Assignment of Local Protein Structure with Different Strategies

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt in englischer Sprache von

Jan Zacharias

aus Hannover

Berlin, Oktober 2014

Die vorliegende Arbeit wurde unter Anleitung von Prof. Dr. E. W. Knapp im Zeitraum von 05.2010 - 08.2014 am Institut für Chemie / Physikalische und Theoretische Chemie der Freien Universität Berlin im Fachbereich Biologie, Chemie und Pharmazie durchgeführt.

1. Gutachter: Prof. Dr. Ernst-Walter Knapp
2. Gutachter: Prof. Dr. Markus Wahl

Disputation am 16.12.2014

Preamble

This thesis summarizes my doctoral research work. It is mainly based on the following two peer-reviewed journal publications:

J. Zacharias and E. W. Knapp, "Geometry motivated alternative view on local protein backbone structures.," *Protein Sci.*, vol. 22, no. 11, pp. 1669–74, Nov. 2013.

http://dx.doi.org/10.1002/pro.2364

J. Zacharias and E.-W. Knapp, "Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC.," *J. Chem. Inf. Model.*, Jun. 2014.

http://dx.doi.org/10.1021/ci5000856

Acknowledgements

This work was carried out at the Freie Universität Berlin in the group of Prof. Ernst-Walter Knapp. I would like to thank him for fruitful discussions and his valuable support.

Arturo Robertazzi for proofreading this manuscript.

Nadia Elghobashi-Meinhardt for proofreading both papers.

All members of the Knapp Group that created a cooperative and friendly working environment.

Meiner Familie für beständige moralische und gelegentliche finanzielle Unterstützung.

Statutory Declaration

I hereby testify that this thesis is the result of my own work and research, except for any explicitly referenced material, whose source can be found in the bibliography. This work contains material that is the copyright property of others which cannot be reproduced without the permission of the copyright owner.

Jan Zacharias

Table of contents

Introduction

Proteins are polymers of amino acids and are essential for all living organisms. They play an important role in virtually all biological reactions—a fact reflected by the vast abundance of proteins in eukaryotic cells, which consist of 70% water and 15% proteins[1]. The broad range of protein functions covers active roles such as immune response, cell signaling, cell reproduction, and catalysis of biochemical reactions as well as passive tasks, like structural functions in the viral envelope or in collagen, and keratin.

The function of a protein is determined by its structure, which is usually described on four different levels of organization: the primary, secondary, tertiary, and quaternary structure.

Proteins consist of polypeptide chains of highly variable size of the twenty different proteinogenic amino acids. The sequence of these amino acids represents the primary structure of the protein. The size of proteins spans the whole range from 20 amino acid residues, as in the case of the synthetic Trp-Cage miniprotein[2], up to 33.000 of Titin, which provides the passive elasticity of muscles[3].

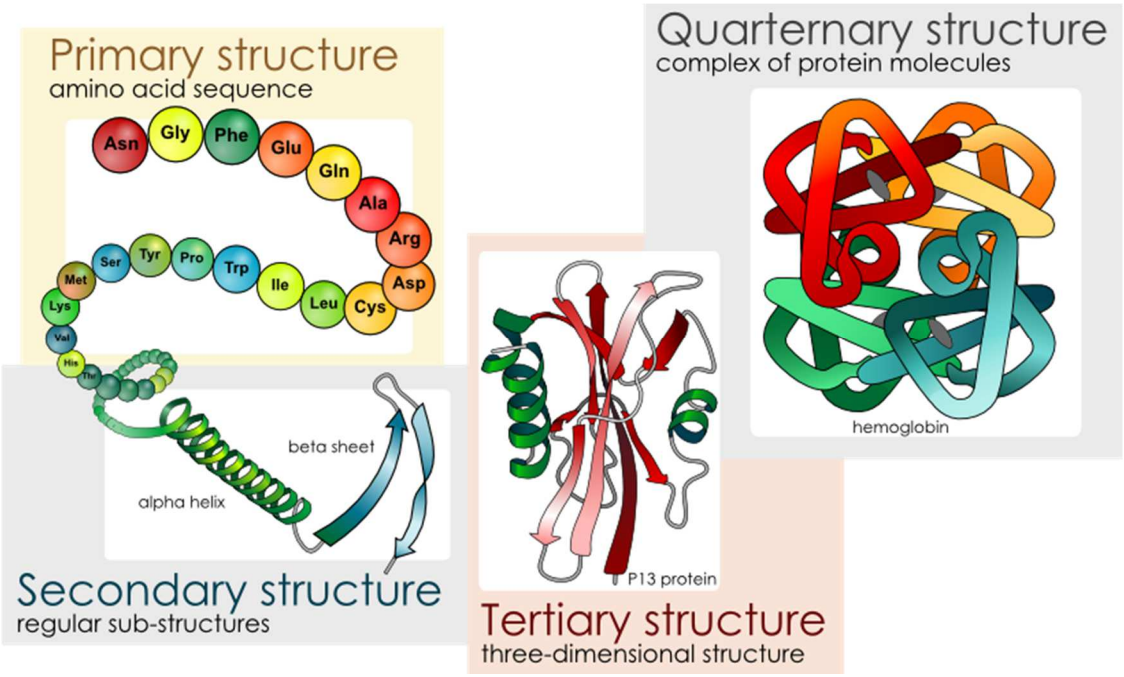

Figure 1: The four levels of biomolecular structure. Minor modifications to original artwork by Mariana Ruiz Villarreal[4].

Of special interest for this work is the protein secondary structure, which describes the local spatial arrangement of shorter protein segments. The most prominent examples of regularly repeating secondary structure motifs are the α-helix and the β-strand. Following the preliminary work by William Astbury in the early 1930s[5] and the prediction by Linus Pauling in 1951[6], [7] the first X-ray structures of myoglobin[8] and hemoglobin[9] were solved three years later, confirming the existence of these structures. In fact, roughly one half[1] of all residues in a protein are either helical or part of a β-strand. The definitions "α-helix" and "β-strand" were derived from the fibrous structural proteins α- and β-keratin, which are both rich in the respective motifs[5]. Neighboring β-strands form the so-called β-pleated-sheets (also called β-sheet). Helices and β-sheets are stabilized by a repeating hydrogen-bond pattern between the protein backbone's C=O- and N-H-groups of different amino acid residues.

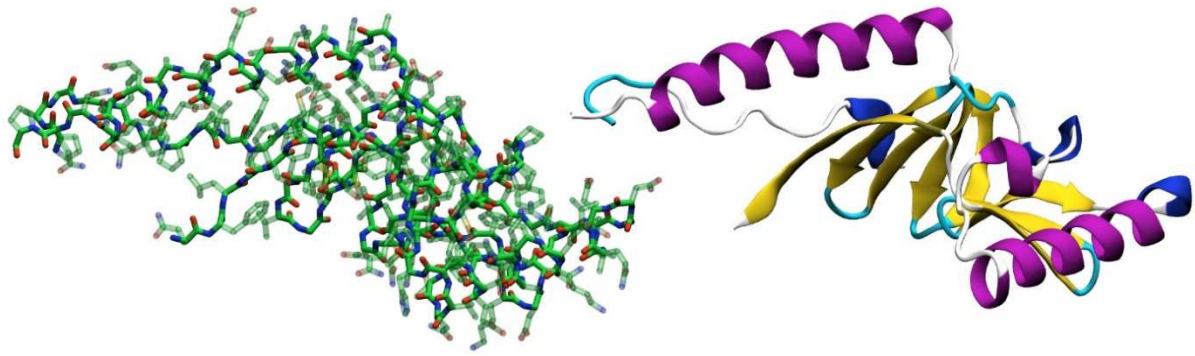

Figure 2: Two visualizations of pepsin inhibitor-3 protein (1F34[10]–[12]*) from Ascaris suum (large round worms of pigs). Left: All-atom representation with backbone atoms in solid and side-chain atoms in transparent mode. Right: Cartoon representation of the same protein. Alpha-helices are in purple, 3_{10} helices in dark blue, strands in yellow, and hydrogen-bonded turns in turquois. Both images have been created with VMD, which uses stride for secondary structure assignment.*

[1] 54% percent for the Astral40 dataset[11] of version 1.75 according to PSSC and DSSP

At the next higher organization level, the tertiary structure describes the complete three-dimensional folding of the protein's peptide chain. Besides covalent disulfide bonds between cysteine side chains, a combination of different non-covalent interactions stabilize the structure of a protein, i.e., the hydrophobic effect of polypeptide-water interactions, salt bridges, and hydrogen bonds, including backbone and side-chain groups.

Several protein chains can form a protein complex. A specific protein may only be functional as such a multimer. As an example, antibodies consist of four chains, i.e., two copies of the immunoglobulin heavy chain and two of the immunoglobulin light chain. The arrangement of protein subunits in space is described by the quaternary structure.

Knowledge about a protein's fold, the arrangement of major secondary structural elements, is a key step towards the understanding of the protein's function. Even though a strong correlation between structure and function exists, structural conservation between functionally similar proteins is more pronounced than conservation of amino acid sequences[13].

The protein folding problem is the task to predict the three-dimensional structure of a protein from its sequence, i.e., secondary and tertiary structures from primary structure. Experimental determination of a protein's structure is mostly carried out with either X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. NMR is usually restricted to smaller water-soluble proteins, while X-ray crystallography requires the protein to be prepared in the crystalline state first—a non-trivial process, which may be challenging or even impossible for some proteins. In contrast to that, full genome sequencing has become a largely automatized process with high throughput at low costs. Hence, the number of known protein sequences raises on a much higher rate than that of known structures. To measure the quality of a conducted secondary structure prediction as well as to train the algorithms employed for this task, a reliable method for secondary structure assignment is crucial.

A versatile tool to describe a protein's structure in a two-dimensional graph is the Ramachandran or (ϕ, ψ) plot, introduced in 1963[14]. In this representation, the backbone torsion angle ψ of a residue is plotted against the torsion angle ϕ, leading to a distinctive scatter plot, where residues of similar secondary structures are found in close proximity, independent from their spatial and sequential distances.

Additionally to two-dimensional visualizations, protein structures are nowadays routinely depicted in a three-dimensional manner on modern computer hardware. The usage of a "cartoon" or ribbon representation[15] has become what can be safely called the most common way of representing protein structures in publications (usually created with tools such as molscript[16], VMD[17], and PyMOL[18]) and online tools for interactive protein visualizing such as Jmol[19], JSmol[20], and GLmol[21] . For such tools to work properly, a solid assignment of helix and β-strand residues is critical to avoid visually unattractive and, most importantly, misleading results.

Interestingly, despite the indisputable importance of secondary structure prediction and hence structure assignment, a widely accepted canonical definition of protein secondary structure has not yet been proposed. Textbooks as well as publications dealing with protein structure almost exclusively focus on idealized motifs that are of infinite length without disruptions and ambiguities. In the interim regions of two secondary, structural motifs residues exist that may be assigned to any of the two interconnected motifs. Especially helices tend to possess contractions and bulges that add 3_{10} helical or π-helical character to residues of an α-helix.

While some publications deal with the problem of helix capping and kinks in longer helices [22]–[26], the majority of available secondary structure assignment software does not take into account the information illustrated in these studies.

The de facto standard for assigning protein secondary structure remains the software DSSP, which was developed in 1983 by Kabsch and Sander[27]. During my work, I developed a fork of this software, named PSSC (Protein Secondary Structure Characterization) that fixes many of the problems of the original software and adds new features such as an identification of mixed secondary classes, left-handed hydrogen-bonded helices, and the polyproline II helix.

Publications

"Geometry motivated alternative view on local protein backbone structures"

Authors

Zacharias, J., Knapp, E.W.

Contribution

- Development of the research question
- Development of the webpage and necessary software tools
- Generation and analysis of the results
- Manuscript preparation

Summary

In this publication, the (d, ϑ)-plot is introduced as an alternative to the well-known Ramachandran plot. Instead of the (ϕ, ψ)-backbone angles, the helix rotation angle ϑ and the helical rise parameter d are displayed in a polar diagram. Both parameters are derived from a description of the local protein backbone structure in terms of a helix that would occur if the (ϕ, ψ) angles were repeated indefinitely. As repeated values of ϕ and ψ always result in a helical symmetry of the backbone structure, this transformation is possible for the whole (ϕ, ψ) space. A helix can be described by the angular rotation step ϑ and the rise d per residue, both with respect to the helical axis.

Assuming standard backbone geometry, the formulas for d and ϑ are then given by:

$$\cos\left(\frac{\vartheta}{2}\right) = -0.8235 \sin\left(\frac{\psi+\varphi}{2}\right) - 0.0222 \sin\left(\frac{\psi-\varphi}{2}\right) \quad (1)$$

$$d \sin\left(\frac{\vartheta}{2}\right) = 2.999 \cos\left(\frac{\psi+\varphi}{2}\right) - 0.657 \cos\left(\frac{\psi-\varphi}{2}\right) \quad (2)$$

The sign of ϑ corresponds to the handedness of the helix (positive for right-handed, negative for left-handed), and the number of residues per full turn is given by $n = 360°/\vartheta$. Hence, a clear discrimination of the handedness of a local structural motif is gained: residues on the left side of the (d, ϑ) correspond to (ϕ, ψ) values that would generate left-handed helices if repeated. For this publication, all parameter pairs (n, r, d, D, ϑ, ϕ, ψ; where $D = d\,n$) were examined, and the combination of d and ϑ was found to be most insightful.

Because in eq. (1) the dominant term is the sine of the sum of ϕ and ψ, the isolines of constant ϑ are almost parallel to the lines for $\phi + \psi = \text{const}$. Firstly, it should be noted that helical residues possess dihedral angles in a way that the sum $\phi + \psi$ is approximately constant[2]; secondly, the boundary between the PII basin and the beta strand region is also diagonally shaped. Both features make the (d, ϑ) plot very appealing for secondary structure assignment.

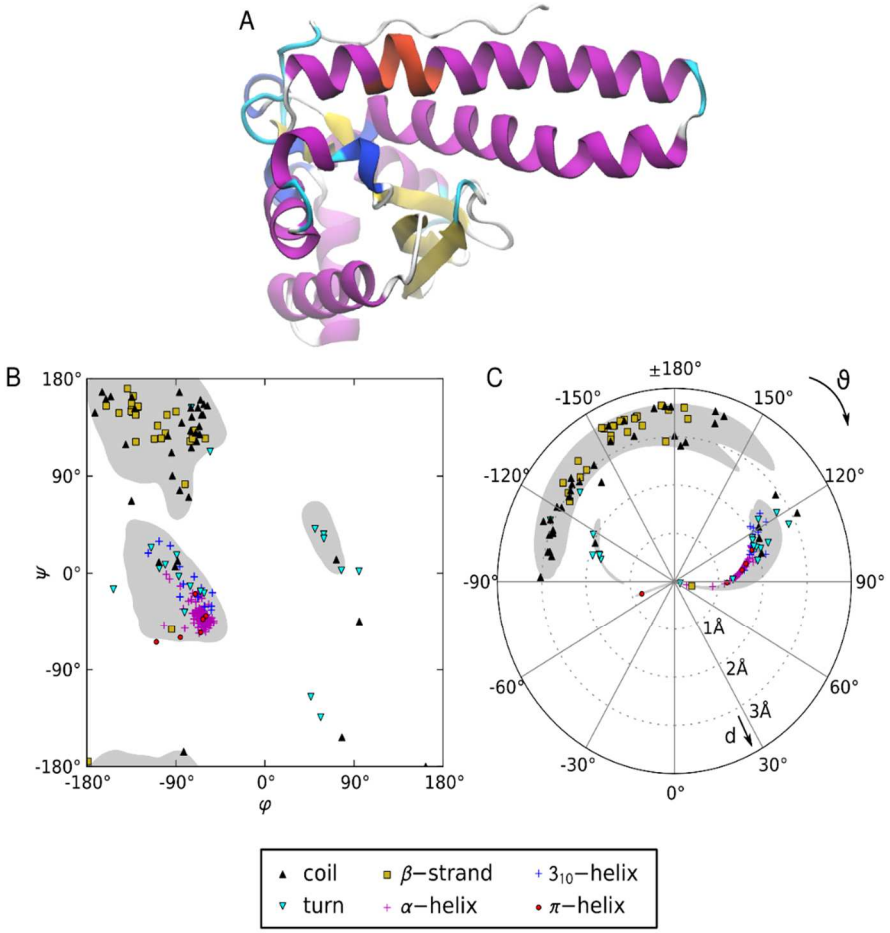

Figure 3: Comparison of protein visualizations. A: Comic representation of the crystal structure of chain A of human superoxide dismutase (PDB id: 1KKC [28]*). B: Ramachandran plot with sterically allowed regions shaded in gray. C: (d, ϑ)-plot of the same data.*

[2] To substantiate this claim, the mean and the standard deviation of the dihedral angles of all residues in the Astral40 protein dataset were evaluated that belong to α-helices according to PSSC. The results are $\phi = -64° \pm 12°$, $\psi = -40° \pm 12°$, and $\phi + \psi = -105° \pm 13°$. If the values ϕ and ψ were uncorrelated, the standard deviation of their sum would be. A clear diagonal trend can also be observed for the allowed regions in the Ramachandran plot of *Figure 3*.

"Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC"

Authors

Zacharias, J., Knapp, E.W.

