Chapter 2

Protein Structures

Proteins are polymers built of amino acids that are covalently joined together by peptide bonds. A single amino acid within a *polypeptide chain* is called a *residue*. Each amino acid has an invariant part, the *backbone* or *main chain*, and a characteristic *sidechain*. The unique sequence of the sidechains gives the protein its individual features. Natural proteins contain a basic repertoire of 20 essential amino acids (see Appendix Table G.1). At physiological temperatures in aqueous solution, a polypeptide chain folds into a form that is globular in most cases. The protein structure can be described at four hierarchical levels:

- **Primary**. The amino acid sequence which is directly determined by the sequence of nucleotides in the gene encoding it (Figure 2.1a).
- **Secondary**. Secondary structure elements (SSEs), either helices or sheets, are regions of the polypeptide chain formed through regular hydrogen bonding interactions between *N-H* and *C=O* groups of the protein backbone (Figure 2.1b).
- **Tertiary**. In the globular form of a single polypeptide chain, the SSEs as well as loops and other irregular structure regions are folded into a *tertiary structure* (Figure 2.1c).
- **Quaternary**. If the protein consists of more than one polypeptide chain, the quaternary structure describes the association of the folded chains (Figure 2.1d).

The *primary* structure of a protein given as its amino acid sequence contains all the information needed to specify the regular repeating patterns of hydrogen bonded backbone conformations (*secondary* structure) such as helices and sheets, as well as the way these elements pack together to form the overall fold of the protein (*tertiary* structure). SSEs that combine in specific geometric arrangements are called *motifs* or *supersecondary* structure elements (small, discrete, commonly observed aggregates of SSEs). Many proteins contain compact globular units within the folding pattern of a single chain that are considered to be stable independently of the rest of the chain. These units are called *domains*. Within the described protein structure hierarchy, domains fall in between motifs and tertiary structure. The conformation of two or more individual polypeptide chains is called *quaternary* structure. Proteins composed of more than one

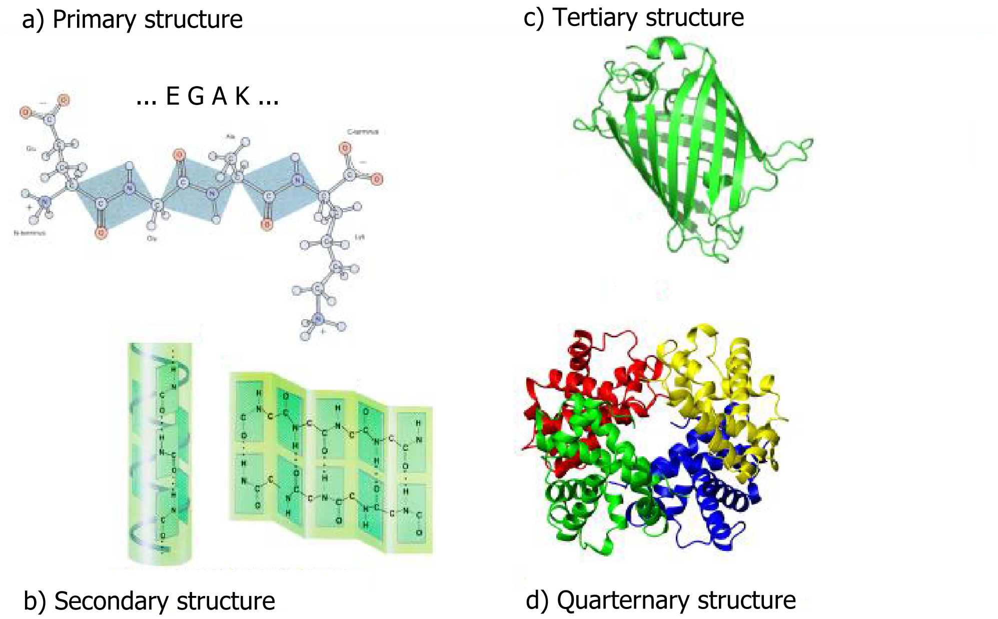


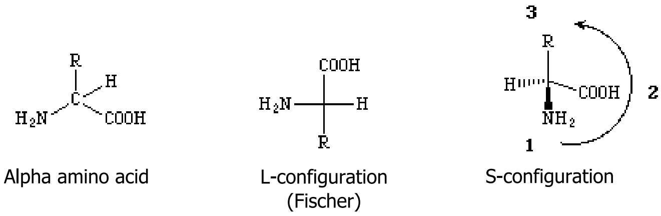
Figure 2.1: **Protein structure levels.** (a) Primary structure. (b) Secondary structure (left: helix, right: sheet). (c) Tertiary structure (Beta barrel). (d) Quarternary structure (single chains in different colors). (Source: `http://www.science.org.au/sats2004/mackay.htm`).

polypeptide chain are called *multimers* or *oligomers*. The most common multimers are dimers. Multimers composed only of a single type of monomer are called *homomers*, e.g., homodimers; multimers composed of monomers encoded by different genes are called *heteromers*.

2.1 Amino Acid Properties

Amino acids are small molecules that contain an *amino* (NH_2) and a *carboxylate* ($COOH$) group (see Figure 2.2). The term *alpha amino acid* refers to molecules having the amino and carboxylate group attached to the same carbon atom, the $C\alpha$-atom. The various alpha amino acids differ in their sidechain (R in Figure 2.2) that is attached to their $C\alpha$-atom. It is this group that distinguishes one amino acid from another and confers the specific chemical properties of a certain amino acid. Additionally, amino acids contain in their backbone a single hydrogen atom attached to the central $C\alpha$-atom.

All 20 standard amino acids (Appendix Table G.1) except glycine have an asymmetric $C\alpha$-atom, i.e., a carbon atom with four different binding partners. The sidechain of glycine consists of one single hydrogen atom, i.e., glycine has effectively no sidechain at all and is therefore, in contrast to the 19 other amino acids, achiral. The configuration of chiral amino acids is often classified according to two mainly used naming conventions, the D/L-nomenclature (Fischer projection) and the R/S-nomenclature (see Figure 2.2). The Fischer projection is a two-dimensional representation of a three-dimensional organic molecule. In mammalian cells all amino acids exist in the L-configuration, where L denotes that the amino acid configuration is similar to L-glyceraldehyde. D-form amino acids are rarely seen in polypeptide chains, but they are a result of direct enzymatic synthesis [133]. The R/S-nomenclature sorts the four different groups on the asymmetric $C\alpha$-atom by assigning priorities according to the Cahn-Ingold-Prelog rules [33]. All natural amino acids, except glycine, are in S-configuration.



Alpha amino acid

L-configuration (Fischer)

S-configuration

Figure 2.2: **Amino acid configurations.** Left: standard alpha amino acid. Middle: Fischer projection. Right: S-configuration. R denotes the sidechain.

The properties of each amino acid are dictated by its sidechain, which can vary in size, shape, charge, reactivity and ability to form hydrogen bonds. The standard amino acids can be loosely grouped into classes based on these chemical properties. Three classes are commonly accepted:

- *Hydrophobic* sidechains are electrically neutral. They are called ’hydrophobic’, because of the thermodynamically unfavorable interaction of hydrocarbons with water molecules.
- *Polar* amino acids like asparagine and glutamine contain amino groups within their sidechains; serine, threonine, and tyrosine contain hydroxyl groups. Polar groups can participate in hydrogen bonding.
- *Charged* sidechains. Aspartic acid and glutamatic acid are negatively charged; lysine and arginine are positively charged. The charged atoms occur at or near the ends of the relatively long and flexible sidechains; the atoms proximal to the backbone are non-polar. Two sidechains with positive and negative charge can approach each other in space to form a ’salt bridge’.

Within these classes, additional sub-classifications are possible, e.g., aromatic or aliphatic, large or small, etc. [224]. Sidechain atoms are identified by their chemical symbol and successive letters from the Greek alphabet, proceeding out from the $C\alpha$-atom. For example, the sidechain atoms of methionine are $C\beta$, $C\gamma$, $S\delta$, and $C\epsilon$. Proline is a special amino acid, because its sidechain forms a bond with its own amino group, causing it to be cyclic. It generally exhibits the properties of an aliphatic nonpolar amino acid, but the cyclic construction limits its flexibility. Some proteins contain amino acids outside the canonical set of 20, but these are produced by chemical modification after the protein is synthesized, or by introduction of a selenocysteine during translation [16], as in gluthathione peroxidase.

