THE UNIVERSITY OF CHICAGO

SOLVATION SIGNATURE IN HYDROGEN BOND GEOMETRY OF PROTEIN HELICES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
TATIANA ORLOVA

CHICAGO, ILLINOIS
MARCH 2017

To my dear parents

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

We thank our adviser L. Ridgway Scott, and the committee members Ariel Fernández Stigliano, and Stuart Kurtz for valuable suggestions regarding this work.

ABSTRACT

As the most abundant type of protein secondary structure helices play an essential role within a protein and in various protein-protein interactions. Thus it is especially important to understand what criteria influences the geometry of helices and helps successfully perform their functions. There are many computational tools used by protein biophysicists. However, it is rare for them to use computer algebra systems. Thus we explored the use of such systems to show how they could be used to study structural properties of proteins. As an example, we chose the geometry of helical structure.

We begin by considering two types of protein helices, α- and 3_{10}-helices stabilized by 1-5 and 1-4 hydrogen bonding pattern respectively and study the relationship between hydrogen bond geometrical requirements and stability of protein helices via mathematical optimization. In particular, we take two major hydrogen bond requirements: linearity and length constraints and ask whether the most common α- and 3_{10}-helix motifs in protein folds result from optimization with respect to a linear combination of these two criteria. We show that these criteria are not sufficient to explain the observed angles. Instead, we suggest that maximizing the solvation of the protein backbone has a significant effect on the observed (φ, ψ) angles.

The above work suggested that a completely unexpected "solvation signature" should be observable in protein structure. There are many tools that can be used to study this suggestion. Since "data science" is a theme of significant current interest, we explored this approach to see what issues arise with these techniques. So, as a next step, we investigated the effects of solvation by collecting and analyzing a high quality protein dataset. We found that helical backbone actively interacts with water irrespective of whether it is located at the surface or buried inside protein. This interaction, as expected, highly correlates with larger φ angles and larger distances between neighboring main-chain carbonyl oxygens. Moreover, we observe a distinct periodic backbone solvation pattern in α-helices, suggesting that most helices have a very specific orientation and position specific residue preferences.

We have seen that new tools can enhance the study of protein biophysics. The success of data mining depends strongly on the quality of the questions being addressed as well as the quality and quantity of the data. This suggests that data science (1) needs to have a firm foundation in basic science and (2) needs to have appropriate analytic tools to examine the data faithfully.

Other tools, such as molecular dynamics and density functional theory have also been used to study protein-water interactions. Given the right questions to ask, these too can be potentially useful and would be interesting to consider in the future.

CHAPTER 1

INTRODUCTION

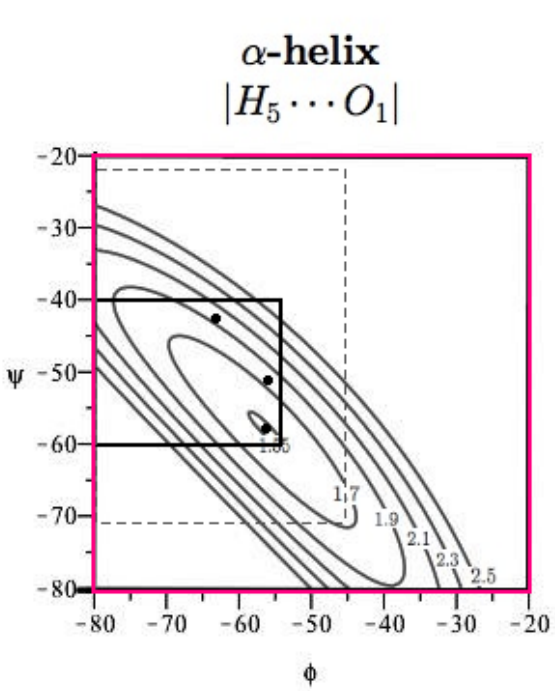
1.1 Hydrogen bonds

The concept of the hydrogen bond was established by 1920 [42], and possibly earlier [38, 20]. Hydrogen bonds are the most important bonds in biochemistry, so there is a compelling reason to understand them in some depth. Unfortunately, there are several challenges. Although hydrogen bonds in proteins have been considered extensively [8, 67, 41], they are not yet fully understood and are still actively studied [33, 34]. In particular, it is difficult to model hydrogen bonds quantitatively with any generality and precision.

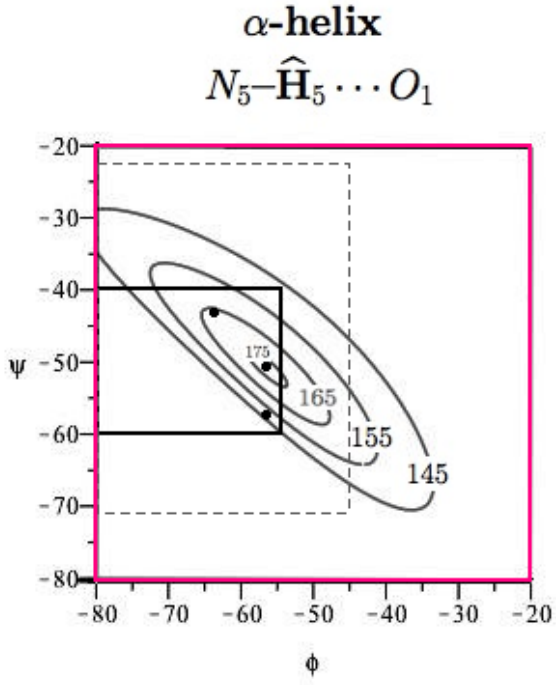
1.1.1 Hydrogen bond models

One study concluded that the hydrogen bond length "is primarily a function of the degree of positive charge on the hydrogen in the H bond" [40]. One might hope that modeling the hydrogen bond as a simple dipole-dipole interaction would be sufficient to capture the angular dependence. But a model of hydrogen bonds based only on partial-charges does not represent well the angular dependence of the energy: "At the distances where H bonding occurs, the dipole moment approximation is a poor one and higher multipoles must be considered" [40]. Following this suggestion, several approaches to modeling hydrogen bonds have involved more sophisticated interactions, including dipole, quadrapole and higher representations of the donor and acceptor groups [16, 17]. One model of hydrogen bonds in water incorporates polarization terms [12, 53]. Unfortunately, the multipole expansion converges slowly for small separations of the donor and acceptor, although these models give useful representations of the asymptotic behavior for large separation distances R. Distributed multipole expansions involving partial charges (and dipoles, etc.) at several positions, have been studied [65, 66].

The authors of [54] proposed a hydrogen bond model based only on the Lennard-Jones potentials of the atomic distances among the donor and acceptor atoms. This model utilizes



(a) Contours for different values of the distance $|H_5 \cdots O_1|$ as a function of φ, ψ are plotted for α-helices. The dot in the smallest contour indicates the (φ, ψ) pair that minimizes the H-O distance in α-helices.



(b) Contours for different values of the angle $\nu = N_5\text{–}\widehat{\mathrm{H}_5 \cdots O_1}$ as a function of (φ, ψ) are plotted for α-helices. The dot in the smallest contour indicates the (φ, ψ) pair that minimizes the NOH angle defined subsequently in Figure 1.2.

Figure 1.1: The dots in the smallest contours in the respective figures indicate the (φ, ψ) pair that minimizes the (a) H-O distance and (b) the NOH angle in α-helices. The remaining dot, to the upper-left, indicates the observed $(\varphi, \psi) = (-63°, -43°)$ pair that occurs most commonly in α-helices. Solid black lines indicated the "allowed regions" of (φ, ψ) space that do not incur any backbone steric clashes. Dashed black lines indicated the "outer limit" of allowed regions of (φ, ψ) space regarding steric clashes as indicated in Table 2.2.

precisely the same positions as a dipole model with a more involved form for the potential energy based on data derived from ab initio quantum chemistry calculations. Curiously, the most sensitive term in their model appears to be a strong repulsion term between the like-charged atoms. Several other models exist [19, 21] but the diversity of models does not imply a broad understanding. Rather, it is a signal of the difficulty of achieving a simple, accurate model.

One way to understand the hydrogen bond is to use data mining to characterize its behaviors in protein structures. We do so here, limiting the study to a large class of hydrogen bonds in which all of the donor and acceptor atoms are the same, namely, those associated with two closely related backbone helical structures. Any useful hydrogen bond model must give predictions consistent with this data, so it provides a useful starting point.

Protein backbone helical structures can be characterized by the values of dihedral angles ϕ and ψ which can vary from unit to unit but in the ideal case can be uniquely determined by a single pair (φ, ψ). We focus our attention here on two common types of protein secondary structure: α- and 3_{10}-helices which are stabilized by hydrogen bonds between residues $i \rightarrow i+4$ and $i \rightarrow i+3$, respectively. Even though both α- and 3_{10}-helices in actual protein structures each represent a large class of individual helices, various studies have shown that dihedral angle pairs $(-63^\circ, -43^\circ)$ and $(-60^\circ, -25^\circ)$ characterize the most common helices within each group respectively. We will see that the constraints that dictate these particular values are not obvious.

1.1.2 Distance and angle optimization

We can imagine several different criteria that would dictate the ideal dihedral angle pairs, e.g., (1) hydrogen bond length or (2) hydrogen bond angles. It turns out that neither of these are decisive for helices in proteins. For example, Figure 1.1 shows that the contours related to different possible hydrogen bond lengths center around a distinctly different (φ, ψ) pair. It appears that it would be possible to form much closer associations between backbone

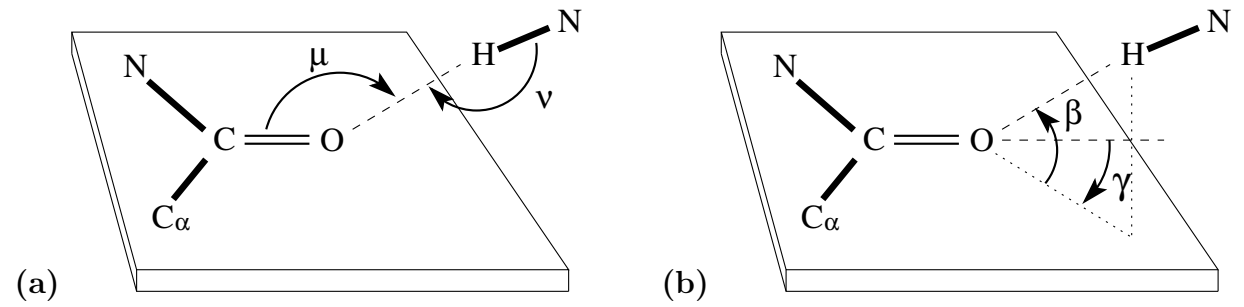


Figure 1.2: Hydrogen bond geometry, styled after [3, Figure 11, page 751] which was reproduced in [8, Figure 22, page 143](a) Angles $\mu = \widehat{COH}$ and $\nu = \widehat{NHO}$ used to define hydrogen bond quality. (b) Angles β (out of plane) and γ (in plane) were defined in [3, Figure 11, page 751] and later used in [8, Figure 22, page 143]

amide and carbonyl groups than actually occurs in nature in α-helices. O-H distances of less than 1.6Å appear possible, instead of the observed distances of about 1.9Å.

There are many angles of interest in defining and analyzing hydrogen bonds, as shown in Figure 1.2. One angle in particular, the $\widehat{NHO}$ angle denoted by ν in Figure 1.2, relates to the linearity of the hydrogen bond. Unfortunately, this angle is also not decisive in determining α-helical structure. Figure 1.1 shows the contours related to different values of this angle, and the "ideal" value of $\nu = 180$ degrees leads us to a (φ, ψ) pair distinctly different from the observed pair. Moreover, it is intuitively clear that optimizing some combination H-O distance and NHO angle will remain somewhere in the lower-right corner of the allowed regions of (φ, ψ) space, not near the empirically observed values. On the other hand, Figure 1.1 confirms that the empirical (φ, ψ) values have a reasonable NHO angle, about 165 degrees. That is, the NHO angle is not very sensitive to choice of (φ, ψ) in this region. Thus we are forced to look for another explanation.

1.2 The role of solvation in proteins

1.2.1 Solvation optimization

It appears that the most decisive criterion for determining helix formation geometry is more subtle. To lead us in the right direction, we recall some history relating to the realization

that hydrogen bonds could actually cause the stabilization of helices in protein structure. Indeed, this was not always believed to be the case.

The role of hydrophobicity in protein chemistry was not firmly established before 1959 [35]. Soon afterward [28, 39], the role of hydrophobicity regarding the strength and stability of hydrogen bonds in proteins was examined. However, the story requires a careful interpretation. The paper [39] studied a model molecule, N-methylacetamiden, that forms the same kind of amide-carbonyl (NH–OC) hydrogen bond formed by the protein backbone. Infrared absorption measurements were performed to assess the strength and stability of these hydrogen bonds in various solvents (including water) with different degrees of polarity. The paper's main conclusion could be misinterpreted as suggesting that hydrogen bonds are not significant for proteins in water [39]: "It seems unlikely, therefore, that interpeptide hydrogen bonds contribute significantly to the stabilization of macromolecular configuration in aqueous solution." However, the authors did confirm the opposite view that in less polar solvents hydrogen bonds could contribute to protein structural determination. We can now say that the paper [39] demonstrated the role of hydrophobic protection of hydrogen bonds in proteins. The key contributor of this hydrophobic protection are the nonpolar carbonaceous group CH_n, $n = 0, 1, 2, 3$, a common constituent of protein sidechains.

A subsequent paper [28] also studied model molecules, including N-methylacetamiden, in solvents based on varying ratios of trans-dichloroethylene and cis-dichloroethylene, via infrared spectroscopy. They concluded that "the free energy and enthalpy of association of the amides can be expressed as a function of the reciprocal of the dielectric constant." Although the variation in dielectric constants achieved with these solvents only reached a level of one-tenth that of water, this paper quantified the effect of dielectric modulation on the strength and stability of hydrogen bonds in systems similar to proteins. Thus it remained only to connect the variation in the dielectric constant to quantifiable variations in protein composition.

Although the energetic role of peptide hydrogen bonds remains a subject of significant

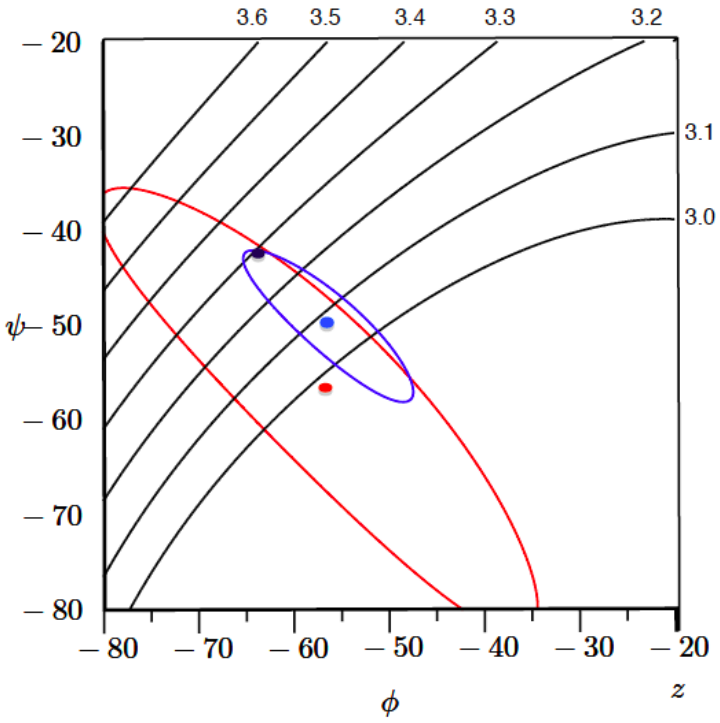


Figure 1.3: Contours for various values of distance $d_{O_iO_{i+1}}$ as function of angles ψ_{i+1} (Y-axis) and φ_{i+1} (X-axis) is shown in black for the case of ideal (Pauling) geometry. It is clear that the distance $d_{O_iO_{i+1}}$ is increasing with increasing φ_{i+1}. Points for (φ, ψ) pairs corresponding to the shortest hydrogen bond (-56.5°,-57.5°)(red), linear hydrogen bond (-56.5°,-50.5°)(blue) and "canonical peak" (-63°,-43°) (black) are also shown. Contours for the hydrogen bond length equal to 2Å(red), and angle $NHO = 165°$ (blue) are shown to illustrate better the point.

interest [9, 10], it now seems clear that the variation in hydrophobicity in proteins has a significant and quantifiable effect on the behavior of proteins [22]. According to [70], "The prevailing view holds that the hydrophobic effect has a dominant role in stabilizing protein structures." Thus we conclude that formation of α-helices is intimately related to the removal of water by hydrophobic groups. But if we apply the rule "if some is good, more is better" we are led to a contradiction. For example, we might think that the strength and stability of α-helices increases monotonically as a function of increasing numbers of nonpolar carbonaceous groups in the sidechains. But it is more complicated than that.

It is known that poly-alanine naturally forms α-helices in solution. Paradoxically, poly-valine does not so readily form α-helices in solution [47, 6, 7]. More precisely, the order of propensity for helix formation is [47, Fig.1, page 4931] as follows:

$$\text{Ala} \approx \text{Leu} > \text{Ile} >> \text{Val} >>> \text{Gly} \ .$$

This can be rationalized by realizing that poly-alanine has the perfect balance: it has some hydrophobic protection but it has fewer hydrophobic groups than valine near the backbone carbonyl oxygen, thus promoting solvation. Such solvation yields additional hydrogen bonds with the backbone (with solvating waters). This is apparently more energetically favorable than the more hydrophobic environment of poly-valine which would better protect its backbone hydrogen bonds. On the other hand, once the hydrophobic protection increases to the level of poly-leucine or poly-isoleucine, helix formation again becomes more viable.

It appears that the (φ, ψ) in typical helices can be characterized by a balance between two competing criteria: maximizing hydrogen bond linearity and maximizing the distance between neighboring main-chain oxygens $d_{O_i O_{i+1}}$, as indicated in Figure 1.3. The latter criteria we think is important for helix solvation in water. As shown in Figure 1.3, and developed in full detail subsequently in Chapter 2, the contours related to Hbond linearity and distance are similar. Thus we could equivalently characterize the optimal (φ, ψ) angles by a competition between maximizing the distance between neighboring main-chain oxygens $d_{O_i O_{i+1}}$ and minimizing the Hbond distance.

What is striking in Figure 1.3 is that the contours for $d_{O_i O_{i+1}}$ are orthogonal to all other contours, making it the leading candidate for explaining the dominant values of (φ, ψ). Many other contours related to other distances and angles will be presented subsequently in Chapter 2. Maximizing $d_{O_i O_{i+1}}$ has a simple interpretation in terms of hydration. The bigger the distance, the more room for a water molecule to enter and make a hydrogen bond.

