

# Description and recognition of regular and distorted secondary structures in proteins using the automated protein structure analysis method

Sushilee Ranganathan,<sup>1</sup> Dmitry Izotov,<sup>1</sup> Elfi Kraka,<sup>1</sup> and Dieter Cremer<sup>1,2\*</sup>

<sup>1</sup>Department of Chemistry, University of the Pacific, Stockton, California 95211

<sup>2</sup>Department of Physics, University of the Pacific, Stockton, California 95211

## ABSTRACT

The Automated Protein Structure Analysis (APSA) method, which describes the protein backbone as a smooth line in three-dimensional space and characterizes it by curvature  $\kappa$  and torsion  $\tau$  as a function of arc length  $s$ , was applied on 77 proteins to determine all secondary structural units via specific  $\kappa(s)$  and  $\tau(s)$  patterns. A total of 533  $\alpha$ -helices and 644  $\beta$ -strands were recognized by APSA, whereas DSSP gives 536 and 651 units, respectively. Kinks and distortions were quantified and the boundaries (entry and exit) of secondary structures were classified. Similarity between proteins can be easily quantified using APSA, as was demonstrated for the roll architecture of proteins ubiquitin and spinach ferredoxin. A twenty-by-twenty comparison of all  $\alpha$  domains showed that the curvature-torsion patterns generated by APSA provide an accurate and meaningful similarity measurement for secondary, super secondary, and tertiary protein structure. APSA is shown to accurately reflect the conformation of the backbone effectively reducing three-dimensional structure information to two-dimensional representations that are easy to interpret and understand.

Proteins 2009; 76:418–438.  
© 2008 Wiley-Liss, Inc.

**Key words:** secondary structure; automated protein structure analysis; curvature; torsion.

## INTRODUCTION

A qualitative and quantitative understanding of protein structure is an essential requirement for unraveling the relationship between protein shape and protein functionality. Numerous investigations have been carried out for this purpose.<sup>1–13</sup> At the more qualitative level, the ribbon representations, made popular by Richardson,<sup>1</sup> have given a visual entry to protein structure. The task of bringing these representations from the qualitative to the quantitative level of understanding requires a tedious analysis of conformational features and their representation in three-dimensional (3D) space in form of symbolic or mnemonic devices. Attempts in this way that describe a specific fold with prior knowledge of its shape and properties do not fulfill the objective of finding a general concept of protein structure directly. Among such investigations is one that describes viral capsid jellyroll topology as wedges<sup>2</sup> and another that obtains orientation angles for the TIM-barrel motif from seven domains.<sup>3</sup> There are other studies that provide detailed accounts of the various types of arrangements of helices<sup>4,5</sup> and  $\beta$ -strands<sup>5</sup> in folds. Though these descriptions throw light on the folding and function of a specific set of proteins, they use different approaches and levels of simplification, preventing their use for automated analysis and classification of proteins in general.

Among those methods that do classify all proteins in the Protein Data Bank (PDB)<sup>6</sup> many are not fully automated, such as CATH<sup>7</sup> and SCOP<sup>8</sup> that require manual intervention for analysis and decision-making. Some of the fully automated methods use more than one criterion [for example, the secondary STRuctural IDentification (STRIDE) method<sup>9</sup> uses  $\phi, \psi$  angles and hydrogen bonding] or arbitrary parameters [for example, the Dictionary of Secondary Structure of Proteins (DSSP)<sup>10</sup> works with arbitrary energy cut-offs that determine the presence of hydrogen bonds] as the basis of analysis. The analysis of the 3D position of individual backbone points such as the  $C_{\alpha}$ -atoms using distance masks (DEFINE<sup>11</sup>) or distance matrices (among other criteria)<sup>12</sup> implies that discrete sets of data points miss important features of protein structure. If the  $\phi$  and  $\psi$  backbone dihedral angles<sup>13</sup> are used as discrete parameters, the nontrivial task emerges to translate a multitude of dihedral angles into a general conformational concept for the purpose of understanding protein structure and functionality. Again, this task has so far not

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

Grant sponsor: NSF; Grant number: CHE 071893.

\*Correspondence to: D. Cremer, 3601 Pacific Ave, Stockton, CA 95211. E-mail: dcremer@pacific.edu

Received 17 July 2008; Revised 17 November 2008; Accepted 16 December 2008

Published online 23 December 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22357

been satisfactorily solved. Hence, a simple, fully automated method that accurately reflects the conformation of the entire polypeptide chain, is easy to interpret, and relates to the 3D shape of the protein is needed.

Recently, we presented a new method for the automated protein structure analysis (APSA) that is based on a two-step approach of describing and categorizing conformational features of proteins.<sup>14</sup> (i) The protein backbone is simplified to a smooth, continuous line in 3D-space. (ii) The curving and twisting of the backbone line is quantified by the curvature and torsion functions  $\kappa(s)$  and  $\tau(s)$ , respectively, where the parameter  $s$  gives the arc length of the backbone line. The diagrams of  $\kappa(s)$  and  $\tau(s)$  adopt typical patterns that make identification of protein secondary structural units easy.<sup>14</sup> In addition, they quantitatively identify all deviations and distortions from the ideal and provide an easy classification and identification of nonregular structural features. A curvature or torsion peak representing the conformation of a residue in a protein reflects also conformational features of the neighboring residues. This complies with the fact that it takes more than one residue (represented in APSA by a  $C_\alpha$  atom) to determine local shapes such as  $\alpha$ -helix and  $\beta$ -strand. APSA works on this principle. Therefore, in the  $\kappa(s)$  and  $\tau(s)$  diagrams, an “ideal helix peak” of a particular  $C_\alpha$  atom reflects the ideal (or close-to-ideal) helix arrangement of the two neighboring  $C_\alpha$  atoms as well thus constituting an ideal conformational environment.

A search for amino acids in ideal conformational environments showed that only 63% of all residues in  $\alpha$ -helices and 49% in  $\beta$ -strands comply with this conformational criterion. This discrepancy between the total number of secondary structural units identified in proteins and the number of ideal helices and  $\beta$ -strands is partly the reason for disparities that occur among the secondary structure assignments of several automated methods discussed in literature.<sup>15</sup> We also demonstrated that how the extended and helical nature of turns is accurately described and identified with the help of their  $\kappa(s)$  and  $\tau(s)$  diagrams.<sup>14</sup> Thus, APSA was shown to be a qualitative as well as quantitative tool for protein structure analysis that projects the 3D conformational features into 2D representations.

In this work, APSA is applied to a set of 77 natural proteins with the objective of quantitatively describing distortions and deviations of helices and  $\beta$ -strands from their ideal conformations. This involves the analysis and categorization of helix caps, entry and exit points of secondary structural units, kinks, bends, and breaks on the basis of the  $\kappa(s)$  and  $\tau(s)$  diagrams. In this connection, the speed of automation, the reliability of the secondary structure assignment, and APSA's versatility in describing varied backbone conformations from diverse proteins will be tested. Throughout the investigation APSA assignments will be compared with DSSP,<sup>10</sup> which is a widely

accepted secondary structure assignment method. A single protein (ubiquitin) will be selected from the set of 77 proteins to demonstrate the application of APSA in detail with respect to the characterization of all secondary structure and turn residues. Similar features seen between ubiquitin and spinach ferridoxin from the  $\kappa(s)$  and  $\tau(s)$  diagrams will guide the way for a simple and effective protein structure comparison based on the treatment of proteins in form of continuous conformational patterns rather than a set of discrete conformational parameter points.

## COMPUTATIONAL PROCEDURES

As described in Ref. 14, APSA is based on the representation of the protein backbone in form of a regularly parameterized smooth curve in 3D space. For this purpose, a coarse-grained image of the backbone is constructed where each residue is represented by its  $C_\alpha$ -atom. These positions are used as anchor points in 3D-space and are connected by a cubic spline function. The cubic spline gives the simplest parameterization of the backbone (compared to higher spline functions); it is computationally robust and easy to implement. Using the methods of differential geometry, the backbone line is described by means of three scalar parameters, curvature  $\kappa$ , torsion  $\tau$ , and the arc length  $s$ . The functions  $\kappa(s)$  and  $\tau(s)$  are generated by APSA for each protein from its coordinates taken from the PDB.<sup>6</sup>

As shown in Ref. 14, curvature and torsion values calculated from the spline are not sensitive to the uncertainties in the atomic coordinates as long as the resolution of the X-ray structural analysis is equal or smaller than 2 Å. The mathematical and physical aspects of the APSA protocol were found to reasonably represent the details of structure and also include global features such as chirality and orientation of structural units in 3D-space. For technical details relating to quantification of sensitivity and properties of the spline fit, we refer to Ref. 14.

A set of 77 proteins (78 chains) listed in the Supporting Information was selected from the PDB<sup>6</sup> including proteins from the four classes of the CATH classification system<sup>7</sup> i.e., “mainly alpha,” “mainly beta,” “mixed  $\alpha$ -beta,” and “few secondary structures.” Only X-ray structures having a resolution of 2.0 Å<sup>-1</sup> or less were selected. Proteins having breaks in the structure, missing amino acids or alternate locations for  $C_\alpha$  atoms were avoided. The proteins used for the APSA description are of different sizes with differing lengths of helices,  $\beta$ -sheets, and loop regions. They have one or more domains on single or several chains and in addition, are monomer or parts of multimeric structure. In the final dataset, the  $\alpha$  class mainly includes 26 different proteins (and 28 domains), the  $\beta$  class mainly includes 24 other proteins (and 26 domains), and the  $\alpha$  and  $\beta$  class includes 23 new pro-

**Table I**

Determination of Working Ranges for Curvature  $\kappa(s)$  and Torsion  $\tau(s)$  Using 5  $\alpha$ -Helices and 8  $\beta$ -Strands with 10 Equidistant Spline Points Between Every  $C_{\alpha}$  Atom<sup>a</sup>

A. $\alpha$ -Helices							
PDB ID	Helix position	$\kappa$			$\tau$		
		Avg	Min	Max	Avg	Min	Max
1A6I	128–148	0.39	0.29	0.59	0.15	0.08	0.22
1BJZ	130–148	0.39	0.29	0.60	0.15	0.07	0.23
1R4M(B)	23–29	0.40	0.30	0.62	0.15	0.08	0.21
1R4M(B)	426–439	0.40	0.29	0.62	0.14	0.07	0.21
1SOD	589–604	0.39	0.29	0.66	0.15	0.07	0.23
Overall	76	0.39	0.29	0.66	0.15	0.07	0.23
B. $\beta$ -Strands							
PDB ID	Strand position	$\kappa$ (Min)	$\kappa$ (Max)	$\tau$ (Min)	$\tau$ (Max)		
1UYL	78–81	<0.07	0.76–1.3	<–2.1	–0.03 to –0.08		
1UYL	89–93	<0.06	0.63–1.2	<–1.8	–0.01 to –0.07		
1ITV	18–21	<0.09	0.85–1.0	<–1.8	–0.09 to –0.07		
1ITV	26–31	<0.18	0.5–0.9	<–1.2	–0.11 to –0.004		
1ITV	74–78	<0.06	0.5–1.0	<–2.8	–0.08 to –0.006		
2PCY	25–30	<0.14	0.7–1.3	<–1.2	–0.12 to –0.07		
2PCY	37–42	<0.18	0.5–1.2	<–0.8	–0.15 to –0.05		
Overall		<0.14	0.4–1.3	<–0.8	–0.15 to –0.004		
1V86	2–6	<0.11	0.48–0.7	>+0.97	+0.02 to +0.14		

<sup>a</sup>Protein structures investigated are given by their PDB identification (ID) number and the residue numbers.

The terms *min* and *max* denote the smallest and largest  $\kappa(s)$  or  $\tau(s)$  values, *avg* the average of all 10 values calculated.

teins (and 30 domains). Two new proteins (and three domains) are included under the few secondary structures class for insights into any standard conformations assumed by these domains. Some popular architectures are represented by a greater number of domains, like the orthogonal bundle, though rare architectures such as the box under the  $\alpha$  and  $\beta$  class are also considered. In addition, various ratios of helices to  $\beta$ -sheets are represented within each architecture. In some cases, sets of identical or very similar proteins are purposely included in the analysis for similarity comparisons and so are some proteins with distinct supersecondary motifs.

## RESULTS AND DISCUSSIONS

In Table I, the average, minimum, and maximum  $\kappa$  and  $\tau$  values of five  $\alpha$ -helices (leaving out the N-terminal and C-terminal residue) and eight  $\beta$ -strands, all of them are free of specific distortions, are recorded. All but one of the  $\beta$ -strands chosen have negative torsion values indicative of a left-handed torsion along the  $\beta$ -strand.<sup>14</sup> The eighth  $\beta$ -strand is an example for one with a right-handed torsion. The average  $\kappa$  values for the  $\alpha$ -helices were determined from 10 equidistant points located along the protein backbone between two successive  $C_{\alpha}$  atoms where the latter were included into the set of equidistant points (they are not simply the average of minimum and maximum value).

