7

METHODS FOR A RAPID AND AUTOMATED DESCRIPTION OF PROTEINS: PROTEIN STRUCTURE, PROTEIN SIMILARITY, AND PROTEIN FOLDING

ZHANYONG GUO AND DIETER CREMER

Computational and Theoretical Chemistry Group (CATCO), Department of Chemistry, Southern Methodist University, Dallas, TX, USA

INTRODUCTION

Protein structure is organized into primary, secondary, tertiary, and quaternary levels expressing in this way its enormous variety and complexity.[1–5] There have been numerous attempts of simplifying measured protein structures for the purpose of identifying unique conformational patterns. For example, Venkatachalam[6] presented a local description of a protein molecule based merely on the polypeptide chain backbone. Furthermore, he showed that just two of the three conformational angles (ϕ and ψ) of the backbone have to be specified for a particular amino acid residue because conjugative effects keep the peptide unit planar (the angle ω at the peptide bond is normally close to 180°).[6] These simplifications do not hinder the valid description of a protein and confirm earlier predictions[1–5] of its periodical structure in agreement with crystallographic data.[7]

A different approach employed by Kabsch and Sander[8] describes secondary structural units (SSUs) in terms of the shape and organization of hydrogen-bonded units

Reviews in Computational Chemistry, Volume 29, First Edition.
Edited by Abby L. Parrill and Kenny B. Lipkowitz.

found along the backbone. They were able to identify helices and β-sheets quickly and precisely. A higher level of organization that involves pairs of SSUs called the supersecondary structures was identified by Rao and Rossman[9] by comparison of protein structures. These and other methods were applied to protein structure description, always with the objective of developing generally applicable conformational rules based on the simplification of protein structure.[10,11]

From an elementary point of view, one can distinguish between three different approaches to fulfill these objectives. One can base protein structure analysis and prediction exclusively on conformational (geometrical) features. Alternatively, one can correlate structural features with amino acid properties such as H-bonding ability, hydrophilicity, hydrophobicity, polarity, etc.[1–4] and use the latter properties for structure classification. Finally, one can combine conformational (geometrical) and physiochemical amino acid properties for the purpose of structure analysis and prediction.

Based on various such descriptions, SSUs like helices and β-sheets, supersecondary structures like hairpins and corners, larger supersecondary motifs like the β-barrel, and folds of domains in globular proteins have been extensively classified.[1–4] Classification of proteins deposited in the PDB (Protein Data Bank)[12,13] can be found in databases such as SCOP (Structural Classification Of Proteins),[14,15] CATH (Class-Architecture-Topology-Homologuous superfamilies),[16,17] DALI (Distance ALIgnment),[18–20] etc., which combine automated and manual sorting of domains. Such domain classifications have been compared and analyzed,[21] and it has been shown that they are frequently conflicting with regard to domain definition and assignment of domain boundaries. Conflicting descriptions can also be obtained when using secondary structure assignment programs such as DSSP (Define Secondary Structure of Proteins),[8,22] STRIDE (STRuctural IDEntification of the secondary structure of proteins),[23] DEFINE (DEFINE protein structure),[24] and KAKSI (Finnish for "two": C_α-distance matrix and (ϕ,ψ)-backbone angles)[25] (for a comparison, see Fourrier and de Brevern,[26] Martin and coworkers,[25] or Offmann and coworkers[27]). Also, it requires detailed and individual analysis of structures at specific localized sites to classify loop regions and identify the simpler units from which they are made.[28,29] More recently, secondary structure has been analyzed on the basis of neural networks.[30–32] In general, modern secondary structure assignment and prediction can refer to a multitude of strategies based on H-bonding,[8] backbone conformation,[6] multiple sequence alignments,[33] or energy-based evaluations.[34,35]

It is generally accepted that one can recognize protein structure from the form of its backbone.[1–6] Once the backbone structure is understood, one can complete the backbone by adding side chains using available procedures.[36] Despite such accomplishments, there is a need for an improved analysis of the backbone structure[37] in connection with protein structure predictions and when elucidating protein functionality. Hence, an automated systematic description of the backbone structure of a protein is still, after many decades of protein studies, a needed tool to relate protein primary structure to protein functionality. The state of the art in protein structure description is often judged in view of its value for protein structure prediction where the efficiency of automation plays a major role.[38,39]

Any useful description of the backbone or the total protein structure has to consider the different levels of structural description, which lead from the SSUs through the supersecondary units and motifs to the tertiary level with folds of domains.[1–5] This hierarchy of structural descriptors implies that for any given point of a protein backbone, an increasingly larger environment has to be taken into account to step up from a purely local to a global description. Clearly, any complete description of the protein backbone must include all levels of the structural hierarchy. One could assume that after 60 years of research in this field, at least the secondary level of protein structure is well understood. However, this is questionable given the many investigations published during the last 20 years, which present a manifold of new SSU assignment methods. The latter can be divided into two categories:

1. Hydrogen bonding-based methods such as DSSP: Beta-Spider (focus on β-sheets),[40] DSSPcont (improved description of H-bonding by considering protein motion and thermal fluctuations),[41] and SECSTR (focus on π-helices).[42]
2. Geometry-based methods: SABA (use of anchor points to obtain coarse-grained structures),[43] PROSIGN (mathematical helix description to distinguish between helices and strands),[44] SEGNO (use of dihedral angles and distances),[45] KAKSI (dihedral angles and distances),[25] PALSSE (description of SSUs as vectors),[46] VoTAP (focusing on contact matrices),[47] *t*-number (contact matrices),[48] and STICK.[49]

These complement or improve the assignment methods from the 1980s and 1990s such as DSSP,[8] STRIDE,[23] and PROMOTIF[50] (all H-bonding based) or DEFINE,[24] P-CURVE,[51] XTLsstr,[52] and P-SEA[53] (all geometry based).

In this connection, special topics such as the description of helices with quaternions,[54] the identification of π-helices,[55,56] the determination of helices with opposite chirality (left-handed helices),[57] the characterization of helix kinking,[45] the classification of turns,[58] the description of SSU distortions, and the precise determination of the SSU termini[45] have been investigated. Some of this work was summarized in recent reviews.[27,59]

In this review, we will describe protein structure in the spirit of early structural simplifications. This implies that all side chains are deleted so that one can focus just on the protein backbone (for an example,[60] see Figure 1a–d). The orientation of the backbone in three-dimensional (3D) space can be determined by a multitude of residue dihedral angles ϕ and ψ. However, such an approach leads to a local and discrete description of the backbone that, only with difficulties, can be applied for the description of tertiary structure. For the purpose of developing a simpler approach to protein structure description, two strategies can be followed[61]:

1. One can sacrifice a fine-grained atom-by-atom description of the backbone leading to excessive detail in favor of a coarse-grained one. The latter can be achieved in various steps with the purpose of including increasingly more nonlocal structural features into the protein description. A common approach is that each residue is represented by just one anchor point (level 1 of coarse

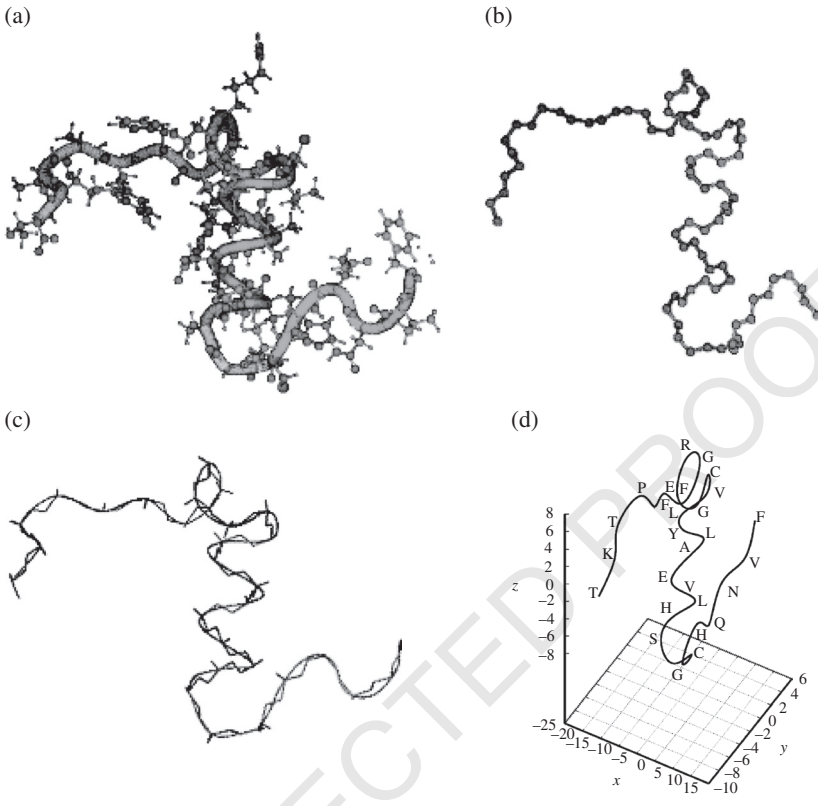

FIGURE 1 Simplification of the protein structure to a smooth line in 3D space in three steps. (a) Perspective drawing of the structure of insulin B chain taken from PDB file 1JCO.[60] (b) After deleting all side chains, the backbone of insulin B chain becomes visible. The C_α atoms of all residues are indicated by dots. (c) The positions of the C_α atoms are connected by a cubic spline fit. (d) The backbone is represented as a smooth line in 3D space.

graining; Figure 1c). At higher levels of coarse graining, one can present supersecondary or tertiary structural units by arrows and planes being derived from multiple anchor points to add more nonlocal features to the protein description (level 2, level 3, etc. of coarse graining). In this way, all features from secondary to tertiary structure become accessible to a purely conformational (geometric) approach of protein structure description.

2. The second feature of a general but simplified approach to protein structure description is based on a continuous rather than discrete representation of the protein backbone.[61] One can convert the latter into a smooth line oriented in 3D space by the anchor points of step (a). Along this line, one can use three mathematical parameters to determine the length and orientation of the backbone line: the arclength s, the (scalar) curvature κ, and the torsion τ. These parameters are sometimes called *Frenet coordinates* according to the

Frenet–Serret frames, which are used in differential geometry to describe the movement of a particle along a curved line in 3D space.[62,63] By expressing κ and τ as a function of the arclength s, one obtains continuous representations of the protein backbone (see Figure 1d), which adopt characteristic patterns for given structural features of a protein. Protein structure can be described in terms of *structure spectra based on the Frenet coordinates* κ *and* τ. The advantages of such a description lie in its ease of automation with the help of a computer program, its general applicability, and the mathematical accuracy of the structure description.[61,64]

There have been various attempts in this direction, which are reviewed briefly in the following.

PROTEIN STRUCTURE DESCRIPTION METHODS BASED ON FRENET COORDINATES AND/OR COARSE GRAINING

Coarse-grained backbone structures based on other than internal coordinates have been repeatedly used. In 1978, Rackovsky and Scheraga[65] suggested to evaluate the Frenet coordinates κ and τ at the C_α positions. The protein backbone was represented by a set of "equidistant" points in space (C_α atoms) connected by virtual bonds. The four points $i-1,\ldots,i+2$ formed an elementary unit at point i, which defined κ_i and τ_i. The graphical representations of κ versus τ were used for the identification of structural elements, and the diagrams giving κ or τ in dependence of the residue number were employed for protein comparison. This method reflected the local properties of the curvature and torsion of the protein backbone. The bond angles and dihedral angles defined by the virtual bonds were related to the traditional angles ϕ and ψ. The virtual bonds, of course, differ significantly from the protein backbone so that $\kappa(s)$ or $\tau(s)$ patterns identifying structural units of a protein could not be obtained. Hence, the method of Rackovsky and Scheraga[65] was limited in its presentation of the protein backbone because a discrete rather than a continuous description in terms of Frenet coordinates leads to shortcomings when analyzing protein structure.

Louie and Somorjai[66,67] considered the native conformation of a protein as the collection of minimal surfaces such as helicoids and catenoids linked by turns and irregular coils, which contain mainly polar groups and are exposed to the solvent. The backbone was defined as a geodesic curve on the minimum energy surface. Only helices and β-strands could be analyzed in this way, whereas an analysis of "nonregular" residues (belonging to turns, loops, etc.) could not be performed. In this approach, curvature κ and torsion τ were uniformly approximated by stepwise functions corresponding to discrete values at C_α atoms. The fitting produced a sequential overlap of all portions of the protein. Recognition of helices was based on the comparison with an average cylinder of given radius and pitch. The bends were defined as changes in axial angles of the fitting helices, whereas turns were recognized when three or more consecutive changes of axial angles were all greater than 40°. The method could not be generally applied and was best suited for helix recognition.

Soumpasis and Strahm[68] also considered describing the protein backbone by curvature and torsion obtained from polygonal nondifferential paths defined by C_α positions. At each vertex, curvature κ was defined utilizing the circumscribed circle of a triangle spanned by three sequential C_α atoms. The torsion τ was defined in a similar way using a tetrahedron spanned by four C_α atoms. The formulas for calculating κ and τ were expressed in terms of Cayley–Menger determinants. In this approach, the definition of τ suffered from an ambiguity and therefore it had to be augmented by an extra condition, which led to a complex torsion with real and imaginary part. Using this approach, the authors obtained specific profiles characterizing secondary and supersecondary structure.[68] One of the drawbacks of the Soumpasis–Strahm approach was that it did not produce an easy-to-analyze picture of the backbone structure.

Hausrath and Goriely suggested a continuous representation of proteins from curvature profiles,[69] which were determined in a way similar to that of Louie and Somorjai.[66,67] Small segments of the protein were described by a piece of a helix, which was then used for the calculation of curvature and torsion. The latter obtained a stepwise character without specific patterns. Curvature and torsion were used for the calculation of amino acid atom coordinates in the local coordinate system. The averaged values of these coordinates were employed to reconstruct the protein conformation for similarity comparisons.

Sklenar, Etchebest, and Laverey[51] suggested a method for smoothing the protein backbone. Each residue was assigned a local helical axis system obtained by a least-squares fit. The latter took into account the differences in the axis systems of the nearest neighbors. Local variations of atomic coordinates were "smoothed" out and a more global picture could be obtained. Again, only a discrete set of points could be produced to represent the protein backbone. Differing orientations of helices were described by the method. The regions with regular secondary structures (helices and β-sheets) were seen as straight segments of a line, which were linked by curved segments corresponding to the nonregular conformations. However, this led to the loss of a visible difference between helices and β-sheets. Additionally, structural details along the protein backbone were lost during the smoothing process. The method contained several parameters, which could not be related directly to ϕ and ψ values or to protein conformation in general.

Zhi and coworkers[70] suggested a smoothing technique for the protein backbone by averaging C_α positions in a seven-residue window. Chain fragments that remained straight after smoothing were denoted as generalized SSUs. The main characteristic was the turning angle along the smoothed backbone. Analysis and comparison of protein structures could be carried out by aligning the arrays of the angles. Though this approach gave a more global view, it was unable to differentiate between SSUs. Can and Wang[71] estimated curvature and torsion of the protein backbone by representing it as a smooth line, which was defined by the set of C_α atoms assumed to be equidistant. A fifth-order spline was applied for the smoothing procedure, but this spline did not pass through the chosen anchor points. Helices were described with high precision, but turns were not. After subsequent normalization, the curvature and torsion were used as signature parameters for structure comparison in different proteins. No secondary structure recognition was attempted.

Finally, it is noteworthy that work on DNA helices was also carried out utilizing a description with curvature and torsion.[72,73]

THE AUTOMATED PROTEIN STRUCTURE ANALYSIS (APSA)

A generally applicable protein structure analysis method based on Frenet coordinates was developed by Cremer and coworkers[61,74,75] and dubbed *Automated Protein Structure Analysis (APSA)* method. The authors derived their approach from methods used in reaction dynamics where the path of a chemical reaction complex is presented as a curved line in multidimensional space. The reaction path is described by Frenet coordinates to obtain a path description in 2D space.[76–78] The techniques and ideas used for the presentation of the path, which the reaction complex would take in a chemical reaction, were applied to the presentation of the protein backbone invoking the following four assumptions[61,64]:

(i) The protein backbone can be presented as a smooth, regularly parameterized space curve and is completely characterized by three parameters including its arclength s, scalar curvature κ, and torsion τ where κ and τ are expressed as functions of s.[63] (ii) Secondary and tertiary structure of a protein backbone can be extracted from characteristic features of $\kappa(s)$ and $\tau(s)$, that is, a local presentation of the backbone can be extended in such a way that nonlocal features are obtained. (iii) SSUs such as helices and β-strands can be presented by arrows with specific properties at the second level of coarse graining. (iv) Successive levels of coarse graining introduce nonlocal features and provide a rapid account of tertiary structure.

In Figure 2, tangent vector **T**, normal vector **N** (in applications also called curvature vector), and binormal vector **B** of a Frenet frame at point P_1 of a left-handed helix curve $\mathbf{r}(s)$ are shown. If the Frenet frame moves along the curve to point P_2, **T**, **N**, and **B** readjust their orientation where the rotation of the Frenet frame around vectors **B** and **T** is given by the Frenet coordinates $\kappa(s)$ and $\tau(s)$. The formulas for these quantities are[62,63]

$$\kappa(s) = \left|\mathbf{r}''(s)\right| \qquad [1]$$

$$\tau(s) = \frac{\left\langle \mathbf{r}'(s), \mathbf{r}''(s), \mathbf{r}'''(s) \right\rangle}{\left|\mathbf{r}''(s)\right|^2} \qquad [2]$$

where $|\cdot|$ and $\langle \cdot \rangle$ denote norm and triple product, respectively.

For the purpose of converting the framework of bonds establishing the backbone into a smooth 3D line, two questions were considered: (i) What are the most suitable anchor points for the representation of the backbone? (ii) What type of spline function should be used?

The first question relates to the coarse graining strategy to be applied for the backbone. If all backbone atoms would be used as anchor points, an excessively detailed description of local conformational features with strongly oscillating scalar curvature and torsion would result. In this way, the chance of describing nonlocal features

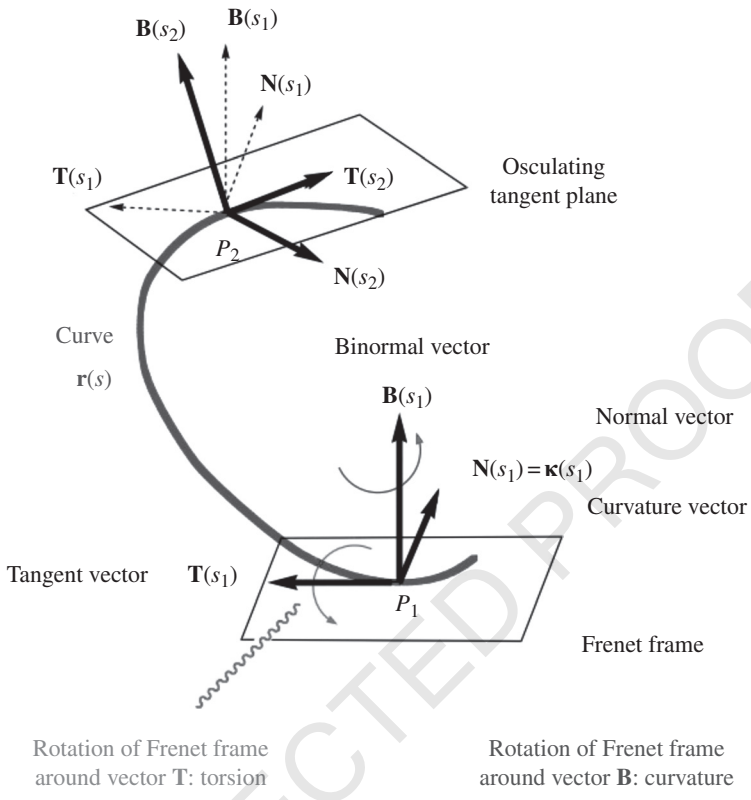

FIGURE 2 Schematic representation of tangent vector **T**(*s*), normal vector **N**(*s*), and binormal vector **B**(*s*) of a Frenet frame at points P_1 and P_2 of curve **r**(*s*) presenting part of a left-handed helix. The normal vector is often also called curvature vector. Movement of the Frenet frame leads to an osculating tangent plane, a rotation at the tangent vector **T**(*s*) (indicated by an arrow around vector **T**), and a rotation at the binormal vector **B**(*s*) (indicated by an arrow around vector **B**) thus specifying torsion and curvature, respectively. For point P_2, the old Frenet frame (dashed arrows) and the new Frenet one (bold arrows) are given, which reveals that the torsion is negative according to a left-handed twist of the Frenet frame around **T**.

of protein structure would be lost. Instead, one has to apply a first level of coarse graining by presenting each residue along the backbone by just one anchor point. This leads to a smoother, more global presentation of the backbone, thus sacrificing unnecessary structural details such as the individual orientation of residue bonds in 3D space.[61]

Various choices for the anchor points exist. These could involve the N or C(=O) atoms of the peptide linkages, alternatively the C_α atoms, or any geometric point of a residue (center of mass, geometric center, etc.). It is easy to see that the residue conformation is best presented at the first level of coarse graining by utilizing the C_α atoms as anchor points. The conformational flexibility of the protein backbone is determined primarily by the C_α atoms and less by the atoms of the peptide linkage.

The impact of the side chains on the protein structure is introduced by the C_α atoms. Hence, any description of protein conformation should reflect the position of the C_α atoms in 3D space. This is in line with the fact that from the early ϕ,ψ-Ramachandran plots[6] to more recent models of protein folding,[79,80] geometrical descriptions of the protein backbone have repeatedly reverted to C_α-based representations. In the same spirit, graphical representations of proteins indicating secondary structure preferentially utilize the C_α positions in 3D space.[81]

It has to be noted that the choice of the anchor points relates to the choice of the polynomial functions selected to represent the protein backbone. Of several possible spline functions that could be used for this purpose, the cubic spline turned out to be most suitable, owing to its simplicity. It is uniquely defined by the requirement of smoothness at the anchor points (C_α atoms) and by two boundary conditions (see the next paragraph). In addition, it corresponds to a curve with minimum deformation energy. Other types of splines do not have this physical property; moreover, the higher-order splines often require extra, nonphysical constraints, as they depend on more adjustable parameters. The cubic spline functions have a well-defined advantage when used with the C_α anchor points, whereas their use in connection with the N, the carbonyl C atoms, or the midpoints of C_α atom pairs does not lead to smooth backbone patterns in terms of curvature and torsion of secondary structures.[61,82]

The cubic spline functions can be used with *natural* boundary conditions at either end, that is, the curvature κ is fixed to zero at the terminal C_α atoms of a protein. This approximation is reasonable considering the high conformational flexibility of the ends of most proteins, which are not associated with a constant κ value. The first and the last two residues of the protein backbone are affected by the boundary conditions, and therefore it is reasonable to exclude them from a backbone analysis in terms of Frenet coordinates.[61]

The accuracy of any representation of a protein backbone by spline functions depends on the accuracy of the coordinates supplied. Therefore, the sensitivity of the cubic spline interpolation to any uncertainty in coordinates at the first level of coarse graining (with the C_α atoms as anchor points) leads to an accuracy in $\kappa(s)$ and $\tau(s)$ values being better than $0.1\,\text{Å}^{-1}$ as long as the resolution R of the protein coordinates is ≤ 2 Å.[61,64]

Once the backbone line of a protein is determined by the APSA method utilizing the Cartesian coordinates of a PDB file,[12] $\kappa(s)$ and $\tau(s)$ are calculated. Based on a set of rules derived from $\kappa(s)$ and $\tau(s)$ values for ideal secondary structure patterns, all secondary structural features of a protein can be analyzed and characterized. Curvature values are always positive, whereas torsion values can be both positive and negative. The sign of the torsion value is defined by the rotation of the binormal vector **B** around the tangent vector **T** of the backbone curve. If it is clockwise (right-handed), the torsion is positive; otherwise it is negative as in the case of the left-handed helix of Figure 2. Equation [2] implicitly includes the direction of the rotation of the binormal vector and, hence, the sign of the torsion. Given that APSA analyzes proteins from the first residue at the N-terminus, any change of sign of $\tau(s)$ correctly reflects any change of chirality of a protein helix or any other structural unit.

THE CURVATURE–TORSION DESCRIPTION FOR IDEALIZED SECONDARY STRUCTURES

APSA was used to compare ideal and real SSUs of proteins.[61,64] Suitable references for the former can be constructed with the help of an 18-residue polyalanine helix and β-strand. For this purpose, ideal ϕ and ψ angles of −57 and −47° for the α-helix, −49 and −26° for the 3_{10}-helix,[83,84] and −57 and −70° for the π-helix[21] were used.[74] A left-handed α-helix was constructed using the ideal backbone angles ϕ and ψ of 57 and 47°.[57] An ideal β-strand is not planar as it is often sterically influenced by neighboring structures, and therefore it resembles a twisted stretched ribbon. This effect was modeled using ϕ and ψ angles of −139 and 135° employing the angles suggested by Hovmoller for antiparallel β-sheets.[85] Paired β-strands were identified by using distance criteria of 6 Å. When the interstrands C_α distance are within this cutoff, the strands were considered as paired.

Table 1 (upper half) contains curvature and torsion values (both in $Å^{-1}$) obtained for the ideal polyalanine SSUs and contrasted with the mean of the curvature and torsion values of 510,525 residues investigated for 73,221 SSUs in 2017 representative proteins taken from the PDB.[12,86] Only X-ray structures having a resolution of about 2 Å or better were selected for this analysis. Structure breaks and proteins with alternate locations provided for C_α positions were avoided though the $\kappa(s)$ and $\tau(s)$

TABLE 1 Curvature κ and Torsion τ of Ideal and Real Secondary Structures the Latter Being Obtained from 2017 Proteins and 510,525 Residues[86]

	Minimum		Maximum		Dihedral Angle	
Ideal SSUs	κ	τ	κ	τ	ϕ	ψ
α-Helix	0.31	0.08	0.55	0.17	−57	−47[a]
3_{10}-Helix	0.29	0.11	0.78	0.29	−49	−26[a]
π-Helix	0.31	0.05	0.47	0.08	−57	−70[b]
β-Strand	0.04	−5.75	0.72	−0.04	−139	135[c]
3_{10} (L) helix	0.29	−0.29	0.78	−0.11	49	26[d]
α (L) helix	0.31	−0.17	0.55	−0.08	57	47[d]
β (R) strand	0.04	0.04	0.72	5.75	139	−135[e]

	Minimum		Maximum		Standard Deviation σ	
Real SSUs with Regular Structure	κ	τ	κ	τ	κ	τ
					Maximum	Maximum
α-Helix	0.30	0.09	0.54	0.19	0.09	0.03
3_{10}-Helix	0.29	0.12	0.65	0.24	0.08	0.09
π-Helix	0.30	0.06	0.45	0.10	0.13	0.17
β-Strand	0.04	−2.26	0.80	−0.07	0.26	1.42

Ideal dihedral angles from [a]Barlow and Thornton,[83] [b]Armen and coworkers,[84] [c]Hovmoller and coworkers,[85] [d]Novotny and Kleywegt,[57] and [e]obtained by sign switch of the dihedral angles of the opposite chirality. Values for real SSUs were obtained from the maxima of the normal distributions shown in Figures 8, 9, and 10.

patterns did not differ significantly in these cases. The set of proteins used for the APSA description contained different sized proteins with various lengths of helices and β-sheets, connected by small- and large-sized loop regions. The CATH classification system[16,17] was followed making sure that the protein examples chosen include all main classes.

