

# Secondary Structure Assignment

Claus A. Andersen<sup>2</sup> and Burkhard Rost<sup>1\*</sup>

<sup>1</sup> CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

<sup>2</sup> Siena Biotech Spa. Via Fiorentina 1, 53100 Siena, Italy

\* Corresponding author: [rost@columbia.edu](mailto:rost@columbia.edu), <http://cubic.bioc.columbia.edu/> Tel: +1-212-305-3773, fax: +1-212-305-7932

*The task.* When looking at how chains of amino acids in proteins are ordered we notice regular macro elements in the 3D structure of practically all known structures: helices and strands (see Chapter 2). These structures were coined protein secondary structure by Linderstrøm-Lang (1952) in the context of a protein's primary, tertiary and quaternary structural definitions. There is no unique inherent physical characteristic to systematically assign secondary structure from 3D coordinates. Instead there are many different assignment schemes each constructed to reflect one or more aspects of protein structure. Here we will address these aspects, the assignment schemes and how they have been used to study proteins.

*Aspects reflected in secondary structure assignments.* When protein secondary structure is assigned the complexity of the all atom 3D structure is reduced dramatically. This level of abstraction is obtained by choosing as determinants for the assignment such aspects as physical interaction energies (e.g. H-bonds and van der Waals), geometrical idealization (e.g. idealized cylinder or  $C_{\alpha}$ -distance masks) and/or other structural descriptions optimized to reflect appealing characteristics (e.g. expert assignment, invariance upon thermal fluctuations or predictability). Another generic aspect of all methods is that they assign secondary structure based on residue independent aspects.

*Notation.* When treating protein secondary structure we will use the terms: *state*, *class* and *regular secondary structure*. In the literature they are not used consistently so we refer to the following notation: *states* are the types of secondary structure defined by a particular method e.g. G refers to  $3_{10}$ -helix in DSSP; *classes* are the groups of similar states e.g. G, H and I all describing helices in DSSP; *regular secondary structure* are all the states belonging to a direct secondary structure assignment. Note that *non-regular structure* thus refers to the remaining states, which on occasion go by names such as random coil or loop. Super secondary structure refers to the relative spatial distances and orientations between two or more secondary structure elements.

*Usage of secondary structure.* The application of secondary structure assignments is very diverse and covers many areas of protein analyses, most noticeably, within structure inspection and visualization (see Chapter 9) and the related task of structure comparison and classification (see Chapters 16,17, and 18). Secondary structure has also been employed in protein modeling and structure prediction (see Chapter 29), as well as in studies of protein folding, dynamics, interactions and function. The one-dimensional nature of the secondary structure description has furthermore been used in sequence alignment. Some applications will be treated in more detail below after the assignment methods have been described.

*History: from expert to automatic assignment of protein secondary structure.* Pauling and colleagues correctly predicted the idealized protein secondary structures of  $\alpha$ -helices (Pauling, Corey, and Branson, 1951),  $\pi$ -helices (Pauling, Corey and Branson, 1951), and of  $\beta$ -sheets (Pauling and Corey, 1951) based on intra-backbone hydrogen bonds. Five decades later, we know that on average about half of the residues in proteins participate in helices or sheets (Berman et al., 2000). Pauling and colleagues incorrectly predicted that  $3_{10}$ -helices would not occur in proteins, due to unfavorable bond angles; however, approximately 4% of the residues are observed in this conformation (Andersen, 2001). Initially, the crystallographers assigned secondary structure by eye from the 3D structures. At the time this was the only way to assign secondary structure. However, it

lacked consistency, since experts occasionally disagree. This was particularly problematic when comparing secondary structure predictions and was actually the primary objective for Kabsch & Sander to automate the assignment in their DSSP program (1983a, 1983b). Originally developed to improve secondary structure prediction, DSSP has remained the standard in the field, most popular for its relatively reliable assignments. Curiously, the prediction method for which Kabsch & Sander originally needed the automatic assignment was never published.

*Experimental investigations of protein secondary structure.* The structure of a protein can be determined at various levels of precision and timescale. X-ray crystallography (see Chapter 4) is widely used and generally provides a static snapshot with all atom resolution, whereas NMR (see Chapter 5) furthermore can measure dynamic motion of proteins in solution, but not below the millisecond regime (Doerr, 2007). Optical spectroscopy is a much faster technique and has been used to inspect H-bond dynamics at a picosecond time scale for a small  $\beta$ -turn peptide (Kolano, 2006). In particular circular dichroism (CD) and Raman spectroscopy are used to characterize overall protein secondary structure dynamics in solution, since the helix and sheet structures give strong characteristic spectra which are highly correlated with X-ray data (Lees, 2006; Tetin, 2003; Janes, 2005). This allows the rapid assessment of conformational changes resulting from ligand binding, macromolecular interactions etc. and conformational assessment of natively unfolded proteins (Pelton, 2000; Maiti, 2004). Spectroscopy resolution can be further enhanced with residue-specific isotope labeling e.g. to dissect the conformation of helical peptides at the residue level (Decatur 2000; Fesinmeyer 2005).

Attempts have been made to determine the protein secondary structure and stability by mass spectrometry (Villanueva, 2002), but the specific technique presented is not likely to be a valuable conformational probe (Beynon, 2004).

---

## SECONDARY STRUCTURE CONCEPTS

The hydrogen bond is used by many methods to describe and assign protein secondary structures so we will introduce this concept and some definitions employed. Using the hydrogen bond spurs from the notion of assigning secondary structure based on the local energy gained in stabilizing the polypeptide chain in a given conformation. Following this notion the energetic calculation can also be extended to the rest of the protein backbone atom interactions calculating electrostatic and van der Waals interaction energies, described here as the backbone-backbone interaction energy.

Likewise mathematical concepts are applied to secondary structure assignment. These may be basic geometrical objects which can be readily comprehended (e.g. a straight line or cylinder) or may require some introduction as done for Voronoï tessellation presented here.

---

### **Hydrogen bond energy**

Pauling (1939) established the hydrogen bond as an important principle in chemistry. The rich network of hydrogen bonds in water creates a very particular environment in which polar molecules participate, while non-polar molecules disrupt the network of hydrogen bonds. This results in missing water-water hydrogen bonds and therefore a relative energy cost compared to the hydrogen bonded case (4 kcal/mol for Isoleucine and Leucine when compared to Glycine (Creighton, 1993)). This energy cost is in the order of two hydrogen bonds (hydrogen bonds are in the range of -2 kcal/mol) and can be avoided or minimized by packing via agglomerating non-polar molecules, thereby resulting in the hydrophobic effect.

The packing of non-polar residues in the core is believed to be the main driving force in tertiary structure formation of proteins, while the specific secondary structures are governed by

intramolecular hydrogen bonds (Hvidt and Westh, 1998). Packing the non-polar residues in the core also means burying the polar backbone atoms and breaking the water-backbone hydrogen bonds. To avoid this heavy energetic cost, the polarities are paired (forming hydrogen bonds) in the protein core, thus fixing the protein conformation. If instead the protein backbone were non-polar, the protein core elements would then be free to move around thus leading to a highly dynamic protein structure and thereby preventing the protein from functioning reliably and efficiently.

Approximately 90% of the backbone C=O and NH groups participate in hydrogen bonds (Baker and Hubbard, 1984). Using the Coulomb hydrogen bond definition (see below), we found that approximately 62% of the backbone C=O and NH groups participate in intra-backbone hydrogen bonds (Andersen, 2001). Pauling defined secondary structure by the intra-backbone hydrogen bonds, and this has later become the prevalent means of assigning secondary structure. Thus, for simplicity, we refer to intra-backbone hydrogen bonds when using the term 'hydrogen bond'.

---

### **Angle-distance hydrogen bond assignment**

There are many different angles and distances that can be measured and used to identify the hydrogen bond. Baker and Hubbard (1984) assigned hydrogen bonds according to the inter-atom angle  $\text{NHO} = \varphi$  and distance  $r_{\text{HO}}$  in the hydrogen bond. A hydrogen bond is assigned when:

$\varphi > 120^\circ$  and  $r_{\text{HO}} < 2.5 \text{ \AA}$

This is similar to other rigid distance and angle constraints published (Bordo and Argos, 1994; Jeffrey and Saenger, 1994), and can be simplified further by considering only the donor-acceptor distance i.e. in this case hydrogen bonds are assigned when  $r_{\text{HO}} < 3.5 \text{ \AA}$ . Although a rather crude way of assigning hydrogen bonds, it has sufficed for several decades and is still being used.

---

### **Coulomb hydrogen bond energy calculation**

One way of finding hydrogen bonds is by calculating the Coulomb energy in the bond, as applied in DSSP (see below) focusing on the electrostatic attraction (Figure 1). The Coulomb energy for the attraction and repulsion is given by:

**Equation 1**

$$E = f\delta^+\delta^-\left(\frac{1}{r_{\text{NO}}} + \frac{1}{r_{\text{HC}'}} - \frac{1}{r_{\text{HO}}} - \frac{1}{r_{\text{NC}'}}\right)$$

where  $f = 332 \text{ \AA kcal}/(e^2 \text{ mol})$  is the dimensional factor and  $\delta^+ = 0.20e$  and  $\delta^- = -0.42e$  are the polar charges given in units of the elementary electron charge  $e$ . A cut-off level has been set for the weakest acceptable hydrogen bond so that the resulting energy is bound by:  $E < -0.5 \text{ kcal/mol}$  in DSSP. In practice, the hydrogen atom position is usually not given in PDB files requiring an extrapolation. For example, the hydrogen atom position that is needed to calculate the two distances  $r_{\text{OH}}$  and  $r_{\text{HC}'}$  in Equation 1 is usually not given in the PDB files. DSSP circumvents this problem by using an approximate position, assuming that the covalent bond between O=C' is parallel to the covalent N-H bond adjacent to the same polypeptide bond. The direction of the O=C' vector is kept while its length is set to  $1 \text{ \AA}$ , i.e., the length of the N-H bond (Chreighton, 1993). The position of the H-atom is extrapolated using the direction of the C'=O vector when starting out from the position of the N-atom. These approximations made by DSSP simplify the calculation of the H-atom position and appear to be rather accurate despite the assumptions that were being made. When compared to the original bond angles and distances (Chreighton, 1993), we found the DSSP approximation to yield an average error around  $0.07 \text{ \AA}$  (Andersen, 2001). The assumption of the trans-peptide bond, giving rise to the rigid peptide plane, was used by DSSP as well as our tests. Partitioning *ab initio* energy calculations of the hydrogen bond into classical components showed that about 75% is electrostatic (Coulombic) and less than 5% comes from polarization and charge-

transfer, for moderate strength bonds (Jeffrey and Saenger, 1994). Note that the Coulomb energy term does not incorporate atom-atom repulsion to penalize steric clashes and does not give rise to a characteristic hydrogen bond length.