Given that hydration is being maximized, this fixes the value of φ. Then the choice of ψ is largely determined by optimizing the angle $\widehat{NHO}$ as shown in Figure 1.1. That is, if we primarily were optimizing the hydrogen bond distance shown in Figure 1.1, then a smaller (more negative) value of ψ would have been obtained.

The choice of (φ, ψ) angles also dictates the distance between the oxygen on the backbone carboxyl (C-O) group and the C^{β} carbon on the succeeding residue (except for glycine). We

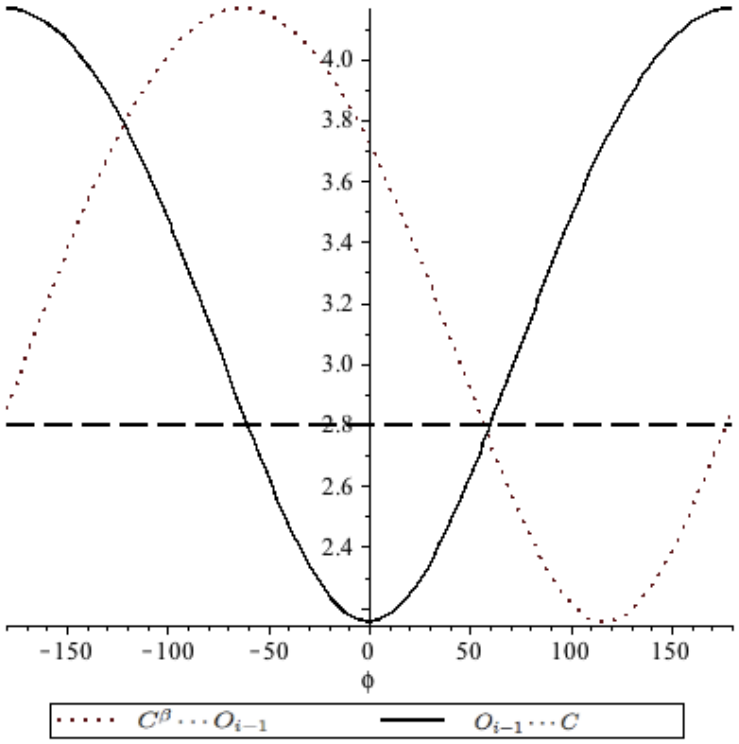


Figure 1.4: Distribution of α-helix interatomic distances (measured in Å) $O_{i-1}\cdots C_i$ (solid line) and $O_{i-1}\cdots C_i^{\beta}$ (dotted line) parametrized by φ. The distance $O_{i-1}\cdots C_i$ is one of the parameters related to a potential clash in a classic view [30], whereas the distance $O_{i-1}\cdots C_i^{\beta}$ is more related to exposure of the backbone to water. Note that the optimal empirical value of $\varphi = -63^{\circ}$ is very close to the maximum of the latter distance.

see in Figure 1.4 that the empirical values for (φ, ψ) angles largely maximize this distance, but variation in φ has little effect on the distance since nearby values of φ are close to the maximum distance. The distance $O_{i-1} \cdots C_i^\beta$ is defined to be the distance between the backbone oxygen O_{i-1} for the peptide $i-1$ and the sidechain carbon C_i^β for peptide i. The optimal empirical value of $\varphi = -63^\circ$ is very close to the maximum of the latter distance as indicated in Figure 1.4. This can be interpreted by saying that the optimal angles maximize the opportunity for backbone hydration, since the hydrophobic C_i^β has been moved as far as possible from the hydrogen bond constituent O_{i-1}.

Figure 1.4 also plots the distance $O_{i-1} \cdots C_i$ between the backbone oxygen O_{i-1} for the $i-1$ peptide and the backbone carbon C_i for peptide i. This is a distance considered related to a potential steric clash in the classic view but it appears to play no significant role [30].

The role of the distance $O_{i-1} \cdots C_i^\beta$ was considered in [30] but only in terms of potential steric clashes. Although it can be argued that the empirically optimal value of $\varphi = -63^\circ$ does minimize the potential clash between these two atoms, the distance is much larger (> 4Å) than needed just to avoid a clash. By contrast, this greater distance is useful to allow the approach of a water molecule.

The optimal hydrogen bond distance depends to a great extent on quantum effects related to nuclear repulsion and other effects. Thus it could well be that hydrogen bond distances less than 1.9Å are not energetically favorable, at least for the relevant $\widehat{NHO}$ angles. Determination of such relationships could be done by quantum computations in the absence of water to see to what extent hydration plays a role in determining the observed (φ, ψ) angles.

1.2.2 Solvation signature

This work suggests that solvation has a significant effect on protein structure. Such a proposal can be investigated in many ways. For example, protein structures could be examined to see if nearby water molecules cause a shift in observed (φ, ψ) angles. This has been studied in detail in Chapter 3. One might ask whether "buried" helices retain the solvation

signature. But even membrane proteins appear to retain attached water molecules, so it is not so clear what "buried" means in terms of isolation from water. From a dynamic point of view, intercolation by water may occur on a very fast time scale compared to the time scale for helix restructuring. In [49] the structure of polyalanine in water was studied using density functional theory calculations. We examined (φ, ψ) angles and other geometrical parameters of the resulting helices in that study and compared to our results in Chapter 3.

The further presentation of material will go as follows: Chapter 2 takes a more detailed look at hydrogen bond geometry in α-helices and contains the results we obtained via mathematical optimization. In Chapter 3 we apply data mining techniques to connect the previous results to the concept of solvation signature in protein helices. At the end of Chapter 3 we apply this concept to some current research in protein secondary structure relevant to amyloid diseases. We conclude by summarizing main results in Chapter 4.

CHAPTER 2

HYDROGEN BOND GEOMETRY IN α-HELICES

2.1 Helical structure

The polypeptide chain has the flexibility to assume many local configurations, as explained in Figure 2.1. The peptide bond typically does not allow rotation, but two other backbone bonds, N–C_α and C_α–C, do rotate freely. Angles φ and ψ measure rotation along bonds N–C_α and C_α–C respectively, and are called *dihedral* or *torsion angles*. The ability of dihedral angles to take a large variety of values allows the backbone to form many different configurations. Each configuration can be fully characterized by a sequence of values (φ_i, ψ_i) for each peptide unit.

Only a small fraction of the local configurations of the polypeptide chain are feasible due to many physiochemical constraints. The local configurations of a polypeptide chain are stabilized by the formation of hydrogen bonds $N - H \cdots O = C$ between backbone groups CO and NH (see Figure 1.2). The most stable configurations are the ones that can be characterized by a regular hydrogen-bonding pattern. Such configurations are helical, meaning the same values (φ,ψ) are repeated from unit to unit. On one hand, we can fully specify a helical configuration by a single pair of values (φ, ψ). On the other hand, a secondary structure can be specified by hydrogen bonding pattern, allowing different helical configurations to fall within the same type. In particular, we focus our attention on two types of secondary structures: α-helix and 3_{10}-helix characterized by regular hydrogen bonding pattern of type $i \to i + 4$ and $i \to i + 3$, respectively. Various studies show that dihedral angle pairs $(-63°, -43°)$ and $(-60°, -25°)$, respectively, result in the most common helices within each group [31].

It is known that a hydrogen bond has specific geometrical requirements. When considering regular hydrogen bonds $N - H \cdots O = C$ formed by protein backbone in a protein helix, we can express its parameters as functions of (φ, ψ). So, it is natural to study the relation-

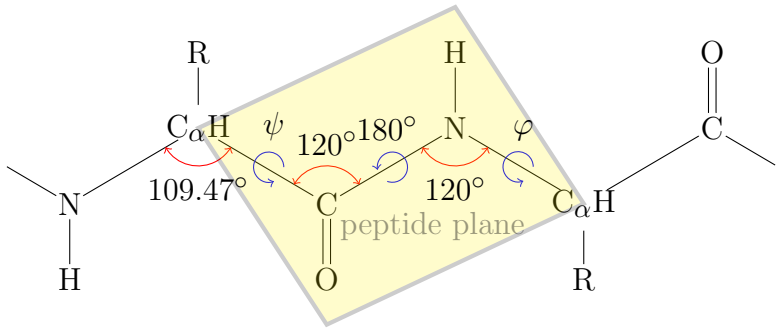


Figure 2.1: *Ideal geometry* of the peptide plane according to Pauling [56]. Angles are in degrees. In the simplified case, the lengths CN, CC_α, NC_α are set to the same parameter L. The physically correct parameters are $CN = 1.33, CC_\alpha = 1.52, NC_\alpha = 1.45, CO = 1.23, NH = 1$ (all units in Ångstroms).

ship between preferred hydrogen bond geometry and stability of protein helical structures. In particular, it is interesting to find out whether two major hydrogen bond requirements: linearity and length constraints serve as a major influence in formation of the most common α- and 3_{10}-helices. We show that these criteria are not sufficient. Moreover, we show that another hydrogen bond parameter plays an important role in the helical formation.

Helical parameters have been studied earlier, using different approaches in [46, 64, 57, 25]. The hydrogen bond geometry has been studied using statistical analysis of small organic molecule data in [59, 55] and globular protein data in [8]. Only one study applied mathematical optimization to linearity property of hydrogen bonds in α-helices [25].

In Section 2.2 we consider all possible helical configurations of backbone atoms in a chain of amino acids. We assume ideal polypeptide chain parameters with values for the covalent bond lengths and angles identified by Pauling in [56] (see Figure 2.1), and allow angles φ and ψ to vary. Any helix is uniquely determined by rotation angle θ, radius r, and rise d. We derive formulas for these parameters as functions of φ and ψ, thus showing that any repeated pair of angles φ and ψ defines a helical structure of a backbone. In Section 2.3 we give an overview of feasible helical configurations. In Section 2.4 we consider groups of atoms $C_i = O_i$ of the i-th amino acid and $N_j - H_j$ of j-th amino acid, where $j - i = 3, 4$ for a 3_{10}

and α-helix respectively. First, we find (φ, ψ) such that the angle

$$N_j\text{–}\widehat{\mathrm{H}_j \cdots O_i} = 180^\circ.$$

This property is called *linearity* of a hydrogen bond. We then minimize the distance $|O_i \cdots H_j|$ as a function of (φ, ψ). We then show that canonical peaks in the (φ, ψ) angle distributions at $(-63^\circ, -43^\circ)$ for α-helices and at $(-60^\circ, -25^\circ)$ in 3_{10}-helices do not result from optimization of hydrogen bonding with respect to any linear combination of these two criteria. We perform further analysis and suggest additional criteria that may play a role in the determination of the empirically optimal angles $(-63^\circ, -43^\circ)$ and $(-60^\circ, -25^\circ)$.

The computations in this chapter were performed by using MAPLE$^{\mathrm{TM}}$[1].

2.2 Protein backbone local regular structures

In this section we study a mathematical model of a protein helix. First, we give a historical overview of the topic including motivation and major results. Second, we describe a mathematical model of a protein helix. We then compute explicit formulas for helical parameters: rotation angle, radius and rise.

2.2.1 History of the topic

In 1930 Meyer and Mark [51] were able to connect chemical universality and rich diversity of protein properties to their molecular structure. Later in the 1930s Astbury, while studying synthetic polypeptides, showed that for the majority of cases polypeptide chains were twined or folded in a certain way [4]. The next major step was made by Huggins when he formulated quantitative geometrical criteria for the theory of polypeptide chain folding. He was the first to consider a hydrogen bond as a major packing force and proposed a helical structure as the best structure to satisfy geometrical criteria for the optimal folding. According to Huggins [32]

"the most stable arrangement for an assemblage of molecules is one in which the component atoms and groups are packed together so that

(a) the distances between neighbors are close to the equilibrium distance,

(b) each atom or group has as many close neighbors as possible, and

(c) there are no large unoccupied regions."

In the same paper he proposed various hypothetical structures that satisfy these requirements and tested them against known X-ray data. Some of those structures had a regular coiling pattern. Following the idea of Astbury [4], he wrote [32]

"a polypeptide chain can be coiled, consistent with the following assumptions:

(1) bond distances and angles are the expected ones;

(2) atoms not directly bonded together are not too close together;

(3) like atoms (or groups) are surrounded equivalently;

(4) adjacent turns are connected by NHO hydrogen bridges."

One of the proposed theoretical structures in that paper [32] was in fact the 3_{10}-helix. Starting with Huggins' research, hydrogen bonds and polypeptide helical structures became very popular research topics. Ten years later, in 1953, the structure of α-helix was proposed by Pauling, Corey and Branson [56, 26].

General interest in helical configurations of chain molecules continued to increase in the 1950s. Some attempts at mathematical treatment of special cases of polypeptide chains were made as early as 1953 [46]. The first rigorous mathematical analysis with torsion angles as parameters was published in 1954 by Mizushima and Simanouchi [64]. They applied an idea of Eyring [27].

In [64] Mizushima and Simanouchi first considered the simplified case, when only one kind of atom constitutes the helix. In this setting bond lengths L, bond angles ξ and internal

rotation angles ζ are the same from unit to unit. They obtained the following formulas for helical rotation angle θ, rise d, and radius r:

$$\cos\theta = \frac{1}{2}(-\cos\zeta + \cos\xi - \cos\xi\cos\zeta - 1) \tag{2.1}$$

$$d^2 = L^2(1-\cos\xi)(1-\cos\zeta)/(3+\cos\zeta-\cos\xi+\cos\zeta\cos\xi) \tag{2.2}$$

$$r^2 = 2L^2(1+\cos\zeta)/(3+\cos\zeta-\cos\xi+\cos\zeta\cos\xi)^2. \tag{2.3}$$

Then the researchers considered a more general case of a stable configuration of a polypeptide chain, where there are three kinds of internal rotation axes -CO-NH-, -CHR-CO-, and -NH-CHR-, with internal rotation angles $\zeta_1, \zeta_2, \zeta_3$, bond distances L_1, L_2, L_3 and bond angles ξ_1, ξ_2, ξ_3 respectively. Again applying the method of Eyring [27] they define

$$A = A_3A_1A_2 \tag{2.4}$$

and

$$B = A_3A_1B_2 + A_3B_1 + B_3, \tag{2.5}$$

where

$$A_i = \begin{pmatrix} -\cos\xi_i & -\sin\xi_i & 0 \\ \sin\xi_i\cos\zeta_j & -\cos\xi_i\cos\zeta_i & -\sin\zeta_i \\ \sin\xi_i\sin\zeta_i & -\cos\xi_i\sin\zeta_i & \cos\zeta_i \end{pmatrix}, \tag{2.6}$$

and

$$B_i = \begin{pmatrix} L_i \\ 0 \\ 0 \end{pmatrix}, \tag{2.7}$$

for $i = 1, 2, 3$. They show how θ, d, and r can be computed in terms of the entries of matrices A and B:

$$\cos\theta = (a_{11} + a_{22} + a_{33} - 1)/2, \tag{2.8}$$

$$d^2 = \frac{[b_1(a_{13}+a_{31}) + b_2(a_{23}+a_{32}) + b_3(a_{33}-a_{11}-a_{22}+1)]^2}{(3-a_{11}-a_{22}-a_{33})(a_{33}-a_{11}-a_{22}+1)}, \tag{2.9}$$

$$r^2 = (b_1^2 + b_2^2 + b_3^2 - d^2)/(3-a_{11}-a_{22}-a_{33}). \tag{2.10}$$

Later in the reference written by Dickerson and Geis [24] published in 1969, we find the following formulas:

$$\cos\left(\frac{\theta}{2}\right) = 0.817\sin\left(\frac{\varphi+\psi}{2}\right) + 0.045\sin\left(\frac{\varphi-\psi}{2}\right) \tag{2.11}$$

$$d\sin\left(\frac{\theta}{2}\right) = -2.967\cos\left(\frac{\varphi+\psi}{2}\right) - 0.664\cos\left(\frac{\varphi-\psi}{2}\right). \tag{2.12}$$

In 1999 Quine [57] computed helical parameters using quaternions. The formula used in [57] for θ is identical to (2.11). The rise is expressed as a scalar product of normalized axis vector $\bar{a}$ and vector $\bar{b}$, which is a virtual bond vector from C_i^α to C_{i+1}^α :

$$d = \bar{b}\cdot\bar{a}. \tag{2.13}$$

The helical radius was computed in vector form:

$$\bar{r} = \frac{1}{2}\cot\frac{\theta}{2}\,\bar{a}\times\bar{b} + \bar{b} - (\bar{b}\cdot\bar{a})\bar{a}. \tag{2.14}$$

We will use a method introduced by Dix in 2002 [25] and compute θ, $\bar{a}$, d, and r symbolically as functions of torsion angles φ, ψ, and a bond length L.

2.2.2 Computations

We assign to each atom a its position $\mathcal{R}_a$ in $\mathbb{R}^3$ space. Any three atoms (a_1, a_2, a_3) with corresponding positions $(\mathcal{R}_{a_1}, \mathcal{R}_{a_2}, \mathcal{R}_{a_3})$ determine a Cartesian coordinate system $E_{(a_1,a_2,a_3)}(\mathcal{R}) =$

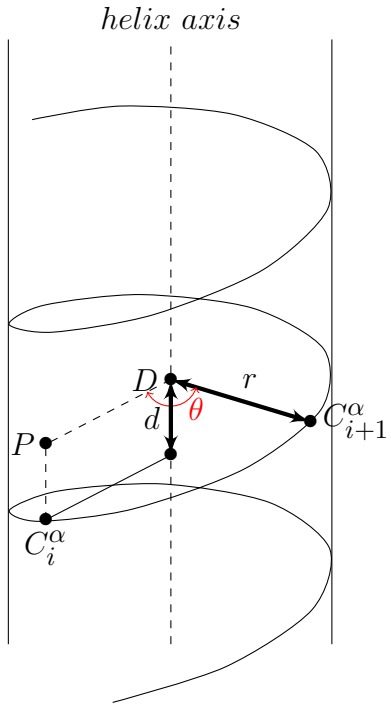


Figure 2.2: Helical parameters: rotation angle θ, radius r, and rise d are shown. Point P is a projection of C^{α}_{i} onto the plane containing C^{α}_{i+1} and orthogonal to the rotation axis.

(e_0, e_1, e_2, e_3) in the following way (as introduced by Dix in [25]):

$$e_0 = \mathcal{R}_{a_1}, \tag{2.15}$$

$$e_1 = \frac{\mathcal{R}_{a_2} - \mathcal{R}_{a_1}}{||\mathcal{R}_{a_2} - \mathcal{R}_{a_1}||}, \tag{2.16}$$

$$e_2 = \frac{(1 - e_1 e_1^T)(\mathcal{R}_{a_3} - \mathcal{R}_{a_1})}{||(1 - e_1 e_1^T)(\mathcal{R}_{a_3} - \mathcal{R}_{a_1})||}, \tag{2.17}$$

$$e_3 = e_1 \times e_2. \tag{2.18}$$

Let a 4×4 matrix $\mathcal{A}$ denote a coordinate transformation matrix from the system $E_{(a_1,a_2,a_3)}$ to $E_{(a'_1,a'_2,a'_3)}$, i.e.

$$E_{(a'_1,a'_2,a'_3)} = E_{(a_1,a_2,a_3)}\mathcal{A}. \tag{2.19}$$

It is of the form

$$\mathcal{A} = \begin{pmatrix} 1 & \bar{0} \\ \bar{b} & A \end{pmatrix}, \tag{2.20}$$

where A is a 3×3 rotation matrix, $\bar{0} = (0, 0, 0)$, and $\bar{b}$ denotes the vector $\overline{a_1 a_1'}$ in $\mathbb{R}^3$.