Figure 3 presents APSA structure spectra of ideal right-handed helices (Figure 3a), an ideal left-handed α-helix (Figure 3b), and an ideal β-strand (Figure 3c). For both curvature and torsion, there is a clear difference between a 3_{10}-helix (large curvature and torsion peaks), an α-helix (medium-sized curvature and torsion peaks), and a π-helix (small curvature and torsion peaks). The differences in the torsion peaks are larger than those in the curvature peaks, which is a general observation.

An ideal helix has constant curvature and torsion values; for example, for a diameter of 1.0 Å, $\kappa(s) = 0.5$ Å^{-1}. The differences in $\kappa(s)$ and $\tau(s)$ values correspond to the differences in pitch and diameter of each helix. Because a 3_{10}-polyalanine helix has an $i \rightarrow i+3$H-bonding pattern (first and third residues are connected by a C═O ⋯ H─N bond), the diameter of a 3_{10}-helix is smaller than that of an α-helix. Only three residues form a turn (each residue corresponds to a 120° turn) giving a translation step of about 2.0 Å (along the helix axis for one loop). Clearly, the tighter 3_{10}-helix must have larger curvature and torsion oscillations than the α-helix (Figure 3a). For the α-helix, the $i \rightarrow i+4$H-bonding pattern results in 3.6 residues per turn where a residue leads to a 100° turn and the translation step is about 1.5 Å. The π-helix is characterized by the $i \rightarrow$ i+5H-bonding pattern, 4.2 residues per turn, a 87° turn per residue, and a translation step of about 1.15 Å. Hence, oscillations in curvature and torsion are smallest for the π-helix and intermediate for the α-helix. Considering the reduction in the diameters and the translation step of 3_{10}, α- and π-helix (in that order), the amplitude of oscillation in the curvature and torsion must decrease as seen from the $\kappa(s)$ and $\tau(s)$ curves (Figure 3a).

The functions $\kappa(s)$ and $\tau(s)$ oscillate between minimal and maximal values, which are an innate property of the helix in question (Table 1, upper half): 0.29/0.78 (curvature) and 0.11/0.29 Å^{-1} (torsion) in case of the 3_{10}-helix indicating strong curving and torsion of the backbone, 0.31/0.55 and 0.08/0.17 Å^{-1} in case of the α-helix (intermediate curvature and torsion), and 0.31/0.47 and 0.05/0.08 Å^{-1} in case of the π-helix (weak curvature and torsion). The curvature maxima coincide with the position of the C_α atoms (dots in Figure 3a) because they are the points of strong bending. For an α-helix, these points lie on the vertices of the approximate "square" formed when the helix is seen end-on. The plane of the amide bond enforces linearity to the region in between the C_α points resulting in curvature minima (Figure 3a, Table 1).

A similar reduction is found for the torsion maxima of the three helices (0.29, 0.17, 0.08 Å^{-1}; Table 1) reflecting the decrease in the axial translation steps from about 2 to 1.5 and 1.15 Å, respectively (Figure 3a, $\tau(s)$). The strongest torsion is found in the peptide units because these point in the direction of the helix axis giving maximum contribution to the rise per amino acid. At the C_α atoms, the torsion is at its minimum (0.11, 0.08, and 0.05 Å^{-1}, Table 1; see also $\tau(s)$ in Figure 3a).

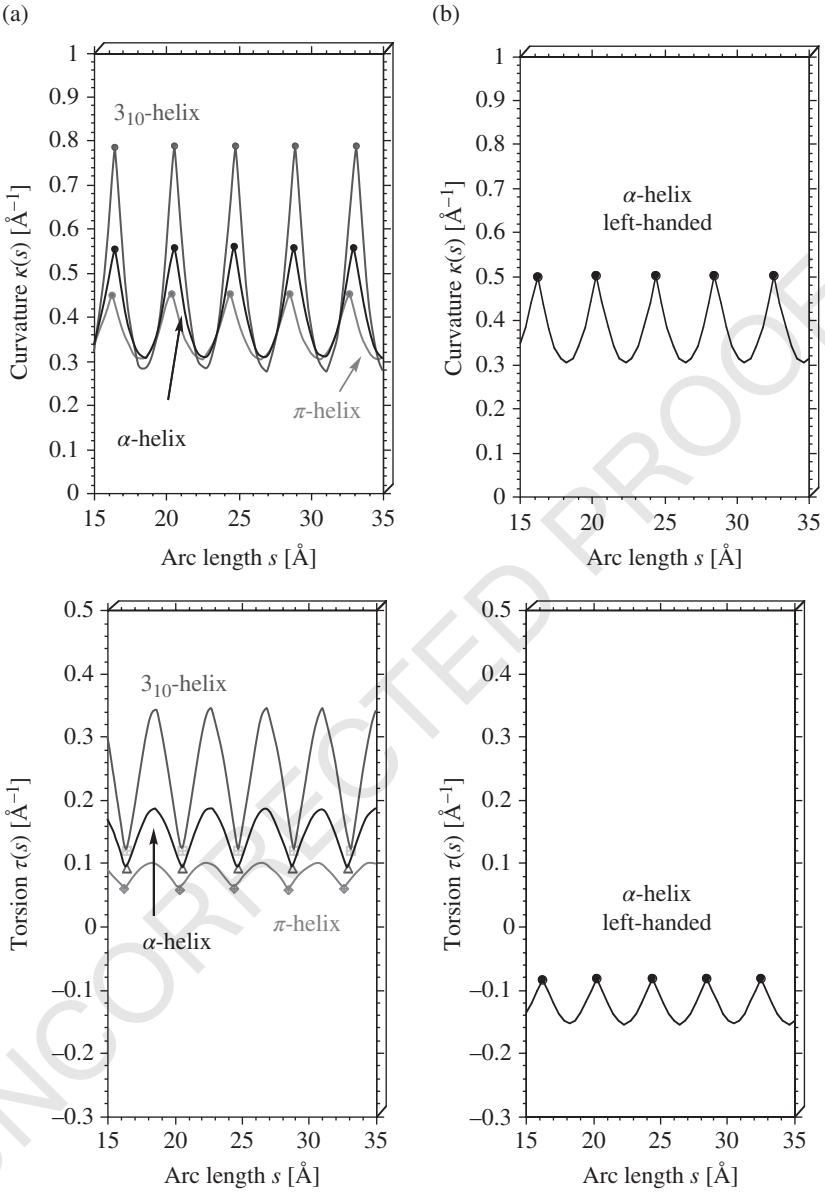

FIGURE 3 Curvature diagrams $\kappa(s)$ (above) and torsion diagrams $\tau(s)$ (below) of (a) right-handed ideal 3_{10}-, α-, and π-helix; (b) left-handed ideal α-helix; (c) ideal β-strand; (d) natural α-helix from N51 to K63 in protein 7AAT[87]; (e) natural β-strand from T17 to K23 in protein 2SOD.[88]

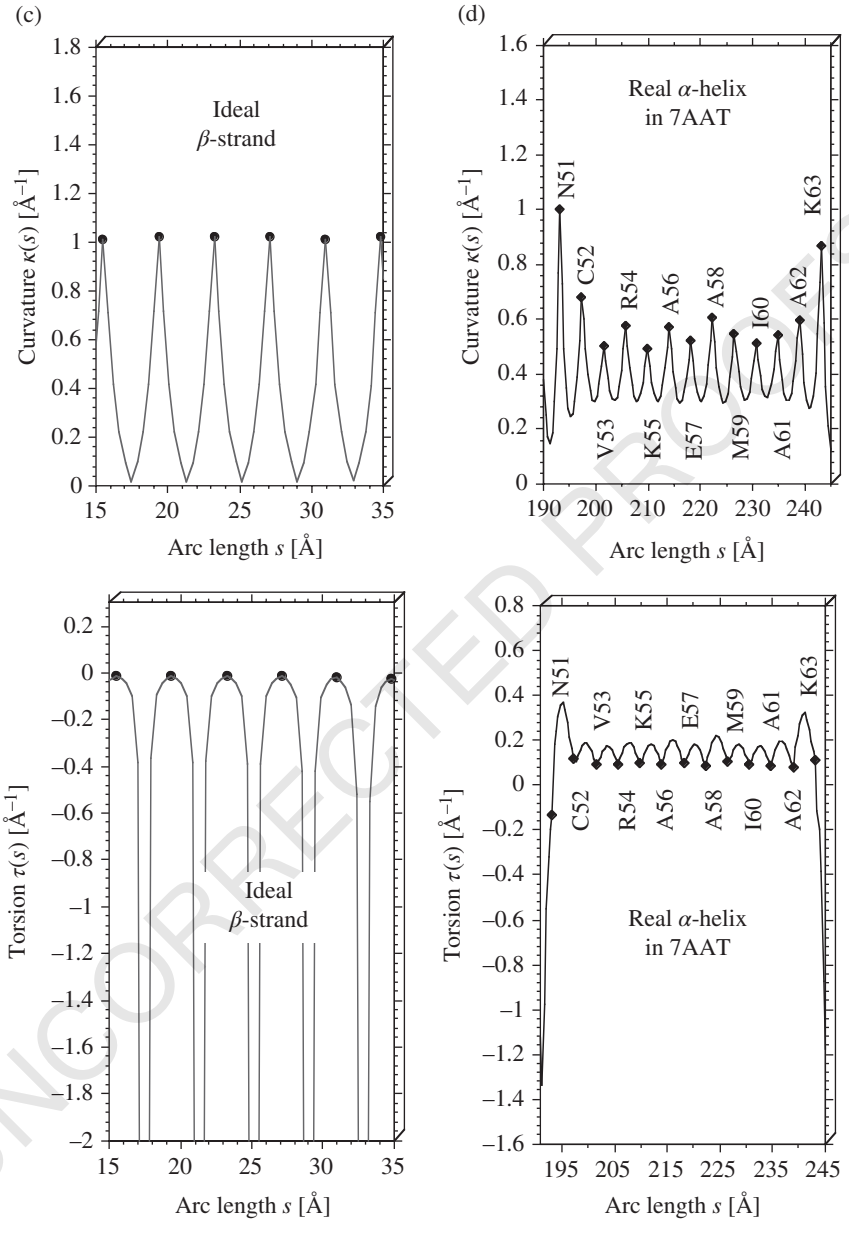

FIGURE 3 (*Continued*)

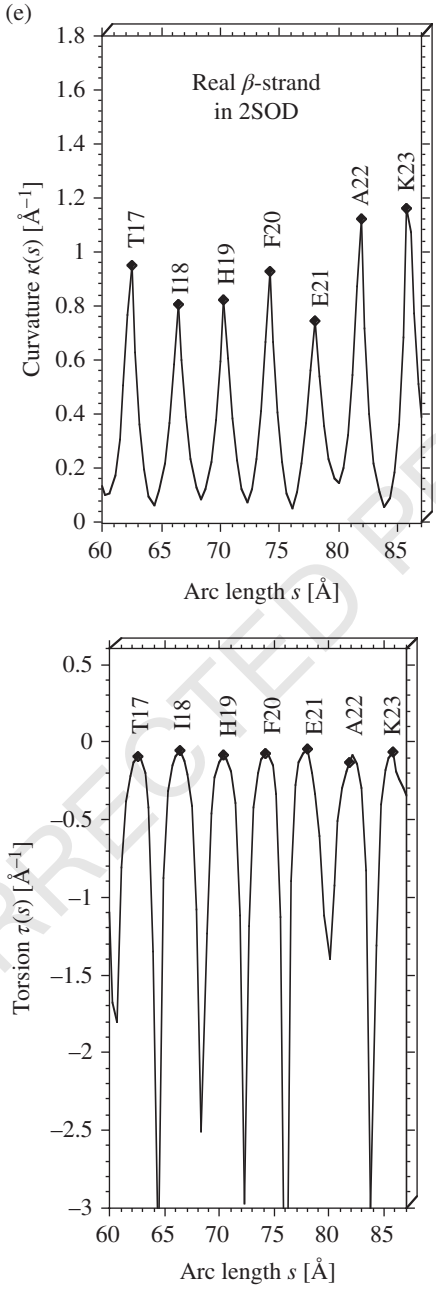

FIGURE 3 (*Continued*)

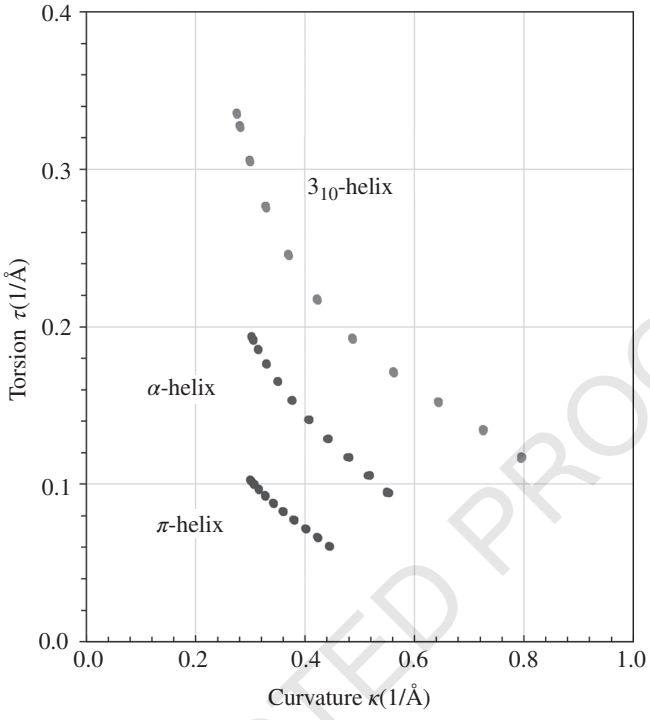

FIGURE 4 The torsion τ is given at 20 points between two consecutive C_α anchor points of the helix backbone in dependence of the corresponding curvature values κ for an ideal 3_{10}-helix, an ideal α-helix, and an ideal π-helix (see text).

Figure 4 gives the relationship between torsion and curvature for ideal helices. The smooth $\kappa(s)$ and $\tau(s)$ curves are obtained by calculating the Frenet coordinates at 20 equally separated points of the backbone line between two C_α anchor points. A small curvature always implies a large torsion and vice versa because the maxima of the $\kappa(s)$ and $\tau(s)$ curves are at different positions: at the C_α (curvature) and between two neighboring C_α points at the peptide bond (torsion). The curve for the 3_{10}-helix has always the largest $\tau(s)$ values and the steepest descent because $\kappa(s)$ and $\tau(s)$ become large because of the tight winding of this helix. The flattest curve is obtained for the π-helix in line with its loose winding.

Right-handed and left-handed helices have differing curvature diagrams (Figure 3a and b), but the differences are not significant. It is more meaningful to investigate the torsion diagram $\tau(s)$ (lower half of Figure 3b), which reveals that a change in chirality leads to a change in sign. For a right-handed helix, torsion values are positive whereas they are negative for a left-handed helix: $-0.17/-0.08\,\text{Å}^{-1}$ (α-(L)-helix) and $-0.29/-0.11\,\text{Å}^{-1}$ (3_{10}-(L)-helix), respectively (see Figure 3a and b and Table 1, upper half).

Ideal β-strands are regular pleats of the protein backbone, rotated along the strand axis where the majority of β-strands are twisted in a right-handed fashion.

This is reflected in the typical patterns of the $\kappa(s)$ and $\tau(s)$ diagrams for the ideal β-strand (Figures 3c). A β-strand can be considered as the sterically stable conformation of a helix that has two amino acids per turn. The C_α atoms are at the positions of a strong curving of the backbone line (high $\kappa(s)$ values) but do not contribute much to the overall rise of the strand ($\tau(s) \approx 0$), whereas the region between the C_α atoms is relatively straight ($\kappa(s) \approx 0$) but responsible for stretching of the strand. The turning of the spline is left-handed resulting in the strong negative torsion minima, which are easily identified in the structure spectrum of a protein. Though the curvature signals of the ideal β-strand and that of the α-helix look similar, it should be noted that the shape and height of the β-peaks are different. The height, given by the difference between the maximum and minimum of each curvature peak, is $0.72 - 0.04 = 0.68\,\text{Å}^{-1}$, whereas for the α-helix it is just $0.55 - 0.31 = 0.24\,\text{Å}^{-1}$.

The structure diagrams of the ideal helices are contrasted with those of some real SSUs taken from proteins of the PDB (Figure 3d and e). The first structure belongs to region N51 to K63 in aspartate aminotransferase (7AAT)[87] and reveals variations in the curvature and torsion maxima that indicate deformations of the helix. The β-strand from region T17 to K23 in superoxide dismutase (2SOD)[88] shows even more fluctuations in curvature and torsion than the ideal structure in Figure 3c. In both cases, the ends of the structures are not as perfect as their ideal counterparts, which is a result of the mutual perturbation at the interface between neighboring SSUs. These examples emphasize that the curvature and torsion patterns are sensitive to any deformation of a real SSU. In addition, they make it possible to exactly describe entries and exits to these units, which opens up various prospects for an automated secondary structure recognition with the help of Frenet coordinates. In this connection, it becomes also obvious that for the $\kappa(s)$-diagrams the differences can only be given on the basis of a quantitative analysis, which seems to indicate that the torsion is more important and easier to analyze quantity than the curvature of the backbone.

The APSA method has been applied to a test set of 2017 proteins (Table 1, lower part).[86] The various residues in the protein backbone can be appropriately determined as part of a helix, β-strand, or random coil by defining lower and an upper limits of $\kappa(s)$ and $\tau(s)$ values in the form of a $\kappa - \tau$ *window* (see the following section) for each specific secondary structure.

IDENTIFICATION OF HELICES, STRANDS, AND COILS

Literature is replete with comparisons of the results of various secondary structure assignment methods. It has been shown that automated methods such as DSSP,[8] DEFINE,[24] and P-curve[25] agree in only 63% of residues among helices, β-strands, and nonregular structures.[89] More recent investigations get to an agreement of 76–85%[25,27] where these numbers depend on the nature of the assignment method, which might be H-bonding or geometry based, fine or coarse grained, or depending on other features of secondary structure. For example, it has been reported that β-strands show greater disparity owing to differing definitions. An automated assignment based on structure comparison involving STRIDE,[23] DSSP,[8] and the PDB[12]

using fuzzy logic[90] revealed that the middle regions of helices show maximum agreement whereas the ends are the disputed regions. In the wake of these problems of structure assignment, it is useful to derive clear definitions of ideal and real helices or strands. In the first case, one can utilize the ideal ϕ, ψ angles published in the literature, convert them to Frenet coordinates with the help of ideal polyalanines, and then use the Frenet coordinates (Table 1, upper half) to identify ideal SSUs in proteins as discussed earlier.

Certainly, one cannot expect that a real SSU of a protein possesses the Frenet coordinates of an ideal polyalanine. The mean values calculated for 73,221 SSUs with 510,525 residues deviate from these ideal values slightly (Table 1, lower half). The structural spectra of the majority of protein SSUs reveal a fluctuation in curvature and torsion values that distinguishes real protein SSUs from ideal ones. These fluctuations are the result of many factors (differences in H-bonding, side chain interactions, environmental influences, etc.) and cause the variation in the Frenet coordinates obtained. The standard deviation of the normal distribution of curvature and torsion values at their maxima and minima has been used to determine a range of values that is typical of regular SSUs and separates them from irregular units with strong conformational distortions such as pronounced bends, kinks, and so on.[74,75] It is useful to contrast the definition of an ideal SSU with that of a real SSU with regular structure (in short *regular SSU*) for which the fluctuations in curvature and torsion do not exceed $\pm\sigma$ where σ is the standard deviation of the normal distribution of the Frenet coordinates calculated for 2017 proteins (Table 1, lower half).

For the test set of 2017 proteins,[86] the total number of helices, strands, and random coils utilizing Frenet coordinates were identified and, among them, the number of helices and strands with regular structures determined. In Table 2, these numbers are compared where the H-bonding-based method DSSP[8] applied to the same set of 2017 proteins was used as a reference.

Difference Between Geometry-Based and H-Bond-Based Methods

The analysis based on Frenet coordinates led to 73,221 SSUs compared to 75,038 SSUs identified by DSSP for 2017 proteins,[86] that is, the number of SSUs differed by 1817 or 2.5 % (the number of residues of helices and strands differed by 14,631 or 5.1%). The largest difference results from the number of random coils (APSA – DSSP: −1745) and the smallest from the number of β-strands (+1054; difference for helices: −1126) indicating that the sum of different SSUs is actually larger (3925 or 5.2%). One reason why less helices were found with the APSA method than with DSSP (difference: −1126) is that the helices identified by APSA are on the average longer, possessing 11.65 residues per helix, whereas DSSP helices had on the average only 10.54 residues. Similar observations could be made for strands (5.54 residues for APSA compared to 5.37 residues for DSSP) but opposite in the situation of random coils (5.80 residues for APSA compared to 5.92 residues with DSSP). APSA recognized more residues as helices or strands. Because the total number of residues analyzed by DSSP and APSA is the same (510,525; actually DSSP excluded 20 terminal residues from the analysis and counted only 510,505 residues; Table 2), the different

TABLE 2 Secondary Structures Identified for 2017 Proteins by APSA or DSSP

	DSSP[a]		APSA Total[a]				APSA Regular[b]			
Type of SSU	# SSU	# Res	# SSU	%	# Res	%	# SSU	%	# Res	%
Helix	16,762	176,745	15,636	93	182,144	103	11,714	75	84,169	46
Strand	20,287	108,954	21,341	105	118,186	108	17,579	82	58,123	49
Coil	37,989	224,806	36,244	95	210,195	94				
α-Helix	15,357	170,003	14,608	95	164,118	97	11,478	79	93,383	57
3_{10}-Helix	1,378	6,600	795	58	3740	57	225	28	741	20
π-Helix	27	142	19	70	87	61	11	58	45	52

[a] The total number of residues is 510,525, of which DSSP missed the last residue in 20 proteins thus yielding only 510,505 residues. Percentages of SSUs are given relative to DSSP numbers. A total of 473 helices confirmed by APSA were not found by DSSP (see text).

[b] Percentages are given relative to the total numbers confirmed by APSA. *APSA regular* is that portion of all SSUs, which have torsion and curvature values in a range being defined by the normal distributions of Figures 8, 9, and 10: *peak of the normal distribution* $\pm\sigma$ where σ is the standard deviation of the normal distribution.

number of SSUs results from characteristic differences in how a geometrically based method (APSA) and an H-bond-based method (DSSP) identify an SSU:

1. H-bonding-based methods have to cope with a number of problems resulting from irregularities of the SSUs of a protein: Any distorted residue can mix up the $i+n$ H-bonding pattern ($n=3$: 3_{10}-; $n=4$: α-; $n=5$: π-helix). In addition, there are bifurcated and trifurcated H-bonds, which make the identification of a helix type even more complicated. Even if methods are applied that assign H-bonding to certain regularities in the preceding part of an SSU, a similar dilemma results at the break points of a helix. A geometry-based method using Frenet coordinates such as APSA does not suffer from these shortcomings and provides well-defined boundaries between SSUs as given by the torsion diagram. The helices identified by APSA are longer and their assignment is closer to that of a crystallographer, whereas an H-bonding-based method such as DSSP is more conservative to set the boundaries of an SSU, thus leading to shorter helices.
2. Frequently, a single winding of an α-helix or just a few residues can switch from the $i+4$ pattern to the $i+3$- or $i+5$ pattern of a 3_{10}- or π-helix. APSA will identify these changes but will also recognize that this is still one helix. Contrarily, DSSP will count multiple helices rather than just one. Accordingly, the number of helices identified by APSA is smaller, whereas the length of the identified helices is on the average longer. This also has consequences for the identification of the helix type. It is reasonable to speak of an α-, 3_{10}-, or π-helix if the character of a given helix is predominated by the properties of this kind of helix. Consequently, the numbers of 3_{10}- or π-helices identified by APSA are reduced significantly compared to those of DSSP. However, this leads to a more realistic and consistent account of the protein structure.
3. H-bonding-based methods such as DSSP will identify a strand only if it pairs with other strands via H-bonding in the fashion of a parallel or antiparallel β-sheet. DSSP searches for residue pairs connecting the strands by H-bonding involving the carbonyl O and the NH groups. Frequently, distortions lead to the fact that H-bonds are missing in the β-sheet pattern. In a geometry-based method such as APSA, both pairing of residues and their conformation are considered with a broad tolerance to imperfections as long as the torsion values indicate the existence of a β-strand. Accordingly, the number of strands identified by APSA is significantly larger than those identified by DSSP. APSA identifies 77,575 residues to occur in 5177 unpaired strands, which are unable to be disclosed by DSSP.
4. H-bonded methods (DSSP) identify more residues as part of random coils. Geometry-based methods (APSA) find that the majority of these controversial residues actually belong to a helix or a strand according to their geometry and Frenet coordinates although they are not identifiable via H-bonding patterns.

The difference between H-bond-based (DSSP) and geometry-based structural analysis methods (APSA) is summarized in points 1–4. This does not imply that one

of these basically different description methods is superfluous: The combination of a geometrical method such as APSA and an H-bond-based method such as DSSP can actually be of advantage. This may be demonstrated for the case of the identification of α-, 3_{10}-, and π-helices. As long as ideal or regular helices have to be described, one can easily distinguish between the three types utilizing Frenet coordinates. With increasing irregularities, the Frenet coordinates of the different types of helices overlap. In this situation, one can combine, for example, APSA and DSSP to obtain more stringent criteria for separating the different helices.

Combination of Geometry-Based and H-Bond-Based Methods

Following the strategy of combining geometry- and H-bond-based methods, one can identify with APSA all helices (Table 2, column *APSA Total*, first row) to make sure that their full length with all residues is recognized correctly. Then, one checks H-bonding for these helices utilizing a procedure also employed by DSSP. The dominant H-bonding pattern (1-3, 1-4, or 1-5) is finally taken as that which determines the dominant helix character. In this way, 14,608 SSUs can be identified according to their Frenet coordinates and H-bonding as being α-helices (i.e., 95% of the α-DSSP helices; Table 2), 795 (58%) as being 3_{10}-helices, and 19 (70%) as being π-helices. This means that of the 15,422 helices identified in total by APSA, 214 helices are not confirmed by DSSP. A simple check with APSA reveals that the majority of the 214 helices are isolated one- or two-turn helices (mostly with four to eight residues) that cannot be identified by DSSP.

Relevant in connection with the ongoing discussion of the appearance of 3_{10}-helices in proteins[91] is the large difference between APSA 3_{10}-helices (795; Table 2) and DSSP 3_{10}-helices (1377; Table 2). It is useful to recognize an SSU only as a helix when at least four residues are arranged in a turn with the typical Frenet coordinates of the helix. If this criterion is reduced to three residues for 3_{10}-helices, 6191 3_{10}-helices will be recognized by DSSP compared to 1174 3_{10}-helices by APSA. The major difference between these numbers results from the exclusion of pseudo 3_{10}-helices interspersed in α-helices by APSA but counted by DSSP. This is also the reason why APSA identifies just 19 of the 27 π-helices as genuine (Table 2) because the other six are located within an α-helix.

It is obvious from the aforementioned discussion and Table 2 that H-bonding is only one descriptor that can be used for specifying the character of an SSU and it is neither necessary (especially at the start or end of an SSU) nor sufficient. However, any geometrically based analysis method such as APSA can benefit from H-bonding information to better distinguish between the various types of helices.