The amino acid sidechains have different tendencies to participate in interactions with each other or with water. These differences influence their contributions to protein folding, protein stability, and protein function. *Hydrophobic* amino acids like glycine, alanine, leucine, isoleucine, proline and valine are aliphatic in nature, i.e., they contain hydrocarbons that are joined together in straight or branched chains. Their tendency to avoid contact with water and pack against each other is the basis for the *hydrophobic effect* (see also Section 2.3) with the exception of glycine and alanine that are too small to contribute to the hydrophobic effect. *Hydrophilic* amino acids are able to form hydrogen bonds with one another, to the peptide backbone, to polar organic molecules, and to water. *Amphiphatic* amino acids have both polar and nonpolar character, making them ideal for forming interfaces. There are three aromatic amino acids, phenylalanine, tyrosine, and tryptophan. These amino acids have sidechains, which contain delocalized π-electrons that can interact with other π-electrons in biomolecules. In addition, the phenolic hydroxyl of tyrosine can ionize under physiological conditions, and thus increase the solubility in water. Two of the amino acids, cysteine and methionine, contain sulfur atoms providing them the possibility to form disulphide bridges which are important for the tertiary structure of proteins. Finally, there are two hydroxyl containing amino acids, serine and threonine. These two amino acids have sidechains that can form hydrogen bonds to water molecules or to other groups on neighboring macromolecules.

2.2 Peptide Bond

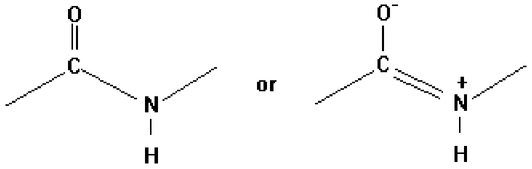


Figure 2.3: **Resonance effect of a peptide bond.**

Two amino acids can be connected via a *peptide bond*. Chemically, the peptide bond is a covalent bond that is built between the carboxylate and the amino group of two amino acids by the loss of a water molecule, a reaction also called condensation (see Figures 2.3 and 2.4). Using the peptide bond, long linear polypeptide chains of amino acids can be generated. Therefore, every polypeptide chain has an amino (N)- and a carboxyl (C)-terminal end providing a sequential ordering of the amino acids from N- to C-terminus. The synthesis of peptide bonds is enzymatically controlled on the ribosomes and directed by the mRNA transcript. Peptide bonds have partial double-bond character and, thus, are very stable due to resonance. The delocalization of electrons over several atoms (see Figure 2.3 right) increases the dipole moment of the bond:

the three non-hydrogen atoms that make the peptide bond (the carbonyl oxygen O, the carbonyl carbon C, and the amide nitrogen N) are almost coplanar (indicated by green planes in Figure 2.4). The peptide group can appear in *cis* or *trans* form, with the trans-isomer being more stable. For all amino acids, except proline, the energy difference between cis and trans is very large. As a result, cis-dipeptides in protein structures appear only between a proline and the residue preceding it in the polypeptide chain.

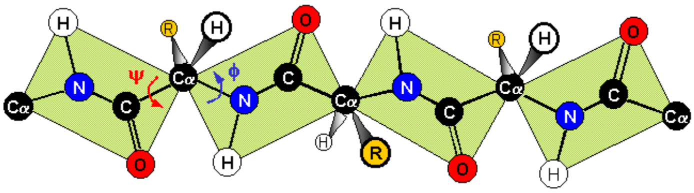


Figure 2.4: **Polypeptide chain.** The backbone atoms are marked by different colors. Sidechains are indicated with R. The torsion angles ϕ and ψ are shown as well as there defining planes. The torsion angle ω between N-$C\alpha$ is not shown. (Source: `http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olunderstandconfo.html`).

The polypeptide backbone allows the rotation only around the single bonds where the Cα-atom is participating. This rotational freedom can be described by torsion angles. Four atoms define torsion angles, better referred to as dihedral angles. Like shown in Figure 2.4, the backbone dihedral angles of a residue i are called ϕ (*phi*, defined by the backbone atoms C_{i-1}-N_i-C_i^α-C_i), and ψ (*psi*, defined by the backbone atoms N_i-C_i^α-C_i-N_{i+1}). There is an additional torsion angle ω (*omega*, defined by the backbone atoms $C_i\alpha$-C_i-N_{i+1}-$C_{i+1}\alpha$), which has only restricted rotational freedom because of the resonance stabilization: the planarity of the peptide bond usually restricts ω to be 180°. Generally, ϕ controls the C_i-C_{i+1} distance, ψ controls the N_i-N_{i+1} distance, and ω the C_i^α-C_{i+1}^α distance. The distance (in Å) between successive atoms on the polypeptide backbone is approximately constant: 1.474 for N-$C\alpha$, 1.53 for $C\alpha$-C, and 1.32 for C-N [206]. The backbone conformation of each residue can be determined by the two angles ϕ and ψ. Rotations around these both bonds are not restricted by the physico-chemical properties of the bond, but only by the steric collisions in possible conformations between the sidechains of the residues and the backbone. A *Ramachandran plot* (a plot of ϕ and ψ angles) maps the entire conformational space of a polypeptide, and thus describes the fold of a polypeptide chain showing the allowed and disallowed conformations [191], which is important, e.g., for model evaluation in protein structure prediction. Figure 2.5 shows the allowed torsion angle combinations in red. The angles allowed depend on the limiting distance chosen for interatomic contacts. If you allow less restricted but still possible interatomic contacts, more conformational space is available (yellow regions in Figure 2.5). There are two main allowed regions, one around $\phi = -57°$ and $\psi = -47°$ (right-handed α-helix in Figure 2.5), and the other around $\phi = -125°$ and $\psi = +125°$ (β sheet in Figure 2.5). The mirror image of the α-helix (left-handed helix in Figure 2.5) is allowed for glycines only.

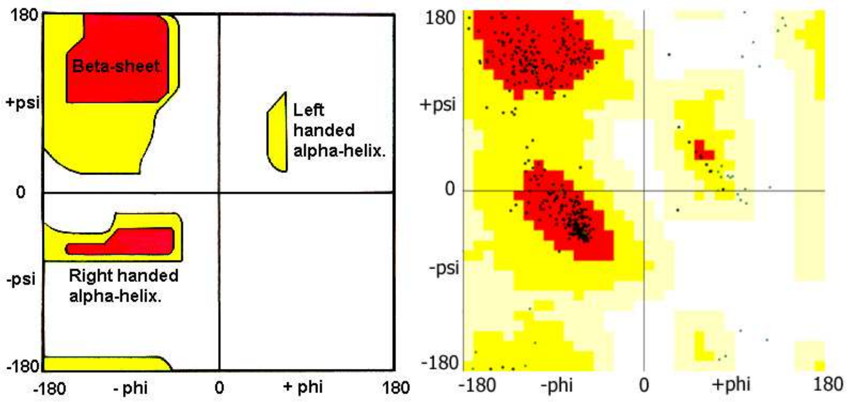


Figure 2.5: **Ramachandran plots.** Left: allowed torsion angle combinations are colored red and yellow (Source: `http://www.cryst.bbk.ac.uk/PP2/course`). Right: torsion angles combinations (black dots) of PDB 1*axs* are plotted with STING [171].

The two major allowed regions correspond to the two major types of secondary structure: helix and sheet (see Section 2.1). Some key exceptions to these conformational limitations can be attributed to glycine and proline. Since glycine has only a single hydrogen as its sidechain, there is a remarkable reduced steric hindrance about the ϕ and ψ angles of this residue. Thus, glycine residues expand the possible conformational space tremendously. Conversely, the cyclic bond present in proline residues reduces the conformational freedom beyond the limitations observed with other amino acids.

Sidechain conformations can also be described by angles of internal rotation, denoted χ_1 to χ_5. Different sidechains have different degree of freedom. The conformations of any sidechain corresponding to different combinations of values of the χ angles are called *rotamers*.