As explained in [25], any coordinate transformation matrix can be expressed as a product of coordinate transformation matrices of 3 types. Namely,

$$T_1(L) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ L & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{2.21}$$

$$T_2(\xi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\xi & \sin\xi & 0 \\ 0 & \sin\xi & -\cos\xi & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{2.22}$$

and

$$T_3(\zeta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\zeta & -\sin\zeta \\ 0 & 0 & \sin\zeta & \cos\zeta \end{pmatrix}. \tag{2.23}$$

These three matrices represent the following basic types of coordinate system transformations:

$$E_{(a_2,a_1,a_3)} = E_{(a_1,a_2,a_3)} T_1(L), \tag{2.24}$$

where $L = L_{a_1,a_2} > 0$ is a length of the $a_1 \to a_2$ bond,

$$E_{(a_1,a_3,a_2)} = E_{(a_1,a_2,a_3)} T_2(\xi), \tag{2.25}$$

where $\xi = \xi_{a_1,a_2,a_3} > 0$ is the angle $\xi = \widehat{a_1 a_2 a_3}$, and

$$E_{(a_1,a_2,a')} = E_{(a_1,a_2,a)} T_3(\pm\zeta), \tag{2.26}$$

where $\zeta = \zeta_{a_1,a_2}$ is the dihedral connecting angle $[a\, a_1\, a_2\, a']$. We limit the subscripts for ζ to the two intervening points since (1) this defines the line around which the angle rotates and (2) the predecessor to a_1 and the successor to a_2 are implicitly given by the protein backbone sequence.

We want to compute a coordinate transformation matrix $\mathcal{A}$ such that

$$E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})} = E_{(C^\alpha_i,C'_i,N_{i+1})} \mathcal{A}, \tag{2.27}$$

or we can also write

$$E_{(C^\alpha_i,C'_i,N_{i+1})} \xrightarrow{\mathcal{A}} E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})}. \tag{2.28}$$

This transformation is a result of a chain of nine transformations consequently applied one after another in the following way:

$$\begin{aligned}
&E_{(C^\alpha_i,C'_i,N_{i+1})} \xrightarrow{\mathcal{A}} E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})} = E_{(C^\alpha_i,C'_i,N_{i+1})} \xrightarrow{T_1(L_{C^\alpha_i,C'_i})} \\
&E_{(C'_i,C^\alpha_i,N_{i+1})} \xrightarrow{T_2(\xi_{C'_i,C^\alpha_i,N_{i+1}})} E_{(C'_i,N_{i+1},C^\alpha_i)} \xrightarrow{T_3(\pm\zeta_{C'_i,N_{i+1}})} \\
&E_{(C'_i,N_{i+1},C^\alpha_{i+1})} \xrightarrow{T_1(L_{C'_i,N_{i+1}})} E_{(N_{i+1},C'_i,C^\alpha_{i+1})} \xrightarrow{T_2(\xi_{N_{i+1},C'_i,C^\alpha_{i+1}})} \\
&E_{(N_{i+1},C^\alpha_{i+1},C'_i)} \xrightarrow{T_3(\pm\zeta_{N_{i+1},C^\alpha_{i+1}})} E_{(N_{i+1},C^\alpha_{i+1},C'_{i+1})} \xrightarrow{T_1(L_{N_{i+1},C^\alpha_{i+1}})} \\
&E_{(C^\alpha_{i+1},N_{i+1},C'_{i+1})} \xrightarrow{T_2(\xi_{C^\alpha_{i+1},N_{i+1},C'_{i+1}})} E_{(C^\alpha_{i+1},C'_{i+1},N_{i+1})} \\
&\xrightarrow{T_3(\pm\zeta_{C^\alpha_{i+1},C'_{i+1}})} E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})}.
\end{aligned} \tag{2.29}$$

We can write $\mathcal{A}$ as a product of nine matrices:

$$\begin{aligned}\mathcal{A} =& T_1(L_{C_i^\alpha,C_i'})T_2(\xi_{C_i',C_i^\alpha,N_{i+1}})T_3(\pm\zeta_{C_i',N_{i+1}})T_1(L_{C_i',N_{i+1}})T_2(\xi_{N_{i+1},C_i',C_{i+1}^\alpha})\\ & T_3(\pm\zeta_{N_{i+1},C_{i+1}^\alpha})T_1(L_{N_{i+1},C_{i+1}^\alpha})T_2(\xi_{C_{i+1}^\alpha,N_{i+1},C_{i+1}'})T_3(\pm\zeta_{C_{i+1}^\alpha,C_{i+1}'}).\end{aligned} \tag{2.30}$$

Let us assume ideal parameters of the peptide chain besides the torsion angles which we denote by φ and ψ:

$$L_{C_i^\alpha,C_i'} = L_{C_i',N_{i+1}} = L_{N_{i+1},C_{i+1}^\alpha} = L,$$

$$\xi_{C_i',C_i^\alpha,N_{i+1}} = \xi_{N_{i+1},C_i',C_{i+1}^\alpha} = 120^\circ, \quad \xi_{C_{i+1}^\alpha,N_{i+1},C_{i+1}'} = 109.47^\circ,$$

$$\zeta_{C_i',N_{i+1}} = 180^\circ, \quad \zeta_{N_{i+1},C_{i+1}^\alpha} = \varphi, \quad \zeta_{C_{i+1}^\alpha,C_{i+1}'} = \psi$$

(see Figure 2.1). Thus,

$$\mathcal{A} = T_1(L)T_2(120^\circ)T_3(180^\circ)T_1(L)T_2(120^\circ)T_3(\pm\varphi)T_1(L)T_2(109.47^\circ)T_3(\pm\psi) \tag{2.31}$$

We performed all symbolic computations in MAPLE 16 [1] and obtained the following results:

$$\mathcal{A} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \begin{pmatrix} \frac{5L}{2} \\ \frac{\sqrt{3}L}{2} \\ 0 \end{pmatrix} & A \end{pmatrix} \tag{2.32}$$

where the rotation sub-matrix of $\mathcal{A}$, which we denote by A, is of the form

$$A = \begin{pmatrix} \frac{1}{3} & -\frac{2\sqrt{2}}{3}\cos(\psi) & \frac{2\sqrt{2}}{3}\sin(\psi) \\ -\frac{2\sqrt{2}}{3}\cos(\varphi) & -\frac{1}{3}\cos(\varphi)\cos(\psi) + \sin(\varphi)\sin(\psi) & \frac{1}{3}\cos(\varphi)\sin(\psi) + \sin(\varphi)\cos(\psi) \\ -\frac{2\sqrt{(2)}}{3}\sin(\varphi) & -\frac{1}{3}\sin(\varphi)\cos(\psi) - \cos(\varphi)\sin(\psi) & \frac{1}{3}\sin(\varphi)\sin(\psi) - \cos(\varphi)\cos(\psi) \end{pmatrix}. \tag{2.33}$$

To see that A is indeed a rotation matrix, we check using Maple that

$$\det(A) = \cos^2(\varphi)\cos^2(\psi) + \sin^2(\varphi)\sin^2(\psi) + \cos^2(\varphi)\sin^2(\psi) + \sin^2(\varphi)\cos^2(\psi) = 1,$$

and

$$A^T A = I.$$

We first compute the trace of A

$$\operatorname{tr}(A) = \frac{1}{3} - \frac{4}{3}\cos(\varphi)\cos(\psi) + \frac{4}{3}\sin(\varphi)\sin(\psi). \tag{2.34}$$

The eigenvalues of A are

$$\lambda_1 = 1, \lambda_{2,3} = \cos(\theta) \pm i\sin(\theta), \tag{2.35}$$

where θ is the rotation angle of A. It follows from (2.35) that

$$\operatorname{tr}(A) = 1 + 2\cos(\theta). \tag{2.36}$$

Then (2.34) and (2.36) give

$$\begin{aligned} \cos(\theta) &= -\frac{1}{3} - \frac{2}{3}\cos(\varphi)\cos(\psi) + \frac{2}{3}\sin(\varphi)\sin(\psi) = \\ &= \frac{-1 - 2\cos(\varphi+\psi)}{3} = \frac{1 - 4\cos^2(\frac{\varphi+\psi}{2})}{3}. \end{aligned} \tag{2.37}$$

The axis of rotation is the eigenvector of A corresponding to the eigenvalue 1. The unnormalized axis vector $\bar{u}$ can be found by the following formula:

$$\begin{aligned} \bar{u} &= (A_{32} - A_{23}, A_{13} - A_{31}, A_{21} - A_{12})^T \\ &= \big(-\frac{4}{3}\sin\big(\varphi - \psi\big), \frac{4\sqrt{2}}{3}\sin\big(\frac{\varphi+\psi}{2}\big)\cos\big(\frac{\varphi+\psi}{2}\big), \frac{4\sqrt{2}}{3}\sin\big(\frac{\varphi+\psi}{2}\big)\sin\big(\frac{\varphi-\psi}{2}\big)\big)^T. \end{aligned} \tag{2.38}$$

Then the normalized axis vector is

$$\bar{a} = \frac{\bar{u}}{2\sin(\theta)}. \tag{2.39}$$

The distance d traveled from C_i^α to C_{i+1}^α parallel to the axis of rotation (also known as a *rise* of a helix) can be found by the formula

$$\begin{aligned}
d &= |\bar{b}\cdot\bar{a}| \\
&= \left|\frac{\left(\frac{5}{2}L, \frac{\sqrt{3}L}{2}, 0\right)\cdot\left(-\frac{4}{3}\sin(\varphi-\psi), \frac{4\sqrt{2}}{3}\sin\left(\frac{\varphi+\psi}{2}\right)\cos\left(\frac{\varphi+\psi}{2}\right), \frac{4\sqrt{2}}{3}\sin\left(\frac{\varphi+\psi}{2}\right)\sin\left(\frac{\varphi-\psi}{2}\right)\right)^T}{2\sin(\theta)}\right| \\
&= \left|\frac{-2L\sin\left(\frac{\varphi+\psi}{2}\right)\left(10\cos\left(\frac{\varphi+\psi}{2}\right)-\sqrt{6}\cos\left(\frac{\varphi-\psi}{2}\right)\right)}{6\sqrt{1-\left(\frac{1-4\cos^2\left(\frac{\varphi+\psi}{2}\right)}{3}\right)^2}}\right| \\
&= \left|\frac{-L\sin\left(\frac{\varphi+\psi}{2}\right)\left(10\cos\left(\frac{\varphi+\psi}{2}\right)-\sqrt{6}\cos\left(\frac{\varphi-\psi}{2}\right)\right)}{2\sqrt{2-2\cos^4\left(\frac{\varphi+\psi}{2}\right)+2\cos^2\left(\frac{\varphi+\psi}{2}\right)}}\right|.
\end{aligned} \tag{2.40}$$

Let point P be a projection of C_i^α onto the plane containing C_{i+1}^α and orthogonal to the rotation axis. Let point D lie on the rotation axis such that $\overline{C_{i+1}^\alpha D}$ is orthogonal to the axis. To find the radius r we consider triangles $\triangle C_i^\alpha C_{i+1}^\alpha P$ and $\triangle C_{i+1}^\alpha PD$ (see Figure 2.2).

Clearly,

$$\frac{|C_{i+1}^\alpha P|^2}{\sin^2(\theta)} = \frac{\bar{b}\cdot\bar{b}-d^2}{\sin^2(\theta)} = \frac{r^2}{(\sin(\frac{180^\circ-\theta}{2}))^2}. \tag{2.41}$$

Thus we have

$$\begin{aligned}
r^2 &= \frac{(\bar{b}\cdot\bar{b}-d^2)\cos^2(\frac{\theta}{2})}{(1-\cos^2(\theta))} = \frac{(\bar{b}\cdot\bar{b}-d^2)(\frac{\cos(\theta)+1}{2})}{(1-\cos(\theta))(1+\cos(\theta))} \\
&= \frac{(\bar{b}\cdot\bar{b}-d^2)}{2(1-\cos(\theta))} = \frac{(\bar{b}\cdot\bar{b}-d^2)}{2(1-\frac{\mathrm{tr}(A)-1}{2})} = \frac{(\bar{b}\cdot\bar{b}-d^2)}{3-\mathrm{tr}(A)} \\
&= \frac{7L^2 - \frac{L^2\sin^2\left(\frac{\varphi+\psi}{2}\right)\left(10\cos\left(\frac{\varphi+\psi}{2}\right)-\sqrt{6}\cos\left(\frac{\varphi-\psi}{2}\right)\right)^2}{8-8\cos^4\left(\frac{\varphi+\psi}{2}\right)+8\cos^2\left(\frac{\varphi+\psi}{2}\right)}}{\frac{8}{3}+\frac{4}{3}\cos(\varphi+\psi)}.
\end{aligned} \tag{2.42}$$

2.2.3 Helical conclusions

In this section we studied a mathematical model of a protein helix. First, we gave a historical overview of mathematical studies of a protein helix. We described a mathematical model of an idealized protein helix, and computed explicit formulas for helical parameters: rotation angle (2.37), rotation axis (2.38), normalized rotation axis (2.39), rise (2.40), and radius (2.42). Formula (2.37) was previously published in [24]. To the best of our knowledge formulas (2.38), (2.40), and (2.42) are not found elsewhere in the literature.

From now on, we will drop the assumption of a common bond length, but we will also not try to give full analytical expressions.

2.3 Feasible protein backbone local regular structures - Allowed (φ, ψ) orientations

It is important to note that not all possible (φ, ψ) orientations are allowed, because of the close contacts between the atoms of the adjacent residues (including side-chain atoms that we consider later in our model). In this section we give an overview of research on feasible helical configurations. We start with a motivation for this direction of research. We then survey some results helpful in our future work.

2.3.1 Motivation

In the 1950's modeling protein structures from X-ray and other data was very active. The structure of collagen was a popular challenge. In 1953 while visiting India, professor J. D. Bernal introduced G. N. Ramachandran to this problem. Ramachandran, being in India, worked on this problem independently while two other groups of British scientists (Rich and Crick in Cambridge and Randall, Cowan, and North in King's College London) worked on it as well. Ramachandran proposed a two-bonded (two hydrogen bonds per 3 residue repeat) triple helix, and his structure was criticized by British scientists (Rich and

Crick) "on the basis of steric hindrance." This criticism inspired a survey of the available crystal structures showing that "short" interatomic distances were allowed in some structures [58]. This information together with already available bond lengths and angles discovered by Pauling in turn inspired the use of distance data in a mathematical study of all possible (or feasible) conformations of a polypeptide chain. In particular the discovery of "shorter" interatomic distances in available data inspired the study of both normal van der Waals radii and shorter ones resulting in "normally allowed" and "outer limit" regions, respectively. Ramachandran's student, Sasisekharan, was the first to study this topic which he published in 1962 [63]. Further developments were done with his graduate adviser G. N. Ramachandran in 1965 [60].

A particularly good rendering of the allowed regions of the Sasisekharan-Ramachandran plot is given in [48, Figure 10, page 4627]. However, for us the main conclusion is that the location of the preferred (φ, ψ) angles does *not* lie on the boundary of the normally allowed region. Thus the empirically observed (φ, ψ) values are not determined by steric hindrance. They could be perturbed substantially and still lie within the normally allowed region.

2.3.2 Allowed (φ, ψ) regions

Sasisekharan treated atoms as simple impenetrable spheres and found three major *normally allowed* (φ, ψ) regions for standard radii of peptide atoms and *outer limit* regions for the smallest radii that can be still considered plausible. The φ, ψ plot with allowed regions marked received the name of *Ramachandran plot*. Here we are mainly concerned with the helical area that we broadly outline in Figure 2.3. We keep the allowed region boundaries in all plots. The outer limit region is important because the 3_{10}-helices belong to it and not to the normally allowed region. It is worth noting that the map was derived when no single protein structure had been solved. It has been tested on known secondary structures such as α-helix and β-sheet and on small peptides (see [58, Figure 6, page 54]).

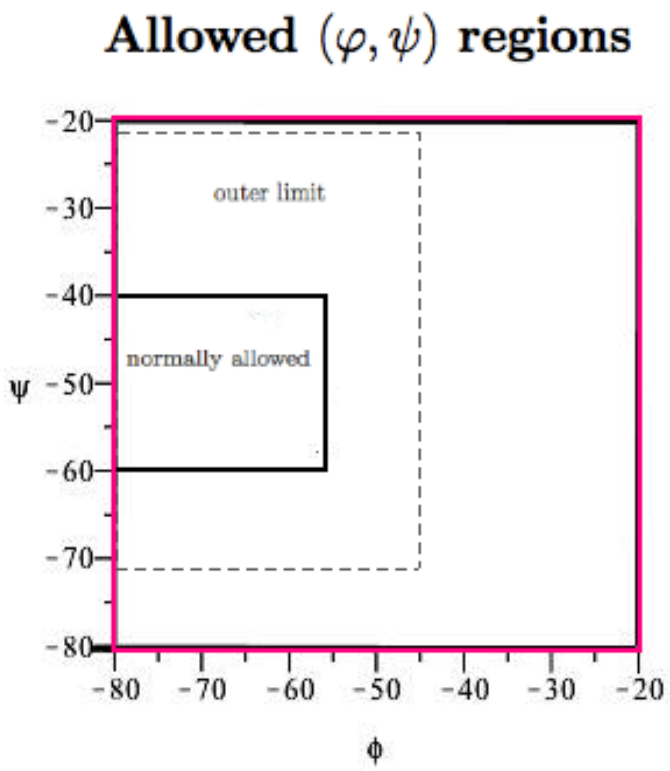


Figure 2.3: (φ, ψ) allowed regions (see highlighted values in Table 2.2).

Contact	**Normally allowed, Å**	**Outer limit, Å**
$C_\alpha \cdots C_\alpha$	3.20	3.00
$C \cdots C$	2.95	2.90
$C \cdots O$	2.80	2.70
$C \cdots N$	2.90	2.80
$C \cdots H$	2.40	2.20
$O \cdots O$	2.70	2.60
$O \cdots N$	2.70	2.60
$O \cdots H$	2.40	2.20
$N \cdots N$	2.70	2.60
$N \cdots H$	2.40	2.20
$H \cdots H$	2.00	1.90

Table 2.1: Minimum contact distances between atoms assumed in [60, Table 2, page 912]

2.3.3 Feasible conclusions

Here we gave an overview of results from [63] and [60] obtained by Sasisekharan and Ramachandran for feasible helical configurations. In particular, in the Table 2.2 we highlight normally allowed and outer limit ranges for (φ, ψ). We use these results in Section 2.4 where we perform the optimization in the domain outlined in Figure 2.3.