Chirality of SSUs

The chirality of a helix or a strand is identified via the sign of their torsion values (Table 1, Figure 3a/b).[61] In this way, one identifies a significant number of residues to possess left-handed helix torsion. However, only few of them are arranged in sequence to provide a left-handed helix (one α-helix, two 3_{10}-helices; all of them with

just a single turn). Their Frenet coordinates differ from the ideal ones listed in Table 1 only for the 3_{10}-helices (curvature: 0.29 and 0.78 (ideal) vs. 0.30 and 0.61 Å^{-1} (real); torsion: −0.29 and −0.11 vs. −0.23 and −0.11 Å^{-1}), whereas the α-helices have similar or identical values (curvature: 0.31 and 0.55 vs. 0.31 and 0.57 Å^{-1}; torsion values: −0.17 and −0.08 vs. −0.18 and −0.09 Å^{-1}). There are 8456 strands with 18,465 residues among 73,221 SSUs that APSA identifies as being left-handed. These strands are usually short with an average of 2.2 residues. The number of left-handed strands with three or more residues is 1269 (4091 residues).

What Is a Regular SSU?

The values given in the lower half of Table 1 serve as the basis for determining regular SSU structures. The mean values of the normal distributions of Frenet coordinates are close to the ideal values (Table 1). Exceptions can be found for the curvature and torsion maxima of the strands, which possess relatively large standard deviations (Table 1). The ϕ,ψ-angles corresponding to the mean Frenet coordinates are −118 and 135°, thus confirming the strong variation in the strand conformations.

A regular helix structure is given when at least three or four residues in sequence possess Frenet coordinates that are all in one of the ranges specified in Table 2. This definition considers SSUs with smaller fluctuations in a helix or strand still as regular, whereas larger distortions lead to an irregular SSU. According to the values given in Table 2, 75% of the helices and 82% of the strands are regular and the rest are irregular according to the definitions given earlier. Furthermore, it becomes apparent that an SSU with a regular conformation possesses a shorter average length: There will be a decrease from 10.5 to 7.0 residues for helices and from 5.5 to 3.3 residues for strands if just regular forms are considered (see the number of regular residues in Table 2). According to APSA, irregular helices and strands have an average length of 13.8 and 7.5 residues, respectively. Most of the irregular helices and strands are split by DSSP (change in H-bonding) in two or more SSUs although there is no geometrical necessity for this.

Among the helices, more α-helices are regular (79%) compared to 3_{10}- (only 28%) and π-helices (58%; Table 2). As mentioned before, many 3_{10}-helices are mixed with α-helices, and therefore their terminal residues are distorted leading to a reduction in length. This implies that many of the one-turn 3_{10}-helices are irregular in view of the length criteria set up in APSA (at least three or four residues are required for a helix).

The strands participating in a β-sheet are not as rigid as helices. This was already discussed in early literature, indirectly, in terms of the inconsistency encountered when calculating and modeling H-bonds in β-ladders.[37] The distortion of β-strands from ideal structure can be made visible by the superposition of one distorted example on an ideal structure as shown for 2POR[92] in Figure 5a. 2POR is a porin with the β-barrel architecture having 16 β-strands with roughly 46% of the residues being ideal. The backbone of a DSSP assigned β-strand from W19 to G33 is an example of a deviation from the ideal structure as a result of an overall curving of the strand and a simultaneous rotation of the backbone (as seen from the staggering of the red carbonyl oxygen atoms). The $\kappa(s)$ and $\tau(s)$ diagrams (Figure 5b and c) reveal

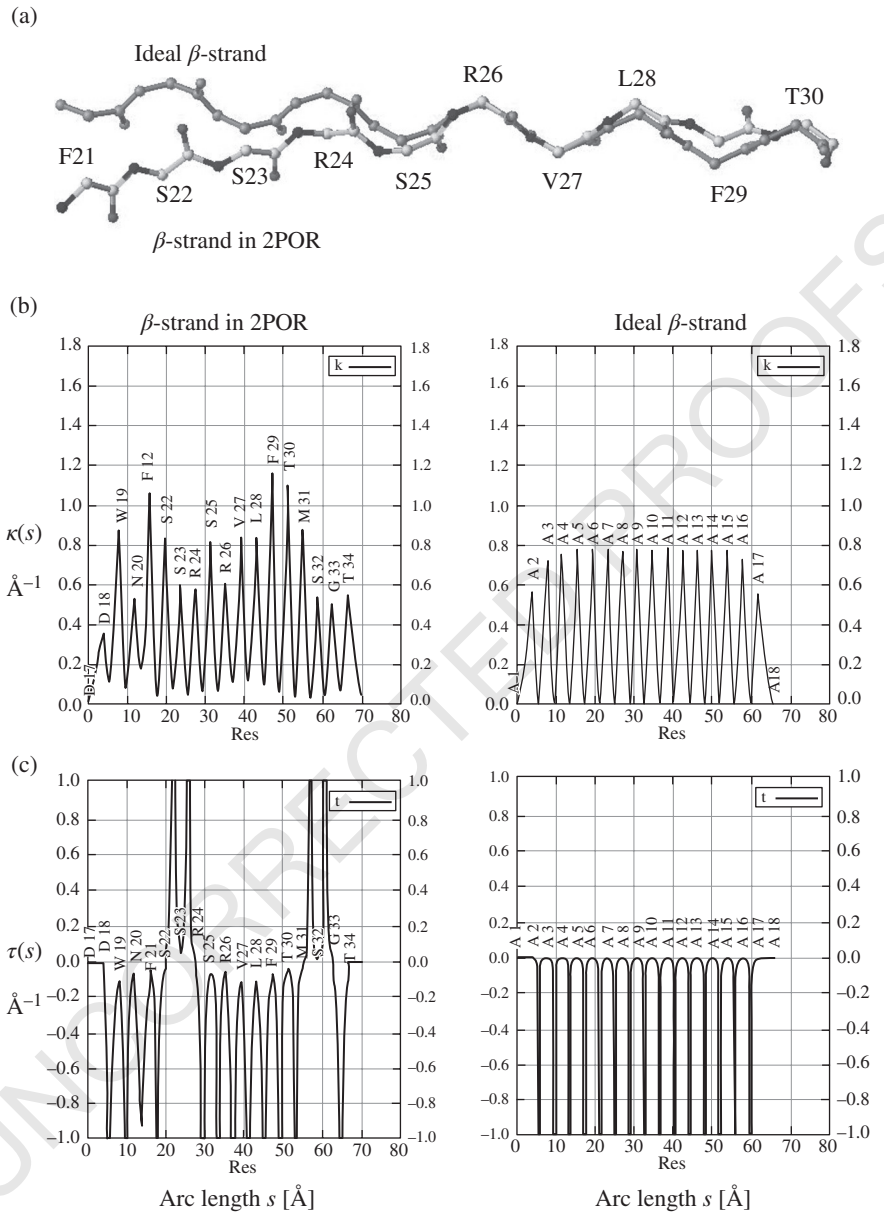

FIGURE 5 (a) A ball-and-stick representation of the backbone region W19 to G33 in 2POR (a porin[92]), which deviates from a strictly ideal β-strand. The position of the carbonyl oxygen indicates torsion of the backbone besides significant backbone curvature. (b) Curvature and (c) torsion diagrams, $\kappa(s)$ and $\tau(s)$, quantify deviations from the ideal β-strand.

these differences clearly (e.g., changes from left to right, back to left, then to right, and finally a left-handed torsion of the strand accompanied by varying curving) and provide a basis for discussing distortions without the need of a 3D comparison (as done in Figure 5a).

A CLOSER LOOK AT HELICES: DISTINCTION BETWEEN α- AND 3_{10}-HELICES

Figure 6 shows the Ramachandran plot obtained from the test set of 2017 proteins consisting of about half a million (510,525) residues.[86] It gives the distribution of SSUs (helices and strands) in terms of the backbone angles ϕ and ψ. Using the ideal values of the SSUs shown in Table 1, one can identify the region of right-handed helices (the high-point-density region in the lower left quadrant) and the strand region (upper left quadrant), which contains also the collagen helix at $\{\phi, \psi\} = \{-66, 158\}$[93] or the polyproline I and II helices at $\{-75, 150\}$ and $\{-75, 160\}$, respectively.[94] Of course, these helices are more extended than a conventional helix. Finally, there are regions with opposite chirality: left-handed helices (right side from $-10 < \psi < 50°$) and left-handed strands (lower right quadrant) where the latter have a low density. Of course, there are many coils of different form interspersed.

Although the helix region in the lower left quadrant has a high-density core, real helices with many types of distortions cover an extended region with densities from 500 (black) to 10 (light-green) samples per ϕ, ψ location. The most populated ϕ, ψ

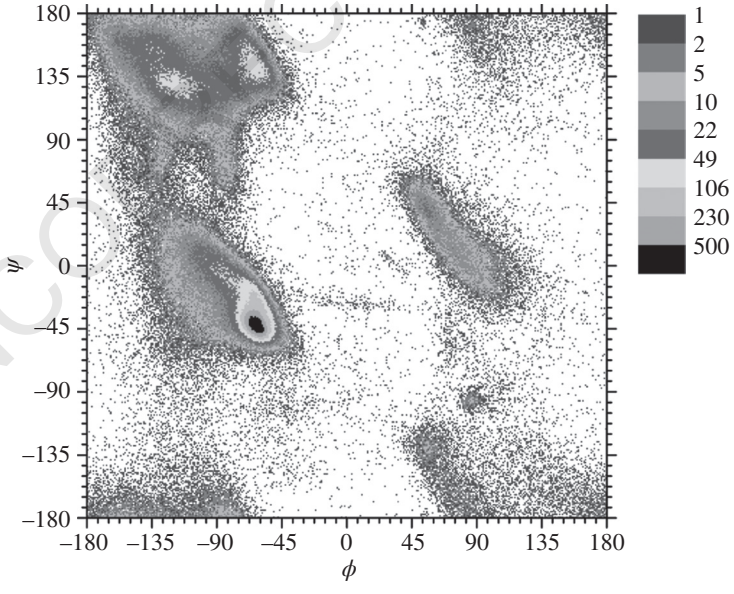

FIGURE 6 Conventional Ramachandran plot given in terms of the backbone angles ϕ and ψ (in degree) for all residues from a set of 2017 proteins.[86] (*See insert for color representation of the figure*.)

locations are at −63.1°, −41.8° for the right-handed helices and −118.1°, 135.1° for the β-strands. These positions differ from the ideal values for the SSUs at −57°, −47° (α-helix) and −139°, 135° (right-handed strands) by 6–20° indicating a larger variation in the strands than in the helices.

Problematic is that in the region between helices and strands, there is no strict boundary or separation zone. On the contrary, there is smooth transition from helices to strands given by a region of low density (Figure 6). For a distorted helix (strand) structure, residues may have coil, helix, or strand properties, which is common when utilizing an atomistic description of the protein backbone based on ϕ and ψ. The right-handed helices are in an elliptically stretched out region given by $40 < \phi < 130°$, $-25 < \psi < 70°$, which suggests a larger deviation from ideal values in these cases. The left-handed strands can be found in low density in a region $145 < \phi < 180°$, $-180 < \psi < -90°$, where of course most of the ϕ, ψ values may relate to coils.

Figure 7 illustrates the curvature–torsion analogue of a Ramachandran plot (which may be dubbed a Ramachandran–Frenet plot) obtained from the sample set of 2017 proteins and half a million residues.[86] In this diagram, two major distribution regions can be identified: (i) In the upper half in a region with low positive torsion and low curvature, there is a high-density region, which corresponds to right-handed helices. (ii) In the lower half in a region with high negative torsion and somewhat higher curvature, there are the extended structures (strands). Similar to the Ramachandran plot, the helix region is highly converged, whereas the extended structures cover a broad region with low density.

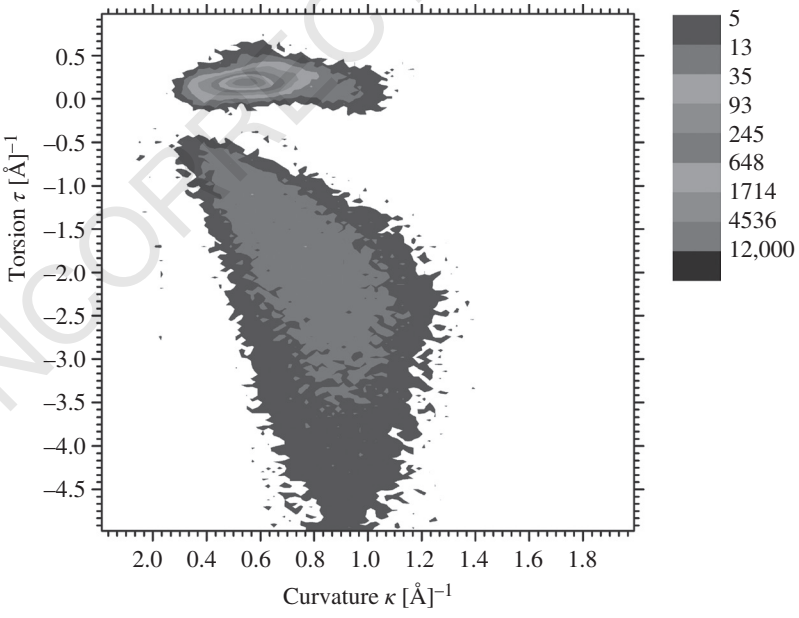

FIGURE 7 Ramachandran–Frenet plot given in terms of curvature κ and torsion τ (both given in Å^{-1}) for all residues from a set of 2017 proteins.[86] (*See insert for color representation of the figure*.)

The distinction between different helices is difficult, as is shown in Figures 8, 9, and 10 as well as Table 3. The figures show distributions of curvature and torsion values where κ and τ are analyzed both at their maximum and minimum values: κ_{base} and κ_p (maximum value positioned at C_α) as well as τ_α (base value at C_α) and τ_p (extreme value corresponding to a peak or trough of $\tau(s)$). For the three secondary structures α-helix, 3_{10}-helix, and β-strand (with the exception of the π-helices for which only a few examples are found), the four distribution diagrams are given in the four subfigures. For all torsion and curvature values considered, α-helices lead to the more narrow and higher peaks in line with the fact that there are more α- than 3_{10}-helices. All distributions are highly converged, that is, data points scatter only moderately.

The distribution functions for the peak values of the 3_{10}-helices are much wider, thus indicating much larger variation and a less regular shape for 3_{10}- than α-helices. This is also reflected by an increased number of examples deviating from the normal distribution fitted to all examples. However, the largest deviation obtained from the normal distributions is found for the β-strands. The torsion trough given by τ_p has a small additional peak close to zero ($\tau_p > 0$). This corresponds to isolated residues with helix structure within a β-strand, often at its ends, thus reflecting the larger flexibility of extended conformations.

By considering the positions and widths of the various peaks, it becomes obvious that the four normal distribution curves of curvature and torsion largely overlap in the

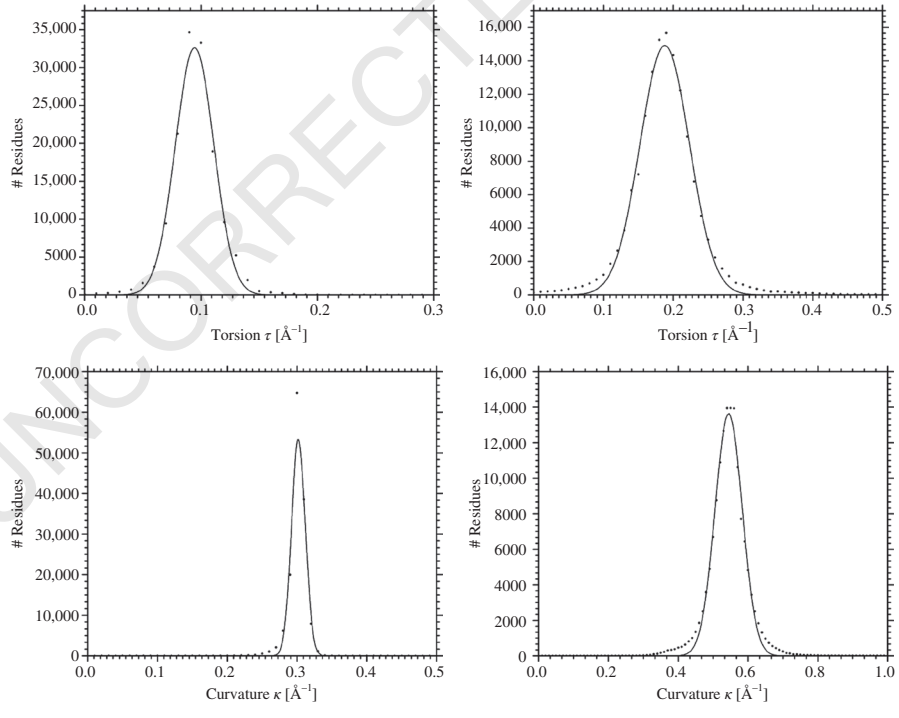

FIGURE 8 Normal distribution curves for all α-helix residues from a set of 2017 proteins. Top left: τ_α; top right: τ_p; bottom left: κ_{base}; bottom right: κ_p.

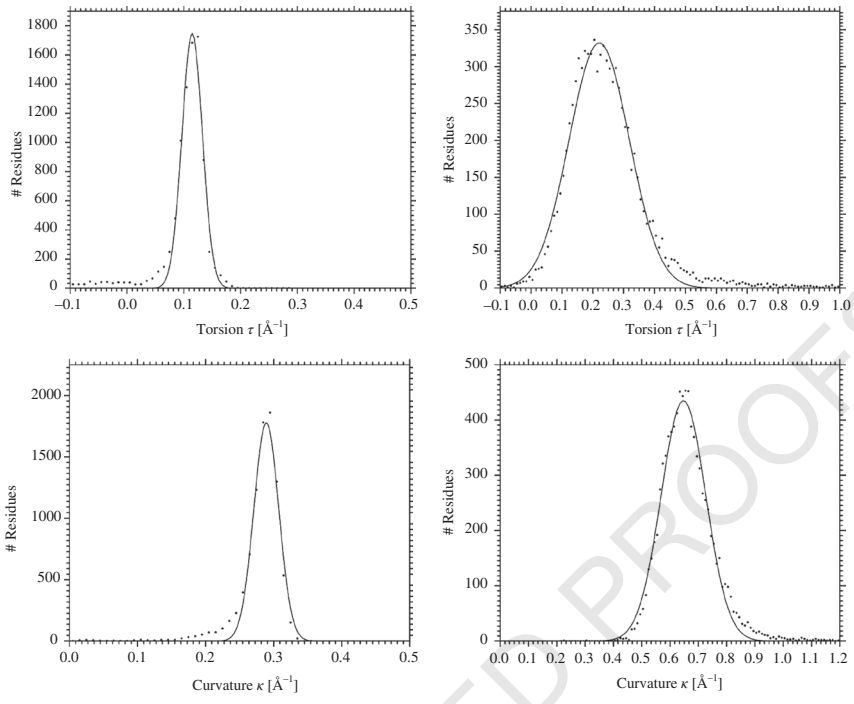

FIGURE 9 Normal distribution curves for all 3_{10}-helix residues from a set of 2017 proteins. Top left: τ_α; top right: τ_p; bottom left: κ_{base}; bottom right: κ_p.

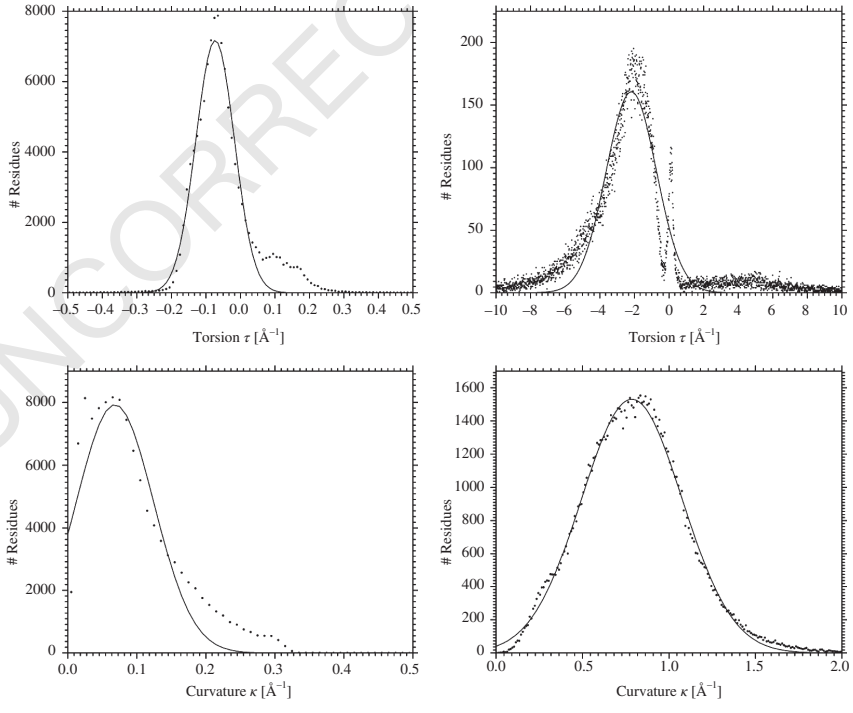

FIGURE 10 Normal distribution curves for all β-strand residues from a set of 2017 proteins. Top left: τ_α; top right: τ_p; bottom left: κ_{base}; bottom right: κ_p.

TABLE 3 Characteristic Values of Curvature κ and Torsion τ Derived from Regular SSUs According to Calculated Normal Distributions for 2017 Proteins[a]

Secondary Structure	κ_{base}		κ_p		τ_α		τ_p	
α-Helix	0.28	0.32	0.45	0.64	0.06	0.13	0.10	0.27
3_{10}-Helix	0.24	0.33	0.46	0.83	0.07	0.16	0	0.45
π-Helix	0.21	0.33	0.14	0.73	–0.06	0.24	–0.28	0.52
β-Strand	0	0.33	0.2	1.44	–0.18	0.05	–32	–0.4

[a]The left entry in a column corresponds to the curvature (torsion) value of the maximum of the normal distributions shown in Figures 8, 9, and 10, whereas the right entry gives the width of the normal distribution. For the function $\kappa(s)$, the peak value κ_p at the position of C_α and the base (minimum) value κ_{base} between two anchor atoms (at the peptide bond) are analyzed. For the function $\tau(s)$, τ_p corresponds to either the peak or the trough of the torsion at the peptide bond, whereas τ_α corresponds to the value at an anchor point.

case of α- and 3_{10}-helices, that is, it is not possible to distinguish between them using their distribution ranges in the Ramachandran–Frenet plot, which is in line with what one finds for a conventional Ramachandran plot (see Figure 6). By using Frenet coordinates, it is possible to identify 3_{10}-helices easily with the help of nonlocal information (as it is always available when analyzing the structure of a protein); however, it is not possible to get a reliable answer if only information for a single residue is available, no matter what kind of coordinates are employed. Therefore, we will take a closer look at helices in the next subsection.

Typical Helix Distortions

In Figure 11, natural helices are compared with the ideal helix of Figure 11a. The body of the helix in 1U4G[95] starting at leucine 135 ($\kappa(s)$ and $\tau(s)$ diagrams of Figure 11b) deviates slightly from the ideal α-helical values where these deviations are not large enough to present a special case of distortion.

Figure 11c displays a typical case of strong helix bending so that a kink results. The kink can be easily realized in the $\tau(s)$-diagram (at L89 of E80–G97 in the E chain of bovine heart cytochrome C oxidase (1V54)[96]), is confirmed by the ribbon representation, but is difficult to identify using the $\kappa(s)$-diagram alone. An even stronger kink shown in Figure 11e for cytochrome P450 (2CPP) can be recognized for both curvature and torsion diagram.

The analysis of helices is often confronted with the problem of ambiguous boundaries as the termini are different from the body. The Frenet coordinates identify these caps directly (see Figure 11f) and reveal how much the caps may differ from the body, even in subtle cases.[61] Often a 3_{10}-conformation occurs at the end of a helix, thus preventing an α-helix from uncoiling and losing its orientation. A 3_{10}-helix occurring at the C-terminus of an α-helix can adopt an α_π-conformation with H-bonding resembling the α-helix pattern and the tilted conformation resembling the π-helices. For example, the region 8–17 of myoglobin (5MBN) has such an α_π-character, which is confirmed by the corresponding κ and τ patterns of Figure 11g. The transition from the 3_{10}-helix into a β-strand is shown in Figure 11h (D18–L37 of 1QTE.[97] A well-extended β-strand can be viewed as a helix with two residues per turn, thus leading to higher curvature peaks than those of a 3_{10}-helix. This trend can be seen for 1QTE at the (positive) τ-peaks corresponding to residues M28, L32, and D34.