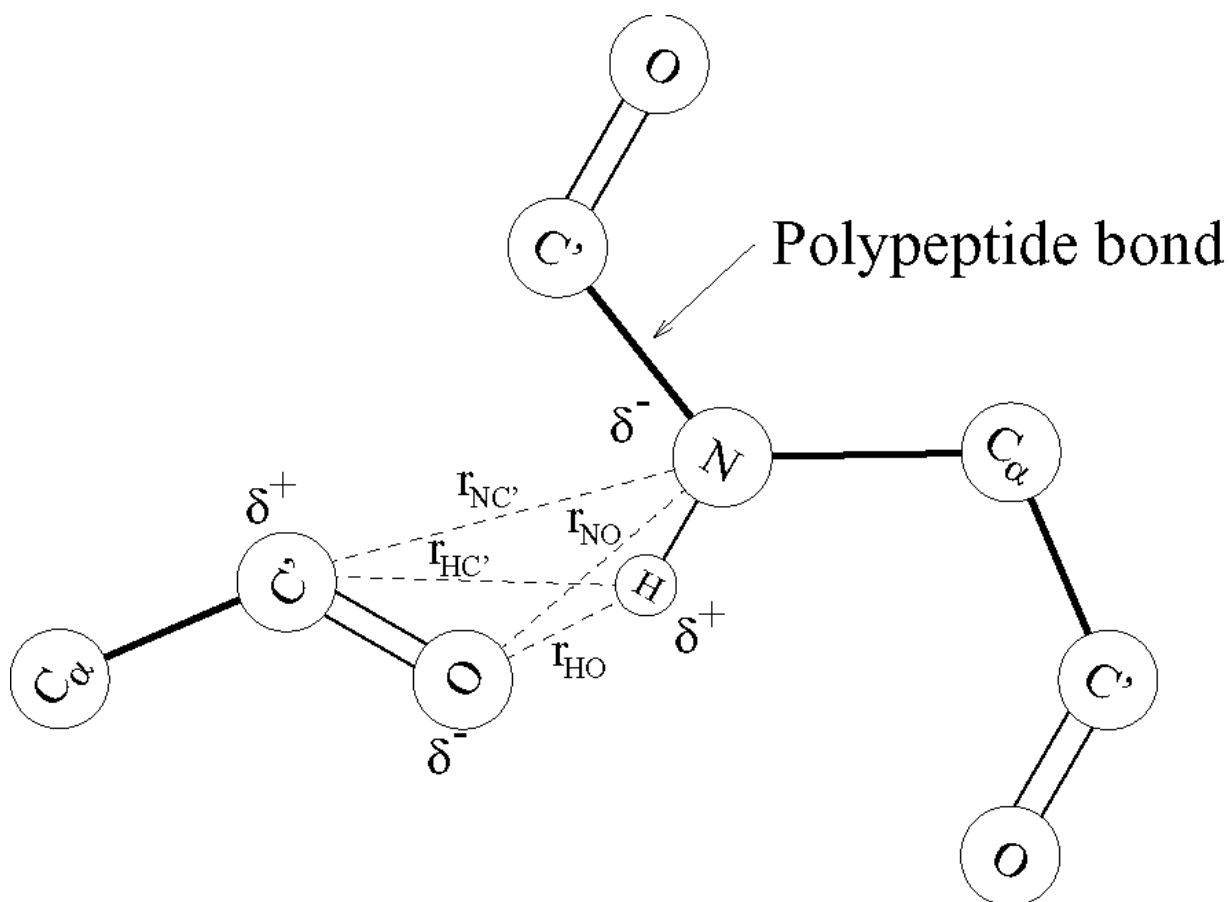


Figure 1: Distances used to calculate the Coulomb hydrogen bond energy.

### ***Empirical hydrogen bond energy calculation***

An empirical hydrogen bond energy calculation can be derived from the hydrogen bond geometry in crystal structures or from polypeptides, peptides, amino acids and small organic compounds (Boobbyer, 1989; Wade, 1993) as applied in STRIDE (see below). The total energy  $E_{hb}$  depends on the NO distance energy  $E_r$ , (reflecting optimal atom distance and atom boundary) and on three bonding angles through the expressions  $E_p$  and  $E_t$  (reflecting favorable hydrogen bond angles extrapolated from electron orbital interactions):

$$\text{Equation 2}$$

$$E_{hb} = E_r \cdot E_t \cdot E_p$$

The distance dependency energy  $E_r$  is similar to the Lennard-Jones potential for the van der Waals interaction, but uses powers of 8 and 6 instead of 12 and 6. Thus reducing the slope of the atom-atom superposition term, whereby the penalty for superpositions is more lenient towards experimental inaccuracies in atom position determination.

$$\text{Equation 3}$$

$$E_r = \left( \frac{4 r_m^6}{r^6} - \frac{3 r_m^8}{r^8} \right) E_m$$

where  $r$  is the NO distance,  $r_m$  is the optimal distance and  $E_m$ , the optimal energy. For intra-backbone hydrogen bonds  $r_m = 3.0 \text{ \AA}$  and  $E_m = -2.8 \text{ kcal/mol}$  is used. The two angular dependent terms are:

$$\text{Equation 4}$$

$$E_p = \cos^2(\theta)$$

$$E_t = \begin{cases} [0.9 + 0.1 \sin(2t_i)] \cos(t_o) & 0^\circ < t_i \leq 90^\circ \\ K_1 [K_2 - \cos^2(t_i)] \cos(t_o) & 90^\circ < t_i \leq 110^\circ \\ 0 & 110^\circ \leq t_i \end{cases}$$

where the angles  $\theta$ ,  $t_i$  and  $t_o$  are specified in Figure 3.

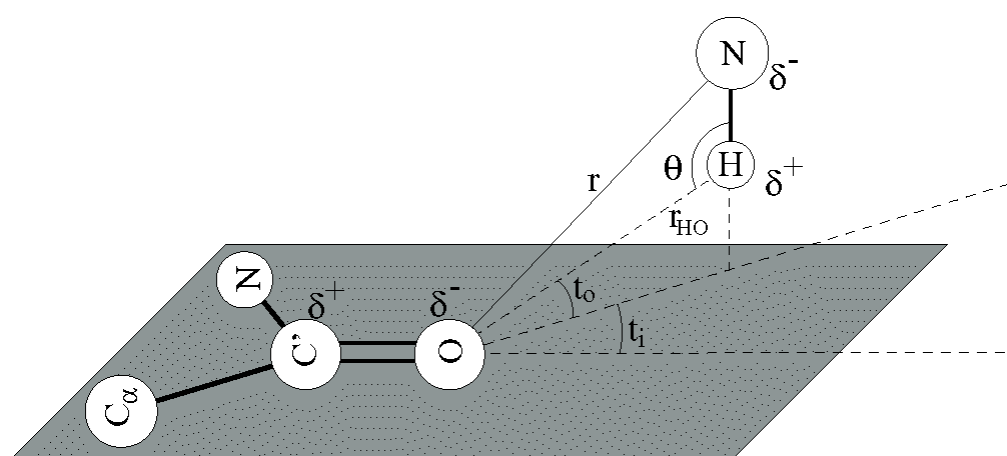


Figure 3: Angles and distances defining the empirical hydrogen bond. Note: figure similar to the one in Frishman and Argos (1995).

### Backbone-backbone interaction energy

The Coulomb and van der Waals interaction energy calculations can also be applied to all backbone atom interactions (i.e. involving the atoms N, C $_{\alpha}$ , C', O, H $_N$  and H $_{\alpha}$ ) as applied in  $\beta$ -spider (see below), thereby covering the two potential backbone hydrogen bonds formed between two residues (described above), as well as the C $_{\alpha}$ -H $_{\alpha}$ ...O=C hydrogen bond and the C=O...C=O dipole. This energy evaluation is calculated for each atom pair ( $E_{ij}^A$ ) and subsequently summed over all pairs between two residues ( $E_{ij}^R$ ).  $E_{ij}^A$  has the same functional form as the Amber force field (Cornell, 1995):

$$\text{Equation 5}$$

$$E_{ij}^A = \epsilon_{ij}^* \left( \frac{R_{ij}^*}{R_{ij}} \right)^{12} - 2\epsilon_{ij}^* \left( \frac{R_{ij}^*}{R_{ij}} \right)^6 + 332 \frac{Q_i Q_j}{R_{ij}}$$

where  $E_{ij}^A$  is the energy between atoms  $i$  and  $j$  with observed distance  $R_{ij}$  [ $\text{\AA}$ ], mixing rules  $\epsilon_{ij}^*$  polar charges  $Q_i$ ,  $Q_j$  and optimal distance  $R_{ij}^*$ . The optimal distance, mixing rules and polar charges were taken from Amber. In the special cases of Glycine and Proline, the backbone constellation has been adjusted accordingly by adding another H $_{\alpha}$  and removing the H $_N$  respectively. In  $\beta$ -spider the hydrogen atom positions were extrapolated geometrically from the N, C $_{\alpha}$ , C' and O coordinates using bond lengths, valence and torsion angles from Amber (Cornell, 1995).

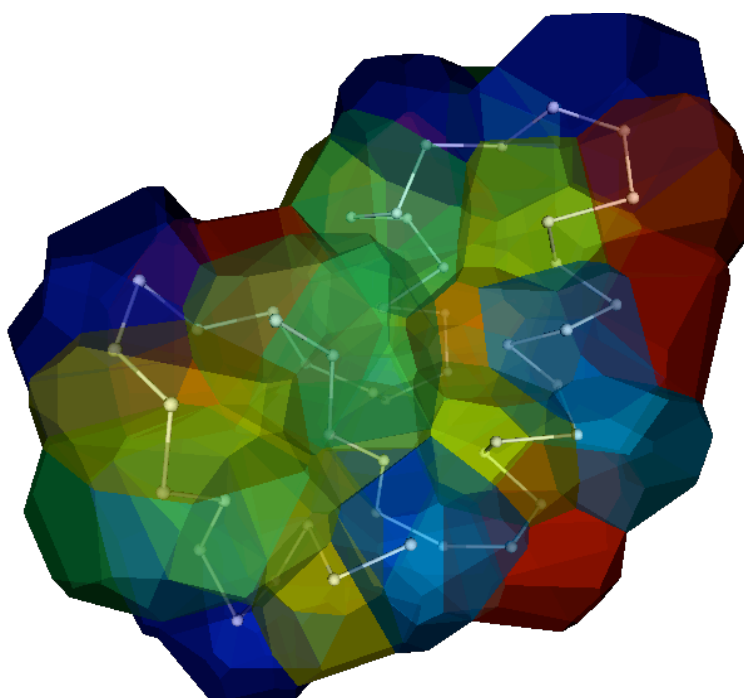
### Equation 6

$$E_{ij}^R = \sum_{\{N, C_\alpha, C, O, H_N, H_\alpha\}}^i \sum_{\{N, C_\alpha, C, O, H_N, H_\alpha\}}^j E_{ij}^A$$

---

## ***Voronoi tessellation for geometrical residue partitioning***

By geometrically deriving a polyhedron around the  $C_\alpha$  of each residue using Voronoi tessellation a new definition of contact maps is described. This is very informative about the relative packing of residues and has been applied to assign secondary structure by VoTAP (see below). Each  $C_\alpha$  is contained within a Voronoi cell which is determined by Delaunay tetrahedral decomposition. This is a unique decomposition with non-overlapping cells that only contain one  $C_\alpha$  atom each and where all internal space inside the protein is contained within a cell as shown in Figure 5.



---

**Figure 5: Voronoi tessellation partitions the protein space into polyhedra surrounding the  $C_\alpha$  atom of each residue (Dupuis, 2004). A unique  $C_\alpha$  atom contact map can thus be defined as the polyhedra sharing a contact surface. The example shown is Crambin (Teeter, 1984;PDB: 1crn) visualized from an angle similar to the one employed in Figure 12A, using the voro3D tool made available to the community by Frank Dupuis (see Table 1).**

---

## ***Converting secondary structure states to three classes***

Several secondary structure assignment methods are presently available, but DSSP continues to be the most widely used method followed by STRIDE. In fact, most prediction methods are based on DSSP assignments. Typically, the 8 DSSP states are converted into three classes using the following convention: [GHI] -> h, [EB] -> e, [TS' ] -> c, which reads:  $3_{10}$ -,  $\alpha$ ,  $\pi$ -helices are grouped into one helix class; Extended  $\beta$ -sheets and  $\beta$ -bridges are grouped into one sheet class; and the remaining secondary structure states turn, bend and “not assigned” are grouped into one coil class.

Usually,  $3_{10}$ -helices and  $\beta$ -bridges constitute short secondary structure segments that have some structural similarity to  $\alpha$ -helix and  $\beta$ -strand, respectively. However, they do have different sequence characteristics. Prediction methods, in general, are more precise in the core of regular secondary structure segments than at the termini (Rost, 1994; Cuff, 1999). Thus,  $3_{10}$ -helices and  $\beta$ -bridges are

more difficult to predict than  $\alpha$ -helices and  $\beta$ -strands. Therefore an alternative conversion that has been used more recently yields a seemingly higher level of prediction accuracy: [H] -> h, [E] -> e, [GIBTS' ] -> c.

---

## ASSIGNMENT METHODS

---

### ***DSSP: H-bond pattern based assignment***

The so-called Dictionary of Secondary Structure of Proteins (DSSP) by Kabsch and Sander (1983a) performs its sheet and helix assignments solely based on the backbone-backbone hydrogen bonds. The DSSP method defines a hydrogen bond when the bond energy is below -0.5 kcal/mol from a Coulomb approximation of the hydrogen bond energy (see SECONDARY STRUCTURE CONCEPTS). The structure assignments are defined such that visually appealing and unbroken structures result. In case of overlaps,  $\alpha$ -helix is given first priority, followed by  $\beta$ -sheet. This procedure does not affect the Coulomb approximation, rather the realisation of 'unbroken structures' addresses the step from individual hydrogen bonds to assigning macro-structures to groups of such bonds.