Utilizing the values listed in Table I, nine ranges of  $\kappa(s)$  and  $\tau(s)$  values arranged in four “windows” were set up to create rules for automated structure recognition (Table II, Scheme 1). The values in Table II were found to strike the right balance by accounting for irregular boundaries of secondary structures without losing geometric details. For  $\alpha$ -helices, the working values differed from the ideal values obtained from earlier evaluations,<sup>14</sup> wherein the  $\kappa(s)$  and  $\tau(s)$  ranged from 0.3 to 0.56  $\text{\AA}^{-1}$  and 0.08 to 0.19  $\text{\AA}^{-1}$ , respectively [Fig. 1(a), Table II]. These ranges have been relaxed for natural helices such that  $\kappa(s)$  ranges from 0.23 to 0.67  $\text{\AA}^{-1}$  and  $\tau(s)$ , from 0.05 to 0.24  $\text{\AA}^{-1}$  (window 2, Table II). For example, the body of the helix in 1U4G starting at leucine 135 [ $\kappa(s)$  and  $\tau(s)$  diagrams of Fig. 1(b)] shows deviations from ideal  $\alpha$  helical values not sufficiently significant to be considered as a special case of distortion. [Slightly distorted helices are shown in Fig. 1(c–m).]

The  $\kappa$  and  $\tau$  values of the first amino acid are different from those of the body of the helix [Fig. 1(b)], which is considered by defining window 1 for the “starter” residue (Table II). The high values of window 1 reflect the fact that the backbone enters into the helix from a relatively straight region by veering sharply into it. Similarly, the  $C_{\alpha}$  atom of the last amino acid belonging to the helix exit (Scheme 2) is at the centre of the smooth transition from a curved helical segment into the relatively straight segment of the following backbone [Fig. 1(b)]. The first  $C_{\alpha}$  point may either lie toward the body of the helix, in which case it has a positive  $\tau$  value, or may lie away

**Table II**Specification of  $\kappa(s)$ ,  $\tau(s)$  Windows for Helices and  $\beta$ -Strands Used for Automation and Comparison with Ideal Helices and Extended Structures<sup>a</sup>

Automation details windows Wn	Working values [ $\text{\AA}^{-1}$ ]	Ideal values [ $\text{\AA}^{-1}$ ]	Minimum length [# of residues]
W1: $\alpha$ -Helix (starter residue)	$0.2 \leq \tau (\text{max}) \leq 0.4$	$0.3 \leq \kappa \leq 0.56$	4
W2: $\alpha$ -Helix; body	$0.25 \leq \kappa \leq 0.67$ $0.05 \leq \tau \leq 0.24$	$0.08 \leq \tau \leq 0.18$	
W3: $\beta$ -Strand: negative $\tau$ : troughs (left-handed)	$0.5 \leq \kappa \leq 1.4$ $0.0 \leq \kappa(\text{min}) \leq 0.02$ $-0.001 \leq \tau(\text{max}) \leq -0.15$ $\tau (\text{min}) < -0.75$	$0.01 \leq \kappa \leq 1.0$ , $\tau (\text{min}) < -2.9$	3
W4: $\beta$ -Strand: positive $\tau$ : peaks (right-handed)	$0.4 \leq \kappa(\text{max}) \leq 0.9$ $0.0 \leq \kappa(\text{min}) \leq 0.02$ $0.001 \leq \tau (\text{min}) \leq 0.15$ $\tau (\text{max}) > 0.75$	$0.01 \leq \kappa \leq 1.0$ , $\tau (\text{max}) > 2.9$	3

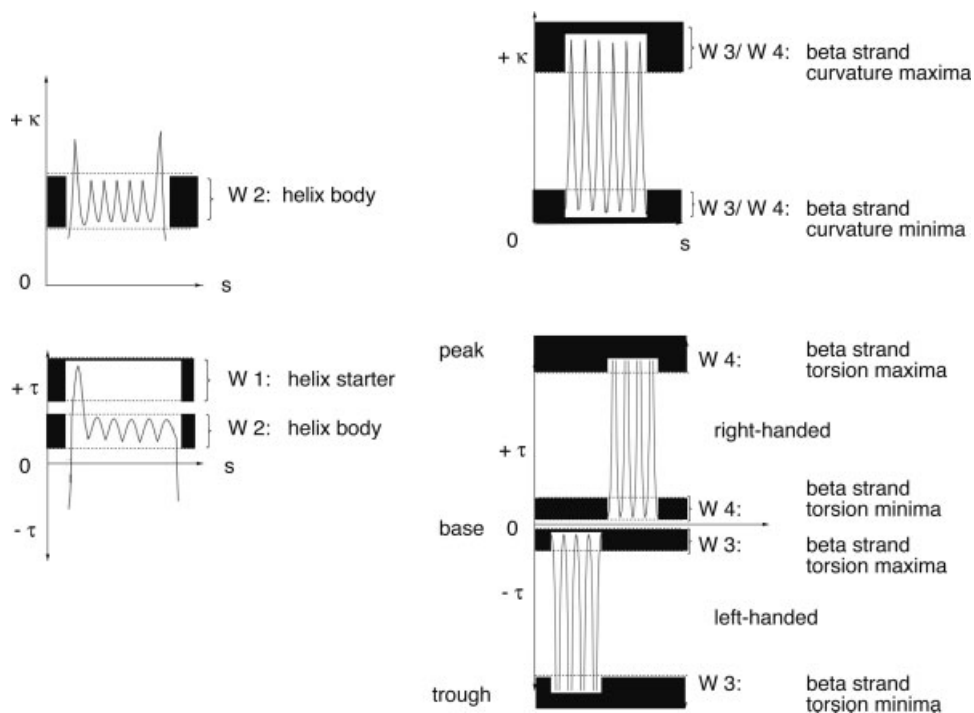
<sup>a</sup>The terms *min* and *max* denote the smallest and largest  $\kappa(s)$  or  $\tau(s)$  values in the range from one  $C_{\alpha}$  atom to the next  $C_{\alpha}$  atom.

from the helix axis, when it shows  $\tau$  values changing from negative to positive (see also Helix entries and exits Section). Both cases lead to  $\tau_{\text{max}} < 0.4 \text{ \AA}^{-1}$  (Table II).  $\kappa$ -Values are not included into Window 1 because they are too unspecific to facilitate identification of the helix starter residue.

Naturally occurring  $\beta$ -strands are mostly twisted or bent and seem to be influenced easily by the surrounding turns and structures. This is especially true in the case of the  $\beta$  sheet occurring in folds such as the roll or the  $\beta$ -barrel. These effects are clearly reflected in the  $\kappa$ - and  $\tau$ -pattern of naturally occurring  $\beta$ -strands (Windows 3 and

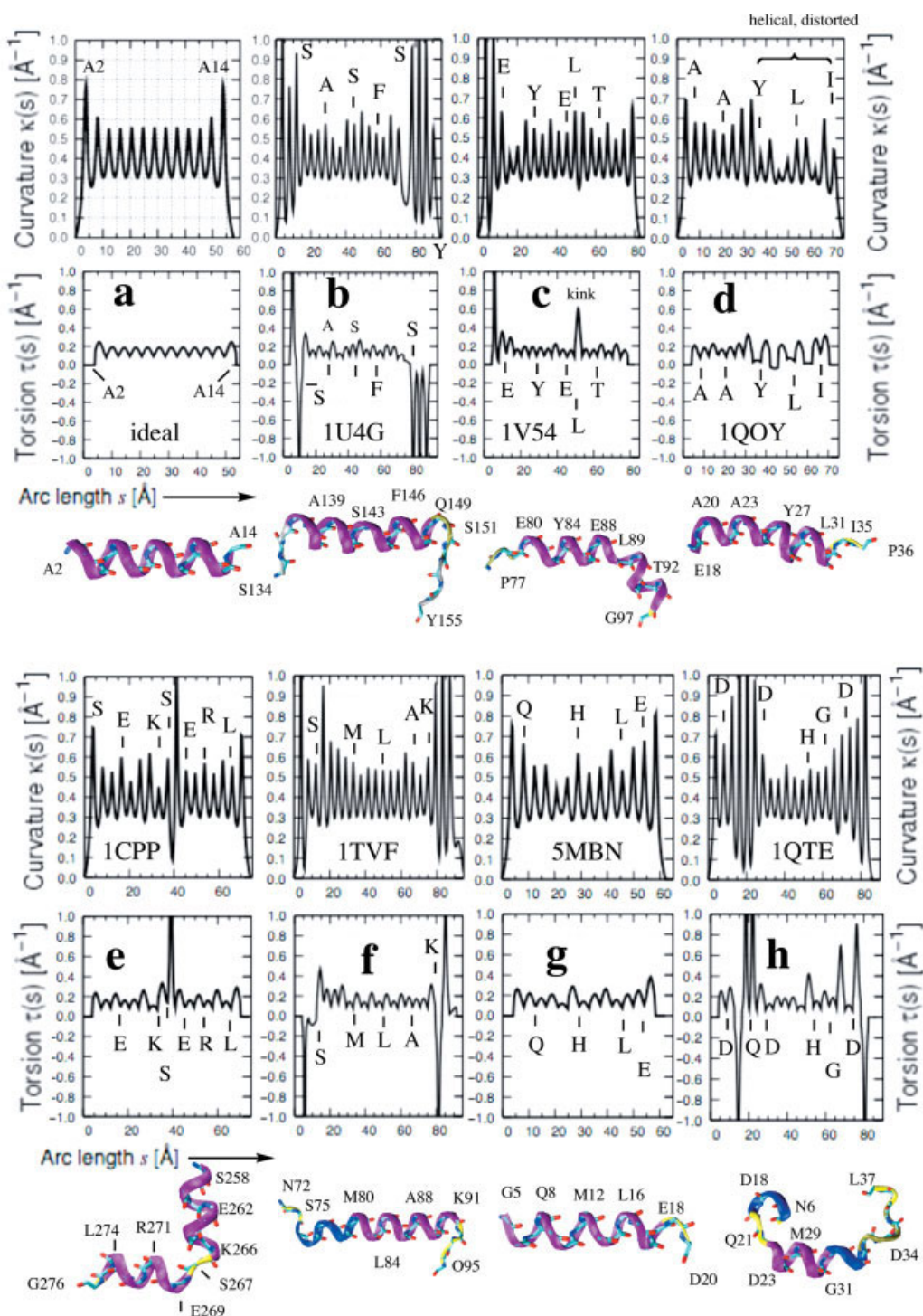
4). The  $\kappa(s)$  peak lengths are large (larger than those of a helix) thus yielding higher peaks ( $0.5\text{--}1.4$ ; helices:  $\kappa < 0.65 \text{ \AA}^{-1}$ , Table II) and a much lower base (see Scheme 1) with values close to zero (helices:  $\kappa > 0.25 \text{ \AA}^{-1}$ ). For the purpose of distinguishing the curvature of  $\beta$ -strands from that of helices, a split window is used (Scheme 1, Table II). It is noteworthy that for the ideal left-handed  $\beta$ -strand [Fig. 1(n)], the curvature values are  $\leq 1.0 \text{ \AA}^{-1}$  (Table II).

The  $\tau(s)$  peaks of the  $\beta$ -strands are also recognized by their base and tip values tested by a split  $\tau$ -window (Tables I and II) where one part accounts for the base