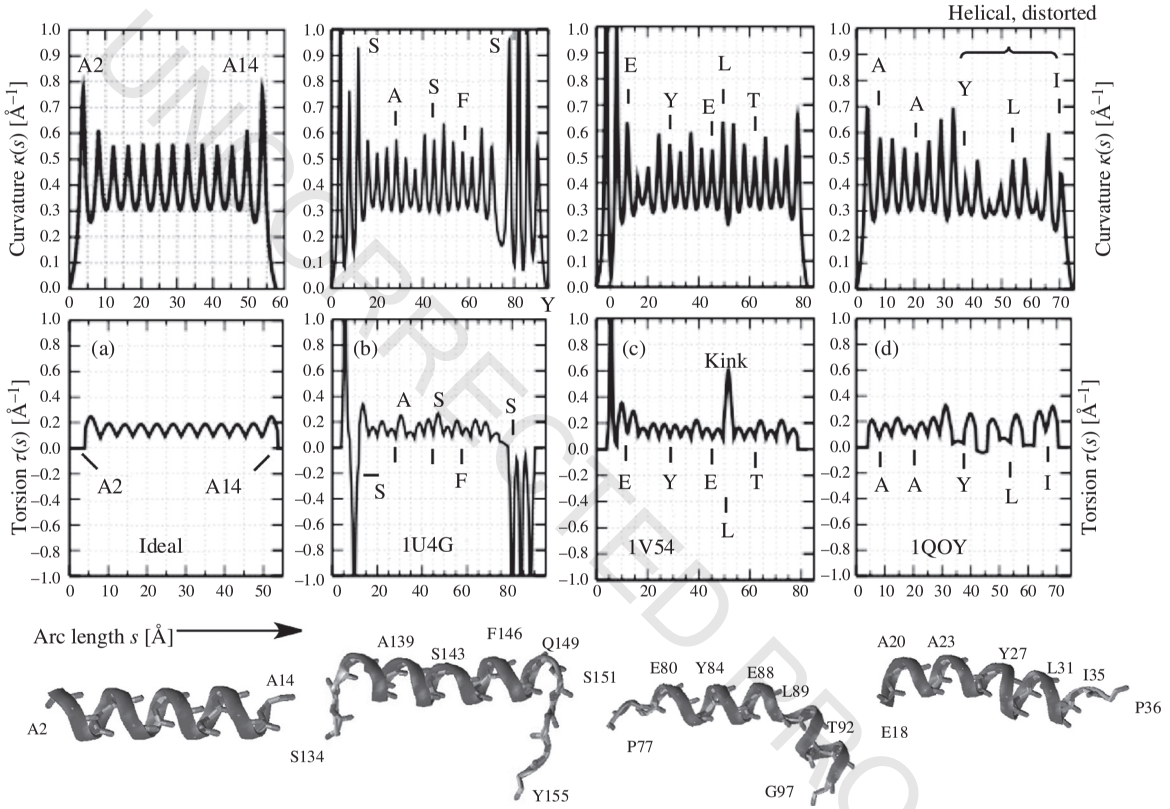

FIGURE 11 Curvature ($\kappa(s)$, above) and torsion diagrams ($\tau(s)$, below) given for ideal and real helices and being complemented by the corresponding ribbon presentation. (a) Ideal 14-residue long polyalanine α-helix. (b) Natural α-helix with small irregularities (1U4G: residues L131–Y155). (c) A kink in an α-helix (1V54: P77–G97). (d) Distortions of an α-helix leading to a looser N-terminus (1QOY: E18–P36). (e) A strong kink leading to a large τ value in a slightly distorted α-helix (2CPP: S258–G276). (f) Difference between body and N-terminus of an α-helix (1TVF: N72–S95). (g) An α_π-cap at the C-terminus of an α-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).[61]

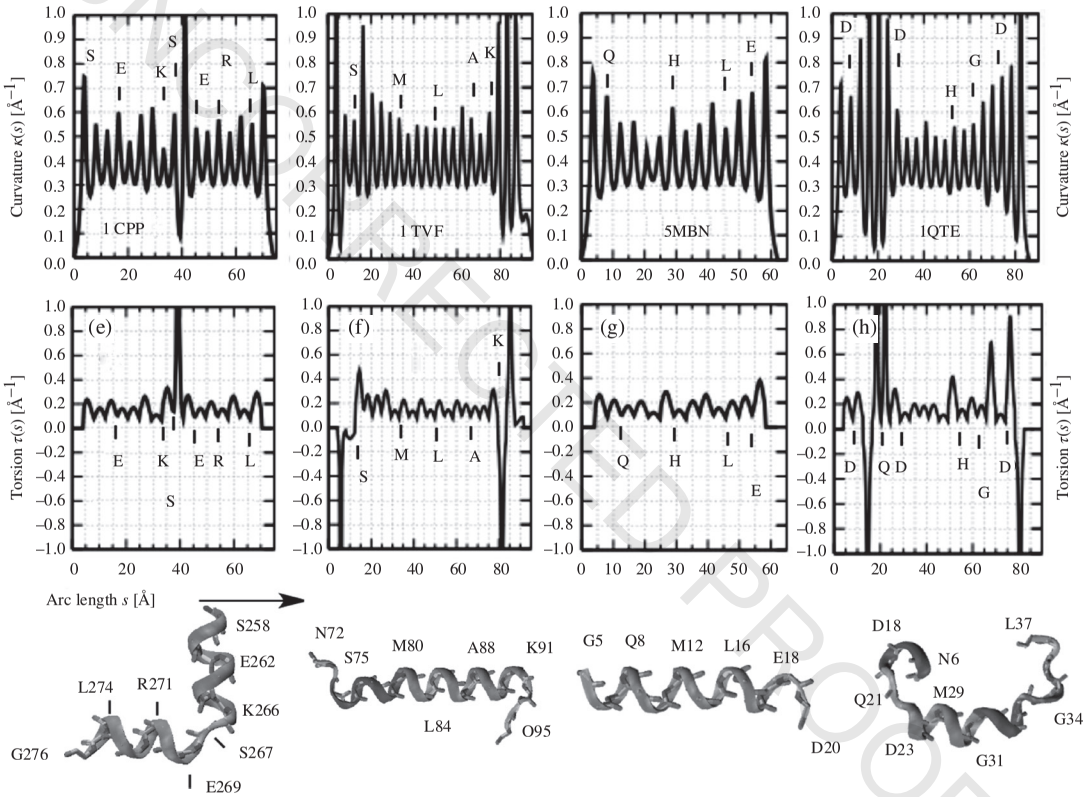

FIGURE 11 (*Continued*)

Level 2 of Coarse Graining: The Curved Vector Presentation of Helices

The representation of a helix as a curved backbone line provides detailed information on distortions, which at level 1 of coarse graining are largely local as they consider the interplay of maximally five residues (in case of π-helices). The impact of the helix on the tertiary structure is not obtained in this way. The latter requires a higher level of coarse graining, which is based on the vector presentation of SSUs as discussed in the following.

A global feature of an ideal helix is its axis, which is defined as the axis of a circumscribed cylinder. This axis defines a vector pointing from the helix start to the helix end, thus indicating its position in the 3D structure of a protein. The presentation of the helix by a vector has been used in the literature for visualization purposes,[98] the investigation of protein structure,[99] the description of helix packing,[100–104] and the similarity analysis of proteins.[105,106] Any bending of the helix axis is usually ignored and the helix axis modeled by a linear vector[107,108] although helix axes are often bent, twisted, or even kinked.[83] Methods of describing linear helix axes and their associated axis vectors have been amply described.[107–111]

More sophisticated representations of the helix axis have to consider a possible bending or twisting of the helix axis.[100] Various methods were published to provide a more realistic representations of the helix including (i) the local axis methods realized in HBEND,[83,112] (ii) HELANAL,[113–116] (iii) P-CURVE,[51] (iv) the QHELIX method[117] based on the algorithms published by Kahn,[109,118] and (v) the MC-HELAN algorithm.[119]

Guo, Kraka, and Cremer[74] developed a method based on Frenet coordinates and a level 2 coarse graining, which absorbs all distortions of a helix into the presentation of the helix axis and thereby provides an accurate assessment of any irregularities of the axis without reverting to the level 1 coarse graining. This method, dubbed HAXIS, implies the following: (i) By using a suitable projection technique, all deviations from an ideal helix are reflected in the projected axis, which thus can adopt an irregular form characterized by bends, torsions, and kinks. (ii) The helix axis is smoothed by appropriate polynomial fitting. (iii) The form of the helix axis, that is, the degree of overall bending and twisting, is quantified in terms of Frenet coordinates. (iv) The overall length and direction of the helix axis are calculated as it changes from the start to the end.

The solution for problem 1 was found by using a mathematically stable procedure of calculating the helix axis piecewise. For this purpose, the unit vectors $\mathbf{T}_i$ and $\mathbf{B}_i$ of a Frenet frame are determined for each residue i of a helix. For a distorted helix, all unit vectors will point in different directions. A relationship between these directions is obtained by moving all unit vectors to a common origin. Then the endpoints of two consecutive vectors $\mathbf{B}_i$ and $\mathbf{B}_{i+1}$ are separated by the distance $b_i = |\mathbf{B}_{i+1} - \mathbf{B}_i|$ and the endpoints of $\mathbf{T}_i$ and $\mathbf{T}_{i+1}$ by $t_i = |\mathbf{T}_{i+1} - \mathbf{T}_i|$. Guo, Kraka, and Cremer showed that Eq. [3] holds[74]

$$\frac{r_b}{r_t} = \frac{b_i}{t_i} \qquad [3]$$

where r_b and r_t are the radii of the circles shown in Figure 12 as defined by anchor points i and $i+1$.

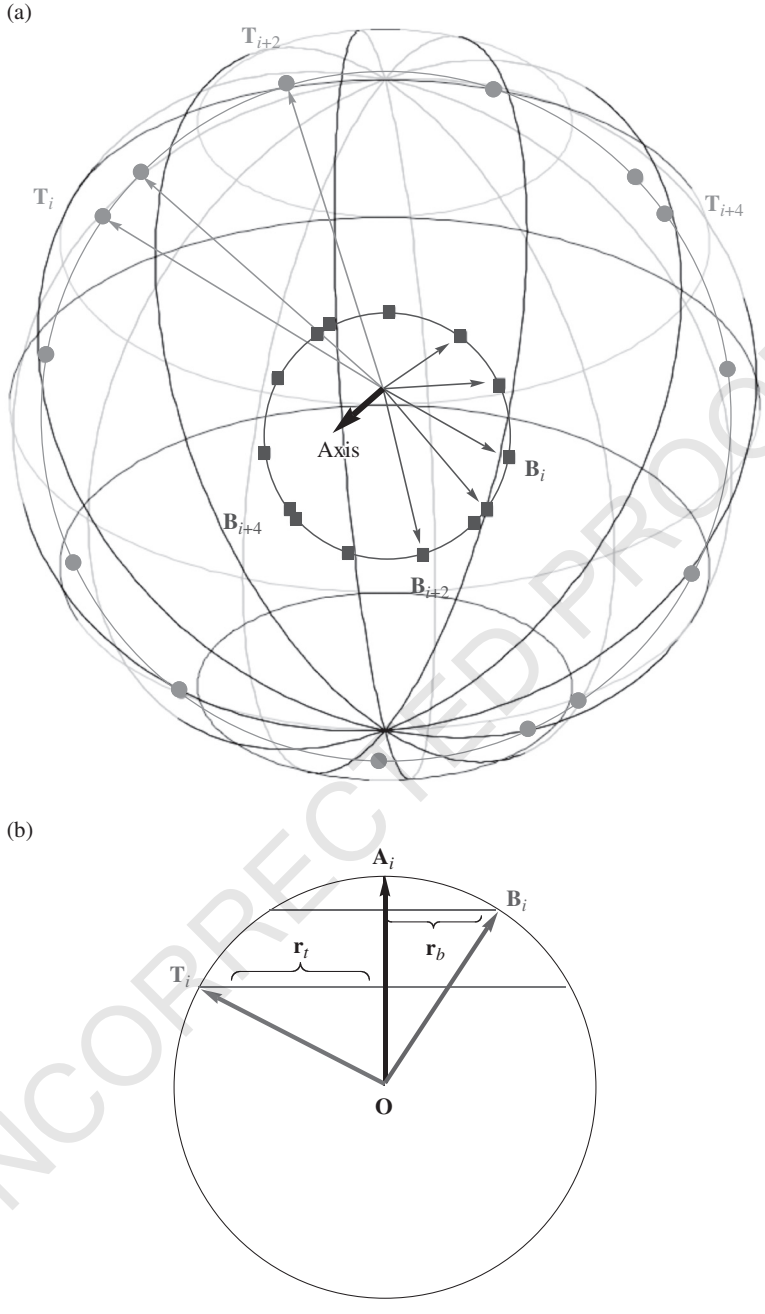

FIGURE 12 (a) If the unit binormal vectors $\mathbf{B}_i$ of all anchor points of a helix are moved to a common origin **O** and the unit tangent vectors $\mathbf{T}_i$ to the same point, their endpoints lie on two circles on the surface of the circumscribed sphere with unit radius and origin **O**. (b) The vectors $\mathbf{T}_i$ and $\mathbf{B}_i$ lie on a slice through the sphere where the trace of the sphere is given as a unit circle and the diameters of the surface circles appear as secants of the unit circle. The parameters r_b and r_t give the radius of binormal and tangent circle, respectively, and make it possible to calculate the direction of the axis vector $\mathbf{A}_i$, which for an ideal helix is the same for all i. For a real helix, the surface circles are distorted and require for the calculation of a circle arc (i.e., the corresponding axis direction $\mathbf{A}_i$) the tangent–binormal pairs of residues i and $i+1$ (see text).[74]

The local direction of the axis is given by an axis vector $\mathbf{A}_i$ according to Eq. [4]:

$$\mathbf{A_i} = r_i\mathbf{B_i} + r_b\mathbf{T_i} \quad [4]$$

In this way, the calculation of **A** by cross products of the unit normal (curvature) vectors $\mathbf{N}_i$ and $\mathbf{N}_{i+1}$ is avoided because this leads to computational instabilities, especially when subsequent curvature vectors point in similar directions.

Using Eqs. [3] and [4] for the m residues of a helix, the total axis is the result of $m-1$ individual axis vectors $\mathbf{A}_i$, which in the case of an ideal helix all point in the same direction whereas for a distorted helix both the scalar vector length a_i (i.e., the *axial translation*) and the direction differ. The axial translation a_i gives the shortest distance between two cor[illegible]d with the anchor points of the corresponding residues ar[illegible]measure for the rise in the helix per residue i.

please hyphen after alpha

In Table 4, properties of ideal α-, 3_{10}-, and π-helices are listed determined by the HAXIS method.[74] Axial translation a per residue, cylinder radius R, pitch p (rise along the axis for one helix turn), and phase angle γ per residue (angle between consecutive unit curvature vectors $\mathbf{N_i}$ and $\mathbf{N_{i+1}}$) are distinguishable for the three types of helices and reflect the tighter winding of a 3_{10}-helix and the looser winding of a π-helix (Table 4). For each of the ideal helices, the axis is a straight line with zero curvature and zero torsion as well as having constant a, p, R, and γ values for each of the residues.

The HAXIS method was used to describe the properties of some selected examples of real helices in proteins (see Table 5).[83,114,120–129] Based on level 2 coarse graining, the following properties are compared: (i) Curvature κ_{av} gives the average bending of the helix axis; (ii) the variation in the axis bending is determined by the curvature difference $\Delta\kappa = \kappa(\max) - \kappa(\min)$ where $\kappa(\max)$ and $\kappa(\min)$ measure the maximum and minimum bending of the axis; (iii) the description of the axis curvature is complemented by the ratio $\eta_k = \Delta\kappa/\kappa_{av}$; (iv) use of the radius of the osculating circle (associated with the curving of the helix axis) as given by $C(s) = 1/\kappa(s)$ to simplify the comparison (s is the arclength of the curved helix axis and defines a position on the axis); and (v) parameters C_{av}, $C(\max)$, $C(\min)$, $\Delta C = C(\max) - C(\min)$, and $\eta_C = \Delta C/C_{av}$ are also given. Because κ is measured in Å^{-1}, C is always given in Å.

Guo, Kraka, and Cremer[74] analyzed 11,761 helices with seven or more residues of which the helices listed in Table 5 constitute some representative examples.[83,114,120–129] The distribution of the κ_{av} values of the 11,761 helices is shown in the form of a bar

TABLE 4 Geometrical Properties of Ideal Helices Using the Cylinder Description

Helix	Translation a (Å)	Radius R (Å)	Angle γ (°)	Residues Per Turn	Pitch p (Å)	ϕ (°)	ψ (°)
α	1.517	2.274	100.1	3.6	5.458	−57	−47[a]
3_{10}	1.955	1.868	121.5	3.0	5.793	−49	−26[a]
π	0.979	2.714	85.2	4.2	4.138	−57	−70[b]

[a]Ideal dihedral angles from Barlow and Thornton[83] and [b]Armen and coworkers.[84]

TABLE 5 Properties of Real Helices Calculated with HAXIS and Compared to Those Obtained with HELANAL and HBEND[a] 74

Protein (PDB)[b]	Helix Residues	κ_{av} (Å^{-1})	$\Delta\kappa$ (Å^{-1})	$\eta_k = \Delta\kappa/\kappa_{av}$	C_{av} (Å)	ΔC (Å)	$\eta_C = \Delta C/C_{av}$	Comp.[c] C (Å)	Type
1RIB	102–129	0.003	0.0001	0.05	333	17	0.05	140	Linear
1MBD	101–118	0.003	0.0001	0.03	311	11	0.03	184	Linear
9PAP	25–42	0.006	0.0005	0.08	167	14	0.08	78*	Linear
1BP2	89–106	0.006	0.0001	0.02	163	3	0.02	71*	Linear
5CPA	174–186	0.007	0.0007	0.10	156	15	0.10	71*	Linear
1BP2	3–11	0.008	0.0002	0.03	131	4	0.03	53*	Linear
2TMN	281–295	0.008	0.0012	0.16	129	20	0.16	100	Linear
1MBD	124–149	0.011	0.0007	0.06	91	6	0.07	88	Curved
1BGE	144–169	0.013	0.0020	0.15	78	12	0.16	73	Curved
1MBD	59–77	0.013	0.0004	0.03	77	2	0.01	85	Curved
4LZT	25–35	0.025	0.0016	0.06	41	3	0.07	58	Curved
7RSA	24–33	0.048	0.0090	0.19	24	4	0.20	34	Curved
1MBD	82–94	0.025	0.0170	0.66	41	27	0.66	24	Kinked
1LIS	44–74	0.030	0.0140	0.47	34	19	0.56	25	Kinked
5CPA	72–89	0.034	0.0110	0.32	31	11	0.37	34	Kinked
2TMN	136–151	0.033	0.0050	0.15	31	5	0.16	35*	Curved

[a]Not all digits are given for the curvature values κ_{av} to obtain the precise η_k and C values.
[b]1RIB: ribonucleotide reductase protein R2[120]; 1MBD: oxymyoglobin[121]; 9PAP: papain[122]; 1BP2: bovine pancreatic phospholipase[123]; 5CPA: carboxypeptidase A[124]; 2TMN: thermolysin[125]; 1BGE: canine and bovine granulocyte colony-stimulating factor (G-CSF)[126]; 4LZT: egg-white lysozyme[127]; 7RSA: ribonuclease A[128]; 1LIS: lysin.[129]
[c]Comparative (Comp.) C values from Bansal and coworkers (HELANAL, first part of table)[114] and Barlow and Thorton (HBEND, second part of table).[83]
*A star indicates a deviation from the classification obtained with HAXIS given in the last column.

diagram in Figure 13. According to the HAXIS analysis, helices can be divided into three major groups:

1. *Linear and quasilinear helices*: They are characterized by C_{av} values larger than 100 Å ($\kappa_{av} < 0.01$ Å, as for example, a C_{av} value of 333 in the case of 1RIB[120] (Table 5). About 18% (2155 helices) of the 11,761 helices investigated are (quasi)linear.[74] According to Figure 13, these are predominantly long, regular helices with more than 13 residues, having a low value of the ratio $\eta_C \leq 0.2$ and displaying a small variation in the axial translation per residue (values a_i) and the pitch values p.
2. *Moderately bent helices*: They have C_{av} values between 30 and 100 Å ($0.01 \leq \kappa_{av} \leq 0.03$ Å^{-1}) and represent the majority of all helices (about 54% or 6321 helices). In this group, helices have an average length of 12.7 residues (but as many as 33 residues are possible).
3. *Strongly bent helices*: Their C_{av} values are smaller than 30 Å ($\kappa_{av} \geq 0.03$ Å^{-1}). Typically, these are short helices with an average length of 10.4 residues (mostly 7–11 residues). The bending of the axis is irregular. They represent the second largest group of all helices with 3285 helices corresponding to 28%.

The regularity descriptors η_κ or η_C seldom correlate with the curving and twisting of the helix axis because large curvature (torsion) also implies large irregularities. However, there is a relationship between relatively long, (quasi)linear, or moderately bent helices, which mostly show high regularity as reflected by $\eta_C \leq 0.15$. About 50% (5889) of all helices were found in this group.[74] The group of helices with moderate irregularities ($0.15 < \eta_C < 0.35$) comprised only 24% (2857) of all helices and involved both moderately and strongly bent helices. Helices with low regularity ($\eta_C \geq 0.35$) were found among the short and strongly curved helices (26% corresponding to 3015 helices).

Noteworthy is a significant torsion of the helix axis, which can adopt τ values as large as 0.08 Å^{-1} (corresponding to a torsion radius of 12.5 Å). The average torsion τ_{av} of the helix axis revealed that the majority of bent helices are also twisted: More than 90% of the helix axes possess a significant torsion value. About 20% of them have low torsion values ($\tau_{av} < 0.04$ Å^{-1}), 60% have values $0.04 < \tau_{av} < 0.28$ Å^{-1}, and 20% have values $\tau_{av} > 0.28$ Å^{-1}, that is, high torsion. Only 9% of all helices were found to have an axis that bends in a plane ($\tau_{av} < 0.01$ Å^{-1}). Hence, a realistic presentation of helices at level 2 coarse graining requires arrows that are curved and twisted.

Identification of Kinked Helices

Helices with kinks are frequent in transmembrane proteins.[130,131] Any identification of a helix kink based exclusively on H-bonding is often impossible because it requires a detailed analysis of the helix geometry.[119] Therefore, Guo, Kraka, and Cremer[74] identified helix kinking utilizing Frenet coordinates to assess the global and local shape of the helix via its axis. According to their investigation, helix kinking is given

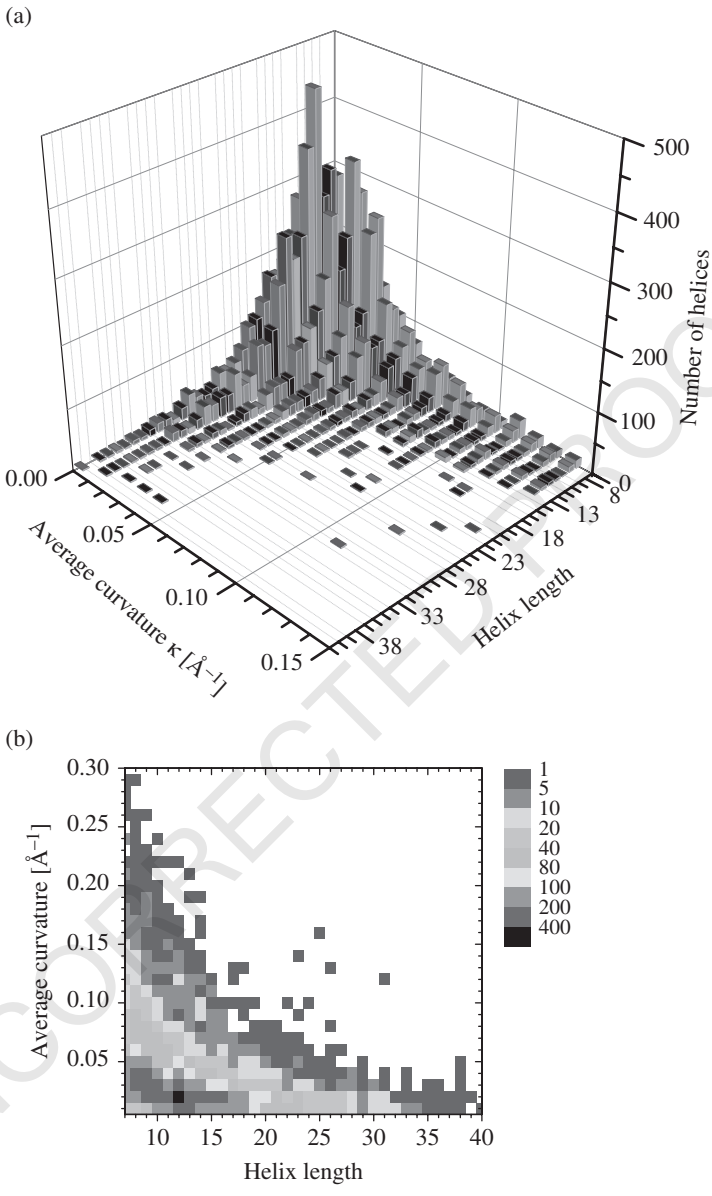

FIGURE 13 Average axis curvature κ_{av} in Å^{-1} of protein helices in dependence of the helix length (i.e., number of residues) given (a) as general distribution in the form of a 3D bar diagram and (b) in a detailed 2D distribution diagram. The curvature of the helix axis is determined after fitting all axis directions with a third-order polynomial. Only helices with 7 or more residues are considered.[74]

when C_{av} is about 40 or less (moderate to strong bending) and complemented by an irregularity descriptor η_C larger than 0.35. In this way, the kinked helices of Table 5 are identified.

Kinking splits a helix into two helices and thereby affects the helix count. Also, a kinked helix has a different function in the protein structure than a normal helix and indicates a strong internal force.[119,132–134] Therefore, the reliable identification of kinked helices is one of the objectives of protein structure analysis.[119,132]

Methods such as HELANAL[113–115] and HBEND[83] assume a regular bending of the helix axis taking place in a plane (any torsion of the axis is assumed to be negligible), and therefore they lead to oversimplified descriptions resulting in a misleading classification of the helix shape. Accordingly, they also miss some of the kinked helices (see starred Comp. C values in Table 5).

Guo, Kraka, and Cremer[74] suggested two complementary methods to describe helix kinking by graphical means. The first is a local description via the axial translation parameters a_i as illustrated in Figure 14 for a part of lysin (PDB id: 1LIS) (from residue 5 to residue 133).[129] For ideal helices, the a-parameter is always below 2 Å (Table 4). Accordingly, 5 helices can be identified in Figure 14 (helix 1: residues 13–37; helix 2: 44–74; helix 3: 82–95; helix 4: 99–107; helix 5: 116–123). These are separated by six coils, which typically have values of $a > 2$ Å. As demonstrated by the example in Figure 14, the a-parameters can be used to determine the start and the end of a helix and to describe coils.

In the center of helix 2, shown in Figure 14, a single a value of 3.16 Å associated with residue 61 of 1LIS denotes the presence of a kink (see also Table 5). Table 6 lists typical axial translation parameter a_i at a kink, which is unusually long between the

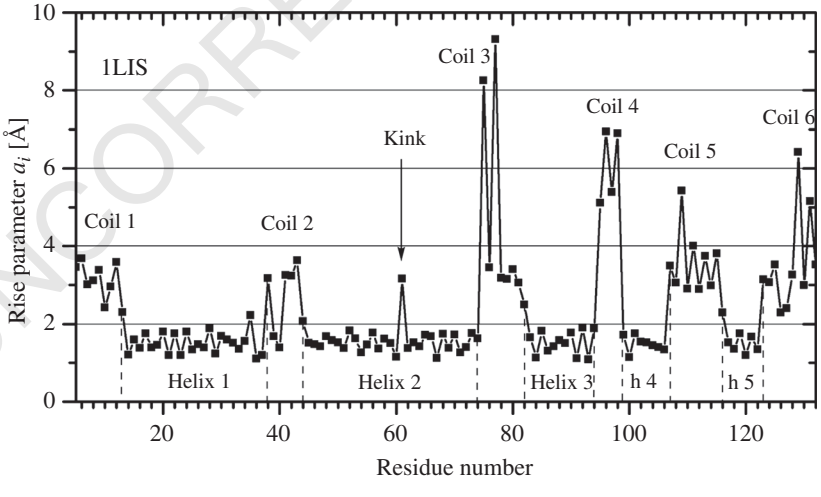

FIGURE 14 Representation of the axial translation parameter a_i for protein 1LIS[129] as a function of the residue number.[74] Five helices (h) and six coils can be identified (dashed lines give start and end of helix). Residue 61 in helix 2 gives the position of a kink. For the identification of residues in helix 2, compare with Figure 15.

please center data in columns 2, 4, and 5

TABLE 6 Identification of Kink Position Using the Maximal Axial Translation Parameter *a*(max) and the Maximum Curvature Value of the Spline-Fitted Helix Axis[74]

Protein (PDB)	Residues	*a*(max) (Å)	Kink Position[a]	Residues Involved in Kink[b]
LIS	44–74	3.16	61	60–63
MBD	82–94	2.58	86	85–88
ECA	52–72	2.19	63	62–65
A6M	83–95	2.66	86	85–88
A8H	379–396	2.44	386, 391	385–388, 390–393
AK0	236–263	2.94	241	240–243
AH7	206–242	2.93	216	215–218
ADE	183–199	3.22	193	192–195
B8O	257–280	3.05	266	267–269
C3D	20–37	2.43	25	24–27

1LIS: lysin[129]; 1MBD: oxymyoglobin[121]; 1ECA: erythrocruorin[135]; 1A6M: myoglobin[136]; 1A8H: Thermus thermophilus methionyl-tRNA synthetase[137]; 1AKO: p1 nuclease[138]; AH7: phospholipase C[139]; 1ADE: adenylosuccinate synthetase[140]; 1B8O: purine nucleoside phosphorylase[141]; 1C3D: human C3d.[142]
[a] According to *a*(max) values.
[b] According to curvature values.

two residues positioned directly at the kink, thus reflecting the extra-strong bending of the helix axis.