2.3 Protein Folding

Proteins can only perform their function, if they fold into a stable state corresponding to the minimum free energy. This so-called *native state* is only stable within narrow ranges of conditions of solvent and temperature. Beyond these boundaries proteins lose their defined compact structure and adopt states with disordered backbone conformation and few residue interactions. The proof that protein structure is dictated by its amino acid sequence alone is based on experiments first carried out by Anfinsen [9, 10], who showed that the denaturation of ribonuclease—the breakup of the native structure by heat or urea—was reversible. If the denatured molecules were returned to normal conditions of temperature and solvent, both structure and enzymatic activity returned. In the absence of large kinetic barriers in the free energy landscape, these results suggest that the native conformations of most proteins are the lowest free energy conformations for their amino acid sequences. Protein folding occurs very rapidly, but there is evidence that one or more partially folded intermediate states exist transiently, along the path to the final structure. These intermediate structures are not as well characterized as the native structures, but have already many of the SSEs of the fully folded protein without the closely packed interior and full complement of weak interactions that characterize the native state. Nevertheless, the 'folding pathway' of a protein is still not very well understood.

In most cases, the entire protein (or at least a large part of it) is necessary for building a stable conformation. Therefore, interactions between different parts of the protein sequence have to be established. Interactions can be formed between residues close in sequence or they can involve parts of the protein that are very distant in sequence, but brought into spatial proximity by the folding process. Peptide bonds are the only covalent bonds that hold the residues together in most proteins. The main contribution of the stabilization energy of a folded protein comes from non-covalent weakly polar interactions. The most important of these interactions stabilizing polypeptides are [185]:

- *electrostatic interactions* are interactions between atoms due to attraction of opposite partial charges and the repulsion of partial charges of same type.

- *hydrogen bonds* are interactions between donor atoms, which are bound to positively polarized hydrogen atoms, and acceptor atoms, which are negatively polarized.

- *salt bridges* are hydrogen bonds in which both donor and acceptor atoms are fully charged. The bonding energy of salt bridges is significantly higher than that of hydrogen bonds.

- *van-der-Waals interactions* are weak attractive forces between two atoms or groups of atoms, arising from the fluctuations in the electron distribution. Van-der-Waals forces are stronger between less electronegative atoms such as found in hydrophobic groups or atoms.

The hydrogen bonding properties of water molecules have important effects on protein stability, because water is potentially both, a donor and acceptor for hydrogen bonds. Water molecules form hydrogen bonds to one another as well as to polar atoms in proteins. Hydrogen bonds also contribute to the binding of cofactors and substrates, e.g., the binding of NAD to an alcohol dehydrogenase. In proteins almost all polar atoms are involved in hydrogen bonds, either between other polar atoms from itself or between water molecules of the solvent environment. In the unfolded state, polar atoms build hydrogen bonds to water. In the folded state, the hydrogen bonding potential of polar backbone atoms buried in the interior of the protein must be satisfied. In the buried protein core, the $C{=}O$ and N-H groups cannot form hydrogen bonds with water, so they tend to form hydrogen bonds with one another. Hereby, they are neutralizing their polar groups leading to regularly hydrogen bonding patterns that are forming the SSEs.

Hydrophobic cores appear to be essential for the stability of globular polypeptide chains and domains. Concentrating hydrophobic groups in the core is energetically favorable, because it minimizes the number of unfavorable interactions of hydrophobic groups with water molecules and maximizes the number of van-der-Waals interactions between hydrophobic groups ('hydrophobic effect') [220]. Therefore, the folding problem can be seen as to maximize the exposure of hydrophilic groups to aqueous solution while minimizing the exposure of its hydrophobic groups.

Many proteins contain covalent chemical bonds in addition of the polypeptide backbone. Disulphide bridges between sulphur atoms in cysteine residues are quite common. They can also link different polypeptide chains, as in insulin or immunoglobulins. Some proteins contain metal ions as integral parts of their structure, either bound directly to sidechains or as organic cofactors, e.g., zinc in 'Zinc-finger' proteins like pig insulin that is covalently bound by sidechains. In other cases, the metal is not bound directly to the protein, but is part of a larger ligand, e.g., the Mg^{2+} ion in bacteriochlorophyll.

2.4 Secondary Structure Elements

Although proteins are linear polymers, the structures of most proteins are not random coils found for synthetic non-natural polymers. A characteristic of folded polypeptide chains is that segments of the chain adopt conformations in which ϕ and ψ torsion angles of the protein backbone repeat in regular patterns, the *secondary structure elements* (SSEs) (see also Section 2.2). The two major types of secondary structure are helices and strands forming beta sheets. The SSEs contribute significantly to the stabilization of the overall protein fold. Helices and sheets consist of extensive networks of hydrogen bonds in which many consecutive residues are involved. These hydrogen bonds provide much of the enthalpy of stabilization that allows the polar backbone groups to exist in the hydrophobic protein core. Helices and sheets account to the majority of secondary structure common in proteins. However, these regular structures are interspersed with regions of irregular structure that are referred to as *loop* (or *coil*) regions. Loop regions are usually present at the surface of the protein. These regions are often simply transitions between regular secondary structures, but can also be the location of the functional part, or the *active site*, of the protein. Because of their irregularity, these elements are difficult to classify. However, some types of irregular structure have been categorized [196]. Beta turns or hairpins, for example, consist of a tight turn between two consecutive strands that reverses the direction of the polypeptide chain, stabilized by one or more backbone hydrogen bonds.

The main SSEs, helices and strands, are built by forming hydrogen bonds. To simplify the description of hydrogen bonding pattern and residues forming SSEs we define a function *hbond* representing a hydrogen bond between two residues:

Definition 1 (Hydrogen Bond)**.** *$Hbond(i, j)$ denotes a hydrogen bond between the C=O group of residue i and the N-H group of residue j. $Hbond$ is a* Boolean *function which is true when there is a hydrogen bond between the residues given as it parameters.*

2.4.1 Helices

Helices always contain a single consecutive stretch of amino acid residues. They are formed by hydrogen bonds between residues in the same helix, as shown in Figure 2.6c. Three different types of helices exist:

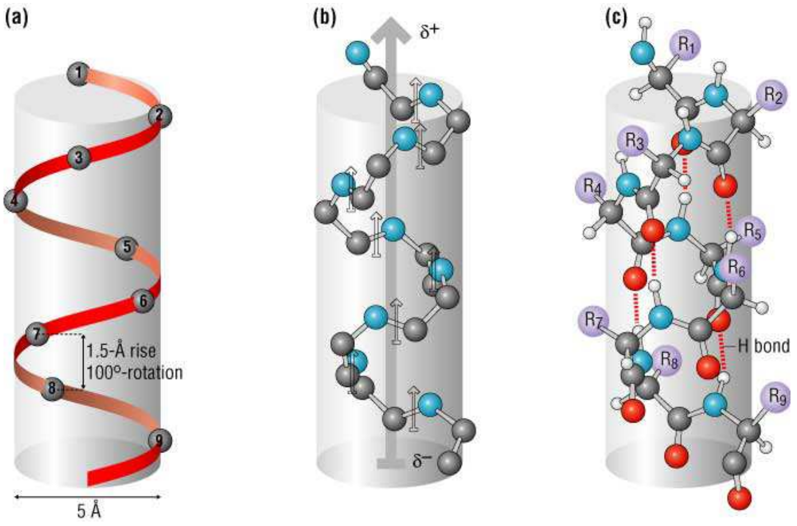


Figure 2.6: **Alpha helix.** (a) Overall shape (Cα-atoms are numbered from N- to C-terminus, backbone is shown in red). (b) Helix dipole moment ($C\alpha$-, C-, N-atoms). (c) Hydrogen bonding network (all backbone atoms). Color coding: C in grey, N in blue, O ind red, sidechain R in violett. (Source: `http://wiz2.pharm.wayne.edu/biochem/prot.html`).

