visually how the differences can be important (see also Supporting Information 2). In the same way, the computation of C_3, that is, the agreement rates between SSAMs (see Methods section), gives also similar results to previous works[16,33,93] (see Fig. 2). Briefly,

Table II. *10 Protein Databanks*

		DB0		DB1		DB2		DB3		DB4	
		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	48.3		47.06		47.10		47.27		48.75	
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		2722		3325		3983		5081	
pc		20		20		25		30		40	
res		1.6		2.5		2.5		2.5		2.5	
R factor		0.25		1.00		1.00		1.00		1.00	
		DB5		DB6		DB7		DB8		DB9	
		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	5.82	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	32.08	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		1425		5847		6823		7141		1630	
pc		50		50		70		80		90	
res		1.6		2.5		2.5		2.5		1.6	
R factor		0.25		1.00		1.00		1.00		0.25	

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 *R* factor.

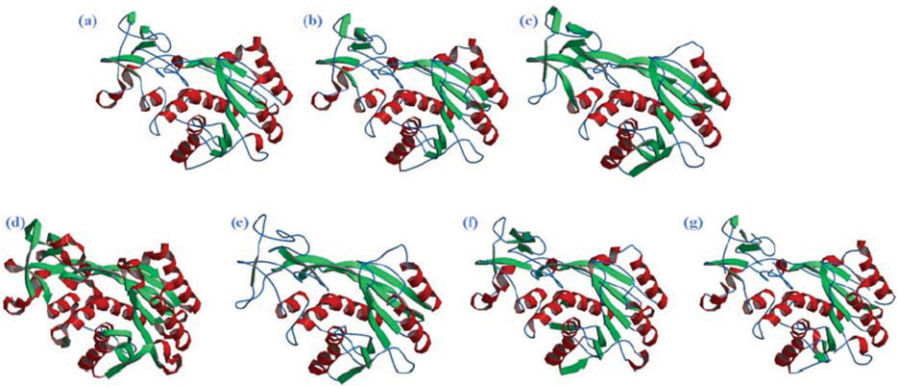

Figure 1. SSAMs of Hhai Methyltransferase. Example of secondary structure assignments for the *Hhai Methyltransferase* (PDB code :10MH[92]) with (a) DSSP, (b) STRIDE, (c) PSEA, (d) DEFINE, (e) PCURVE, (f) XTLSSTR, and (g) SECSTR. All the methods have been reduced to three states with the helical states in red ribbons, the extended state in green arrows, and the coil in blue line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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