Guo, Kraka, and Cremer[74] suggested using the curvature diagram of the spline-fitted helix axis as a second method of identifying the kinks of a helix, thus developing an idea originally based on the work of Ranganathan and coworkers.[61] By determining curvature and torsion of a helix axis, the kinking, which always involves more than one helix residue, can be determined accurately. This is demonstrated by the curvature diagram of the axis of helix 2 in 1LIS shown in Figure 15. Four large curvature peaks arranged in two doublets are found for T60, H61, W62, and A63, thus confirming that a helix kink involves at least the two neighboring residues on each side of the kink. Since the curvature of the axis line depends on the second derivative with regard to the arclength s[143] it is more sensitive to kinking than the axial translation parameter a_i (Figures 14 and 15). In this way, a kinking of the helix is accurately identified, as is shown in Table 6 for some typical examples.[135–142]

The analysis of a curved helix axis can reveal whether a helix is strongly bent. Kinking is characterized by at least one curvature peak of the quadruple configuration being larger than 1.0 Å^{-1}.[74] In this way, 708 kinked helices of a total of 11,761 helices investigated were identified by Guo, Kraka, and Cremer, thus suggesting that on the average 6.0% of all helices are kinked and an additional 4.8% (564 helices) are strongly curved (0.6–1.0 Å^{-1}). Another 5.8% (677 helices) reveal small local irregularities with peak values of 0.4–0.6 Å^{-1}. Other methods[114,119] fail to provide accurate numbers on the percentage of kinked helices as an accurate description of the helix shape is needed to detect all kinked helices and to distinguish them from strongly bent helices.

Figure 16a shows that 44% of the kinked helices have curvature peaks of 1–1.5 Å^{-1}, 19% are in the range of 1.5–2 Å^{-1}, and 37% exceed 2 Å^{-1} where the number of helices

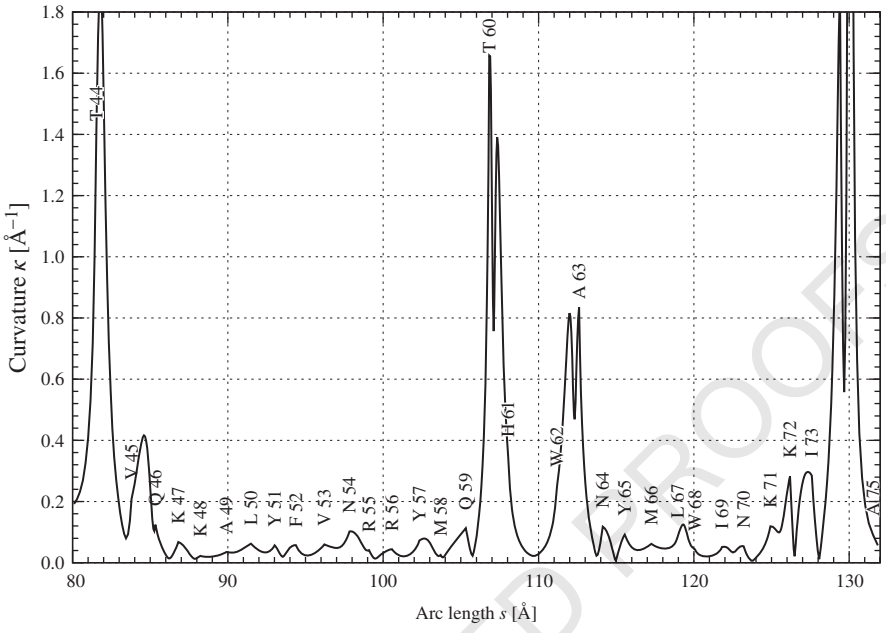

FIGURE 15 Curvature diagram of the spline-fitted axis of helix 2 in 1LIS.[129] Bending as well as kinking involves always several residues: T60 to A63 for the kink in 1LIS. Note that at the ends of the helix the transition to a coil or turn causes an increase in the curvature.[74]

with stronger kinks decreases exponentially. In about one third of all cases, the kink is near the end of the helix, that is, one turn from the end. Helices with a kink are normally short (8–16 residues; Figure 16b). This observation directly relates to the fact that long helices prefer a linear or quasilinear structure. Short helices are less stable and respond to forces exerted on the helix backbone first by bending and then by kinking.

ANALYSIS OF TURNS

The analysis of SSUs such as helices and strands is straightforward with geometry-based methods like APSA. However, turns are more variable and, despite several attempts to describe them,[29,144] remain elusive to any form of strict classification. Note that the term loop refers to the irregular regions of a protein not recognized as turns by the respective classification system used for structure description. These regions comprise about 50% of a protein.[29] Turns have been considered in the literature[29] as a combination of helical and extended geometries. Given that protein structures are reliably described in both ideal and real forms by a level 1 coarse-grained method based on Frenet coordinates, the description of turns in terms of curvature $\kappa(s)$ and torsion $\tau(s)$ is straightforward.

Examples of a type I[145] and a type II turn[146] are shown in Figure 17. The first example is taken from ubiquitin (1UBQ, residues 18–21; darker shading, Figure 17a).

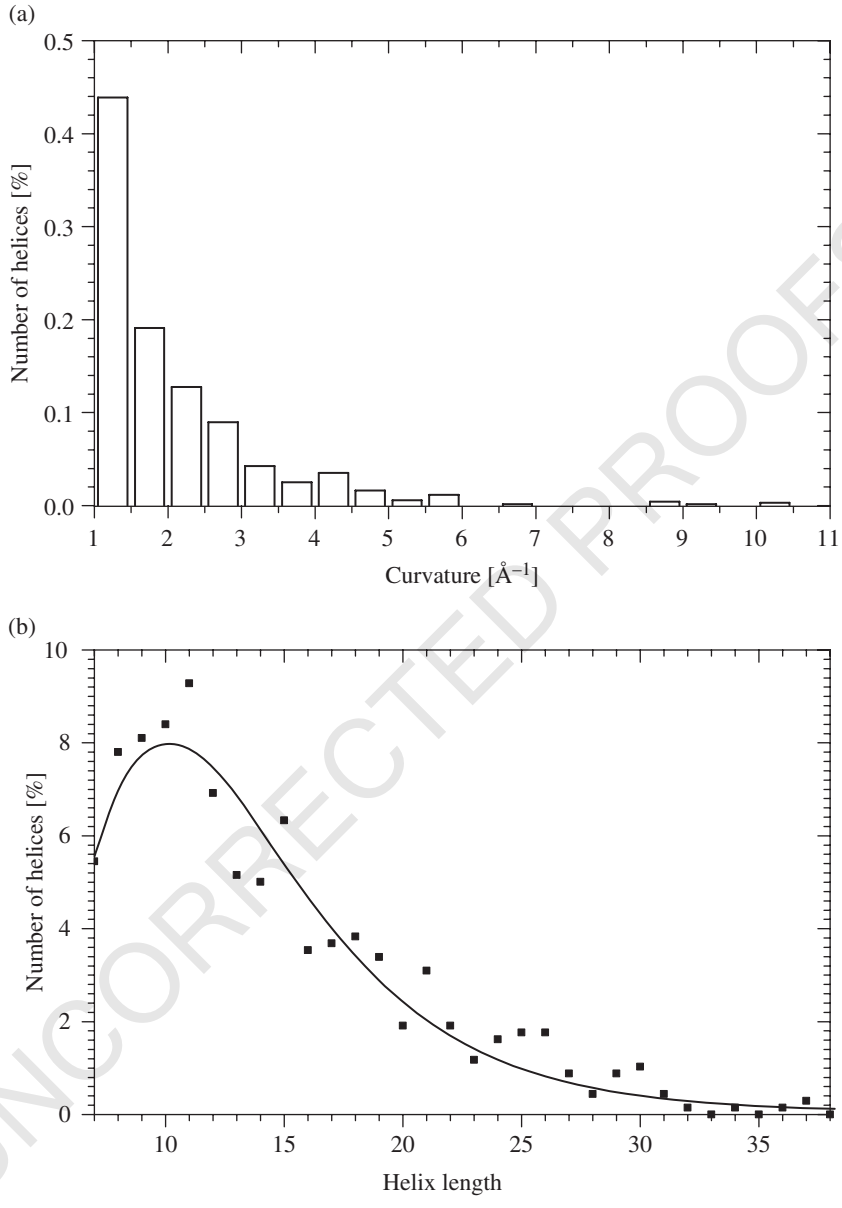

FIGURE 16 Number of kinked helices in percent given in dependence of (a) average curvature values larger than 1 Å^{-1} and (b) the helix length given in terms of the number of residues.[74]

In the $\kappa(s)$ and $\tau(s)$ diagrams of this region, the peaks of the amino acids of the turn are highlighted in bold. The pattern of the turn is not a regular repeating SSU; however, the helical τ-peak at proline 19 and the β-like τ-peaks at E18, S20, and D21 are recognizable. It is evident from the backbone representation that the P19 is mainly

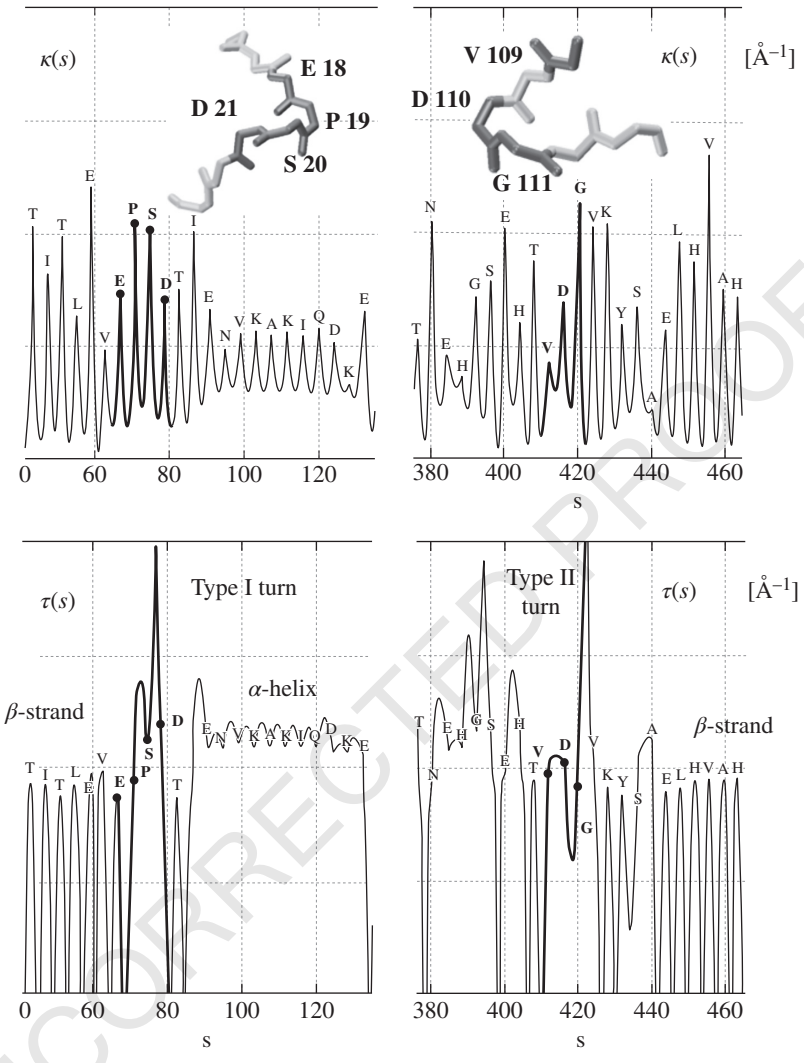

FIGURE 17 Calculated $\kappa(s)$ and $\tau(s)$ diagrams of (a) a type I turn as occurring in ubiquitin (1UBQ)[145] region E18–D21 and (b) a type II turn from carbonic anhydrase form B (2CAB)[146] region V109–V112. Insets: the backbone is darkened in the turn regions.

responsible for the turning of the backbone and that the backbone part going "downward" (inset picture of Figure 17a) raises a little at S20 before it starts going down again corresponding to the sign changes in the torsion. The turn occurs between a regular β-strand to the left and a regular α-helix to the right (Figure 17a). Similarly, in Figure 17b, the type II turn from amino acid V109 to G111 (protein: carbonic anhydrase form B (2CAB)[146]) occurs mainly at the aspartate residue at D110 where the torsion peak is helical. In contrast to the turn example in Figure 17a, the helix is

left-handed as shown by the negative torsion peak at G111 and by the direction of the turn in the backbone picture. Two turn regions (both darkened) can be recognized in the inset of Figure 17b, and the part of the $\tau(s)$ diagram preceding the V109–G111 turn reveals neither helix nor strand pattern. From the $\kappa(s)$ and $\tau(s)$ diagrams, it can be seen that the type II turn is followed by a β-strand distorted in the beginning.

Due to the fact that the $\kappa(s)$ and $\tau(s)$ diagrams of the type shown in Figure 17 are detailed and contain all the conformational information needed, similarity between proteins or parts of proteins in 3D can be assessed easily. To address this issue, six segments from the proteins 1JZB (variant 2 scorpion toxin[147]), 2PAB chains A and B (prealbumin[97]), 2TPI (trypsinogen[148]), 5PTI (bovine pancreatic trypsin inhibitor[149]), and 7RXN (rubredoxin[150]), known to contain the same turn, were analyzed. The $\kappa(s)$ and $\tau(s)$ diagrams of these segments cut out of the respective proteins are shown in Figure 18 with a representative backbone rendering from 5PTI (inset of Figure 18). The $\kappa(s)$ and $\tau(s)$ diagrams of the turn regions are similar in all proteins qualitatively and, to a large extent, even quantitatively.

Of the seven residues shown in each turn of Figure 18, the first is always β-like and extends up to the C_α point of the second residue (see $\kappa(s)$ diagram), after which it turns helical as seen from the high minima (and short peak height) of $\kappa(s)$ and the helix-like shape of the $\tau(s)$ peak. This shape continues through the following residues up to the C_α of the fourth. The fifth and sixth residues are again extended but pointing up and down as shown by the backbone rendering and by the $\tau(s)$ sign changes. The C_α atom of the fifth residue (a glycine in all cases shown) is at the point of a sharp change in backbone orientation as shown by the strong $\kappa(s) > 1.4$ by the $\tau(s)$ changing sign from strongly negative to high positive values. The turn is mostly right-handed as revealed by the positive torsion values in the helical region (residues 2 and 3; Figure 18).

In summary, the characterization of protein structure in terms of Frenet coordinates is not limited to regularly wound SSUs but can be applied in general. This is the basis for a more general and simplified description of protein structure.

INTRODUCTION OF A STRUCTURAL ALPHABET

Chemists have invented an elaborate vocabulary to describe the 3D shape of molecules and how this shape is changed. These terms establish the language of conformational analysis and help chemists to quickly inform each other about the conformation of molecules without reverting to computer-generated 3D images or solid ball-and-stick models of molecules. The usefulness and applicability of the conformational language has its limits. In the case of biomolecules, especially proteins, the manifoldness of possible 3D forms is so huge and the interconversion so complex that without efficient computer representations of biomolecules an understanding of their conformation and 3D shape is hardly possible.

Extending the language of conformational analysis to proteins requires the introduction of a suitable conformational alphabet, which here is based on elementary structural features, thus enabling a simplification of the description of protein structure

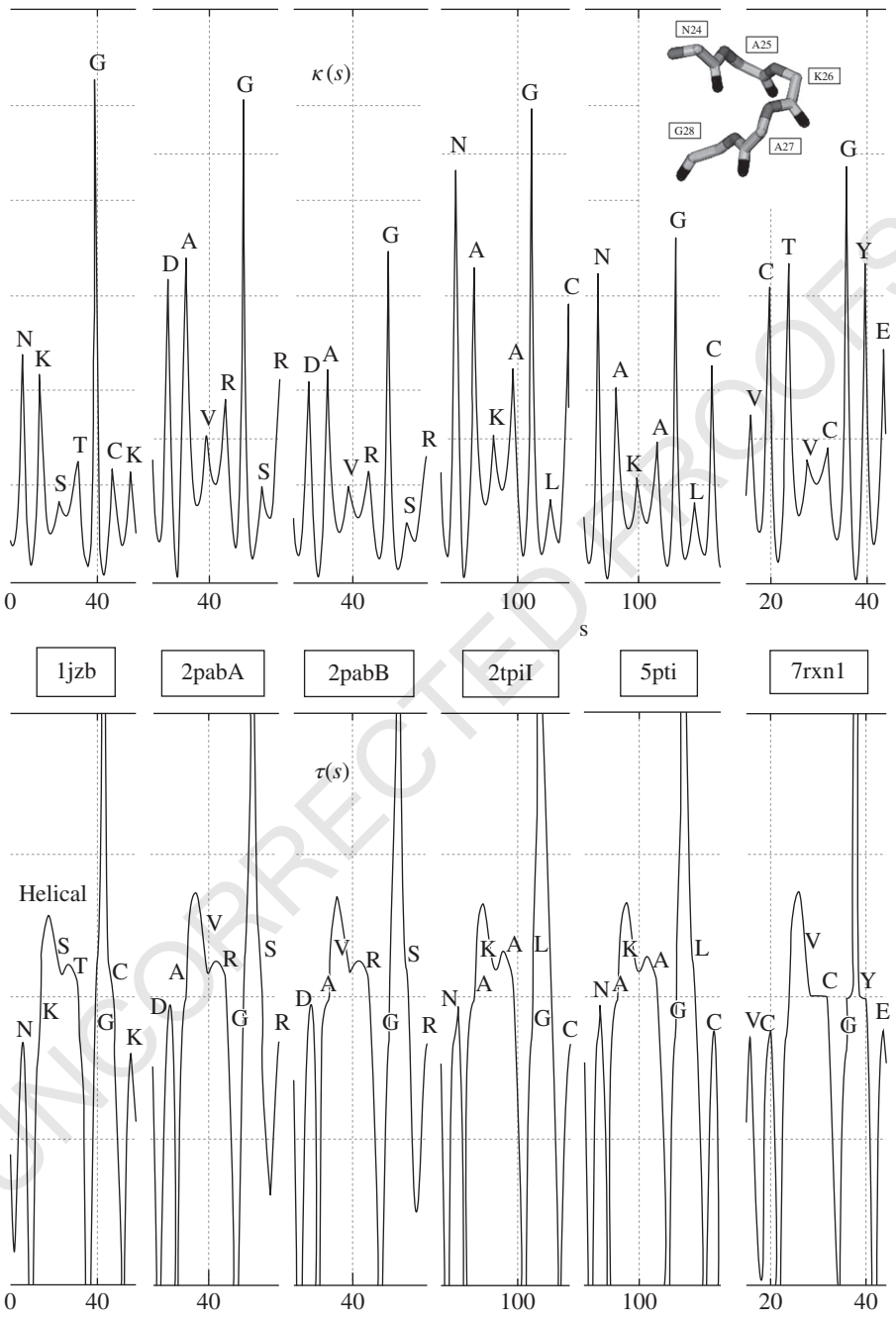

FIGURE 18 Similar turn regions have similar $\kappa(s)$ and $\tau(s)$ patterns. The $\kappa(s)$ (above) and $\tau(s)$ patterns (below) of a turn region (seven residues long) that occurs in six different proteins (PDB notation: 1JZB,[147] 2PAB(A),[97] 2PAB(B),[97] 2TPI(I),[148] 5PTI(I),[149] 7RXN[150]) are shown. The inset gives the turn region for the backbone rendering of protein 5PTI.

while guaranteeing at the same time that the huge variety in protein structure is represented correctly. This is accomplished by following a three-step strategy:

Step 1: The structural features of a protein structure are reduced to those of the protein backbone, which is presented in a coarse-grained form (level 1) properly spline fitted to a smooth line in 3D space.

Step 2: The geometrical features of the 3D protein backbone are described in terms of the Frenet coordinates of the APSA method.[61] This leads to the structural diagrams $\kappa(s)$ and $\tau(s)$ that can be considered as a projection of protein structure from 3D to 2D space. Special features of the 2D structural diagrams can be associated with the SSUs of a protein.

Step 3: Typical curvature–torsion features are expressed in a 24-letter code, which simplifies the identification of these features. In this way, the 2D structural diagrams can be converted into a 1D string of structural letters that represent the 3D structure of a protein.

The advantages of a structural alphabet become immediately obvious. A 1D code of protein structure provides the basis for a rapid description and comparison of 10^5 proteins using moderate computational facilities. Some of the holy grails of protein chemistry can be systematically pursued in this way. For example, protein similarity and protein folding can be described quantitatively, thus leading to a better understanding of protein structure. The elementary question of *ab initio* methods[79,151,152] in computational protein chemistry can be tackled: How does one obtain the 3D structures of millions of sequenced proteins from their 1D sequence without employing time-consuming X-ray diffraction or NMR-based structural analyses? Clearly, if all proteins stored in the PDB are converted into their 1D codes and the relationship between residue sequences and structural 1D codes is statistically analyzed, the discovery of structural rules would be accelerated.

A 1D letter code of a protein can be improved upon by including both geometrical and H-bond information, the simple letter code of a protein can be arranged in *words* (corresponding to SSUs), and words can be combined into *phrases* of increasing complexity (corresponding to supersecondary structures, motifs, folds, families, domains, etc.) until finally *complete sentences* corresponding to the tertiary structure of a protein result. Hence, a suitable 1D letter code can improve computational techniques such as homology modeling,[153–155] threading,[156–158] or *ab initio* prediction methods[159,160] in general.

Now that steps 1 and 2 have been sketched in the previous sections, we show in the following the derivation of a 24-letter code suitable for the description of protein structure.

Derivation of a Protein Structure Code

A geometry-based description of the protein structure in terms of Frenet coordinates is suitable for developing a structural code. Of the two Frenet coordinates curvature and torsion, the latter turns out to be more sensitive to conformational changes than

the curvature, thus making it easy to keep track of any structure variation taking place along the protein backbone. This is a direct consequence of the fact that the curvature κ is a second derivative quantity of the backbone line, whereas the torsion τ is a third derivative quantity and therefore more nonlocal than κ.

The curvature measures the rate of change of the tangent (the direction) of the spline-fitted backbone whereas the torsion measures the rate of change of the osculating plane, that is, the changes of both the curvature and the tangent vector are registered by the torsion, which justifies considering the torsion as a "3D curvature." Accordingly, SSUs are more easily identified with the help of the torsion than with the curvature diagrams. Helices, strands, and other SSUs possess 3D rather than 2D structures. Therefore, the torsion τ is more effective than is the curvature κ when describing protein structure.

The torsion parameter has the additional advantage of identifying the chirality of any helical or (twisted) ribbonlike structure (positive τ values: right-handed twist; negative τ values: left-handed twist). Hence, the torsion parameter is well suited for deriving a structural code. For this purpose, the continuous $\tau(s)$-diagrams are replaced by a (continuous) sequence of τ-windows where each window is associated with a residue (given by its C_α position) and represented by a letter. This is possible because each residue of the protein backbone leads to a τ-peak, τ-trough, or τ-base value. Each window includes information that identifies each peak, trough, or base in a τ-diagram. These are the values τ_α and τ_p (in the region up to the next C_α atom) where the latter corresponds to the height (depth) of the τ-peak (trough) located in this region.

The ratio τ_p/τ_α can become very large because the base value of $|\tau_\alpha|$ is often in the range 0.02–0.1. So, it is more appropriate to compare τ_α and $\log(\tau_p)$. When the two torsion values deviate significantly from one another, both sets of information help differentiating between different windows and, by this, different conformations.[75] An interesting alternative exists, which does not use τ_p because of the problems arising when comparing it with τ_α. Instead of τ_p, the angle γ is used. This is the angle between two consecutive binormal unit vectors $\mathbf{B}_i$ and $\mathbf{B}_{i+1}$ at the anchor points of residues i and $i+1$ (see Figure 2).

Changes in the osculating plane and thereby changes in the torsion are reflected by changes in the binormal vector $\mathbf{B}$. The derivative $d\mathbf{B}/ds = -\tau\mathbf{N} \approx \Delta\mathbf{B}/\Delta s$. However, if the Δs is too large, an average τ value (τ_{av}) that scales the length of the curvature vector $\mathbf{N}$ has to be considered. Angle γ measures the change in the $\mathbf{B}$ direction and by this also the change in the osculating plane thus providing another information on τ. Angle γ can adopt angles between −180 and +180° and its sign provides chirality information, which is identical to that given by the torsion τ. Accordingly, only the sign combinations +,+ and −,− are possible (combinations +,− and −,+ are forbidden) in a γ,τ-coordinate system based on the whole range of possible γ values. For the purpose of simplifying the situation, the absolute value of γ is used so that the sign of τ alone determines the chirality.

In Figure 19, the τ_α–γ coordinate system defined in the way described is used to identify the position of the residues of 33 proteins (the first 31 of the 2017[86] plus 1UBQ and 1A70) in terms of their structural properties expressed directly or

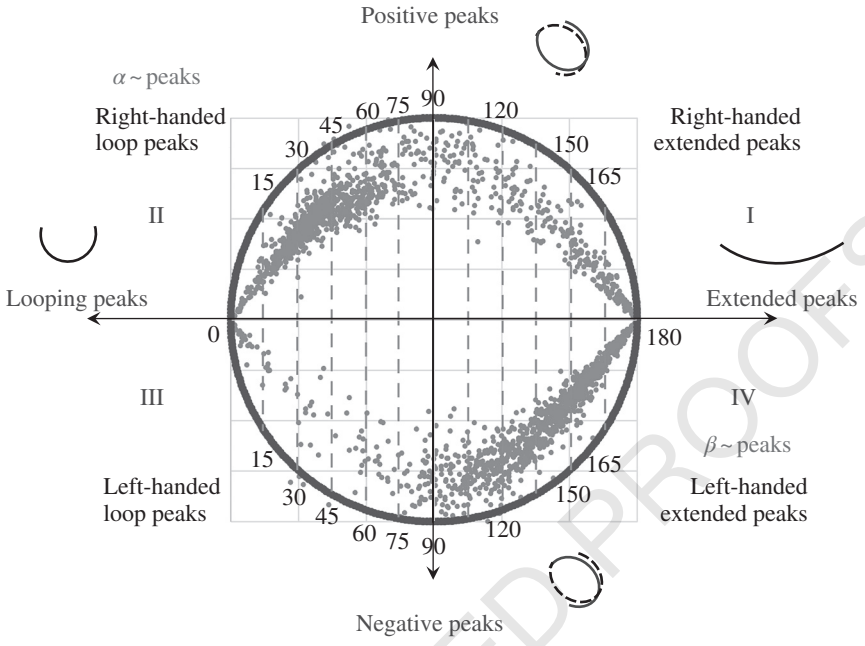

FIGURE 19 Distribution of the torsion characteristics of the residues of 33 proteins in a two-dimensional phase diagram spanned by the parameters γ and τ_α, which reflect properties of the torsion peaks assigned to each residue. The majority of $\gamma \cdot \tau_\alpha$ points are found within the area of a circle indicated by the bold line. Specifically, looping or extended conformations and, utilizing the torsion peak sign, left- (–) or right-handed (+) twisting of the backbone can be distinguished (see text).

indirectly via their Frenet coordinates.[86] The horizontal axis is chosen to be that of γ. Because γ is an angle, only values in the range $-180° \leq \gamma \leq 180°$ are possible, which limits the available 2D space. The vertical axis is taken as the $\pm\tau$-direction. By normalizing the $\pm\gamma$ distance on the horizontal axis and the largest $\pm\tau$ span to ± 1, all τ_α, γ data points of 33 proteins are found (with only a few exceptions) in the circle area indicated by the bold circumference. The distribution of γ, τ_α points is similar when 2017 proteins leading to half a million residue conformations are used to generate Figure 19.