- α-helix is made by successive hydrogen bonds:

$$Hbond(i, i+4), Hbond(i+1, i+5) \ldots$$

- 3_{10}-helix is made by successive hydrogen bonds:

$$Hbond(i, i+3), Hbond(i+1, i+4) \ldots$$

- Π-helix is made by successive hydrogen bonds:

$$Hbond(i, i+5), Hbond(i+1, i+6) \ldots$$

The most common type of helix in proteins is the α-helix, where two sequential neighbored residues are rotated by approximately 100° around the helix axis and translated along the axis by 1.50Å (Figure 2.6a). The atomic distance between N and O atoms measures 2.8Å. The hydrogen bonds are almost parallel to the helix axis and the total dipole moment points from the C-terminus (-) to the N-terminus (+) (Figure 2.6b). This helix dipole is important in the interaction of neighboring helices in the packing of secondary structure motifs into the 3D structure, where the backbone atoms are in van-der-Waals contact with each other across the helix axis. α-helices in known protein structures are almost without exception right-handed. The bonds forming helices restrict the torsion angles. The idealized angles for 'geometrically' correct α-helix are $\phi = -57°$ and $\psi = -47°$. The ideal structural parameters of helices (and sheets) are given in Table 2.1 and the regions for these parameters are illustrated in the Ramachandran diagram (Figure 2.5 left). However, the real angles usually

Table 2.1: **Structural parameters for protein secondary structures.** ϕ and ψ are the conformational angles of the backbone, n the number of residues per turn, d the displacement between sucessive residues along the SSE axis, and p the distance along the helix axis of a complete turn. (Source: [142]).

Structure	ϕ	ψ	n	d [Å]	p [Å]
α-helix	$-57°$	$-47°$	3.6	1.5	5.5
3_{10}-helix	$-49°$	$-26°$	3.0	2.0	6.0
π-helix	$-57°$	$-70°$	4.4	1.1	5.0
Polyproline II helix	$-79°$	$+149°$	3.0	3.1	9.4
Parallel β-sheet	$-119°$	$+113°$	2.0	3.2	6.4
Antiparallel β-sheet	$-139°$	$+135°$	2.0	3.4	6.8

deviate. Many α-helices in globular proteins present a hydrophobic face to the external aqueous solvent, and, on the opposite site, a hydrophobic face to the interior. Helix conformation satisfies the main chain hydrogen bonding potential of amino acid residues within the helix, except for those at the ends. Often special sequences and conformations appear at the so-called 'helix caps' stabilizing the helix termini. Proline residues in helices must interrupt the hydrogen bonding pattern, because proline does not have an N-H group. Some proline-rich helices can form special conformations, e.g., the polyproline II helix.

2.4.2 Strands and Sheets

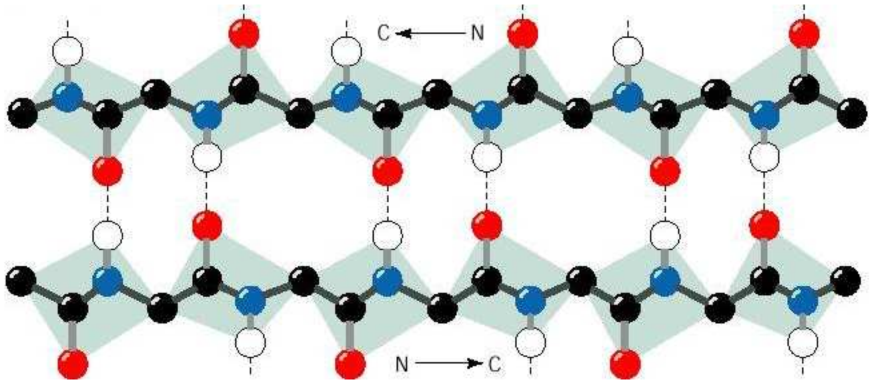


Figure 2.7: **Antiparallel β-sheet.** Color coding: black C, white H, blue N, red O. (Source: `http://cmgm.stanford.edu`).

An idealized β-strand should have torsion angles of approximately $\phi = -120°$ and $\psi = +120°$. Single β-strands are not stable structures but occur in association with neighboring strands. A β-sheet is formed from separate β-strands, which may arise from regions distant in the sequence. Sheets are formed by successive hydrogen bonds between residues, as shown in Figure 2.7.

The backbone hydrogen bonding groups *N-H* and *C=O* are planar, with the bonding groups from successive residues pointing in opposite directions.

Let residue i be in one strand, and residue j in another. Then the bonding of the two strands can be either parallel or antiparallel:

- Parallel bonding pattern is formed by each residue forming hydrogen bonds with two residues on two parallel strands of the sheet. Between residues of two neighbored strands, this means successive hydrogen bonding patterns:

$$Hbond(i,j), Hbond(j+2,i+2), Hbond(i+4,j+4), \ldots \ .$$

- Antiparallel bonding pattern is formed by each residue forming hydrogen bonds with two residues on two antiparallel strands of the sheet. Between residues of two neighbored strands, this means successive hydrogen bonding patterns (Figure 2.7):

$$Hbond(i,j), Hbond(j{+}2,i{-}2), Hbond(j{-}2,i{+}2), Hbond(i{+}4,j{-}4), \ldots \ .$$

The ideal structural parameters for β-strands and β-sheets are given in Table 2.1. Sheets can either be parallel, antiparallel or mixed (with both parallel and antiparallel hydrogen bonds). The relationship between the positions of the strands in a sheet in space, and their position in sequence is quite variable. It is possible to form an antiparallel sheet in which strands appear successively in the sequence, connected via turns or β-hairpins. In contrast, two adjacent parallel strands require a bridging segment, usually an α-helix. Many proteins contain a succession of β-α-β units.

The β-bulge is an irregularity in the hydrogen bonding pattern of a sheet often observed in an edge strand of sheets. If one imagines the edge strands of a β-sheet to form a hydrogen bond with each other, a closed structure is created, called a β-barrel, e.g., the eight parallel strands linked by eight helices in triose phosphate isomerase, also known as 'TIM barrel'.

2.4.3 Identifying Secondary Structure Elements

Although helices and strands are regular elements within protein structures, there is no unique definition or method to assign SSEs from atom coordinates. The term 'secondary structure element' is not used consistently in the literature. Irregularities in real structures make the identification of the SSEs difficult. Especially the ends or caps of helices and strands are hard to define, i.e., to exactly identify the residue where an SSE starts and ends. However, there does not exist a precise general definition for SSEs. The few definitions that exist are biased by the author's view. On the other hand, automatic methods for identification do exist. Most methods for SSE assignment use hydrogen bond definitions, because it is believed that the specific formation of SSEs is governed by intraprotein hydrogen bonds [110]. There are different methods for the assignment of hydrogen bonds [6]:

- **Angles**. Hydrogen bonds can be assigned using the angle θ between the *N-H* and *C=O* backbone groups of two distinct residues and the distance r_{HO} [19, 28, 112]:

$$\theta > 120^\circ \ and \ r_{HO} < 2.5\text{Å} \ .$$

- **Coloumb energy**. Hydrogen bonds can be determined by calculating the Coloumb energy of the bond, as applied in DSSP [119], using a purely electrostatic model. It assigns charges of $\pm q_1 \equiv 0.42e$ to the carbonyl carbon and oxygen atoms, respectively, and charges of $\pm q_2 \equiv 0.20e$ to the amide nitrogen and hydrogen atoms, respectively. The electrostatic energy is then given by

$$E = q_1 q_2 \left[\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right] * 332 kcal/mol \enspace ,$$

  where r are atom distances (the hydrogen atom is not included in most PDB files, i.e., it has to be extrapolated). According to DSSP, a hydrogen bond exists if and only if:

$$E < 0.5 kcal/mol \enspace .$$

- **Empirical calculations**. Hydrogen bonds are defined by empirical hydrogen bond calculations that can be derived form hydrogen bond geometry in crystal structures [27, 235] and is applied in STRIDE [74].