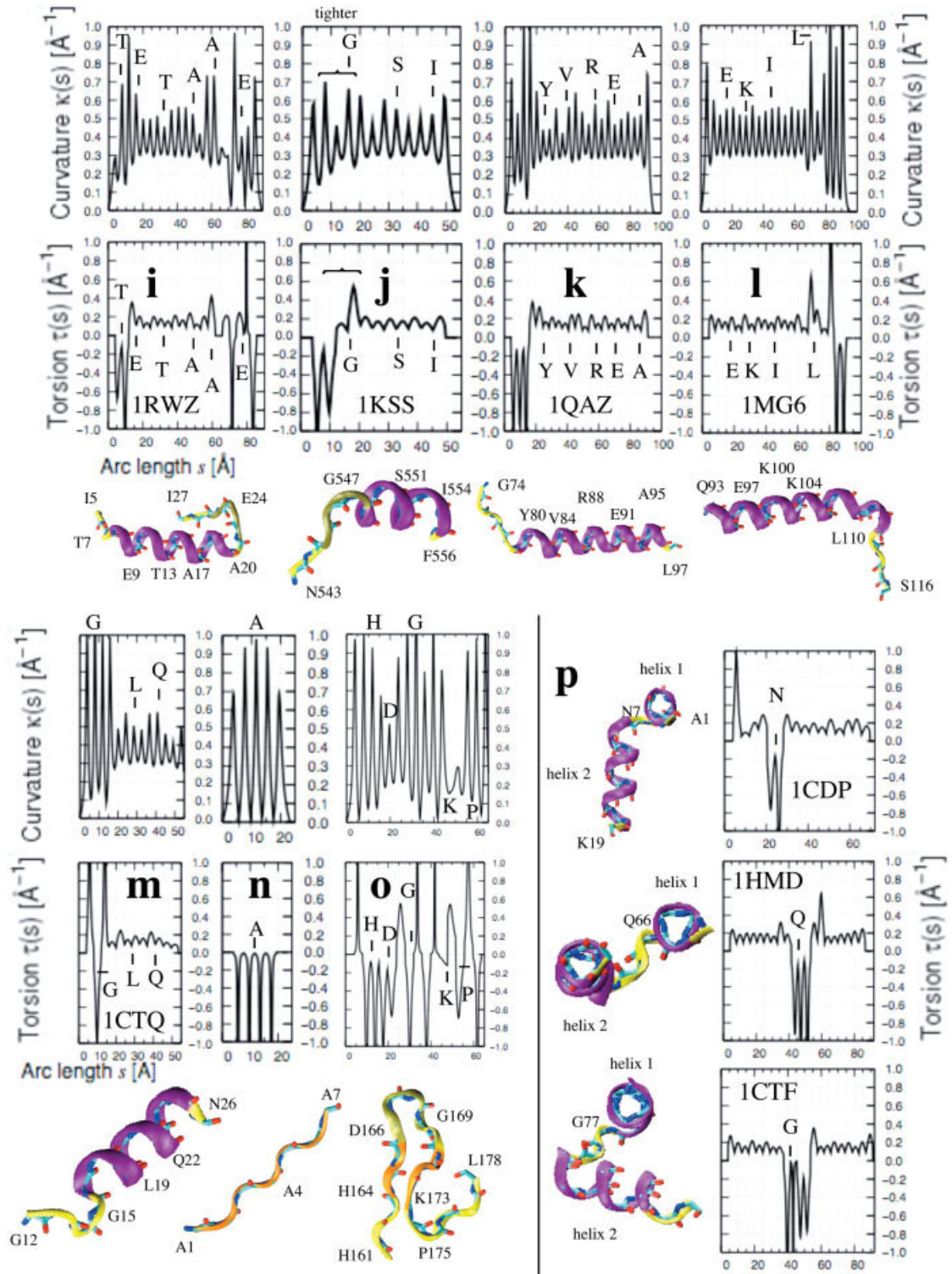
**Scheme 1**

The four windows W1, W2, W3, and W4 are schematically shown presenting the curvature  $\kappa(s)$  and torsion ranges of  $\tau(s)$  for each Wn, the peak (trough) forms, and the terms used in the text for describing the windows.



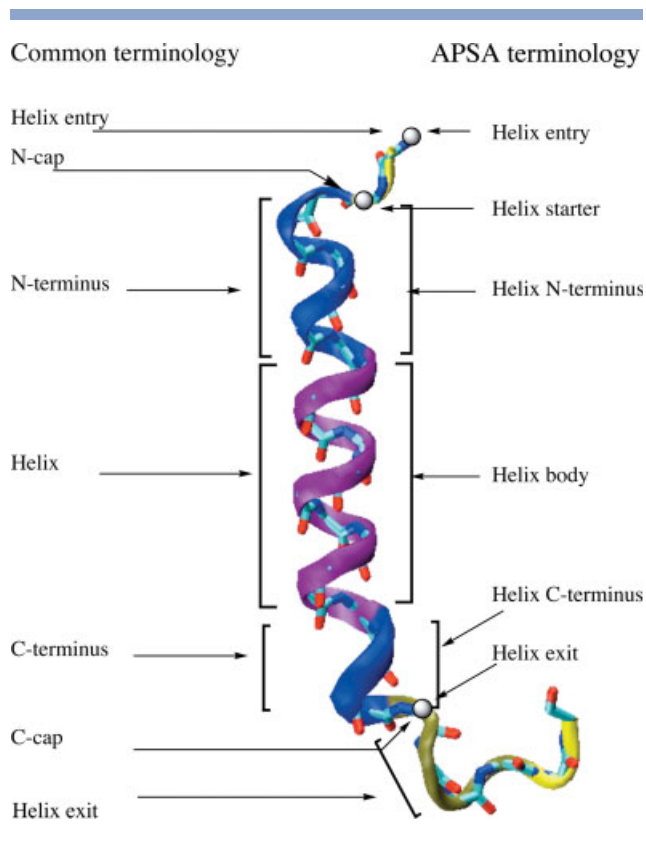
**Figure 1**

Curvature (above) and torsion diagrams (below),  $\kappa(s)$  and  $\tau(s)$ , for typical secondary structural units given in form of ribbon presentations. Every peak in a diagram reflects the conformation of a residue in the analyzed segments. (a) Ideal 14-residue long polyalanine  $\alpha$ -helix. (b) Natural  $\alpha$ -helix with small irregularities (1U4G: residues L131–Y155). (c) A kink in an  $\alpha$ -helix (1V54: P77–G97). (d) Distortions of an  $\alpha$ -helix leading to a looser N-terminus (1QOY: E18–P36). (e) A strong kink leading to a large  $\tau$ -value in a slightly distorted  $\alpha$ -helix (2CPP: S258–G276). (f) Difference between body and N-terminus of an  $\alpha$ -helix (1TVF: N72–S95). (g) An  $\alpha_\pi$  cap at the C-terminus of an  $\alpha$ -helix leading to higher curvature (5MBN: G5–D20). (h) Transition from a helix to a turn region with gradually increasing curvatures and interspersed high torsions (1QTE: W17–L37). (i) C-terminal caps of  $\alpha$ -helix (1RWZ: I5–I27). (j) N-terminal caps of  $\alpha$ -helix (1KSS: N543–F556). (k) N-terminal caps of  $3_{10}$ -helix (1QAZ: G74–L97). (l) C-terminal caps of  $3_{10}$ -helix (1MG6: Q93–S116). (m) N-terminal caps of  $\alpha$ -helix with positive entry (1CTQ: G12–N26). (n) Ideal four-residue long polyalanine  $\beta$ -strand. (o) Distorted  $\beta$ -strand (1RIE: H161–L178). (p) Degree of rotation of a second helix with regard to a first as reflected by the number of  $\beta$ -troughs in the  $\tau$ -diagram of the connecting turn [1CDP 1–19 to 1HMD 55–66 to 1CTF 67–90 (below)]; each additional  $\beta$ -trough indicates a left-handed  $90^\circ$ -rotation of the second helix.



**Figure 1**  
(Continued)



**Scheme 2**

A hypothetical helix with an  $\alpha$ -helix body and two caps at either end is shown. The helix is started by the “starter,” ended by the “exit” and the “entry” is defined by the residue just prior to the starter. A distinct 3D structure formed by the few residues into the helix from either end is termed as a “terminus” by APSA. On the left side the terms are given that are given in the literature<sup>1</sup> where helix entry and helix exit are added according to Efimov.<sup>28</sup> The APSA terminology tries to follow the terminology used in the literature however considers at the same time the exact definition of terms via curvature and torsion.

with values close to zero and the other for the extremes. In addition, one has to consider the sign of the  $\tau(s)$ -peaks, which indicates a left-handed ( $-$  sign, troughs; Window 3 in Scheme 1, Table II) or right-handed strand ( $+$  sign, peaks; Window 4). The troughs of the negative  $\tau(s)$  have so low-lying minima that it is sufficient to give just an upper boundary of  $\tau$  ( $-0.75 \text{ \AA}^{-1}$ , Table II). The positive peaks of the right-handed  $\beta$ -strands (Window 4) have somewhat different  $\kappa$ -ranges ( $0.4\text{--}0.9 \text{ \AA}^{-1}$ ) and a different  $\tau$ -window. The values of the peak bases are found between  $0.001$  and  $0.15 \text{ \AA}^{-1}$  and the peak maxima are  $>0.75 \text{ \AA}^{-1}$  (Table II, Scheme 1).

Additional windows can be defined for left-handed  $\alpha$ -helices, right-handed  $3_{10}$ -, and  $\pi$ -helices. However, because of the fact that these structures occur relatively seldom in the ideal conformations, we refrain at this stage from setting up suitable  $\kappa$  and  $\tau$  ranges. Instead, we operate with the curvature and torsion values obtained for the ideal structures described in our previous work.<sup>14</sup>

## Application of the APSA windows

A summary of all  $\alpha$  helices and  $\beta$  strands found among the 77 proteins (78 chains) investigated (see also Computational procedures Section and Supporting Information) is presented in Table III. Special forms of secondary structural units are shown in Figure 1(a–p), which contain the calculated  $\kappa(s)$  and  $\tau(s)$  diagrams and a VMD<sup>16</sup> representation of the 3D structure. The 77 proteins possess a total of 547  $\alpha$ -helices and 656  $\beta$ -strands according to DSSP assignments<sup>10</sup> that are based on hydrogen bonding patterns. APSA identifies 543 helices and 654  $\beta$ -strands where the numbers do not necessarily indicate a close match with the DSSP assignments. As detailed in Table III, differences are found for 20 helices and 46  $\beta$ -strands, which leads to an agreement in 96 and 93% of all helical and  $\beta$ -strand environments, respectively.

APSA, contrary to DSSP, clearly identifies all distorted shapes via their graphic patterns of  $\kappa$  and  $\tau$  values falling outside the strict windows of Table II. For example, 20 split or kinked helices are recognized, though in these cases DSSP had labeled them as one continuous helix (Table III). The distortion in  $\tau(s)$  at leucine 89 in 1V54 corresponds to a change in the helix orientation as is confirmed by the ribbon representation [Fig. 1(c)]. Four DSSP-labeled H-bonded turns appear  $\alpha$ -helical in the  $\kappa$ - $\tau$  diagrams and so do other regions that have no secondary structures assigned. Twelve structures recognized as “ $\alpha$ -helices” are found to deviate from the regular pattern and hence, are commented as being “distorted” in Table III, though they show overall helical shapes [see, e.g., Fig. 1(d)]. Structures identified as “ $3_{10}$ -helices” by DSSP do not possess a unique APSA pattern, which is in line with descriptions of variable  $3_{10}$ -helix geometries in peptides as given in the literature.<sup>1,17</sup>

APSA finds 654 undistorted  $\beta$ -strands. These correspond to  $\beta$ -strands alone; “isolated  $\beta$ -bridges” of DSSP are not included for reasons of simplicity (excluding the need to analyze loops and turns). Twenty-two new  $\beta$ -strands are found (with start and end residue ranges slightly shifted along the backbone; Table III) and in 22 cases, the strands are geometrically distorted because of different bending and twisting of amino acids not typical of ideal  $\beta$ -strands [see Fig. 1(o) for a distorted  $\beta$ -strand in 1RIE]. In two situations, adjacent strands merged into one continuous strand resulting in the loss of one strand in each case.

## The APSA description of helices and their distortions

The  $\kappa(s)$  and  $\tau(s)$  diagrams precisely reflect the range of distortions for the helices investigated. There are some regions of the backbone where the geometry, though close to an  $\alpha$ -helix, is intermediate between the  $\alpha$  and  $3_{10}$  conformation. Such distortions can occur both in the

**Table III**APSA Results for a Dataset of 77 Proteins<sup>a</sup>

CATH Architecture	S. No	PDB ID	# of $\alpha$ DSSP	# of $\alpha$ APSA	Comments	# of $\beta$ DSSP	# of $\beta$ APSA	Comments
Class: mainly $\alpha$								
$\alpha$ Horse-shoe	1	1M8Z(A)	28	28		0	0	
	2	1QTE(A01)	37	37	1(Split)	0	2	2 Strand+
	3	1V54(E)	5	5	1(Split)	0	1	1 Strand+
$\alpha$ Solenoid	4	1PPR(M)	16	16	2(Split)	0	0	
$\alpha/\alpha$ Barrel	5	1QAZ(A)	12	12		0	2	2 Strand+
Orthogonal bundle	6	1ECA(A)	8	8		0	0	
	7	1GM8(B03)	14	14	1(Split)	42	41	1 Strand-
	8	1HC0(A)	7	7		3	3	
	9	1KSS(A02)	20	20	1(Split)	19	19	
	10	1LMB(3)	5	5	1(Split)	0	0	
	11	1NG6(A)	7	7		0	0	
			1QTE(A02,3)					
	12	1U4G(A02)	8	8		10	10	
	13	1UTG(A)	4	4		0	0	
	14	2CPP(A)	19	18	2(Split), 1(distort)-	14	12	2 strand-
	15	2CTS(A)	20	20	4(Split)	2	2	
16	2LZM(A)	10	10	1(Split)	3	3	1strand+, 1strand-	
17	2MHB(A)	8	8	1(Split)	0	0		
18	3WRP(A)	6	6		0	0		
19	5MBN(A)	9	9		0	0		
20	5PAL(A)	7	7		2	2		
Up and down bundle	21	17GS(A02)	10	10		4	4	
	22	1AA7(A)	10	9	1(Distort)-	0	0	
	23	1MG6(A)	4	4		2	2	
	24	1O83(A)	6	6		0	0	
	25	1QOY(A)	9	8	1(Distort)-	2	2	
	.	1V54(A)	21	20	3(Split), [2]	1	1	
26	1VKE(B)	5	5		0	0		
Class: mainly $\beta$								
3 Solenoid	27	1EZG(A)	0	0		6	6	
	28	1QRE(A)	2	2		22	24	2 Strand+
3 Layer Sandwich	29	1NYK(A)	1	1		11	9	1 Strand-, [2 Strand]
	30	1RIE(A)	1	2	1+ (:3 <sub>10</sub> )	10	9	1 Strand-
4 layer Sandwich	.	1GM8(B01)						
4 Propeller	31	1ITV(A)	3	4	1+	17	17	
$\beta$ Barrel	32	1EY0(A)	3	3		8	7	1 Strand-
	33	2POR(A)	3	3		16	16	
	34	4PEP(A)	7	7		24	21	3 Strand-
	35	1AQ2(A02)	16	17	1(:H-bo Turn)+	28	28	
$\beta$ Complex	36	1TGX(A)	0	0		5	5	
Ribbon	37	1G79(A)	6	6		10	9	1 Strand-
Roll	38	1GCQ(A)	0	0		5	5	
	.	1GM8(B02)						
39	1TVF(A02)	11	10	1(Distort)-	16	16		
40	1ZX6(A)	0	0		5	5		
Orthogonal prism Sandwich	41	1B2P(A)	0	0		12	12	
	42	1GCS(A)	1	1		14	13	1 Strand-
	43	1REI(A)	0	0		10	10	
	44	2AZA(A)	2	2	1(Cap)+, 1(distort)-	8	8	
	45	2PAB(A)	1	1		9	9	
	46	2PCY(A)	0	0		8	8	
	47	2SOD(O)	0	0		9	8	1 Strand-
Single sheet	49	7RXN(A)	0	0		3	2	1 Strand-
Trefoil	50	1WBA(A)	0	0		13	11	1 Strand-
	51	2FGF(A)	0	0		10	10	
Class: $\alpha$ and $\beta$								
2 Layer sandwich	52	1B4V(A02)	12	11	1(Distort)-	19	19	1(split)+
	53	1B8S(A02)	12	11	1(Distort)-	19	19	
	54	1CF3(A03)	17	16	1(Distort)-	20	20	2 Strand-
	55	1CRN(A)	3	3		2	2	
	56	1CTF(A)	3	3		3	3	
	57	1TGS(I)	NA	1	1 Helix +	NA	2	
	58	2CI2(I)	1	1		3	4	1 Strand+