By using the 2D coordinate system of Figure 19, two important properties can be described: torsion and chirality of the protein backbone at the position of each residue where the latter property is given by the sign of τ. The population of the four different quadrants of the circle numbered from I to IV in Figure 19 are identified with the help of Table 7: in quadrant II (–/+ signs of coordinates) the right-handed structures with typically loop peaks, that is, especially the normal α-helix cluster, and in quadrant IV (+/–) the structures with left-handed extended peaks, that is, the normal right-handed β-strands. Left-handed β-strands are found in quadrant I (+/+) and the few left-handed helices in quadrant III (–/–).

TABLE 7 The 24-Letter Code for the Description of Protein Structure (τ_α,γ System)[a]

		γ		τ_α			
Code	Group	Min	Max	Min	Max	Average	Comments
D–	R-strands	180	165	–0.049	0	–0.023	Most ext. R-strands
E–		165	150	–0.092	–0.034	–0.063	ext. R-strands
T–		150	135	–0.135	–0.061	–0.098	R-strands
R–		135	120	–0.172	–0.086	–0.129	R-strands
B–		120	105	–0.200	–0.098	–0.149	R-strands
J–		105	90	–0.231	–0.092	–0.162	Coil
I–	L-helix	90	75	–0.266	–0.066	–0.166	Coil
G–		75	60	–0.287	–0.031	–0.159	Coil, L-helix
V–		60	45	–0.258	–0.019	–0.139	Coil, L-helix
A–		45	30	–0.220	–0.003	–0.111	L-helix
H–		30	15	–0.139	–0.002	–0.071	L-helix
N–		15	0	–0.058	0	–0.023	Coil, L-helix
N+	R-helix	0	15	0	0.061	0.027	Distorted helix
H+		15	30	0.030	0.112	0.071	Loose helix
A+		30	45	0.068	0.130	0.099	α-Helix
V+		45	60	0.065	0.183	0.124	3_{10}-helix
G+		60	75	0.061	0.232	0.146	Helix entry
I+		75	90	0.067	0.270	0.168	Distorted helix, coil
J+	L-strands	90	105	0.080	0.258	0.169	Coil
B+		105	120	0.076	0.233	0.155	L-strands
R+		120	135	0.068	0.190	0.129	L-strands
T+		135	150	0.047	0.148	0.097	L-strands
E+		150	165	0.025	0.095	0.060	ext. L-strands
D+		165	180	0	0.046	0.020	Most ext. L-strands

[a]The abbreviations *min*, *max*, and *average* denote the smallest, largest, and average γ or torsion value of a given γ,τ_α-box associated with a specific letter code.

Because the scattering of data points is along the lines of a square (after an appropriate adjustment of the axes), it is straightforward to define 12 γ-increments of 15°, which partition the range from 0 to 180° (Figure 19) where each of the subranges corresponds to a specific residue conformation. Rather than using the whole $\Delta\gamma$ region, it is reasonable to exclude the less or not populated area closer to the center of the bold circle in Figure 19 and instead define vertical boxes by setting upper and lower τ_α limits so that the densely populated areas along the square are included. These γ,τ-boxes defining the different types of residue conformations are shown in Figure 20a.

There are 12 boxes for positive τ_α values (upper half of diagram) and 12 for negative τ_α values (lower half of diagram) where the $\Delta\tau$ ranges differ because of the greater scattering of data points for the less frequent chiralities. Each of the 24 rectangular boxes in Figure 20a is identified by a letter and a sign, thus leading to a 24-letter code for the rapid description of residue structures along the protein backbone. The letter code is explained in detail in Table 7. For reasons of comparison,

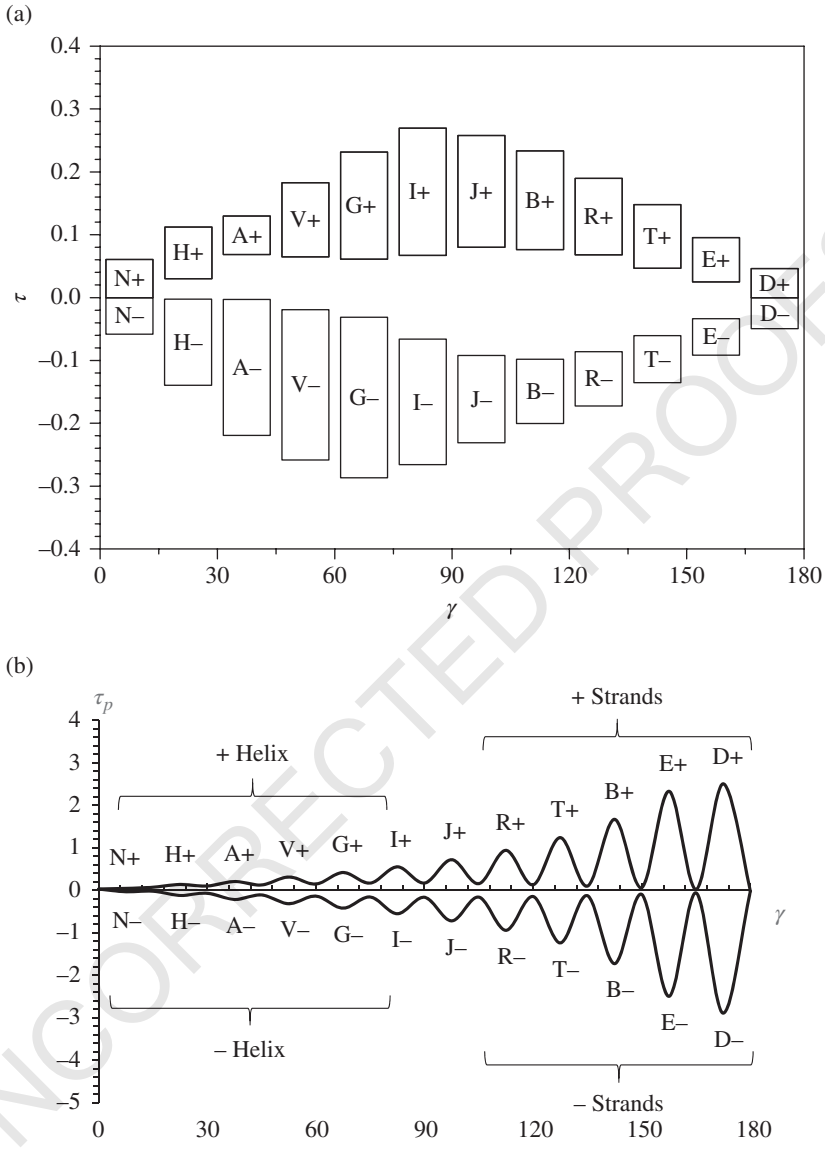

FIGURE 20 (a) Schematic representation of the 24 boxes (identified by 24 letters and the appropriate signs) that partition the γ, τ_α space in frequently populated structure areas. Compare with Figure 19. (b) The torsion peaks (upper half; troughs: lower half) associated with the 24-letter code. They are derived as averages from 510,525 residues of 2017 proteins.[86] All torsion values in Å and angles in degree. For additional explanations, see Table 7.

Figure 20b schematically gives for each letter the average τ_p values obtained from all residues. For example, a residue in a regular α-helix is denoted by "A+" and has a γ value between 30 and 45°, a τ value between 0.068 and 0.130 Å^{-1}, and a τ_p value of 0.205 Å^{-1}, whereas a residue in a regular right-handed β-strand denoted by "B-" is

found in the range $-120 \leq \gamma \leq -105^{\circ}$ and $-0.200 \leq \tau \leq -0.098^{\circ}$ Å^{-1} having a τ_p value of -0.944 Å^{-1}. Using the character code of Table 7, a 3D structure of a protein can be converted into a 1D character string, providing a number of advantages when analyzing protein similarity and protein folding.

DESCRIPTION OF PROTEIN SIMILARITY

Protein similarity[161] encloses both sequence similarity and structure similarity.[162] The ultimate goal of protein research is to connect both types of similarity with protein functionality. In view of the limited role of sequence similarity for functional similarity,[162–165] the purpose of this review is to focus on structural similarity to establish a basis for predicting functional similarity. A low-cost but reliable procedure of determining protein similarity can be used for a multitude of purposes such as the finding of common motifs, the location of similar active sites for different proteins interacting with a given pharmacophore, the classification of proteins, enzyme specificity, malfunctions of proteins causing diseases, or the description of folding just to mention some of them.

A commonly used tool in connection with protein structure comparison is the superposition of two or more structures, which is carried out in the way that the root-mean-square deviation (RMSD) measuring the average distance between all residues of the superimposed proteins is minimized.[166,167] Then, the minimized RMSD value is used as a measure of (dis)similarity.

Protein structure comparison (also called structural alignment) is essentially a *largest common point set (LCP)* problem,[168,169] which is considered in complexity theory to be *NP-complete* (nondeterministic polynomial time complete). This means that the needed computational time for solving the problem of the complexity class NP has as an upper bound a polynomial expression of the size of the input for the algorithm used but such an algorithm leading to an exact solution is so far not known.[170] Dynamic programming and recently linear programming[171] are often used to find the approximate optimal solution. Alternatively, heuristic methods are employed to find the local best solution. For this purpose, protein structure description has to be quantified and the (dis)similarity between two proteins has to be measured by some quantity such as a distance. The distance/similarity measure could be the distance matrix of a fragment as done in DALI method,[18,19] CE (combinatorial extension),[172] or certain scoring methods based on a segment of local geometry such as SSAP (Sequential Structure Alignment Program),[173–176] GDT_TS (Global Distance Test_Total Score),[177] Maxsub (maximum subarray) fit (i.e., the "maximum number of residues that fit well"),[178] or treatment of secondary structure as vectors such as in VAST (Vector Alignment Search Tool)[106,179]. A variety of programs have been developed to improve structure comparison, as, for example, MAMMOTH,[180] TOPOFIT,[180,181] SALIGN,[182] TM-align,[183] SABERTOOTH,[184] GANGSTA+,[185] or RAPIDO.[186]

Although the structure alignment quality has been improved by these methods, the superposition itself based on RMSD or other distance measures remains the basic

problem of any similarity algorithm. The distance measures are global (dis)similarity descriptors, which fail to disclose and specify local differences. These are often important as, for example, in the study of folding pathways (see the following text). A method that circumvents the problem of structure superposition is APSA.[61,74] It is suitable for the rapid similarity analysis of larger numbers of protein structures without scarifying the reliability and accuracy of structure comparison.

Qualitative and Quantitative Assessment of Protein Similarity

In Figure 21, the first 38 residues of ubiquitin (1UBQ) and ferredoxin (1A70) are compared because in both cases they correspond to the same α,β-roll topology[187] leading to large similarity ($\beta\beta\alpha$ motif). The curvature diagram does reveal only in a limited way this similarity. However, this is obvious when comparing the two torsion diagrams, which both show the structural sequence of strand β1, turn T1, strand β2, turn T2, and helix α1 (Figure 21). The difference between the two proteins arises predominantly from turns T1 and T2. For 1A70, T1 is presented by a single, broad

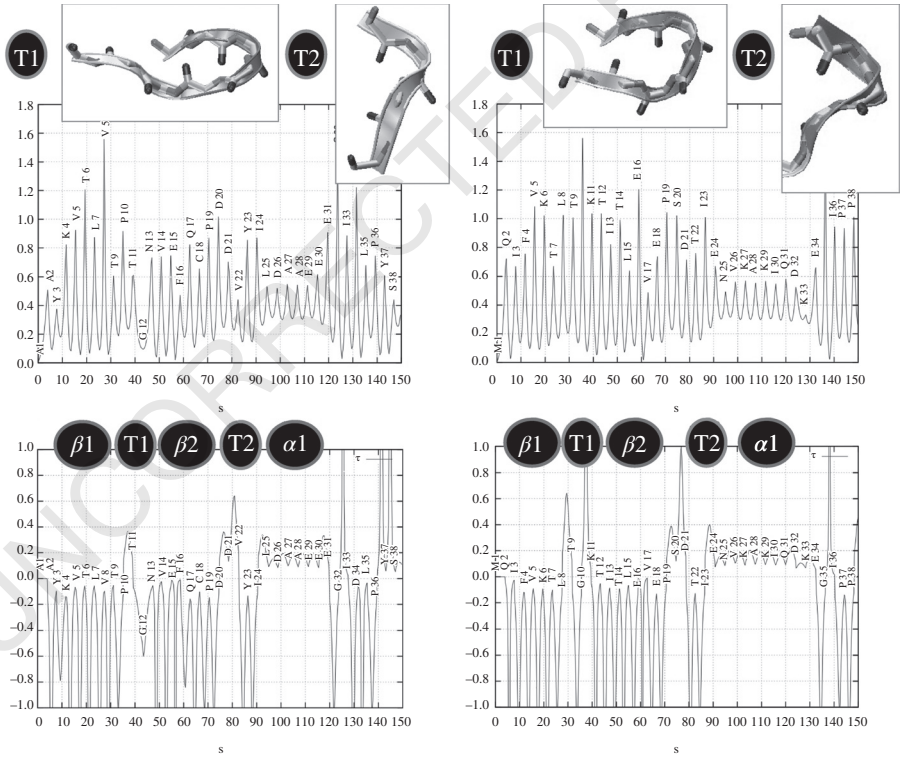

FIGURE 21 Comparison of curvature (above) and torsion diagrams (below) of the first 38 residues of ferredoxin (1A70) on the left and ubiquitin (1UBQ) on the right. The first five SSUs (β1T1β2T2α1 corresponding to the $\beta\beta\alpha$ motif) are indicated. Ribbon diagrams for T1 and T2 are shown as insets for each protein above the curvature diagrams.

torsion peak, whereas T1 of 1A70 splits up into two torsion peaks. Accordingly, T1 of 1A70 is characterized by one strong bend, whereas T1 of 1UBQ includes two strong bends, as is verified by the pictorial presentations given in the insets for T1. T2 is in both cases presented by two torsion peaks. For 1UBQ (right side), the second peak is somewhat higher. In line with this, the ribbon diagrams for T2 give two windings in both cases where the second for 1UBQ is indeed stronger. Hence, the torsion diagrams provide a detailed account on the (dis)similarity of T1 and T2 in the two proteins, which is quantitatively documented by the torsion diagrams.

Utilizing the character code of Table 7, one can use graphical means to describe protein similarity. This is demonstrated for the two examples shown in Figures 22 and 23. The diagram in Figure 22 describes the structural similarity of 40 structures of the protein GB1 domain (GB1)[188] by using the structural alphabet of Table 7. The horizontal axis indicates the sequence of residues (excluding the first and last one), whereas the vertical axis gives the 24-letter code of the structural alphabet. The structural characters are equally separated to provide a qualitative impression of (dis) similarity of the 40 structures of GB1. The conformation of each residue is presented by a dot. The size of a dot measures the population density, that is, it indicates how

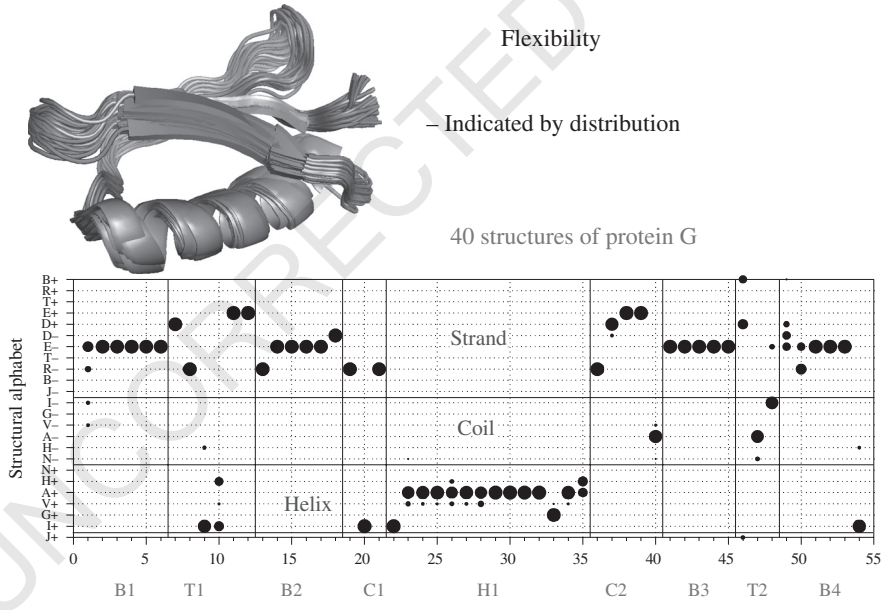

FIGURE 22 Forty structures of domain GB1[188] are compared using a similarity diagram. The horizontal axis gives the residue numbers of GB1 and the vertical axis the structural code of each residue using the structural alphabet of Table 7. Each residue structure is given by a black dot. The size of a dot is proportional to the population of the residue structure in question. Vertical lines separate SSUs of the protein whereas horizontal lines separate helix, coil, and strand regions of the structural alphabet. Structural flexibility increases with increasing distribution of dots in a given residue column. In the upper left, the 40 structures of GB1 presented as ribbon diagrams are superimposed.

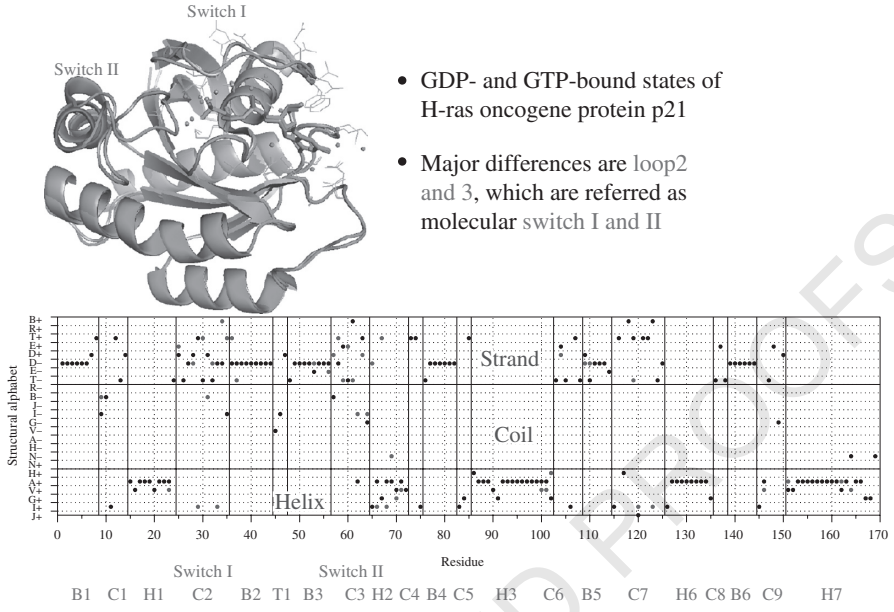

FIGURE 23 Qualitative similarity comparison between two states of H-ras oncogene protein p21,[189,190] one bound to GNP (5-guanylylimidodiphosphate) and one bound to GDP (guanosine diphosphate). Horizontal and vertical axes are defined as in Figure 22. Black dots give the residue structures of the GNP-bound complex. Red dots indicate that the corresponding residue has in the GDP-bound complex a different structure. A superposition of the two structures in the form of ribbon diagrams is shown in the upper left (see text). (*See insert for color representation of the figure*.)

many proteins have for a particular residue identical shape. If two or multiple dots are shown for one residue position in the diagram, different shapes for that particular residue are found in different proteins and the lower is its conformational stability (the higher its flexibility).

The diagram is divided by vertical and horizontal lines. The space marked by the vertical lines indicates individual SSUs, which can be easily identified via abbreviations such as B1, T1, etc. The three zones marked by horizontal lines indicate from bottom to top helices, coils, and strands where this simplification (compare with Table 7) is justified because regular left-handed helices are not present in domain GB1.

The flexibility of each residue can be easily identified from the diagram in Figure 22. Helix H1 varies slightly in the helix body and close to its ends. Overall, the β-strands are rigid with the exception of B4, which at its start can adopt three different conformations. The coil C1 connecting B1 and the head of H1 is very stable. Coil C2 between H1 and strand B3 is slightly variable. The largest flexibility of GB1 is found for the turns T1 and T2, which have different shapes. The two central residues of T1 are in the right-handed helical region, and for T2, they are in the left-handed coil region.

Figure 23 presents a qualitative similarity comparison between two states of H-ras oncogene protein p21,[189,190] one bound to GNP (5-guanylylimidodiphosphate), which is a GTP (guanosine triphosphate) analogue, and one to GDP (guanosine diphosphate). The GTP (GNP)-bound crystal structure has been taken from Pai and coworkers,[189] whereas the GDP-bound structure is from Tong and coworkers.[190]

Ras p21 is a product of H-ras oncogene, which is known to cause cancer.[189] This protein is a small GTPase that is likely to be involved in cellular processes such as signal transduction, protein transport, and secretion as well as polypeptide chain elongation.[191] Like other small GTPases, Ras p21 acts as a molecular switch. The GTP-bound conformation (*on state*) is biologically active and rapidly deactivated to the GDP-bound conformation (*off state*) through interaction with the GTPase-activating protein.[192] Mutant proteins that have been identified in human tumors are effectively locked in the active GTP conformation and are unable to be recycled quickly enough to the inactive GDP-bound state.[189,192]

Black dots in Figure 23 give the structure of the GNP-bound complex. A red dot indicates where the structure of the corresponding residue in the GDP-bound complex differs. The comparison of the GDP- and GTP(GNP)-bound states of Ras p21 reveals that the largest dissimilarity is found in regions *loop2* (coil C2 or Switch 1) and *loop3* (coil C3 or Switch II; Figure 23), as is revealed by the increased occurrence of the red dots. Helix H2 in the GDP complex is almost dissolved and fused with loop3. Obviously, the differences are due to the different shapes of the two ligands. The extra phosphate group of GTP (GNP) acts as a template. If this template is missing, the terminal part of H2 is more disordered.

Figures 22 and 23 show how a qualitative analysis of different proteins can be carried out utilizing the structural alphabet of Table 7 and graphical means. However, once the structural alphabet in the form of the 24-letter code of Table 7 is determined for two or more proteins, a quantitative comparison is also straightforward. For this purpose, the distance d between two residues in the 2D conformational γ, τ space is calculated. Comparable d values are obtained after normalizing the maximal τ, γ values to ± 1 so that γ and τ both range from -1 to $+1$, that is, the maximum d is always 2 considering that the majority of all data points of Figure 19 will be located after normalization within a circle of radius 1 (indicated in Figure 19 by the bold circle). Maximal similarity (100%) is given for a distance of zero by $(2-d)/2$, whereas maximal dissimilarity implies a distance of 2 divided by 2 equal to 100% (similarity 0%).

By determining the similarity for each pair of residues, the similarity of the entire protein is calculated as the average similarity of each matched residue. This approach is used to quantitatively determine the similarity between proteins 1UBQ and 1A70 in Table 8.

For the first 39 residues shown with their structural diagrams in Figure 21, a similarity of 78.85% is obtained whereas the similarity of the entire proteins is only slightly lower (76.22%). If a matching residue cannot be found (*gap*), the similarity for the gap will be 0%, which, when included into the similarity calculation, leads to somewhat lower values (for the first 39 residues, there are 3 gaps; Table 8).

TABLE 8 Similarity of Ubiquitin and Ferredoxin for the first $\beta\beta\alpha$ Motif Based on the γ, τ System

Ubiquitin (1UBQ)						Ferredoxin (1A70)						
Residue		γ	τ_α	Code	SSU	Residue		γ	τ_α	Code	SSU	Similarity
2	Q	169.8	−0.025	T−	B	3	Y	101.0	−0.136	J−	B	0.60
3	I	130.1	−0.117	R−	B	4	K	151.5	−0.066	E−	B	0.89
4	F	143.3	−0.090	T−	B	5	V	157.5	−0.058	E−	B	0.86
5	V	141.5	−0.093	T−	B	6	T	155.0	−0.056	E−	B	0.86
6	K	138.8	−0.100	T−	B	7	L	142.4	−0.109	T−	B	1.00
7	T	115.0	−0.136	B−	B	8	V	157.5	−0.053	E−	B	0.71
						9	T	113.3	−0.173	B−	B	0.[illegible]
8	L	86.8	0.127	I+	T	10	P	58.0	0.149	G+	T	0.[illegible]
9	T	121.6	−0.138	B−	T	11	T	51.9	−0.490	I−	T	0.47
10	G	148.4	0.069	E+	T	12	G	73.0	−0.069	G−	B	0.47
11	K	161.5	−0.053	E−	B	13	N	170.6	−0.023	D−	B	0.93
12	T	147.1	−0.084	T−	B	14	V	162.8	−0.054	E−	B	0.89
13	I	133.8	−0.127	R−	B	15	E	177.4	−0.009	D−	B	0.68
14	T	155.1	−0.066	E−	B	16	F	107.0	−0.156	J−	B	0.66
15	L	133.9	−0.137	R−	B	17	Q	139.3	−0.102	T−	B	0.93
16	E	176.4	−0.011	D−	B	18	C	115.3	−0.145	B−	B	0.60
17	V	130.1	−0.126	R−	B	19	P	117.7	−0.129	B−	B	0.95
18	E	122.8	−0.108	R−	C							
19	P	59.9	0.118	G+	T	20	D	56.9	0.129	G+	T	0.95
20	S	112.0	0.156	B+	T	21	D	90.7	0.209	J+	T	0.89
21	D	133.2	−0.126	R−	C	22	V	129.7	−0.131	R−	C	1.00
22	T	117.1	−0.131	B−	C	23	Y	120.9	−0.119	B−	C	0.95
23	I	62.4	0.123	G+	< H	24	I	55.4	0.128	V+	< H	1.00
24	E	27.8	0.084	V+	H	25	L	27.3	0.083	A+	H	1.00
25	N	36.3	0.106	A+	H	26	D	34.8	0.098	A+	H	0.95
26	V	32.9	0.088	A+	H	27	A	32.1	0.089	A+	H	1.00

(*Continued*)

TABLE 8 (*Continued*)

Ubiquitin (1UBQ)						Ferredoxin (1A70)						
Residue		γ	τ_α	Code	SSU	Residue		γ	τ_α	Code	SSU	Similarity
27	K	35.4	0.096	A+	H	28	A	33.9	0.090	A+	H	1.00
28	A	34.1	0.093	A+	H	29	E	34.2	0.082	A+	H	0.95
29	K	33.1	0.091	A+	H	30	E	54.0	0.113	V+	H	0.90
30	I	32.0	0.090	A+	H	31	E	115.6	–0.144	B–	H>	0.40
31	Q	39.0	0.120	V+	H							
32	D	19.8	0.068	H+	H							
33	K	24.4	0.056	H+	H>	32	G	168.7	0.026	D–	C	0.20
34	E	127.8	–0.128	R–	C	33	I	142.3	–0.109	T–	C	0.93
35	G	173.2	0.014	D+	C	34	D	168.0	–0.032	D–	C	0.89
36	I	113.4	–0.139	B–	C	35	L	115.2	–0.178	B–	C	0.95
37	P	125.9	–0.125	R–	C	36	P	163.9	0.058	E+	C	0.61
38	P	69.4	0.129	G+	C	37	Y	162.2	0.048	E+	C	0.43
39	D	21.4	0.062	H+	C	38	S	22.5	0.045	H+	C	0.95
Similarity in %								78.85% (76.22% entire protein)				

It is interesting to note that the similarity of turns T1(L8,T9) and T1(P10,T11) is low (15 and 47%; Table 8), whereas the similarity of turns T2(P19,S20) and T2(D20,D21) is 95% and 89%, respectively, as discussed on the basis of the torsion diagrams shown in Figure 21. Helices α1 of Figure 21 have a similarity exceeding 90% on the average, whereas that of the β-strands is significantly lower (see Table 8).