Contribution

- Software development
- Generation and analysis of the results
- Manuscript preparation

Summary

Today's most widely used program for secondary structure assignment is DSSP, despite the fact that this software has been introduced in 1983 and only a few algorithmic changes have been proposed. The structure characterization carried out with DSSP is divided into eight distinct classes; this represents the basis for several approaches for protein secondary structure prediction and is used for learning, prediction, and evaluation.

In this publication, an alternative concept is introduced, representing the internal structure characterization of DSSP as an eight-character string that is human-interpretable and easy to parse by software. This protein secondary-structure characterization (PSSC) code allows for inspection of complicated structural features.

In order to evaluate the introduced changes in interpreting DSSP information, it is shown that a better clustering of secondary structures in (ϕ, ψ) dihedral angle space can be obtained with the PSSC method.

The possible definition of new secondary structure classification schemes with PSSC is demonstrated, and classifications are performed for a number of examples. The approach presented in this work, which was applied without modifying the DSSP source code, enables a more detailed protein characterization. DSSP's original one-letter code is easily derivable from the eight-letter PSSC representation.

Conclusion

In the first publication "Geometry motivated alternative view on local protein backbone structures", a transformation from (ϕ, ψ)-space to (d, ϑ)-space was introduced; the usability of this transformation was demonstrated mainly by graphical representations. In the second publication "Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC", the transformation from (ϕ, ψ) to (d, ϑ) is applied to discriminate between left- and right-handed helices, a distinction not possible with the original DSSP program. Particular attention is paid on the importance of the order in which secondary structure classes are considered. Especially the three different helix types are often found following directly after another. Hence, single residues in the transition regime could be assigned to two different helices of different types. However, after reduction to a single letter code information about ambiguities is lost. While the full output of DSSP still carries this information, it cannot be used without actually reimplementing larger parts of DSSP's algorithm. In the PSSC code, the ambiguity of mixed classes is clearly displayed.

As the version of the PSSC program presented in this publication still relies on the original software DSSP, a new version has been developed, with a complete redesign of the underlying DSSP algorithm, but includes the PSSC code generation. While algorithms and results for hydrogen-bond pattern assignment are similar to those of DSSP, a number of features have been added, among these, the most important being the (d, ϑ)-diagram, which is now the basis for a secondary structure classification that is largely orthogonal to the hydrogen bond-based classification. In the web frontend to the PSSC software, the secondary structure assignment based on the current PSSC version is displayed together with a classical Ramachandran plot side-by-side to the new (d, ϑ)-plot.

Protein Backbone Geometry

A protein chain is a linear polymer chain of amino acids that are bonded together by peptide bonds. The formation of a peptide bond (or amide bond) between two amino acids occurs through a dehydration reaction, i.e., the carboxyl group (–COOH) of one amino acid forms a covalent bond with the amino group (NH_2) of another amino acid by releasing a water molecule (H_2O) in the process. Hence, the building blocks of a protein chain are usually referred to as amino acid residues or just residues. The name C-terminus refers to that end of the chain with the free carboxyl group, while the N-terminus is the end with the free amino group. Residues are usually numbered sequentially from the C- to the N-terminus as this is the direction in which proteins are synthesized in the ribosome[1].

The 22 different proteinogenic amino acids (including the two cotranslationally inserted amino acids selenocysteine and pyrrolysine[29]) have identical backbone atoms and only differ in their side chains. The side chain is attached to the Cα atom, i.e., the carbon atom between the amino and the carboxyl group. Because Cα is a chiral atom, all amino acids possess handedness, with the exception of glycine, where a single hydrogen atom is bonded to Cα instead of a side chain. Naturally occurring amino acids are usually L-α-stereoisomers (left-handed isomers), even though right-handed d-α-amino acids have been shown to exist in some microorganisms, plants, and fish [30]–[32].

The backbone of a protein consists of the N, Cα, C, and O atoms of its amino acids residues. This backbone structure is relatively rigid and possesses only a limited number of degrees of freedom. While bond angles and distances are virtually constant, the N-Cα-C angle τ_1 can vary by about ± 5°[33].

Bond lengths, bond angles, and rotational angles calculated for a protein can be used for secondary structure description or may serve as a quality measure for validation of a proposed new structure. When employed for structural validation, the values calculated for a single protein are compared against the standard measures derived from a large dataset of reference structures[34]–[36]. It should be noted that a dihedral angle can formally range between -180° and +180°, while an angle between two vectors is restricted to values from 0° to 180°. Besides these bond-geometry parameters, pseudo-geometry parameters can be calculated that are based solely on Cα atom positions. Such geometry parameters are usually referred to as pseudo-bond angles or pseudo-torsion angles. Cα atoms are always a suitable selection as an

additional source of information on the backbone structure, because these chemically inert atoms are located at the center of their residue.

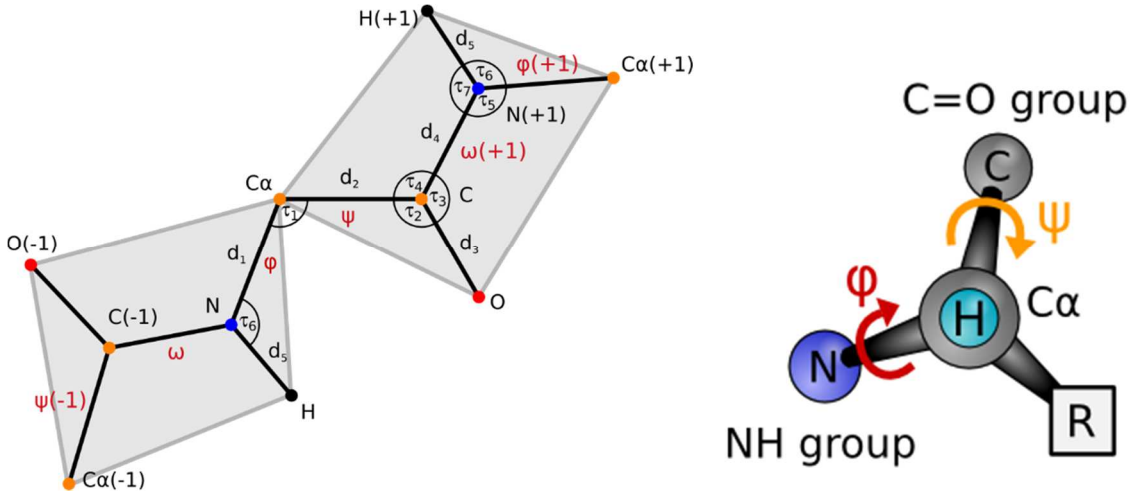

Figure 4: Left side: Protein backbone geometry. Right side: Backbone torsion angles ϕ and ψ.

bond length		value [degree]
d_1	N-CA	1.460(12)
d_2	CA-C	1.524(12)
d_3	C-O	1.233(12)
d_4	C-N	1.330(09)
d_5	N-H	1.003(04)

parameter		value [Å]
τ_1	N-CA-C	111.1(25)
τ_2	CA-C-O	120.5(11)
τ_3	O-C-N	122.7(11)
τ_4	CA-C-N	116.7(14)
τ_5	C-N-CA	121.4(17)
τ_6	CA-N-H	117.4(14)
τ_7	C-N-H	120.9(14)

Table 1: Geometric parameters of the protein backbone. The values were measured on 105332 residues from 550 proteins with no more than 40% sequence identity. Only structures were taken from the PDB that had been resolved via X-ray diffraction with a resolution between 0.5Å and 1.5Å. Hydrogen atoms were added with CHARMM; hence, the parameters d_5, τ_6 and τ_7 are essentially reflecting properties of the CHARMM force field. The only parameter that is considerably influenced by the angle omega is the angle τ_5 (C-N-CA). For residues in the cis conformation (i.e. the omega dihedral angle around the enclosed C-N axis being around 0°; this is the case for 327 residues in this dataset), the corresponding value for t_5 becomes 126.7(32)°.

Of the 88369 protein structures available in the Protein Data Bank[37] (PDB) as of November 2013, the majority (89%) is resolved by means of X-ray crystallography and 75% have been determined with a resolution higher than 2.5 Å. At such resolution, the three-dimensional co-ordinates of most heavy atoms are known with sufficient accuracy to infer hydrogen bonds between polar groups. For those proteins with very high resolution, even the coordinates of the backbone hydrogen atoms can be resolved. However, because the single electron of the hydrogen atom only leads to a weak reflection in the X-ray scattering pattern, and the hydrogen has a higher mobility than heavy atoms, hydrogen atoms usually need to be added by modeling[38].

Backbone dihedral angles

The main source for protein backbone flexibility stems from the dihedral angles ϕ and ψ, corresponding to the rotation around the N-Cα and the Cα-C bonds of the same Cα atom, respectively (Figure 4). These two angles are not mutually independent, but restricted by steric hindrances[39]. Both angles ϕ and ψ are typically defined to be 180° for a fully extended, planar conformation and increase in a right-handed (clockwise) sense when viewed from Cα to N for angle ϕ, or from Cα to C for ψ. Early studies used a definition that was offset by 180° compared to the modern convention.

The fact that the whole backbone flexibility can be described in a good approximation by only the two parameters ϕ and ψ allowed Ramachandran[14] to introduce the (ϕ, ψ)-plot that is now commonly known under his name.

The third dihedral angle ω describes the rotation around the C-N bond and is restricted to angles around 180° (the trans conformation) or to angles around 0° (the rare cis conformation). Nearly all residues in a protein adopt the trans conformation with ω around +/-180°. In the Astral40 dataset only 0.2% of all residues are found in the cis conformation state (Figure 5). The majority (85%) of the cis residues is proline peptides. These figures are in line with previous findings [40], [41]. Because the existence of a cis instead of a trans peptide bond may exert a strong influence on the structure (for instance, the distance between the Cα atoms of consecutive residues is reduced by roughly 1 Å), a reliable assignment of the adopted conformation is important.

As shown in Figure 5, ω is strictly restrained to values close to either the ideal cis or the ideal trans conformation so that the two regions around +/-90° are left nearly empty. Hence, the value of 90° is a reasonable threshold for cis/trans discrimination, i.e., residues with $|\omega| < 90°$ may safely be considered as cis and those with $|\omega| \geq 90°$ as trans.

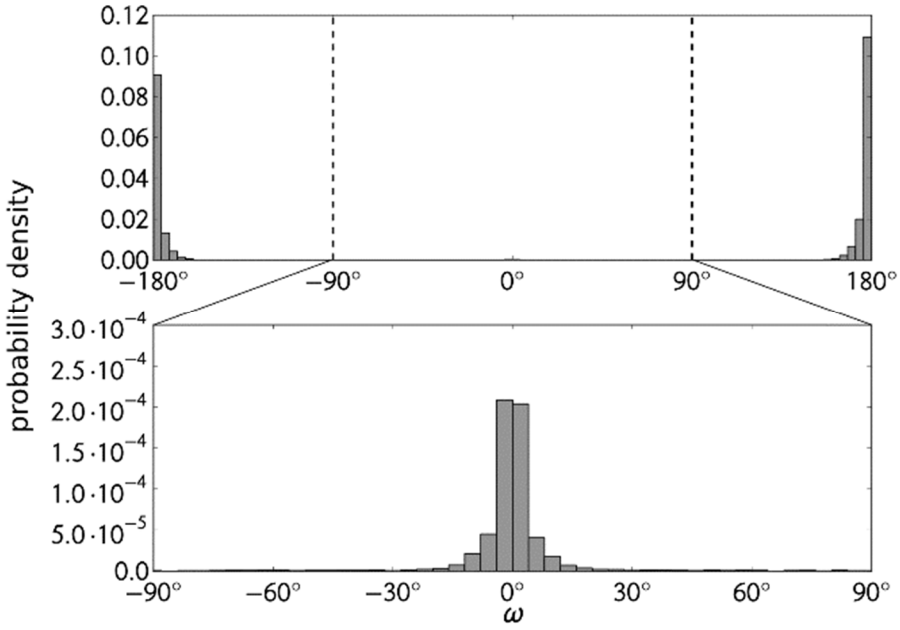

Figure 5: Histogram of the dihedral angle ω in the Astral40 dataset. The normalized density distribution of ω values is also shown (i.e. the integral over all bins equals 1). The lower part displays an enlarged view on the minority class of cis residues around ω = 0° that is almost completely hidden when displayed together with the vast majority of trans residues.

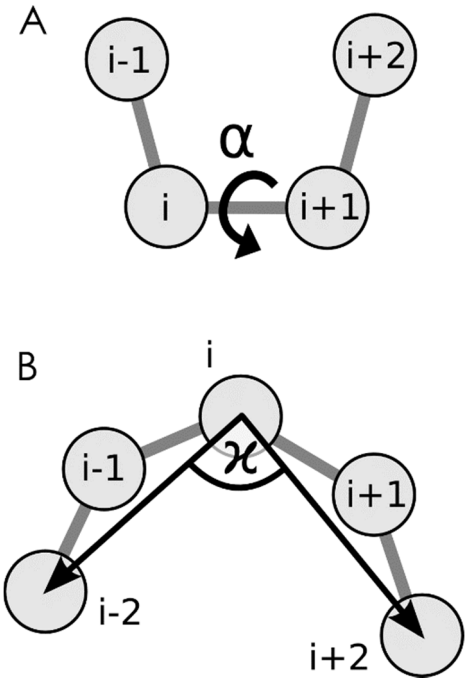

Figure 6: Definition of pseudo-torsion angle α and pseudo-bond angle ϰ

Pseudo-Bond angles

The pseudo-bond angle $\varkappa$ and the pseudo-torsion angle α (Figure 6) are both employed in DSSP and are defined in terms of Cα coordinates. For residue i, the pseudo-bond angle $\varkappa$ is calculated as:

$$\varkappa(i) = \measuredangle C\alpha(i - 2)C\alpha(i)C\alpha(i + 2) \qquad (3)$$

and the pseudo-torsion angle α as:

$$\alpha(i) = \sphericalangle C\alpha(i - 1)C\alpha(i)C\alpha(i + 1)C\alpha(i + 2). \qquad (4)$$

While the definition for $\varkappa$ is symmetric with respect to the atom C$\alpha(i)$, four points are needed for the calculation of the dihedral angle $\alpha(i)$, so that C$\alpha(i)$ cannot be placed at the center of the residue window under consideration.

Similarly to the ϕ and ψ angles, the $\varkappa$ and α angles are not independent from each other. For a pair of angles defined in a similar manner (but on a narrower sequence window), this was used for analog diagrams of the Ramachandran plot in [42], [43].

DSSP uses the pseudo-bond angle $\varkappa$ to mark residues as bent, when $\varkappa < 110°$. However, residues in helices and beta-turns are virtually always bent according to this criterion. The bent class S is used as the lowest priority class in DSSP: Only when no regular secondary structure (such as helix, strand, or even turn) is assigned to a residue, it might be assigned to the class S. The value of the torsion angle α is not used for class assignment, and only the sign of α is given to mark right- or left-handed regions in the protein in the full output of DSSP.

angle	definition	comment
ω	Torsion angle	cis (0°) or trans (180°)
ϕ	Torsion angle	Abscissa in Ramachandran-Plot
ψ	Torsion angle	Ordinate in Ramachandran-Plot
$\varkappa$	Cα-pseudo bond angle	Bend class S in DSSP when $\varkappa < 110°$
α	Cα-torsion angle	Handedness in DSSP
ϑ	Helix angular step	Used in the (d, ϑ)-plot

Table 2: Summary of various angles in protein geometry

Hydrogen Bonds

Properties

A hydrogen bond X–H…Y–Z is a directed interaction between an electronegative atom (acceptor, Y) that is covalently bond to a relatively electropositive atom (Z) and a hydrogen atom (H) attached to another electronegative atom (donor, X). Hydrogen bonds are well described by electrostatic forces between two interacting permanent dipoles as only a small wave-function overlap occurs. This negligible partial covalent-bond character stems from the charge transfer between the donor and acceptor[44]. Typical acceptors are oxygen, nitrogen, and fluorine. With typical energies in the range of 2–6 kcal/mol (8–25 kJ/mol) [45]–[47], hydrogen bonds are weaker then covalent bonds, but usually stronger than dipole-dipole or van der Waals interactions. As for covalent bonds, the length of a hydrogen bond is relatively fixed. Moreover, the van der Waals radii of the donor and acceptor atoms typically overlap.

The strength of the bond is stronger as the X–H…Y angle approaches linearity (180°), i.e., strong hydrogen bonds are often linear.[44]

Hydrogen bonds in proteins

Hydrogen bonds play a critical role in protein folding and protein dynamics. They give rise to the hydrophobic effect that drives protein folding by forcing hydrophobic amino acids to be buried into the protein's interior and hydrophilic amino acids to the solvent-exposed surface. On the surface, the potential of hydrogen-bond formation facilitates the specificity and strength of binding modes to other proteins or smaller molecules.

Additional stabilization of the protein structure stems from the formation of hydrogen bonds within the protein core. Backbone amide and carbonyl groups, the side chains of polar residues, and buried water molecules can form hydrogen bonds in the inner region of the protein. Repeated backbone-to-backbone hydrogen bonding pattern are the crucial feature of protein secondary structure formation and stabilization.