An  $\alpha$ -helix assignment (DSSP state 'H') starts when two consecutive amino acids have  $i \rightarrow i+4$  hydrogen bonds, and ends likewise with two consecutive  $i \rightarrow i-4$  hydrogen bonds. This definition is also used for  $3_{10}$ -helices (state 'G' with  $i \rightarrow i+3$  hydrogen bonds) and for  $\pi$ -helices (state 'I' with  $i \rightarrow i+5$  hydrogen bonds) as well. The helix definition does not assign the edge residues having the initial and final hydrogen bonds in the helix. A minimal size helix is set to have two consecutive hydrogen bonds in the helix, leaving out single helix hydrogen bonds, which are assigned as turns for all three helices (state 'T').

$\beta$ -sheet residues (state 'E') are defined as either having two hydrogen bonds in the sheet, or being surrounded by two hydrogen bonds in the sheet. This implies three sheet residue types: anti-parallel and parallel with two hydrogen bonds or surrounded by hydrogen bonds. The minimal sheet consists of two residues at each partner segment. Isolated residues fulfilling this hydrogen bond criterion are labelled as  $\beta$ -bridge (state 'B'). The recurring H-bonding patterns connecting the partnering strands in a  $\beta$ -sheet are occasionally interrupted by one or more so-called  $\beta$ -bulge residues. In DSSP these residues are also assigned as  $\beta$ -sheet 'E' and may comprise up to four residues on one strand and up to one residue on the partnering strand. These interruptions in the  $\beta$ -sheet H-bonding pattern are only assigned as sheet if they are surrounded by H-bond forming residues of the same type, i.e. either parallel or anti-parallel. The remaining two DSSP states 'S' and ' ' (space) indicate a bend in the chain and the unassigned/other state, respectively.

---

PDB:1crn

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N
....										
15	15	V	H << S+	0	0	99	-4,-1.7	3,-1.3	2,-0.2	-2,-0.2
16	16	c	H 3<>S+	0	0	18	-4,-2.5	5,-0.8	1,-0.3	-2,-0.2
17	17	R	H ><5S+	0	0	94	-4,-2.0	3,-1.6	1,-0.2	-1,-0.3
18	18	L	T <<5S+	0	0	144	-3,-1.3	-1,-0.2	-4,-0.6	-2,-0.2
19	19	P	T 3 5S-	0	0	107	0, 0.0	-1,-0.3	0, 0.0	-2,-0.1
20	20	G	T < 5 +	0	0	53	-3,-1.6	-3,-0.2	1,-0.2	-2,-0.1
21	21	T	< -	0	0	37	-5,-0.8	-1,-0.2	1,-0.1	5,-0.1
22	22	P	>> -	0	0	81	0, 0.0	4,-2.2	0, 0.0	3,-0.7
23	23	E	H 3> S+	0	0	70	1,-0.2	4,-2.5	2,-0.2	5,-0.1
24	24	A	H 3> S+	0	0	63	1,-0.2	4,-1.7	2,-0.2	-1,-0.2
25	25	I	H <> S+	0	0	99	-3,-0.7	4,-1.8	2,-0.2	-1,-0.2
26	26	c	H X S+	0	0	0	-4,-2.2	4,-1.9	2,-0.2	6,-0.4
27	27	A	H X S+	0	0	12	-4,-2.5	4,-2.7	-5,-0.2	5,-0.5
28	28	T	H < S+	0	0	120	-4,-1.7	-1,-0.2	1,-0.2	-2,-0.2
29	29	Y	H < S+	0	0	176	-4,-1.8	-1,-0.2	-5,-0.2	-2,-0.2
30	30	T	H < S-	0	0	24	-4,-1.9	-2,-0.2	-3,-0.2	-3,-0.2
31	31	G	S < S+	0	0	35	-4,-2.7	-3,-0.2	1,-0.4	-4,-0.1
32	32	b	-	0	0	5	-5,-0.5	-1,-0.4	-6,-0.4	2,-0.3
33	33	I	E -A	3	0A	51	-30,-2.8	-30,-2.4	-3,-0.1	2,-0.5
34	34	I	E -A	2	0A	78	-2,-0.3	-32,-0.2	-32,-0.2	3, 0.0
....										

Figure 4: Explanation of DSSP output example segment from Crambin (Teeter, 1984). The first two columns contain the unique DSSP residue number and the corresponding PDB residue number. The third column (here empty) indicates the chain identifier if there are multiple chains. Then follows the amino acid 'AA' in one letter codes (note: lower case letters are all Cysteines, in order to match up Cysteine-bridges, e.g. residue 16 has a disulfide bond to residue 26). The 'STRUCTURE' section starts with the secondary structure synopsis (HBEGITS listed in order of priority in case of overlaps) and is followed by helix hydrogen bond indications for  $3_{10}$ -,  $\alpha$ - and  $\pi$ -helix hydrogen bonds, where '>' indicates an acceptor, '<' a donor and 'X' both. The bend and chirality are each given a column followed by the  $\beta$ -bridge label columns (lower case labels are parallel  $\beta$ -bridges and upper case are anti-parallel). The DSSP numbers of their partners are written in the 'BP1' and 'BP2' columns. Each  $\beta$ -sheet is also given a label (independent of the  $\beta$ -bridge labels) indicated in the adjacent column. The 'ACC' column contains the solvent accessible surface measured in  $\text{\AA}^2$  by estimating the number of water molecules in contact with the present residue. The two strongest backbone-backbone hydrogen bonds are then listed, where 'N-H-->O' are donor hydrogen bonds and 'O-->H-N' acceptor hydrogen bonds. The format indicates the relative sequence position of the hydrogen bond partner followed by the energy in kcal/mol (e.g. '-5,-0.8' means that the partner residue's DSSP number is 5 less than the present one and that the hydrogen bond energy is -0.8 kcal/mol). The DSSP output also contains the  $C_\alpha$  coordinates, phi/psi angles and other angles which were left out in this figure due to limitations of space.

### **DSSPcont: Continuous DSSP assignment reflecting protein motion**

The aim of the continuous assignment was to reflect the structural variability due to thermal motion in a way so that regions which do not vary upon thermal motion have crisp assignments (almost discrete values), while regions which undergo thermal motion should reflect this in their continuous assignment and ideally reflect the occupancy of each secondary structure state. We estimated this by following the energy based secondary structure assignment of DSSP and letting the strength of the hydrogen bond reflect thermal motion in the assignment (Andersen et al., 2001). This concept led us to develop a continuous extension of DSSP. This continuous assignment is based upon multiple runs of DSSP with different hydrogen bond thresholds. Then, we compile a weighted

average over the individual DSSP assignments to assign secondary structure to each residue. We determined the weights by applying the above criterion for 'good' assignments starting with structural homologues from the FSSP (Holm, 1998) database. Inspecting the structural alignments in detail, we noted a number of possible reasons for observed structural differences:

1. Different solution composition, spatial grouping and/or environment of the proteins.
2. Uncertainties/errors in the experimental structure determination setup.
3. Minor thermal fluctuations (even though mostly averaged out).
4. Local amino acid substitutions causing the structural change.
5. Insertions/deletions adjacent to the local stretch in question.
6. Non-local changes forcing a new local conformation.
7. Other less likely causes e.g. prion-like switching.

Our objective was a secondary structure assignment method de-emphasising the effects of 1-3 while capturing differences caused by sequence changes. However, for structural alignments of homologues, we cannot separate these effects as illustrated by a comparison between two related structures: periplasmic binding protein (PDB: 4MBP; Quioco, 1997) and putrescine binding protein (PDB: 1POT; Sugiyama, 1996). The structural alignment was obtained from FSSP with a Z-score of 23.2 and an RMSD of 3.6 Å over 303 residues. We will focus on a small ten-residue segment (Figure 5A) that has spiralling structure ( $\alpha$ -helix,  $3_{10}$ -helix or turn) and a  $\beta$ -bridge at the penultimate position (Figure 5A). Based on the assignment alone one might characterise the differences as problems in the assignment process, since both segments have  $3_{10}$ -helix hydrogen bonds over the entire stretch. On the other hand, 1POT has no  $\alpha$ -helix hydrogen bonds resulting in the assignment of  $3_{10}$ -helix. The results from three high-quality prediction methods (Figure 5B) suggest that the structural differences resulted from the sequence divergence. This means that the secondary structure assignments of the two segments should not necessarily be the same. This line of reasoning can be extended from short helices to short sheets and to the N- or C-terminal ends of helices and strands (caps). Therefore, we chose to optimise the weights for DSSPcont based on the comparisons between different NMR models for the same protein.

---

## A

PDB:4mbp											
AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N			
82	D	<b>S</b>	>>	-	0	0	103	-2,-0.4	4,-1.9	1,-0.1	3,-0.8
83	K	<b>F</b>	3>	S+	0	0	155	1,-0.2	4,-1.5	2,-0.2	-1,-0.1
84	A	<b>I</b>	>4	S+	0	0	63	2,-0.2	3,-0.7	1,-0.2	4,-0.4
85	F	<b>E</b>	X>	S+	0	0	8	-3,-0.8	3,-2.5	1,-0.3	4,-0.6
86	Q	<b>H</b>	><	S+	0	0	65	-4,-1.9	3,-0.9	1,-0.3	-1,-0.3
87	D	<b>T</b>	<<	S+	0	0	95	-4,-1.5	-1,-0.3	-3,-0.7	-2,-0.2
88	K	<b>T</b>	<4	S+	0	0	111	-3,-2.5	217,-2.2	-4,-0.4	218,-0.4
89	L	<b>B</b>	<<	S-G	304	OC	2	-3,-0.9	215,-0.2	-4,-0.6	214,-0.1
90	Y	<b>T</b>	>>	-	0	0	75	213,-1.5	3,-1.4	-2,-0.2	4,-0.8

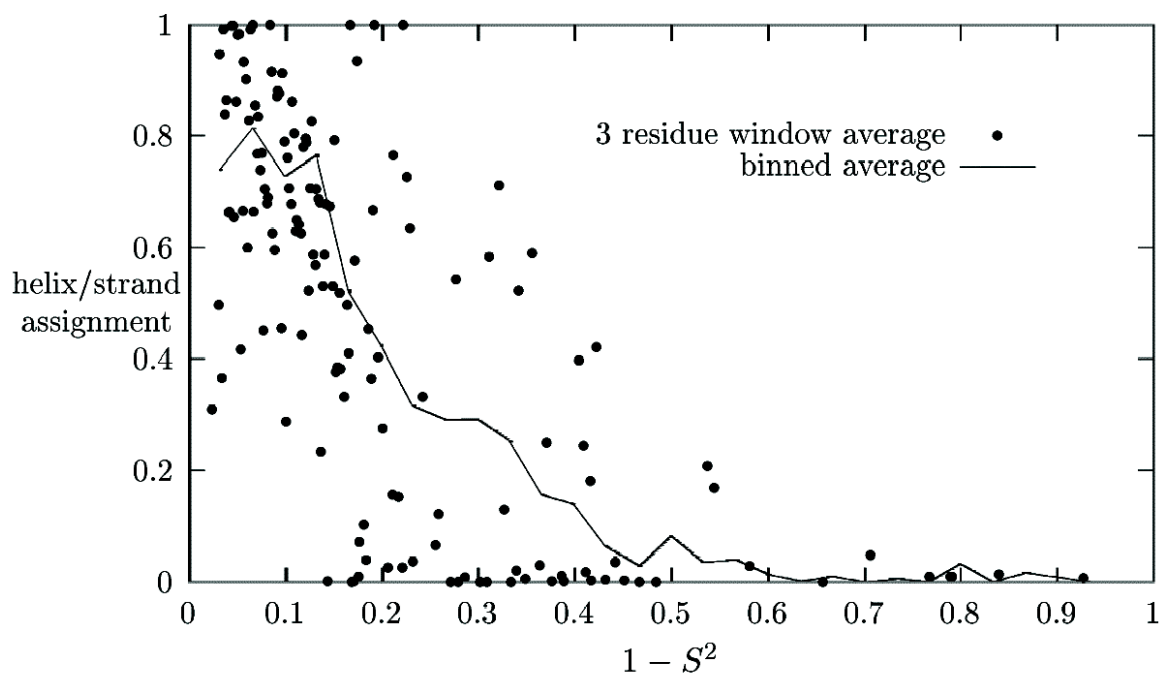
  

PDB:1pot											
AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N			
103	K	<b>S</b>	<	S+	0	0	117	-3,-1.5	2,-0.8	1,-0.2	-1,-0.2
104	L	<b>T</b>	>	+	0	0	2	-3,-0.4	3,-1.5	1,-0.1	-1,-0.2
105	T	<b>T</b>	3	+	0	0	97	-2,-0.8	3,-0.2	1,-0.2	-1,-0.1
106	N	<b>T</b>	>	S+	0	0	38	1,-0.1	3,-2.6	2,-0.1	-1,-0.2
107	F	<b>G</b>	X	+	0	0	34	-3,-1.5	3,-2.0	1,-0.3	-1,-0.1
108	S	<b>C</b>	3	S+	0	0	104	1,-0.3	-1,-0.3	-3,-0.2	-2,-0.1
109	N	<b>G</b>	<	S+	0	0	35	-3,-2.6	189,-1.8	170,-0.1	190,-0.8
110	L	<b>B</b>	<	S-L	272	OE	8	-3,-2.0	187,-0.2	187,-0.2	5,-0.1
111	D	<b>T</b>	>	-	0	0	35	185,-3.0	3,-1.6	-2,-0.2	-1,-0.1