2.4 Stable protein backbone structures and hydrogen bond geometry

Recall that all backbone local regular structures (i.e., helices) can be defined by a pair (φ, ψ) for $-180^\circ \leq \varphi, \psi \leq 180^\circ$, but not all such structures are feasible (see Table 2.2). Only some feasible structures are stable, namely the ones that allow for regular hydrogen bonding pattern. In this section we consider common protein backbone structures and study the geometry of main chain – main chain hydrogen bonds formed within these structures. The relationship between hydrogen bond geometrical requirements and stability of various protein secondary structures has been extensively studied by applying statistical analysis to protein experimental data [8, 55]. We study this relationship via mathematical optimization and use the method introduced in 2002 by Dix in [25].

We start by giving a precise definition of a hydrogen bond together with its major geometrical parameters. The strength of a hydrogen bond depends on certain geometrical constraints with respect to linearity, length and various associated angles. Stronger hydrogen bonds in turn should result in the formation of more stable helices. We focus our attention on two types of protein secondary structure: α-helix and 3_{10}-helix characterized by regular hydrogen bonding pattern of type $i \to i+4$ and $i \to i+3$, respectively. Even though both α- and 3_{10}-helices each represent a group of helices, various studies show that dihedral angle pairs $(-63^\circ, -43^\circ)$ and $(-60^\circ, -25^\circ)$ characterize the most common helices within each group respectively [31]. Using mathematical models of protein α- and 3_{10}-helices, in-

corporating empirical angles and bond lengths, we compute various geometrical parameters of a main chain – main chain hydrogen bond in each structure as functions of (φ, ψ) and study their influence on helical stability. We optimize with respect to hydrogen bond linearity and length in α- and 3_{10}-helix and show that canonical (φ, ψ) angles are not optimal with respect to any linear combination of these two criteria. We then study some other hydrogen bond parameters and show that they play an important role in the determination of optimal (φ, ψ) angles.

Previously, mathematical optimization was only applied to the α-helix, and only the linearity of the hydrogen bond was considered [25]. We successfully reproduce that result and use this as a verification that our computational model is correct. Our original contribution to the topic starts in Sections 2.4.3 and 2.4.4 where we compute a similar linearity result for the 3_{10}-helix. It then continues in Sections 2.4.5, 2.4.6, and 2.4.7 where we minimize the length of a hydrogen bond in both α- and 3_{10}-helices, compute other relevant parameters, and give numerical results respectively. In particular, our results in Section 2.4.7 show that canonical peaks at $(-63°, -43°)$ in α-helix and at $(-60°, -25°)$ in 3_{10}-helix do not result from optimization of hydrogen bonding with respect to any linear combination of two major hydrogen bond requirements: linearity and length constraints. We conclude by giving further analysis in Section 2.4.9. First, we notice a clustering of the in-plane component $\gamma_{C_i\widehat{-O_i\ldots}H_j}$ around $12-13°$ for canonical peak values in both α- and 3_{10}-helices (see Figure 1.2(b) for the definition of the γ angle), and then show how the relationship between angles $C_i - \widehat{O_i \ldots} H_j$, $N_j - \widehat{H_j \ldots} O_i$, and the distance $|H_j \ldots O_i|$ play a very important role in formation of most common α- and 3_{10}-helices.

2.4.1 A hydrogen bond and its geometrical parameters

An attractive interaction between two electronegative atoms X and Y through an intervening hydrogen atom H is called a *hydrogen bond* and denoted by X–H$\cdots$Y–Z, where dots denote the bond. Atom X is called a *donor* and Y is called an *acceptor*. The hydrogen bond is

N–$\widehat{C_\alpha \cdots C}$ (°)	**Normally allowed range** φ(°)	ψ(°)	**Outer limit range** φ(°)	ψ(°)
(a) non-glycyl residues				
105	[-158,-65]	[-59, -51]∪[95,141]	[-180,-60]∪[177, 180]	[-71,-38]∪[38,163]
110	**[-158, -53]**	**[-60,-40]** ∪[92,179]	**[-180,-45]** ∪[45,61]	**[-71,-22]** ∪[20,180]
115	[-153,-41]	[-59,-30]∪[98,177]	[-176,-33]∪[33,57]	[-70,188]
(b) glycyl residues				
110	[-180, -53]∪[53,180]	[-180,-40]∪[40,180]	[-180,-45]∪[45,180]	[-180,-22]∪[22,180]
115	[-180,-41]∪[41,180]	[-180,-30]∪[30,180]	[-180,-33]∪[33,180]	no restriction

Table 2.2: Ranges of φ and ψ allowed by the contact distances between the atoms N_1, C_2, and C_β and the remaining backbone atoms in the two residues for non-glycyl residues and between the atoms N_1, C_α, and H_2 and the remaining backbone atoms for glycyl residues (see [60, Table 3, page 916]) for details. Regions relevant to our study are highlighted.

strong enough to hold molecules XH and YZ together at normal temperatures and directional enough so that this association is orientationally specific [23]. Here we consider hydrogen bonds formed within a chain of amino acids. Namely, we study backbone-backbone hydrogen bonds N_j–$H_j \cdots O_i = C_i$, formed between groups $C_i = O_i$ of the i-th amino acid and N_j–H_j of the j-th amino acid, for $j - i = 3, 4$, with N_j as a donor and C_i as an acceptor. The following parameters characterize geometry of these hydrogen bonds (see Figure 1.2):

1) distance $|O_i \cdots H_j|$;

2) angle $\nu = N_j - \widehat{H_j \cdots O_i}$;

3) angle $\mu = C_i - \widehat{O_i \cdots H_j}$, along with its two components:

4) in-plane angle $\gamma_{C_i - \widehat{O_i \cdots H_j}}$ and

5) out-of-plane angle $\beta_{C_i - \widehat{O_i \cdots H_j}}$.

Chain	ID	AA	$\varphi(^\circ)$	$\psi(^\circ)$
A	111	Gln	-64.15	-43.26
	112	Lys	-61.02	-44.53
	113	Val	-60.18	-42.60
	114	Ser	-63.65	-43.56
	115	Glu	-64.30	-40.59
	116	Ala	-64.76	-41.70
	117	Cys	-63.19	-41.83

Table 2.3: TRIOSEPHOSPHATE ISOMERASE 2VXN [2]. An example of a globular protein α-helix with φ and ψ values tightly clustered around canonical peak $(-63^\circ,-43^\circ)$. Structure was solved in 2010 via X-Ray chrystallography with resolution 0.82Å. Data is taken from Protein Geometry Database [13].

2.4.2 Most common α- and 3_{10}-helices (significance of canonical peaks)

Dihedral angle pairs $(-63^\circ, -43^\circ)$ and $(-60^\circ, -25^\circ)$ are known as *canonical peaks* ((φ, ψ) values that occur most commonly) for α- and 3_{10}-helix respectively. The significance of these peaks can be clearly seen in [31, Figure 2, page 15].

Protein helices with (φ, ψ) values near the canonical peaks can be found in globular proteins in recently solved high resolution structures. See for example the 0.82Å resolution structure of triosephosphate isomerase PDB file 2VXN. The (φ, ψ) values for the α-helix fragment 2VXN.A.111-117 are shown in Table 2.3.

Also, in hydrophobic environments α-helices are the most stable with (φ, ψ) values clustering around the canonical peak $(-63^\circ, -43^\circ)$ [37]. An example of such helices in membrane proteins is found in the fragment of a transmembrane protein integrin depicted in PDB file 2K9J. The (φ, ψ) values for an α-helix fragment 2K9J.A.12-30 are shown in Table 2.4.

2.4.3 Computational models and transformation matrices for α- and 3_{10}-helix

Here we are interested in two coordinate transformations:

$$E_{(H_j,N_j,C'_{j-1})} \xrightarrow{\mathcal{A}_{j-i}} E_{(O_i,C'_i,C^\alpha_i)} \tag{2.43}$$

for $j-i=3,4$. These transformations connect atoms H_j and O_i corresponding to a backbone-backbone hydrogen bond N_j–$H_j \cdots O_i$ formed in 3_{10}-helices for $j-i=3$ and in α-helices for $j-i=4$.

The transformation matrix $\mathcal{A}_4$ was computed by Dix in [25] via the IMIMOL computer program. IMIMOL provides a graphical two-dimensional interface that facilitates representation of a three-dimensional molecular structure as a collection of labeled graphs. Such representations make it possible to automate some geometry calculations. In particular, if two atoms are selected in the given structure, then using IMIMOL one can automatically create a Maple procedure to compute symbolically a transformation matrix between two Cartesian coordinate systems originated in the selected atoms respectively. In [25] a proper graph representation of three-dimensional molecular structure of an α-helix was created in IMIMOL and a Maple procedure to compute $\mathcal{A}_4$ symbolically was exported. Then optimization computation was carried out in Maple using the IMIMOL-exported procedure.

Our original plan was to reproduce the computation for $\mathcal{A}_4$ exactly as it was done in Section 5.4 [25]. However, when we used the procedure exported by IMIMOL for computing $\mathcal{A}_4$ to compute the optimal (φ,ψ) values for the linearity of a hydrogen bond in α-helix exactly the way described by Dix we have received values different from his result. Unable to immediately find the mistake that led us to the incorrect answer we decided to compute $\mathcal{A}_4$ symbolically "from scratch" without using IMIMOL. We then did the same optimization as before and this time our answer was correct. By comparing our result for $\mathcal{A}_4$ with the procedure exported via IMIMOL we identified that we set up the initial model for α-helix in IMIMOL incorrectly and those mistakes carried into the exported procedure. We realized that the interface provided in IMIMOL allows for incorrect labeling and object selection. In order to avoid another possibility for mistakes we decided to do all computations without using IMIMOL.

We will explain our computation of $\mathcal{A}_4$ and compute $\mathcal{A}_3$ in a similar manner. Recall, that the ideal parameters of the peptide chain are $\xi_{C_i',C_i^\alpha,N_{i+1}}=120^\circ$, and $\xi_{N_{i+1},C_i',C_{i+1}^\alpha}=120^\circ$,

Chain	**ID**	**AA**	$\varphi(^\circ)$	$\psi(^\circ)$
A	12	Trp	-61.1	-38.0
	13	Val	-62.4	-44.3
	14	Leu	-61.9	-43.9
	15	Val	-64.2	-44.9
	16	Gly	-64.0	-38.4
	17	Val	-65.1	-44.3
	18	Leu	-62.3	-41.8
	19	Gly	-61.6	-34.9
	20	Gly	-64.7	-38.2
	21	Leu	-62.2	-41.4
	22	Leu	-62.5	-41.1
	23	Leu	-61.1	-44.8
	24	Leu	-65.7	-42.4
	25	Thr	-63.2	-45.7
	26	Ile	-63.4	-44.6
	27	Leu	-63.7	-43.2
	28	Val	-64.2	-44.6
	29	Leu	-62.0	-42.7
	30	Ala	-64.3	-45.4

Table 2.4: INTEGRIN ALPHAIIb-BETA3 TRANSMEMBRANE COMPLEX 2K9J [43]. An example of a membrane protein α-helix with φ and ψ values tightly clustered around canonical peak $(-63^\circ,-43^\circ)$. Structure was solved in 2009 via solution NMR. Data is taken from Protein Data Bank [13].

$\xi_{C^\alpha_{i+1},N_{i+1},C'_{i+1}} = 109.47^\circ$, $\zeta_{C'_i,N_{i+1}} = 180^\circ$. As before, we denote the torsion angles by φ and ψ: $\zeta_{N_{i+1},C^\alpha_{i+1}} = \varphi$, $\zeta_{C^\alpha_{i+1},C'_{i+1}} = \psi$. To compute $\mathcal{A}_3$ and $\mathcal{A}_4$ as functions of (φ,ψ) we use the same peptide chain model as we used for computing helical parameters with one exception. Instead of using parameter L as the length of all backbone bonds we use the actual bond distances [56] (Å)

$$L_{C^\alpha_i,C'_i} = 1.52,\ L_{C'_i,N_{i+1}} = 1.33,\ L_{N_{i+1},C^\alpha_{i+1}} = 1.45$$

(see Figure 2.1), corresponding to the ideal α-helical geometry. We will also use two additional bond lengths $L_{O_i,C'_i} = 1.23, L_{H_j,N_J} = 1$ also described in [56]. To simplify notation we can set $i = 1$ without any loss of generality. Then from (2.43) for an α-helix

$$\begin{aligned}
&E_{(H_5,N_5,C'_4)} \xrightarrow{\mathcal{A}_4(\varphi,\psi)} E_{(O_1,C'_1,C^\alpha_1)} = E_{(H_5,N_5,C'_4)} \xrightarrow{T_1(1)} E_{(N_5,H_5,C'_4)}\\
&\xrightarrow{T_2(120)} E_{(N_5,C'_4,H_5)} \xrightarrow{T_3(0)} E_{(N_5,C'_4,C^\alpha_4)} \xrightarrow{T_1(1.33)} E_{(C'_4,N_5,C^\alpha_4)}\\
&\xrightarrow{T_2(120)} E_{(C'_4,C^\alpha_4,N_5)} \xrightarrow{T_3(\psi)} E_{(C'_4,C^\alpha_4,N_4)} \xrightarrow{T_1(1.52)} E_{(C^\alpha_4,C'_4,N_4)}\\
&\xrightarrow{T_2(109.27)} E_{(C^\alpha_4,N_4,C'_4)} \xrightarrow{T_3(\varphi)} E_{(C^\alpha_4,N_4,C'_3)} \xrightarrow{T_1(1.45)} E_{(N_4,C^\alpha_4,C'_3)}\\
&\xrightarrow{T_2(120)} E_{(N_4,C'_3,C^\alpha_4)} \xrightarrow{T_3(180)} E_{(N_4,C'_3,C^\alpha_3)} \xrightarrow{T_1(1.33)} E_{(C'_3,N_4,C^\alpha_3)}\\
&\xrightarrow{T_2(120)} E_{(C'_3,C^\alpha_3,N_4)} \xrightarrow{T_3(\psi)} E_{(C'_3,C^\alpha_3,N_3)} \xrightarrow{T_1(1.52)} E_{(C^\alpha_3,C'_3,N_3)}\\
&\xrightarrow{T_2(109.27)} E_{(C^\alpha_3,N_3,C'_3)} \xrightarrow{T_3(\varphi)} E_{(C^\alpha_3,N_3,C'_2)} \xrightarrow{T_1(1.45)} E_{(N_3,C^\alpha_3,C'_2)}\\
&\xrightarrow{T_2(120)} E_{(N_3,C'_2,C^\alpha_3)} \xrightarrow{T_3(180)} E_{(N_3,C'_2,C^\alpha_2)} \xrightarrow{T_1(1.33)} E_{(C'_2,N_3,C^\alpha_2)}\\
&\xrightarrow{T_2(120)} E_{(C'_2,C^\alpha_2,N_3)} \xrightarrow{T_3(\psi)} E_{(C'_2,C^\alpha_2,N_2)} \xrightarrow{T_1(1.52)} E_{(C^\alpha_2,C'_2,N_2)}\\
&\xrightarrow{T_2(109.27)} E_{(C^\alpha_2,N_2,C'_2)} \xrightarrow{T_3(\varphi)} E_{(C^\alpha_2,N_2,C'_1)} \xrightarrow{T_1(1.45)} E_{(N_2,C^\alpha_2,C'_1)}\\
&\xrightarrow{T_2(120)} E_{(N_2,C'_1,C^\alpha_1)} \xrightarrow{T_3(180)} E_{(N_2,C'_1,C^\alpha_1)} \xrightarrow{T_1(1.33)} E_{(C'_1,N_2,C^\alpha_1)}\\
&\xrightarrow{T_2(120)} E_{(C'_1,C^\alpha_1,N_2)} \xrightarrow{T_3(180)} E_{(C'_1,C^\alpha_1,O_1)} \xrightarrow{T_2(120)} E_{(C'_1,O_1,C^\alpha_1)}\\
&\xrightarrow{T_1(1.23)} E_{(O_1,C'_1,C^\alpha_1)},
\end{aligned} \tag{2.44}$$

and for 3_{10} helix

$$
\begin{aligned}
&E_{(H_4,N_4,C'_3)} \xrightarrow{\mathcal{A}_3(\varphi,\psi)} E_{(O_1,C'_1,C^\alpha_1)} = E_{(H_4,N_4,C'_3)} \xrightarrow{T_1(1)} E_{(N_4,H_4,C'_3)} \\
&\xrightarrow{T_2(120)} E_{(N_4,C'_3,H_4)} \xrightarrow{T_3(0)} E_{(N_4,C'_3,C^\alpha_3)} \xrightarrow{T_1(1.33)} E_{(C'_3,N_4,C^\alpha_3)} \\
&\xrightarrow{T_2(120)} E_{(C'_3,C^\alpha_3,N_4)} \xrightarrow{T_3(\psi)} E_{(C'_3,C^\alpha_3,N_3)} \xrightarrow{T_1(1.52)} E_{(C^\alpha_3,C'_3,N_3)} \\
&\xrightarrow{T_2(109.27)} E_{(C^\alpha_3,N_3,C'_3)} \xrightarrow{T_3(\varphi)} E_{(C^\alpha_3,N_3,C'_2)} \xrightarrow{T_1(1.45)} E_{(N_3,C^\alpha_3,C'_2)} \\
&\xrightarrow{T_2(120)} E_{(N_3,C'_2,C^\alpha_3)} \xrightarrow{T_3(180)} E_{(N_3,C'_2,C^\alpha_2)} \xrightarrow{T_1(1.33)} E_{(C'_2,N_3,C^\alpha_2)} \\
&\xrightarrow{T_2(120)} E_{(C'_2,C^\alpha_2,N_3)} \xrightarrow{T_3(\psi)} E_{(C'_2,C^\alpha_2,N_2)} \xrightarrow{T_1(1.52)} E_{(C^\alpha_2,C'_2,N_2)} \\
&\xrightarrow{T_2(109.27)} E_{(C^\alpha_2,N_2,C'_2)} \xrightarrow{T_3(\varphi)} E_{(C^\alpha_2,N_2,C'_1)} \xrightarrow{T_1(1.45)} E_{(N_2,C^\alpha_2,C'_1)} \\
&\xrightarrow{T_2(120)} E_{(N_2,C'_1,C^\alpha_1)} \xrightarrow{T_3(180)} E_{(N_2,C'_1,C^\alpha_1)} \xrightarrow{T_1(1.33)} E_{(C'_1,N_2,C^\alpha_1)} \\
&\xrightarrow{T_2(120)} E_{(C'_1,C^\alpha_1,N_2)} \xrightarrow{T_3(180)} E_{(C'_1,C^\alpha_1,O_1)} \xrightarrow{T_2(120)} E_{(C'_1,O_1,C^\alpha_1)} \\
&\xrightarrow{T_1(1.23)} E_{(O_1,C'_1,C^\alpha_1)}.
\end{aligned}
\tag{2.45}
$$

We can then express $\mathcal{A}_4$ and $\mathcal{A}_3$ as products of T_1, T_2, T_3 matrices:

$$
\begin{aligned}
&\mathcal{A}_4(\varphi,\psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27) \\
&T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27) \\
&T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27) \\
&T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(180)T_2(120)T_1(1.23),
\end{aligned}
\tag{2.46}
$$

and

$$
\begin{aligned}
&\mathcal{A}_3(\varphi,\psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27) \\
&T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27) \\
&T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(180)T_2(120)T_1(1.23).
\end{aligned}
\tag{2.47}
$$

2.4.4 Optimizing with respect to linearity of a hydrogen bond

Optimization with respect to linearity of a hydrogen bond in an α-helix was done by Dix in [25]. We will explain his result and compute a similar result for the case of 3_{10}-helix.