Observed donor (H) to acceptor (O) distances in α-helices of proteins range between 1.7 and 2.4 Å. The hydrogen bond donor angle (N-H…O) ranges between 130° and 170°, while the acceptor angle (C=O…H) is slightly more restricted to values around 150°. [48], [49].

Definition in DSSP and PSSC

The mostly electrostatic characteristics of a hydrogen bond interaction complicates the decision about the existence of a hydrogen bond between two potential hydrogen bond partners. There is no sharp cutoff, neither for distance, nor for angle or energy, to identify hydrogen bonds; one or several of these criteria are usually applied.

DSSP assumes a backbone hydrogen bond to be present between two residues if the electrostatic energy E is below -0.5 kcal/mol. An approximation of E in kcal/mol is calculated by the formula

$$E = f \times q_1 q_2 \left(+ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{CH}} - \frac{1}{r_{CN}} \right) \qquad (5)$$

With charges $q_1 = 0.24e$ and $q_2 = 0.20e$. The factor $f = 332$ converts the value of e to units in kcal/mol.

The energy function that is in use in PSSC considers the electrostatic interaction between donor and acceptor atoms with the following more realistic expression:

$$E = f \times \left(\frac{q_H q_O}{r_{HO}} + \frac{q_H q_C}{r_{HC}} + \frac{q_N q_C}{r_{NC}} + \frac{q_N q_O}{r_{NO}} + \frac{q_C q_{C\alpha}}{r_{CC\alpha}} + \frac{q_O q_{C\alpha}}{r_{OC\alpha}} \right) \qquad (6)$$

The values for the partial charges are taken from the CHARMM force field and given in Table 3. The value of 0.16 for $q_{C\alpha}$ charge units for the Cα atom was chosen to guarantee a net zero overall charge. A hydrogen bond is assumed to exist when the energy e is below a user-defined threshold. As a reasonable default value for PSSC the threshold value of -0.75 kcal/mol has been set.

value	atom	charge
$\boldsymbol{q_H}$	H	0.31
$\boldsymbol{q_N}$	N	-0.47
$\boldsymbol{q_C}$	C	0.51
$\boldsymbol{q_O}$	O	-0.51
$\boldsymbol{q_{C\alpha}}$	Cα	0.16

Table 3: Atomic partial charges as derived from CHARMM.

Secondary Structure Types

Beside β-strands and the three right-handed helical types, some less prominent, but still frequently occurring secondary structure motifs, exist. Additional motifs like the 2.2_7 ribbon, the gamma-helix and left-handed helices have been proposed in the past, but are now known not to exist in longer, repeating segments[50].

3_{10}, α-, and π-helix

The three helix types 3_{10}, α-, and π-helix are characterized by their repeated $(i, i + n)$ hydrogen-bonding pattern: A hydrogen bond exists between the accepting C=O group of residue i and the donating N-H group of the residue at position $i + n$, where the value of n is 3 for 3_{10} helices, 4 for α, and 5 for π helices (Figure 7). To ensure such a repeating network of hydrogen bonds, the backbone of the protein must follow a tightly packed helical path with the amino acid side chains pointing outwards.

Following the same nomenclature as for the 3_{10} helix, the α-helix has historically been named 3.6_{13} and the π-helix as 4.4_{16} helix. In this notation, the number of residues per helical turn is given first, and the following subscript refers to the number of atoms contained in the ring formed by the hydrogen bond. The vast majority of all helical residues is in the α-helical conformation; π-helices exists only rarely in their pure form. Usually they appear as a bulge in a longer α-helical segment[50]–[52]. 3_{10} helices can appear at the beginning or end of a longer α-helix, but are also often found alone; they are usually significantly shorter than a typical α-helix.

The IUPAC-IUB suggests two alternative definitions of a helical segment, the first being based on the ϕ, ψ angles and the second based on the hydrogen-bond pattern[53].

According to the (ϕ, ψ)-based definition, an α-helix is a stretch of residues with (ϕ, ψ) angles close to $(-57°, -47°)$. The first and last residues in this stretch are also the first and last residues of the helix segment[53].

Following the second definition, the first residue of a helix segment is the first residue whose C=O group partakes in the regular $(i, i + n)$ hydrogen-bonding pattern and the last residue is the last residue whose N-H group is acting as donor in the pattern. Residues with irregular hydrogen bonds are not be considered to be helical. It should be noted that helical segments

defined by the (ϕ, ψ)-rule may be up to two residues shorter than the same segment defined by the hydrogen-bonding rule.

Because the DSSP secondary structure definition does not count the first and last residues involved in the helix hydrogen-bonding pattern as helical, the DSSP definition is in closer agreement with the (ϕ, ψ)-based IUPAC rule.

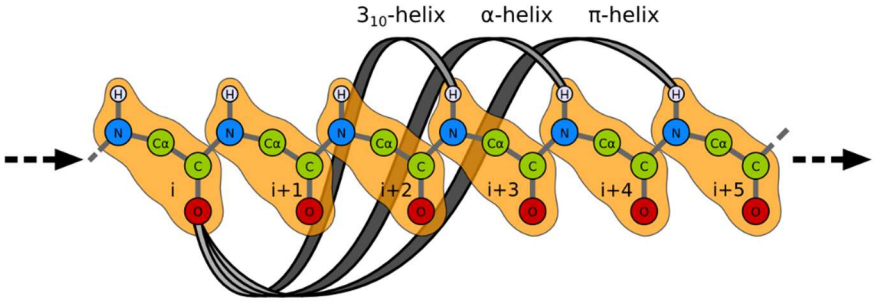

Figure 7: Hydrogen-bonding pattern in helices.

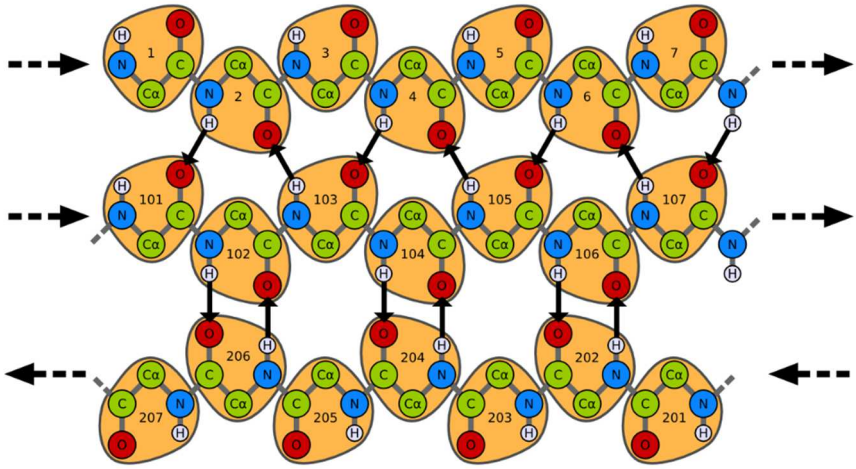

Figure 8: The hydrogen-bonding pattern in an idealized beta-sheet.

β-Strands

When a number of consecutive residues are close to their fully extended conformation with ϕ and ψ both close to 180°, they form a longer polypeptide stretch that is called a β-strand. Such strands are usually hydrogen bonded to other strands, either in a parallel or antiparallel manner, forming the so-called β-sheets. The anti-parallel arrangement allows for linear hydrogen-bonds between two strands, while two parallel strands lead to smaller hydrogen-bond angles.

β-sheets were introduced by William Astbury[5] and later refined by Linus Pauling and Robert Corey[54], based on the proposed existence of the hydrogen-bonding pattern between two or more β-strands. Thus, the concept of an isolated strand without any neighbor strands is not covered by this definition.

It has been pointed out by Fitzkee et al.[55] that it is a common misconception to believe that the inter-strand hydrogen bonds are necessary to stabilize a β-strand. In fact, even though they are usually not considered to form a hydrogen bond, the N-H and C=O groups of the same residue are almost parallel. As Maccallum and Ho demonstrated, the β-strand basin can be explained with the energetic minimum of a combined Lennard-Jones and electrostatic potential even without considering hydrogen bonds.[56]–[58]

Avbelj argued that electrostatic interactions are more relevant than the hydrophobic effect and conformational entropy in determining the secondary structure of a protein[59]. However, also the chain entropy favors the stretched conformation, again without the need for inter-strand hydrogen bonds[55]. The preference of the extended backbone conformation for β-branched amino acids (such as Val, Ile and Thr) and aromatic residues (such as Tyr, Phe and Trp) is due to steric clashes between backbone and possible side-chain conformers that may occur in a helical conformation, leading to a loss of conformational entropy[60]–[62].

Isolated β-strands are only rarely assigned by secondary structure assignment tools. If they are assigned, the discrimination between them and the PII conformation (a motif with similar ϕ, ψ angles and missing backbone-backbone hydrogen bonds; see next section) ambiguous.

It is important not to confuse isolated β-strands with isolated β-bridges. Isolated β-bridges consist of two hydrogen bonds that match all criteria for regular inter β-strand hydrogen bonds, but they are not repeated. In contrast, isolated strands do not possess any such hydrogen bonds by definition.

Polyproline Helix

The polyproline PII helix (PII) is a left-handed helical motif that differs from the better known α-, 3_{10}-, and π-helices in that it is not stabilized by backbone-backbone hydrogen bonds. Owing to a side chain covalently bonded to the backbones nitrogen atom, proline cannot act as a hydrogen-bond donor. This makes proline incompatible with the hydrogen-bonded helices;

however, proline has a high propensity to be found in a PII helix. Despite the name "polyproline", all amino acids can adopt this conformation.

Maximization of the chain entropy[55], [63] and minimization of the bulk solvent disturbance[64], [65] are both considered to be a source of the stability of the PII helix. Because side chain-to-backbone interactions in PII conformations were found to be mostly non-local and due to their high solvent-exposure, it has been suggested that PII helices are involved in protein-protein interaction[66], [67].

As will be shown in section "Isolated Strands and Polyproline Helix" in more detail, the PII helix and the isolated β-strand are very similar as both conformations are not backbone hydrogen-bonded and both possess backbone torsion angles corresponding to the upper left area of the Ramachandran plot.

Turns

A turn can be defined as a short motif, where the protein backbone's direction changes. Such a sharp turn in the polypeptide overall direction occurs naturally when two antiparallel β-strands are connected by a link of two or three residues, but is not restricted to these locations. The existence of a backbone hydrogen bond ($i + n \rightarrow i$) between the N-H($i + n$) and the C=O(i) group of two sequentially close residues i and $i + n$ is a frequent feature in turns. The existence of such an hydrogen-bond pattern is required by the definition of "turn", i.e., non hydrogen-bonded turns are not always considered a valid secondary structure class. If they are, these hydrogen-bond free motifs are referred to as "open" turns[68]. Based on the value of n, turns can be further classified (Table 4).

n	turn-type
-5	π
-4	α
-3	β
-2	γ
1	δ
2	ε

Table 4: Turn-Names

The backbone dihedral angles can be used to further discriminate turns into several subclasses. What should be noted is the fact that an α-, 3_{10}-, or π-helix can be considered as multiple conjoined turns of the type α-, β-, or π-, respectively. However, a turn is not necessarily a valid building block of a longer helix, as the existing hydrogen bond still allows for larger variations in ϕ, ψ angle pairs than those occurring in helical configurations.

Coil

The coil class comprises all residues that have not been assigned to any of the defined secondary structure classes. Thus, it should not be considered as a real class, but merely as the "other" category. Therefore, what is actually contained in the coil category depends on what secondary structure motifs are assigned to the regular classes. PII helices and turns are two motifs that are regularly labeled as coil. As a prominent example, the Protein Coil Library[69] defines as coil every residue with ϕ, ψ angles other than those from strand or helix.

Amino Acid Preferences

Different amino acids possess diverse physicochemical behavior, resulting in a specific propensity to assume a given secondary structure. This is extensively exploited in secondary structure prediction. However, a prediction solely based on the single-residue level yields poor accuracy.

Especially the two amino acids glycine and proline have a very distinct influence on a protein's secondary structure. Both are known for their potential of starting or ending a helix, rather than being part of one. The lack of a side chain allows glycine to adopt a larger range of backbone torsion angles. Hence, glycine is often found in turn regions. In contrast, proline is not capable of acting as a backbone hydrogen-bond donor due to its missing polar hydrogen atom N-H. Instead, its side chain is covalently bound to the backbone atom N, this reducing the degree of freedom for the ϕ, ψ angles of proline as well as for the preceding residue.

Helices tend to contain many methionine, alanine, leucine, glutamate, and lysine residues, while β-strands rich with tryptophan, tyrosine and phenylalanine, isoleucine, valine, and threonine[70]. Proline is often found in isolated β-strands or in the PII conformation due its limited number of allowed (ψ, ϕ) combinations and its inability to act as a backbone hydrogen-bond donor.

Secondary Structure Assignment

Several approaches to assign secondary structure to a given protein structure have been proposed. They can be loosely separated into the following categories based on hydrogen bonding, backbone dihedral angle, Cα, or, more generally, geometry. Usually, a combination of these approaches is used. It should be noted that the methods based on hydrogen bonding and backbone dihedral angles require input data of high-resolution protein structure. For instance, the dihedral angles can only be calculated when the coordinates of all backbone heavy atoms are known. Due to the low flexibility of the amide plane, these coordinates are also sufficient to infer the location of the hydrogen atoms and subsequently of the backbone hydrogen bonds. Methods that are only considering Cα atoms may be applied on protein structures resolved with lower resolution.

Hydrogen bond-based

The approach used by DSSP[27] and stride[71] closely resembles the original definition of α-helices and β-strands. First, the probable positions of hydrogen atoms are generated from backbone heavy-atom coordinates. After this initial step, an electrostatic (for DSSP) or empirical (for stride) energy function is used to estimate the energy of possible hydrogen bonds. The network formed by all hydrogen bonds with energies below a given threshold is then analyzed for repeating patterns indicating helical (3_{10}, α-, or π-helical) or β-sheet residues.

Dihedral-Angle based

PROSS[62], [72] is the most noteworthy software for secondary structure assignment based on dihedral angles, which only uses the two angles ϕ and ψ for its assignment. Depending on the user's selection, either a course or a fine grained grid can be employed. Each (ϕ, ψ) pair is mapped to a specific character code by this map. A stretch of five or more contiguous residues that belong to helical grid bins are considered helical. In the same way, strands are assigned when at least three residues belong to strand-like bins. A larger number of (ϕ, ψ)-quadruplets (i.e., ϕ, ψ angles of two consecutive residues) is considered as indicating a turn. In a last step, unassigned residues are considered as PII-helical, if they possess angles from the PII basin. The coil class consists of remaining unassigned residues.

Cα-based

Because for some protein structures only low-resolution X-ray data exists and for results of longer molecular dynamics simulations often only Cα atoms are stored, occasionally it may be necessary or desired to assign secondary structure by only using information that can be derived from Cα atoms. One example of such tools is SABA[73], whose authors claim an impressive agreement with DSSP assignment by extrapolating donor and acceptor groups in a protein from Cα coordinates. DSSP itself uses two pseudo-bond angles based on Cα atoms in addition to the hydrogen bond information.

VoTAP (Voronoï Tessellation Assignment Procedure) [74] uses an interesting approach for secondary structure assignment that circumvents the frequent problem of setting cutoffs for distances or energies by defining a contact between two residues in terms of a shared face of the two residue's Voronoï cells. A residue contact map is constructed according to this neighborhood definition, and secondary structure may then be assigned in a very similar way to that used in DSSP or stride.

Geometry-based

The software P-Curve[75] calculates a helicoidal axis based on differential geometry. The helical parameters (radius, tilting, rolling, and twisting) are then used to find the secondary structure motif that best fits the calculated values. While mainly using Cα coordinate information, this software is usually considered to be using a rather extraordinary approach to structural assignment. Another example for a closely related approach that uses quaternions was proposed by Hanson[76].

Discussion

In the left part of Figure 9, a selection of 500 randomly selected data points for each of the four helix classes is shown. Here, it can be seen that the three right-handed helix types share the same maximum density at the location of the α-helix conformation around $(-64°, -40°)$. For the α-helix, the distribution is strictly located to this region. 3_{10} helical residues are more widely spread to lower ϕ and higher ψ values, while the π-helix distribution has a high density at lower ϕ and lower ψ values. However, due to the large overlap of these three classes, a reliable separation by ϕ, ψ values alone cannot be expected. In contrast, left-handed helices need to be identified by some other means than only the hydrogen-bond pattern. The majority (86%) of

the left-handed helices possesses 3_{10}-helical hydrogen bonds and are therefore assigend to the class G by DSSP.