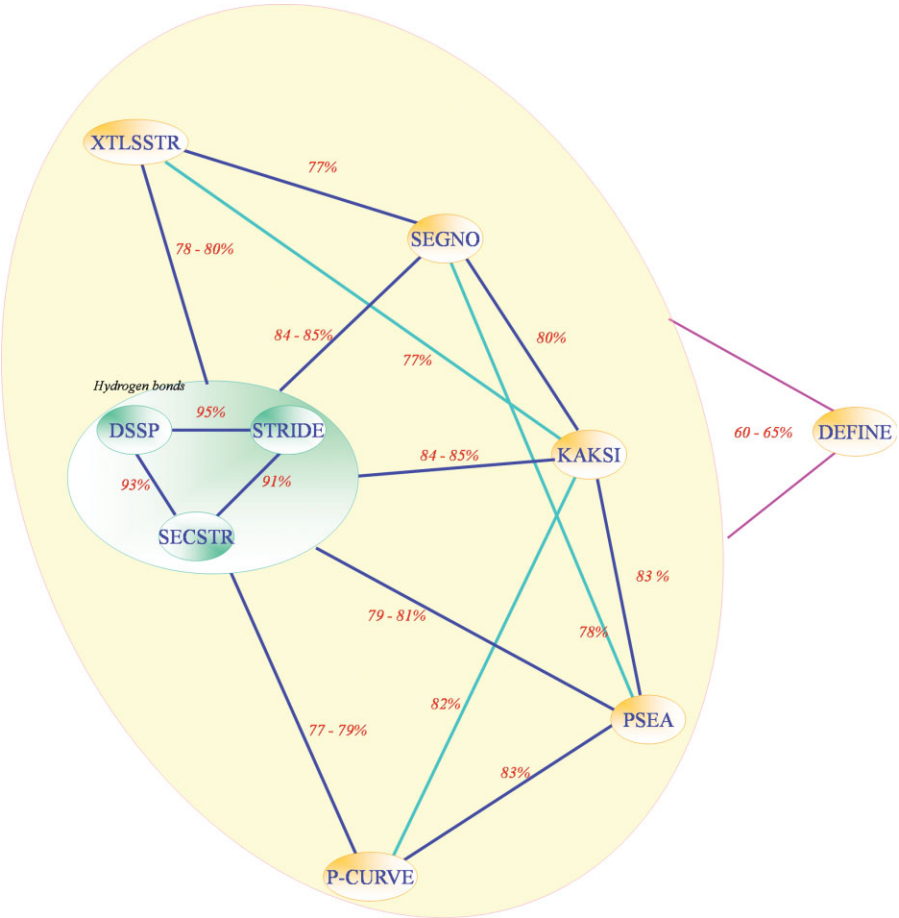

Figure 2. C_3 values for different SSAMs (DB0 dataset). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

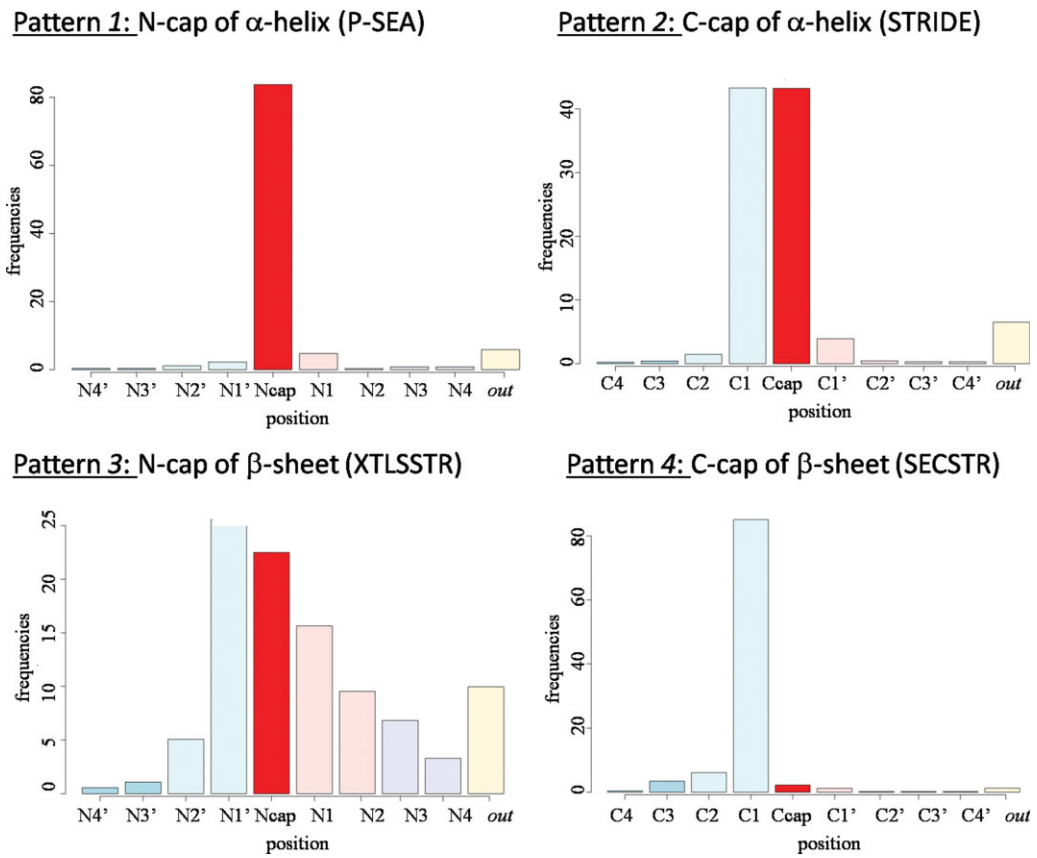

Figure 3. Examples of discrepancies between N or C cap positions assigned by DSSP with other SSAMs. Examples of the four kinds of differences are shown. (*x*-axis) the position of the capping region, (*y*-axis) frequencies of N or C cap central positions of SSAMs according to DSSP. Central positions are in red color. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