Similarity tests require a proper alignment of two or more proteins, which can be done by suitable optimization procedures,[59,161,193–195] where in the cases shown in this review dynamic programming[193] was used. If proteins are very different, the major problem is to find the 3D correspondence of their structures. This task can be facilitated by using level 2 coarse graining, which is based on a vector presentation of SSUs (see the preceding text).[74] As soon as an alignment at this level has been accomplished, level 1 coarse graining is applied with a residue-by-residue structural analysis, which leads to the actual similarity value.

In all similarity investigations, the specification of the gap penalty is decisive.[193] In the similarity analysis presented in Table 8, a linear gap penalty was used.[193,196] There are many other gap penalty methods of which just the affine gap penalty in sequence alignment is mentioned here, which has a high gap opening penalty but a low gap extension penalty.[197,198]

The Secondary Code and Its Application in Connection with Protein Similarity

In Table 9, the structural alphabet (*primary code*) is converted into a *secondary code*, which collects all residues belonging to a particular SSU and thus simplifies the identification of the SSUs of a protein while sacrificing structural details contained in the primary code. The advantages of the secondary code are evident when a more qualitative similarity comparison is needed. This is, for example, the case when the objective of a protein structure analysis is the search for certain motifs.

Figure 24 shows the ribbon diagrams for the 5-bladed β-propeller[199] and the β-helix.[200] For both motifs, the primary code leads to a similarity of 93% between the blades and 83% between the 16 β-strands (including the loop areas at the end) of the β-helix shown. Inspection of the primary codes leads to the same conclusion when using the letter code. Therefore, it is much simpler to convert the primary code into a secondary code according to Table 9, even though details of structural dissimilarities are lost, but revealing the two motifs as a sequence of β-strands, turns, and coils interrupted by a sequence of single, short helices in the case of the blades. The level 2 coarse graining with the vector presentation of strands and helices then provides a direct image of the tertiary structure and reveals bendings and torsions of the SSUs in a simpler and understandable way.

DESCRIPTION OF PROTEIN FOLDING

By using the similarity measure of the previous section, one can easily define a measure of *folding completeness* that facilitates the analysis of a calculated folding trajectory. The trajectory analysis of a folding process is commonly carried out with the

TABLE 9 Rules for the Encoding of Secondary Structures Using the 24-Letter Code

SSU	Recognition Rule by Primary Code: Structure Label	Recognition Rule by Primary Code: Pattern	Sec. Code	Comments
α-Helices				
Starter	G+, V+, I+, A+, H+, N+		<H	Starter + body + exit: minimum of 4 residues equal to 1 full turn of α-helix
Body	A+, H+, V+, N+, G+	Minimum: two consecutive residues	H	
Exit	First nonhelical residue		H>	
β-Strands				
Regular	D–, E–, B–, T–, R– D+, E+, B+, T+, R+	Minimum: two consecutive residues	B	β-Strand
Distorted	J+, J–, I+, I–	Single residue		
Extended	D–, E–, B–, T–, R–, D+, E+, B+, T+, R+	Minimum: two consecutive residues	S	Extended single strand
Other helices				
3_{10}-Helix	V+, H+, A+	Consecutive V+ or mixture	3	
π-Helix	N+, H+, A+, V+	Alternate N+ with other codes	P	
Turn			T	Involving <6 residues, ends separated by <7 Å
Coil			C	Remaining unclassified positions

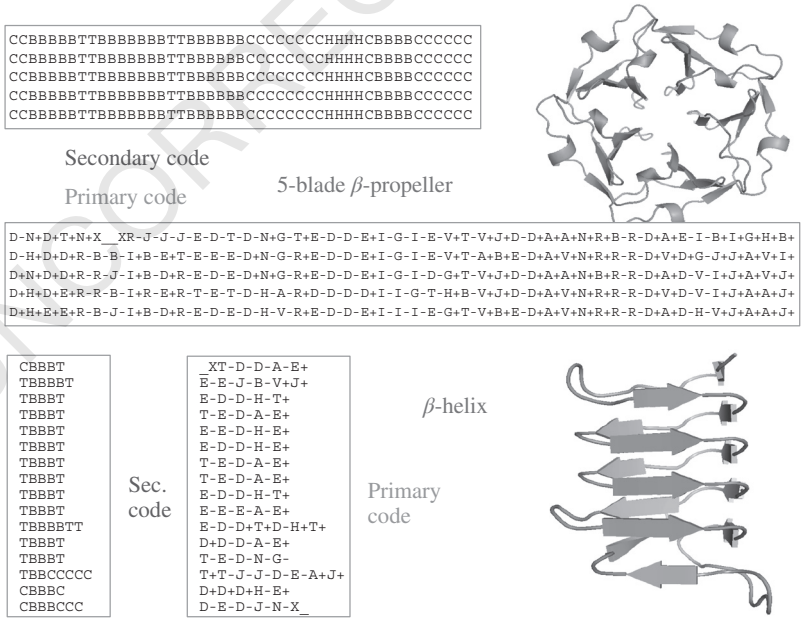

FIGURE 24 Ribbon diagrams for the 5-bladed β-propeller[199] (above) and the β-helix[200] (below). The repeating motif is shown in each case in the form of the primary code and the secondary code, respectively.

help of the minimization of the RMSD,[166,167] but as mentioned previously, the RMSD value is a global measure that does not provide information about local details. In particular, the RMSD parameter barely gives any information on how the folding of a protein propagates during a molecular dynamics (MD) simulation. Very different conformations may lead to the same RMSD value, and therefore any analysis based on RMSD values can be misleading.[201,202]

The structural alphabet of APSA provides an effective but simple way of comparing local residue structures directly without the need of protein superposition. The primary code of APSA can be used qualitatively and quantitatively. As such, the overall similarity of a calculated snapshot during the folding process can be quickly compared with the native (target) structure of the protein in question and a quantitative measure of folding completeness determined. An example is shown in Figure 25, which analyzes the folding trajectory of the antimicrobial peptide 2L24 (13 residues: IFGAIAGFIKNIWX)[203] given in the upper left of the figure in the form of a ribbon diagram. Since an X-ray diffraction structure of 2L24 is known,[203] its primary code can be derived and used for the similarity analysis. If the folding process is finished, the folding completeness is 1 and the MD simulation stopped.

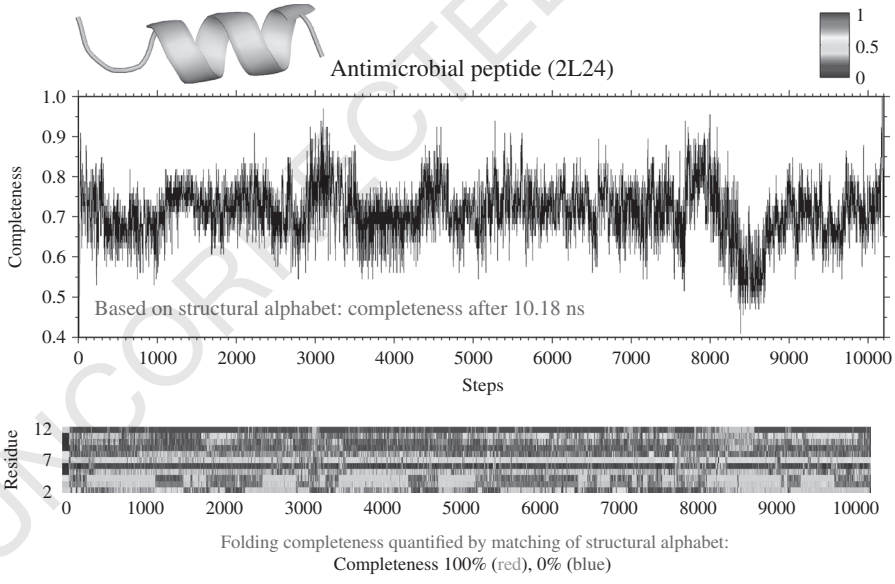

FIGURE 25 Upper part: folding trajectory of the antimicrobial peptide 2L24 (shown as ribbon diagram in the upper left corner) obtained from a 10.18 ns MD simulation. For each picosecond step, the folding completeness is given as derived from a similarity analysis. Lower diagram: folding spectrum in the form of a color-coded similarity test for each of the residues analyzed (vertical axis) at the 10,180 time steps (horizontal axis). The color code is given in the upper right: dark blue—unfolded; dark red—folded (see text). (*See insert for color representation of the figure*.)

The folding trajectory of 2L24 obtained by an MD calculation is shown in the upper diagram of Figure 25. The horizontal axis counts 10,180 MD picosecond steps until folding completeness is reached, which is given by the vertical axis with completeness values between 0 (unfolded) and 1.0 (folded into the native structure: 100% completeness). Because the similarity tests require little time, they are carried out after each MD step and the calculation is terminated once the folding completeness is reached. Each step of the 10.18 ns simulation can be evaluated with the completeness diagram. No regular pattern suggests a specific mechanism leading to folding. For more than a dozen MD steps, completeness is above 90% but does not lead in the subsequent steps to complete folding.

In the lower diagram of Figure 25, all snapshots of the folding trajectory are represented as a *folding spectrum*. Again, the horizontal axis gives the 10,180 MD steps, whereas the vertical axis gives the 13 residues of 2L24 (the first and last residues are excluded because of the analysis in terms of Frenet coordinates) where the structure of each residue is color-coded according to its similarity to the structure in the folded (native) peptide (see upper right corner of Figure 25). Dark blue (dark red) indicates the peptide is unfolded (folded). If all residues of an MD snapshot are given by a dark red column, folding completeness is reached. This is the situation at step 10,180 shown at the far right of the diagram (see Figure 25).

The folding spectrum given in the lower part of Figure 25 suggests a folding mechanism, which was not obvious in the upper diagram. The C-terminal segment (residue 12) of the helix folds rapidly and remains, with one exception after 8500 steps, stable. Contrary to the C-terminal, the N-terminal segment tends to unfold from time to time (green color of residues 2 and 3). However, the folding barrier is associated with residues 6 and 7 (G and F), which remain in an unfolded structure most of the time. Hence, folding can be seen as the difficulty of combining the flexibility of glycine (G) with the conformational preferences of phenylalanine (F). The completeness peaks larger than 0.9 in the upper diagram correspond to those MD steps for which the folding barrier has been surmounted (indicated by the red lines crossing through the blue horizontal bar). Hence, the secret of the folding mechanism is to combine the folding of the N-terminal segment with that of the residue pair GF. Once this combination is understood, an accelerated folding process can be initiated by freezing G and F in the conformations they adopt in the folded structure.

CONCLUDING REMARKS

The description of protein structure by its backbone is a well-accepted concept. This review describes the possibilities of protein structure analysis when using coarse graining and describing the protein backbone as a smooth line in 3D space rather than a collection of discrete backbone points. The natural choice for the anchor points leading to a smooth backbone line is the C_α atoms because they represent the hinge joints of the backbone, which directly reflect the conformational and steric influences of the side chains. By using a continuous rather than a discrete representation of the backbone and describing it in terms of Frenet coordinates, the 3D structure of

the protein can be converted into the 2D Frenet structure diagrams of curvature $\kappa(s)$ and torsion $\tau(s)$, which reveal characteristic and easy-to-distinguish patterns for all SSUs of a protein. Protein structure analysis is transferred thereby from a qualitative or semiquantitative to a quantitative level.

Ideal and regular helices (3_{10}-, α-, or π-helix), β-strands, coils, turns, etc. can be easily distinguished by their $\kappa(s)$ and $\tau(s)$ diagrams. Their chirality can be determined and the degree of distortions quantified. A classification of turns and loops is possible.

By introducing a structural alphabet consisting of a 24-letter code, a quantification of protein similarity is realized. The 24-letter code is a primary code of sufficient detail and robustness because it is based on the torsion properties of half a million residues (510,525 from 2017 proteins). It is used to convert the 3D structure of a protein via the 2D Frenet structure diagrams into a 1D character string that can be used for rapid structure classification and similarity analysis. There are qualitative and quantitative measures of assessing protein similarity; to do this, the γ, τ-quantification scheme can be used.

A primary letter code is best suited for a quantitative description of protein similarity, whereas a secondary letter code provides a rapid assessment of all SSUs in a protein. This leads to an identification of the termini of an SSU, helps to establish the second level of coarse graining, and rapidly handles tasks such as motif identification. Either the primary code or the secondary code can be used to study the folding of a protein, and one can elucidate the folding mechanism by calculating folding spectra, which indicate the folding completeness during an MD simulation in a quantitative way.

Once a basis for a coarse-grained geometrical description of the protein structure is laid in terms of Frenet coordinates, there are several possibilities of extending its application repertoire. Shape features of the tertiary protein structure are easily described using Frenet coordinates by a level 2 coarse-grained approach. Helices and strands are represented by curved and twisted arrows, which quantitatively reflect major distortions of these SSUs. By keeping the backbone line for turns and coils, supersecondary structures, folds, or tertiary structure can be described easily. In general, the use of Frenet coordinates in methods such as APSA is based on the idea of subsequent steps of coarse graining the protein backbone so that more and more nonlocal features are included into the description and the description of tertiary protein structure is facilitated.

Combining a geometry-based description in terms of Frenet coordinates with an H-bond-based description as used in DSSP is also possible. This can improve the detail description of both secondary and tertiary structure. Obviously, there are shortcomings of the latter when kinks in helices have to be properly identified, which is important for the correct count of helices, for example, in similarity studies.

The list of protein structures described by APSA leads to a library of protein subunits or building blocks that can be used in protein structure prediction and modeling. The analysis of 2D Frenet structure diagrams provides a meaningful and systematic way of breaking down loop regions to substructures that contribute to a library of κ-τ or γ, τ patterns. Such a library can be used to search and classify turns, supersecondary structures, and folds in a systematic way.[61] Noteworthy is also that Frenet coordinates can be applied to polymers, DNA strands, carbohydrates, or any molecules with long strands.

please insert: 1152357 and Grant CHE (so that it reads: Grant CHE 1152357 and Grant CHE 1464906.)

ACKNOWLEDGMENT

We thank S. Ranganathan, D. Izotov, and E. Kraka for important preliminary contributions to this review. This work was financially supported by the National Science Foundation, Grant CHE 1464906. We thank SMU for providing computational resources.

REFERENCES

1. Pauling, J. L.; Corey, R.; Branson, H. The Structure of Proteins; Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci. U.S.A.* 1951, **37**, 205–211.
2. Nelson, D. L.; Cox, M. M. *Lehninger Principles of Biochemistry*, 5th Edition; Freeman, New York, 2008.
3. Mathews, C.; van Holde, K.; Ahern, K. *Biochemistry*; Addison, Wesley Longman, New York, 1999.
4. Andersen, C. A. F.; Rost, B. In *Structural Bioinformatics*, Vol. **44**; Bourne, P. E., Weissig, H., Eds.; Wiley-Liss, Hoboken, NJ, 2003; p 1.
5. Cox, M. M.; Phillips, G., Jr. *Handbook of Proteins, Structure, Function and Methods*, Vol. **1–2**; John Wiley & Sons, New York, 2007.
6. Venkatachalam, C. Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of Three Linked Peptide Units. *Biopolymers* 1968, **6**, 1425–1436.
7. Kendrew, J.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H.; Phillipsa, D. A Three-Dimensional Model of the Myoglobin Molecule Obtained by x-Ray Analysis. *Nature* 1958, **181**, 662–666.
8. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 1983, **22**, 2577–2637.
9. Rao, S.; Rossman, M. G. Comparison of Super-Secondary Structures in Proteins. *J. Mol. Bio.* 1973, **76**, 211–256.
10. Zaki, M.; Bystroff, C. *Protein Structure Prediction, Methods in Molecular Biology*; Humana Press, Totowa, 2010.
11. Rangwala, H.; Karypos, G. *Wiley Series in Bioinformatics, Introduction to Protein Structure and Prediction: Methods and Algorithms*; John Wiley & Sons, New York, 2010.
12. Westbrook, Z.; Feng, G.; Gilliland, T. N.; Bhat, H.; Weissig, I.; Shindyalov, P. The Protein Data Bank. *Nucleic Acids Res.* 2000, **28**, 235–242.
13. Berman, H. M. The Protein Data Bank: A Historical Perspective. *Acta Crystallogr.* 2008, **A64**, 88–95.
14. Murzin, A.; Brenner, S.; 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.
15. Andreeva, A.; Howorth, D.; Chandonia, J.-M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Data Growth and Its Impact on the SCOP Database: New Developments. *Nucleic Acids Res.* 2007, **36**, D419–D425.
16. Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. CATH—A Hierarchic Classification of Protein Domain Structures. *Structure* 1997, **5**, 1093–1108.

17. Cuff, A. L.; Sillitoe, I.; Lewis, T.; Clegg, A. B.; Rentzsch, R.; Furnham, N.; Pellegrini-Calace, M.; Jones, D.; Thornton, J.; Orengo, C. A. Extending CATH: Increasing Coverage of the Protein Structure Universe and Linking Structure with Function. *Nucleic Acids Res.* 2011, **39**, D420–D426.
18. Holm, L.; Sander, C. DALI: A Network Tool for Protein Structure Comparison. *Trends Biochem. Sci.* 1995, **20**, 478–480.
19. Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. A Database of Protein Structure Families with Common Folding Motifs. *Trends Biochem. Sci.* 1992, **1**, 1691–1698.
20. Dietmann, S.; Holm, L. Identification of Homology in Protein Structure Classification. *Nat. Struct. Biol.* 2001, **8**, 953–957.
21. Day, R.; Beck, D.; Armen, R.; Daggett, V. A Consensus View of Fold Space: Combining SCOP, CATH, and the DALI Domain Dictionary. *Protein Sci.* 2003, **12**, 2150–2160.
22. Andersen, C. A.; Palmer, A. G.; Brunak, S.; Rost, B. Continuum Secondary Structure Captures Protein Flexibility. *Structure* 2002, **10**, 175–184.
23. Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins* 1995, **23**, 566–579.
24. Richards, F.; Kundrot, C. Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins* 1988, **3**, 71–84.
25. Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Brevern, A.; Gibrat, J.-F. Protein Secondary Structure Assignment Revisited: A Detailed Analysis of Different Assignment Methods. *BMC Struct. Biol.* 2005, **5**, 17.
26. Fourrier, L.; de Brevern, A. G. Use of a Structural Alphabet for Analysis of Short Loops Connecting Repetitive Structures. *BMC Bioinf.* 2004, **5**, 58.
27. Offmann, B.; Tyagi, M.; de Brevern, A. G. Local Protein Structures. *Curr. Bioinf.* 2007, **2**, 165–202.
28. Efimov, A. Standard Structures in Proteins. *Prog. Biophys. Mol. Biol.* 1993, **60**, 201–239.
29. Richardson, J. The Anatomy and Taxonomy of Protein Structure. *Adv. Protein Chem.* 1981, **34**, 167–339.
30. Raveh, B.; Rahat, O.; Basri, R.; Schreiber, G. Rediscovering Secondary Structures as Network Motifs—An Unsupervised Learning Approach. *Bioinformatics* 2007, **23**, e163–e169.
31. Mottalib, M. A.; Mahdi, S. R.; Haque, Z. A. B. M.; Al-Mamun, S. M.; Al-Mamun, A. H. Protein Secondary Structure Prediction Using Feed-Forward Neural Network. *J. Converg. Inf. Technol.* 2007, **1**, 64–68.
32. Cheng, J.; Tegge, A.; Baldi, P. Machine Learning Methods for Protein Structure Prediction. *IEEE Rev. Biomed. Eng.* 2008, **1**, 41–49.
33. Salamov, A.; Solovyev, V. Prediction of Protein Secondary Structure by Combining Nearest-Neighbor Algorithms and Multiple Sequence Alignments. *J. Mol. Biol.* 1995, **247**, 11–15.
34. Xiang, Z.; Soto, C. S.; Honig, B. Evaluating Conformational Free Energies: The Colony Energy and Its Application to the Problem of Loop Prediction. *Proc. Natl. Acad. Sci. U.S.A.* 2002, **99**, 7432–7437.
35. Jager, M.; Deechongkit, S.; Koepf, E. K.; Nguyen, H.; Gao, J.; Powers, E. T.; Gruebele, M.; Kelly, J. Understanding the Mechanism of Beta-Sheet Folding from a Chemical and Biological Perspective. *Biopolymers* 2008, **90**, 751–758.

36. Bower, M.; Cohen, F.; Dunbrack, R. Sidechain Prediction from a Backbone-Dependent Rotamer Library: A New Tool for Homology Modeling. *J. Mol. Biol.* 1997, **267**, 1268–1282.
37. Mattos, C.; Petsko, G.; Karplus, M. Analysis of Two-Residue Turns in Proteins. *J. Mol. Biol.* 1994, **238**, 733–747.
38. Jones, D. Critically Assessing the State-of-the-Art in Protein Structure Prediction. *Pharmacogenomics J.* 2001, **1**, 126–134.
39. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round x. *Proteins* 2014, **82**, 1–6.
40. Parisien, M.; Major, F. A New Catalog of Protein Beta-Sheets. *Proteins* 2005, **61**, 545–558.
41. Carter, P.; Andersen, C. A. F.; Rost, B. DSSPcont: Continuous Secondary Structure Assignments for Proteins. *Nucleic Acids Res.* 2003, **31**, 3293–3295.
42. Fodje, M. N.; Al-Karadaghi, S. Occurrence, Conformational Features and Amino Acid Propensities for the p-Helix. *Protein Eng.* 2002, **15**, 353–358.
43. Park, S. Y.; Yoo, M.-J.; Shin, J.; Cho, K.-H. SABA (Secondary Structure Assignment Program Based on only Alpha Carbons): A Novel Pseudo Center Geometrical Criterion for Accurate Assignment of Protein Secondary Structures. *BMB Rep.* 2011, **44**, 118–122.
44. Hosseini, S.-R.; Sadeghi, M.; Pezeshk, H.; Eslahchi, C.; Habibi, M. PROSIGN: A Method for Protein Secondary Structure Assignment Based on Three-Dimensional Coordinates of Consecutive C(Alpha) Atoms. *Comput. Biol. Chem.* 2008, **32**, 406–411.
45. Cubellis, M.; Cailliez, F.; Lovell, S. Secondary Structure Assignment That Accurately Reflects Physical and Evolutionary Characteristics. *BMC Bioinf.* 2005, **6**, S8.1–S8.9.
46. Majumdar, I.; Krishna, S. S.; Grishin, N. V. PALSSE: A Program to Delineate Linear Secondary Structural Elements from Protein Structures. *BMC Bioinf.* 2005, **6**, 202.1–202.24.
47. Dupuis, F.; Sadoc, J.-F.; Mornon, J.-P. Protein Secondary Structure Assignment Through Voronoi Tessellation. *Proteins* 2004, **55**, 519–528.
48. Taylor, T.; Rivera, M.; Wilson, G.; Vaisman, I. I. New Method for Protein Secondary Structure Assignment Based on a Simple Topological Descriptor. *Proteins* 2005, **60**, 513–524.
49. Taylor, W. R. Defining Linear Segments in Protein Structure. *J. Mol. Biol.* 2001, **310**, 1135–1150.
50. Hutchinson, E. G.; Thornton, J. M. PROMOTIF: A Program to Identify and Analyze Structural Motifs in Proteins. *Protein Sci.* 1996, **5**, 212–220.
51. Sklenar, H.; Etchebest, C.; Laverey, R. Describing Protein Structure: A General Algorithm Yielding Complete Helicoidal Parameters and a Unique Overall Axis. *Proteins* 1989, **6**, 46–60.
52. King, S. M.; Johnson, W. C. Assigning Secondary Structure from Protein Coordinate Data. *Proteins* 1999, **35**, 313–320.
53. Labesse, G.; Colloch, N.; Pothier, J.; Mornon, J. P. P-SEA: A New Efficient Assignment of Secondary Structure from C_α. *Comput. Appl. Biosci.* 1997, **13**, 291–295.
54. Hanson, R. M.; Kohler, D.; Braun, S. G. Quaternion-Based Definition of Protein Secondary Structure Straightness and Its Relationship to Ramachandran Angles. *Proteins: Struct. Funct. Bioinf.* 2011, **79**, 2172–2180.

55. Cooley, R. B.; Arp, D. J.; Karplus, P. A. Evolutionary Origin of a Secondary Structure: pi-Helices as Cryptic but Widespread Insertional Variations of Alpha-Helices That Enhance Protein Functionality. *J. Mol. Biol.* 2010, **404**, 232–246.
56. Riek, R. P.; Graham, R. M. The Elusive π-Helix. *J. Struct. Biol.* 2011, **173**, 153–160.
57. Novotny, M.; Kleywegt, G. J. A Survey of Left-Handed Helices in Protein Structures. *J. Mol. Biol.* 2005, **347**, 231–241.
58. Koch, O.; Cole, J.; Block, P.; Klebe, G. Secbase: Database Module to Retrieve Secondary Structure Elements with Ligand Binding Motifs. *J. Chem. Inf. Model.* 2009, **49**, 2388–2402.
59. Andersen, C. A. F.; Rost, B. Secondary Structure Assignment. In *Structural Bioinformatics*; Gu, J.; Bourne, P. E., Eds.; Wiley-Blackwell, Hoboken, NJ, 2009, pp 459–484.
60. Keller, D.; Clausen, R.; Josefsen, K.; Led, J. J. Flexibility and Bioactivity of Insulin: An NMR Investigation of the Solution Structure and Folding of an Unusually Flexible Human Insulin Mutant with Increased Biological Activity. *Biochemistry* 2001, **40**, 10732–10740.
61. Ranganathan, S.; Izotov, D.; Kraka, E.; Cremer, D. Description and Recognition of Regular and Distorted Secondary Structures in Proteins Using the Automated Structure Analysis Method. *Proteins* 2009, **76**, 418–438.
62. Guggenheimer, H. *Differential Geometry*; Dover Publications, London, 1977.
63. Kreyszig, E. *Differential Geometry*; Dover Publications, London, 1991.
64. Ranganathan, S.; Izotov, D.; Kraka, E.; Cremer, D. Automated and Accurate Protein Structure Description: Distribution of Ideal Secondary Structural Units in Natural Proteins. *arXiv* 2008, **0811**, 3587.1–3587.27.
65. Rackovsky, S.; Scheraga, H. Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations. *Macromolecules* 1978, **11**, 1168–1174.
66. Louie, A.; Somorjai, R. Differential Geometry of Proteins: Helical Approximations. *J. Mol. Biol.* 1983, **168**, 143–162.
67. Louie, A.; Somorjai, R. Differential Geometry of Proteins: A Structural and Dynamical Representation of Patterns. *J. Theor. Biol.* 1982, **98**, 189–209.
68. Soumpasis, D.; Strahm, M. Efficient Identification and Analysis of Substructures in Proteins Using the Kappa-Tau Framework: Left Turns and Helix c-Cap Motifs. *J. Bimol. Struct. Dyn.* 2000, **17**, 965–979.
69. Hausrath, A. C.; Goriely, A. Continuous Representations of Proteins: Construction of Coordinate Models from Curvature Profiles. *J. Struct. Biol.* 2007, **158**, 267–281.
70. Zhi, D.; Krishna, S.; Cao, H.; Pevzner, P.; Godznik, A. Representing and Comparing Protein Structures as Curves in Three-Dimensional Space. *BMC Bioinf.* 2006, **7**, 460.1–460.15.
71. Can, T.; Wang, Y.-F. Protein Structure Alignment and Fast Similarity Search Using Local Shape Signatures. *J. Bioinf. Comp. Biol.* 2004, **2**, 215–239.
72. Goyal, S.; Perkins, N.; Lee, C. Nonlinear Dynamics and Loop Formation in Kirchhoff Rods with Implications to the Mechanics of DNA and Cables. *J. Comp. Phys.* 2005, **209**, 371–389.
73. Goyal, S.; Lillian, T.; Blumberg, S.; Meiners, J.-C.; Meyhofer, E.; Perkins, N. Intrinsic Curvature of DNA Influences LacR-Mediated Looping. *Biophys. J.* 2007, **93**, 4342–4359.
74. Guo, Z.; Kraka, E.; Cremer, D. Description of Local and Global Properties of Protein Helices. *J. Mol. Model.* 2013, **19**, 2901–2911.