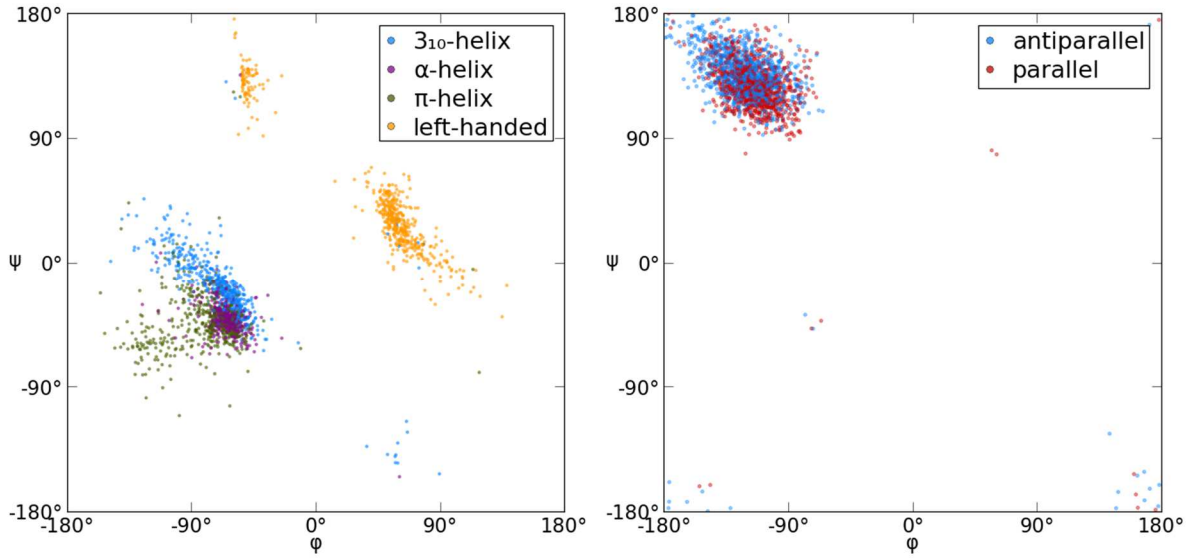

Figure 9: Ramachandran plots of helix and strand residues. Left side: Ramachandran plot of helical residues from the ASTRAL40 dataset. The same number of residues (500) has been chosen for each class to visualize the spread of the distribution rather than the absolute occurrence. Right side: Ramchandran plot of 500 parallel and 500 antiparallel resdiues randomly chosen from the same dataset. Only residues with two parallel or two antiparallel neighbors are shown.

The right part of Figure 9 shows a Ramachandran plot for residues with two parallel or two antiparallel neighbor β-strands. From both sets, 500 residues where selected randomly. While there is a tendecy for antiparallel β-strand residues to be shifted to the upper left corner of the (ϕ, ψ)-map compared to the parallel β-strand residues, the two distributions are clearly not separable by means of the (ϕ, ψ)-angles. For residues at the edge of a β-sheet that only posses one β-strand neighbor as well as for mixed residues with one parallel and one antiparallel β-strand neighbor, the trend is even less pronounced (data not shown).

The polyproline helix PII is a secondary structure motif that, due to a lack of hydrogen bonds, clearly needs to be assigend by means of (ϕ, ψ)-angles or possibly some pseudo-bond geometric criteria.

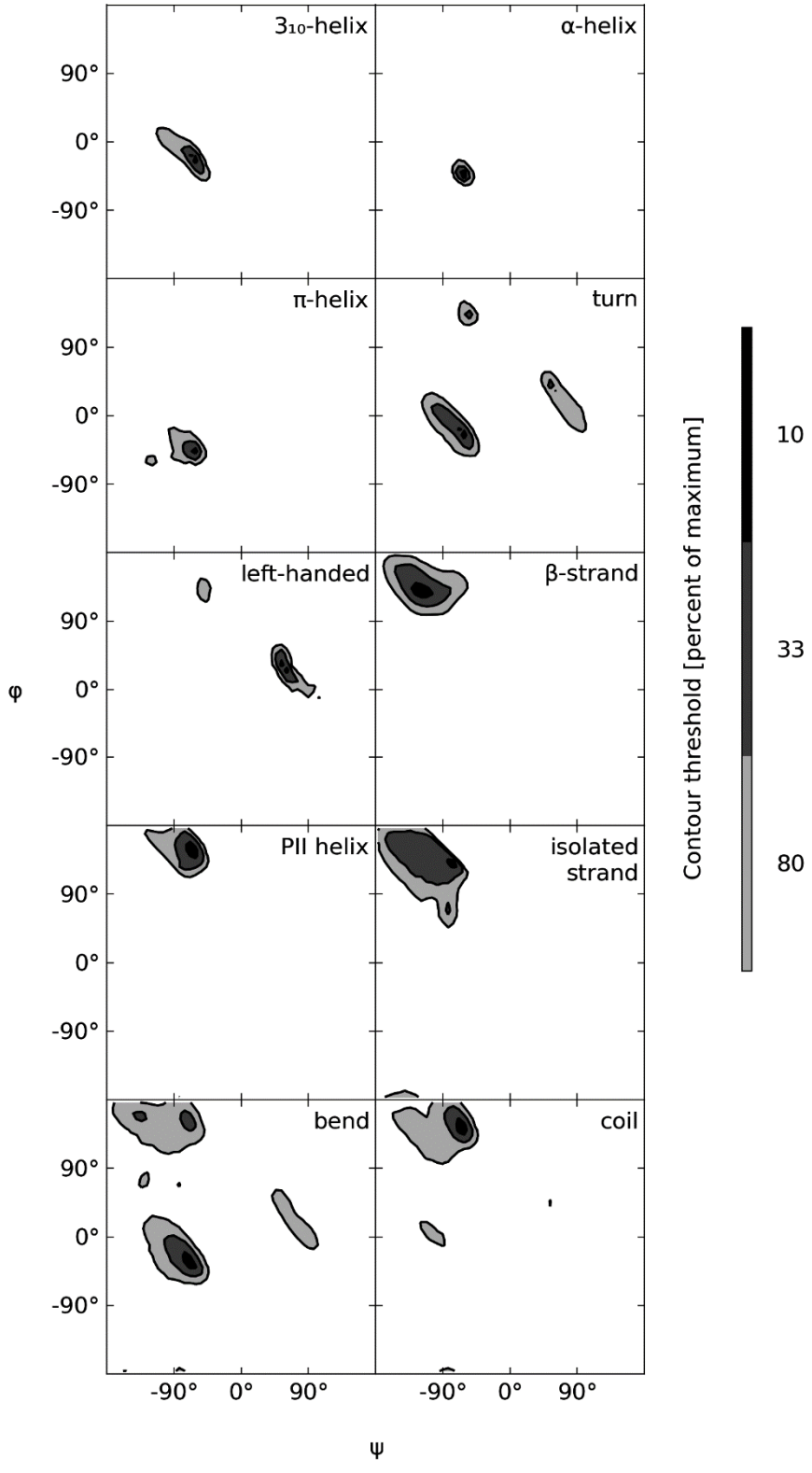

Figure 10: Most probable Ramachandran regions for the various secondary structure elements. The complete ASTRAL40 dataset has been classified and assigned to the ten default classes of PSSC.

Development of PSSC

Even though DSSP has been introduced 30 years ago, it is still the most cited and employed assignment tool, followed by the stride software. In the past years, a number of alternative secondary structure assignment tools have been developed, and a comparison of different methods has been carried out by Martin et al in 2005[77]. However, most of these tools are no longer actively developed; many are not even available, and, for some programs, the source code has never been made accessible. In contrast, DSSP has been rewritten in 2011 by Maarten Hekkelman using C/C++, including the Boost library[78], and released under the Boost open-source license.

PSSC exists in three different versions. The version used in publication[79] is a python script that parses the output from the original DSSP. This approach was the most convenient way to analyze new classification themes based on the same hydrogen-bond definition as used by DSSP. The usage of a script language from an integrated development environment that allows for instantaneous plotting and visualization of intermediate results proved to be a sensible choice during the development process. For an end-user however, the usability as well as the execution time of the software are the primary concerns; hence, a second version of PSSC was created as a so-called fork of DSSP.

Due to the development in a modern, standardized, object-oriented and generic language in combination with the countless additional features that the Boost library provides, the DSSP source is reasonably well-structured and maintainable, allowing for an easy application of changes on the original approach as well as addition of new functionalities. Thus, the second version of PSSC uses the source code of the established DSSP software as the starting point, but offers additional options such as the possibility to use the more realistic electrostatic energy term with arbitrary energy cutoffs (eq. 6).

The final version of PSSC is a complete new development that reuses much of the original DSSP algorithm, but addresses some issues in protein secondary structure assignment such as interrupted β-sheet ladders in a more elegant way.

Differences between DSSP and PSSC

DSSP and PSSC first analyze the probable hydrogen-bonding pattern of a given protein structure. Two classes of hydrogen-bond motifs exist that lead to the assignment of hydrogen-bonded secondary structure: the turn-like hydrogen bonds and the β-bridges. For both classes, the assignment ignores which one are the hydrogen-bond donor and acceptor residues and can only be applied to the residues that are bridged by the hydrogen bond under consideration.

Because the consistent approach of DSSP for assigning β-strands and α-helices by hydrogen-bond patterns is considered by the community as the standard, this part of DSSP was very carefully modified in PSSC. Except for the introduction of the electrostatic energy function for the hydrogen bonds and the possibility to use different cutoff energies, this part of the algorithm remains unchanged. However, the reduction to a single secondary class by DSSP conceals the fact that combinations of different helix types are frequent in proteins. By not addressing the backbone torsion angles (ϕ, ψ) during the secondary structures assignment, not or only weakly hydrogen bonded capping motifs of regular structures can stay unnoticed.

Hydrogens and Hydrogen Bonds

As the assignment of hydrogen bonds is the basic step for secondary structure assignment with DSSP and PSSC, this part has been redesigned with particular care in the PSSC versions that are not dependent on the DSSP's output.

In macromolecular-structure data derived from X-ray diffraction, the coordinates of hydrogen atoms are usually not resolved. This is obviously a serious problem for the identification of hydrogen bonds. DSSP and PSSC tackle this problem by placing explicit polar backbone hydrogen atoms in the protein structure. For every residue (except for proline, which does not possess such a hydrogen due to its cyclic side chain), the vector $\vec{h} := \overrightarrow{OC}/|\overrightarrow{OC}|$, connecting the backbone oxygen and the C'-atom of the sequentially preceding residue, is calculated. When $\vec{h}$ is translated so that it starts at the nitrogen atom, it points approximately to the coordinates of the polar hydrogen. Due to the very rigid bond angles and ω dihedral angles in the peptide plane, this is a reasonable approximation.

Note, however, that this procedure is only valid for residues in trans conformation. For non-proline cis residues, this places the hydrogen too far from the nitrogen. These non-proline residues in cis conformation are quite rare, but they do occur often enough [41], [80] to be treated correctly. When PSSC needs to assign a hydrogen atom to a residue that possesses a ω-dihedral angle below 90°, this residue is considered a cis residue[3]. The vector $\vec{h} := \overrightarrow{OC}/|\overrightarrow{OC}|$, is thus replaced with the vector $\overrightarrow{h'} := \overrightarrow{OC\alpha}/|\overrightarrow{OC\alpha}|$, and the calculation is carried out as previously discussed.

A new feature of PSSC provides the possibility of using the hydrogen atoms that are already present in the PDB file. This can be useful for high-resolution X-ray data and molecular dynamics simulation results. This feature further allowed to investigate the influence of more advanced hydrogen atom placement methods in addition to the previously described default procedure.

Efficient evaluation of Hydrogen-bonded residue pairs

In order to find pairs of interacting residues in a protein with given coordinates, the well-known fixed-radius near-neighbor search problem needs to be solved. Various algorithms can be applied that differ in algorithmic simplicity and run time. Each residue can interact with any residue in its neighborhood. A distance threshold of 10 Å is a reasonable value for the maximal distance between the Cα atoms of two interacting residues. All pairs found to be in close contact by this criterion need to be evaluated in more detail. In PSSC, an electrostatic energy calculation of the backbone atoms is performed to discriminate between hydrogen-bonded and non-hydrogen-bonded pairs.

The brute force or naïve approach calculates all distances between all possible pairs of the n residues. Thus, a number of $n(n-1)/2$ distances needs to be calculated, including those between residues being arbitrarily far from each other. This approach is used in the original version of DSSP. Being clearly the easiest algorithm to implement, the number of distances to be calculated grows quadratically with the number of residues.

[3] The value of 90° has been chosen because it lies in the very low populated region of the angle ω. The large majority of residues possesses ω close to 180° for trans and 0° for cis conformations.

The k-d tree[81] is a data structure that is employed by many programs and is included in numerous libraries. This structure organizes k-dimensional data points by means of a binary-space partitioning tree. The worst-case time complexity of a range search in a three-dimensional k-d tree is of the order $O(3N^{2/3})$, and construction can be done in $O(3\ N \log N)$ time[82]. Arbitrary ranges (i.e., distances for neighbor search) can be used for any given k-d tree.

A very elegant and surprisingly simple algorithm (Figure 11) is given in ref. [83]. Here, the maximal distance l for the neighbor search has to be given in advance. The x, y, and z coordinates are quantized to form a grid of cells of width by computing $i = \text{floor}(x / l)$, $j = \text{floor}(y / l)$, and $k = \text{floor}(z / l)$, where i, j, and k are integers, and the floor function maps a real number to the largest previous integer.

The integer grid coordinates i, j, k are then mapped to a fixed-size list (the hash table) of length n via the hash function $h(i, j, k) = (i \times p_1 \text{ xor } j \times p_2 \text{ xor } k \times p_3) \bmod n$.

Here p_1, p_2, p_3 are large, arbitrarily chosen prime numbers[4]. Every bucket of the hash table stores the real coordinates of the points for which the hash function h gave the integer value corresponding to the bucket's position in the list. Due to the so-called hash-collisions, a bucket may store points that are not neighbored in real space.

A neighbor search for distance l for any given coordinates x, y, z is performed by first calculating the corresponding grid cell i, j, k. For this grid cell and all its 26 direct neighbors in the grid, the hash value $h(i, j, k)$ is calculated, and the precise distances from all points found in these cells to the point (x, y, z) are computed.

Because multiple grid cells may lead to the same hash value, it is theoretically possible that all points are stored in the same hash bucket. Hence, the worst-case complexity for a single search has the complexity $O(n)$. However, the given hash function is known to work steadily enough to ensure amortized constant time. $O(n)$ time is also needed for the creation of the hash map[83].

[4] The alleged prime numbers 73856093, 19349663, and 83492791 were suggested in the original publication[83]. However, 41 × 471943 = 19349663 holds, so the value proposed for p_2 is not prime. Most likely, the last digit has been misprinted, as 19349669 is the nearest larger prime number. However, this does not seem to significantly influence the efficiency of the algorithm.

The implementation in C++ proves to be rather simple, as the new C++ Standard Library provides the std::unordered_map template, an unordered associative container that allows for custom user defined hash functions.

This algorithm is also applied for the calculation of the solvent excluded area (for further details, see next chapter.)

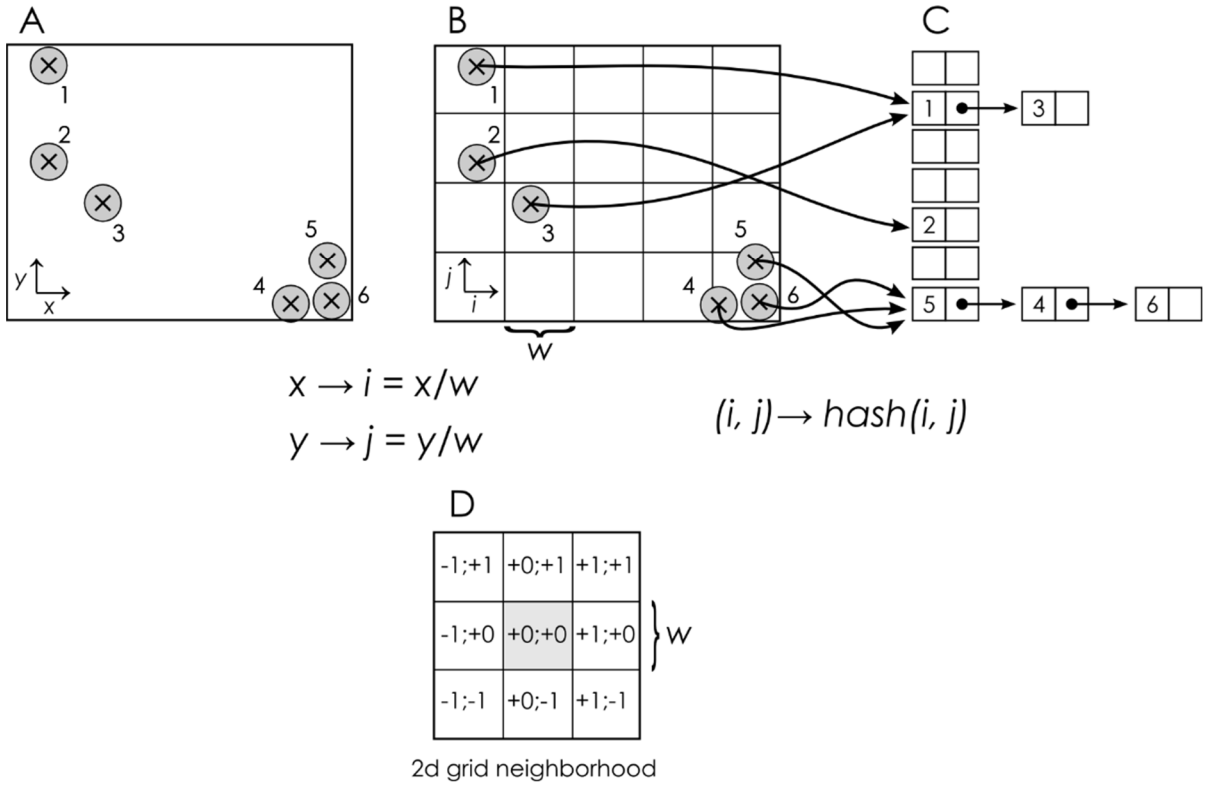

Figure 11: Geometric hash algorithm displayed for two dimensional data. The maximal search distance equals the lattice constant w. A: 6 points with given (x, y)-coordinates. B: After transformation to grid coordinates (i, j), every data point is associated with a grid cell. C: For every non-empty grid cell the value of the hash function determines in which bucket of the list of size n the points are stored. Here, points 1 and 3 are stored in different buckets, but for point 1 and 3, a hash collision occurs: both points are stored in the same bucket, despite their large separation in real space. Points 4, 5, and 6 belong to the same grid cell and thus are automatically stored in the same bucket. Implementation as a singly linked list is indicated in this scheme, but, clearly, every dynamically growing data structure is suitable. D: Neighborhood in terms of grid coordinates. For every neighbor search, the hash values of the nine directly neighbored grid cells (including the gray center cell) have to be calculated. Finally, all points stored in the corresponding buckets of the hash list need to be evaluated.