and locating their extremities.[87,88,94-98] 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 six residues have been used). Figure 3 shows some examples of this systematic comparison (see Supporting Information 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 Fig. 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 Supporting Information 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, that is, 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, that is, the correspondence between SSAMs are quite complex. Surprisingly, even the

SSAM related to DSSP are not strictly equivalent to it, for example, 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 III shows the over- and under-representation of amino acid of the different SSAMs in terms of *Z*-scores.[99] Thus, at each position of each SSAM is given the important amino acids. *KLd*[100] values were also computed to locate the most informative positions (see Supporting Information 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 III) 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 III) 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 III. Positions C_2, C_1, and C_2' have a strong amino acid distribution associated with high *KLd* values, whereas the boundary region, that is, 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 Supporting Information 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 $C_{3'}$ 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.[16,32,33,34,101] 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, that is, positions N_{cap}, N_1, N_2 and positions C_2, C_1, C_{cap}.[99] 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

Table III. *Amino Acid Over- and Under-Representation at Capping Regions*

C cap beta[a]		C_2	C_1	C_{cap}	C_1'	C_2'	C_3'
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	PSTAND	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
SECSTER	(−)			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
C cap alpha[a]		C_2	C_1	C_{cap}	C_1'	C_2'	C_3'
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	IVLAF		A
DEFINE	(−)						
KAKSI	(−)	VPGT	VPGD	IVP		IVL	
SEGNO	(−)	PGSTND	VPGTD	IVFPGD	IVPT	IVF	IVLFYT
PBs	(−)	PGST	VPGTD	VGT	IVG	IVG	IVLMAFYWTE
N cap beta[b]		N_3'	N_2'	N_1'	N_{cap}	N_1	N_2
DSSP	(+)	PGND	PGND	PGND	IVFYT	IVLFY	IVFY
STRIDE	(+)	PGSND	PGND	PGND	IVFYT	IVLFY	IVFY
SECSTR	(+)	PGN	PGND	PGND	IVFYWT	IVLY	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
XTLLSTR	(−)	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
C cap beta[b]		C_2	C_1	C_{cap}	C_1'	C_2'	C_3'
DSSP	(+)	IVLFYW	IVFYWC	IVFYD	GSND	PGSND	GSND
STRIDE	(+)	IVFYW	IVFYWC	IVYD	GND	PGSND	GSND
SECSTR	(+)	IVLFYW	IVFYWCT	IVFY	GND	PGSND	GSND
XTLLSTR	(+)	IVF	IVFYWT	IVFYTD	PGND	PGSD	PGSND
PSEA	(+)	IVLFY	IVFYCT	PSTD	PND	GSND	GSD

(*Continued*)

TABLE III. *Amino Acid Over- and Under-Representation at Capping Regions (Continued)*

C cap beta[a]		C_2	C_1	C_{cap}	C_1'	C_2'	C_3'
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

[a] 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 DBo.
[b] 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 DBo.

possess a true amino acid patterns (see Table III), 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 20–50 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, for example, DSSP and KAKSI helices. These results are in agreement with the results of Kruus and coworkers[89] 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,[38] namely the Protein Blocks.[93,99,101-112] Their results are a bit different from the SSAMs, for example, N_{cap} and C_{cap} have always lower *KLd* values than other positions. Contrary to the SSAMs, they approximate even the nonrepetitive states, that is, 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,[113] "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.[114]" Indeed, the discrepancies are often found at the extremities of repetitive structures and loop boundaries are essential in loop conformation prediction.[65] Nonetheless, we have shown here that systematically differences do not appear in terms of sequence. This result reinforce the results of Kruus and coworkers.[89] This study is also related to the elegant research done by Zhang and coworkers.[34] 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 emphasized the idea of Grishin with PALSSE, that focus on the sequence property as on the structure properties to assign the repetitive structure.[25]

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).[115] Recently, Zhang and coworkers 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%.[116] 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 coworkers showed that loop boundary prediction methods relying on sequence specificities seem to be more efficient that methods based on physical properties of amino acids.[117] 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[90,91] 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 II presents all the details of these databanks.