75. 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* 2008, **0811**, 3258.1–27.
76. Konkoli, Z.; Kraka, E.; Cremer, D. Unified Reaction Valley Approach: Mechanism of the Reaction $CH_3 + H_2 \rightarrow CH_4 + H$. *J. Phys. Chem. A.* 1997, **101**, 1742–1757.
77. Kraka, E.; Cremer, D. Computational Analysis of the Mechanism of Chemical Reactions in Terms of Reaction Phases: Hidden Intermediates and Hidden Transition State. *Acc. Chem. Res.* 2010, **43**, 591–601.
78. Cremer, D.; Kraka, E. From Molecular Vibrations to Bonding, Chemical Reactions, and Reaction Mechanism. *Curr. Org. Chem.* 2010, **14**, 1524–1560.
79. Dill, K.; Ozkan, S.; Shell, M.; Weikl, T. The Protein Folding Problem. *Ann. Rev. Biophys.* 2008, **37**, 289–316.
80. Fedyukina, D.; Cavagnero, S. Protein Folding at the Exit Tunnel. *Ann. Rev. Biophys.* 2011, **40**, 337–359.
81. Tung, C.-H.; Huang, J.-W.; Yang, J.-M. Kappa-Alpha Plot Derived Structural Alphabet and BLOSUM-Like Substitution Matrix for Fast Protein Structure Database Search. *Genome Biol.* 2007, **8**, R31.1–R31.16.
82. Ranganathan, S.; Izotov, D.; Kraka, E.; Cremer, D. Classification of Supersecondary Structures in Proteins Using the Automated Protein Structure Analysis Method. *arXiv* 2008, **0811**, 3464.1–3464.40.
83. Barlow, D. J.; Thornton, J. M. Helix Geometry in Proteins. *J. Mol. Biol.* 1988, **201**, 601–619.
84. Armen, R.; Alonso, D. O. V.; Daggett, V. The Role of α-, 3_{10}-, and π-Helix in Helix $\rightarrow$ Coil Transitions. *Protein Sci.* 2003, **12**, 1145–1157.
85. Hovmoller, S.; Zhou, T.; Ohlson, T. Conformations of Amino Acids in Proteins. *Acta Crystallogr. D Biol. Crystallogr.* 2002, **58**, 768–776.
86. Guo, Z.; Cremer, D. APSA12, Automated Protein Structure Analysis. see https://sites.smu.edu/dedman/catco/apsa.html and references cited therein, 2012; Southern Methodist University, Dallas, TX.
87. McPhalen, C. A.; Vincent, M. G.; Jansonius, J. N. X-Ray Structure Refinement and Comparison of Three Forms of Mitochondrial Aspartate Aminotransferase. *J. Mol. Biol.* 1992, **225**, 495–517.
88. Tainer, J. A.; Getzoff, E. D.; Beem, K. M.; Richardson, J. S.; Richardson, D. C. Determination and Analysis of the 2 A-Structure of Copper, Zinc Superoxide Dismutase. *J. Mol. Biol.* 1982, **160**, 181–217.
89. Colloc'h, N.; Etchebest, C.; Thoreau, E.; Henrissat, B.; Mornon, J. P. Comparison of Three Algorithms for the Assignment of Secondary Structure in Proteins: The Advantages of a Consensus Assignment. *Protein Eng.* 1993, **6**, 377–382.
90. Arab, S.; Didehvar, F.; Saleghi, M. Helix Segment Assignment in Proteins Using Fuzzy Logic. *Iranian J. Biotech.* 2007, **5**, 93–99.
91. Enkhbayar, P.; Hikichi, K.; Osaki, M.; Kretsinger, R. H.; Matsushima, N. 3_{10}-Helices in Proteins Are Parahelices. *Proteins* 2006, **15**, 691–699.
92. Weiss, M. S.; Schulz, G. E. Structure of Porin Refined at 1.8 A Resolution. *J. Mol. Biol.* 1992, **227**, 493–509.
93. Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. Crystal Structure of the Collagen Triple Helix Model [(Pro-Pro-Gly)(10)](3). *Protein Sci.* 2002, **11**, 262–270.

94. Adzhubei, A.; Sternberg, M. J.; Makarov, A. A. Polyproline-II Helix in Proteins: Structure and Function. *J. Mol. Biol.* 2013, **425**, 2100–2132.

95. Bitto, E.; McKay, D. B. Elastase of Pseudomonas Aeruginosa with an Inhibitor. *RCSB Protein Data Bank* 2004, 1U4G.

96. Tsukihara, T.; Shimokata, K.; Katayama, Y.; Shimada, H.; Muramoto, K.; Aoyama, H.; Mochizuki, M.; Shinzawa-Itoh, K.; Yamashita, E.; Yao, M.; Ishimura, Y.; Yoshikawa, S. The Low-Spin Heme of Cytochrome c Oxidase as the Driving Element of the Proton-Pumping Process. *Proc. Natl. Acad. Sci. U.S.A.* 2003, **100**, 15304–15309.

97. Asselt, E. J.; Thunnissen, A. M.; Dijkstra, B. W. High Resolution Crystal Structures of the *Escherichia coli* Lytic Transglycosylase Slt70 and Its Complex with a Peptidoglycan Fragment. *J. Mol. Biol.* 1999, **291**, 877–898.

98. Lopera, J. A.; Sturgis, J. N.; Duneau, J. P. Ptuba: A Tool for the Visualization of Helix Surfaces in Proteins. *J. Mol. Graph. Model.* 2005, **23**, 305–315.

99. Hu, C.; Koehl, P. Helix-Sheet Packing in Proteins. *Proteins* 2010, **78**, 1736–1747.

100. Tatulian, S. Determination of Helix Orientations in Proteins. *Computat. Biol. Chem.* 2008, **32**, 370–374.

101. Chothia, C.; Levitt, M.; Richardson, D. Helix to Helix Packing in Proteins. *J. Mol. Biol.* 1981, **145**, 215–250.

102. Walther, D.; Eisenhaber, F.; Argos, P. Principles of Helix-Helix Packing in Proteins: The Helical Lattice Superposition Model. *J. Mol. Biol.* 1996, **255**, 536–553.

103. Lee, S.; Chirikjian, G. Interhelical Angle and Distance Preferences in Globular Proteins. *Biophys. J.* 2004, **86**, 1105–1117.

104. Dalton, J. A.; Michalopoulos, I.; Westhead, D. R. Calculation of Helix Packing Angles in Protein Structures. *Bioinformatics* 2003, **19**, 1298–1299.

105. Singh, A. P.; Brutlag, D. L. Hierarchical Protein Structure Superposition Using both Secondary Structure and Atomic Representations. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1997, **5**, 284–293.

106. Gibrat, J. F.; Madej, T.; Bryant, S. H. Surprising Similarities in Structure Comparison. *Curr. Opin. Struct. Biol.* 1996, **6**, 377–385.

107. Åqvist, J. A Simple Way to Calculate the Axis of an α-Helix. *Comput. Chem.* 1986, **10**, 97–99.

108. Enkhbayar, P.; Damdinsuren, S.; Osaki, M.; Matsushima, N. HELFIT: Helix Fitting by a Total Least Squares Method. *Comput. Biol. Chem.* 2008, **32**, 307–310.

109. Kahn, P. Defining the Axis of a Helix. *Comput. Chem.* 1989, **13**, 185–189.

110. Christopher, J. A.; Swanson, R.; Baldwin, T. O. Algorithms for Finding the Axis of a Helix: Fast Rotational and Parametric Least-Squares Methods. *Comput. Chem.* 1996, **20**, 339–345.

111. Nievergelt, Y. Fitting Helices to Data by Total Least Squares. *Comput. Aided Geom. Des.* 1997, **14**, 707–718.

112. Blundell, T.; Barlow, D.; Borkakoti, N.; Thornton, J. Solvent Induced Distortion and Curvature of α-Helices. *Nature* 1983, **306**, 281–283.

113. Kumar, S.; Bansal, M. Geometrical and Sequence Characteristics of α-Helices in Globular Proteins. *Biophys. J.* 1998, **75**, 1935–1944.

114. Bansal, M.; Kumar, S.; Velavan, R. HELANAL: A Program to Characterize Helix Geometry in Proteins. *J. Biomol. Struct. Dyn.* 2000, **17**, 811–819.

115. Kumar, S.; Bansal, M. Structural and Sequence Characteristics of Long α-Helices in Globular Proteins. *Biophys. J.* 1996, **71**, 1574–1586.

116. Sugeta, H.; Miyazawa, T. General Method for Calculating Helical Parameters of Polymer Chains from Bond Lengths, Bond Angles, and Internal-Rotation Angles. *Biopolymers* 1967, **5**, 673–679.

117. Lee, H. S.; Choi, J.; Yoon, S. QHELIX: A Computational Tool for the Improved Measurement of Inter-Helical Angles in Proteins. *Protein J.* 2007, **26**, 556–561.

118. Kahn, P. Simple Methods for Computing the Least Squares Line in Three Dimensions. *Comput. Chem.* 1989, **13**, 191–195.

119. Langelaan, D. N.; Wieczorek, M.; Blouin, C.; Rainey, J. K. Improved Helix and Kink Characterization in Membrane Proteins Allows Evaluation of Kink Sequence Predictors. *J. Chem. Inf. Model* 2010, **50**, 2213–2220.

120. Nordlund, P.; Eklund, H. Structure and Function of the *Escherichia coli* Ribonucleotide Reductase Protein R2. *J. Mol. Biol.* 1993, **232**, 123–164.

121. Phillips, S. E.; Schoenborn, B. P. Neutron Diffraction Reveals Oxygen-Histidine Hydrogen Bond in Oxymyoglobin. *Nature* 1981, **292**, 81–82.

122. Kamphuis, I. G.; Kalk, K. H.; Swarte, M. B.; Drenth, J. Structure of Papain Refined at 1.65 A Resolution. *J. Mol. Biol*. 1984, **179**, 233–256.

123. Dijkstra, B. W.; Kalk, K. H.; Hol, W. G.; Drenth, J. Structure of Bovine Pancreatic Phospholipase A2 at 1.7A Resolution. *J. Mol. Biol.* 1981, **147**, 97–123.

124. Rees, D. C.; Lewis, M.; Lipscomb, W. N. Refined Crystal Structure of Carboxypeptidase A at 1.54 A Resolution. *J. Mol. Biol.* 1983, **168**, 367–387.

125. Tronrud, D. E.; Monzingo, A. F.; Matthews, B. W. Crystallographic Structural Analysis of Phosphoramidates as Inhibitors and Transition-State Analogs of Thermolysin. *Eur. J. Biochem.* 1986, **157**, 261–268.

126. Lovejoy, B.; Cascio, D.; Eisenberg, D. Crystal Structure of Canine and Bovine Granulocyte-Colony Stimulating Factor (G-CSF). *J. Mol. Biol.* 1993, **234**, 640–653.

127. Walsh, M. A.; Schneider, T. R.; Sieker, L. C.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. Refinement of Triclinic Hen Egg-White Lysozyme at Atomic Resolution. *Acta Crystallogr. D Biol Crystallogr.* 1998, **54**, 522–546.

128. Wlodawer, A.; Svensson, L. A.; Sjolin, L.; Gilliland, G. L. Structure of Phosphate-Free Ribonuclease A Refined at 1.26 A. *Biochemistry* 1988, **27**, 2705–2717.

129. Shaw, A.; McRee, D. E.; Vacquier, V. D.; Stout, C. D. The Crystal Structure of Lysin, a Fertilization Protein. *Science* 1993, **262**, 1864–1867.

130. Hall, S. E.; Roberts, K.; Vaidehi, N. Position of Helical Kinks in Membrane Protein Crystal Structures and the Accuracy of Computational Prediction. *J. Mol. Graph. Mod.* 2009, **27**, 944–950.

131. Kauko, A.; Illergård, K.; Elofsson, A. Coils in the Membrane Core Are Conserved and Functionally Important. *J. Mol. Biol.* 2008, **380**, 170–180.

132. Riek, R. P.; Rigoutsos, I.; Novotny, J.; Graham, R. M. Non-α-Helical Elements Modulate Polytopic Membrane Protein Architecture. *J. Mol. Biol.* 2001, **306**, 349–362.

133. Harris, T.; Graber, A. R.; Covarrubias, M. Allosteric Modulation of a Neuronal K+ Channel by 1-Alkanols Is Linked to a Key Residue in the Activation Gate. *Am. J. Physiol. Cell. Physiol.* 2003, **285**, C788–C796.

134. Reddy, T.; Ding, J.; Li, X.; Sykes, B. D.; Rainey, J. K.; Fliegel, L. Structural and Functional Characterization of TM IX of the NHE1 Isoform of the Na^+/H^+ Exchanger. *J. Biol. Chem.* 2008, **283**, 22018–22030.

135. Steigemann, W.; Weber, E. Structure of Erythrocruorin in Different Ligand States Refined at 1.4 A Resolution. *J. Mol. Biol.* 1979, **127**, 309–338.

136. Vojtechovsky, J.; Chu, K.; Berendzen, J.; Sweet, R. M.; Schlichting, I. Crystal Structures of Myoglobin-Ligand Complexes at Near-Atomic Resolution. *Biophys. J.* 1999, **77**, 2153–2174.

137. Sugiura, I.; Nureki, O.; Ugaji-Yoshikawa, Y.; Kuwabara, S.; Shimada, A.; Tateno, M.; Lorber, B.; Giege, R.; Moras, D.; Yokoyama, S.; Konno, M. The 2.0 A Crystal Structure of Thermus thermophilus Methionyl-tRNA Synthetase Reveals Two RNA-Binding Modules. *Structure* 2000, **8**, 197–208.

138. Romier, C.; Dominguez, R.; Lahm, A.; Dahl, O.; Suck, D. Recognition of Single-Stranded DNA by Nuclease P1: High Resolution Crystal Structures of Complexes with Substrate Analogs. *Proteins* 1998, **32**, 414–424.

139. Hough, E.; Hansen, L. K.; Birknes, B.; Jynge, K.; Hansen, S.; Hordvik, A.; Little, C.; Dodson, E.; Derewenda, Z. High-Resolution (1.5 A) Crystal Structure of Phospholipase C from *Bacillus cereus*. *Nature* 1989, **338**, 357–360.

140. Silva, M.; Poland, B.; Hoffman, C.; Fromm, H.; Honzatko, R. Refined Crystal Structures of Unligated Adenylosuccinate Synthetase from Refined Crystal Structures of Unligated Adenylosuccinate Synthetase from *Escherichia coli*. *J. Mol. Biol.* 1995, **254**, 431–446.

141. Fedorov, A.; Shi, W.; Kicska, G.; Fedorov, E.; Tyler, P. C.; Furneaux, R. H.; Hanson, J. C.; Gainsford, G. J.; Larese, J. Z.; Schramm, V. L.; Almo, S. C. Transition State Structure of Purine Nucleoside Phosphorylase and Principles of Atomic Motion in Enzymatic Catalysis. *Biochemistry* 2001, **40**, 853–860.

142. Nagar, B.; Jones, R. G.; Diefenbach, R. J.; Isenman, D. E.; Rini, J. M. X-Ray Crystal Structure of C3d: A C3 Fragment and Ligand for Complement Receptor 2. *Science* 1998, **280**, 1277–1281.

143. Kühnel, W. *Differential Geometry: Curves - Surfaces - Manifolds*; Dover, New York, NY, 2002.

144. Koch, O.; Klebe, G. Turns Revisited: A Uniform and Comprehensive Classification of Normal, Open, and Reverse Turn Families Minimizing Unassigned Random Chain Portions. *Proteins* 2009, **74**, 353–367.

145. Vijay-Kumar, S.; Bugg, C.; Cook, W. Structure of Ubiquitin Refined at 1.8 A Resolution. *J. Mol. Biol.* 1987, **194**, 531–544.

146. Kannan, K. K.; Ramanadham, M.; Jones, T. A. Structure, Refinement, and Function of Carbonic Anhydrase Isozymes: Refinement of Human Carbonic Anhydrase I. *Ann. N. Y. Acad. Sci.* 1984, **429**, 49–60.

147. Cook, W. J.; Zell, A.; Watt, D. D.; Ealick, S. E. Structure of Variant 2 Scorpion Toxin from Centruroides sculpturatus Ewing. *Protein Sci.* 2002, **11**, 479–486.

148. Walter, J.; Steigemann, W.; Singh, T.; Bartunik, H.; Bode, W.; Huber, R. On the Disordered Activation Domain in Trypsinogen. Chemical Labelling and Low-Temperature Crystallography. *Acta Crystallogr., Sect. B* 1982, **38**, 1462–1472.

149. Wlodawer, A.; Walter, J.; Huber, R.; Sjolin, L. Structure of Bovine Pancreatic Trypsin Inhibitor. Results of Joint Neutron and X-Ray Refinement of Crystal Form II. *J. Mol. Biol*. 1984, **180**, 301–329.

150. Adman, E.; Sieker, L. C.; Jensen, L. Structure of Rubredoxin from Desulfovibrio vulgaris at 1.5 A Resolution. *J. Mol. Biol*. 1991, **217**, 337–352.

151. Bonneau R.; Baker, D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* 2001, **30**, 173–189.

152. Zhang, Y. Progress and Challenges in Protein Structure Prediction. *Curr. Opin. Struct. Biol.* 2008, **18**, 342–348.

153. Chothia, C.; Lesk, A. M. The Relation Between the Divergence of Sequence and Structure in Proteins. *EMBO J.* 1986, **5**, 823–826.

154. Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 1993, **234**, 779–815.

155. Zhang, Y.; Skolnick, J. The Protein Structure Prediction Problem Could Be Solved Using the Current PDB Library. *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**, 1029–1034.

156. Bowie, J. U.; Lüthy, R.; Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* 1991, **253**, 164–170.

157. Jones, D. T.; Taylor, W. R.; Thornton, J. M. A New Approach to Protein Fold Recognition. *Nature* 1992, **358**, 86–89.

158. Bryant, S. H.; Altschul, S. F. Statistics of Sequence-Structure Threading. *Curr. Opin. Struct. Biol.* 1995, **253**, 236–244.

159. Lee, L.; Wu, S.; Zhang, Y. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*; Rigden, D. J., Ed.; Springer, London, 2009, pp. 1–26.

160. Floudas, C.; Fung, H.; McAllister, S.; Mönnigmann, M.; Rajgaria, R. Advances in Protein Structure Prediction and De Novo Protein Design: A Review. *Chem. Eng. Sci. Biomol. Eng.* 2006, **61**, 966–988.

161. Hasegawa, H.; Holm, L. Advances and Pitfalls of Protein Structural Alignment. *Curr. Opin. Struct. Biol.* 2009, **19**, 341–348.

162. Marti-Renom, M. A.; Capriotti, E.; Shindyalov, I. N.; Bourne, P. E. In *Structural Bioinformatics*, 2nd Edition; Gu, J., Bourne, P. E., Eds.; John Wiley & Sons, Ltd, Chichester, UK, 2009; pp. 397–417.

163. Lancia, G.; Sorin, I. Protein Structure Comparison: Algorithms and Applications. In *Protein Structure Analysis and Design*; Guerra, C., Istrail, S., Eds.; Springer, Berlin, 2003, pp. 1–33.

164. Sierk, M. L.; Pearson, W. R. Sensitivity and Selectivity in Protein Structure Comparison. *Protein Sci.* 2004, **13**, 773–785.

165. Nayeem, A.; Sitkoff, D.; Krystek, S. A Comparative Study of Available Software for High-Accuracy Homology Modeling: From Sequence Alignments to Structural Models. *Protein Sci.* 2006, **15**, 808–824.

166. Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystall.* 1976, **922**, 702–710.

167. Coutsias, E. A.; Seok, C.; Dill, K. A. Using Quaternions to Calculate RMSD. *J. Comput. Chem.* 2004, **25**, 1849–1857.

168. Ambühl, C.; Chakraborty, S.; Gärtner, B. Algorithms—ESA 2000, Computing Largest Common Point Sets Under Approximate Congruence. *Lect. Notes Comput. Sci.* 2000, **1879**, 52–64.

169. Akutsu, T.; Halldorsson, T. A. On the Approximation of Largest Common Subtrees and Largest Common Point Sets. *Theor. Comput. Sci.* 2000, **233**, 33–50.

170. van Leeuwen, J. *Handbook of Theoretical Computer Science*; Elsevier, New York, 1998.

171. Wohlers, I.; Domingues, F.; Klau, G. W. Towards Optimal Alignment of Protein Structure Distance Matrices. *Bioinformatics* 2010, **26**, 2273–2280.

172. Shindyalov, I. N.; Bourne, P. E. Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. *Protein Eng.* 1998, **11**, 739–747.

173. Orengo, C.; Taylor, W. SSAP: Sequential Structure Alignment Program for Protein Structure Comparison. *Methods Enzymol.* 1996, **266**, 617–635.

174. Taylor, W. R.; Orengo, C. A. Protein Structure Alignment. *J. Mol. Biol.* 1989, **208**, 1–22.

175. Orengo, C. A.; Taylor, W. R. A Rapid Method of Protein Structure Alignment. *J. Theor. Biol.* 1990, **147**, 517–551.

176. Orengo, C. A.; Taylor, W. R. A Local Alignment Method for Protein Structure Motifs. *J. Mol. Biol.* 1993, **233**, 488–497.

177. Zemla, A. LGA: A Method for Finding 3-D Similarities in Protein Structures. *Nucleic Acids Res.* 2003, **31**, 3370–3374.

178. Siew, A., Elofsson, N.; Rychlewsk, L.; Fischer, D. MaxSub: An Automated Measure for the Assessment of Protein Structure Prediction Quality. *Bioinformatics* 2000, **16**, 776–785.

179. Madej, T. G. J.; Bryant, S. H. Threading a Database of Protein Cores. *Proteins* 1995, **23**, 356–369.

180. Leslin, C. M.; Abyzov, A.; Ilyin, V. A. TOPOFIT-DB, a Database of Protein Structural Alignments Based on the TOPOFIT Method. *Nucleic Acids Res.* 2007, **35**, D317–D321.

181. Ilyin, V. A.; Abyzov, A.; Leslin, C. M. Structural Alignment of Proteins by a Novel TOPOFIT Method, as a Superimposition of Common Volumes at a Topomax Point. *Protein Sci.* 2004, **13**, 1865–1874.

182. Madhusudhan, M. S.; Webb, B. M.; Marti-Renom, M. A.; Eswar, N.; Sali, A. Alignment of Multiple Protein Structures Based on Sequence and Structure Features. *Protein Eng. Des. Sel.* 2009, **22**, 569–574.

183. Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* 2005, **33**, 2302–2309.

184. Teichert, F.; Bastolla, U.; Porto, M. SABERTOOTH: Protein Structural Alignment Based on a Vectorial Structure Representation. *BMC Bioinf.* 2007, **8**, 425.1–425.17.

185. Guerler, A.; Knapp, E. W. Novel Protein Folds and Their Non-Sequential Structural Analogs. *Protein Sci.* 2008, **17**, 1374–1382.

186. Mosca, R.; Schneider, T. R. RAPIDO: A Web Server for the Alignment of Protein Structures in the Presence of Conformational Changes. *Nucleic Acids Res.* 2008, **36**, W42–W46.

187. Bomar, M. G.; D'Souza, S.; Bienko, M.; Dikic, I.; Walker, G. C.; Zhou, P. Unconventional Ubiquitin Recognition by the Ubiquitin-Binding Motif Within the Y Family DNA Polymerases ι and Rev1. *Mol. Cell* 2010, **37**, 408–417.

188. Ross, S.; Sarisky, C.; Su, A.; Mayo, S. Designed Protein G Core Variants Fold to Native-Like Structures: Sequence Selection by ORBIT Tolerates Variation in Backbone Specification. *Protein Sci.* 2001, **10**, 450–454.

189. Pai, E. F.; Krengel, U.; Petsko, G. A.; Goody, R.S.; Kabsch, W.; Wittinghofer, A. Refined Crystal Structure of the Triphosphate Conformation of H-ras p21 at 1.35Å Resolution: Implications for the Mechanism of GTP Hydrolysis. *EMBO J.* 1990, **9**, 2351–2359.

190. Tong, L. A.; de Vos, A. M.; Milburn, M. V.; Kim, S. H. Crystal Structures at 2.2Å Resolution of the Catalytic Domains of Normal Ras Protein and an Oncogenic Mutant Complexed with GDP. *J. Mol. Biol.* 1991, **217**, 503–516.

191. Malumbres, M.; Barbacid, M. RAS Oncogenes: The First 30Years. *Nat. Rev. Cancer* 2003, **3**, 459–465.

192. Goodsell, D. S. The Molecular Perspective: The Ras Oncogene. *Oncologist* 1999, **4**, 263–264.

193. Mount, D. W. In *Bioinformatics: Sequence and Genome Analysis*; Mount, D. W., Ed.; Cold Spring Harbor Laboratory Press, New York, 2004.

194. Kolodny, R.; Petrey, D.; Honig, B. Protein Structure Comparison: Implications for the Nature of 'Fold Space', and Structure and Function Prediction. *Curr. Opin. Struct. Biol.* 2006, **16**, 393–398.

195. Wohlers, I.; Malod-Dognin, N.; Andonov, R.; Klau, G. W. CSA: Comprehensive Comparison of Pairwise Protein Structure Alignments. *Nucleic Acids Res.* 2012, **40**, W303–W309.

196. Vingron, M.; Waterman, M. S. Sequence Alignment and Penalty Choice: Review of Concepts, Case Studies and Implications. *J. Mol. Biol.* 1994, **235**, 1–12.

197. Altschul, S. F.; Erickson, B. W. Optimal Sequence Alignment Using Affine Gap Costs. *Bull. Math. Biol.* 1986, **48**, 603–616.

198. Zachariah, M. A.; Crooks, G. E.; Holbrook, S. R.; Brenner, S. E. Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy. *Proteins* 2005, **58**, 329–338.

199. Beisel, H.; Kawabata, S.; Iwanaga, S.; Huber, R.; Bode, W. Tachylectin-2: Crystal Structure of a Specific GlcNAc/GalNAc-Binding Lectin Involved in the Innate Immunity Host Defense of the Japanese Horseshoe Crab Tachypleus tridentatus. *EMBO J.* 1999, **18**, 2313–2322.

200. Leinala, E.; Davies, P.; Jia, Z. Crystal Structure of Beta-Helical Antifreeze Protein Points to a General Ice Binding Model. *Structure* 2002, **10**, 619–627.

201. Maiorov, V. N.; Crippen, G. M. Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins. *J. Mol. Biol.* 1994, **235**, 625–634.

202. Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* 2004, **57**, 702–710.

203. Zhu, S.; Aumelas, A.; Gao, B. Convergent Evolution-Guided Design of Antimicrobial Peptides Derived from Influenza A Virus Hemagglutinin. *J. Med. Chem.* 2011, **54**, 1091–1095.