Solvent Accessible Area Calculation

The Solvent Accessible Area (SASA) is the surface area of a protein that is accessible to solvent molecules. For residues in the outer region of the protein, this surface is roughly proportional to the number of water molecules in the first hydration shell[27].

To calculate the SASA (Figure 12), a number of points is drawn on the surface of a sphere around each atom with radius R, where R is the sum of the atom's van der Waals radius and

the effective radius of a water molecule (usually 1.4 Å is used for the latter). Points that are inside the volume of another atom of the protein are removed, and the remaining points sample the SASA.

Several algorithms exist to arrange points approximately equidistant on the surface of a sphere. The current version of DSSP uses a version of the golden section spiral algorithm (see for example ref. [84]. It should be noted that DSSP originally used a different approach based on recursive divisions of a polyhedron[27].) The same point distribution is used in PSSC in order to gain the same numerical values for solvent accessibility. To guarantee agreement, the same atomic radii for backbone and side chain atoms need to be used. In PSSC, the possible intersections of surface points with other atoms are checked with the spatial hashing algorithm.

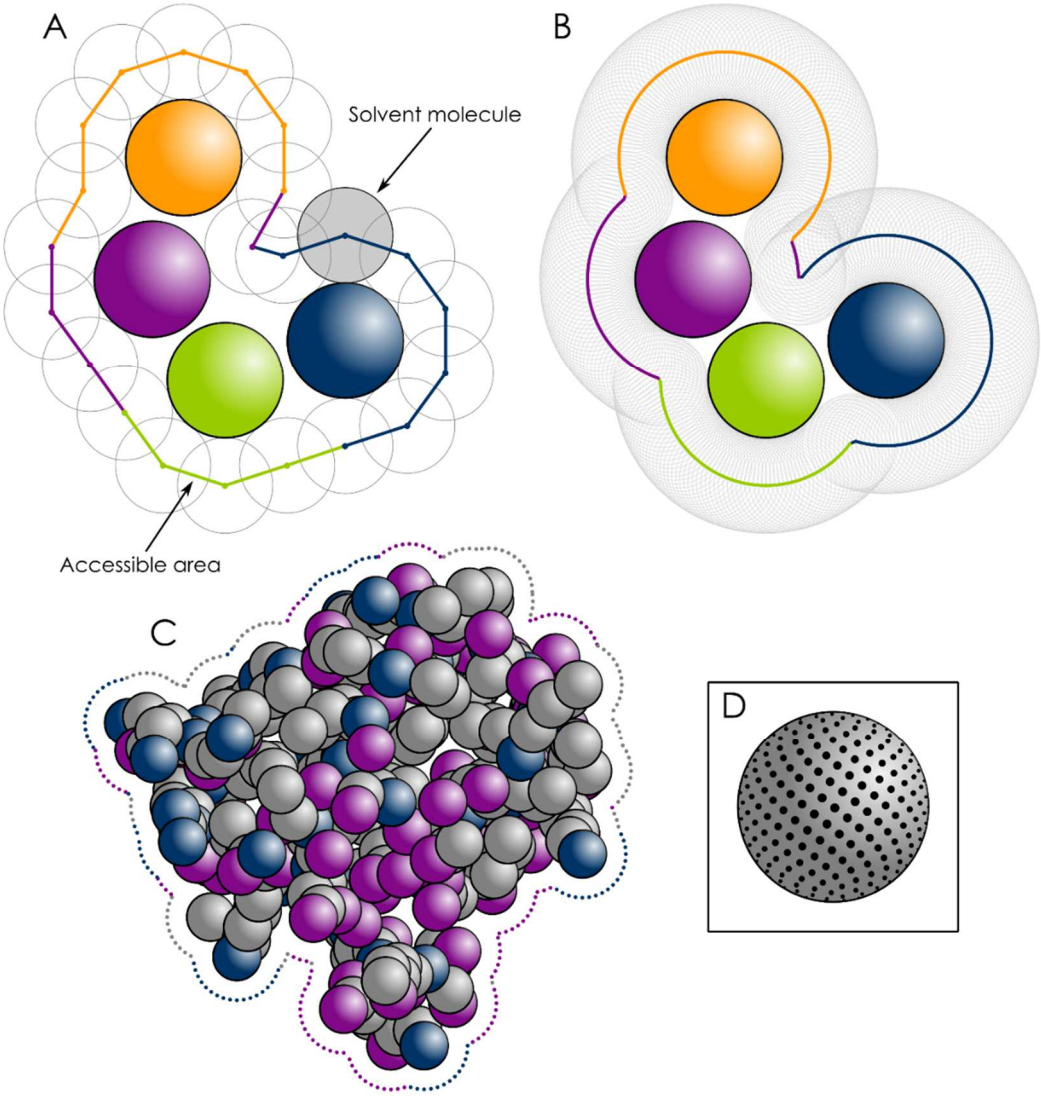

Figure 12: Accessible Area calculation.

Secondary Structure Assignment with PSSC

Turns and Helices

Backbone hydrogen bonds connecting two residues sequentially neighbored at positions and i (with C=O) and $i + n$ (with N-H), where n is 3, 4, or 5, are referred to as turn-like hydrogen bonds. They correspond to the three possible helix-types 3_{10}, α, and π. Two turn-like hydrogen bonds of the same type (i.e., same n-value) give rise to the assignment of the respective helix class for the residues that are bridged by these hydrogen bonds. If a turn-like hydrogen bond appears isolated, DSSP assigns the bridged residues to the general turn class that does not discriminate regarding the different n-values. To account for irregularities in longer helices, DSSP allows for missing turn-like hydrogen bonds as long as no more than three residues are affected.

Bridges and Strands

Similarly to turns, consecutive β-bridges are combined and give rise to the assignment of an extended β-strand, while isolated β-bridges give rise to the assignment to the β-bridge class. The assignment applies to the residues that are "bridged". Hence, it is not mandatory that a strand- or β-bridge residue is partaking in any hydrogen bond itself, as this is only required for its direct preceding and following residues.

Seven Building Blocks of Hydrogen-Bonded Secondary Structure

It is illustrative to discuss in more detail the basic hydrogen-bond pattern that leads to secondary structure assignment. The two motif classes n-turns and β-bridges can be described by seven distinct patterns, if n is restricted to the three values 3, 4, and 5. The patterns are shown schematically in Figure 13, with the hydrogen bonds being represented by arrows from donor to acceptor residues.

β-bridges are a cooperative pattern of two hydrogen bonds between two consecutive patches of three residues, i.e., an internal connectivity via covalent backbone bonds is mandatory for both involved polypeptide segments separately. A connection between these, other than by hydrogen bonding, is not needed. Hence, the two patches can also belong to different chains. A distinction between inter or intrachain β-bridges has been proposed in some methods[85].

The two polypeptide segments can be oriented parallel or antiparallel with respect to each other—corresponding to parallel or antiparallel β-strand arrangements. The repeated β-bridge

pattern can be regarded as a cyclic path through the structure: If the hydrogen bonds are traversed from donor to acceptor and the backbone is followed from N to C terminus, a full cycle with 2, 4, or 6 nodes can be described.

In the antiparallel and parallel cases, two different types of β-bridges need to be considered. In the antiparallel case, the distinction is obvious, as in the first case (termed here as "antiparallel long cycle"), the outer four residues (termed as a, c, s, and u in Figure 13) are hydrogen-bonded, and in the second antiparallel motif only two residues (c and s in Figure 13) of the two polypeptide segments are involved in hydrogen bonds. Hence, the latter case is termed as "short antiparallel cycle".

The two possible cycles in parallel β-bridges consist of four residues, with one outgoing and one ingoing hydrogen bond from the central residue of one strand. Interestingly, the two parallel cycles only differ in their starting points. If in the parallel cycle 1 (Figure 13 P), the first residue visited is q instead of a, this cycle is considered to be of type 2.

Data in Figure 13 show that in a longer strand, a long cycle must be followed by a short cycle in the antiparallel case, and a parallel cycle of type 2 follows a cycle of type 1 (and vice versa). Exactly one hydrogen bond is shared by two consecutive cycles of different types. In DSSP and PSSC, such consecutive cycles are summed up into β-ladders. Only the two residues in the middle of the two three-residue windows (displayed in bold in Figure 13) are later considered to be of β-strand or β-bridge class. Thus, each of the four different cycles corresponds to two β-strand or β-bridge residues—one on both connected β-strands. Ladders consisting of a single β-bridge lead to a DSSP and PSSC assignment of the β-bridge class B, and longer ladders lead to the assignment of the β-strand class E. Because of the special treatment of singular bridges and consecutive bridges overlap, singular residues of class E are not feasible, i.e., a residue of class E is always followed or preceded by another β-strand residue.

A precise assignment of the β-strand or isolated β-bridge class to the two "bridged" residues in the center of a cycle can be motivated by considering that all cycles involve two ϕ and two ψ angles, one pair on each β-strand. These dihedral angles only belong to the center residues. The (ϕ, ψ) dihedral angles of the outer residues are not directly restricted by the constraint of

an existing hydrogen bond. Thus, this hydrogen-bond-based definition reflects the conformational properties of β-strand residues. This is also clear from Figure 8, where all atoms participating in the hydrogen bond pattern are shown.

Compared to β-strand cycles, the *n*-turn is a much simpler motif to define and detect. In particular, *n*-turns differ from β-bridges, as they involve only a single hydrogen bond. Each hydrogen bond from a donor residue $i + n$ to an acceptor residue i is a turn-like hydrogen bond. Also in this case, the donating and accepting residues are not assigned to the turn class, only the residues "bridged" by the turn are. Two consecutive turns of the same type give rise to the assignment of the corresponding helix class (classes G, H, and I for 3_{10}, α-, and π-helix for n = 3, 4, and 5, respectively) instead of the turn class. This helix definition results in a minimal helix length of n. Hence, α-helices of length 3 cannot exist. However, shorter helix segments may be assigned by PSSC, if different helix motives overlap in the sequence. In contrast, DSSP may neglect helices in such cases[52].

Every intrachain hydrogen bond may be clearly considered as a *n*-turn, but, due to geometric and energetic reasons, proteins cannot form true helix structures for values of $n = 2$ or $n \geq 6$. The occurrence of weak intra-residue hydrogen bonds ($n = 0$) is known[86], [87] to stabilize β-strands, but this feature is not needed for the current definition of β-strands.

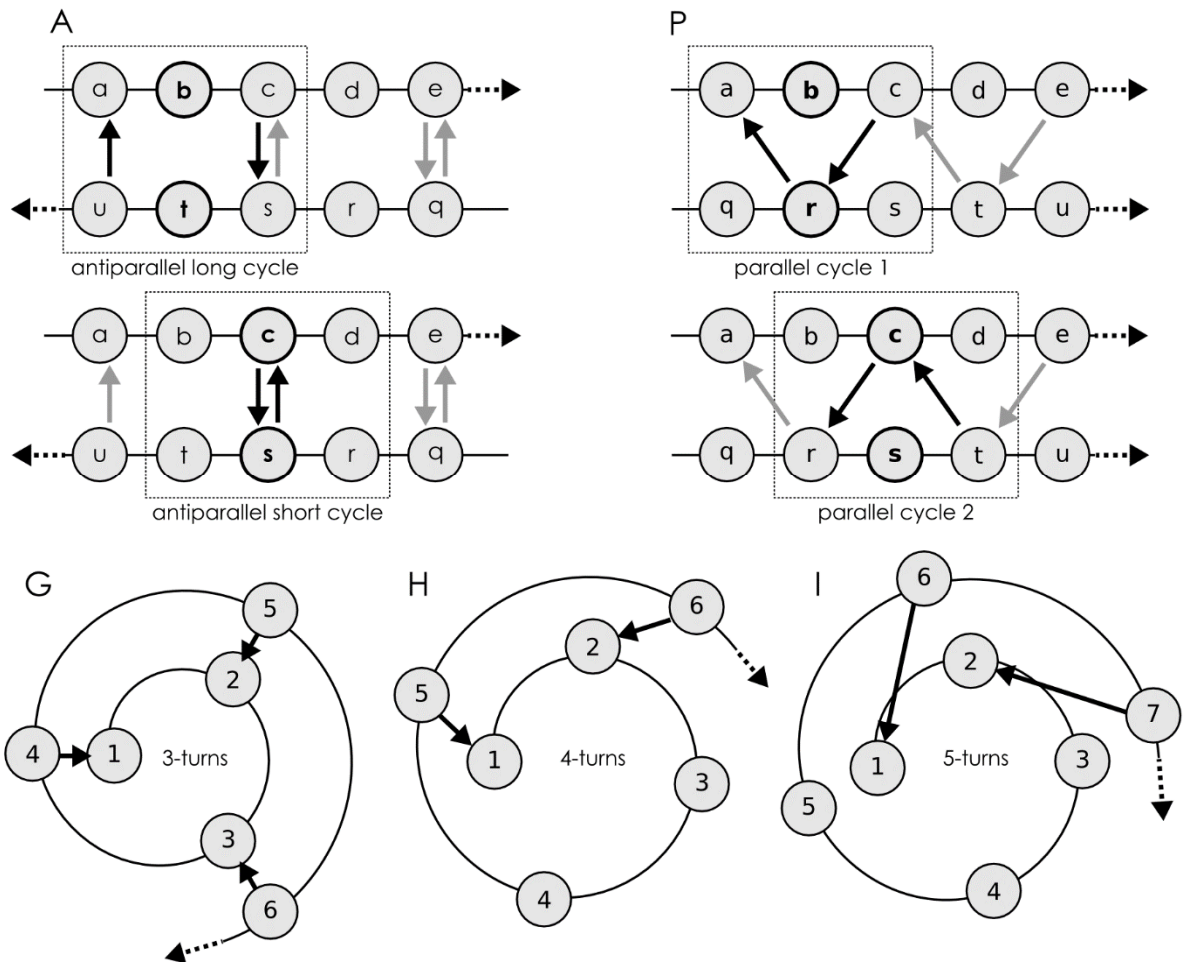

Figure 13: The seven basic Hydrogen-Bond Patterns.

Coils and Bents

The previously introduced pseudo-bond angle $\varkappa$ (Figure 5) is calculated for all residues. Values of $\varkappa < 110°$ indicate a bent structure and, residues that have not been assigned to any hydrogen-bonded class, are consequently assigned to the bent class (B).

It is arguable whether the coil class should be considered as a true secondary structure class. The authors of the original DSSP suggest that this is not the case. Residues that do not possess any of the structural features used by the DSSP-assignment strategy, are left with a blank symbol in the DSSP output. However, in many studies that use DSSP for structural assignment, the blank is replaced by the letter C. Although this may just be done to increase readability, this suggests the interpretation as a real class. This may be justified by the fact that in DSSP only those residues that are *not* bent are assigned to the coil class. Hence, the coil class represents residues that are part of a rather stretched conformation.

Assessment of Dihedral Angles

DSSP calculates the dihedral angles ϕ and ψ and includes the values in its output, but does not actually use them. However, the dihedral angles may be particularly useful to assign PII helices and/or isolated β-strands and for an explicit treatment of left-handed helices. The sparse left-handed helices share the same hydrogen-bonding pattern of the common right-handed helices, but differ in the handedness. Additionally, β-strands may be extended by one or two residues that are not hydrogen-bonded, but have (ϕ, ψ) angles in the β-strand region. This assignment strategy idea has been employed at least in the secondary structure assignment software stride[71].

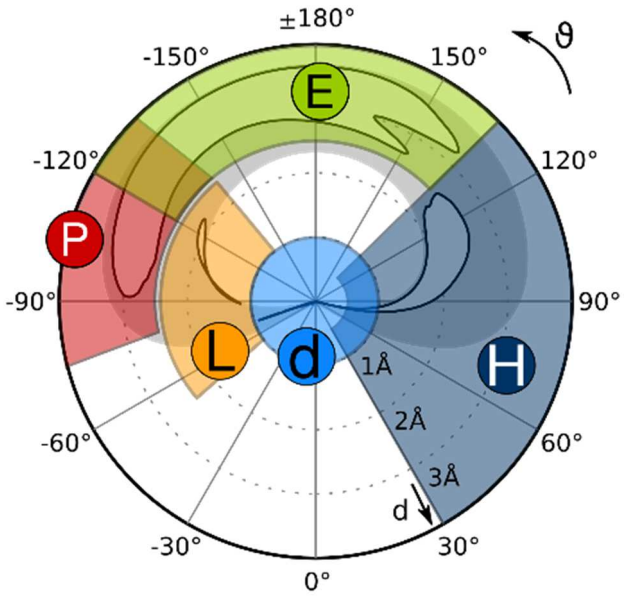

Figure 14: The (d, ϑ)-plot in use for secondary structure assignment. The five colored regions correspond to the respective secondary structure class: H: helical, E: β-strand, P: PII helix, L: left handed helices: d: small d values. Backbone dihedral angles corresponding to the d-region can occur in helices disturbed regions. Overlaps exist between the PII helix and the β-strand regions as well as between the helical and the d-region.