Secondary structure assignments

They have been done with five distinct software: DSSP[6] (CMBI version 2000), STRIDE,[11], SECSTR[10] (version 0.2.3-1), XTLSSTR,[18], PSEA[17] (version 2.0), DEFINE[14] (version 2.0), KAKSI[16] (version 1.0.1), and SEGNO[20] (version 3.1). PBs[93] have been assigned using in-house software (available at http://www.dsimb.inserm.fr/~debrevern/DOWN/LECT/), it follows similar rules to assignment done by PBE web server (http://bioinformatics.univ-reunion.fr/PBE/).[118] 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 ($n = 1$–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.[33]

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:

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

with $n_{i,j}^{obs}$ the observed occurrence number of amino acid i in position j for a given secondary structure and $n_{i,j}^{\text{th}}$ the expected number. The product of the occurrences in position j with the frequency of amino acid i in the entire databank equals $n_{i,j}^{\text{th}}$. Positive *Z*-scores (respectively negative) correspond to over-represented 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,[100] denoted by *KLd,* evaluates the contrast between two amino acid distributions, that is, 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:

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

where $P(\mathrm{aa}_j = i|S_x)$ is the probability of observing the amino acid i in position j ($j = -w, \ldots, 0, \ldots, +w$) of the sequence window (15 residue long, $w = 7$) given a secondary structure S_x, and, $P(\mathrm{aa}_j = i|DB)$ 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.[99]

References

1. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 37:251–256.
2. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37: 205–211.
3. Hutchinson EG, Thornton JM (1993) The Greek key motif: extraction, classification and analysis. Protein Eng 6:233–245.
4. Hutchinson EG, Thornton JM (1990) HERA–a program to draw schematic diagrams of protein secondary structures. Proteins 8:203–212.
5. Michalopoulos I, Torrance GM, Gilbert DR, Westhead DR (2004) TOPS: an enhanced database of protein structural topology. Nucleic Acids Res 32:D251–D254.
6. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.
7. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242.
9. Andersen CA, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. Structure (Camb) 10:175–184.
10. Fodje MN, Al-Karadaghi S (2002) Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng 15:353–358.
11. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579.
12. Smith D (1989) SSTRUC: A program to calculate a secondary structural summary. Department of Crystallography, Birkbeck College, University of London, Cambridge, UK.
13. Hutchinson EG, Thornton JM (1996) PROMOTIF–a program to identify and analyze structural motifs in proteins. Protein Sci 5:212–220.
14. Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 3:71–84.
15. Levitt M, Greer J (1977) Automatic identification of secondary structure in globular proteins. J Mol Biol 114: 181–239.
16. Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17.
17. Labesse G, Colloc'h N, Pothier J, Mornon JP (1997) P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci 13:291–295.
18. King SM, Johnson WC (1999) Assigning secondary structure from protein coordinate data. Proteins 35: 313–320.
19. Srinivasan R, Rose GD (1999) A physical basis for protein secondary structure. Proc Natl Acad Sci USA 96: 14258–14263.
20. Cubellis MV, Cailliez F, Lovell SC (2005) Secondary structure assignment that accurately reflects physical and evolutionary characteristics. BMC Bioinformatics 6 (Suppl 4):S8.
21. Dupuis F, Sadoc JF, Jullien R, Angelov B, Mornon JP (2005) Voro3D: 3D Voronoi tessellations applied to protein structures. Bioinformatics 21:1715–1716.
22. Dupuis F, Sadoc JF, Mornon JP (2004) Protein secondary structure assignment through Voronoi tessellation. Proteins 55:519–528.
23. Taylor T, Rivera M, Wilson G, Vaisman, II (2005) New method for protein secondary structure assignment based on a simple topological descriptor. Proteins 60: 513–524.
24. Sklenar H, Etchebest C, Lavery R (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 6:46–60.
25. Majumdar I, Krishna SS, Grishin NV (2005) PALSSE: A program to delineate linear secondary structural elements from protein structures. BMC Bioinformatics 6: 202.
26. Hosseini S, Sadeghi M, Pezeshk H, Eslahchi C, Habibi M (2005) PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. Comput Biol Chem 32:406–411.
27. Parisien M, Major F (2005) A new catalog of protein beta-sheets. Proteins 61:545–558.
28. Kumar S, Bansal M (1996) Structural and sequence characteristics of long alpha helices in globular proteins. Biophys J 71:1574–1586.
29. Kumar S, Bansal M (1998) Geometrical and sequence characteristics of alpha-helices in globular proteins. Biophys J 75:1935–1944.
30. Bansal M, Kumar S, Velavan R (2000) HELANAL: a program to characterize helix geometry in proteins. J Biomol Struct Dyn 17:811–819.
31. Woodcock S, Mornon JP, Henrissat B (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein Eng 5:629–635.
32. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 6: 377–382.
33. Fourrier L, Benros C, de Brevern AG (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 5:58.
34. Zhang W, Dunker AK, Zhou Y (2007) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins 71: 61–67.
35. Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339.