(Continued)



**Table III**  
(Continued)

CATH Architecture	S. No	PDB ID	# of $\alpha$ DSSP	# of $\alpha$ APSA	Comments	# of $\beta$ DSSP	# of $\beta$ APSA	Comments	
3 Layer (aba) sandwich	.	17GS(A01)							
	59	1CTQ(A)	6	7	1(Split), 1(Cap)+	6	6		
	.	1TVF(A01)							
	60	1WPU(A)	4	4		4	4		
	61	2AK3(A)	8	8	1(Split)	7	7		
	62	2FOX(A)	5	6	1+	6	6		
3 Layer (bba) sandwich	63	5CPA(A)	9	9		8	12	4 Strand+	
	.	1B4V(A01)							
	.	1B8S(A01)							
	.	1CF3(A01)							
	.	1KSS(A03)							
	64	3GRS(A)	14	14		23	23	2 Strand+, 2 Strand-	
	65	1H61(A)	11	11		14	12	2 Strand-	
	$\alpha$ - $\beta$ Barrel	66	1B8P(A02)	12	11	1(Distort)-	14	13	1 Strand-
	$\alpha$ - $\beta$ Complex	67	1F7L(A)	4	4		5	5	
	.	1KSS(A01)							
68	2CDV(A)	4	3	1(Distort)-	4	3	1 Strand-		
69	3HSC(A)	12	13	1:(H-bo Turn)+	18	19	1 Strand+		
70	7AAT(A)	16	16		13	13			
71	9PAP(A)	5	5		8	8			
$\alpha$ - $\beta$ Horseshoe	72	1OZN(A)	2	1	1(Distort)-	18	17	1 Strand+, 1 strand-, [2 strand]	
Box	73	1RWZ(A)	4	4		18	19	1 Strand+	
Roll	.	1U4G(A01)							
74	1UBQ(A)	1	2	1(:3 <sub>10</sub> )	5	5			
75	2CA2(A)	4	3	1(Distort)-	15	15			
76	2CAB(A)	3	3		16	16			
Class: few secondary structures									
Irregular	.	1CF3(A02)							
77	1HIP(A)	2	2		3	4	1 strand+		
78	5PTI(A)	1	1		2	4	2 strand+		
			547	543		656	654		

<sup>a</sup>PDB ID denotes the Protein Data Bank Identifier including chain and domain IDs where appropriate. The symbols # of  $\alpha$  and # of  $\beta$  denote the number of  $\alpha$ -helices and  $\beta$ -strands, respectively, recognized by either DSSP (Dictionary of Secondary Structure of Proteins [10]) or the APSA (curvature-torsion based) method. The secondary structures are also commented as being split, distorted, and labeled as a "hydrogen-bonded turn" (H-bo Turn) by DSSP; an  $\alpha$ -helical cap (cap) is labeled as 3<sub>10</sub> helix by DSSP (:3<sub>10</sub>). The + (-) signs following each expression indicate that the secondary structure was added to (subtracted) from the total APSA count of  $\alpha$ - or  $\beta$ -structures. The symbol [.] indicates that two secondary structures are merged by APSA.

body of helices and toward their ends. Splits and kinks in the body of helices have been extensively studied and accounted in literature.<sup>18–20</sup> The degree of kink can be quantified using the  $\kappa$  and  $\tau$  values. An example is the helix between E80 and G97 in the E chain of bovine heart cytochrome C oxidase (1V54)<sup>6</sup> [Fig. 1(c)]. The corresponding  $\tau$  diagram identifies the amino acid (L 89) responsible for the kink. The height of the  $\tau$  peak (just over 0.5 A<sup>-1</sup>) indicates that the kink is still helical, but the  $\kappa$  and  $\tau$  values are close to those of a 3<sub>10</sub>-helix.<sup>14</sup> A more drastic kink, as in cytochrome P450 (2CPP), produces a corresponding strong disturbance in both  $\kappa(s)$  and  $\tau(s)$  [see Fig. 1(e)].

Helical distortions can be considered as regions where the backbone is still helical, but does not belong to the well-defined conformations of the 3<sub>10</sub>-,  $\alpha$ -, or  $\pi$ -helices. The  $\tau(s)$  diagrams of these regions are interspersed with extended peaks indicating stretching of the helix. These distended helices regions have traces of overall helicity and the coiling of the entire backbone into a helix becomes visible only as a global characteristic.

The analysis of secondary structures in proteins is often confronted with the problem of ambiguous boundaries. Early investigation<sup>1</sup> have documented that the ends of helices are different from the body. For this reason, secondary structure assignment methods must treat the amino acids belonging to these "cap"-like structures with some caution. For example, some dihedral angle-based methods<sup>21,22</sup> analyzed helices by discarding amino acids that took up any set of values lying outside predefined regions of the Ramachandran plot. A detailed geometry-based analysis of such regions would throw more light on this problem and also suggest a systematic and uniform way of classifying and handling them in future. This is possible using calculated  $\kappa$  and  $\tau$  values of these regions. In some cases, the difference between the body and the termini of a helix is so strong that it might be considered as a turn rather than as an extension of the helix whereas in other cases it may be very subtle [Fig. 1(f)]. Therefore, we will explicitly discuss this problem in Helix termini as described by APSA and in Helix entries and exits Sections.

### The $3_{10}$ -helix conformation

$3_{10}$ -helices were first reported in 1941.<sup>23</sup> The N termini of  $3_{10}$ - and  $\alpha$ -helices have been studied and compared<sup>24</sup> with specific amino acid propensities and preferences. The latter were related to functionality and probable progression of protein folding along the  $\alpha$ -helical axis.<sup>25</sup> It has been proposed<sup>1</sup> that the occurrence of  $3_{10}$ -conformations at the ends of helices serve the purpose of tightening the  $\alpha$ -helix from uncoiling and losing its orientation. It has also been documented that  $3_{10}$ -helices could smoothly uncoil into  $\alpha$ -helices and vice versa because the corresponding Ramachandran regions are allowed for this transformation.<sup>26</sup> This observation suggests the possibility of functional importance to these regions. Thus amino acids in a  $3_{10}$ -cap, whether at the N or C terminus of the helix, fulfill the purpose of a tighter coiling and stabilizing the ends of an  $\alpha$ -helix.

The  $3_{10}$ -helices occurring at the C termini of  $\alpha$ -helices can have an  $\alpha_{\pi}$ -conformation ( $\pi$ -conformation mixed into an  $\alpha$ -helix), with H-bonding resembling the  $\alpha$ -helix pattern and the slightly tilted conformation resembling the  $\pi$  helices.<sup>1</sup> For example, the region 8–17 of myoglobin (5MBN) has such an  $\alpha_{\pi}$ -character, which is confirmed by the corresponding  $\kappa$  and  $\tau$  patterns [Fig. 1(g)]. The difference between an  $\alpha$  and a  $3_{10}$  N-terminus is that in the former case  $\tau$  reaches up to  $0.4 \text{ \AA}^{-1}$  whereas in the latter case it ranges from 0.4 to  $0.56 \text{ \AA}^{-1}$ ,<sup>14</sup> thus reflecting the different rise per amino acid of both structures along the helix axis. From the  $\kappa(s)$  and  $\tau(s)$  diagrams of APSA, a smooth transition is often seen from the  $\alpha$ -helix through the  $3_{10}$ -helix into the extended regions of turns or  $\beta$ -strands. The  $\alpha$ -helix (3.6 amino acids per turn) possesses an average curvature peak length of  $0.56 - 0.3 = 0.26 \text{ \AA}^{-1}$  (Table II) and therefore is more relaxed than a  $3_{10}$ -helix (3 amino acids per turn) with an average  $\kappa$  peak length of  $0.81 - 0.28 = 0.53 \text{ E}^{-1}$ .<sup>14</sup> The transition from the  $3_{10}$ -helix conformation into the  $\beta$ -strand can be understood on the basis that the well-extended  $\beta$ -strand can be viewed as a helix with 2-amino acids per turn thus leading to higher  $\kappa$ -peaks than those of a  $3_{10}$ -helix (up to 1.0 compared to  $0.8 \text{ \AA}^{-1}$  in the latter case; Table II and Ref. 14). This trend can be partly seen in 1QTE [Fig. 1(h)] at the (positive)  $\tau$ -peaks corresponding to amino acid methionine 28, leucine 32, and aspartate 34.

### The $\pi$ -helix conformation

Though it has been known over the years that  $\pi$ -helices are rare, there are conflicting results<sup>27</sup> that indicate their occurrence to be as frequent as one out of every 10 helices in the PDB.<sup>6</sup> It is also discussed how H-bonding and amino acid preferences can be used to characterize  $\pi$ -helices and enumerates important associated functionalities such as specific ligand binding.<sup>27</sup> Some studies<sup>26</sup> consider the  $\pi$  (and the  $3_{10}$ -helix as folding intermediates in the

formation of the  $\alpha$ -helix; the  $\alpha$ - and the  $3_{10}$ -helices have been described to share a common initiation paths while folding.<sup>25</sup> In the set of 77 proteins investigated by APSA, a pure  $\pi$ -region was not observed although  $\pi$ -character was found to be mixed into some of the helix caps (see The  $3_{10}$ -helix confirmation Section).

### Helix termini as described by APSA

In literature<sup>24</sup> the term helical cap is used for the last helical amino acid, whereas helix terminus (and in other literature<sup>1</sup> the same term cap) is used to denote a few amino acids towards the end of the helix (see Scheme 2) indicating that they are not always sharply defined. It should be noted that the terms cap, terminus, and end are used interchangeably and thus become loosely defined in literature. In the APSA investigation, the terms become equivalent because the spline fitting ensures that the  $\kappa(s)$ - and  $\tau(s)$ -functions at every amino acid reflect the conformation of the neighboring amino acids. Amidst all the discussion about the occurrence, distribution, property, and details of helices and helix caps, there is no systematic classification of these structures based on just the geometry. APSA considers caps as a special case of “distortions” occurring toward the termini of helices. From the  $\kappa(s)$  and  $\tau(s)$  diagrams of various protein segments it becomes evident that the cap at the terminus conformationally spreads over neighboring amino acids in either direction, and can be identified using torsion  $\tau(s)$  alone. Utilizing the APSA results, the termini can be broadly divided into three different types.