## B

	PSIPred	SSpro	PROF
4mbp	DKAFQDKLY hhhhhhccc	DKAFQDKLY chhhhhccc	DKAFQDKLY chhhhhccc
1pot	KLTFNSNLD cccccccc	KLTFNSNLD cccccccc	KLTFNSNLD cccccccc

**Figure 5: DSSP assignments for similar structures: 4MBP and 1POT.** (A) The DSSP assignment for two segments taken from two structurally similar proteins (periplasmic binding protein 4MBP (Quioco, 1997) and putrescine binding protein 1POT (Sugiyama, 1996)) illustrates that the observed differences between these segments may originate from sequence differences. The boxed letters shown in the column next to the amino acid sequence give the final DSSP assignment: G =  $3_{10}$ -helix, H =  $\alpha$ -helix, T = turn, B =  $\beta$ -bridge, and S = bend. The next column shows the hydrogen bonds (>: hydrogen bond acceptor, <: hydrogen bond donor and X: both), with indications of the hydrogen bond length, i.e.  $i \rightarrow i+(3,4)$  for  $3_{10}$  and  $\alpha$ -helices, respectively. (B) All the predictions from PSIPRED (Jones, 1999), SSpro (Baldi, 1999) and PROFphd (Rost, 1996) (see Chapter 29) correctly spot the  $\alpha$ -helix signal in 4MBP, while missing this signal for 1POT. This may indicate that the altered sequence changed the structure significantly in this region. Here, 'h' refers to the DSSP class helix (H or G) and 'c' to the DSSP non-regular class. Note: the predictions are cut out from those for the entire protein.



**Figure 7: Protein motion and secondary structure.** Using one set of coordinates from an ensemble of NMR models, the continuous DSSP assignment reproduces the segments in proteins that experimentally had a high degree of motion due to thermal fluctuations in water. Protein motion has been independently measured by the order parameter  $1-S^2$ , by the tumbling of the N-H backbone bond-vector.  $1-S^2$  is low when the amino acid is fixed as in the protein core, and it is high when the residue fluctuates.  $1-S^2$  is shown versus the continuous DSSP assignment grouping helices (GHI) and strands (EB). The points are averages over a window segment of three consecutive residues; the line gives an average of helix/strand assignments. Figure reproduced from Andersen et al. (2001)

We found that the single residue RMSD between models of high-quality NMR structures correlated well with thermal fluctuations in water as independently measured by the order parameter. The resulting continuous DSSP assignments were constructed to reflect the differences between NMR models of the same protein, so that the assignments reflect segments with thermal fluctuations (Figure 7). This means that the more a sequence segment fluctuates the lower the probability for the assigned helix/sheet will become. Information of this type can also be obtained directly from crystal structures. Overall, we found that the continuous assignment of secondary structure reflected the average occupancy of secondary structure assignments. In particular, our continuous assignment for a single NMR structure is similar to the average obtained over all models.

### ***STRIDE: H-bond and phi/psi-angle based assignment mimicking experts***

The secondary STRuctural IDentification method (STRIDE) by Frishman and Argos (1995) uses an empirically derived hydrogen bond energy (see SECONDARY STRUCTURE CONCEPTS) and phi-psi torsion angle criteria to assign secondary structure. Torsion angles are given  $\alpha$ -helix and  $\beta$ -sheet propensities according to how close they are to their regions in Ramachandran plots (see Chapter 2 Figure 6) (Ramachandran and Sasisekharan, 1968). The method fixes five internal parameters for  $\alpha$ -helix and four for  $\beta$ -sheets. The parameters are optimised to mirror visual assignments made by crystallographers for a set of proteins. However, crystallographers often disagree in their assignment of secondary structure, which STRIDE aims to even out by averaging over many structures. Since the secondary structure categories have different parameters, their assignment thresholds are independent for the hydrogen bond and phi-psi torsion angles. By

construction, the STRIDE assignments agreed better with the expert assignments than DSSP, at least for the data set used to optimise the free parameters. In particular, the authors reported that every 11th  $\beta$ -sheet and every 32nd  $\alpha$ -helix were more in register with the expert assignments for the data set used.

Like DSSP, STRIDE assigns the shortest  $\alpha$ -helix ('H') if it contains at least two consecutive  $i \rightarrow i+4$  hydrogen bonds. In contrast to DSSP, helices are elongated to comprise one or both edge residues if they have acceptable phi-psi angles, similarly a short helix can be removed if the phi-psi angles are unfavourable. This implies that hydrogen bond patterns may be ignored if the phi-psi angles are unfavourable. The sheet category does not distinguish between parallel and anti-parallel sheets. The minimal sheet ('E') is composed of two residues each in one of five possible hydrogen bond conformations, i.e. two more than for DSSP. The dihedral angles are incorporated into the final assignment criterion as was done for the  $\alpha$ -helix.  $\beta$ -sheet bulges are accepted applying the same criterion as DSSP. Single residue sheets, i.e.  $\beta$ -bridges are labelled as 'B' for the three DSSP hydrogen bond conformations and as 'b' for the remaining two.  $3_{10}$ - ('G'),  $\pi$ -helices ('I') are implemented according to the DSSP scheme, but with the empirical hydrogen bond criterion. Turns are assigned according to the phi-psi angles of residue  $i+1$  and  $i+2$  as described in Wilmot and Thornton (1990). The 'C' symbol is used whenever none of the above structure requirements are met.

```

PDB:1crn
REM |---Residue---| |--Structure--| |-Phi-| |-Psi-| |-Area-| 1CRN
....
ASG VAL - 15 15 H AlphaHelix -69.24 -41.22 93.8 1CRN
ASG CYS - 16 16 H AlphaHelix -56.67 -36.00 18.4 1CRN
ASG ARG - 17 17 H AlphaHelix -77.07 -16.13 94.1 1CRN
ASG LEU - 18 18 H AlphaHelix -53.21 -46.17 143.0 1CRN
ASG PRO - 19 19 C Coil -77.19 -7.60 108.9 1CRN
ASG GLY - 20 20 C Coil 106.26 7.31 52.1 1CRN
ASG THR - 21 21 C Coil -52.67 136.34 38.4 1CRN
ASG PRO - 22 22 C Coil -56.98 146.62 81.9 1CRN
ASG GLU - 23 23 H AlphaHelix -56.41 -36.19 68.9 1CRN
ASG ALA - 24 24 H AlphaHelix -63.43 -34.86 61.3 1CRN
ASG ILE - 25 25 H AlphaHelix -74.77 -37.89 98.2 1CRN
ASG CYS - 26 26 H AlphaHelix -64.95 -31.69 0.0 1CRN
ASG ALA - 27 27 H AlphaHelix -62.04 -54.03 11.6 1CRN
ASG THR - 28 28 H AlphaHelix -68.78 -25.49 121.1 1CRN
ASG TYR - 29 29 H AlphaHelix -67.59 -36.30 174.0 1CRN
ASG THR - 30 30 H AlphaHelix -108.96 -18.47 23.4 1CRN
ASG GLY - 31 31 C Coil 91.82 -3.07 36.1 1CRN
ASG CYS - 32 32 C Coil -69.52 164.38 4.6 1CRN
ASG ILE - 33 33 E Strand -129.76 157.03 51.0 1CRN
ASG ILE - 34 34 E Strand -111.56 129.59 78.0 1CRN
....

```

Figure 8: Explanation of STRIDE output. The STRIDE output for Crambin is shown to explain the format and for comparison to Figure 5. The format is simple and easily parsed, with 'ASG' as the first word in the lines used for assignment. The residue columns comprise the three-letter amino acid code, the chain identifier ('-' for single chains), the PDB residue number and the STRIDE residue number, which starts from one for every new chain. The two structure columns contain the one-letter structure assignments (HGIEBbTC) and its short description. The columns with phi and psi are followed by the column with solvent accessibility (measured in  $\text{\AA}^2$ )

### **STICK: continuous assignment based on line segments**

The standard method used to define line segments is to fit an axis through each secondary structure element (e.g. DEFINE). This approach has difficulties, both with inconsistent definitions of secondary structure and the problem of fitting a single straight line to a bent structure. STICK avoids these problems by finding a set of line segments independently of any external secondary

structure definition (Taylor, 2001). This allows the segments to be used as a novel basis for secondary structure definition by taking the average rise/residue along each axis to characterise the segment. This practice has the advantage that secondary structures are described by a single (continuous) value that is not restricted to the conventional classes of  $\alpha$ -helix,  $3_{10}$ -helix, and  $\beta$ -strand. This latter property allows structures without "classic" secondary structures to be encoded as line segments that can be used in comparison algorithms. When compared over a large number of pairs of homologous proteins, the current method was found to be slightly more consistent than a widely used method based on hydrogen bonds.

---

### ***Beta-Spider: packing energy assignment***

The stabilizing factors maintaining the secondary structure were used in  $\beta$ -spider (Parisien, 2005) as the primary parameters for assignment. This was done by calculating the packing energy of the backbone interactions in the form of Coulomb electrostatic and van der Waals forces (see SECONDARY STRUCTURE CONCEPTS). For possible  $\beta$ -sheet interacting strands the packing energy was calculated for tri-peptide pairs and defined as 'favored' if at least -5 kcal/mol (-1.67 kcal/mol per residue). This may seem as more than three times the interaction energy required by DSSP H-bonds (-0.5 kcal/mol per residue), but the number of possible polar atom interactions per residue has also tripled (two H-bond donors/acceptors and one dipole), and furthermore the van der Waals energy is included. Two geometrical considerations are also used, which require that the sheet residue pairs are maximally 6 Å apart and that the torsion  $\left\langle \begin{matrix} \rightarrow & \rightarrow & \rightarrow & \rightarrow \\ C_{\beta i}, C_{\alpha i}, C_{\alpha j}, C_{\beta j} \end{matrix} \right\rangle \leq 90^\circ$ . So, if  $C_{\alpha i}$  and

$C_{\alpha j}$  are aligned, the projected angle spanned between the two  $C_{\beta}$  must be less than 90 degrees, i.e. point either towards each other or away from each other (Glycine is exempted).  $\beta$ -bulges up to three residues in length are also allowed within the sheet. In summary a  $\beta$ -spider sheet must contain at least one energetically 'favored' tri-peptide pair and start/end with a residue pair following the geometrical requirements and be at least two residues long. This setup increases the set of tri-peptide H-bonding motifs allowed within a  $\beta$ -sheet, resulting in approximately 11 and 6 percent increase in number of motif matches for parallel and anti-parallel sheets as compared to DSSP, respectively. At present only a  $\beta$ -sheet assignment scheme has been published.

---

### ***XTLsstr: circular dichroism driven assignment***

The circular dichroism (CD) of a protein in the far ultraviolet range is determined principally by amide-amide interactions of the backbone (King, 1999), which has driven the authors to develop a protein secondary structure assignment scheme: XTLsstr. The program calculates two backbone dihedral angles and three distances (two of which are simple two atom H-bond distances), which are used for the assignment. By visual inspection the authors have developed range definitions for each secondary structure ( $3_{10}$ -,  $\alpha$ -helices,  $\beta$ -sheets and two turn types) that would be consistent with the amide-amide interactions observed in CD.