In the following we will shortly describe the two commonly most accepted methods for SSE assignment, DSSP [119] and STRIDE [74] that are used throughout this thesis. Beside these two methods, there exist many other methods like DEFINE [194], which uses $C\alpha$ distances, or P-Curve [213] that is based on mathematical analysis of protein curvature. There are also methods based on geometry, for example, Segno [52] that defines SSEs on a number of geometric parameters for backbone atoms. STICK [225] defines SSEs as linear line segments, independently of any external SSE definition. DSSPCont [5, 39] uses multiple runs of DSSP to introduce more flexibility into the SSE assignment and to capture geometric differences. These differences are due to thermal fluctuations, experimental uncertainties, or different solution or environment conditions.

Andersen and Rost [6] found that DSSP and STRIDE agree in 96% of all residues (see also Table 2.2), where disagreement is mainly related to helix assignment, whereas DEFINE and P-Curve agree only in 74% and 79% of all residue assignments with DSSP, respectively.

DSSP

The most commonly used program to define SSEs in protein structures is probably DSSP (Define Secondary Structure of Proteins) [119] that is mainly based on hydrogen bonding patterns. DSSP defines one of eight different secondary structure states for every residue of a polypeptide chain. Here, hydrogen bonds are described by an electrostatic model (see above).

A helix assignment (states H,G,I) needs two consecutive amino acids with the hydrogen bond patterns $Hbond(i-1, i+n-1)$ and $Hbond(i, i+n)$ for $n = (3, 4, 5)$ representing the three different helix types (see Section 2.4.1). Longer helices are defined by overlaps of shorter helices. Note that for residue i nothing is required about the hydrogen bond state of residue $i+1$, so fragments can be defined as helices without all the involved residues satisfying the hydrogen bonding criteria. β-sheet residues (state E) are defined as either having two hydrogen bonds in a sheet, or being surrounded from two hydrogen bonds in a sheet. One or more β-bulge residues that are also assigned as sheet residues can interrupt the sheet hydrogen bonding pattern. Parallel and antiparallel sheets are distinguished. A minimal sheet consists of two residues at each partner strand. β-bridges (state B) are isolated residues satisfying the hydrogen bonding property. All other residues are assigned with S (bend) or nothing, indicating loops and turns.

Stride

STRIDE (STRuctural IDEntification) [74] uses empirically derived hydrogen bond energies and backbone torsion angle criteria to assign SSE states. The torsion angles for helices and strands are defined according to allowed standard regions in Ramachandran plots [191]. A helix assignment (states H,G,I) starts, like in DSSP, with two consecutive hydrogen bonds, but is elongated only, in contrast to DSSP, if one of the edge residues has a correct torsion angles combination, i.e., hydrogen bonding pattern could be ignored, if the torsions angles are not acceptable. The minimal sheet (state E) needs two residues in one of five different hydrogen bond conformations with acceptable torsion angles. Bulges and bridges are also assigned according to similar criteria than in DSSP. Turns (state T) are assigned separately, as described in [241]. In all other cases the residue is assigned with C.

2.4.4 Amino Acid Distributions in SSEs

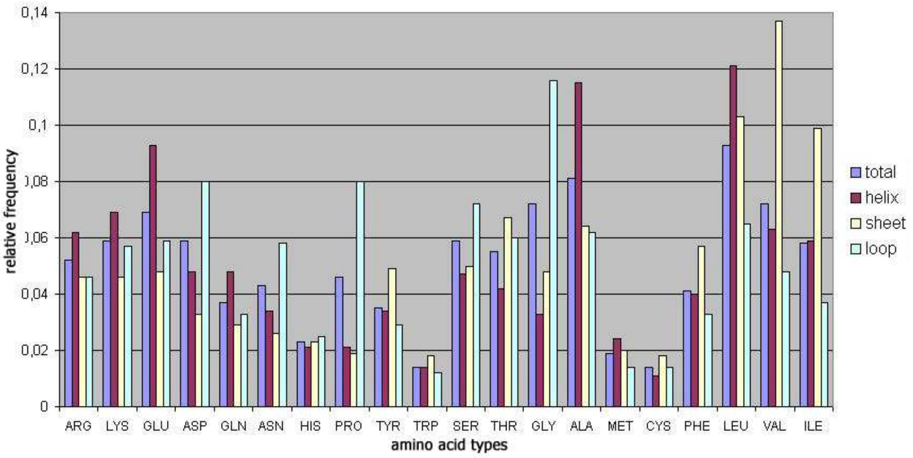


Figure 2.8: **Amino acid distributions.** Relative amino acid frequencies for the ASTRAL Scop40 dataset [41] using Stride [74] for SSE assignment.

It is well known that certain amino acid types are more preferred in α-helices, others in β-sheets, or loop segments. Since we are analyzing protein structures in the context of secondary structure and contact maps, it is interesting to see what amino acids are favorable in which type of SSE. Moreover, the amino acid type distributions are important to explore contact type propensities between certain types of amino acids in secondary structures. To evaluate the amino acid distributions of globular polypeptide chains we have examined the ASTRAL Scop40 dataset (see Section D.1 for details) using Stride [74] and DSSP [119] for SSE identification. This yields $5,397$ non-redundant globular protein domains (Table 2.2).

The resulting amino acid distributions for Stride are shown in Figure 2.8. We found similar propensities like studies conducted before [240]. Amino acids with long sidechains such as leucine, glutamine, and glutamatic acid can often be found in helices, because their extended sidechain can project far away from the crowded helix cylinder. In contrast, residues with sidechains branching at the $C\beta$-atom, such as valine, isoleucine, and phenylalanine, are found more often in strands than in helices, because every other sidechain in a sheet is pointing to the opposite direction, leaving space for $C\beta$-branched sidechains to pack. In most cases, proline is disfavored in both, helices and strands, because of its cyclic structure that prevent hydrogen bonding to other backbone atoms. Because of its missing sidechain, glycine is also less commonly present in SSEs. Both residue types are strongly associated with loops, especially beta turns. Table 2.2 gives the total residue distributions for the two different SSE assignment methods, DSSP and Stride. It is well known, see [6], that DSSP tends to assign helices too short, because the hydrogen bonding pattern of helix cap residues are incomplete. Therefore, Stride assigns much more helix residues and SSE residues in comparison to DSSP. The number of strand assignments of both methods is rather consistent.

Table 2.2: **Astral Scop40 dataset [41].** Total numbers for the Astral Scop40 dataset comprising totally $5,397$ domains with $991,784$ residues using Stride [74] and DSSP [119] for SSE assignment. *sse* gives the number of residues in helices and strands.

no residues	DSSP	Stride
loop	$459,198$	$431,507$
strand	$201,063$	$207,780$
helix	$331,523$	$352,497$
sse	$532,586$	$560,277$

2.5 Protein Domains and Structural Motifs

Most of the protein structures are globular resulting in compact shapes. Since globular proteins range in molecular weight from a thousand to over a million Dalton, one could think that the size of these folds given as its diameter would increase with molecular weight, but this is not the case at all. Proteins whose molecular weights are less than about $20,000$ Dalton often have a simple globular shape, with an average molecular diameter of 20-30Å, but larger proteins usually fold into two or more independent globular structures [35, 59] called structural *domains*. The notion that domains of large proteins are independently stable has been verified by cloning the corresponding DNA sequences and expressing them independently [99]. A domain is a compact region of protein structure that is often, but not always, made up of a continuous segment of the amino acid sequence, and is often capable to maintain its characteristic structure, even if it is separated from the overall protein. Domains vary in size but are usually not larger than the largest single-domain proteins, i.e., about 250 amino acids. Most domains have about 200 amino acids or less [59, 169] with an independent hydrophobic core.

Motifs (also referred to as *supersecondary structure*) are small substructures that are not necessarily structurally independent. Generally, they consist of only a few SSEs and the interactions between them. Supersecondary structure elements include the α-helix hairpin, the β-hairpin, the β-α-β units, or the Greek-key motif (see also Chapter 3). Specific structural motifs are seen repeatedly in many different protein structures. Most often they are integral elements of protein folds. Further, motifs often have a functional significance, and in these cases represent a minimal functional unit within a protein. Several motifs can combine to specific domains.