- i.  *$\alpha$ -Terminus*: This is a segment of  $\alpha$  helix broken off from its body. About three or four amino acids of the  $\alpha$ -helix are cut off from the rest and oriented toward a direction different from that of the helix.  $\alpha$ -Termini can show some standard distortions and resemble the  $\alpha$ -helix only by average  $\kappa$  and  $\tau$  values. They include the  $\alpha_{\pi}$ -type of structures [Fig. 1(g)].
- ii. *Tighter terminus*: Such a terminus has a larger  $\kappa$  value, thus including  $3_{10}$ -caps and the distortions that are narrower in diameter than an  $\alpha$ -helix loop. Some caps of mixed geometry are distorted with only the bare remnants of helicity resembling a completely stretched spring. In these cases, defining the cap and differentiating it from a loop region becomes difficult. The  $\kappa(s)$ - $\tau(s)$  diagrams reflect the true state of the backbone in a graphical way that aids the analysis and recognition of complicated patterns. Some examples of tighter termini are presented in Figure 1(i–l).
- iii. *Looser terminus*: This terminus is more relaxed with a larger  $\alpha$ -helix diameter and therefore includes a typical  $\pi$  cap or related distortions. Figure 1(d) shows the ending of the helix in 1QOY (hemolysin E) with a looser terminus starting at leucine 24. The larger diameter of the terminus increases the flexibility of the backbone to

some extent introducing alternating high and low  $\tau(s)$  values typical of helical yet more planar curves. Looser termini and  $\alpha$ -termini appear to occur much less frequently than tighter helix termini.

### Helix entries and exits

Some helix entries and exits have been described in literature based on  $\phi$ - $\psi$  values and amino acid properties.<sup>28,29</sup> By APSA, the  $C_\alpha$  atom of the amino acid prior to the starting of the helix is considered to be the “entry” (Scheme 2). The polypeptide chain can enter into the helix in either a left- or right-handed fashion. The left-handed entry is found more frequently [Fig. 1(b,f,i-k)] and is the point of chain reversal from strongly negative  $\tau(s)$  (left-handed torsion) through  $\tau(s) \sim -0.1 \text{ \AA}^{-1}$  at  $C_\alpha$  to positive  $\tau(s)$  values (right-handed helix torsion). The right-handed entry [Fig. 1(c,h,m)] leads to no chain reversal and therefore the  $\tau(s)$  remains positive. They have large values for curvature peak heights [with low minima; Fig. 1(m)] resembling the peaks of extended conformations, whereas the following  $\kappa$ -minima are relatively large and slightly helical giving the impression as if the helix has been stretched to increase its pitch [Fig. 1(m)].

Helix exits, much like the entries, can have positive or negative torsion, where again the latter are more frequent [Fig. 1(f,m)]. The positive exit in Figure 1(h,l) continues in the same overall direction of the helix whereas the negative exit appears to “peel away” from the helical formation [see inset of Fig. 1(b)].

### The APSA description of extended structures and their distortions

In Application of the APSA windows Section, we showed that the series of  $\tau$  peaks representing the  $\beta$ -regions can be either positive or negative where the sign gives the overall orientation of the strand in 3D (left- or right-handed twist). On a more detailed note, a  $\beta$ -strand could be considered to have “local” and “global” twisting. The “local” twist is given by the arrangement of  $C_\alpha$  atoms along the strand and the “global” twist refers to the twisting of the whole  $\beta$ -ribbon. Both local and global twisting of the strand contributes to the torsion, the former being dominant. The global twisting is relatively small and does not produce any noticeable impact on the overall torsion value. The sign of the strand itself is indicative of the direction it points in 3D with respect to the last point in the preceding structure (strand, turn, loop, helix).

Figure 1(p) shows three pairs of helices from different proteins and demonstrates the handedness of local twisting in  $\beta$ -strands. The first pair from parvalbumin (1CDP, 1–17) has two helices separated by two  $\beta$ -troughs, the second pair from hemerythrin (1HMD, 55–77) by three, and the third pair from ribosomal protein (1CTE, 67–90)

by four. For the addition of every  $\beta$ -trough, the second helix does not only undergo a translation, but also a rotation: the relative orientations of the helices reveals that the first would rotate into the second, which would rotate into the third in a left-handed fashion. An addition of one more  $\beta$ -trough in the turn region would point the second helix in the same direction as in 1CDP, hence indicating pattern repetition for the addition of every fourth  $\beta$ -trough. The positive  $\beta$ -peaks (not shown) were found to have the same effect in the opposite direction of rotation, confirming that extended regions have local twisting and are not flat ribbons. Application of APSA to extended regions reveals that they are separated by numerous one residue-long kinks that bend the strands by less than  $90^\circ$ , though these are sometimes considered as supersecondary structures.<sup>28</sup> It is interesting to note that a range of  $\tau$ -peaks can be obtained for all intermediate structures ranging from a planar  $90^\circ$  strand ( $\tau(s)$  close to 0, large  $\kappa(s)$ ) to a strand that is bent strongly out of plane of the  $\beta$ -ribbon (close to the torsion of a  $3_{10}$ -helix). When looking end-on (along the axis of a helix or  $\beta$ -strand), a helix looks like a circle and a  $\beta$ -strand like an ellipsis (rather than just a straight line as is often shown in textbooks for reasons of simplification). The plane of the  $\beta$ -ribbon refers to that defined by the strand axis and the major axes of the ellipsis.

### $\beta$ -Strand entries and exits

The positive entries into  $\beta$ -strands often have sharp reorientations of the backbone and are accompanied by high curvature whereas the negative entries are usually those that enter from left-handed loop regions, as there is no need for the backbone to reverse the torsion. Excluding several kinks within  $\beta$ -strands, the exits lead either smooth into the next loop regions (in case of negative exits) or into well-defined turns. In the latter case, the exit is either positive or negative depending on the nature of the turn.

The discussion of the APSA results listed in Table III reveals that the number of  $\alpha$ -helices and  $\beta$ -strands assigned by APSA is comparable to those suggested by existing methods such as DSSP, the disparities being further analyzed and found meaningful. APSA can also be used to quantify and systematically classify the regular as well as irregular structures leading to a more manageable and uniform structure description system, as all conformations are analyzed in the same way when they are classified. Turns are more variable among the secondary structures and owing to their non-repeating regularity, they are difficult to describe and categorize. It was shown in an earlier study<sup>14</sup> that turns that are similar (different) in 3D, indeed have similar (different)  $\kappa(s)$ - $\tau(s)$  patterns. The detail present in the  $\kappa(s)$ - $\tau(s)$  plots can be used to analyze kinks and distortions, which is sufficient proof that they contain extensive information regarding

**Table IV**

Comparison of the Conformational (Structural) Features of Ubiquitin (1UBQ) as Described by APSA and Three Other Methods taken from the Literature

VBC <sup>a</sup> (PDB)r	Secondary structure ranges			Approx. s values APSA		APSA Secondary structure	Comments
	DSSP <sup>b</sup>	Folding analysis <sup>c</sup>	APSA (This work)	Start	End		
M1-T7	Q2-T7	~β-1	Q2-T7	0	25	β-Strand 1	τ Plot: regular patterns that resemble the ideal beta strand
T7-G10 (β-Turn)	L8-T9	~Turn 1	L8-G10	25	38	Turn 1	τ Plot: successive and even number of sign changes shows a flat turn region.
			G10	36	Glycine pivot	Sharp reorientation of backbone at G10 causes the steady sign change through 0 in τ and a strong κ.	
G10-V17	T12-E16	~β-2	K11-E18	38	68	β-Strand 2	τ Plot: regular patterns that resemble the ideal beta strand
E18-D21	P19-S20	~Turn 2	P19-T22	68	84	Turn 2	κ & τ Plot: partial helix character
	T22 (Turn)		T22		Helix entry <sup>d</sup>	τ Plot: sign changes. τ Plot: extended conformation <sup>c</sup> at helix entry	
I23-E34	I23-E34	α-Helix	I23-E34	84	134	α-Helix 1	κ and τ Plot: end of helix shows slight distortion <sup>e</sup>
			G35	132	Glycine pivot	κ and τ Plot: sharp reorientation as in G10.	
			I36	134	148	β Conformation <sup>d</sup> Helix entry of the turn	τ Plot: β peak
			P37	144			τ Plot: extended conformation at helix entry in T22. <sup>d</sup>
P37-Q40 (Type III turn)	P38-Q40 (3 <sub>10</sub> helix)	P38-Q40 (short helix1)	P38-Q40	148	160	Turn 3	κ & τ Plot: shows 3 <sub>10</sub> helix character of a Type III turn.
Q40-F45	Q41-F45	~β-3	Q41-I44	160	174	β-Strand 3	τ Plot: short distorted segment.
F45-K48 (Type III' turn)	A46-G47	~Turn 3	F45-G47	174	186	Turn 4	κ and τ Plot: turn has partial helix character.
K48-L50	K48-Q49	~β-4	K48-E51	186	202	β-Strand 4	κ Plot: short and bent <sup>e</sup>
E51-R54	D52-G53	~Turn 4, ~Turn 5	D52-T55	202	220	Turn 5	1. Mixed helix and β- character in κ & τ peaks, as in Turn 2. 2. Extended conformation (helix entry) at T55, as in T22 (τ-Plot) <sup>d</sup> 3. T55-D58 region shows slight 3 <sub>10</sub> character of Type III turn (κ-Plot)
T55 (β-Bridge)							DSSP's 3 <sub>10</sub> -helix is an α-helix by torsion and a 3 <sub>10</sub> helix by κ.
T55-D58 (Type III turn)							
L56-Y59	L56 (Turn)	L56-Y59 (Short helix2)	L56-Y59	220	234	Helical segment	
	S57-Y59 (3 <sub>10</sub> helix) N60 (Turn)		N60 I61			β-Conformation <sup>d</sup>	Very strong κ at N60 shows strong bending of the backbone.
Q62-S65	K63-E64	~Turn 6	Q62-E64	243	258	Turn 6	κ Plot: partial helix character. τ Plot: sign changes.
E64-R72	T66-L71	~β-5	S65-L73	258	283	β-Strand 5	κ Plot: long and relatively flat. <sup>e</sup>
		~Turn 7	R72	284	End	Turn 7	κ Plot: R72 shows strong κ within the β-strand.

<sup>a</sup>Ref. 30.<sup>b</sup>Ref. 10.<sup>c</sup>Ref. 31.<sup>d</sup>The term β (or extended) conformation has been used based on this work; it refers to the conformation of the individual amino acid (as taken up in a β-strand) as determined from κ(s)- τ (s) plots. – Turns are given according to Ref. 5. Turns can be viewed as combinations of helix- and strand-amino acid conformations<sup>5</sup> and this is reflected in the κ(s)- τ (s) plots.<sup>e</sup>Comparison of κ patterns to the ideal (Figure 1n) shows slight strand distortion.

the direction and structure of turns. Thus, an analysis of a single protein is undertaken in APSA description of ubiquitin Section to show that the span of α-helices and β-strands as well as the nature of all loops and turns is accurately described by APSA.

### APSA description of ubiquitin

The results of the application of APSA to ubiquitin (1UBQ) are summarized in Table IV. Ubiquitin<sup>30</sup> is an α-and-β class protein with a roll topology according to



the CATH<sup>7</sup> classification. It is a single chain protein with 76 amino acids that assume approximately 14 recognizable secondary structures, including an  $\alpha$ -helix, 2 short helical segments, 5  $\beta$ -strands, and 6 turns. Table IV compares the APSA assignment of the structural units of IUBQ [for  $\kappa(s)$  and  $\tau(s)$  plots see Fig. 2(a)] with (i) a H-bonding- and  $\phi$ ,  $\psi$ -based method used by Vijayakumar, Bugg, and Cook (VBC),<sup>30</sup> (ii) the H-bonding-based DSSP method,<sup>10</sup> and (iii) the secondary structure assignment used for the description of IUBQ folding.<sup>31</sup> The terminology used in Table IV, assignment criteria, and the number of amino acids (span) of each structure are as stated in the original literature.<sup>10,30–32</sup> For example, the type III turns, as assigned by VBC,<sup>30</sup> have been well defined in literature as turns that have repeating  $\phi, \psi$  values of  $-60^\circ$ ,  $-30^\circ$ , identical with those of the  $3_{10}$ -helix. The type III' turn would be its mirror image. A “ $\beta$ -turn” (turn 1) refers to the turn connecting two successive antiparallel  $\beta$ -strands.

The terms used in connection with APSA are (if not discussed in the previous sections): (i) “ $\beta$ -Trough (peak),” which is a single strongly negative (positive)  $\tau$ -trough (peak) of a  $\beta$ -strand; (ii) “helical segment,” which is used when the segment is helical, but the exact conformation is not typically an  $\alpha$ -,  $3_{10}$ - or  $\pi$ -segment; (iii) “ $\beta$ -conformation,” which refers to the  $\beta$ -peaks (or extended peaks) occurring at the respective amino acids.

The  $\alpha$ -helix from I23 to E34 was identified unambiguously by all assignment methods, and so were the five  $\beta$ -strands. Of the two helical segments, the 38–40 [148–158 Å, Fig. 2(a)] one was variously described as a turn, a  $3_{10}$ -helix, or a short helix whereas the  $\kappa(s)$ - $\tau(s)$  diagrams clearly indicate  $3_{10}$  character. The second helical segment from 56 to 59 (right after turn 4 at the N-terminus). was described as type III turn by VBC; DSSP assigned a  $\beta$ -bridge, a turn, and a  $3_{10}$ -helix in succession whereas the folding analysis considered two turns followed by a “short helix.” As can be seen from the  $\kappa(s)$  and  $\tau(s)$  diagrams [Fig. 2(a)], the region can be split in any of the ways mentioned. However, an accurate APSA-based description of this region is that amino acids 52 and 53 of turn 4 forms a loop and then extend into 54 and 55 where a  $3_{10}$ -helix starts from the latter amino acid.