For every embedding $\mathbf{R}$ of our model of α-helix into $\mathbb{R}^3$ we have

$$E_{(O_1,C_1',C_1^\alpha)}(\mathbf{R}) = E_{(H_5,N_5,C_4')}(\mathbf{R})\mathcal{A}_4(\varphi,\psi). \tag{2.48}$$

We want to find optimal φ, ψ that will result in angle N_5–$\widehat{\mathrm{H}}_1\cdots O_1$ being 180°. To satisfy this we need atoms N_5, H_5, and O_1 to be collinear. This means that $\mathbf{R}_{O_1}$ should lie on the X-axis of $E_{(H_5,N_5,C_4')}$. Since atom O_1 is located at the origin with respect to $E_{(O_1,C_1',C_1^\alpha)}(\mathbf{R})$ we have

$$\mathbf{R}_{O_1} = E_{(O_1,C_1',C_1^\alpha)}(\mathbf{R})(1,0,0,0)^T = E_{(H_5,N_5,C_4')}(\mathbf{R})\mathcal{A}_4(\varphi,\psi)(1,0,0,0)^T. \tag{2.49}$$

On the other hand

$$\mathbf{R}_{O_1} = E_{(H_5,N_5,C_4')}(\mathbf{R})(1,x_{O_1},y_{O_1},z_{O_1})^T, \tag{2.50}$$

where $(x_{O_1},y_{O_1},z_{O_1})^T$ are the coordinates of O_1 with respect to $E_{(H_5,N_5,C_4')}(\mathbf{R})$. This gives us the following equation

$$(1,x_{O_1},y_{O_1},z_{O_1})^T = \mathcal{A}_4(\varphi,\psi)(1,0,0,0)^T \tag{2.51}$$

from which we can deduce $y_{O_1} = \mathcal{A}_4(\varphi,\psi)_{3,1}$ and $z_{O_1} = \mathcal{A}_4(\varphi,\psi)_{4,1}$. Collinearity of atoms O_1, N_1 and H_1 implies $y_{O_1} = 0$ and $z_{O_1} = 0$. Thus we obtain two equations in two unknowns φ and ψ

$$\mathcal{A}_4(\varphi,\psi)_{3,1} = 0, \text{and } \mathcal{A}_4(\varphi,\psi)_{4,1} = 0. \tag{2.52}$$

Solving (2.52) simultaneously will give us the optimal (φ,ψ) values in α-helix favoring lin-

earity of the hydrogen bond N_5–$H_5 \cdots O_1$. By the exact same argument solving

$$\mathcal{A}_3(\varphi, \psi)_{3,1} = 0, \text{and } \mathcal{A}_3(\varphi, \psi)_{4,1} = 0 \tag{2.53}$$

will give us optimal (φ, ψ) values for a hydrogen bond N_4–$H_4 \cdots O_1$ in a 3_{10}-helix with angle $\widehat{\mathrm{N_4H_1O_1}} = 180°$. Details regarding both solutions are summarized in Table 2.5.

2.4.5 Minimizing the length of a hydrogen bond

Let $\ell_{j-i}(\varphi, \psi)$ denote the length of a hydrogen bond N_j–$H_i \cdots O_i$. It can be computed as follows

$$\ell_{j-i}(\varphi, \psi) = |H_j \cdots O_i| = \left\|\overline{H_j O_i}\right\| = \left\|(\mathcal{A}_{j-i}(\varphi, \psi)_{2,1}, \mathcal{A}_{j-i}(\varphi, \psi)_{3,1}, \mathcal{A}_{j-i}(\varphi, \psi)_{4,1})\right\|. \tag{2.54}$$

Then solving

$$\frac{\partial \ell_{j-i}(\varphi, \psi)}{\partial \varphi} = 0, \text{and } \frac{\partial \ell_{j-i}(\varphi, \psi)}{\partial \psi} = 0 \tag{2.55}$$

simultaneously will give us a pair (φ, ψ) that minimizes the length of a hydrogen bond N_j–$H_i \cdots O_i$.

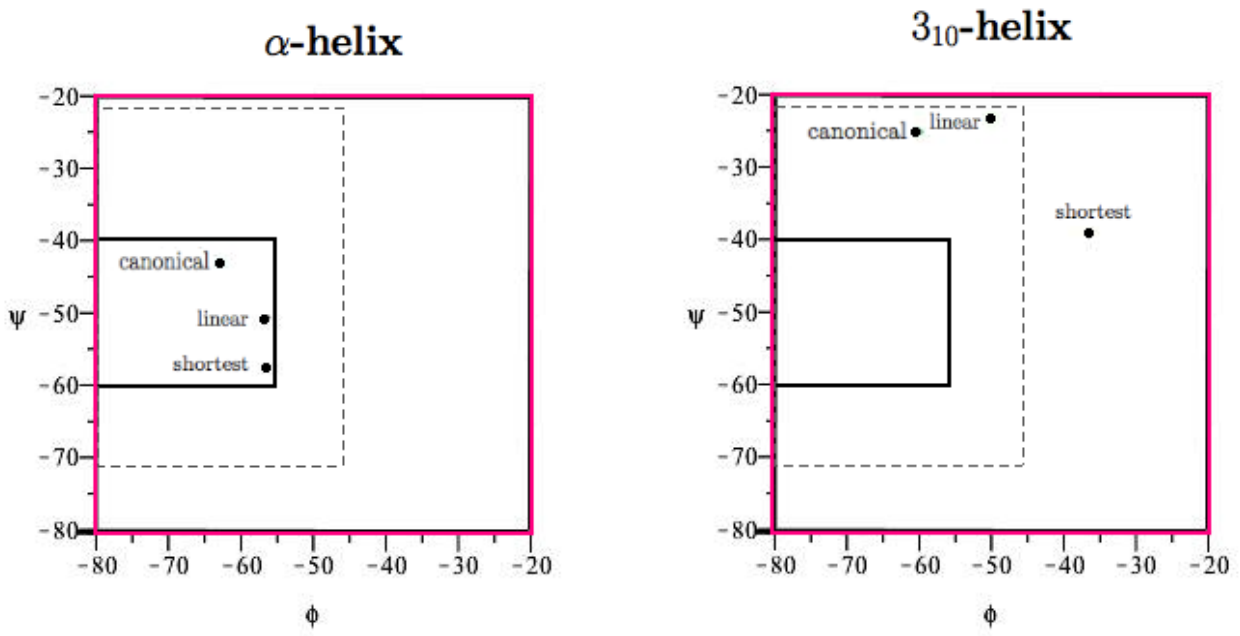
It follows from (2.55), that solving

$$\frac{\partial \ell_4(\varphi, \psi)}{\partial \varphi} = 0, \text{and } \frac{\partial \ell_4(\varphi, \psi)}{\partial \psi} = 0 \tag{2.56}$$

simultaneously will give us a pair (φ, ψ) that minimizes the length of a hydrogen bond N_5–$H_1 \cdots O_1$ in α-helix. And solving the system of equations

$$\frac{\partial \ell_3(\varphi, \psi)}{\partial \varphi} = 0, \text{and } \frac{\partial \ell_3(\varphi, \psi)}{\partial \psi} = 0 \tag{2.57}$$

will give us a pair (φ, ψ) that minimizes the length of a hydrogen bond N_4–$H_1 \cdots O_1$ in a 3_{10}-helix. Details regarding both solutions are summarized in Table 2.5.



(a) (φ, ψ) pairs corresponding to α-helix. **(b)** (φ, ψ) pairs for 3_{10}-helix.

Figure 2.4: (φ, ψ) pairs corresponding to the **shortest** hydrogen bond, **linear** hydrogen bond and **canonical** peak.

2.4.6 Computing other geometrical parameters of a hydrogen bond

Clearly, for an α-helix

$$\begin{aligned}\cos(N_5\text{–}\widehat{H_5\cdots O_1}) &= \frac{\overline{H_5N_5}\cdot\overline{H_5O_1}}{\left\|\overline{H_5N_5}\right\|\left\|\overline{H_5O_1}\right\|} = \\ &= \frac{(1,0,0)(\mathcal{A}_4(\varphi,\psi)_{2,1},\mathcal{A}_4(\varphi,\psi)_{3,1},\mathcal{A}_4(\varphi,\psi)_{4,1})^T}{\ell_4(\varphi,\psi)} = \frac{\mathcal{A}_4(\varphi,\psi)_{2,1}}{\ell_4(\varphi,\psi)}\end{aligned} \tag{2.58}$$

and for a 3_{10}-helix

$$\begin{aligned}\cos(N_4\text{–}\widehat{H_4\cdots O_1}) &= \frac{\overline{H_4N_4}\cdot\overline{H_4O_1}}{\left\|\overline{H_4N_4}\right\|\left\|\overline{H_4O_1}\right\|} = \\ &= \frac{(1,0,0)(\mathcal{A}_3(\varphi,\psi)_{2,1},\mathcal{A}_3(\varphi,\psi)_{3,1},\mathcal{A}_3(\varphi,\psi)_{4,1})^T}{\ell_3(\varphi,\psi)} = \frac{\mathcal{A}_3(\varphi,\psi)_{2,1}}{\ell_3(\varphi,\psi)}.\end{aligned} \tag{2.59}$$

We will be also interested in computing angle $C_1 - \widehat{O_1\cdots H_5}$ along with its two com-

ponents: in-plane $\gamma_{C_1-\widehat{O_1\cdots}H_5}$, and out-of-plane $\beta_{C_1-\widehat{O_1\cdots}H_5}$, for an α-helix. Similarly, for a 3_{10}-helix we will compute $C_1-\widehat{O_1\cdots}H_4$ together with in-plane $\gamma_{C_1-\widehat{O_1\cdots}H_4}$, and out-of-plane $\beta_{C_1-\widehat{O_1\cdots}H_4}$ components, respectively. Let

$$\begin{aligned}&\mathcal{C}_4(\varphi,\psi)=T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)\\&T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)\\&T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)T_2(120)\\&T_3(180)T_1(1.33).\end{aligned}\tag{2.60}$$

Clearly

$$\overline{H_5C_1}=(\mathcal{C}_4(\varphi,\psi)_{2,1},\mathcal{C}_4(\varphi,\psi)_{3,1},\mathcal{C}_4(\varphi,\psi)_{4,1}).\tag{2.61}$$

Then

$$\begin{aligned}&\overline{C_1O_1}=\overline{H_5O_1}-\overline{H_1C_1}=\\&=(\mathcal{A}_4(\varphi,\psi)_{2,1}-\mathcal{C}_4(\varphi,\psi)_{2,1},\mathcal{A}_4(\varphi,\psi)_{3,1}-\mathcal{C}_4(\varphi,\psi)_{3,1},\mathcal{A}_4(\varphi,\psi)_{4,1}-\mathcal{C}_4(\varphi,\psi)_{4,1}),\end{aligned}\tag{2.62}$$

and

$$\cos(C_1-\widehat{O_1\cdots}H_5)=\frac{\overline{O_1H_5}\cdot\overline{O_1C_1}}{\left\|\overline{O_1H_5}\right\|\left\|\overline{O_1C_1}\right\|},\tag{2.63}$$

where $\left\|\overline{O_1C_1}\right\|=1.23$.

In order to compute the in-plane $\gamma_{C_1-\widehat{O_1\cdots}H_5}$, and out-of-plane $\beta_{C_1-\widehat{O_1\cdots}H_5}$ components of $C_1-\widehat{O_1\cdots}H_5$ we need to compute

$$\begin{aligned}&\mathcal{N}_4(\varphi,\psi)=T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)\\&T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)\\&T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45).\end{aligned}\tag{2.64}$$

Then

$$\overline{H_5N_1}=(\mathcal{N}_4(\varphi,\psi)_{2,1},\mathcal{N}_4(\varphi,\psi)_{3,1},\mathcal{N}_4(\varphi,\psi)_{4,1}),\tag{2.65}$$

and

$$\begin{aligned}&\overline{N_1C_1} = \overline{H_5C_1} - \overline{H_5N_1} = \\ &= (\mathcal{C}_4(\varphi,\psi)_{2,1} - \mathcal{N}_4(\varphi,\psi)_{2,1}, \mathcal{C}_4(\varphi,\psi)_{3,1} - \mathcal{N}_4(\varphi,\psi)_{3,1}, \mathcal{C}_4(\varphi,\psi)_{4,1} - \mathcal{N}_4(\varphi,\psi)_{4,1}).\end{aligned} \tag{2.66}$$

Atoms N_1, C_1, and O_1 lie in the same peptide plane. Let $\overline{N}$ be the normal vector of that plane

$$\overline{N} = \frac{\overline{N_1C_1} \times \overline{C_1O_1}}{\left\|\overline{N_1C_1} \times \overline{C_1O_1}\right\|}, \tag{2.67}$$

and let $\overline{O_1H_5}^{Proj}$ be a projection of $\overline{O_1H_5}$ onto that plane

$$\overline{O_1H_5}^{Proj} = \overline{O_1H_5} - (\overline{N} \cdot \overline{O_1H_5})\overline{N}. \tag{2.68}$$

Then

$$\cos(\gamma_{C_1\text{–}O_1\cdots H_5}) = \frac{\overline{C_1O_1} \cdot \overline{O_1H_5}^{Proj}}{1.23 \left\|\overline{O_1H_5}^{Proj}\right\|}, \tag{2.69}$$

and

$$\cos(\beta_{C_1\text{–}O_1\cdots H_5}) = \frac{\overline{O_1H_5} \cdot \overline{O_1H_5}^{Proj}}{\ell_4(\varphi,\psi) \left\|\overline{O_1H_5}^{Proj}\right\|}. \tag{2.70}$$

In the case of a 3_{10}-helix,

$$\begin{aligned}&\mathcal{C}_3(\varphi,\psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi) \\ &T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45) \\ &T_2(120)T_3(180)T_1(1.33),\end{aligned} \tag{2.71}$$

and thus

$$\begin{aligned}&\overline{C_1O_1} = \overline{H_4O_1} - \overline{H_1C_1} = \\ &= (\mathcal{A}_3(\varphi,\psi)_{2,1} - \mathcal{C}_3(\varphi,\psi)_{2,1}, \mathcal{A}_3(\varphi,\psi)_{3,1} - \mathcal{C}_3(\varphi,\psi)_{3,1}, \mathcal{A}_3(\varphi,\psi)_{4,1} - \mathcal{C}_3(\varphi,\psi)_{4,1}),\end{aligned} \tag{2.72}$$

$$\cos(C_1 - \widehat{O_1 \cdots} H_4) = \frac{\overline{O_1H_4} \cdot \overline{O_1C_1}}{\left\|\overline{O_1H_4}\right\| \left\|\overline{O_1C_1}\right\|} = \frac{\overline{O_1H_4} \cdot \overline{O_1C_1}}{1.23 \left\|\overline{O_1H_4}\right\|}, \tag{2.73}$$

and

$$\begin{aligned} &\mathcal{N}_3(\varphi,\psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi) \\ &T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45). \end{aligned} \tag{2.74}$$

Then

$$\overline{H_4N_1} = (\mathcal{N}_3(\varphi,\psi)_{2,1}, \mathcal{N}_3(\varphi,\psi)_{3,1}, \mathcal{N}_3(\varphi,\psi)_{4,1}), \tag{2.75}$$

and

$$\begin{aligned} &\overline{N_1C_1} = \overline{H_4C_1} - \overline{H_4N_1} = \\ &= (\mathcal{C}_3(\varphi,\psi)_{2,1} - \mathcal{N}_3(\varphi,\psi)_{2,1}, \mathcal{C}_3(\varphi,\psi)_{3,1} - \mathcal{N}_3(\varphi,\psi)_{3,1}, \mathcal{C}_3(\varphi,\psi)_{4,1} - \mathcal{N}_3(\varphi,\psi)_{4,1}). \end{aligned} \tag{2.76}$$

Let $\overline{O_1H_4}^{Proj}$ be a projection of $\overline{O_1H_4}$ onto that plane

$$\overline{O_1H_4}^{Proj} = \overline{O_1H_4} - (\overline{N} \cdot \overline{O_1H_4})\overline{N}. \tag{2.77}$$

Then

$$\cos(\gamma_{C_1-O_1\cdots H_4}) = \frac{\overline{C_1O_1} \cdot \overline{O_1H_4}^{Proj}}{1.23 \left\|\overline{O_1H_4}^{Proj}\right\|}, \tag{2.78}$$

and

$$\cos(\beta_{C_1-O_1\cdots H_4}) = \frac{\overline{O_1H_4} \cdot \overline{O_1H_4}^{Proj}}{\ell_3(\varphi,\psi) \left\|\overline{O_1H_4}^{Proj}\right\|}. \tag{2.79}$$

Details regarding all of these parameters for the optimal solutions are summarized in Table 2.5.

2.4.7 Optimization results and balance between two criteria

We considered the model of a protein helix described above with angles φ and ψ as free parameters and optimized hydrogen bonds N_5–$H_5 \cdots O_1$ and N_4–$H_4 \cdots O_1$, first to maximize linearity, and then separately to minimize distances $H_5 \cdots O_1$ and $H_4 \cdots O_1$. The results are summarized in first two columns of Table 2.5 respecively. Note that values (φ, ψ)=$(-56.5°, -50.5°)$ and $|O_1 \ldots H_5| = 1.76$Å are the same as obtained by Dix in [25] which verifies the correctness of our computational model. In the third column we computed hydrogen bond parameters for (φ, ψ) pairs known as *canonical peaks* ((φ, ψ) values that occur most commonly) for α- and 3_{10}-helix respectively. The significance of these peaks can be clearly seen in [31, Figure 2, page 15]. The fourth column contains results taken from statistical analysis of hydrogen bond geometry in secondary structures of globular proteins performed in 1984 by Hubbard and Baker [8] and included in the table for comparison with the results in the third column. The fact that the computed parameters for canonical peaks are very close to the statistical averages validates the correctness of the computation. We used MAPLE 16 for computation mostly because that was the choice of Dix in [25] (see Sections 5.3-5.4 [25] for details).