36. Fitzkee NC, Fleming PJ, Gong H, Panasik N, Jr, Street TO, Rose GD (2005a) Are proteins made from a limited parts list? Trends Biochem Sci 30:73–80.
37. Fitzkee NC, Fleming PJ, Rose GD (2005b) The protein coil library: a structural database of nonhelix, nonstrand fragments derived from the PDB. Proteins 58:852–854.
38. Offmann B, Tyagi M, deBrevern AG (2007) Local Protein Structures. Curr Bioinformatics 3:165–202.
39. Rose GD, Seltzer JP (1977) A new algorithm for finding the peptide chain turns in a globular protein. J Mol Biol 113:153–164.
40. Rose GD, Gierasch LM, Smith JA (1985) Turns in peptides and proteins. Adv Protein Chem 37:1–109.
41. Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci 3:2207–2216.
42. Fuchs PF, Alix AJ (2005) High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59:828–839.
43. Bornot A, deBrevern AG (2006) Protein beta-turn assignments. Bioinformation 1:153–155.
44. Street TO, Fitzkee NC, Perskie LL, Rose GD (2007) Physical-chemical determinants of turn conformations in globular proteins. Protein Sci 16:1720–1727.
45. Edwards MS, Sternberg JE, Thornton JM (1987) Structural and sequence patterns in the loops of beta alpha beta units. Protein Eng 1:173–181.
46. Thornton JM, Sibanda BL, Edwards MS, Barlow DJ (1988) Analysis, design and modification of loop regions in proteins. Bioessays 8:63–69.
47. Ring CS, Kneller DG, Langridge R, Cohen FE (1992) Taxonomy and conformational analysis of loops in proteins. J Mol Biol 224:685–699.
48. Wintjens RT, Rooman MJ, Wodak SJ (1996) Automatic classification and analysis of alpha alpha-turn motifs in proteins. J Mol Biol 255:235–253.
49. Boutonnet NS, Kajava AV, Rooman MJ (1998) Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. Proteins 30: 193–212.
50. Wintjens R, Wodak SJ, Rooman M (1998) Typical interaction patterns in alphabeta and betaalpha turn motifs. Protein Eng 11:505–522.
51. Efimov AV (2008) Structural trees for proteins containing phi-motifs. Biochemistry (Mosc) 273:23–28.
52. Espadaler J, Querol E, Aviles FX, Oliva B (2006) Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics 22: 2237–2243.
53. Rekha N, 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.
54. Fernandez-Fuentes N, Hermoso A, Espadaler J, Querol E, Aviles FX, Oliva B (2004) Classification of common functional loops of kinase super-families. Proteins 56: 539–555.
55. Srinivasan N, Bax B, Blundell TL, Parker PJ (1996) Structural aspects of the functional modules in human protein kinase-C alpha deduced from comparative analyses. Proteins 26:217–235.
56. Panchenko AR, Madej T (2004) Analysis of protein homology by assessing the (dis)similarity in protein loop regions. Proteins 57:539–547.
57. Panchenko AR, Madej T (2005) Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol Biol 5:10.
58. Panchenko AR, Wolf YI, Panchenko LA, Madej T (2005) Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 61:535–544.
59. Madej T, Panchenko AR, Chen J, Bryant SH (2007) Protein homologous cores and loops: important clues to evolutionary relationships between structurally similar proteins. BMC Struct Biol 7:23.
60. Wolf Y, Madej T, Babenko V, Shoemaker B, Panchenko AR (2007) Long-term trends in evolution of indels in protein sequences. BMC Evol Biol 7:19.
61. Jiang H, Blouin C (2007) Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. BMC Bioinformatics 8: 444.
62. Huang Z, Wong CF, Wheeler RA (2008) Flexible protein-flexible ligand docking with disrupted velocity simulated annealing. Proteins 71:440–454.
63. Nabuurs SB, Wagener M, deVlieg J (2007) A flexible approach to induced fit docking. J Med Chem 50: 6507–6518
64. Wong S, Jacobson MP (2008) Conformational selection in silico: loop latching motions and ligand binding in enzymes. Proteins 71:153–164.
65. Lessel U, Schomburg D (1999) Importance of anchor group positioning in protein loop prediction. Proteins 37:56–64.
66. Miyazaki S, Kuroda Y, Yokoyama S (2002) Characterization and prediction of linker sequences of multi-domain proteins by a neural network. J Struct Funct Genomics 2:37–51.
67. Wohlfahrt G, Hangoc V, Schomburg D (2002) Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. Proteins 47:370–378.
68. Rohl CA, Strauss CE, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with rosetta. Proteins 55:656–677.
69. Boomsma W, Hamelryck T (2005) Full cyclic coordinate descent: solving the protein loop closure problem in Calpha space. BMC Bioinformatics 6:159.
70. Monnigmann M, Floudas CA (2005) Protein loop structure prediction with flexible stem geometries. Proteins 61:748–762.
71. Fernandez-Fuentes N, Fiser A (2006) Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol 6:15.
72. Fernandez-Fuentes N, Oliva B, Fiser A (2006a) A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 34:2085–2097.
73. Fernandez-Fuentes N, Zhai J, Fiser A (2006b) ArchPRED: a template based loop structure prediction server. Nucleic Acids Res 34:W173–W176.
74. Zhu K, Pincus DL, Zhao S, Friesner RA (2006) Long loop prediction using the protein local optimization program. Proteins 65:438–452.
75. Kanagasabai V, Arunachalam J, Prasad PA, Gautham N (2007) Exploring the conformational space of protein loops using a mean field technique with MOLS sampling. Proteins 67:908–921.
76. Olson MA, Feig M, Brooks CL, III (2008) Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. J Comput Chem 29:820–831.
77. Prasad PA, Kanagasabai V, Arunachalam J, Gautham N (2007) Exploring conformational space using a mean field technique with MOLS sampling. J Biosci 32: 909–920.

78. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B (2008) Loop modeling: sampling, filtering, and scoring. Proteins 70:834–843.
79. Reetz MT, Carballeira JD, 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.
80. Hu X, Wang H, Ke H, Kuhlman B (2007) High-resolution design of a protein loop. Proc Natl Acad Sci USA 104:17668–17673.
81. Rapp CS, Strauss T, Nederveen A, Fuentes G (2007) Prediction of protein loop geometries in solution. Proteins 69:69–74.
82. Efimov AV (1991b) Structure of coiled beta-beta-hairpins and beta-beta-corners. FEBS Lett 284:288–292.
83. Efimov AV (1991a) Structure of alpha-alpha-hairpins with short connections. Protein Eng 4:245–250.
84. Rufino SD, Donate LE, Canard LH, Blundell TL (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.
85. Wojcik J, Mornon JP, 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.
86. Michalsky E, Goede A, Preissner R (2003) Loops In Proteins (LIP)–a comprehensive loop database for homology modelling. Protein Eng 16:979–985.
87. Presta LG, Rose GD (1988) Helix signals in proteins. Science 240:1632–1641.
88. Aurora R, Rose GD (1998) Helix capping. Protein Sci 7: 21–38.
89. Kruus E, Thumfort P, Tang C, Wingreen NS (2005) Gibbs sampling and helix-cap motifs. Nucleic Acids Res 33:5343–5353.
90. Wang G, Dunbrack RL, Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19:1589–1591.
91. Wang G, Dunbrack RL, Jr (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33:W94–W98.
92. Sheikhnejad G, Brank A, Christman JK, Goddard A, Alvarez E, Ford H, Jr, Marquez VE, Marasco CJ, Sufrin JR, O'Gara M (1999) Mechanism of inhibition of DNA (cytosine C5)-methyltransferases by oligodeoxyribonucleotides containing 5,6-dihydro-5-azacytosine. J Mol Biol 285:2021–2034.
93. deBrevern AG (2005) New assessment of a structural alphabet. In Silico Biol 5:283–289.
94. Doig AJ, Baldwin RL (1995) N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. Protein Sci 4:1325–1336.
95. Mandel-Gutfreund Y, Zaremba SM, Gregoret LM (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.
96. Mandel-Gutfreund Y, Gregoret LM (2002) On the significance of alternating patterns of polar and non-polar residues in beta-strands. J Mol Biol 323:453–461.
97. Bang D, Gribenko AV, Tereshko V, Kossiakoff AA, Kent SB, Makhatadze GI (2006) Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis. Nat Chem Biol 2:139–143.
98. Rose GD (2006) Lifting the lid on helix-capping. Nat Chem Biol 2:123–124.
99. deBrevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41:271–287.
100. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86.
101. deBrevern AG, Benros C, Hazout S. Structural alphabet: from a local point of view to a global description of protein 3d structures. In: Yan PV, Ed. (2005a) Bioinformatics: New Research. New York: Nova Publishers, pp 127–169.
102. deBrevern AG, Camproux AC, Hazout S, Etchebest C, Tuffery P. Protein structural alphabets: beyond the secondary structure description. In: Sangadai S, Ed. (2001) Recent research developments in protein engineering. Trivandrum: Research Signpost, pp 319–331.
103. deBrevern AG, Valadie H, Hazout S, Etchebest C (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 11:2871–2886.
104. deBrevern AG, Hazout S (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19:345–353.
105. deBrevern AG, Benros C, Gautier R, Valadie H, Hazout S, Etchebest C (2004) Local backbone structure prediction of proteins. In Silico Biol 4:381–386.
106. deBrevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C, 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.
107. Benros C, deBrevern AG, Etchebest C, Hazout S (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62:865–880.
108. deBrevern AG, Etchebest C, Benros C, 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.
109. Etchebest C, Benros C, Bornot A, Camproux AC, deBrevern AG (2007) A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36:1059–1069.
110. Benros C, deBrevern AG, Hazout S (2009) Analyzing the sequence-structure relationship of a library of local structural prototypes. J Theor Biol 256:215–226.
111. Bornot A, Etchebest C, deBrevern AG (2009) A new prediction strategy for long local protein structures using an original description. Proteins 76:570–587.
112. Faure G, Bornot A, deBrevern AG (2009) Analysis of protein contacts into protein units. Biochimie 91: 876–887.
113. Lesk AM (2005) Introduction to Bioinformatics. Oxford: Oxford University Press.
114. Robson B, Garnier J (1986) Introduction to proteins and protein engineering. Amsterdam: Elsevier Press.
115. Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34:508–519.
116. Zhang W, Dunker AK, Zhou Y (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins 71: 61–67.
117. Dovidchenko NV, Bogatyreva NS, Galzitskaya OV (2008) Prediction of loop regions in protein sequence. J Bioinform Comput Biol 6:1035–1047.
118. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, deBrevern AG, Offmann B (2006) Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 34:W119–W123.