2.6 Protein Structure Evolution

Evolution is the major process responsible for the generation of biological diversity. The primary events are *mutation* and *deletion* of single nucleotides in DNA sequences, or transpositions of larger parts of genetic material. *Selection* reacts on the level of protein function as determined by protein structure. If a gene is coding for a functional protein, a mutant gene can produce either an alternative protein of equivalent function, a protein that carries out the same function but with an altered rate or specificity profile, a protein with an altered function, or a protein that does not function—or even fold—at all. Evolutionary events within populations, i.e., the change in distribution of DNA sequences among individual organisms, may occur through positive or negative selection, or by random fixation of variants without selective advantage. Examination of homologous genes and proteins in different species has shown that evolutionary variation and divergence occur very generally at the molecular level. The same proteins from related species have similar but not identical protein structures. Families of related proteins tend to retain similar folding patterns over ranges of sequence homology from near identity down to below 20%, e.g., myoglobin and its distant relative hemoglobin. The general folding pattern is preserved, but there are distortions that increase with the amount of sequence divergence. Generally, often only the protein core is conserved within protein families, and as the sequences diverge the structures progressively deform. In both closely and distantly related proteins the general response to mutation is conformational change. A simple way of altering a protein fold without significant destabilization is to change the sequential order of its constituting SSEs while maintaining their spatial arrangement. This can be done by the internally swapping of similar SSEs or by reversing the direction of some of its SSEs. There are many proteins with similar secondary structures and architectures but different topologies that could be related in such a way, e.g., an immunoglobulin domain and the plastocyanin fold. Residues active in function, such as the proximal histidine of the globins or the catalytic serine, histidine, and aspartate in serine proteases, are resistant to mutations because changing them would accompanied with loss of function. It is the ability of protein structures to accommodate mutations in non-functional residues. Therefore, loop and surface residues not involved in function are usually free to mutate.

Theoretically, SSEs can be combined to form a complete protein fold in an almost unlimited way. The universe of protein folds is called the fold space. Interestingly, currently available proteins structure data suggest that the fold space is in fact quite limited relative to the possible range of folds [46, 93] (see also Figure 1.1). The structural similarity between sequentially unrelated proteins are often explained by convergence to a stable fold as opposed to divergence from a common ancestor. The convergence implies not only that proteins are of independent origin but also that they had different original folds. With no evidence for their difference in original fold, they are referred as having undergone parallel evolution. Proteins that have descended from the same ancestor most often have similarity with that ancestor in sequence, structure, and function. Generally, strong sequence similarity alone is considered to be sufficient evidence for common ancestry. Close structural and functional similarity together is also accepted as sufficient evidence for distance homology between proteins that lack significant sequence similarity. But neither structural nor functional similarity alone is considered to be strong evidence for a common ancestor [168].

2.6.1 Divergent Evolution

In the case of *divergent* evolution, the number of protein folds is limited because they are derived from a relatively small group of shared common ancestor proteins. These early ancestor proteins would have discovered a stable fold, which has then been duplicated and reused by organisms for many other functions over the course of evolution. Presumably, modification of an existing fold is more likely to occur than the spontaneous generation of a new fold [228]. Although protein structure leads to protein function, similar protein structures will not always have similar functions. Many cases exist where two proteins have similar sequences and structures, but differ by a few key amino acid residues in an active site and hence have very different functions. Thus, it is important to consider the overall protein fold as a guide to the function of that protein, rather than a definition of the function. This functional versatility suggests the possibility that many protein folds will never seen in nature because organisms have simply not required or developed them [142].

2.6.2 Convergent Evolution

In the case of *convergent* evolution, the number of protein folds is limited, because certain folds are much more biophysically favored [90], and so have been created independently in multiple cases. Certain folds are clearly over-represented in the set of known structures, e.g., TIM barrel structures. In some cases, there is no detectable sequence similarity between proteins sharing the similar fold, suggesting that they have converged on a similar structure independently and do not share a common ancestor [105]. Therefore, it is possible that many possible folds will never be seen, because they are not structurally favored, and other more favorable folds can be adapted to the needed functions.

2.7 Protein Structure Representations

The fundamental 3D protein structure description consists of the specification of the coordinates for each atom, as given in the PDB [22]. The coordinates are determined by either X-ray crystallography or NMR spectroscopy. In structure comparison it is common to let one or two atoms represent one residue, often the Cα-atom. Sometimes the coordinates of Cβ-atoms or coordinates representing the 'mean' atom of the sidechain are used, in order to include some information on the orientation of the sidechains. A very common representation is the torsion angle representation of the backbone that is described in Section 2.2.

Generally, we can say that a protein structure consists of elements: atoms, residues, or SSEs. A protein structure description, therefore, describes different features of these elements [65]:

- The spatial arrangement of elements is called *architecture*. When the elements are atoms or residues, the architecture is sometimes called *geometry*.
- The ordering of the elements along the backbone together with their spatial arrangement is often referred as *topology* (see Chapter 3).
- The properties of elements can be given by, e.g., the physico-chemical properties of single amino acids or types of SSEs (see Section 2.1).

Labeled graphs can be used to represent all element-based features, with nodes representing the elements and the edges representing the relations between these elements. For instance, nodes can represent residues or SSEs, and the edges can represent contacts, distances, or angular relationships between residues, or number of contacts, geometric relationships, or type of parallism between SSEs.

In this work, we use contact maps and protein graphs as the main protein representations. Both concepts are introduced within the next two subsections. More explanations will follow in the subsequent chapters.

First of all we want to give a general definition of graphs that is used throughout this thesis:

Definition 2 (Graph)**.** *An undirected labeled graph $G = (V, E)$ is defined by a finite vertex set V and a set of undirected edges $E \subseteq \mathcal{P}_2(V)$. $\mathcal{P}_2(V)$ is the set of all subsets of V with exactly two different vertices.*

2.7.1 Contact Maps

A contact is a concept that has been introduced to state that two amino acids that are very close in 3D space are possibly able to form some kind of chemical bond [88]. A *contact map* of a protein structure shows the pairwise interactions or contacts between elements, in our case residues (Figure 2.9 (top)). In this way it is a detailed 2D representation of the 3D fold of a polypeptide chain [136].]. One or more atoms commonly represent a residue.

Definition 3 (Contact Map)**.** *The contact map of a polypeptide sequence $S = (r_1 r_2 \ldots r_n)$ containing n residues can be represented for a given contact definition by an undirected graph $G(S) = (V_s, E_s)$ with $E_s \subseteq V_s^2$ and V_s as the set of all residues of S, i.e.,*

$$V_s = \{r_i | 1 \leq i \leq n\} \ , \tag{2.1}$$

and E_s as the set of edges:

$$\forall r_i, r_j \in V_s : e_{ij} = \begin{cases} 1 & \textit{if } r_i \textit{ and } r_j \textit{ are in contact,} \\ 0 & \textit{else } . \end{cases} \tag{2.2}$$

The contact map graph can also be represented as a $n \times n$-Matrix C, where C_{ij} is defined as:

$$C_{i,j} = \begin{cases} 1 & \textit{if } e = (r_i, r_j) \in E_s \textit{ exists,} \\ 0 & \textit{else } . \end{cases}$$

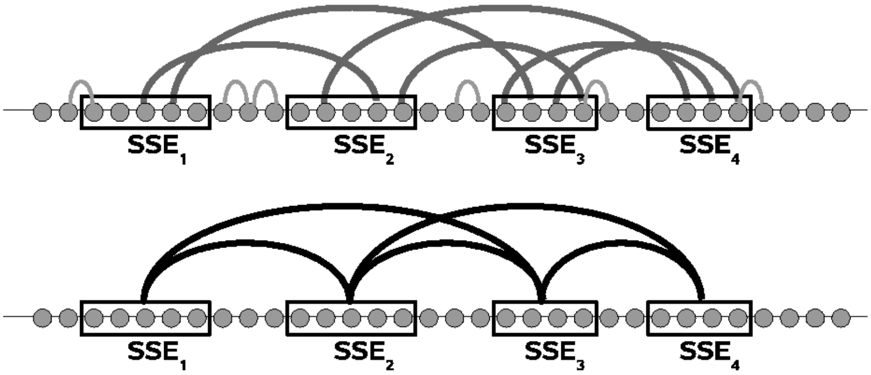


Figure 2.9: **Contact map and protein graph**. In the contact map (top) each residue of the polypeptide chain is represented by a circle, each SSE is marked as a box framing all residues located in that SSE. All residue interactions residues are expressed by an edge between these two residues. In the protein graph (bottom) these edges are replaced by only one edge connecting two SSEs if there is a contact according to the used contact definition for SSEs.