It is noteworthy that APSA is able to recognize single  $\beta$ -peaks (troughs) for DSSP's “isolated  $\beta$ -bridges,” although this is not the topic of this investigation because loop regions are not analyzed here. These are examples of the effects of tertiary structure on secondary structure. The  $\beta$ -bridge H-bond imposes the “ $\beta$ -peak” conformation on the isolated amino acid as reflected in the  $\tau(s)$  diagram [at 215 Å, Fig. 2(a)]. Among other proteins of the dataset though, this peak was found alongside other neighboring  $\beta$ -peaks leading to continuous  $\beta$ -strand assignments. The fact that turns can be viewed as combinations of extended and helix conformations has been documented.<sup>1</sup> This feature is seen in several of the

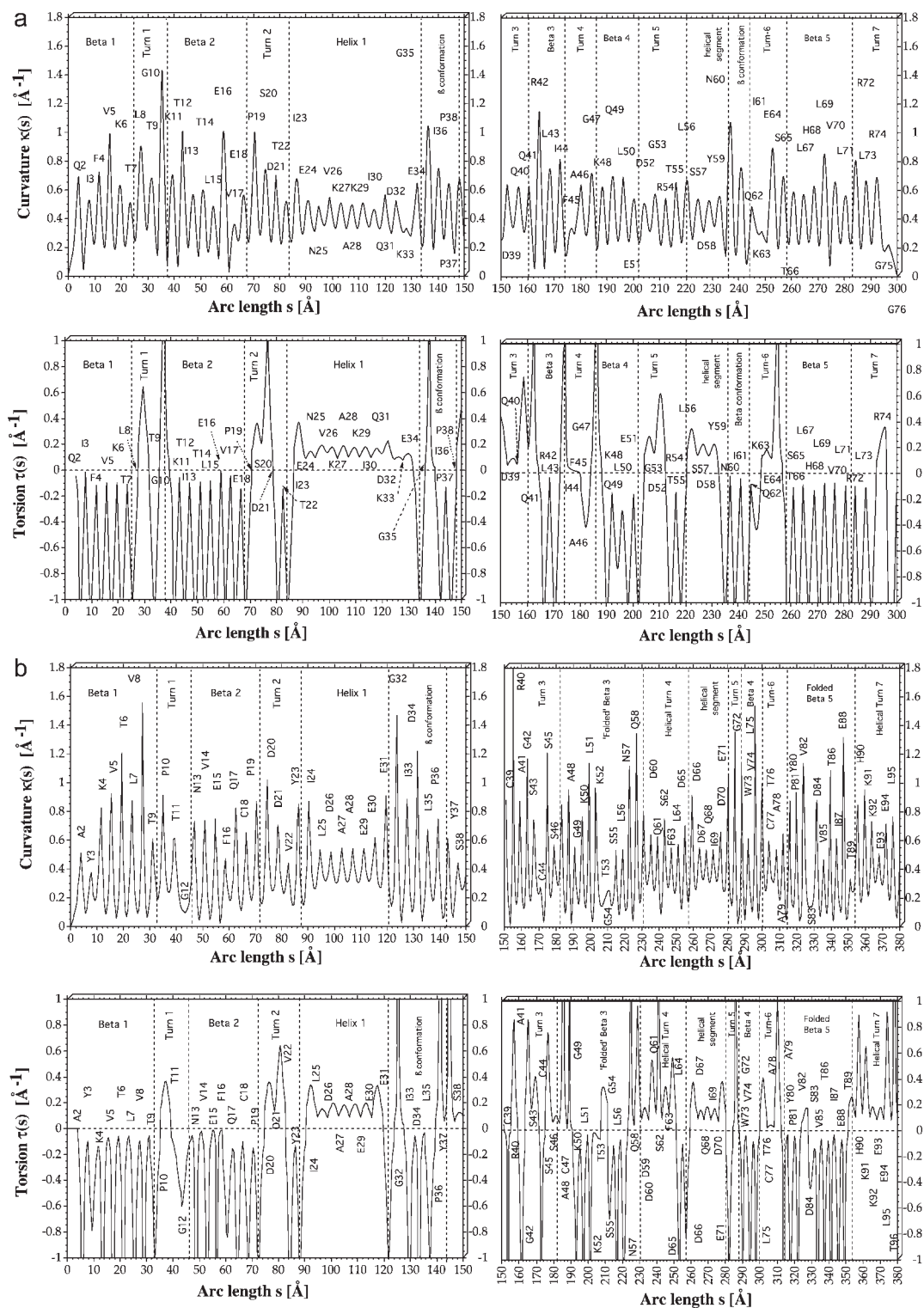
turn segments. In the IUBQ segment from  $s = 68$  to 84 Å [Fig. 2(a)], the P19 peak in  $\tau$  is helical (compare with the first peak of the helix at 84 Å) whereas the other three amino acids have  $\beta$ -peaks. Similar features are recognizable with the other turns. The entry (and exit) of the polypeptide chain into (and out of) the helix at threonine 22, proline 37, (asparagine 60), etc. due to local unwinding results in the characteristic  $\beta$ -peaks. In addition, the way it reorients its general direction using glycines 10, 35 and asparagine 60  $C_\alpha$  atoms as pivots have been shown [Table IV, Fig. 2(a)]. The advantage of a graphical representation is exploited to visualize that the  $\beta$ -strands of IUBQ, as seen from its  $\kappa(s)$  and  $\tau(s)$  patterns, are not perfectly flat [compare with ideal  $\beta$ -strand in Fig. 1(n)].

It can be seen that there are differences in structure assignment among the various methods. These differences arise not only due to the difference in the criteria used for assignment, but also due to the differing sensitivities in detecting the boundaries of the secondary structures. Early, it has been documented<sup>33</sup> that “ambiguity” is an intrinsic property of the protein, especially with respect to the turn regions that connect the boundaries of adjacent secondary structures (see Table IV). However, the similarity of turn 2 to turn 4 and its difference from  $\beta$ -turn-1 gives an idea to construct turn templates for loop regions. With respect to the choice of criteria, it should be remembered that the definition of the H-bond according to DSSP with respect to energy and distance is arbitrary and that the  $\phi$ - $\psi$  angle description of the polypeptide chain backbone is both discrete and local. As stated above, the deviation of the  $\beta$ -strands from the ideal is explicit and recognizable, especially as it is represented graphically. One can also relate to the specific parts of the secondary structure that is likely to deviate from the ideal. For example, the lysine 33 in the  $\alpha$ -helix deviating from the rest of the helix that stretches from isoleucine 23 to glutamate 34 is evident from the  $\kappa(s)$  and  $\tau(s)$  diagrams [compare ideal structure in Fig. 1(a)].

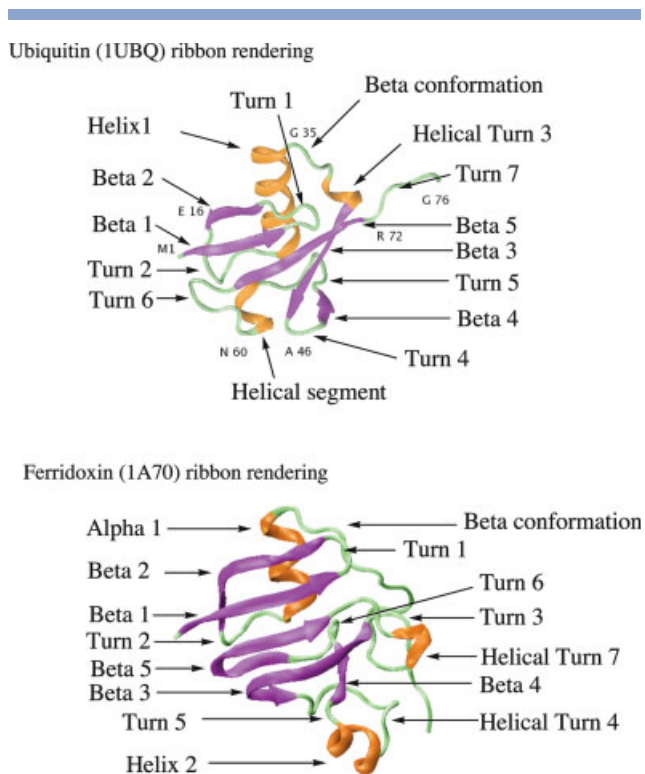
Though the  $\kappa$  and  $\tau$  information in Table III is mainly about  $\alpha$ -helices and  $\beta$ -strands, it is well known that many more intermediate structures exist to allow many conformations to occur (among the loop regions). Analysis and classifications of these regions will be the topic of a forthcoming paper.<sup>34</sup> With APSA, longer loop regions can be quantitatively described as having helical and extended regions alone.

### Recognizing common architectures—An APSA similarity test

Similar structural features have similar patterns in the  $\kappa(s), \tau(s)$  diagrams. Figure 2(b) shows the  $\tau(s)$  diagram of spinach ferredoxin 1A70, an iron-sulfur protein. It is 97 amino acids long and has the same roll architecture as


**Figure 2**

Torsion and curvature diagrams,  $\kappa(s)$  and  $\tau(s)$ , of (a) ubiquitin (1UBQ) and (b) spinach ferridoxin (1A70) also having the roll architecture of ubiquitin. Structural regions obtained from APSA are separated by vertical dashed lines and identified by a short term. Compare with the ribbon diagrams of 1UBQ and 1A70 given in Figure 3. See text and Table IV (Table V) for more details.

**Figure 3**

Ribbon diagrams of 1UBQ (top) and 1A70 (bottom).

ubiquitin (1UBQ, 76 residues) by CATH<sup>7</sup> classification. In the  $\kappa(s)$  and  $\tau(s)$  diagrams of Figure 2(b), the 2 helices, 5  $\beta$ -strands, and six turns that resemble 1UBQ are indicated to aid comparison (see also Fig. 3). As torsion  $\tau$  is an important and highly sensitive parameter, it is sufficient to use just  $\tau$  for the comparison of 1A70 and 1UBQ.

Inspection of the  $\tau(s)$  diagrams in Figure 2 immediately reveals the similarity of the two protein structures with regard to  $\beta$ -strands 1, 2, 3, 5, and helix 1. This can also be concluded when comparing the ribbon diagrams in Figure 3. However, the APSA diagrams of Figure 2 also reveal (dis)similarities in the non-regular structures such as the turns. For example turn 1 in 1UBQ is much more (right-left) twisted [larger  $\pm\tau$ -values, Fig. 2(a)] than that in 1A70 [Fig. 2(b)]. The same applies to turn 2. Protein 1A70 has 2 additional features labeled 'helical turn segments 4 and 7', which differ from turns 4 and 7 in 1UBQ [Fig. 2(a,b)]. These loop regions are only slightly helical and account for the fact that 1A70 is longer. The helical segment of 1UBQ at  $s = 220 \text{ \AA}$  [which is barely one turn of a  $3_{10}$ -helix; see curvature diagram in Fig. 2(a)] is longer and  $\alpha$ -helical in 1A70 (labeled  $\alpha$ -helix 2), occupying approximately an equivalent 3D position. The short and crooked  $\beta$ -strand 4 is found in both proteins, but is arranged differently in the sequence of secondary structural elements with respect to  $\alpha$ -helix 2.

In Table V, regular and non-regular structures of the two proteins are compared by complementing the APSA information from Figure 2 (and Table IV) by appropriate 3D pictures. The similarities of turns 1, 2, 6, and the  $\beta$ -conformation as reflected by the  $\tau(s)$  diagrams are confirmed by the 3D-pictures (see comments in Table V). In summary the  $\tau(s)$  diagrams (optionally complemented by the  $\kappa(s)$  diagrams) provide a rapid, accurate, and detailed analysis of the structures of the two proteins, which is confirmed by appropriate ribbon diagrams.




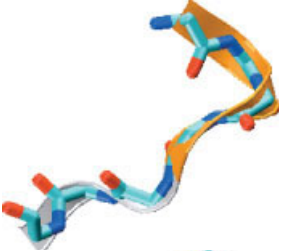
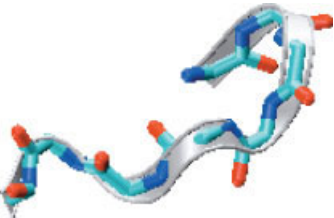



### Comparison of domain similarity

A fully automated and accurate method that can compare and classify proteins and protein segments at secondary, supersecondary, and tertiary levels without the need for manual intervention is not yet available. Though there are several databases of classified structures based on the proteins deposited in the PDB<sup>6</sup> such as CATH,<sup>7</sup> SCOP,<sup>8</sup> Dali,<sup>35</sup> TOPS,<sup>36</sup> etc., each of them uses a different approach to judge similarity among proteins. CATH and SCOP databases need manual analysis to complete the judgment of similarity. The update is sometimes accompanied by a rearrangement of previously classified structures when new structures are included into the database. The TOPS database makes the overall connectivity and folds visible by a rather drastic simplification of representing the secondary structures as cartoons. These methods are rigid in their assignment of secondary structures in the way that once a structure does not satisfy any of the limited definitions, the entire region is treated as "loop." Without further attempt to characterize the geometry in these regions, they are simply compared with the aim of getting differences.