In Figures 2.4 (a) and (b) we plotted (φ, ψ) pairs corresponding to the shortest hydrogen bond, linear hydrogen bond and canonical peaks for α-helix and for 3_{10}-helix. It is clear that in both cases the canonical peak could not result solely from the balance (a linear combination) of two optimization criteria we considered. This is easily proven for the α-helix by considering the contours shown in Figures 1.1 (a) and (b). Any optimization with a linear combination of the two objective functions would give an objective function defined by a linear combination of these two contours. The two sets of contours are essentially parallel, and thus the resulting objective function would have contours that would be just a translation from one of set of contours to the other. Therefore the minimum would lie on an essentially straight line joining the two minima for the two separate objective functions. The corresponding argument for the 3_{10}-helix is the same but using Figures 2.9 (a) and 2.10

α-helix	N_5–$\widehat{\mathrm{H}}_5\cdots O_1 = 180°$	$\min\lvert O_1\cdots H_5\rvert$	canonical peak	data from [8]
(φ,ψ) (°,°)	$(-56.5, -50.5)$[25]	$(-56.5, -57.5)$	$(-63,\ -43)$ [31]	N/A
$\varphi+\psi$ (°)	107	114	106	N/A
$\lvert O_1\cdots H_5\rvert$ (Å)	1.76 [25]	1.54	1.92	2.05 ± 0.15
N_5–$\widehat{\mathrm{H}}_5\cdots O_1$ (°)	180	146	166	157 ± 9
C_1–$\widehat{\mathrm{O}}_1\cdots H_5$ (°)	168	147	156	147 ± 7
$\gamma_{C_1\text{–}\widehat{\mathrm{O}}_1\cdots H_5}$ (°)	-3.8	-25	-11.8	-18 ± 9
$\beta_{C_1\text{–}\widehat{\mathrm{O}}_1\cdots H_5}$ (°)	10.6	22	20.5	28 ± 8
3_{10}-helix	N_4–$\widehat{\mathrm{H}}_4\cdots O_1 = 180°$	$\min\lvert O_1\cdots H_4\rvert$	canonical peak	data from [8]
(φ,ψ) (°,°)	$(-50.3, -23.3)$	$(-36.6, -39)$	$(-60,\ -25)$ [31]	N/A
$\varphi+\psi$ (°)	73.6	75.6	85	N/A
$\lvert O_1\cdots H_4\rvert$ (Å)	1.76	1.62	1.93	2.17 ± 0.16
N_4–$\widehat{\mathrm{H}}_4\cdots O_1$ (°)	180	158	155	153 ± 10
C_1–$\widehat{\mathrm{O}}_1\cdots H_4$ (°)	144	159	122	114 ± 10
$\gamma_{C_1\text{–}\widehat{\mathrm{O}}_1\cdots H_4}$ (°)	-20.8	-5.9	-13	-30
$\beta_{C_1\text{–}\widehat{\mathrm{O}}_1\cdots H_5}$ (°)	30	19.2	56	60

Table 2.5: With φ and ψ as free parameters hydrogen bonds N_5–$H_5\cdots O_1$ in α-helix and N_4–$H_4\cdots O_1$ in 3_{10}-helix were optimized to maximize linearity (first column), and then to minimize their lengths (second column). In the third column we computed hydrogen bond parameters for (φ,ψ) pairs known as *canonical peaks* ((φ,ψ) values that occur most commonly) for α- and 3_{10}-helix respectively (values were taken from [31]). In the fourth column contains results from statistical analysis of hydrogen bond geometry in secondary structures of globular proteins performed in 1984 by Hubbard and Baker [8]. Values in column three are close to statistical values in column four which is evidence of correctness.

α-helix	N_5–$\widehat{\mathrm{H}}_5\cdots O_1 = 180°$	$\min\lvert O_1\cdots H_5\rvert$	canonical peak
(φ,ψ) (°,°)	$(-56.5, -50.5)$ [25]	$(-56.5, -57.5)$	$(-63,\ -43)$ [31]
θ (°)	97.95	93.56	98.6
d (Å)	1.45	1.31	1.49
r (Å)	2.33	2.46	2.31
3_{10}-helix	N_4–$\widehat{\mathrm{H}}_4\cdots O_1 = 180°$	$\min\lvert O_1\cdots H_4\rvert$	canonical peak
(φ,ψ) (°,°)	$(-50.3, -23.3)$	$(-36.6, -39)$	$(-60,-25)$ [31]
θ (°)	121.44	120	113
d (Å)	1.98	1.94	1.86
r (Å)	1.87	1.90	1.99

Table 2.6: Rotation angle θ, rise d, and radius r computed for the following (φ,ψ)-helices: optimized to maximize linearity of a hydrogen bond (first column), and then to minimize a length of a hydrogen bond (second column). The third column contains (φ,ψ) pairs known as *canonical peaks* ((φ,ψ) pairs that occur most commonly) for α- and 3_{10}-helix respectively (peak values (φ,ψ) values were taken from [31]).

(b) instead.

2.4.8 Hydrogen bond geometry in small molecules

Small molecules offer some indications of "expected" hydrogen bond geometries in the absence of any restrictions due to the protein structure [8]. Ramakrishnan and Prasad in 1971 [59] analyzed various parameters associated with N–H$\cdots$O type of hydrogen bonds using data from reported crystal structures of amino acids and simple peptides. They found that the directions C–O and O$\cdots$N and between C–O and O$\cdots$H, when the location of a hydrogen atom was known, tends to lie between two cones about C–O with semi-vertical angles 40° and 70°. This means that (when the location of a hydrogen atom was known),

$$40^\circ \leq (N - \widehat{H \cdots O}) - (C - \widehat{O \cdots H}) \leq 70^\circ.$$

Also the authors found that distribution of $\beta_{C\text{–}O\cdots H}$ is between 0° and 40°, with 90% having an angle less that 50°.

Olovsson and Jonsson in 1976 [55] studied geometry of bonds N–H$\cdots$O using available X-ray diffraction data of small molecules. They found that hydrogen bonds deviating by 10° to 15° from linearity seem to occur as frequently as more linear bonds. Also, the angle $C - \widehat{O \cdots H}$ is clustered around 125° with some large deviations for longer bonds.

The facts above suggest taking a closer look at the relationship between angles $N_5 - \widehat{H_5 \cdots O_1}$ and $C_1 - \widehat{O_1 \cdots H_5}$, and between $N_4 - \widehat{H_4 \cdots O_1}$ and $C_1 - \widehat{O_1 \cdots H_4}$.

2.4.9 Relationship between geometrical parameters

We created detailed contour plots for geometrical parameters of hydrogen bonds N_5–$H_5 \cdots O_1$ (see Figures 1.1 (a) and (b), and Figures 2.8 (a) and (b)), and N_4–$H_4 \cdots O_1$ (see Figures 2.9 (a) and (b), and Figures 2.10 (a) and (b)) as functions of (φ, ψ). We then combined some contours in one plot for α- and one for 3_{10}-helix, to show graphically a possible role that

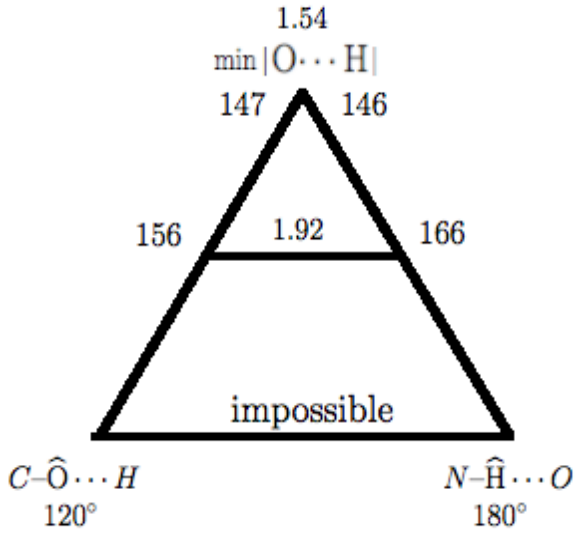


Figure 2.5: Relationship between α-helix hydrogen bond parameters $H_5 \cdots O_1$, $C_1 - \widehat{O_1 \cdots H_5}$ and $C_1 - \widehat{O_1 \cdots H_5}$ is shown in the diagram. It suggests that minimizing the bond distance $H_j \cdots O_i$, maximizing linearity and minimizing the angle $C_i - \widehat{O_i \cdots H_j}$ simultaneously moves us closer to the canonical peaks for α-helix.

the difference $(N_j - \widehat{H_j \cdots O_i}) - (C_i - \widehat{O_i \cdots H_j})$, for $(j-i) = 3, 4$, might play in formation of peaks (see Figures 2.6 and 2.7). In particular, plots suggest that minimizing the bond distance $H_j \cdots O_i$, maximizing linearity and minimizing the angle $C_i - \widehat{O_i \cdots H_j}$ simultaneously moves us closer to the canonical peaks for both α-helix and 3_{10}-helix. We show this relationship graphically in the diagram (see Figure 2.5).

It has been noted in [8] that the hydrogen bond energy is very sensitive to the angle $N\widehat{–H \cdots O}$, but is hardly affected by the angle $C - \widehat{O \cdots H}$. Correlation between linearity and length of a hydrogen bond can be seen in Figures 2.11 (a) and (b), where we plotted contours for $N_j - \widehat{H_j \cdots O_i}$ and $H_j \cdots O_i$ for $j - i = 3, 4$ simultaneously. We also plotted contours for $C_i - \widehat{O_i \cdots H_j}$ and $H_j \cdots O_i$ for $j - i = 3, 4$ (see Figures 2.12 (a) and (b)). It is clear that within allowed helical region a shorter hydrogen bond distance is favored for larger values of $N - \widehat{H \cdots O}$. On the other hand, a much broader range of $C - \widehat{O \cdots H}$ values is allowed when the hydrogen bond distance is restricted to smaller values.

2.4.10 Other geometrical restrictions

Reference [37] defines δ as the "peptide plane tilt angle" with respect to the helical axis, with $\delta = 0$ being the parallel case [37, Figure 1, page 2086]. The preferred range is $4^\circ \leq \delta \leq 12^\circ$. The figure [37, Figure 2, page 2087] presents peptide plane tilt angles with respect to the helical axis (δ) diagramed as a function of φ, ψ torsion angles from uniform helical models leading to the development of this Ramachandran-delta diagram. Superimposed on this diagram are φ, ψ values from 500 high-resolution ($\geq$ 1.8Å resolution) crystal structures. The figure [37, Figure 3, page 2088] shows δ contours φ, ψ values for a membrane protein.

2.4.11 Hydrogen bond conclusions

In this section we reviewed the definition of a hydrogen bond and related geometrical parameters. We explained the significance of (φ, ψ) values known as canonical peaks: $(-63^\circ, -43^\circ)$ in α-helix and $(-60^\circ, -25^\circ)$ in 3_{10}-helix [31]. Following the method of Dix [25] we expressed main chain – main chain hydrogen bond parameters in α- and 3_{10}-helix as functions of dihedral angles (φ, ψ). We successfully reproduced the optimization result from [25], where optimization was done with respect to the linearity of a hydrogen bond in α-helix. We obtained the same values (φ, ψ)=$(-56.5^\circ, -50.5^\circ)$ as in [25]. We used it as a verification that our computational model is correct. We then produced a similar linearity result for 3_{10}-helix and obtained values (φ, ψ)=$(-50.3^\circ, -23.3^\circ)$. We then minimized the length of a hydrogen bond in both α- and 3_{10}-helices obtaining values (φ, ψ)=$(-56.5^\circ, -57.5^\circ)$ and $(-36.3^\circ, -39^\circ)$ respectively. For the four optimal pairs mentioned above and for the canonical peaks $(-63^\circ, -43^\circ)$ in α-helix and $(-60^\circ, -25^\circ)$ in 3_{10}-helix, we computed hydrogen bond geometrical parameters and presented numerical results in Table 2.5, where for a better comparison we also included results from statistical analysis of hydrogen bond geometry in secondary structures of globular proteins performed in 1984 by Hubbard and Baker [8]. In Figures 2.4 (a) and (b), we showed that canonical peaks $(-63^\circ, -43^\circ)$ in α-helix and $(-60^\circ, -25^\circ)$ in 3_{10}-helix do not result from optimization of hydrogen bonding with respect

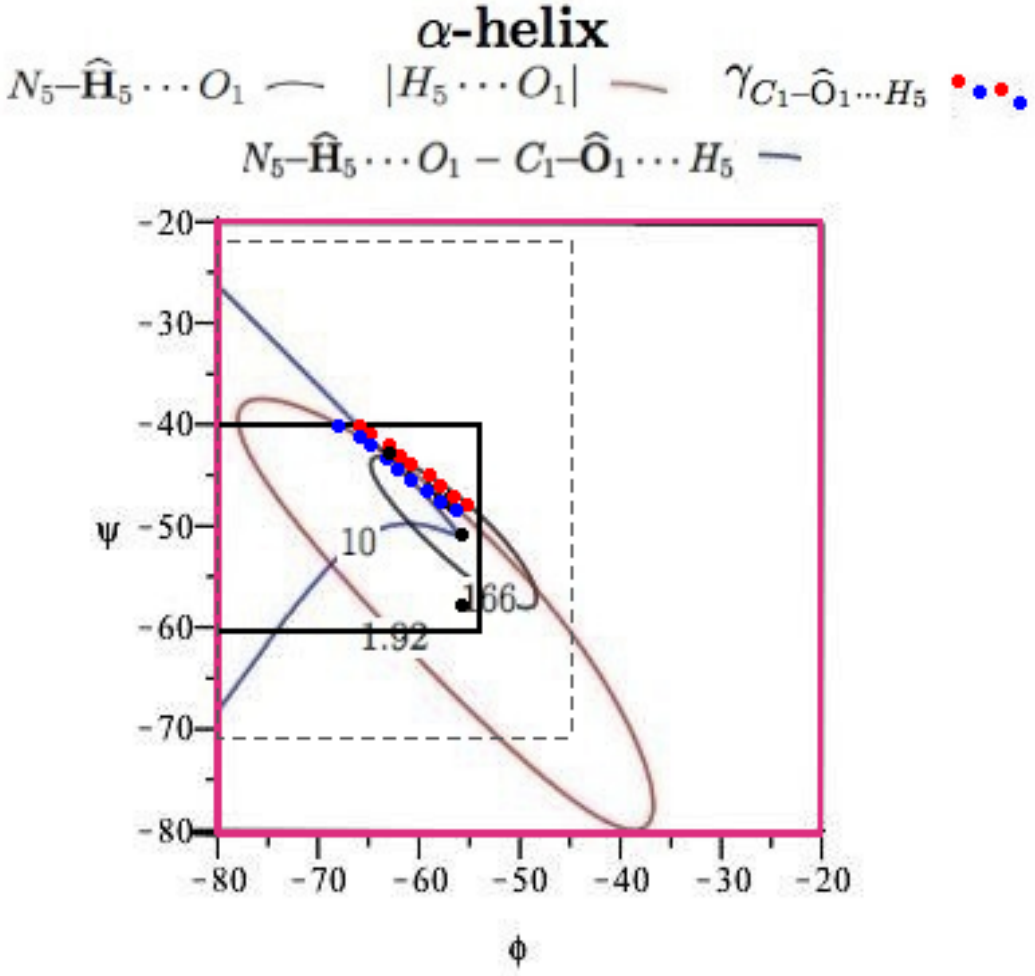


Figure 2.6: This plot suggests that minimizing the bond distance $H_5 \cdots O_1$, maximizing linearity and minimizing the angle C_1–$\widehat{O_1 \cdots H_5}$ simultaneously moves us closer to the canonical peak for α-helix: $\gamma_{C_1\text{–}\widehat{O_1 \cdots H_5}} = -15^\circ$ is shown in red and $= -10^\circ$ in blue dots.

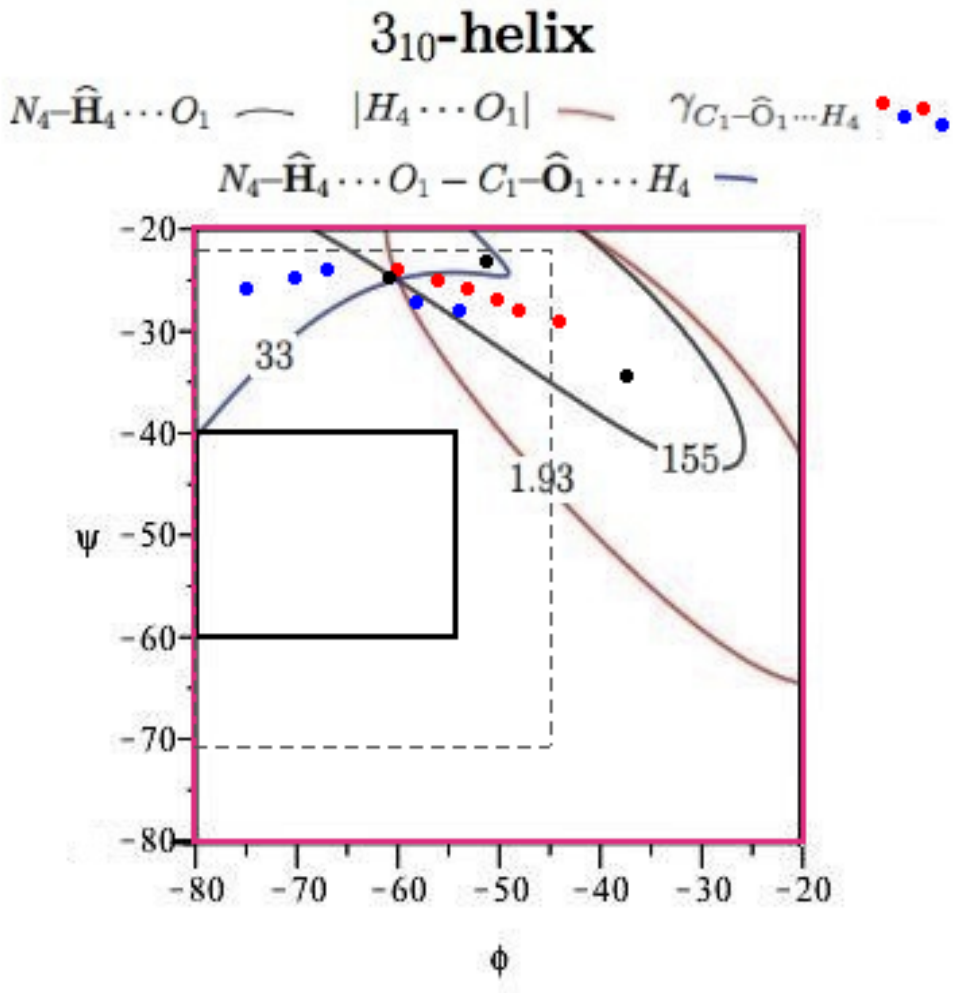
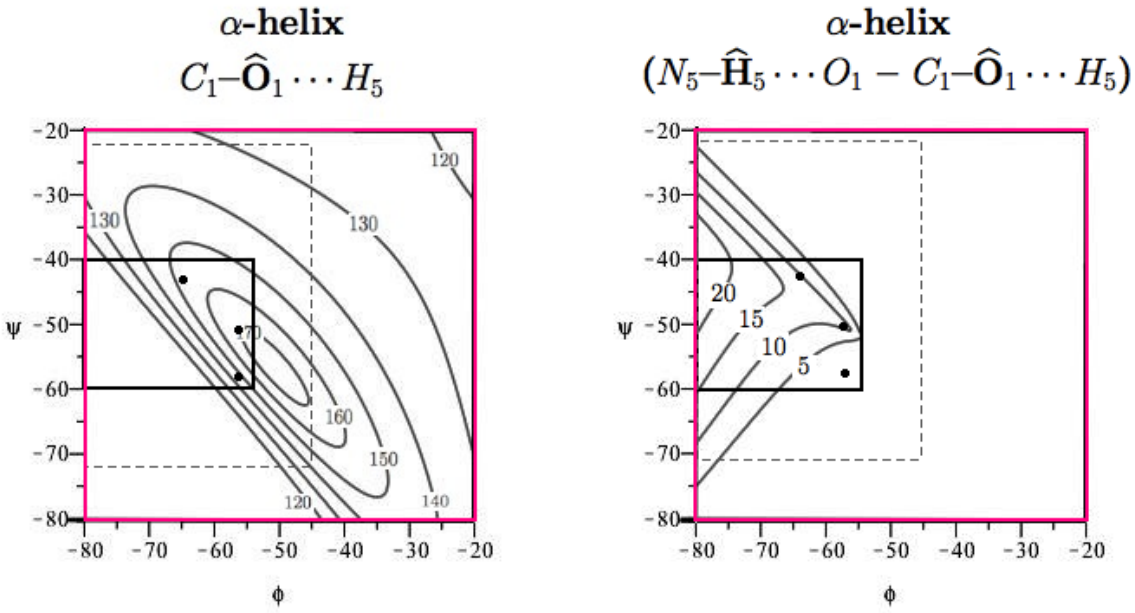