PSSC does not use the (ϕ, ψ) dihedral angles directly. However, the previously described and published[88] (*d*, ϑ) values are used. In the (*d*, ϑ) space, a separation between the regions typically occupied by specific secondary structure motifs is more obvious then with the traditional (ϕ, ψ)-plot. Additionally, the pseudo-dihedral *α* (Figure 6) with its problematic non-symmetric definition is obsolete, as the handedness information is readily provided on a per-residue level. Figure 14 provides the five different regions of (*d*, ϑ) values that PSSC uses for its dihedral angle-based secondary structure assignment. Each residue is assigned a dihedral code (DSSH) according to the area of the corresponding (*d*, ϑ) values. A lower-case letter is given by default, and an upper-case letter is assigned if the preceding and the following residues in

the sequence belong to the same region. The *d*-region corresponds to residues with very small *d*-values, and hence handedness information is not trustworthy for these residues. Residues belonging to π-helix bulges are frequently found here. Hence, an exception is made for the assignment with the large-case letter "H": if a residue is located in the helical area, its neighbors may also be from the *d*-region to trigger an upgrade to the upper-case letter assignment.

It should be noted that the PII region P overlaps with the β-strand region E as well as the helical overlaps with the *d*-region. For residues found in the overlap areas, the two corresponding characters are set in the DSSH.

Isolated Strands and Polyproline Helix

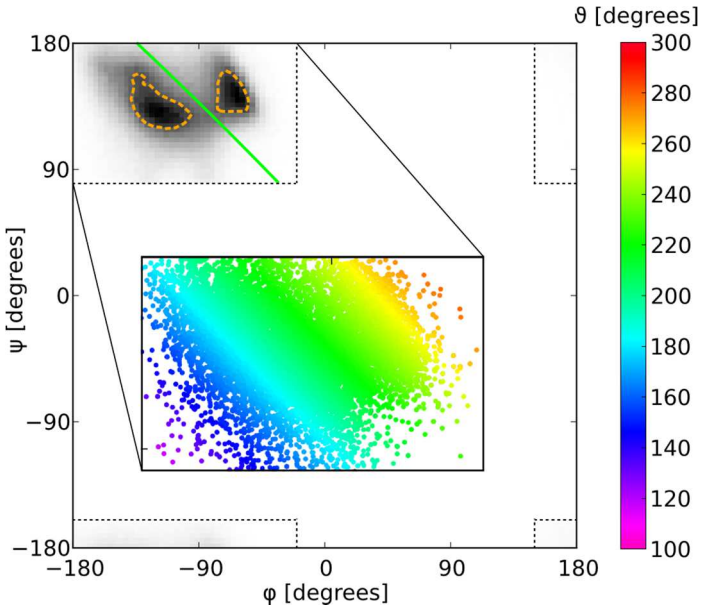

Figure 15: Ramachandran Plot of the β-strand and PII region. Also shown, the density of (φ, ψ) angles of all residues in the Astral40 dataset for 100×100 bins. The dashed orange lines represent the contour lines at 2/3 of the maximum density. The green solid line indicates the n=2.5 or ϑ=216° isoline. In the blow-up, a scatter plot of 10000 randomly selected residues is colored by the ϑ value. Note that values of ϑ > 180° represent left-handed conformations.

The upper left area in the classical Ramachandran plot with angles of about $\varphi \in [-210°; -20°]$, and $\psi \in [80°; 180°]$ is often referred to as "the β-strand region". This oversimplified view is supported by many figures in textbooks and publications that display this area of allowed (φ, ψ) angles as a broad, contiguous region, sometimes not even indicating the PII configuration on the right side[89].

In fact, this area is covered by two rather distinct point clusters as shown in Figure 15. There the regions of highest density are highlighted by two orange dotted lines that connect the points where the density is 2/3 of the maximum value. While the left cluster consists of mostly β-strand residues, the right cluster corresponds to residues with (φ, ψ) angles in the PII conformation. Both clusters are elongated along the diagonal. The $\vartheta = (360° - 144° = 216°)$ line (green solid line in Figure 15), which corresponds to $n = 2.5$ residues per turn, clearly separates the clusters. This can be easily rationalized as β-strands are expected to possess n values close to 2, whereas the ideal PII possesses $n = 3$ residues per turn.

β-strand residues are readily defined (and assigned by PSSC) by their backbone-backbone hydrogen-bond pattern. However, even without strand-like hydrogen bonds, there is an energetic minimum for stretched polypeptide backbone conformations. This has given rise to the definition of isolated β-strands (here referred to as β_0 conformation) by several authors[90], [91]. As the PII and the β_0 conformations both lack hydrogen bonds and can only be defined by means of their torsion angles, the close proximity of the two clusters is clearly problematic. The (φ, ψ) values of the regularly hydrogen-bonded β-strands cannot be simply used to assign residues to the β_0 class as these two classes may be not bound to the same (φ, ψ) region.

Discriminating between Strands, Isolated Strands, and PII Helices

Figure 16 shows histograms that compare the distributions of the r, n, d, ϑ, φ, and ψ values of all stretched non-glycine coil-like residues with residues that are part of β-strands. Here, "stretched" refers to the fact that the residues are taken from the upper-left corner of the Ramachandran map as described in the previous section. Residues are referred to as coil-like when they are not in any hydrogen-bonded class (helical, turn, or β-strand) and are not bent. This last restriction rules out 8.8% of the residues from this region. It should be noted that the plot for n has a non-continuous axis, as values between -2 and 2 are not allowed for n.

For class E, all six distributions of possible strand-selection parameters show a single, pronounced peak at the "optimal" strand location and an unpronounced shoulder at the PII location. The main peaks of all coil distributions (class C) are found at the PII location. Only the n-distribution shows a second maximum corresponding to the isolated strand conformation. This clearly indicates that these parameters are not suitable to separate the two underlying distributions. Hence, only the n and ϑ distributions were further investigated.

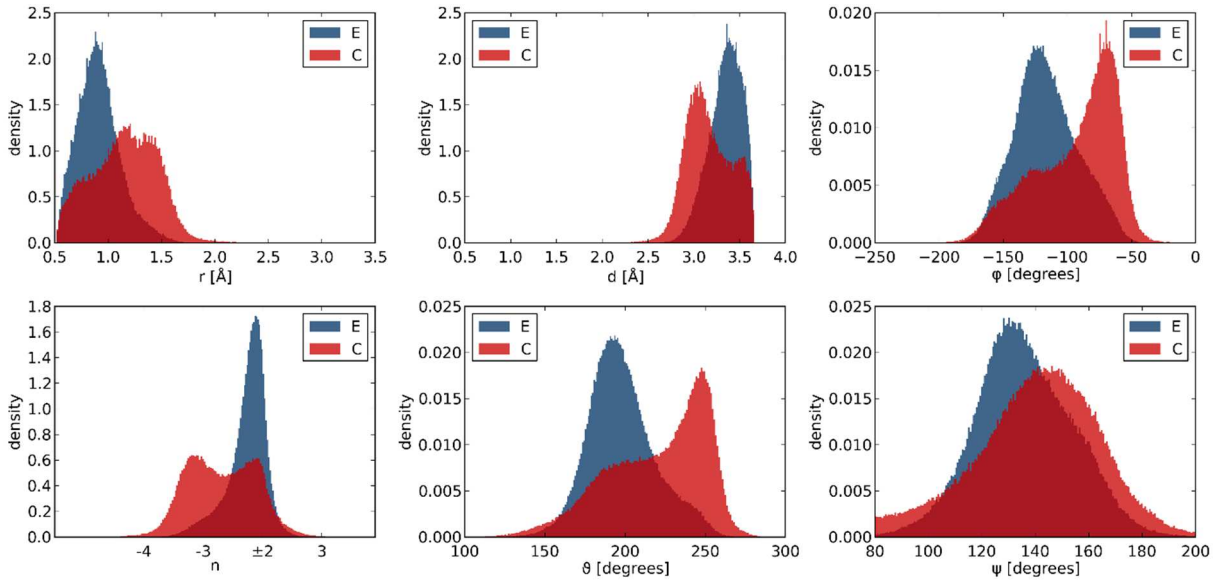

Figure 16: Histograms of the various parameters r, n, d, ϑ, ϕ, and ψ of coil-like and strand residues for the upper left corner Ramachandran plot of the Astral40 dataset. Blue corresponds to β-strand residues (E) and red to coil residues (C). The values are scaled so that the sum over all bins times the bin-width equals to unity. Note that the n-axis is non-continuous at n = 2 as values between -2 and +2 are impossible by construction.

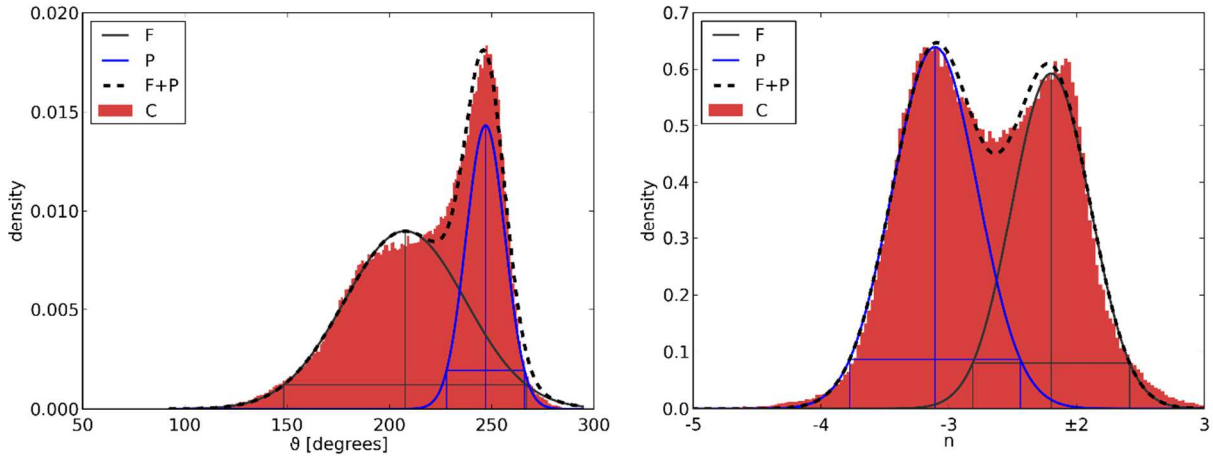

Figure 17: Left: Fitted curves for coil residues in extended conformations. Right: The same data after transformation of ϑ to n. The fit routine was applied after transformation.

Using the non-linear fitting routines from the Gnuplot software[92], Gaussian distributions $f_\alpha(\vartheta)$ and $g_\alpha(n)$ were fitted to the distribution data for ϑ:

$$f_\alpha(\vartheta) = N_\alpha \exp\left(-\frac{(\vartheta - \vartheta_\alpha)^2}{2\sigma_\alpha^{\vartheta}}\right)$$

where α is a placeholder for the secondary structure type considered (E for hydrogen-bonded strands, F for isolated strand residues, or P for PII -helical residues), ϑ_α is the mean value, $\sigma_\alpha^{\vartheta}$ is the standard deviation, and N_α is a normalizing factor. Normalization of the underlying data ensures that the sum over all bins times the bin-width equals to unity. The distribution of the ϑ values for stretched coil residues is described by a sum of two Gaussian distributions:

$$f_C(\vartheta) = f_F(\vartheta) + f_P(\vartheta)$$

Similarly, the n-distributions are fitted with the functions:

$$g_\alpha(n) = N_\alpha^n \exp\left(-\frac{(n - n_\alpha)^2}{2\sigma_\alpha^n}\right)$$

$$g_C(n) = g_F(n) + g_P(n)$$

To address the non-continuous axis problem, values of $n > 2$ are shifted by -4 before fitting. The results of the fitting procedure are given in Table 5, and the resulting function plots shown in Figure 17; the standard errors estimated by the fitting routine are indicated. No significant changes are observed when the bin sizes is varied, or when glycine residues are included in the data. Because the simple two-Gaussian model applied to the measured distributions fits remarkably good (for both the ϑ-distribution and n-distribution), it can be concluded that two separate secondary structural motifs are present, which, however, overlap.

The obtained results show that the broad ϑ-distribution for the isolated strands covers the whole PII distribution within the 2σ-interval. In contrast, two distinct peaks in the n-distribution can be observed that result in two Gaussian functions, which only partially overlap. As the value n gives the number of residues per helical turn, the clustering of the strand-residues around the values of $n = -2.2$ and $n = -3.1$ for the PII residues is close to the expected values for an ideal helical conformation.

For the generation of the PSSC dihedral code, the $(n_\alpha \pm 2\sigma_\alpha^n)$ values are transformed into the corresponding ϑ values: The regular β-strand residues (class E) are expected to be found in the ϑ-regime ranging between -139° and 158.2°; the non-hydrogen-bonded singular strand (class F) appears in the broader range from -128° to 149.1°, and the polyproline residues are expected to range between -148° and -95.4°.

Parameter	Value	Parameter	Value
N_E^ϑ	$2.16(2) \times 10^{-2}$	N_E^n	1.64(3)
ϑ_E	194(1)	n_E	-2.17(1)
σ_E^ϑ	16.7(1)	σ_E^n	0.218(4)
N_F^ϑ	$8.93(9) \times 10^{-2}$	N_F^n	0.593(6)
ϑ_F	208(1)	n_F	-2.20(1)
σ_F^ϑ	29.7(5)	σ_F^n	0.306(1)
N_P^ϑ	$1.43(3) \times 10^{-2}$	N_E^n	0.640(6)
ϑ_P	247(1)	n_P	-3.11(1)
σ_P^ϑ	9.5(2)	σ_P^n	0.334(1)

Table 5: Results for the distribution parameters for the fitted distributions.

Results

Besides the three helix types 3_{10}, α-, and π-helix; the hydrogen-bonded turn; the β-strand and the bend class, the proposed secondary structure classes from PSSC include the PII helix, the isolated strand and the left-handed helix. In case a reduction to fewer categories is needed, the latter three classes, together with the bend class, should be merged with the coil class. Statistics of the occurrence of these secondary structure elements have been carried out for the ASTRAL40 dataset. The most probable regions for each class are displayed in Figure 10; the observed frequencies are given in Table 6.

Description	Code	Occurrence
3_{10}-helix	**G**	5.3%
α-helix	**H**	33%
π-helix	**I**	0.43%
β-strand	**E**	21%
Turn	**T**	11%
Bend	**S**	8.5%
Left-handed helix	**L**	0.081%
Isolated Strand	**F**	4.0%
PII helix	**P**	2.7%
Coil	**C**	14%

Table 6: Frequencies of the secondary structure motifs assigned by PSSC for the ASTRAL40 dataset.

Development of a Web Frontend for PSSC

Usually the results of a secondary structure assignment program are to be subsequently processed by some other software. Only rarely does the end-user access the output of the assignment software directly. Even if the secondary structure for one protein is to be evaluated in more detail, the user most likely will employ additional tools to visualize the protein or its secondary structure motifs. This task is usually performed by another program. Hence, the problem of inter-program communication needs to be addressed.

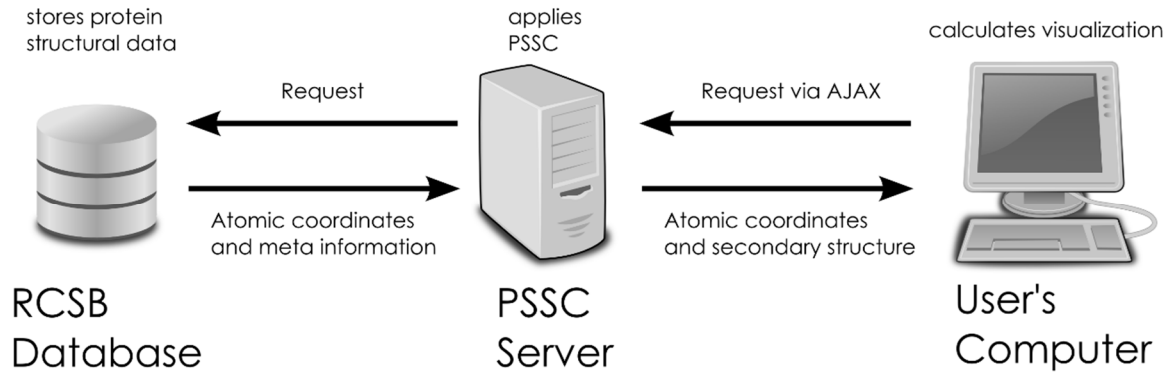

Figure 18: PSSC web frontend.