Instead of contact maps often distance matrices [104] are used to represent protein structures. A distance matrix shows the pairwise distances between residues instead of the non-dimensionally contact definition in contact maps. Certain atoms, e.g., Cα-atoms, represent the residues. The distances between these residues are given in Å rounded to integers, and only distances smaller than a certain threshold are used.

Contact maps as well as distance matrices contain enough information to reconstruct the 3D structure [231], except for handedness or chirality [65]. Contact maps are often used in protein structure alignment [38,88,136], but also used as basic information to elucidate protein structures from NMR spectroscopy [95].

2.7.2 Protein Graphs

The protein graph of a protein chain describes which SSEs (or *cores*) are in spatial relationship to each other in the 3D protein structure (see Figure 2.9 (bottom)). Distances, numbers of contacts, or types of parallism, can describe this relationship, for example.

Definition 4 (General Protein Graph)**.** *The protein graph PG is defined as labeled undirected graph $PG = (V_c, E_c)$ with $E_c \subseteq V_c^2$. Here, V_c is defined as the set of m SSEs $V_c = \{c_i | 1 \leq i \leq m\}$ and E_c is the set of edges with*

$$\forall c_i, c_j \in V_c (i \neq j) : e_{ij}^c = \begin{cases} 1 & \textit{if } \exists \textit{ contact between } c_i \textit{ and } c_j \ , \\ 0 & \textit{else} \ . \end{cases} \tag{2.3}$$

Each node in PG represents an SSE of consecutive residues within the protein chain while each edge represents an interaction between the two SSEs according to the used contact definition between SSEs.

In fold recognition algorithms [158, 243], also all neighboring SSEs are often connected via edges to represent the sequential ordering of the SSEs explicitly. Protein graphs are useful for motif descriptions within protein structures (see chapter 3) and protein structure alignment (see Chapters 5 and 6).

2.7.3 Contact Definitions

Contacts between residues can be defined in various ways. In most applications, distance-based contacts are used. Most often, the contact distance R is obtained by observing only the Cα- or Cβ-atoms [231]. Other methods use the centers of mass to represent every residue [100]. *All atom* contacts are based on the smallest distance R between any atom of two residues: the residues are in contact if R is smaller than a given threshold [101, 161]. Additionally, there are geometry-based contact definitions, like Voronoi [250] or Delaunay [222] tessellation to determine neighboring contacts, overlapping van-der-Waals radii [125], or convex hulls [53]. Here, we introduce in more detail contacts defined by Voronoi tessellation and van-der-Waals radii that are used throughout this thesis.

Voronoi Contacts

Residue contacts within protein structures can be defined by determining spatially directly neighbored residues using Voronoi tessellation [233] of the protein structure, a contact definition, often used for protein structure comparison [26, 111, 197]. The main benefits of using Voronoi defined contacts are, in contrast to distance-based definitions, that the nearest-neighbor contacts are well-defined containing only those residues that share a common face in their Voronoi cells, and that residues not only have to be close in space but must also be directly neighbored without any other residue in between.

For any discrete set S of points in Euclidean space, we use the Cβ-atoms (for glycine the Cα-atom) of the protein, a Voronoi decomposition partitions the space into convex polyhedrons, called Voronoi cells. Each residue cell contains by definition all points that are closer to the corresponding Cβ-atom than to all other points. All polyhedrons, which are directly neighbored in space, share a common face corresponding to a residue contact between two residues. A detailed description how to calculate Voronoi contacts can be found in [26, 250].

Van-der-Waals Contacts

Contacts between residues can also be determined using van-der-Waals radii of corresponding atoms. Two residues are said to be in contact, if the van-der-Waals radii of at least two atoms are overlapping in space. The van-der-Waals radius of an atom is the radius of an imaginary hard sphere. From quantum-mechanics it is well known that atoms are no hard spheres at all, but, on the other hand, the location of atomic nuclei are accessible by experimental methods, like X-ray spectroscopy, for example. The van-der-Waals radius is named after Johannes Diderik van der Waals, winner of the 1910 Nobel Prize in Physics. Van-der-Waals radii are determined from contact distances between non-bonded atoms. Reliable values cannot be given for these radii because they depend on the measuring technique and the molecule (see [174]). The most common used values are listed in Table G.2 in the Appendix.

Distribution of Different Contact Types

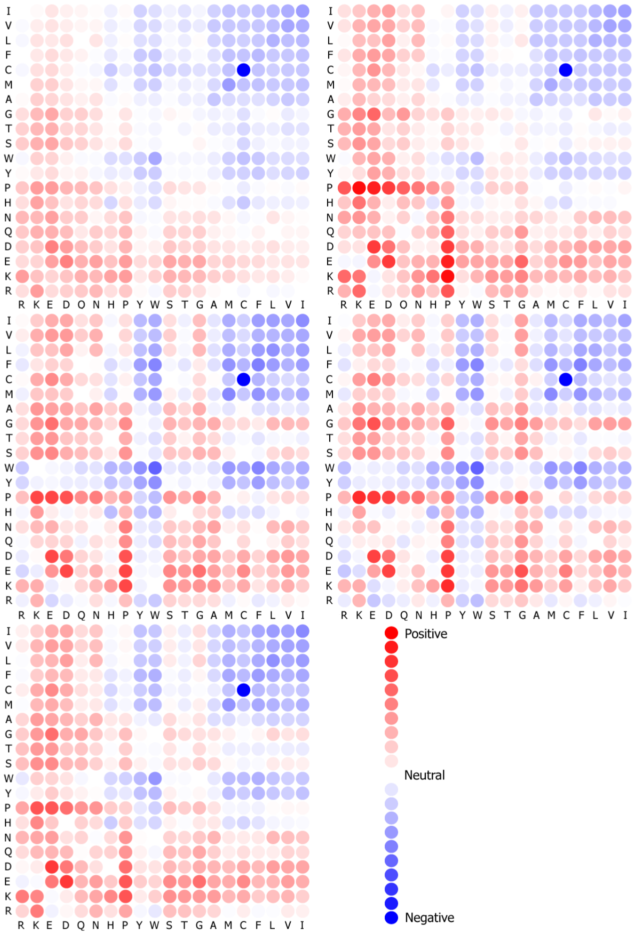


Figure 2.10: ***Total* contact potentials**. All residue potentials for the five contact definitions using Stride [74] for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.

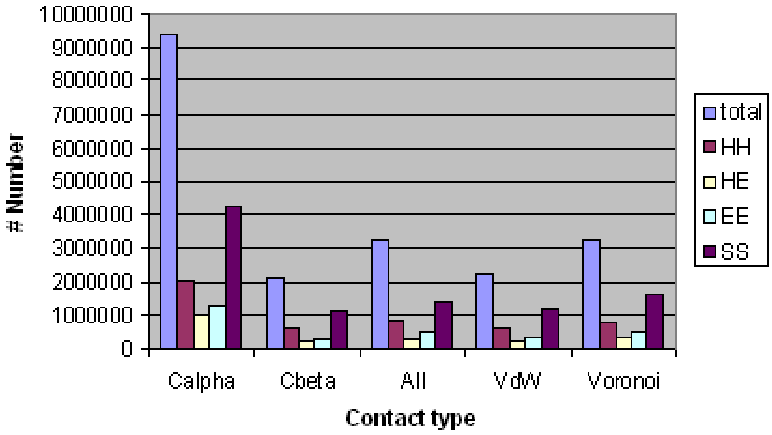


Figure 2.11: **Contact type distributions using Stride [74]**. Total numbers for the five contact type subsets defined in the text (HH: helix-helix, HE: helix-strand, EE: strand-strand).

Certain pairwise amino acids contacts are more probable between two helices than between two strands or between strands and helices, respectively, especially for different contact definitions. To evaluate the contact propensities for pairwise amino acid contacts within globular polypeptide chains we have examined the ASTRAL Scop40 dataset (see Section D.1 for details) using both Stride [74] and DSSP [119] for SSE identification. Only contacts between residues were considered that were separated at least two residues in the amino acid sequence. We have used five different contact definitions:

1. *all* contacts: a contact is defined, if the distance between any two atoms of two different residues is smaller than 5Å.