---

### ***VoTAP: residue contact assignment based on geometry***

By first dividing the protein volume into residue specific polyhedra using Voronoï tessellation a residue contact map is automatically defined (Dupuis, 2004) (see SECONDARY STRUCTURE CONCEPTS). This contact map is unique and does not depend on a distance threshold, here two residues are in contact if they share one face of their polyhedra. The contacts were divided into strong and normal contacts based on the contact surface area size in a residue specific manner, thus

yielding three residue-residue contact designations 0,1 and 2 for no, normal and strong contact, respectively. VoTAP assigns residues into the three overall classes helix ( $3_{10}$ ,  $\alpha$  and  $\pi$ ), sheet (anti-parallel, parallel and  $\beta$ -bridge) and coil. The assignment was performed by fitting contact pattern statistics to the consensus assignment of DSSP, PSEA, DEFINE and STRIDE in two steps. The first step focuses on the diagonal of the contact matrix by comparing residue contact quintuplets. A lookup table for each quintuplet was created for the consensus assignment and a probability for each secondary structure class stored. A secondary structure class probability is subsequently assigned to each residue using a sliding window. In the second step the off-diagonal contacts are used to assign sheet residues if they follow the parallel or anti-parallel  $\beta$ -plated sheet pattern. By VoTAP helices and sheets are constrained to be at least 3 residues in length.

---

### ***DEFINE: Idealized $C_{\alpha}$ -distance mask based assignment***

The algorithm DEFINE (Richards and Kundrot, 1988) assigns secondary structures by matching  $C_{\alpha}$ -coordinates with a linear distance mask of the ideal secondary structures. First, strict matches are found, which subsequently are elongated and/or joined allowing moderate irregularities or curvature. The algorithm locates the starts and ends of  $\alpha$ - and  $3_{10}$ -helices,  $\beta$ -sheets, sharp turns and omega-loops. With these classifications, the authors are able to assign 90-95% of all residues to at least one of the given secondary structure classes.

To assign  $\alpha$ -helices the linear mask is matched with each row in the distance matrix of the query protein and the root-mean-square difference between the distances in the mask and the ones observed in the query protein is calculated as a measure of cumulative discrepancy. If a segment longer than four residues matches the mask within the allowed cumulative discrepancy limit (default  $\epsilon = 1 \text{ \AA}$ ), an  $\alpha$ -helix is assigned to the segment. Assigned  $\alpha$ -helices are checked whether they start or end with a  $3_{10}$ -helix, but individual  $3_{10}$ -helices and  $\pi$ -helices are not investigated.

In order to assign  $\beta$ -sheets as a single category, the authors have applied a linear distance mask taken from ideal anti-parallel sheets. The problems associated with backbone bendability inside sheets and curvatures in larger sheets have been 'solved' by excluding non-rigid sheets from the definition. The minimum length of sheets is set to be four residues. According to Pauling's definition of a  $\beta$ -sheet, each strand must pair to another strand to form a sheet. In contrast, DEFINE may assign unpaired strands.

---

### ***P-Curve: Idealized protein curvature based assignment***

Sklenar, Etchebest and Lavery (1989) based their assignment scheme P-Curve on a mathematical analysis of protein curvature. Using differential geometry, they calculated a helicoidal axis on the basis of the fixed axis systems of a series of peptide planes. The secondary structure assignments are performed by motif matching, where the parameters in the motif are the radius of the helicoidal system along with a series of tilting, rolling and twisting measures describing geometrical differences between two peptide planes. This parameter analysis is achieved mainly by the use of the  $C_{\alpha}$ -coordinates. The P-Curve assignment differs significantly from those performed from phi/psi angles or hydrogen bonds, since different parameters are used (e.g. helicoidal radius, tilting, rolling, twisting). Furthermore, the degrees of freedom allowed when matching a P-Curve motif are quite different from those allowed when matching a DEFINE linear distance mask. For example, while the linear distance mask of DEFINE fits poorly to a curved  $\beta$ -strand, the local P-Curve parameters are likely to fit better. The assigned secondary structures are recognised by matching known structural motifs. These motifs are based on idealized values for helicoidal parameters. The following motifs are used: right- and left-handed  $\alpha$ -helix,  $3_{10}$ - and  $\pi$ -helix, parallel and anti-parallel

$\beta$ -sheets and some other structures of little interest here. Note that like DEFINE, P-Curve may assign the category sheet to unpaired strands.

### ***PALSSE: Linear element assignment for structure comparison***

With the objective to describe, in a vector form, the two major classes of secondary structure for structure comparison PALSSE performs a three class assignment based on  $C_\alpha$ -coordinates (Majumdar, 2005). The helix class assigned includes  $\alpha$ -helices,  $3_{10}$ -helices,  $\pi$ -helices and turns that show a helical propensity in view of the observation that many  $\alpha$ -helices start and end with tighter ( $3_{10}$ ), looser ( $\pi$ ) or non-backbone H-bonded turns/helices. Similarly the  $\beta$ -strand class used includes parallel-sheets, anti-parallel-sheets,  $\beta$ -bridges,  $\beta$ -bends and  $\beta$ -hairpins. This results in many regular structure assignments where approximately 80% of residues are reported in the helix or sheet classes. The high coverage is important for structure comparison and similarity searches where the secondary structure elements are represented as vectors, thus allowing a higher degree of differentiation between proteins.

### ***KAKSI: $C_\alpha$ and phi/psi based assignment mimicking experts***

The KAKSI assignment has been designed to best fit the secondary structure assignments done by experts in the PDB file header (Martin, 2005). This is done by defining allowed  $C_\alpha$  distance measures and phi/psi angle values using a single sliding window for helices and two sliding windows for  $\beta$ -sheets to ensure partnering strands in the  $\beta$ -sheet. First helices are assigned followed by  $\beta$ -sheet assignment on the remaining residues, with minimal lengths of 5 and 3 residues, respectively. When comparing the KAKSI helix class to DSSP and STRIDE the authors map  $3_{10}$ -,  $\alpha$ - and  $\pi$ -helices into one helix class and  $\beta$ -sheets and  $\beta$ -bridges into one sheet class.

### ***Other secondary structure assignment methods***

P-SEA (Labesse, 1997) assigns helices and  $\beta$ -strands using only the  $C_\alpha$ -coordinates. This is primarily done using a short range  $C_\alpha$  distance mask ( $i \rightarrow (i+2, i+3, i+4)$ ) and two angle criteria for each secondary structure. The helix class assigned covers  $\alpha$ -,  $3_{10}$ - and  $\pi$ -helices, and the strand class covers parallel and anti-parallel  $\beta$ -strands, with minimal lengths of 5 and 3 residues, respectively.

SEGNO (Cubellis, 2005) performs its assignments based on  $C_\alpha$ -coordinates, phi/psi angles and an angle-distance hydrogen bond. For helix assignment the  $C_\alpha$ -atoms must primarily reside within an imaginary cylinder helix, inspired from Richardson and Richardson (1988). The axis is defined by the mean of a sliding window of four  $C_\alpha$ -atoms and the cylinder radius is 1.7-3Å. The  $\beta$ -strand assignment is based on favourable phi/psi angles of at least three residues, and strands are associated into sheets using the angle-distance hydrogen bond. The authors report that this gives a stronger amino acid trend at the helix caps and that it improves secondary structure guided sequence alignments. At present the method is not available so it wasn't possible to compare it directly to the other methods presented, but its web site is reported in Table 1.

SECSTR (Fodje, 2002) was developed to identify and study the rare  $\pi$ -helices. It uses a DSSP like hydrogen bond definition and a Pauling  $i \rightarrow i+5$  hydrogen bond  $\pi$ -helix assignment scheme requiring at least two consecutive bonds. Approximately 10 times more  $\pi$ -helices were found compared to DSSP by giving priority to the strongest hydrogen bond instead of giving priority to  $\alpha$ -helices and thus  $i \rightarrow i+4$  bonds in the assignment. This amounted to 104 overlooked  $\pi$ -helices extracted from a non-homologous set of high quality X-ray structures, which were verified by manual inspection.

Local protein structure analyses investigating frequently reoccurring small segments of the polypeptide chain also describe protein structure at the same scale as secondary structure and is being used for structure prediction by ab-initio methods (see Chapter 32), covered nicely in a recent review (Offmann, 2007).

---

## Availability of secondary structure assignment programs

Table 1

Program	Internet
DSSP	<a href="http://www.cmbi.kun.nl/gv/dssp">http://www.cmbi.kun.nl/gv/dssp</a>
STRIDE	<a href="http://webclu.bio.wzw.tum.de/stride/">http://webclu.bio.wzw.tum.de/stride/</a>
DSSPcont	<a href="http://cubic.bioc.columbia.edu/services/DSSPcont">http://cubic.bioc.columbia.edu/services/DSSPcont</a>
VoTAP	<a href="http://www.lmcp.jussieu.fr/%7Emornon/voronoi.html">http://www.lmcp.jussieu.fr/%7Emornon/voronoi.html</a>
Beta-spider	<a href="http://www-lbit.iro.umontreal.ca/bSpider/">http://www-lbit.iro.umontreal.ca/bSpider/</a>
XTLsstr	<a href="http://oregonstate.edu/dept/biochem/faculty/johnson.html">http://oregonstate.edu/dept/biochem/faculty/johnson.html</a>
KAKSI	<a href="http://migale.jouy.inra.fr/mig/mig_fr/servlog/kaksi/">http://migale.jouy.inra.fr/mig/mig_fr/servlog/kaksi/</a>
PALSSE	<a href="http://prodata.swmed.edu/palsse/palsse.php">http://prodata.swmed.edu/palsse/palsse.php</a>
SEGNO	<a href="http://www.bioinf.man.ac.uk/~lovell/segno.shtml">http://www.bioinf.man.ac.uk/~lovell/segno.shtml</a>
SecStr	<a href="http://www.mbfys.lu.se/Services/SecStr/">http://www.mbfys.lu.se/Services/SecStr/</a>

## Secondary structure statistics and comparison

### Secondary structure frequency and length

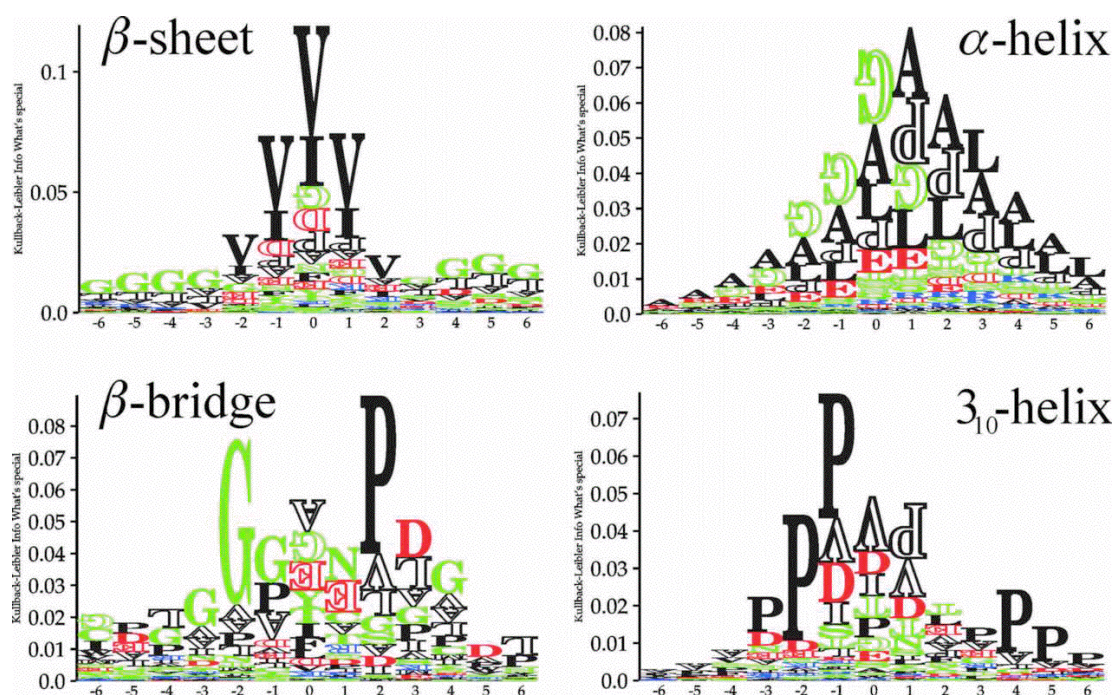
The regular secondary structures (helices and sheets), as defined by DSSP, comprise a bit more than half the protein residues (see Table 3), where the  $\alpha$ -helix is the most abundant with 31.3% followed by anti-parallel  $\beta$ -sheets with 15.7% and parallel  $\beta$ -sheets with 5.7%. The remaining residues thus comprise a bit less than 50% (depending on the definition of regular structure used (see below)), which introduces a natural skewness of relevance within secondary structure prediction. The average lengths of the  $\alpha$ -helix and  $\beta$ -sheet are 11.2 and 4.4 residues, respectively, which in terms of physical length interestingly enough are quite similar 16.8 Å and 14-15 Å, respectively (using physical distances for an  $\alpha$ -helix of 1.5 Å/res. (Branden and Tooze, 1999) and residue span in fully extended strands is 3.2 Å/res. in parallel  $\beta$ -sheets and 3.4 Å/res. in anti-parallel beta sheets (Creighton 1993)).