In the light of the above need of having a more efficient and meaningful protein structure comparison method, it can be shown that in order to compare domains, averaging and simplification could be done without the loss of details at the secondary level. A direct comparison of the  $\kappa(s)$ ,  $\tau(s)$  diagrams of two proteins of the same architecture (Recognizing common architectures—An APSA similarity test Section) reveals how domain similarity can be ascertained by the locations of the secondary structures and the overall similarity of the turns. The closer the folds of the two proteins, the more identical their  $\kappa$ - $\tau$  patterns become.

For the purpose of providing further proof for the fact that APSA is perfectly suited to quantitatively determine the (dis)similarity of protein structure, different domains are compared in the following way. A set of 20 domains was selected from the "all alpha" class, 15 belonging to the "orthogonal bundle" and 5 to the "up and down Bundle" and these were compared with each other. The CATH tree is shown and numbered in Figure 4 representing a sampling at all levels of the CATH classification. As a measure of the relationship of these domains,

**Table V**Some Regular and NonRegular Structural Features of Spinach Ferredoxin (1A70) Compared with Those of Ubiquitin (1UBQ)<sup>a</sup>

1UBQ features APSA	Pictures	1A70 features APSA	Pictures
Turn 1 ( $s = 25-38$ )		<b>P10-G12</b> More planar than turn 1 in 1UBQ: the $\tau$ averages to 0. ( $s = 33-46$ )	
Turn 2 ( $s = 68-84$ )		<b>D20-Y23</b> The first part of the turn is partly helical and resembles the turn in 1UBQ ( $s = 72-88$ )	
$\beta$ -Conformation ( $s = 134-148$ )		<b>G32-Y37</b> loop region twists approx at right angles at every sign change of $\tau$ like a 'staircase'. ( $s = 122-144$ )	
$\beta 3$ ( $s = 160-174$ )	See ribbon diagram Figure 3	<b>C47-N57</b> folded beta strand ( $s = 182-230$ )	See ribbon diagram Figure 3
Turn 4 ( $s = 174-186$ )		<b>D59-D65</b> stretched right-handed loop; ( $s = 230-257$ )	
Helical segment ( $s = 220-236$ )		<b>D66-E71</b> distorted in 1UBQ wellformed $\alpha$ -helix in 1A70. ( $s = 257-280$ )	
Turn 4-small $\beta 4$ -Turn 5- helical segment ( $s = 174-236$ )		Turn 4 - Helix 2 - Turn 5 - small $\beta 4$ : rearranged relative to 1UBQ ( $s = 230-300$ )	
Turn 6 ( $s = 244-258$ )		<b>L75-A78</b> Partly helical as in 1UBQ. The first turn residue has positive $\tau$ pointing the rest of the turn downward (residues Q62 + K63 are oriented the same). ( $s = 300-314$ )	
$\beta 5$ ( $s = 258-283$ )	Figure 3	<b>A79-E88</b> long and folded like $\beta 3$ . ( $s = 314-354$ )	Figure 3
Turn 7 ( $s = 283-end$ )	Figure 3	<b>T89-A97</b> A turn within the $\beta$ -strand in 1UBQ becomes a long helical loop; ( $s = 354-end$ )	Figure 3

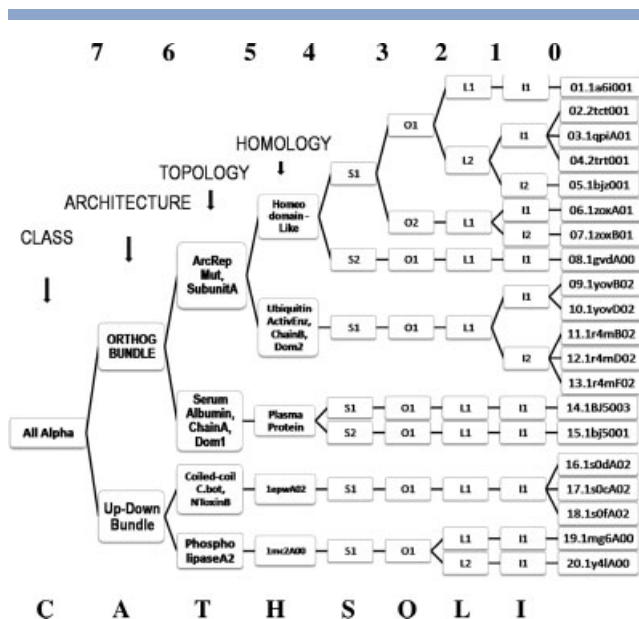
<sup>a</sup>Arc length values  $s[\text{\AA}^{-1}]$  are taken from Table 4 and Figure 2. The 3D structural units of 1UBQ and 1A70 have been prepared with VMD and oriented in such a way that the first two residues always point in the same direction for a pair.

an "order of relationship" was set up by counting (from right to left in Fig. 4) the number of CATH nodes separating two domains. Two domains of 0-order relationship belong to the same "I level"; the "D" level, the final level of CATH containing identical proteins, is not considered in the order scale. The highest number in terms of order

is 8, signifying different classes. Thus, fourth order relationship domains belong to the same homology, but not to the same "S level." This class contains identical proteins and therefore is not considered in the order scale.

As a measure of similarity, a grading scheme was set up with letters ranging from A (identical) to F (dis-





**Figure 4**

CATH<sup>7</sup> similarity relationships for 20 domains that are shown on the far right. The CATH levels are given on the bottom and the orders of relationship on the top. See text for details.

similar) signifying decreasing similarity of domains within the same all alpha class (see Figs. 4 and 5). The criteria A to F used were based on the number and ordering of sec-

ondary structures,  $\kappa(s)$ ,  $\tau(s)$  patterns of the turns, types of entries and exits, the nature of loop regions, the size of the domains, and the overall ordering of the secondary structures with respect to each other (see Table VI). Allowance was given for variation; for example, some loop regions that appeared to be distorted helices were recognized similar to an  $\alpha$ -helix (Table VI). A correlation of this similarity index was combined with the order index creating a similarity matrix (see Fig. 5). It is to be expected from such a correlation, that the smaller the order, the closer the relationship of the domains by CATH, the higher should be the grade of similarity assigned.

For an “A” grade similarity of two domains (see Fig. 5) 99% of all amino acids have to have similar  $\tau(s)$  patterns according to the properties listed in Table VI. An example is shown in Figure 6(a) where the  $\tau(s)$  values of domains 1 and 2 having an order of relationship of 2 are identical. A reference to the length of the domain is made to accommodate greater flexibility in the longer loops of larger domains (Table VI), as in the case of domain 16, 17, and 18 that are about 300 amino acids long. Distortions in helices are permitted along with some minor variations. A grade “B” similarity (Table VI, Fig. 5) implies stronger distortions in secondary structures and/or differences in parts of turns such as 2–3 negative  $\tau$ -troughs instead of positive ones. Domain 6 differs from domain 2 at amino acids 18 to 20, at the C-terminus of the last helix and at the arrangement of the last few amino acids. Stronger distortions that evidently bend

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	2	2	2	2	3	3	4	5	5	5	5	5	6	6	7	7	7	7	7
2	A	-	0	0	1	3	3	4	5	5	5	5	5	6	6	7	7	7	7	7
3	A	A	-	0	1	3	3	4	5	5	5	5	5	6	6	7	7	7	7	7
4	A	A	A	-	1	3	3	4	5	5	5	5	5	6	6	7	7	7	7	7
5	A	A	A	A	-	3	3	4	5	5	5	5	5	6	6	7	7	7	7	7
6	B	B	B	B	B	-	1	4	5	5	5	5	5	6	6	7	7	7	7	7
7	B	B	B	B	B	A	-	4	5	5	5	5	5	6	6	7	7	7	7	7
8	C	C	C	C	C	C	C	-	5	5	5	5	5	6	6	7	7	7	7	7
9	D	D	D	D	D	D	D	D	-	0	1	1	1	6	6	7	7	7	7	7
10	D	D	D	D	D	D	D	D	A	-	1	1	1	6	6	7	7	7	7	7
11	D	D	D	D	D	D	D	D	A	A	-	0	0	6	6	7	7	7	7	7
12	D	D	D	D	D	D	D	D	A	A	A	-	0	6	6	7	7	7	7	7
13	D	D	D	D	D	D	D	D	A	A	A	A	-	6	6	7	7	7	7	7
14	E	E	E	E	E	E*	E*	D	E	E	E*	E*	E*	-	4	7	7	7	7	7
15	E*	E*	E*	E*	E*	E	E	E	E	E	E	E	E	C	-	7	7	7	7	7
16	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	-	0	0	6	6
17	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	A	-	0	6	6
18	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	A	A	-	6	6
19	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	E	E	E	-	2
20	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	E	E	E	A	-

**Figure 5**

A similarity matrix constructed for 20 domains (see Fig. 4) with the order of relationship taken from CATHSOLID classification system on the upper half of the matrix and the graded APSA similarity on the lower half. The order of relationship is given by the number of nodes separating the domains as counted from the relationship chart shown in Figure 4. Note that E\* is between D and E.

**Table VI**Grading system A-F Used for the Similarity Test Shown in Figure 5<sup>a</sup>

Grade	Same length <sup>a</sup> of proteins	Different length property criterion	Comment
A	99% APSA agreement	Residues agree for P1–P8	Similarity test is simplified by the agreement in length
B	≥93% APSA agreement	Residues to be comparable agree for P1–P8	In case of different protein length, residues without partners have to be eliminated
C	Note within the 20 domains	Comparable residues agree for P1–P6, differ for P7,P8	Overall shape is maintained. “Acceptable” differences are those where loop regions resemble distorted secondary structures.
D	No example	Comparable residues have similar P1–P3 and P5	Different loop lengths may change sec. structure orientation. “Acceptable” differences are discounted & $\kappa(s)$ shows more similarity than $\tau(s)$
E	No example	Comparable residues have similar P1–P2; P3 may differ	Differences in ordering of sec. structures cause different folds or at least topologies; some scattered turns contain resemblances
F	No example	Comparable residues have similar P1	Different fold; some turns scattered in the domain are still similar with respect to shape and sign indicating similar super secondary structures
G	No example	P1–P8 are different	Different classes

<sup>a</sup>Two proteins will be considered to be of the same length if the calculated arc lengths  $s$  agree within  $\pm 5 \text{ \AA}$ . The following 8 structural properties P (ordered according to increasing detail) derived from the  $\tau(s)$  diagrams of APSA are determined: (P1) Ratio of helices and  $\beta$ -strands according to  $\tau$ -patterns; (P2) Number secondary structural units according to  $\tau$ -patterns; (P3) Order of secondary structural units according to  $\tau$ -patterns; (P4) Lengths of turns/loops connecting secondary structures according to arc length  $s$ ; (P5) Overall sign of torsion at the turns/loops according to  $\tau(s)$  ( $\tau$  is averaged over the whole turn by calculating area under the curve of the  $\tau$  peaks); (P6)  $\tau$ -Sign of individual residues in the turns/loops; (P7) Nature of the turns/loops whether the  $\tau$  patterns resemble helical or extended conformations; (P8) Match within assigned secondary structures with respect to distortions according to  $\tau(s)$ .

helices to orient them differently in 3D space are graded with a “C” similarity (Table VI), which also includes significant differences in the turns and loops owing to the different sizes of the domains being compared. In the case of domains 2 and 8, the first 3 helices of domain 2 strongly resemble the whole of domain 8. Thus, even though both domains are “3 helix bundles,” the presence of extra helices in domain 2 can be clearly seen.