Interestingly, the output of the current DSSP version still resembles the output format of the initial release of DSSP, which is in human-readable, column-oriented text. Any software that is to use the results from DSSP needs to parse this output text. This strategy is error-prone as there are many details that need to be taken into account such as the non-numerical residue id's in the PDB data format, non-standard residues, gaps in the sequence that lead to incoherent numbering, and very large proteins that need more space for numbering than offered by the column-oriented format.

The parsing of a tab-separated file is slightly simpler to implement and, most importantly, libraries for this task are readily available, and standard office tools can import tab-separated data file. To provide an easily accessible way of receiving the secondary structure of a protein stored in the PDB, a web frontend to PSSC was developed. This is written as a JavaScript application that visualizes the protein's secondary structure in a tabular and in a two-dimensional comic representation of the assigned motifs. Additionally, a Ramachandran plot and the corresponding (d, ϑ)-plot are shown. In addition, an interactive three-dimensional representation is shown using JSmol.

To allow for a smooth interaction between JavaScript and PSSC, an alternative output function was added to PSSC that returns the calculated results in the JSON[93] format. Because JSON is in fact a subset of the JavaScript language, parsing of such files is readily implemented in all current browsers.

Modeling of Hydrogen Positions

PSSC employs two different energy functions for possible hydrogen bonds:

- The classic DSSP energy function (command line parameter "-u DSSP") [code: D]
- The more realistic full electrostatic (-u E) [code: E]

Additionally, PSSC can use two different strategies to model hydrogen atom positions:

- Use of hydrogen coordinates as given in the PDB file (-b PDB) [code: C for CHARMM[94] as these hydrogen atoms are modelled by that software in this work]
- Modeling of hydrogen coordinates as that in DSSP (-b DSSP) [code: N for naïve]

The energy threshold of -0.5 kcal/mol in DSSP has been carefully adjusted considering the energy function given in eq. (5) employed by the authors of DSSP[27] to generate structural assignments that lead to visually convincing results. This value cannot be changed by end-users of DSSP as it is hardcoded in the source code. There is no conclusive energy threshold to establish whether a generic electrostatic interaction can be considered a hydrogen bond.

PSSC maintains many of the DSSP's strategies to deal with interruptions in otherwise regular hydrogen-bonding patterns. The inclusion of energetically non-optimal hydrogen bonds may hence help to generate a structural assignment that agrees with visual inspection. The energy threshold of PSSC was thus optimized in order to generate results as close to the DSSP results as possible.

It should be noted that DSSP allows for rather large β-bulges (up to a size of five residues); by definition β-bulges lack the characteristic inter-β-sheet hydrogen bonds.

Data preparation

A randomly chosen list of 100 non-homologous proteins was generated with the online tools provided by the Research Collaboratory for Structural Bioinformatics (RCSB) homepage[12], [95]. Only protein structures that have been resolved with X-ray crystallography, with a resolution larger than 1.5 Å, and a molecular weight between 10 kDa and 30 kDa were considered. Homologue proteins with 30% sequence identity were removed. For each cluster of sequentially similar proteins, the RCSB search tool automatically selects the cluster representative protein with the highest quality score, which is in a first approximation the inverse of the resolution. In total, these 100 proteins consist of 239 polypeptide chains with 16399 residues, for which a number of 15887 polar backbone hydrogen atoms needed to be modeled. To estimate the error, all calculations have been repeated ten times on a randomly chosen subset (with repetition) of 5000 residues.

1ATG	1B0B	1C9O	1DI6	1G8A	1HZT	1IJU	1J0P	1J98	1JG1
1KYF	1NNX	1OI7	1RCF	1SX7	1TGR	1TQG	1V6P	1VE4	1WYX
1X8Q	1YBK	1Z1S	1ZMA	2BJD	2DKO	2DLB	2EWH	2F5T	2FG1
2G7S	2HAX	2HIN	2HS1	2IC6	2IMS	2OV0	2OZH	2PQX	2UU8
2VIF	2WDS	2WUJ	2XJP	2XU3	2Y0O	3ACH	3BD1	3C6A	3D06
3D4E	3DQP	3DQY	3E8O	3F04	3F40	3FSO	3FZ4	3H0N	3H0O
3HWU	3IP0	3JUD	3K21	3KPE	3KUS	3KWU	3M97	3MST	3PLW
3RHB	3SD2	3SD6	3T47	3TC5	3TE4	3TQ5	3U62	3UEJ	3ZJA
3ZRX	3ZUC	4A3P	4AE7	4ANN	4AU1	4D8B	4EKF	4FZO	4G2E
4G7X	4GOF	4GUC	4H4J	4IPC	4J8C	4JVU	4KQP	4MZC	4N13

Table 7: Proteins used for hydrogen experiment.

Preparation of Structural Data

It is well-known that PDB files often contain problematic or erroneous data. Problems can for example derive from flexible and disordered regions of a protein that were not resolved in the X-ray study. Coordinates for some residues may thus be only partial or completely missing. In addition to the 20 proteinogenic amino acids, proteins may contain non-standard amino-acids that have been formed by post-translational modification. For instance, selenomethionine is typically added by crystallographers to solve the phase problem. Some residues may be able to adapt alternative conformations; for such residues, multiple sets of atomic coordinates (altlocs) are given, sometimes with different probabilities.

For many software tools that can process PDB files, non-standard amino acids and altlocs are problematic. Prior to any further calculations, the PDB files from both datasets used (ASTRAL40[11] and the high-resolution set of 100 proteins) were thus processed by a script that selects the altloc conformation with the highest probability (if not possible, the first altloc is used) and exchanged all selenomethionines with standard methionine.

Adding Hydrogen Atoms

In order to place the hydrogen atoms, the CHARMM[94] software was used. First, all missing hydrogen atoms (including those attached to side chain atoms, that are not considered for secondary structure assignment) were added. They were then optimized using 1000 iterations with steepest descent minimization, followed by 5000 iterations of the adopted basis Newton-Raphson algorithm. Hydrogen atoms that were already present in the crystallographic data were included in this minimization procedure. Data generated with this strategy are marked with the letter C indicating the use of CHARMM.

The default method of adding hydrogen atoms in PSSC employs the same algorithm used by DSSP: A simple geometric generation, where the N-H vector is assumed to be exactly antiparallel to the C=O vector in the same peptide plane. This procedure is here named as the "naïve" strategy (N). The following shorthand scheme is used: [D or E][N or C][threshold in kcal/mol], for example ND-0.5 for naively placed hydrogens with the DSSP energy function at the -0.5 kcal/mol threshold.

E-func. / **H-pos.**	**Naïve (N)**	**CHARMM (C)**
DSSP (D)	ND	CD
Electrostatic (E)	NE	CE

Results

When the DSSP-like hydrogen-bond energy function with the original -0.5 kcal/mol threshold is replaced by the full electrostatic energy function of CHARMM (ND-0.50 ⇒NE*x*, *x* being the variable threshold), the best agreement in secondary structure assignment with DSSP is reached when an energy threshold of -0.75 kcal/mol is chosen, the difference between the two assignments being 1.8%. When the original energy threshold of -0.50 kcal/mol is kept, the disagreement increases to 3.4%. This suggests that the DSSP-energy function may systematically underestimate the energetic gain of probable hydrogen bonds, a fact that may partially explain

the satisfactory results on secondary structure assignment that are produced with a very low energy threshold that allows for very weak hydrogen bonds.

The threshold value of -0.75 kcal/mol also leads to the best agreement when additionally to the energy function the hydrogen atoms are energetically minimized (CD-0.50 ⇒CE*x*). The -0.50kcal/mol threshold of DSSP is optimized for the naïvely placed hydrogen atoms. Nevertheless, the disagreement between ND-0.50 and CD*x* is also minimal (0.86%), when an energy threshold of x = -0.50kcal/mol is used. When x is decreased or increased by 0.1 kcal/mol, the disagreement additionally increases by 1.7% or 0.7%, respectively. Thus, it is not surprising that the same optimal threshold is gained, when the CHARMM energy minimization combined with the electrostatic function (CE) is compared against the naïve hydrogen placement and the DSSP energy function (ND-0.50 ⇒CE-0.75). The difference in this case is 1.87%.

A critical point to investigate is whether a significant change in structural assignment occurs, if a computationally expensive energy minimization is used. Interestingly, the difference between these strategies is only 0.8% (CE-0.75 ⇒NE-0.75).

Discussion

A systematic comparison of the results obtained by using the two different hydrogen-bond energy functions and the two different approaches to add hydrogens. The electrostatic energy function will be preferentially used in PSSC over the classic DSSP energy function, as it is expected to deliver more realistic results. However, for PSSC, the energy threshold should be -0.75 kcal/mol to be in line with the results of the widely excepted DSSP secondary structural assignment. Using PSSC with these settings, the secondary structure differs by 1.8% +/- 0.5% from the DSSP results—a relatively small, but significant difference.

Hydrogen atoms in PSSC are assigned with the same simple algorithm as in DSSP, except for the corrected handling of non-proline cis residues in PSSC. It has been shown that the use of the more sophisticated CHARMM energy function for hydrogen atom placement only changes the results by 0.8%. Such a small difference justifies the use of the naïve algorithm to place the hydrogen atoms.

Outlook

Current modern methods for prediction of protein secondary structure reach about 80% accuracy[96]–[98] for the three-class problem where only β-strands (E), a general helix class (H) and a coil class (C) are considered. This high prediction accuracy was reached by introducing protein sequence alignment profiles into the learning process[99]–[101]. Prior to this approach, only about 60% of the residues could be assigned to the correct secondary structure class. The secondary structure assignment by PSSC will be used for secondary structure prediction with the software SPARROW[97] (**S**econdary structure **P**redicting **ARR**ays of **O**ptimized **W**eights) that has been previously developed in our group. For this purpose, the statistical learning procedures will be trained with the data generated by PSSC and then evaluated against independent test data.

The eight-letter representation of PSSC enables a straightforward translation of the complete structural description into the reduced three-class scheme E, H, and C. For instance, composite helices may be considered as half α- and half 3_{10}-helix during the learning process. Additionally, the different strategies used to assign residues into eight or three classes of secondary structure presented in our contribution[79] will be evaluated with regard to their influence on the overall prediction results.

Similarly to protein secondary structure prediction, the protein-protein docking problem has recently received much attention. Since the secondary structure of solvent accessible residues will have an important influence on the occurrence of interface regions and binding modes[102]–[104] , it will also be useful to perform structural assignment with PSSC to address the protein-protein docking problem. Such studies are currently being performed in our laboratory.

Summary

Secondary structure is one of the most prominent features of a protein; it describes the local backbone conformations of its residues. Accurate and unambiguous representations of structural data are of great importance when dealing with individual proteins or with large protein databases. Two-dimensional graphs like the Ramachandran plot provide insight into the overall appearance of the whole protein structure as well as into conformations of individual residues. During my PhD work, I analyzed such structural features using the (d, ϑ)-plot that interprets the local protein backbone structure in terms of a helix of infinite length. The specific pair of helical parameters (d, ϑ) proved to be most insightful parameters compared to the other possible combinations of helix parameters and torsion angles. The formulas to calculate d and ϑ from given ϕ and ψ backbone dihedral angles used in the Ramachandran plot were calculated for the most recent values of protein backbone geometries in terms of bond length and bond angles.

Related with the work to visualize secondary structure, the program PSSC, a new tool for the characterization of the secondary structure, was developed. In particular, I developed the software PSSC to overcome the known problems of DSSP, the standard tool for secondary structure assignment. Secondary structure assignment with PSSC leads to a better clustering in the (ϕ, ψ) space. Ambiguities in the secondary structure of proteins, especially at the intersection of different helix types, are represented in a comprehensible format, allowing for mixed secondary structure classes. I further demonstrated the abundance of such mixed helical regions in proteins, underlining the necessity to introduce such mixed classes. As a first application of the transformation from the (ϕ, ψ) in the (d, ϑ) space, the software PSSC was extended to use this information for assignment of additional secondary structure.

An interactive web page displaying the results of PSSC together with a visualization in form of a (d, ϑ) plot is available at http://agknapp.chemie.fu-berlin.de/secsass.

Zusammenfassung auf Deutsch

Die Sekundärstruktur ist ein wesentliches Merkmal von Proteinen. Sie beschreibt die lokalen Backbone-Konformere der Aminosäurereste. Eine verläßliche und eindeutige Darstellung von Sekundärstrukturdaten ist unverzichtbar, sowohl bei der Analyse einzelner Proteine, als auch bei großen Proteindatenbanken. Zweidimensionale Graphen wie das Ramachandran-Diagramm bieten eine Übersicht über die gesamte Proteinstruktur sowie die Konformere der einzelnen Residuen. In meiner Dissertation habe ich Proteinsekundärstrukturen mithilfe des (d, ϑ)-Graphen analysiert. Hierbei wird die lokale Geometrie einzelner Residuen als unendlich ausgedehnte Helix interpretiert. Die Wahl des Variablenpaars (d, ϑ) hat sich hierfür gegenüber allen anderen möglichen Kombinationen von zwei Helixparametern oder Torsionswinkeln überlegen gezeigt. Die Beziehungen, um d und ϑ aus gegebenen Torsionswinkeln ϕ und ψ zu bestimmen, wurden mit den besten gegenwärtig verfügbaren Werten für die Winkel und Bindungslängen im Proteinrückgrat berechnet.

Bei dem Programm PSSC handelt es sich um ein von mir entwickeltes Werkzeug zur Charakterisierung von Proteinsekundärstrukturen. PSSC dient insbesondere dazu, vorhandene Probleme in dem weitverbreiteten Programm DSSP zu überwinden. Die von PSSC erzeugte Sekundärstrukturzuordnung liefert eine geringere Streuung im (ϕ, ψ)-Raum. Mehrdeutigkeiten durch überlappende Sekundärstrukturmotive, die vor allem in längeren Helices auftreten, werden in einem übersichtlichen Code dargestellt, der die Zuordnung von gemischten Sekundärstrukturklassen gestattet.

Die Notwendigkeit solcher gemischten Klassen habe ich durch die Analyse der Häufigkeit des Auftretens von gemischten Helices in Proteinen zeigen können. Eine direkte Anwendung findet die Transformation vom (ϕ, ψ)- in den (d, ϑ)-Raum im PSSC-Programm, um zusätzliche, nicht über Wasserstoffbrücken identifizierbare Sekundärstrukturklassen zu definieren

Eine interaktive Web-Applikation, die die Sekundärstrukturzuordnung von PSSC mithilfe eines (d, ϑ)-Graphen visualisiert, ist unter http://agknapp.chemie.fu-berlin.de/secsass verfügbar.

References

[1] A. L. Lehninger, *Biochemistry: The Molecular Basis of Cell Structure and Function (Second Edition)*. Worth Pub, 1978.

[2] H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. J. Smeenk, A. J. Kettelarij, J. H. van Maarseveen, P. Timmerman, P. G. Bolhuis, and S. Woutersen, "Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations.," *J. Phys. Chem. B*, vol. 117, no. 39, pp. 11490–501, Oct. 2013.

[3] C. A. Opitz, M. Kulke, M. C. Leake, C. Neagoe, H. Hinssen, R. J. Hajjar, and W. A. Linke, "Damped elastic recoil of the titin spring in myofibrils of human myocardium.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 22, pp. 12688–93, Oct. 2003.

[4] "File:Main protein structure levels en.svg - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg. [Accessed: 27-Feb-2014].

[5] W. T. Astbury and H. J. Woods, "X-Ray Studies of the Structure of Hair, Wool, and Related Fibres. II.--The Molecular Structure and Elastic Properties of Hair Keratin," *Proc. R. Soc. B Biol. Sci.*, vol. 114, no. 788, pp. 314–316, Jan. 1934.

[6] L. Pauling and R. B. Corey, "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 11, pp. 729–40, Nov. 1951.

[7] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain," *Proc. Natl. Acad. Sci.*, vol. 37, no. 4, pp. 205–211, Apr. 1951.

[8] J. C. KENDREW, R. E. DICKERSON, B. E. STRANDBERG, R. G. HART, D. R. DAVIES, D. C. PHILLIPS, and V. C. SHORE, "Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 □. Resolution," *Nature*, vol. 185, no. 4711, pp. 422–427, Feb. 1960.

[9] A. F. Cullis, H. Muirhead, M. F. Perutz, M. G. Rossmann, and A. C. T. North, "The Structure of Haemoglobin. IX. A Three-Dimensional Fourier Synthesis at 5.5 Formula Resolution: Description of the Structure," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 265, no. 1321, pp. 161–187, Jan. 1962.

[10] K. K. Ng, J. F. Petersen, M. M. Cherney, C. Garen, J. J. Zalatoris, C. Rao-Naik, B. M. Dunn, M. R. Martzen, R. J. Peanasky, and M. N. James, "Structural basis for the inhibition of porcine pepsin by Ascaris pepsin inhibitor-3.," *Nat. Struct. Biol.*, vol. 7, no. 8, pp. 653–7, Aug. 2000.

[11] J.-M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL Compendium in 2004.," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D189–D192, Jan. 2004.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

[13] K. Illergård, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence--a study of structural response in protein cores.," *Proteins*, vol. 77, no. 3, pp. 499–508, Nov. 2009.

[14] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” *J. Mol. Biol.*, vol. 7, no. 1, pp. 95–99, Jul. 1963.