Figure 2.7: This plot suggests that minimizing the bond distance $H_3 \cdots O_1$, maximizing linearity and minimizing the angle C_1–$\widehat{O_1 \cdots H_4}$ simultaneously moves us closer to the canonical peak for 3_{10}-helix: $\gamma_{C_1\text{–}\widehat{O_1 \cdots H_4}} = -15°$ is shown in red and $= -10°$ in blue dots.

to any linear combination of the two major hydrogen bond requirements: linearity and length constraints. We noticed a clustering of the in-plane component $\gamma_{C_i\text{–}\widehat{O_i \ldots H_j}}$ around $12° - 13°$ for canonical peak values in both α- and 3_{10}-helix (see Figures 2.6 and 2.7). And we showed how the relationship between angles $C_i - \widehat{O_i \ldots H_j}$, $N_j - \widehat{H_j \ldots O_i}$, and the distance $|H_j \ldots O_i|$ play a very important role in formation of most common α- and 3_{10}-helices (see the diagram in Figure 2.5, and contour plots in Figures 1.1 – 2.12).

Our results show the clustering of the in-plane component $\gamma_{C_i\text{–}\widehat{O_i \ldots H_j}}$ at $12° - 13°$ for canonical peak values in both α- and 3_{10}-helices which we believe to be significant (see



(a) Contours for different values of C_1–$\widehat{O_1 \cdots H_5}$ as a function of (φ, ψ) are plotted.

(b) Contours for the difference $(N_5$–$\widehat{H_5 \cdots O_1} - C_1$–$\widehat{O_1 \cdots H_5})$ as a function of (φ, ψ) are plotted.

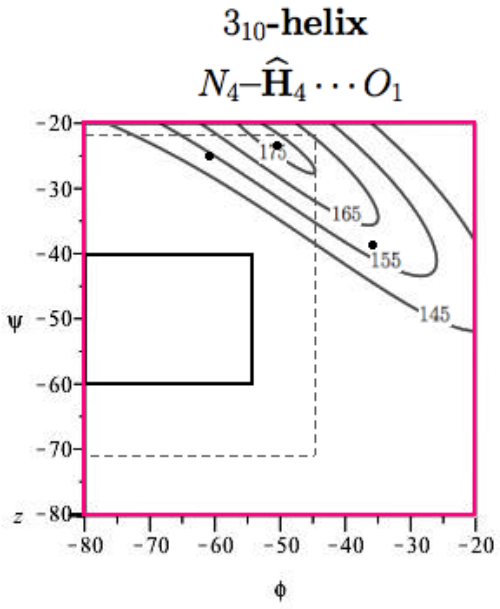
Figure 2.8: Contours for different angles as a function of (φ, ψ) for α-helix.

Figures 2.6 and 2.7).

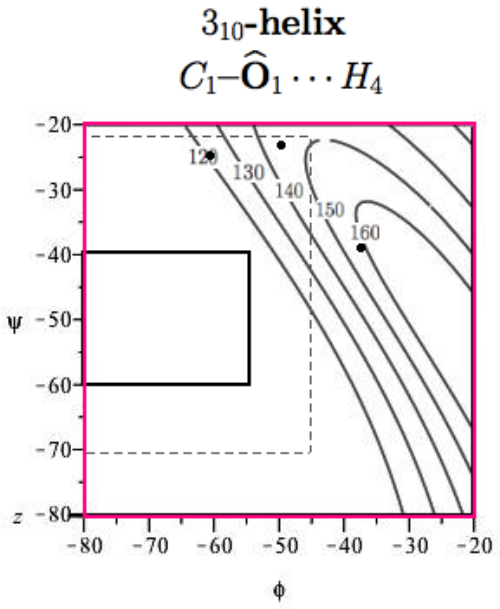
2.5 Summary and conclusions

In this chapter we studied hydrogen bond geometry in protein helices. First, we considered all possible helical configurations of backbone atoms in a chain of amino acids. We assumed ideal polypeptide chain parameters [56] with all backbone bonds of length L and derived formulas for rotation angle θ, rotation axis $\bar{u}$, radius r, and rise d in terms of φ, ψ and L. Thus we showed that any repeated pair of angles φ and ψ defines a helical structure of a backbone. Formula (2.37) for rotation angle was previously published in [24]. To the best of our knowledge explicit formulas for rotation axis (2.38), rise (2.40), and radius (2.42) as functions of φ, ψ and L can not be found in the current literature.

Since not all helical backbone configurations are feasible, we continued by reviewing

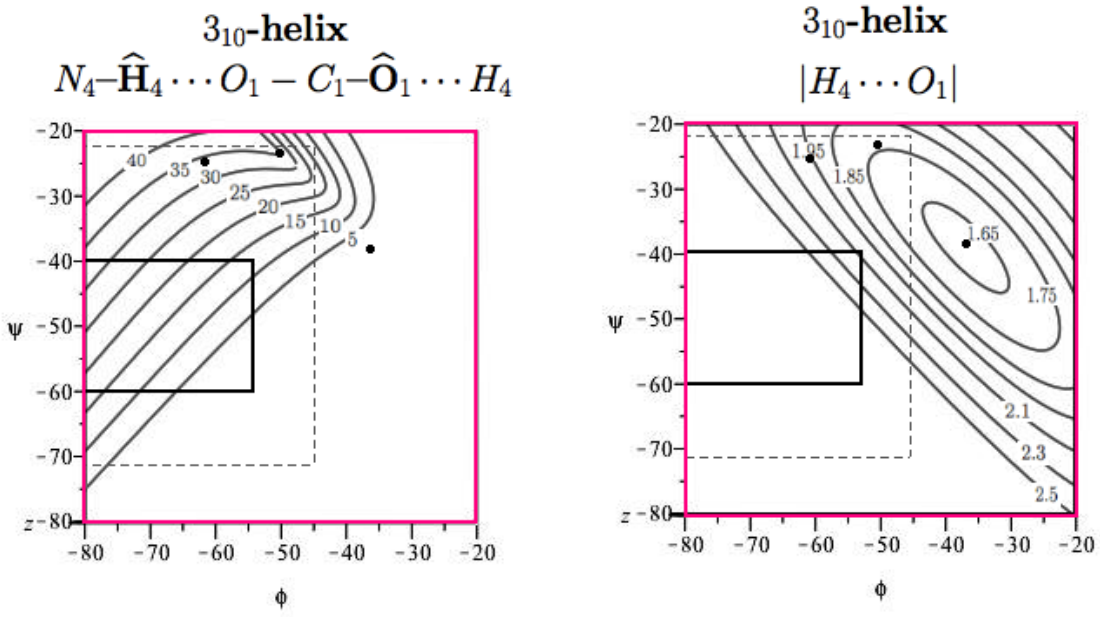


(a) Contours for different values of N_4–$\widehat{H_4 \cdots O_1}$ as a function of (φ, ψ) are plotted.



(b) Contours for different values of C_1–$\widehat{O_1 \cdots H_4}$ as a function of (φ, ψ) are plotted.

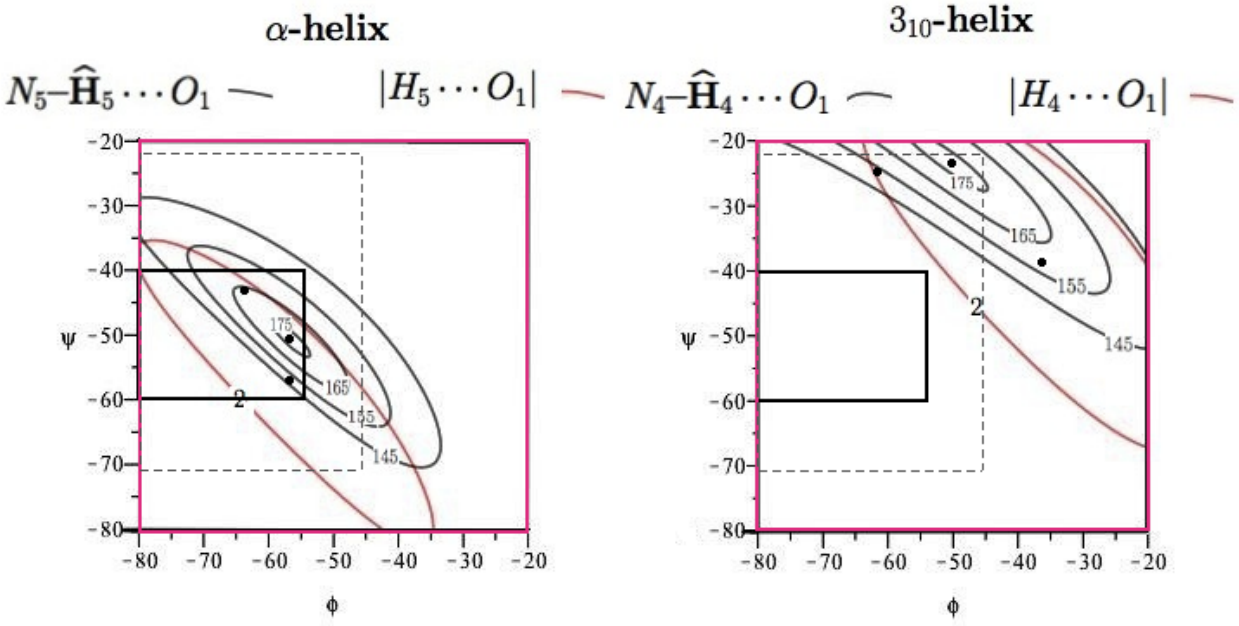
Figure 2.9: Contours for different angles as a function of (φ, ψ) for 3_{10}-helix.



(a) Contours for different values of the difference $(N_4\text{–}\widehat{H_4 \cdots O_1} - C_1\text{–}\widehat{O_1 \cdots H_4})$ as a function of φ, ψ.

(b) Contours for different values of the distance between atoms $|H_4 \cdots O_1|$ as a function of φ, ψ.

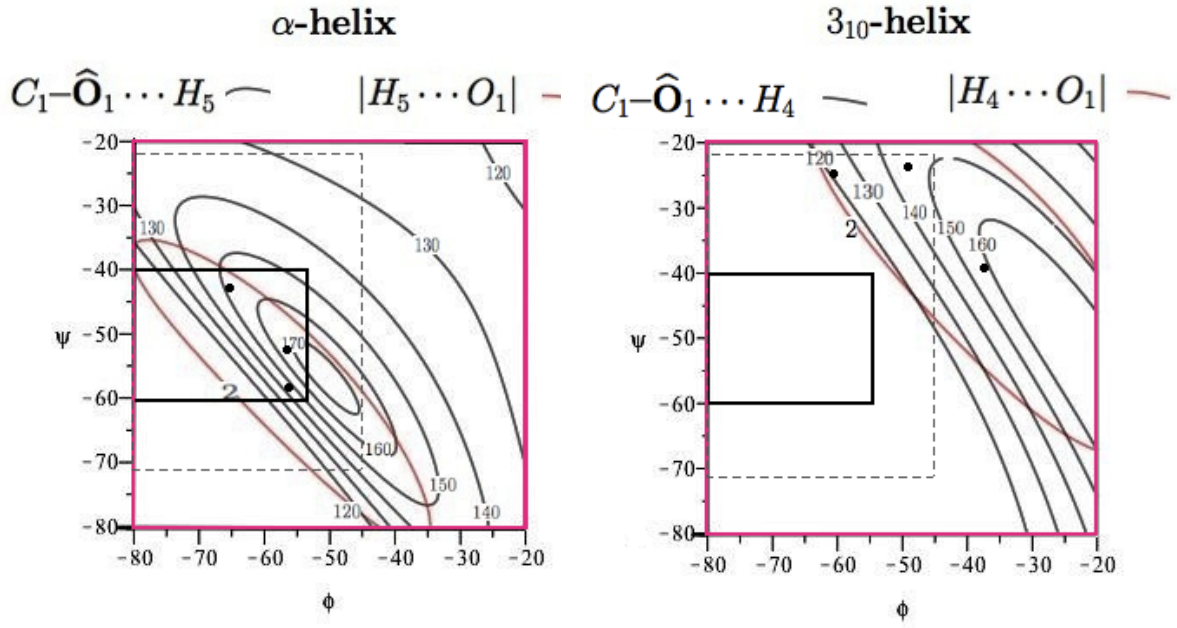
Figure 2.10: Contours for angle differences and distances as a function of φ, ψ for 3_{10} helices.



(a) Contours for different values of N_1–$\widehat{H_1 \cdots O_5}$ and $|H_5 \cdots O_1|$ as functions of (φ, ψ) are plotted.

(b) Contours for different values of N_1–$\widehat{H_1 \cdots O_4}$ and $|H_4 \cdots O_1|$ as functions of (φ, ψ) are plotted.

Figure 2.11: Contours for different values of N_1–$\widehat{H_1 \cdots O_k}$ where $k = 4$ for α-helix and $k = 3$ for 3_{10}-helix. Clearly, a shorter hydrogen bond distance is favored for larger values of N–$\widehat{H \cdots O}$.



(a) α-helix: contours for different values of C_1–$\widehat{O_1 \cdots H_5}$ and $|H_5 \cdots O_1|$ as functions of (φ, ψ).

(b) 3_{10}-helix: contours for different values of C_1–$\widehat{O_1 \cdots H_4}$ and $|H_4 \cdots O_1|$ as functions of (φ, ψ).

Figure 2.12: Contours for different values of C_1–$\widehat{O_1 \cdots H_k}$ and $|H_k \cdots O_1|$ as functions of (φ, ψ), where $k = 5$ for α-helix and $k = 4$ for 3_{10}-helix. A much broader range of C–$\widehat{O \cdots H}$ values is allowed when the hydrogen bond distance is restricted to smaller values.

results on feasible backbone helical configurations from [63] and [60] obtained by Sasisekharan and Ramachandran. These results helped us to identify appropriate optimization domain $-20^\circ \leq \varphi \leq -80^\circ$, and $-20^\circ \leq \psi \leq -80^\circ$ which we used later in Section 2.4 (see Figure 2.3).

Finally, we studied main chain – main chain hydrogen bonds in two types of common helical configurations: α- and 3_{10}-helices. We started by defining a hydrogen bond and its geometrical parameters. We then explained the significance of (φ, ψ) values known as canonical peaks: $(-63^\circ, -43^\circ)$ in α-helix and $(-60^\circ, -25^\circ)$ in 3_{10}-helix [31]. We wanted to know how hydrogen bond geometry influenced the formation of canonical peaks. We computed main chain – main chain hydrogen bond parameters in α- and 3_{10}-helix as functions of dihedral angles (φ, ψ). Optimization with respect to the linearity of a hydrogen bond in α-helix was done by Dix in [25], where the pair (φ, ψ)=$(-56.5^\circ, -50.5^\circ)$ was obtained. We reproduced this result to test the correctness of our computation. Then following the method in [25] we computed a linearity result for 3_{10}-helix and obtained values (φ, ψ)=$(-50.3^\circ, -23.3^\circ)$. Another important geometrical parameter we studied was the length of a hydrogen bond. We computed lengths of main chain – main chain hydrogen bonds in α- and 3_{10}-helices in terms of (φ, ψ). By minimizing these lengths we obtained values (φ, ψ)=$(-56.5^\circ, -57.5^\circ)$ and = $(-36.3^\circ, -39^\circ)$ for α- and 3_{10}-helix respectively. For the pairs $(-56.5^\circ, -50.5^\circ)$, $(-56.5^\circ, -57.5^\circ)$, $(-63^\circ, -43^\circ)$ in α-helix and $(-50.3^\circ, -23.3^\circ)$, $(-36.3^\circ, -39^\circ)$, $(-60^\circ, -25^\circ)$ in 3_{10}-helix, we computed other hydrogen bond geometrical parameters and compared numerical results (see Table 2.5). We then showed that canonical peaks $(-63^\circ, -43^\circ)$ in α-helix and $(-60^\circ, -25^\circ)$ in 3_{10}-helix do not result from optimization of hydrogen bonding with respect to any linear combination of the two major hydrogen bond requirements: linearity and length constraints (see Figures 2.4 (a) and (b)). We also showed that the clustering of the in-plane component $\gamma_{C_i\widehat{-O_i\ldots}H_j}$ around $12^\circ - 13^\circ$ for canonical peak values in both α- and 3_{10}-helix and the relationship between angles $C_i\widehat{-O_i\ldots}H_j$, $N_j\widehat{-H_j\ldots}O_i$ and the distance $|H_j \ldots O_i|$ play a very important role in formation of most common α- and 3_{10}-helices. However, the most obvious contributor to the canonical peak

values in both α-helix and 3_{10}-helix appears to be maximizing solvation of the back bone as indicated in Figure 1.3 and Figure 1.4. We study this in more detail in the next chapter.