2. *ca* contacts: a contact is defined, if the distance between any two Cα-atoms of two different residues is smaller than 11Å.

3. *cb* contacts: a contact is defined, if the distance between any two Cβ-atoms of two different residues is smaller than 7Å. For glycine the Cα atom is considered.

4. *vdW* contacts: a contact is defined, if the distance between any two atoms of two different residues is smaller than $(R_1 + R_2)$ with R_i as the van-der-Waals radius for the atom type of atom i as given in Table G.2.

5. *vor* contacts: Voronoi contacts are calculated as described above using the program as described in [26].

We calculated for every contact definition type all contacts for the following five contact subsets:

1. *total*: all contacts between all residues.

2. *SS*: all contacts between residues within SSEs.

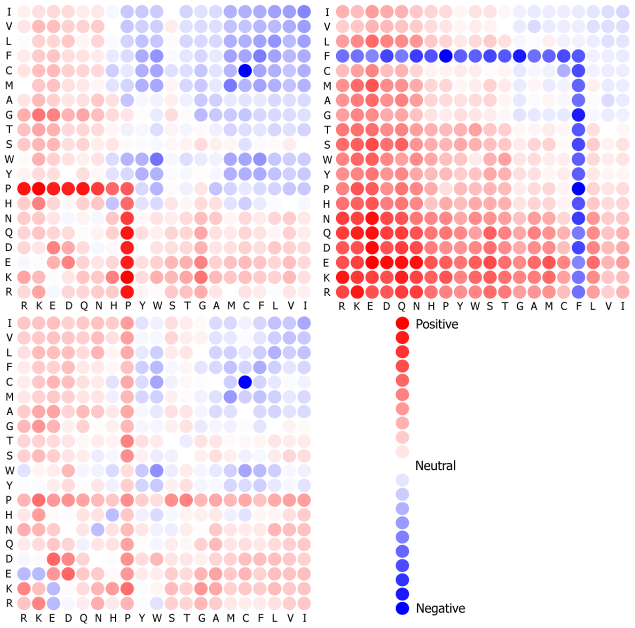


Figure 2.12: **SSE *ca* contact potentials using Stride [74].** The *helix-helix* (top left), *helix-strand* (top right), and *strand-strand* (bottom left) residue potentials for the *ca* contact definition using Stride for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red).

3. *helix-helix*: all contacts between residues within helices.

4. *strand-strand*: all contacts between residues within strands.

5. *helix-strand*: all contacts between residues where one residue is part of a strand and the other residue part of a helix, or vice versa.

Statistical potentials are a common way to describe amino acid contact propensities. They are used for simulating protein folding, judging the quality of proposed protein models, as well as in protein structure prediction methods including pairwise residue interactions [150, 158, 243]. In Appendix B we give a short introduction into statistical potentials, as well as a description how the different potentials were calculated. We used 5 different contact definitions together with 5 different contact subsets and 2 different SSE assignment methods yielding 50 different statistical pairwise contact potentials in total. The pairwise contact potentials are shown in Figures 2.10, 2.13 and 2.12 or in the Appendix F.1. The propensities using Stride and DSSP were very similar, so we decided to discuss here only examples using Stride for the SSE assignment. Since the propensities are given as pseudo energies, negative energy values show frequent occurences, and positive energy values rare occurences of certain contacts, respectively. For all pairwise contact potentials we use the following color coding: negative energies are shown in blue, positive in red, neutral in white;

the higher the absolute value of an energy, the more intensive the coloring, see, for example, Figure 2.10.

The resulting contact distributions are shown in Figure 2.11. Using Cα-atom contacts resulted in about 3 times to 4 times more contacts than with any other contact definition. The reason for that is the relatively large distance cutoff of 11Å, which can assign residues contacts that are clearly separated in 3D. In contrast, $C\beta$-atom and van-der-Waals contacts show much less contacts, because of the low distance cutoff values that were used. All atom and Voronoi cell contacts show approximately the same contact distributions over all subsets because of their much stricter neighborhood definition. Figure 2.10 shows the pairwise residue propensities for all residues in the dataset for the five contact definitions. Cysteine-cysteine contacts have always the best energy values. Cysteines are not very frequent in proteins, but in case they are present they often form disulphide-bridges. For all interaction types contacts between hydrophobic residues are highly favored, as well as contacts where residues with long polar sidechains are involved, like tryptophan and tyrosine. Proline is for all contact definitions the most disfavored amino acid type, especially for helix-helix interactions (see Figure 2.12), since it is known as 'helix breaker'. For Cα-atom contacts proline is more often involved in contacts as for the other four contact types, again because of the large distance cutoff. The statistical potentials for contacts between SSEs in general, illustrated in Figure 2.13, show some qualitative variations. Again, contacts between hydrophobic residues are favored, but additionally for the more sequence-based contact types contacts between residues with opposite charges are more favored than in the total contact distributions (see, e.g., the *all* contacts in Figure 2.13). Almost all amino acid types are disfavored for helix-strand contacts, except hydrophobic residues. A special case is phenylalanine that appears more often in sheets than in helices and that is able, due to its branched sidechain, to come in contact distance with residues of the other SSE type (Figure 2.12). But it could also be an artifact of the used dataset. All these propensities reflect that hydrophobic residues are highly engaged in van-der-Waals interactions or hydrogen bonding within the tertiary structure, as well as polar or charged residues have to compensate their dipole moments or charges, respectively, during the folding process. Generally, the Cα-atom contacts are more flexible because of their larger number and the used distance criterion. Since all other contact definitions using sidechain information or nearest-neighbor definition, they are much more sequence-specific in terms of possible interaction partners.

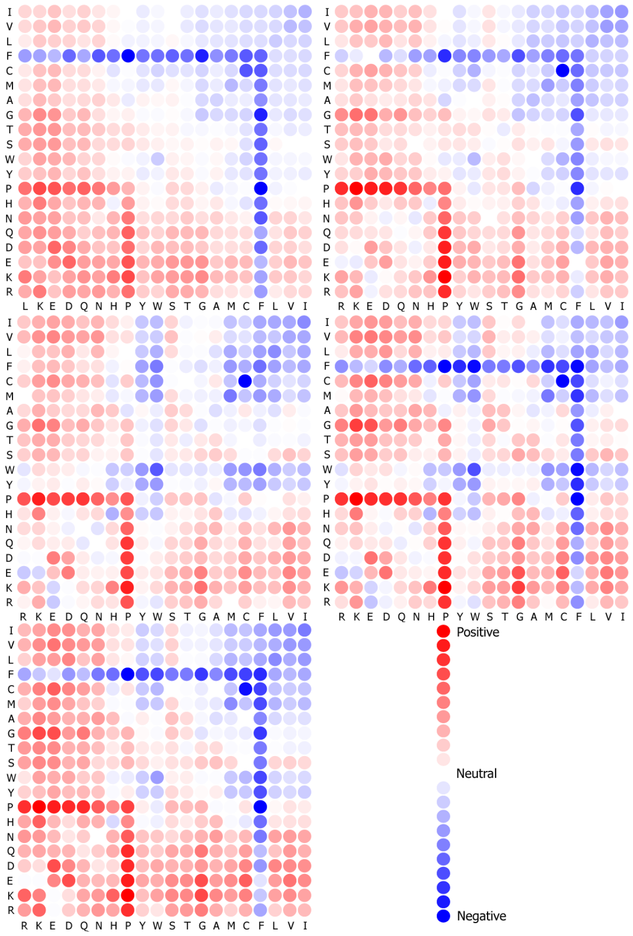


Figure 2.13: ***SS* contact potentials using Stride [74].** The SSE residue potentials for the five contact definitions using Stride for SSE assignment. The color spectrum goes from the most negative energy (blue) over neutral (white) to the most positive energy value (red). Top left: *ca*, top right: *cb*, middle left: *all*, middle right: *vdW*, bottom left: *vor*.