[15] J. S. Richardson, “Early ribbon drawings of proteins.,” *Nat. Struct. Biol.*, vol. 7, no. 8, pp. 624–5, Aug. 2000.

[16] P. J. Kraulis, “MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures,” *J. Appl. Crystallogr.*, vol. 24, no. 5, pp. 946–950, Oct. 1991.

[17] W. Humphrey, A. Dalke, and K. Schulten, “VMD: Visual molecular dynamics,” *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, Feb. 1996.

[18] L. Schrödinger, “The PyMOL Molecular Graphics System,” Aug. 2010.

[19] “Jmol: an open-source Java viewer for chemical structures in 3D.” [Online]. Available: Jmol: an open-source Java viewer for chemical structures in 3D.

[20] “JSmol | Free software downloads at SourceForge.net.” [Online]. Available: http://sourceforge.net/projects/jsmol/. [Accessed: 27-Feb-2014].

[21] “GLmol - Molecular Viewer on WebGL/Javascript.” [Online]. Available: http://webglmol.sourceforge.jp/index-en.html. [Accessed: 27-Feb-2014].

[22] G. D. Rose, “Lifting the lid on helix-capping.,” *Nat. Chem. Biol.*, vol. 2, no. 3, pp. 123–4, Mar. 2006.

[23] J. Prieto and L. Serrano, “C-capping and helix stability: the Pro C-capping motif.,” *J. Mol. Biol.*, vol. 274, no. 2, pp. 276–88, Nov. 1997.

[24] J. Segura, B. Oliva, and N. Fernandez-Fuentes, “CAPS-DB: a structural classification of helix-capping motifs.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D479–85, Jan. 2012.

[25] D. P. Leader and E. J. Milner-White, “The structure of the ends of α-helices in globular proteins: effect of additional hydrogen bonds and implications for helix formation.,” *Proteins*, vol. 79, no. 3, pp. 1010–9, Mar. 2011.

[26] R. Aurora and G. D. Rose, “Helix capping.,” *Protein Sci.*, vol. 7, no. 1, pp. 21–38, Jan. 1998.

[27] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.,” *Biopolymers*, vol. 22, no. 12, pp. 2577–637, Dec. 1983.

[28] S. Flückiger, P. R. E. Mittl, L. Scapozza, H. Fijten, G. Folkers, M. G. Grütter, K. Blaser, and R. Crameri, “Comparison of the crystal structures of the human manganese superoxide dismutase and the homologous Aspergillus fumigatus allergen at 2-A resolution.,” *J. Immunol.*, vol. 168, no. 3, pp. 1267–72, Feb. 2002.

[29] A. Ambrogelly, S. Palioura, and D. Söll, “Natural expansion of the genetic code.,” *Nat. Chem. Biol.*, vol. 3, no. 1, pp. 29–35, Jan. 2007.

[30] M. L. Ruiz del Castillo and G. Dobson, “Varietal differences in terpene composition of blackcurrant (Ribes nigrum L) berries by solid phase microextraction/gas chromatography,” *J. Sci. Food Agric.*, vol. 82, no. 13, pp. 1510–1515, Oct. 2002.

[31] J. Kobayashi, Y. Shimizu, Y. Mutaguchi, K. Doi, and T. Ohshima, “Characterization of d-amino acid aminotransferase from Lactobacillus salivarius,” *J. Mol. Catal. B Enzym.*, vol. 94, pp. 15–22, Oct. 2013.

[32] R. MARCHELLI, “The potential of enantioselective analysis as a quality control tool,” *Trends Food Sci. Technol.*, vol. 7, no. 4, pp. 113–119, Apr. 1996.

[33] P. A. Karplus, “Experimentally observed conformation-dependent geometry and hidden strain in proteins.,” *Protein Sci.*, vol. 5, no. 7, pp. 1406–20, Jul. 1996.

[34] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, “PROCHECK: a program to check the stereochemical quality of protein structures,” *J. Appl. Crystallogr.*, vol. 26, no. 2, pp. 283–291, Apr. 1993.

[35] R. W. W. Hooft, C. Sander, and G. Vriend, “Objectively judging the quality of a protein structure from a Ramachandran plot,” *Bioinformatics*, vol. 13, no. 4, pp. 425–430, Aug. 1997.

[36] R. J. Read, P. D. Adams, W. B. Arendall, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart, “A new generation of crystallographic validation tools for the protein data bank.,” *Structure*, vol. 19, no. 10, pp. 1395–412, Oct. 2011.

[37] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, “The Protein Data Bank: a computer-based archival file for macromolecular structures.,” *J. Mol. Biol.*, vol. 112, no. 3, pp. 535–42, May 1977.

[38] M. Elias, D. Liebschner, J. Koepke, C. Lecomte, B. Guillot, C. Jelsch, and E. Chabriere, “Hydrogen atoms in protein structures: high-resolution X-ray diffraction structure of the DFPase.,” *BMC Res. Notes*, vol. 6, no. 1, p. 308, Jan. 2013.

[39] G. N. Ramachandran, C. M. Venkatachalam, and S. Krimm, “Stereochemical Criteria for Polypeptide and Protein Chain Conformations. III. Helical and hydrogen-bonded polypeptide chains.,” *Biophys. J.*, vol. 6, no. 6, pp. 849–872, 1966.

[40] M. S. Weiss, A. Jabs, and R. Hilgenfeld, “Peptide bonds revisited.,” *Nat. Struct. Biol.*, vol. 5, no. 8, p. 676, Aug. 1998.

[41] D. E. Stewart, A. Sarkar, and J. E. Wampler, “Occurrence and role of cis peptide bonds in protein structures.,” *J. Mol. Biol.*, vol. 214, no. 1, pp. 253–60, Jul. 1990.

[42] G. J. Kleywegt, “Validation of protein models from Calpha coordinates alone.,” *J. Mol. Biol.*, vol. 273, no. 2, pp. 371–6, Oct. 1997.

[43] T. J. Oldfield and R. E. Hubbard, “Analysis of Cα geometry in protein structures,” *Proteins Struct. Funct. Genet.*, vol. 18, no. 4, pp. 324–337, 1994.

[44] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, and D. J. Nesbitt, “Definition of the hydrogen bond (IUPAC Recommendations 2011),” *Pure Appl. Chem.*, vol. 83, no. 8, pp. 1637–1641, Jul. 2011.

[45] P. Muller, “Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994),” *Pure Appl. Chem.*, vol. 66, no. 5, pp. 1077–1184, Jan. 1994.

[46] S. Lifson, A. T. Hagler, and P. Dauber, “Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the C:O.cntdot..cntdot..cntdot.H-hydrogen bonds,” *J. Am. Chem. Soc.*, vol. 101, no. 18, pp. 5111–5121, Aug. 1979.

[47] T. E. Creighton, *Proteins: Structures and Molecular Properties*. W. H. Freeman, 1992.

[48] R. E. Hubbard and M. Kamran Haider, “Hydrogen Bonds in Proteins: Role and Strength,” *Encycl. Life Sci.*, Feb. 2010.

[49] E. N. Baker and R. E. Hubbard, “Hydrogen bonding in globular proteins.,” *Prog. Biophys. Mol. Biol.*, vol. 44, no. 2, pp. 97–179, Jan. 1984.

[50] S. A. Hollingsworth, D. S. Berkholz, and P. A. Karplus, “On the occurrence of linear groups in proteins.,” *Protein Sci.*, vol. 18, no. 6, pp. 1321–5, Jun. 2009.

[51] R. P. Riek and R. M. Graham, “The elusive π-helix.,” *J. Struct. Biol.*, vol. 173, no. 1, pp. 153–60, Jan. 2011.

[52] R. B. Cooley, D. J. Arp, and P. A. Karplus, “Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality.,” *J. Mol. Biol.*, vol. 404, no. 2, pp. 232–46, Nov. 2010.

[53] “IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969).,” *Biochem. J.*, vol. 121, no. 4, pp. 577–85, Feb. 1971.

[54] L. Pauling and R. B. Corey, “The Pleated Sheet, A New Layer Configuration of Polypeptide Chains,” *Proc. Natl. Acad. Sci.*, vol. 37, no. 5, pp. 251–256, May 1951.

[55] N. C. Fitzkee, P. J. Fleming, H. Gong, N. Panasik, T. O. Street, and G. D. Rose, “Are proteins made from a limited parts list?,” *Trends Biochem. Sci.*, vol. 30, no. 2, pp. 73–80, 2005.

[56] P. H. Maccallum, R. Poet, and E. James Milner-White, “Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of α-helices and antiparallel β-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includ,” *J. Mol. Biol.*, vol. 248, no. 2, pp. 361–373, Jan. 1995.

[57] P. H. Maccallum, R. Poet, and E. J. Milner-White, “Coulombic attractions between partially charged main-chain atoms stabilise the right-handed twist found in most beta-strands.,” *J. Mol. Biol.*, vol. 248, no. 2, pp. 374–84, Apr. 1995.

[58] B. K. Ho, A. Thomas, and R. Brasseur, “Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix.,” *Protein Sci.*, vol. 12, no. 11, pp. 2508–22, Nov. 2003.

[59] F. Avbelj and L. Fele, “Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins.,” *J. Mol. Biol.*, vol. 279, no. 3, pp. 665–84, Jun. 1998.

[60] T. R. Lezon, J. R. Banavar, A. M. Lesk, and A. Maritan, “What determines the spectrum of protein native state structures?,” *Proteins*, vol. 63, no. 2, pp. 273–7, May 2006.

[61] R. L. Baldwin and G. D. Rose, “Is protein folding hierarchic? I. Local structure and peptide folding.,” *Trends Biochem. Sci.*, vol. 24, no. 1, pp. 26–33, Jan. 1999.

[62] R. Srinivasan and G. D. Rose, “A physical basis for protein secondary structure.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 25, pp. 14258–63, Dec. 1999.

[63] R. V Pappu and G. D. Rose, “A simple model for polyproline II structure in unfolded states of alanine-based peptides.,” *Protein Sci.*, vol. 11, no. 10, pp. 2437–55, Oct. 2002.

[64] M. Mezei, P. J. Fleming, R. Srinivasan, and G. D. Rose, “Polyproline II helix is the preferred conformation for unfolded polyalanine in water.,” *Proteins*, vol. 55, no. 3, pp. 502–7, May 2004.

[65] A. N. Drozdov, A. Grossfield, and R. V Pappu, “Role of solvent in determining conformational preferences of alanine dipeptide in water.,” *J. Am. Chem. Soc.*, vol. 126, no. 8, pp. 2574–81, Mar. 2004.

[66] M. V Cubellis, F. Caillez, T. L. Blundell, and S. C. Lovell, “Properties of polyproline II, a secondary structure element implicated in protein-protein interactions.,” *Proteins*, vol. 58, no. 4, pp. 880–92, Mar. 2005.

[67] A. A. Adzhubei, M. J. E. Sternberg, and A. A. Makarov, “Polyproline-II helix in proteins: structure and function.,” *J. Mol. Biol.*, vol. 425, no. 12, pp. 2100–32, Jun. 2013.

[68] K. C. Chou, “Prediction of tight turns and their types in proteins.,” *Anal. Biochem.*, vol. 286, no. 1, pp. 1–16, Nov. 2000.

[69] N. C. Fitzkee, P. J. Fleming, and G. D. Rose, “The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB.,” *Proteins*, vol. 58, no. 4, pp. 852–4, Mar. 2005.

[70] P. Y. Chou and G. D. Fasman, “Prediction of protein conformation,” *Biochemistry*, vol. 13, no. 2, pp. 222–245, Jan. 1974.

[71] M. Heinig and D. Frishman, “STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W500–2, Jul. 2004.

[72] “PROSS: Dihedral Angle-Based Secondary Structure Assignment.” [Online]. Available: http://folding.chemistry.msstate.edu/utils/pross.html. [Accessed: 10-Sep-2014].

[73] S. Y. Park, M.-J. Yoo, J. Shin, and K.-H. Cho, “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.*, vol. 44, no. 2, pp. 118–22, Feb. 2011.

[74] F. Dupuis, J.-F. Sadoc, and J.-P. Mornon, “Protein secondary structure assignment through Voronoï tessellation.,” *Proteins*, vol. 55, no. 3, pp. 519–28, May 2004.

[75] H. Sklenar, C. Etchebest, and R. Lavery, “Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis.,” *Proteins*, vol. 6, no. 1, pp. 46–60, Jan. 1989.

[76] R. M. Hanson, D. Kohler, and S. G. Braun, “Quaternion-based definition of protein secondary structure straightness and its relationship to Ramachandran angles.,” *Proteins*, vol. 79, no. 7, pp. 2172–80, Jul. 2011.

[77] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A. G. de Brevern, and J.-F. Gibrat, “Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.,” *BMC Struct. Biol.*, vol. 5, p. 17, Jan. 2005.

[78] B. Schäling, *The Boost C++ Libraries*. XML Press, 2011.

[79] J. Zacharias and E.-W. Knapp, “Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC.,” *J. Chem. Inf. Model.*, Jun. 2014.

[80] A. Jabs, M. S. Weiss, and R. Hilgenfeld, “Non-proline cis peptide bonds in proteins.,” *J. Mol. Biol.*, vol. 286, no. 1, pp. 291–304, Feb. 1999.

[81] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.

[82] D. T. Lee and C. K. Wong, “Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees,” *Acta Inform.*, vol. 9, no. 1, 1977.

[83] M. G. Matthias Teschner, Bruno Heidelberger, Matthias Mueller, Danat Pomeranets, “Optimized Spatial Hashing for Collision Detection of Deformable Objects.”

[84] Á. González, “Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices,” *Math. Geosci.*, vol. 42, no. 1, pp. 49–64, Nov. 2009.

[85] Y. Dou, P.-F. Baisnée, G. Pollastri, Y. Pécout, J. Nowick, and P. Baldi, “ICBS: a database of interactions between protein chains mediated by beta-sheet formation.,” *Bioinformatics*, vol. 20, no. 16, pp. 2767–77, Nov. 2004.

[86] F. Cordier, L. Nisius, A. J. Dingley, and S. Grzesiek, “Direct detection of N-H[...]O=C hydrogen bonds in biomolecules by NMR spectroscopy.,” *Nat. Protoc.*, vol. 3, no. 2, pp. 235–41, Jan. 2008.

[87] E. Tüchsen and P. E. Hansen, “Hydrogen bonding monitored by deuterium isotope effects on carbonyl 13C chemical shift in BPTI: intra-residue hydrogen bonds in antiparallel beta-sheet.,” *Int. J. Biol. Macromol.*, vol. 13, no. 1, pp. 2–8, Feb. 1991.

[88] J. Zacharias and E. W. Knapp, “Geometry motivated alternative view on local protein backbone structures.,” *Protein Sci.*, vol. 22, no. 11, pp. 1669–74, Nov. 2013.

[89] A. Light, *Proteins: structure and function*. Prentice-Hall, 1974.

[90] N. Eswar, C. Ramakrishnan, and N. Srinivasan, “Stranded in isolation: structural role of isolated extended strands in proteins,” *Protein Eng. Des. Sel.*, vol. 16, no. 5, pp. 331–339, May 2003.

[91] M. V. Cubellis, F. Cailliez, and S. C. Lovell, “Secondary structure assignment that accurately reflects physical and evolutionary characteristics.,” *BMC Bioinformatics*, vol. 6 Suppl 4, p. S8, Dec. 2005.

[92] T. Williams, C. Kelley, and many others, “Gnuplot 4.4: an interactive plotting program,” 2010. [Online]. Available: http://gnuplot.sourceforge.net.

[93] T. Bray, “The JavaScript Object Notation (JSON) Data Interchange Format.” [Online]. Available: http://tools.ietf.org/html/rfc7159. [Accessed: 10-Sep-2014].

[94] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, “CHARMM: the biomolecular simulation program.,” *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–614, Jul. 2009.

[95] “RCSB Protein Data Bank - RCSB PDB.” [Online]. Available: http://www.pdb.org/pdb/home/home.do. [Accessed: 10-Sep-2014].

[96] J. L. Leopold and R. L. Frank, “Protein secondary structure prediction using BLAST and exhaustive RT-RICO, the search for optimal segment length and threshold,” in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2012, pp. 35–42.

[97] F. Bettella, D. Rasinski, and E. W. Knapp, “Protein secondary structure prediction with SPARROW.,” *J. Chem. Inf. Model.*, vol. 52, no. 2, pp. 545–56, Feb. 2012.

[98] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices.,” *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[99] B. Rost and C. Sander, “Improved prediction of protein secondary structure by use of sequence profiles and neural networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 16, pp. 7558–62, Aug. 1993.

[100] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy.,” *J. Mol. Biol.*, vol. 232, no. 2, pp. 584–99, Jul. 1993.

[101] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–402, Sep. 1997.

[102] C. Yan, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, “Characterization of protein-protein interfaces.,” *Protein J.*, vol. 27, no. 1, pp. 59–70, Jan. 2008.

[103] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, “Protein-protein docking dealing with the unknown.,” *J. Comput. Chem.*, vol. 31, no. 2, pp. 317–42, Jan. 2010.

[104] S. Mukherjee and Y. Zhang, “Protein-protein complex structure predictions by multimeric threading and template recombination.,” *Structure*, vol. 19, no. 7, pp. 955–66, Jul. 2011.

Figures

Tables