Table 3: DSSP secondary structure statistics from a set of 707 non-homologous protein chains. (a) The following DSSP residues were counted: sheet H-bonded, middle and single bulge. (b) The sum of the parallel and anti-parallel sheet frequencies is higher than the total  $\beta$ -sheet residues, since a residue may be in two sheets of different type. (c) The average length of sheets reported is different from the average length of connected 'E' stretches (5.2 residues), since overlapping sheets are counted as one.

Secondary structure statistics using DSSP assignments				
	$\alpha$ -helix: 'H'	$\beta$ -sheet: 'E'	par. $\beta$ -sheet: 'E' (a)	anti-par. $\beta$ -sheet: 'E' (a)
Frequency	<b>31.3%</b>	<b>20.4%</b> (b)	5.7%	15.7%
Average length	11.2 residues	4.4 residues (c)	4.0 residues	4.6 residues
	$3_{10}$ -helix: 'G'	$\pi$ -helix: 'I'	$\beta$ -bridge: 'B'	other: 'C', 'T', 'S'
Frequency	<b>3.7%</b>	<b>0.04%</b>	<b>1.3%</b>	<b>43.3%</b>
Average length	3.4 residues	5.2 residues	1 res. per definition	6.8 residues

## Residue distributions for secondary structure

The amino acids typically found in  $\alpha$ -helices differ considerably from those found in  $\beta$ -sheets (Figure 9). Alanine and Leucine often occur in  $\alpha$ -helices, while Proline and Glycine are rare. In  $\beta$ -sheets Valine and Isoleucine are over-represented, while Glycine, Aspartic acid and Proline are under-represented. Shorter structures such as  $3_{10}$ -helices and  $\beta$ -bridges have distinct residue distributions. For  $3_{10}$ -helices the Alanine and Leucine signal has disappeared, instead the sequences are dominated by Proline which often is observed as a helix initiator and breaker. For  $\beta$ -bridges, we no longer find a preference for Valine and Isoleucine. This indicates the role of the side chain in defining secondary and tertiary structure. In general, these preferences have long been the basis of secondary structure prediction methods.

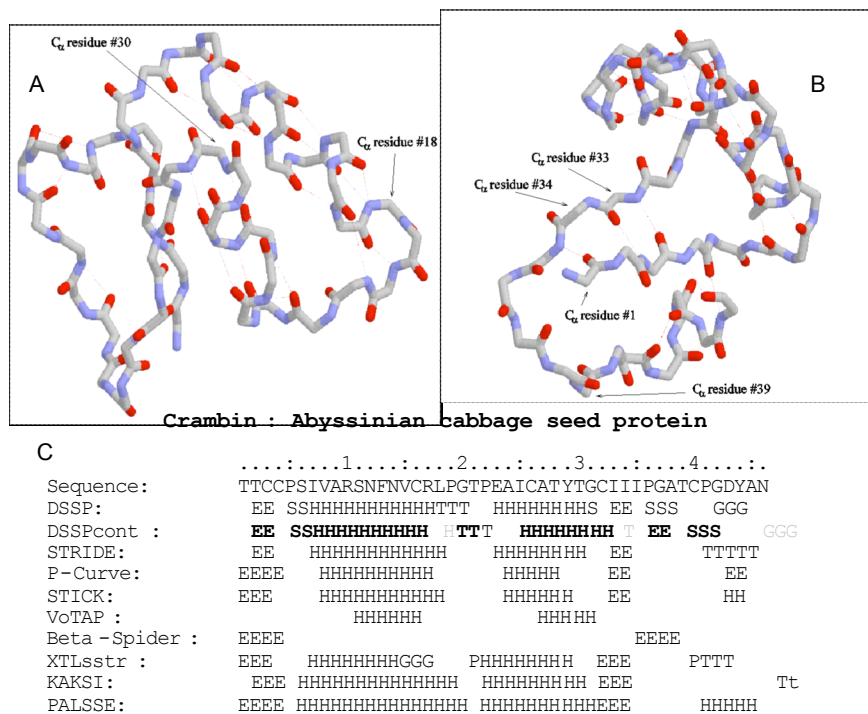


**Figure 9: Sequence distributions for secondary structure.** The four graphs show alignment statistics for  $\beta$ -sheets,  $\alpha$ -helices,  $\beta$ -bridges and  $3_{10}$ -helices, by the Kullback-Leibler information at positions surrounding the one assigned (position 0). The number of aligned segments are: ( $\alpha$ -helix) 41803, ( $3_{10}$ -helix) 4952, ( $\beta$ -sheets) 27320, ( $\beta$ -bridges) 1851. These segments were retrieved from a data set of 707 non-homologous protein chains using the DSSP assignment. At a given position, we therefore observed the 20 amino acids with a certain frequency; the Kullback-Leibler information calculates the information content of the observed frequencies with respect to the background frequencies (irrespectively of the structure). The more an observed set of frequencies differs from the background, the higher the respective letter. If an amino acid at a given position is observed less frequently than in the background, it is drawn upside-down and hollow.

## Comparison

The secondary structure assignment schemes delineated above have each been designed using one or more aspects of protein structure and with one or more applications in mind. Good quality assignments applied within CD spectroscopy analyses may not necessarily be optimal when applied within structure comparison or vice versa. It is therefore inherently difficult to perform a qualitative comparison between protein secondary structure assignments, so the main focus of the comparison presented will be quantitative. In essence one should choose the secondary structure assignment scheme which is most consistent with the investigation where it is applied.

We used the simple structure of Crambin as an example to point out differences in the assignment schemes (Figure 12, note that the P-Curve assignment was taken from the original publication (Sklenar and Etchebest, 1989)). When comparing the three class assignments ( $\alpha$ -helix,  $\beta$ -sheet and other), STRIDE and DSSP are identical except for one residue at the C-cap of an  $\alpha$ -helix, whose last H-bond is weak as reported by DSSPcont (reported in the figure as a grey letter). This assignment was confirmed by XTLsstr and KAKSI with some variation on cap assignments. P-Curve, STICK and PALSSE also assigned the same elements with capping variations, but report an additional helix or sheet towards the protein's C-term. Looking at the sheet region in detail (Figure 9B), we see that the residues 39 and 40, assigned sheet by P-Curve only, are distant from any residue on a putatively pairing strand. According to Pauling, such an assignment would not be valid. The two sheets have been extended by beta-spider showing that high backbone-backbone packing energy not caught by standard H-bonds is keeping the strands together. Longer sheets were also identified by XTLsstr, KAKSI, PALSSE and partially by STICK and P-Curve indicating that the backbone conformation is extended also for these amino acids, so the extension of the two DSSP/STRIDE sheets appears reasonable. VoTAP identifies the two primary  $\alpha$ -helices, but is the only method which, for this particular protein, does not assign the two sheets observed by other methods.



**Figure 12: Protein secondary structure for Crambin.** The structure of the small protein Crambin (Teeter, 1984, PDB:1CRN) is shown from two angles: (A) the image of the two helices, and (B) the central short sheet. (C) The overall assignment of secondary structure is shown for the different assignment methods, where the overall elements were found to agree between most methods shown. Please note that beta-spider only assigns sheets and the DSSPcont assignments have been discretized to crisp (100%:bold or space), high (>90%:normal), mixt (>50%:grey) variants of the DSSP assignment. Each assignment method was used with its default/standard settings supplied.

Are the discrepancies observed for Crambin representative? Secondary structure capping assignment do indeed constitute the major differences between methods as the exact specification of where a regular secondary structure ends is not well defined as reported by Colloc'h et al. (1993) for DSSP, P-Curve and DEFINE. Investigating DSSP in particular we also found variability of where the helix or sheet starts and ends between high quality NMR models for the same protein, where the NMR model variability was found to correlate with inherent protein motion (Andersen,

2002). Helix and sheet core segments are also the assignments which correlate best with circular dichroism spectra (Sreerama, 1999).

Martin et al. (2005) have performed a comparison of some of the methods described above on a high quality X-ray data set (resolution  $< 1.7\text{\AA}$  and R-factor  $< 0.19$ ) containing 689 protein structures. They used a comparison score called  $C_3$  which compares two assignments as the number of identical regular secondary structure residue assignments relative to the total number of residues assigned to a regular structure, where the helix and sheet class assignments were compared. They find that DSSP and STRIDE are very similar ( $C_3=95\%$ ) followed by the expert assignments in PDB ( $C_3 > 87\%$ ). KAKSI is then the nearest neighbour to the group with  $C_3$  in the 81-84% range followed by PSEA and XTLsstr in the 78-81% range. XTLsstr differs the most from the other assignment methods with a  $C_3$  score down to 76% when compared to PSEA. Similar to what Colloch et al. (1993) has reported for DSSP, Martin et al. (2005) found that STRIDE assigns many short helices.

Another observation from the Crambin example is that the methods do not mix helix and sheet assignments on the same residue. This has generally been observed to hold true for DSSP, P-Curve and DEFINE, where only 0.32% of the residues showed conflicts between helix and sheet assignments (Colloch, 1993).

---

## APPLICATIONS OF SECONDARY STRUCTURE

Secondary structure is being used in many areas of structural bioinformatics, from structure visualization, classification and comparison to predictions, all covered in this book. Here we will delineate some areas where secondary structure is applied in investigations within different contexts, for example in understanding biological processes and disease.

### Secondary structure and protein function

Circular Dichroism (CD) spectroscopy is often used to measure differences in protein secondary structure contents under different conditions such as mutated residues or changing environments. An example is the thimet metallo endo-peptidase (THOP1) which is known to be modulated by changes in calcium concentration. To study how this change comes about, key Aspartic acid residues believed to bind calcium have been mutated, and the calcium induced change in  $\alpha$ -helix content studied by CD (Oliviera, 2005). See also Chapter 21 describing functional inference from structure.

### Secondary structure and disease

Protein aggregation plays an important part in several well known diseases such as Alzheimer's, Parkinson's, Huntington's and prion disease (Kajava, 2006). The internal structure of amyloid fibrils in Alzheimer's disease and type 2 diabetes is a ladder of  $\beta$ -sheet structure arranged in a cross- $\beta$  conformation (Stromer, 2005). Whether e.g. the amyloid- $\beta$  plaques are causing Alzheimer's disease (AD) or are mere agglomerates of excess amyloid- $\beta$  is debated (Watson, 2005), but anti-aggregates breaking the  $\beta$ -sheet formation are investigated to prevent AD (Rzepecki, 2004; Chacon, 2004). The formation of amyloid aggregation has been studied in close detail by Cerda-Costa et al. (2007), who found that the C-terminus of the molecule (comprising the last and edge  $\beta$ -strand) is the major contributor to amyloid fibril formation. 3D structure analysis revealed the stability of amyloid fibrils, their self-seeding characteristic and their tendency to form polymorphic structures. (Nelson, 2005).

A conserved N-capping box has been found to be important for the structural autonomy of the prion  $\alpha$ -helix, where the disease associated D202N mutation destabilizes the helical conformation (Gallo, 2005).

---

### Secondary structure mimicking compounds

To develop a drug that can inhibit protein-protein interactions, compounds are being specifically synthesized to mimic helices and strands (Antuch, 2006;Kutzki, 2002;Song, 2001). Antuch et al. report  $\alpha$ -helix mimetic compounds disrupting the Bcl-w/Bak protein-protein interaction, which is important in cancer (Wagner, 2005). Song et al. describe non-peptidic  $\beta$ -strand mimetic compounds that inhibit the HIV-1 protease dimerization necessary for its enzymatic activity. Helices represent one of the most common recognition motifs in proteins (Che, 2007), therefore characterization and analysis of the secondary structures involved in protein-protein interactions becomes important. See also Chapter 27 for a description of protein ligand design.