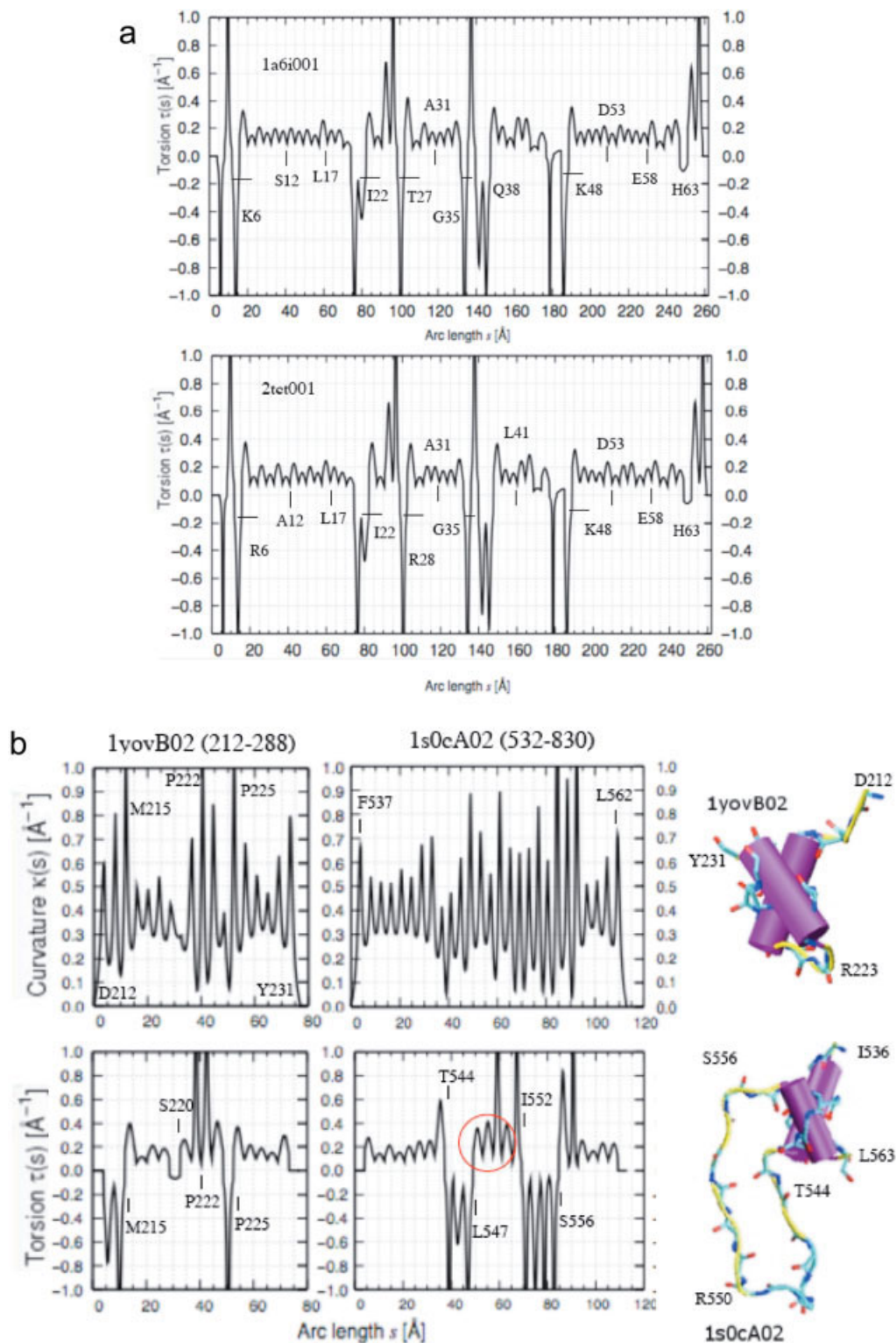
A grade that would interpret as “different” is “D” (Table VI, Fig. 5). When given a grade “E,” the secondary structures of the domains are present in totally different supersecondary arrangement making the fold of the domain significantly different. However, similarities can be seen between different parts of the protein. Some of the supersecondary structures are similar; however, they occur in a “jumbled” order, thus differing in topology. A greater difference leads to a grade of “F.” It is noteworthy that, though the domain as a whole (boundaries as prescribed by CATH) is considered to be very different by this index, similar supersecondary structures and folds can still be recognized at various parts. For example, among the first few helices of domain 17 (1socA02) ranging approximately from amino acid 536 to 644, several secondary structure and turn features can be identified belonging to the orthogonal bundle architecture. A comparison with domain 9 as shown in Figure 6(b) clarifies the fact that though the loop in domain 17 is more meandering resulting in the many oscillations in  $\tau(s)$ , the patterns are equivalent. The two positive  $\beta$ -peaks at 221 and 222 in 1YOV correspond to the same at 549 and 551

in 1S0C. As discussed in The APSA description of extended structures and their distortions Section, one  $\beta$ -peak (at 244, 1YOV) is equivalent, by rotation to four  $\beta$ -peaks (at 552, 1S0C). After accounting for these rotations and translations, the equivalence of the two segments can be seen in the 3D inset. The long helices and the turns that appear after amino acid 644 in 1S0C, however, clearly reflect a different arrangement, namely the up-and-down bundle. As the similarity assignment from A to F is done only for domains within the same class (all alpha), greater differences that occur beyond the all alpha class of domains are not documented.

## CONCLUSIONS

The performance of APSA being based on the determination of curvature and torsion of the protein backbone has been demonstrated in this work. Previous protein structure descriptions, which have taken an approach related in some way to APSA, have been discussed in Ref. 14 and the advantages of APSA with regard to these approaches have been worked out there and do not need to be repeated here.

A systematic analysis performed on five  $\alpha$  helices and eight  $\beta$  strands (Table I) resulted in the derivation of a working definition for the same secondary structures in terms of curvature and torsion patterns,  $\kappa(s)$  and  $\tau(s)$ . An automated analysis of 77 proteins carried out with APSA led to a secondary structure assignment that was compared to that of DSSP. A total of 533  $\alpha$ -helices and

**Figure 6**

(a) Domains 1 (1a6i001) and 2 (2tct001) both ranging from amino acids 2 to 66 of the respective proteins whose  $\tau$  are identical; order of relationship = 2, similarity index = A. (b) A comparison of the orthogonal helix pairs from domains 9 (1yovB02) and 17 (1s0cA02) show resemblances in  $\tau$  and 3D arrangement. See text for details.



644  $\beta$  strands were recognized by APSA, whereas DSSP's assignments (536  $\alpha$ -helices and 651  $\beta$ -strands) differed for 20  $\alpha$ -helices (12 more, 8 less) and 46  $\beta$ -strands (24 more, 22 less). Though the approaches are vastly different, the total number of structures was thus found comparable. In addition, the conformational features in 3D space were accurately described in the 2D  $\kappa(s)$  and  $\tau(s)$  diagrams. From  $\tau(s)$  alone, in most cases, kinks and distortions could be recognized and quantified. A list of distortions was also discussed as occurring in the body and termini of  $\alpha$ -helices and  $\beta$ -strands. A way of describing distorted helical termini based on whether the diameter of the region was larger or smaller than the  $\alpha$ -helix, as deduced from low or high  $\kappa(s)$  values and variations in  $\tau(s)$  was presented.

Similar structural features between any two proteins also become evident in APSA's  $\kappa(s)$  and  $\tau(s)$  diagrams. The roll architecture of ferridoxin (1A70) and ubiquitin (1UBQ) were compared. Two extra loop regions of the former protein between residues 32–44 and 57–64 that correspond to an increase in overall length of the fold were shown. In the wake of such a comparison, the degree of CATH relationship and index of similarity was correlated in an analysis that compared twenty all  $\alpha$  domains with each other. It was shown that these 2D  $\kappa(s)$ ,  $\tau(s)$  patterns could be used for similarity comparisons at any level whether secondary, super-secondary, or tertiary. Accordingly, domains of different homologous superfamily, topology, and architecture were shown to have increasingly different  $\kappa$  and  $\tau$  profiles.

The APSA method accurately reflects the conformation of the backbone effectively reducing 3D information to a 2D representation. The method is mathematically well founded and computationally robust, describing each secondary structure with a unique  $\kappa(s)$ ,  $\tau(s)$  pattern reflecting its 3D properties. Analysis of the 78 protein chains investigated in this work with APSA requires about 1 sec computer time. Hence, APSA is well-suited for the rapid structure analysis of the 50,000 proteins of the PDB. It provides a complete conformational analysis and identification of all residues of a protein.

It is a continuous representation where a global trend in conformation can be seen for all amino acids, whether they are in the helical, extended or loop regions of proteins. The speed and the simplicity of the analysis are due to the use of a simplified backbone representation. It was demonstrated that APSA can be easily applied to the analysis of supersecondary and tertiary structure.<sup>34</sup>

APSA is exclusively based on conformational (structural) protein data as reflected by the positions of the  $C_\alpha$  atoms in the protein backbone whereas the DSSP description strongly depends on the types and arrangements of H-bonding in the protein. APSA does not need any charge or energy information, which are essential for DSSP. This is a clear advantage over DSSP's assessment of backbone structure because H-bonding patterns do not supply information on the distortions and orienta-

tions of backbone structures. Otherwise, APSA and DSSP should complement each other where APSA should take the lead in the structural analysis because of its rapid description and DSSP should come in with additional information, especially on H-bonding.

## ACKNOWLEDGMENTS

DC and EK thank the University of the Pacific for support.

## REFERENCES

- Richardson JS. The anatomy and taxonomy of protein structure. *Adv Protein Chem* 1981;34:167–339.
- Chelvanayagam G, Heringa J, Argos P. Anatomy and evolution of proteins displaying the viral capsid jellyroll topology. *J Mol Biol* 1992;228:220–242.
- Scheerlinck JPY, Lasters I, Claessem M, DeMaeyer M, Pio F, Delhaise P, Wodak SJ. Recurrent  $\alpha\beta$  loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins* 1992;12:299–313.
- Murzin AG, Finkelstein AV. General architecture of the  $\alpha$ -helical globule. *J Mol Biol* 1988;204:749–769.
- Richardson JS.  $\beta$ -Sheet topology and the relatedness of proteins. *Nature* 1977;268:495–500.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank 2000. *Nucleic Acid Res* 2000;28:235–242.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. *Structure* 1997;5:1093–1108.
- Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
- Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins* 1985;23:566–579.
- Sanger K. Dictionary of secondary structure of proteins: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
- Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary and first level supersecondary structure. *Proteins* 1988;3:71–84.
- Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. *J Mol Biol* 1977;114:181–293.
- Venkatachalam C. Stereochemical criteria for polypeptides and proteins: conformation of a system of three linked peptide units. *Biopolymers* 1968;6:1425–1436.
- Ranganathan S, Izotov D, Kraka E, Cremer D. Automated and accurate protein structure description: Distribution of ideal secondary structural units in natural proteins. arXiv: 0811.3252 [q-bio.QM].
- Arab S, Didehvar F, Eslahchi C, Sadeghi M. Helix segment assignment in proteins using fuzzy logic. *Iran J Biotechnol* 2007;5:93–99.
- Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. *J Mol Graph* 1996;14:33–38.
- Benedetti E, Blasio BD, Pavone V, Pedone C, Santini A, Crisma M, Toniolo C. Molecular Conformations and Biological Interactions: The 3-10 and alpha-helical conformation in peptides. *Ind Acad Sci* 1991;3:497–502.
- Barlow DJ, Thornton JM. 3Helix geometry in proteins. *J Mol Biol* 1988;201:601–619.
- Kumar S, Bansal M, Velavan R. HELANAL: a program to characterize helix geometry in proteins. *J Biomol Struct Dyn* 2000;17:811–819.
- Cartailler JP, Luecke H. Structural and functional characterization of pi bulges and other short interhelical deformations. *Structure* 2004;12:133–144.



21. Sun Z, Blundell T. The pattern of common supersecondary structure (motifs) in protein databases. In: System science, 1995;5:312–318. Proceedings of the 28th Annual Hawaii International Conference on System Sciences, Wailea, HI, January 3–6, 1995.
22. Sun Z-R, Zhang C-T, Wu F-H, Peng L-W. A vector projection method for predicting supersecondary motifs. *J Protein Chem* 1996;15:721–729.
23. Taylor H. Large molecules through atomic spectacles. *Proc Am Phyl Soc* 1941;85:1–12.
24. Doig AJ, MacArthur MW, Stapely BJ, Thornton JM. Structure of N-termini of helices in proteins. *Protein Sci* 1997;6:147–155.
25. Karpen ME, De-Aseth PL, Neet KE. Differences in amino acid distribution of 3[10]-helices and alpha-helices. *Protein Sci* 1992;1:1333–1342.
26. Toniolo C, Crisma M, Formaggio F, Peggion C, Broxterman Q, Kaptein B. Peptide  $\beta$ -bend and  $3_{10}$  helix: from 3D structural studies to applications as templates. *J Inclusion Phenom Nad Macrocyc chem* 2005;51:121–136.
27. Fodje MN, Al-Karadaghi S. Occurrence, conformational features, and amino acid propensities for the pi-helix. *Protein Eng* 2002;15:533–358.
28. Sowdhamini R, Srinivasan N, Ramakrishana C, Balaram P. Orthogonal  $\beta\beta$  motifs in proteins. *J Mol Biol* 1992;223:845–851.
29. Efimov AV. Standard structures in proteins. *Prog Biophys Mol Biol* 1993;60:201–239.
30. Vijayakumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution. *J Mol Biol* 1987;194:531–544.
31. Zhang J, Quin M, Wang W. Multiple folding mechanisms of protein ubiquitin. *Proteins* 2005;59:565–579.
32. Protein Data Bank. [www.rcsb.org/pdb](http://www.rcsb.org/pdb), December 2008.
33. Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. *Adv Prot Chem* 1985;37:1–109.
34. Ranganathan S, Izotov D, Kraka E, Cremer D. Projecting three-dimensional protein structure into a one-dimensional character code utilizing the automated protein structure analysis method. arXiv: 0811.3258 [q-bio.QM].
35. Holm L, Sander C. Mapping the protein universe. *Science* 1996; 273:595–603.
36. Michalopoulos I, Torrance GM, Gilbert DR, Westhead DR. An enhanced database of protein structural topology. *Nucl Acid Res* 2004;32:D251–D254.