### Secondary structure and alignment

Taking local structural characteristics into account improves the detection of remote homologs and the alignment quality (Wallqvist, 2000; Shi, 2001; Qiu, 2006). In FUGUE, Shi et al. take the secondary structure, solvent accessibility, and hydrogen bonding status into account when aligning structures and sequences, where the alignment gap penalties are dependent on secondary structure and its conservation. SSALN, a similar method by Qiu and Elber (2006), was found to outperform CLUSTALW and GenThreader using the Fisher's benchmark.

### Secondary structure capping

Helix capping has recently been studied in greater detail, for example the capping dynamics of a glycine  $\alpha_L$  C-capping motif has been studied in detail (Bang, 2006). Bang et al., using chemical synthesis, X-ray crystallography and thermodynamic data, determined that local conformational strain is responsible for most of the energy penalty in an  $\alpha_L$  C-capping motif (Rose, 2006). Helix caps are often stabilized by H-bonds other than the ones used for assignment (described above), as shown by Manikandan and Ramakumar (2004) for the C-H $\cdots$ O H-bond. The donor carbon studied was C $_{\alpha}$ , C $_{\beta}$ , C $_{\gamma}$ , C $_{\delta}$  or C $_{\epsilon}$ , depending on the residue, and the acceptor was the backbone C'=O oxygen. The  $\alpha$ -helices were assigned using PROMOTIF (Hutchinson, 1996) and their ends defined using the Aurora-Rose nomenclature (Aurora, 1998) with N-terminal residues as Ncap to N5 and C-terminal residues as C5 to C'''. They find on average 1.5 and 3.9 C-H $\cdots$ O H-bonds per N- and C-term, respectively, thus indicating their relevance in stabilizing  $\alpha$ -helix caps.

---

## CONCLUSION

Assigning secondary structure from 3D coordinates is an important problem. Many successful solutions have been proposed over the past 25 years. One of the first automated assignment schemes was DSSP, which has become the standard in the field, followed by STRIDE. In fact, secondary structure assignment may be one of the exceptional examples of tools in structural biology and bioinformatics that have not been revolutionised by the explosion of data. For most residues, most of the available methods agree in their assignment. Methods tend to differ mainly in locating the caps of regular secondary structure segments and in distinguishing between more subtle differences (e.g.  $\alpha$ -,  $3_{10}$ -, or  $\pi$ -helix). These differences reflect the aspects used for the assignment and the application(s) for which the method was designed. There are furthermore indications that some fuzziness in the caps is inherent to protein motion (observed by NMR and CD) and that structural homologs also tend to differ more frequently there. Since the applications of secondary structure vary, their assignment schemes will as well, due to the differences in objectives. Optimization for structure comparison will aim at a high regular structure content, while optimization for prediction will aim at clear capping signals and optimization for spectroscopists will aim at reflecting the experimental readout, so the story will continue.

---

## Abbreviations used

<b>3D</b>	three-dimensional
<b>CD</b>	Circular dichroism
<b>DEFINE</b>	method assigning secondary structure from 3D co-ordinates based on linear distance masks of ideal secondary structure (Richards, 1988)
<b>DSSP</b>	program and database assigning secondary structure and solvent accessibility for proteins of known 3D structure from hydrogen bonding patterns (Kabsch and Sander, 1983)
<b>DSSPcont</b>	continuous assignment of secondary structure for proteins of known 3D structure (Andersen, 2002)
<b>H-bond</b>	Hydrogen bond
<b>NMR</b>	nuclear magnetic resonance
<b>P-Curve</b>	curvature based assignment of secondary structure from 3D (Sklenar, 1989)
<b>PDB</b>	Protein Data Bank of experimentally determined 3D structures of proteins (Berman, 2000)
<b>RMSD</b>	root-mean square deviation
<b>STRIDE</b>	secondary STRuctural IDentification method to assign secondary structure from 3D using hydrogen bonds and torsion angles (Frishman and Argos, 1995)
<b>VoTAP</b>	Voronoi tessellation based protein secondary structure assignment (Dupuis, 2004)

---

## Acknowledgements

Thanks to Jinfeng Liu (CUBIC, Columbia) for computer assistance and Jenny Gu for helpful comments on the manuscript. The work of BR was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

## References

- Andersen, C. A. (2001). "Protein Structure and the Diversity of Hydrogen Bonds." The Technical University of Denmark, Ph.D. thesis.
- Andersen, C. A., A. G. Palmer, et al. (2002). "Continuum secondary structure captures protein flexibility." Structure **10**(2): 175-84.
- Antuch, W., S. Menon, et al. (2006). "Design and modular parallel synthesis of a MCR derived alpha-helix mimetic protein-protein interaction inhibitor scaffold." Bioorg Med Chem Lett **16**(6): 1740-3.
- Aurora, R. and G. D. Rose (1998). "Helix capping." Protein Sci **7**(1): 21-38.
- Baker, E. N. and R. E. Hubbard (1984). "Hydrogen bonding in globular proteins." Prog Biophys Mol Biol **44**(2): 97-179.
- Baldi, P., S. Brunak, et al. (1999). "Exploiting the past and the future in protein secondary structure prediction." Bioinformatics **15**(11): 937-46.
- Bang, D., A. V. Gribenko, et al. (2006). "Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis." Nat Chem Biol **2**(3): 139-43.
- Barton, G. J. (1995). "Protein secondary structure prediction." Curr Opin Struct Biol **5**(3): 372-6.

- Beynon, R. J. (2004). "Sequential exoproteolysis as a structural probe: a cautionary note." J Mass Spectrom **39**(2): 188-92.
- Bhat, T. N., P. Bourne, et al. (2001). "The PDB data uniformity project." Nucleic Acids Res **29**(1): 214-8.
- Blundell, T. L. and K. Mizuguchi (2000). "Structural genomics: an overview." Prog Biophys Mol Biol **73**(5): 289-95.
- Boobbyer, D. N., P. J. Goodford, et al. (1989). "New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure." J Med Chem **32**(5): 1083-94.
- Bordo, D. and P. Argos (1994). "The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins." J Mol Biol **243**(3): 504-19.
- Branden, C. a. T., J (1991). Introduction to Protein Structure. New York, Garland Publishing.
- Burley, S. K., S. C. Almo, et al. (1999). "Structural genomics: beyond the human genome project." Nat Genet **23**(2): 151-7.
- Bystroff, C. and D. Baker (1998). "Prediction of local structure in proteins using a library of sequence-structure motifs." J Mol Biol **281**(3): 565-77.
- Cerda-Costa, N., A. Esteras-Chopo, et al. (2007). "Early kinetics of amyloid fibril formation reveals conformational reorganisation of initial aggregates." J Mol Biol **366**(4): 1351-63.
- Chacon, M. A., M. I. Barria, et al. (2004). "Beta-sheet breaker peptide prevents Abeta-induced spatial memory impairments with partial reduction of amyloid deposits." Mol Psychiatry **9**(10): 953-61.
- Che, Y., B. R. Brooks, et al. (2007). "Protein recognition motifs: design of peptidomimetics of helix surfaces." Biopolymers **86**(4): 288-97.
- Colloc'h, N., C. Etchebest, et al. (1993). "Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment." Protein Eng **6**(4): 377-82.
- Cornell, W. C., P., Bayly, C.I., Gould, I.R., Merz Jr, K.M., Ferguson, D. M, Spellmeyer, and F. D. C., T., Caldwell, J.W. and Kollman, P.A (1995). "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules." J. Am. Chem. Soc. **117**: 5179-5197.
- Creighton, T. (1993). Protein: Structures and Molecular Properties. New York, W.H. Freeman.
- Cubellis, M. V., F. Cailliez, et al. (2005). "Secondary structure assignment that accurately reflects physical and evolutionary characteristics." BMC Bioinformatics **6 Suppl 4**: S8.
- Cuff, J. A. and G. J. Barton (1999). "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction." Proteins **34**(4): 508-19.
- de la Cruz, X. and J. M. Thornton (1999). "Factors limiting the performance of prediction-based fold recognition methods." Protein Sci **8**(4): 750-9.
- Decatur, S. M. (2000). "IR spectroscopy of isotope-labeled helical peptides: probing the effect of N-acetylation on helix stability." Biopolymers **54**(3): 180-5.
- Di Francesco, V., P. J. Munson, et al. (1999). "FORESST: fold recognition from secondary structure predictions of proteins." Bioinformatics **15**(2): 131-40.
- Doerr, A. (2007). "Ultrafast spectroscopy: timing is everything." Nat Methods **4**(2): 111.
- Dupuis, F., J. F. Sadoc, et al. (2005). "Voro3D: 3D Voronoi tessellations applied to protein structures." Bioinformatics **21**(8): 1715-6.
- Dupuis, F., J. F. Sadoc, et al. (2004). "Protein secondary structure assignment through Voronoi tessellation." Proteins **55**(3): 519-28.
- Fain, B. and M. Levitt (2001). "A novel method for sampling alpha-helical protein backbones." J Mol Biol **305**(2): 191-201.
- Fasman, G. (1989). The development of the prediction of protein structure. New York, Plenum Press.

- Fesinmeyer, R. M., E. S. Peterson, et al. (2005). "Studies of helix fraying and solvation using  $^{13}\text{C}$  isotopomers." *Protein Sci* **14**(9): 2324-32.
- Finkelstein, A. V. (1997). "Protein structure: what is it possible to predict now?" *Curr Opin Struct Biol* **7**(1): 60-71.
- Fischer, D. and D. Eisenberg (1996). "Protein fold recognition using sequence-derived predictions." *Protein Sci* **5**(5): 947-55.
- Fodje, M. N. and S. Al-Karadaghi (2002). "Occurrence, conformational features and amino acid propensities for the pi-helix." *Protein Eng* **15**(5): 353-8.
- Frishman, D. and P. Argos (1995). "Knowledge-based protein secondary structure assignment." *Proteins* **23**(4): 566-79.
- Gallo, M., D. Paludi, et al. (2005). "Identification of a conserved N-capping box important for the structural autonomy of the prion alpha 3-helix: the disease associated D202N mutation destabilizes the helical conformation." *Int J Immunopathol Pharmacol* **18**(1): 95-112.
- Gorodkin, J., O. Lund, et al. (1999). "Using sequence motifs for enhanced neural network prediction of protein distance constraints." *Proc Int Conf Intell Syst Mol Biol*: 95-105.
- Hogue, C. W. and S. H. Bryant (1998). "Structure databases." *Methods Biochem Anal* **39**: 46-73.
- Holm, L. and C. Sander (1998). "Touring protein fold space with Dali/FSSP." *Nucleic Acids Res* **26**(1): 316-9.
- Hutchinson, E. G. and J. M. Thornton (1996). "PROMOTIF--a program to identify and analyze structural motifs in proteins." *Protein Sci* **5**(2): 212-20.
- Hvidt, A. a. W., P (1998). "Different views on the stability of protein conformations and hydrophobic effect." *J. Solution Chem.* **27**: 395-402.
- Janes, R. W. (2005). "Bioinformatics analyses of circular dichroism protein reference databases." *Bioinformatics* **21**(23): 4230-8.
- Jaroszewski, L., L. Rychlewski, et al. (1998). "Fold prediction by a hierarchy of sequence, threading, and modeling methods." *Protein Sci* **7**(6): 1431-40.
- Jeffrey, G. a. S. W. (1994). *Hydrogen Bonding in Biological Structures*. Berlin, Springer-Verlag.
- Jennings, A. J., C. M. Edge, et al. (2001). "An approach to improving multiple alignments of protein sequences using predicted secondary structure." *Protein Eng* **14**(4): 227-31.
- Jones, D. T. (1999). "GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences." *J Mol Biol* **287**(4): 797-815.
- Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." *J Mol Biol* **292**(2): 195-202.
- Jones, D. T., M. Tress, et al. (1999). "Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure." *Proteins Suppl* **3**: 104-11.
- Jones, D. T. a. O., CA and Thornton, JM (1996). *Protein Folds and Their Recognition from Sequence*. Oxford, Oxford University Press.
- Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* **22**(12): 2577-637.
- Kabsch, W. and C. Sander (1983). "How good are predictions of protein secondary structure?" *FEBS Lett* **155**(2): 179-82.
- Kajava, A. V., J. M. Squire, et al. (2006). "Beta-structures in fibrous proteins." *Adv Protein Chem* **73**: 1-15.
- King, S. M. and W. C. Johnson (1999). "Assigning secondary structure from protein coordinate data." *Proteins* **35**(3): 313-20.
- Koh, I. Y., V. A. Eylich, et al. (2003). "EVA: Evaluation of protein structure prediction servers." *Nucleic Acids Res* **31**(13): 3311-5.
- Kolano, C., J. Helbing, et al. (2006). "Watching hydrogen-bond dynamics in a beta-turn by transient two-dimensional infrared spectroscopy." *Nature* **444**(7118): 469-72.

- Kolinski, A., P. Rotkiewicz, et al. (1999). "A method for the improvement of threading-based protein models." *Proteins* **37**(4): 592-610.
- Kutzki, O., H. S. Park, et al. (2002). "Development of a potent Bcl-x(L) antagonist based on alpha-helix mimicry." *J Am Chem Soc* **124**(40): 11838-9.
- Labesse, G., N. Colloc'h, et al. (1997). "P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins." *Comput Appl Biosci* **13**(3): 291-5.
- Lees, J. G., A. J. Miles, et al. (2006). "A reference database for circular dichroism spectroscopy covering fold and secondary structure space." *Bioinformatics* **22**(16): 1955-62.
- Lesk, A. M. (1991). *Protein Architecture - A practical approach*. Oxford, Oxford University Press.
- Lesk, A. M., L. Lo Conte, et al. (2001). "Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts." *Proteins Suppl* **5**: 98-118.
- Lesk, A. M. and G. D. Rose (1981). "Folding units in globular proteins." *Proc Natl Acad Sci U S A* **78**(7): 4304-8.
- Lindahl, E. and A. Elofsson (2000). "Identification of related proteins on family, superfamily and fold level." *J Mol Biol* **295**(3): 613-25.
- Linderstrøm-Lang, K. (1952). *Proteins and Enzymes*, Stanford University Press.
- Liu, J. and B. Rost (2001). "Comparing function and structure between entire proteomes." *Protein Sci* **10**(10): 1970-9.
- Liu, J. and B. Rost (2002). "Target space for structural genomics revisited." *Bioinformatics* **18**(7): 922-33.
- Lo Conte, L., B. Ailey, et al. (2000). "SCOP: a structural classification of proteins database." *Nucleic Acids Res* **28**(1): 257-9.
- Lupas, A. (1996). "Coiled coils: new structures and new functions." *Trends Biochem Sci* **21**(10): 375-82.
- Maiti, N. C., M. M. Apetri, et al. (2004). "Raman spectroscopic characterization of secondary structure in natively unfolded proteins: alpha-synuclein." *J Am Chem Soc* **126**(8): 2399-408.
- Majumdar, I., S. S. Krishna, et al. (2005). "PALSSE: a program to delineate linear secondary structural elements from protein structures." *BMC Bioinformatics* **6**: 202.
- Manikandan, K. and S. Ramakumar (2004). "The occurrence of C--H...O hydrogen bonds in alpha-helices and helix termini in globular proteins." *Proteins* **56**(4): 768-81.
- Marchler-Bauer, A., K. J. Address, et al. (1999). "MMDB: Entrez's 3D structure database." *Nucleic Acids Res* **27**(1): 240-3.
- Marti-Renom, M. A., A. C. Stuart, et al. (2000). "Comparative protein structure modeling of genes and genomes." *Annu Rev Biophys Biomol Struct* **29**: 291-325.
- Martin, J., G. Letellier, et al. (2005). "Protein secondary structure assignment revisited: a detailed analysis of different assignment methods." *BMC Struct Biol* **5**: 17.
- Murzin, A. G. (1996). "Structural classification of proteins: new superfamilies." *Curr Opin Struct Biol* **6**(3): 386-94.
- Nelson, R., M. R. Sawaya, et al. (2005). "Structure of the cross-beta spine of amyloid-like fibrils." *Nature* **435**(7043): 773-8.
- Offmann, B. a. T., M and de Brevern, AG (2007). "Local Protein Structure." *Current Bioinformatics* **2**(3): 165-202.
- Orengo, C. A., F. M. Pearl, et al. (1999). "The CATH Database provides insights into protein structure/function relationships." *Nucleic Acids Res* **27**(1): 275-9.
- Parisien, M. and F. Major (2005). "A new catalog of protein beta-sheets." *Proteins* **61**(3): 545-58.
- Pauling, L. (1939). *The Nature of the Chemical bond*. New York, Cornell University Press.
- Pauling, L. and R. B. Corey (1951). "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets." *Proc Natl Acad Sci U S A* **37**(11): 729-40.

- Pauling, L., R. B. Corey, et al. (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain." *Proc Natl Acad Sci U S A* **37**(4): 205-11.
- Pearl, F. M., N. Martin, et al. (2001). "A rapid classification protocol for the CATH Domain Database to support structural genomics." *Nucleic Acids Res* **29**(1): 223-7.
- Pelton, J. T. and L. R. McLean (2000). "Spectroscopic methods for analysis of protein secondary structure." *Anal Biochem* **277**(2): 167-76.
- Przytycka, T., R. Aurora, et al. (1999). "A protein taxonomy based on secondary structure." *Nat Struct Biol* **6**(7): 672-82.
- Qiu, J. and R. Elber (2006). "SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs." *Proteins* **62**(4): 881-91.
- Quioco, F. A., J. C. Spurlino, et al. (1997). "Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor." *Structure* **5**(8): 997-1015.
- Ramachandran, G. N. and V. Sasisekharan (1968). "Conformation of polypeptides and proteins." *Adv Protein Chem* **23**: 283-438.
- Rice, D. W. and D. Eisenberg (1997). "A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence." *J Mol Biol* **267**(4): 1026-38.
- Richards, F. M. and C. E. Kundrot (1988). "Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure." *Proteins* **3**(2): 71-84.
- Richardson, J. S. and D. C. Richardson (1988). "Amino acid preferences for specific locations at the ends of alpha helices." *Science* **240**(4859): 1648-52.
- Richardson, J. S. a. R. D. C. (1989). *Principles and patterns of protein conformation*. New York, Plenum Press.
- Rose, G. D. (2006). "Lifting the lid on helix-capping." *Nat Chem Biol* **2**(3): 123-4.
- Rost, B. (1995). "TOPITS: threading one-dimensional predictions into three-dimensional structures." *Proc Int Conf Intell Syst Mol Biol* **3**: 314-21.
- Rost, B. (1996). "PHD: predicting one-dimensional protein structure by profile-based neural networks." *Methods Enzymol* **266**: 525-39.
- Rost, B. (1998). "Marrying structure and genomics." *Structure* **6**(3): 259-63.
- Rost, B. (2001). "Review: protein secondary structure prediction continues to rise." *J Struct Biol* **134**(2-3): 204-18.
- Rost, B. and V. A. Eylich (2001). "EVA: large-scale analysis of secondary structure prediction." *Proteins Suppl* **5**: 192-9.
- Rost, B. and S. O'Donoghue (1997). "Sisyphus and prediction of protein structure." *Comput Appl Biosci* **13**(4): 345-56.
- Rost, B. and C. Sander (1996). "Bridging the protein sequence-structure gap by structure predictions." *Annu Rev Biophys Biomol Struct* **25**: 113-36.
- Rost, B. and C. Sander (2000). "Third generation prediction of secondary structures." *Methods Mol Biol* **143**: 71-95.
- Rost, B., C. Sander, et al. (1994). "Redefining the goals of protein secondary structure prediction." *J Mol Biol* **235**(1): 13-26.
- Rost, B., R. Schneider, et al. (1997). "Protein fold recognition by prediction-based threading." *J Mol Biol* **270**(3): 471-80.
- Rost, B. a. S., C (1994). *1D secondary structure prediction through evolutionary profiles*. Amsterdam, IOS Press.
- Russell, R. B., R. R. Copley, et al. (1996). "Protein fold recognition by mapping predicted secondary structures." *J Mol Biol* **259**(3): 349-65.

- Rzepecki, P., L. Nagel-Steger, et al. (2004). "Prevention of Alzheimer's disease-associated A $\beta$  aggregation by rationally designed nonpeptidic beta-sheet ligands." J Biol Chem **279**(46): 47497-505.
- Sali, A. (1998). "100,000 protein structures for the biologist." Nat Struct Biol **5**(12): 1029-32.
- Sauder, J. M., J. W. Arthur, et al. (2000). "Large-scale comparison of protein sequence alignment algorithms with structure alignments." Proteins **40**(1): 6-22.
- Schulz, G. E. (1988). "A critical evaluation of methods for prediction of protein secondary structures." Annu Rev Biophys Biophys Chem **17**: 1-21.
- Shapiro, L. and T. Harris (2000). "Finding function through structural genomics." Curr Opin Biotechnol **11**(1): 31-5.
- Shi, J., T. L. Blundell, et al. (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." J Mol Biol **310**(1): 243-57.
- Sippl, M. J. (1995). "Knowledge-based potentials for proteins." Curr Opin Struct Biol **5**(2): 229-35.
- Sippl, M. J. and H. Flockner (1996). "Threading thrills and threats." Structure **4**(1): 15-9.
- Sklenar, H., C. Etchebest, et al. (1989). "Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis." Proteins **6**(1): 46-60.
- Song, M., S. Rajesh, et al. (2001). "Design and synthesis of new inhibitors of HIV-1 protease dimerization with conformationally constrained templates." Bioorg Med Chem Lett **11**(18): 2465-8.
- Sreerama, N., S. Y. Venyaminov, et al. (1999). "Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy." Protein Sci **8**(2): 370-80.
- Srinivasan, R. and G. D. Rose (1999). "A physical basis for protein secondary structure." Proc Natl Acad Sci U S A **96**(25): 14258-63.
- Stec, B., U. Rao, et al. (1995). "Refinement of puorothionins reveals solute particles important for lattice formation and toxicity. Part 2: structure of beta-puorothionin at 1.7 Å resolution." Acta Crystallogr D Biol Crystallogr **51**(Pt 6): 914-24.
- Sternberg, M. J., P. A. Bates, et al. (1999). "Progress in protein structure prediction: assessment of CASP3." Curr Opin Struct Biol **9**(3): 368-73.
- Stromer, T. and L. C. Serpell (2005). "Structure and morphology of the Alzheimer's amyloid fibril." Microsc Res Tech **67**(3-4): 210-7.
- Sugiyama, S., Y. Matsuo, et al. (1996). "The 1.8-Å X-ray structure of the Escherichia coli PotD protein complexed with spermidine and the mechanism of polyamine binding." Protein Sci **5**(10): 1984-90.
- Taylor, W. R. (2001). "Defining linear segments in protein structure." J Mol Biol **310**(5): 1135-50.
- Teeter, M. M. (1984). "Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin." Proc Natl Acad Sci U S A **81**(19): 6014-6018.
- Tetin, S. Y., F. G. Prendergast, et al. (2003). "Accuracy of protein secondary structure determination from circular dichroism spectra based on immunoglobulin examples." Anal Biochem **321**(2): 183-7.
- Villanueva, J., V. Villegas, et al. (2002). "Protein secondary structure and stability determined by combining exoproteolysis and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." J Mass Spectrom **37**(9): 974-84.
- Wade, R. C., K. J. Clark, et al. (1993). "Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds." J Med Chem **36**(1): 140-7.
- Wagner, G. (2005). "Ending the prolonged life of cancer cells." Nat Chem Biol **1**(1): 8-9.

- Wallqvist, A., Y. Fukunishi, et al. (2000). "Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases." Bioinformatics **16**(11): 988-1002.
- Watson, D., E. Castano, et al. (2005). "Physicochemical characteristics of soluble oligomeric A $\beta$  and their pathologic role in Alzheimer's disease." Neurol Res **27**(8): 869-81.
- Wilmot, C. M. and J. M. Thornton (1990). "Beta-turns and their distortions: a proposed new nomenclature." Protein Eng **3**(6): 479-93.
- Xu, Y., D. Xu, et al. (1999). "Protein threading by PROSPECT: a prediction experiment in CASP3." Protein Eng **12**(11): 899-907.
- Yang, A. S. and B. Honig (2000). "An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments." J Mol Biol **301**(3): 691-711.
- Young, M., K. Kirshenbaum, et al. (1999). "Predicting conformational switches in proteins." Protein Sci **8**(9): 1752-64.
- Zemla, A., C. Venclovas, et al. (1999). "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment." Proteins **34**(2): 220-3.