CHAPTER 3

SOLVATION SIGNATURE IN α-HELICES

3.1 Proteins interact with water.

Proteins actively interact with hydration water at the surface as well as with water molecules buried inside [44]. Such interactions are essential for their structure and function [61, 14]. Despite the importance, relationship between solvent molecules and peptide structure is still poorly understood and remains an active research topic [44, 71]. α-helix is the most abundant type of protein secondary structure. It plays an essential role within a protein and in various protein-protein interactions [72]. Thus it is especially important to understand how solvation influences the geometry of α-helices within a protein and helps successfully perform its functions.

Water-induced distortion of alpha-helices in three refined X-ray structures of proteins was first detected by Blundell et al. in 1983 [15]. Later some other studies of α-helix interactions with the solvent and side-chains were performed in [11, 68]. More detailed analysis of hydrated alpha-helices was done by Sundaralingam and Sekharudu in 1989 [62]. In [62] authors identified different types of hydrogen bonding between peptide backbone and water molecules, among which the interaction of a backbone carbonyl oxygen with water was the most common. In this type of interaction a water molecule is hydrogen bonded to the backbone carbonyl oxygen of an α-helix (see Figure 3.1). A nice summary on the topic was written in 2003 by McColl et. al. [50] where the importance of the main-chain carbonyl oxygen hydration was reemphasized as the most common and "especially relevant to aqueous solution studies." In this chapter we focus our study on the hydration of main-chain carbonyl oxygens, examining changes in a local α-helical geometry of a dipeptide as a function of the distance between backbone carbonyl oxygen and the closest water molecule.

Our earlier analysis of α-helix geometry in the previous chapter suggests that the values of dihedral angles $(\varphi, \psi) = (-63^\circ, 43^\circ)$, also known as the *canonical* peak [31], in most common

helices result as a balance between two competing criteria: maximizing the main-chain hydrogen bond linearity, which in ideal α-helix is achieved at values $(\varphi, \psi) = (-56.5°, -50.5°)$, and maximizing the distance between neighboring main-chain oxygens $d_{O_i O_{i+1}}$, see Figure 1.3. The linearity constraint is essential for strong hydrogen bonds that allow specific helical geometry to hold. The latter criteria is important because it favors the necessary hydration of backbone carbonyl oxygens. We confirm this by analyzing a large dataset of high resolution protein structures rich in α-helical and water content.

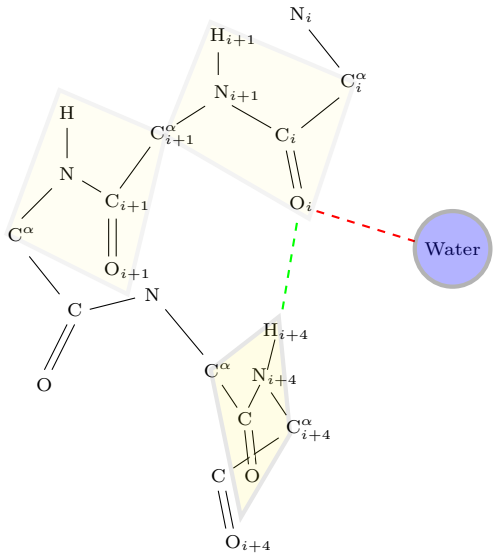


Figure 3.1: **Helical backbone carbonyl oxygen interacts with water molecule**: water molecule (blue) forms a hydrogen bond with the main-chain carbonyl oxygen O_i (red dashed line) which is also hydrogen bonded to the main-chain hydrogen H_{i+4} (green dashed line), following the regular α-helical hydrogen bonding pattern.

3.2 Measuring backbone solvation in α-helices

3.2.1 Helix nomenclature

We use helix nomenclature according to [5]:

$$\cdots - N'' - N' - Ncap - N1 - N2 - N3 - \cdots - C3 - C2 - C1 - Ccap - C' - C'' - \cdots,$$

Where $Ncap - N1 - N2 - N3$ and $C3 - C2 - C1 - Ccap$ are corresponding N- and C-terminus, and $N4 - N5 - \cdots - C5 - C4$ comprise the middle residues of the helix.

3.2.2 Computing geometrical parameters

In order to capture instances of interaction of a helical backbone carbonyl oxygen with water we consider main-chain carbonyl oxygens in α-helices and look for the closest water molecule. To do this for each residue i, such that the i-th and $(i+1)$-th residues are both part of an α-helix, we measure distances between carbonyl oxygen O_i and all water oxygens available in the experimental structure and record the shortest such distance $d_i = \min_{\forall W} d_{O_i O_W}$, with the cut-off value of 10 Å. We also consider relevant α-helical geometrical parameters, namely, distances $d_{O_i C^{\beta}_{i+1}}$ and $d_{O_i O_{i+1}}$, and dihedral angles ψ_i, $\varphi_{(i+1)}$, ψ_{i+1}, see Figure 3.2. From ideal geometry computation we expect $d_{O_i C^{\beta}_{i+1}}$ not to change significantly, see Figure 3.3 **(a)**. On the other hand, distance $d_{O_i O_{i+1}}$ is increasing with increasing φ_{i+1}, see Figure 3.3 **(b)**. We perform a detailed study of the changes in these angles and distances when helical backbone carbonyl oxygen interacts with water molecule.

3.2.3 Dataset

We considered all `.pdb` files available in Protein Data Bank [13] (as of July 23, 2016) with resolution < 1.5 Å. We found 8754 such structures. After selecting files containing at least

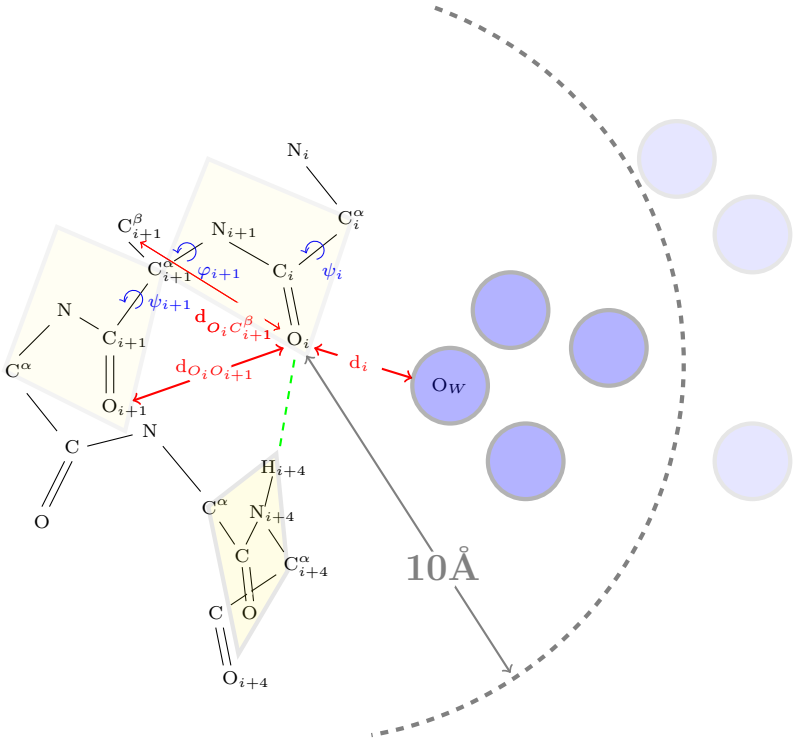


Figure 3.2: We consider a pair of consecutive i-th and $(i+1)$-th residues in an α-helix. For each such pair we measure dihedral angles $\psi_i, \varphi_{i+1}, \psi_{i+1}$, and distances $d_{O_iO_{i+1}}, d_{O_iC^{\beta}_{i+1}}$ (given that (i+1)th residue is not a GLY), and d_i, where distance $d_i = \min_{\forall W, d_{O_iO_W} < 10\mathring{A}}(d_{O_iO_W})$, i.e. the distance from the carbonyl oxygen O_i to the oxygen of the closest water molecule is within 10Å range.

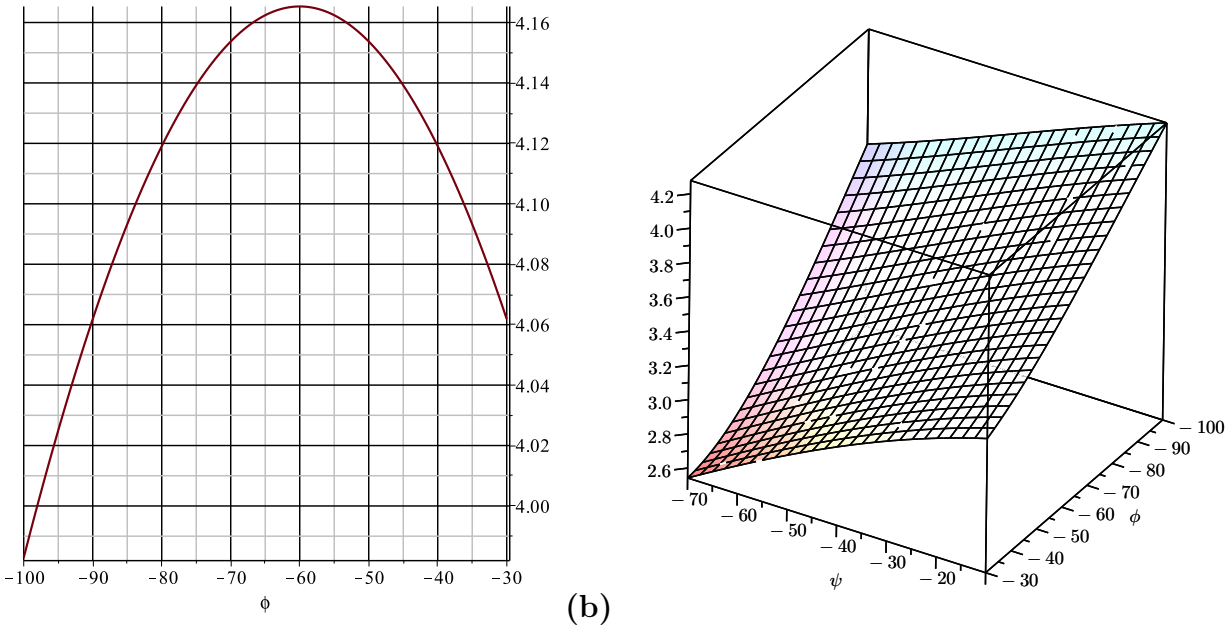


Figure 3.3: **(a)** Dependence of distance $d_{O_iC^{\beta}_{i+1}}$ (Y-axis) on angle φ_{i+1} (X-axis); **(b)** Dependence of distance $d_{O_iO_{i+1}}$ (Z-axis) on angles ψ_{i+1} (Y-axis) and φ_{i+1} (X-axis) are shown for the case of ideal geometry [56].

HR dataset (Res $<$ 1.5Å, R$\leq$30)	**Protein Structures**	**Number of α-helices**	**Qualified Dipeptides**			**Waters**
			Nter	**Mid**	**Cter**	
total	2176	20838	70065	109410	53482	990258

Table 3.1: Number of relevant protein structures, α-helices and waters in high resolution protein dataset. In the column **Qualified Dipeptides**: **Nter** corresponds to dipeptides Ncap-N1, N1-N2, N2-N3, and N3-N4, found at the N-terminus of a helix; **Mid** corresponds to the dipeptides N4-N5, $\cdots$, C5-C4, C4-C3, located in the middle of a helix; and **Cter** corresponds to dipeptides C3-C2, C2-C1, C1-Ccap at the C-terminus.

one α-helix, and information about the location of (at least one) water molecule we were left with 8117 structures. We then culled these files for identical strictures using PISCES server [69], restricting the R-factor, a measure of the agreement between the crystallographic model and the experimental X-ray diffraction data (or in other words a measure how well the refined structure predicts the observed data) to < 0.3, and sequence percentage identity to ≤ 25. We obtained a dataset of 2176 structures with the total of 20838 α-helices, and 990258 waters. Amino-acids with alternative C_α positions were omitted from computations.

For high resolution protein dataset statistics see Table 3.1.

3.3 Solvation effects on backbone geometry in α-helices

3.3.1 Results for high resolution protein dataset.

Basic dataset statistics is given in Table 3.2. We found that helical backbone actively interacts with water irrespective of whether it is located at the surface or buried inside protein. As can be seen in Figure 3.4, almost every backbone carbonyl oxygen has a water molecule within 10Å, and more than a half – within 5Å. This explains why helical amino acid propensities computed in [29, Table 2 on p.5], do not differ for exposed versus buried residues. It is important to note that three peaks for $d_i \geq 5$Å at ≈ 5.1Å, ≈ 5.8 Å, and ≈ 6.7Å appear as an artifact of our definition of d_i and in many cases correspond to the primary water contacts with O_{i-1}, O_{i+1}, and O_{i+2} accordingly. These peaks are removed when we only consider unique shortest contact with any specific water, see the updated histogram in Figure 3.7.

Helical backbone interaction with water, as expected, highly correlates with larger $|\varphi|$ angles and larger distances between neighboring main-chain carbonyl oxygens $d_{O_iO_{i+1}}$. The details of the relationship between parameters d_i, φ_{i+1}, and $d_{O_iO_{i+1}}$ are summarized in Figures 3.6-3.17.

In Figure 3.6 histograms of φ_{i+1} and $d_{O_iO_{i+1}}$ are plotted for the full dataset, and for cases organized by d_i as primary, secondary, and no water contacts. We see a clear shift for both parameters in the case of the primary water contact with respect to the cases of secondary and no water contact. We also see that the canonical peak $\varphi = -63^\circ$ in the total dataset results from the combination of two different peaks: $\varphi = -65^\circ$ corresponding to the case of the primary water contact, and $\varphi = -62^\circ$ for other cases. This effect is preserved when we only consider unique shortest contact with any specific water, see Figure 3.5.

To better indicate the changes in φ_{i+1} and $d_{O_iO_{i+1}}$ parametrized by d_i we made scatter plots with the least squares fit lines shown, see Figure 3.8. To see the landscape of data points distribution we plotted 3D histograms and heatmaps in Figures 3.9, 3.10, 3.11.

The correlation between the proximity of water and the type of amino acids in the

d_i (Å)	[0,3.5)	[3.5,5)	[5,10)	Total
Number of dipeptides	31990	27239	50182	109410
mean φ_{i+1} (°)	-65.4 ± 6.5	-62.2 ± 4.7	-62.2 ± 4.6	-63.1 ± 5.4
mean $d_{O_iO_{i+1}}$ (Å)	3.48 ± 0.20	3.42 ± 0.16	3.42 ± 0.15	3.44 ± 0.17
Unique water contacts only	30360	19778	16896	67034
mean φ_{i+1} (°)	-65.4 ± 6.3	-62.2 ± 4.4	-62.2 ± 4.6	-63.7 ± 5.6
mean $d_{O_iO_{i+1}}$ (Å)	3.48 ± 0.19	3.42 ± 0.16	3.43 ± 0.17	3.45 ± 0.18

Table 3.2: Results for high resolution protein dataset for middle residues.

dipeptide can be seen in Figure 3.12. This suggests that the geometry of backbone carbonyl oxygen interaction with water could be different for different amino acids. In particular, the distance $d_{O_iO_{i+1}}$ depending of the value of φ_{i+1} can grow differently for different types of amino acids. Indeed, if we plot the least squares fit lines to the data points broken into groups corresponding to the total dataset and for primary, secondary and no water contact cases, only the slight difference is observed, see Figure 3.13, as well as when considering similar plots but for different amino acids at ith position in $(i, i+1)$ dipeptide, see Figures 3.15, 3.16 (left), and 3.17 (left). On the contrary, when we do a similar plot grouping data by the kind of amino acid at $(i+1)$th position we see more diversity in φ_{i+1} and $d_{O_iO_{i+1}}$ dependence, see Figure 3.14, 3.16 (right), and 3.17 (right). Especially for the case of the primary water contact in Figure 3.14, where the lines corresponding to different amino acids are noticeably well ordered. It is very interesting that this order provides a nice classification of amino acids connecting secondary structure propensities with geometrical possibilities for backbone solvation, as demonstrated in Table 3.3.

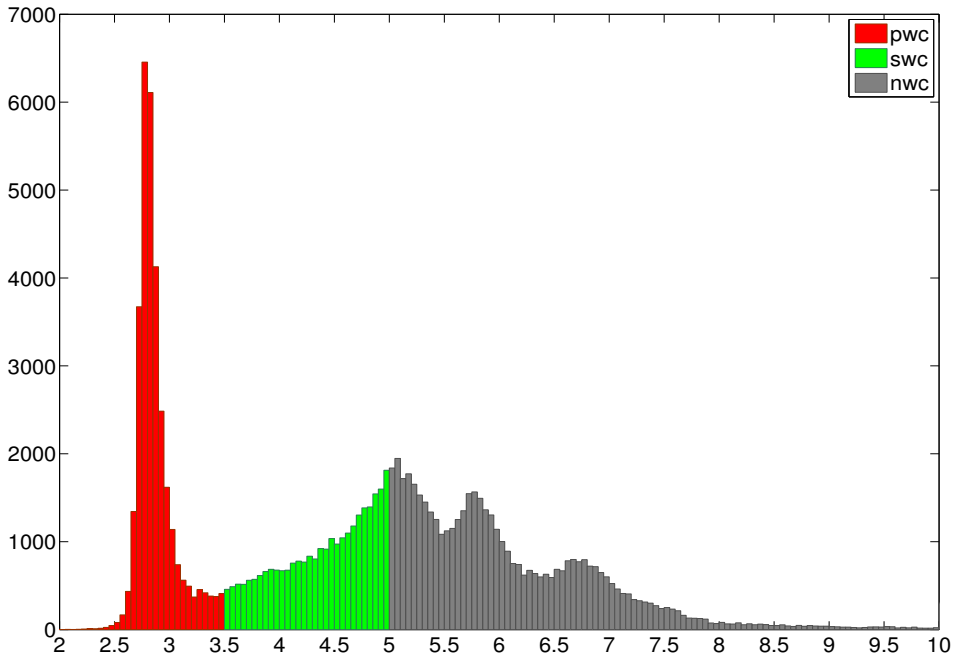


Figure 3.4: Histogram of distances d_i for middle residues in α-helices in high resolution protein dataset colored by the intervals corresponding to: primary water contacts [0,3.5) (red), secondary water contacts [3.5,5) (green) and no water contacts [5,10) (grey). Note, that the three peaks in the grey area at $\approx$ 5.1Å, $\approx$ 5.8 Å, and $\approx$ 6.7Å appear as an artifact of our definition of d_i and in many cases are formed by the primary water contacts with O_{i-1}, O_{i+1}, and O_{i+2} accordingly. Also, the choice of the 10Å cut off guarantees the inclusion of almost all middle residues (there is a very insignificantly small amount of residues that has water farther than 10Å.)

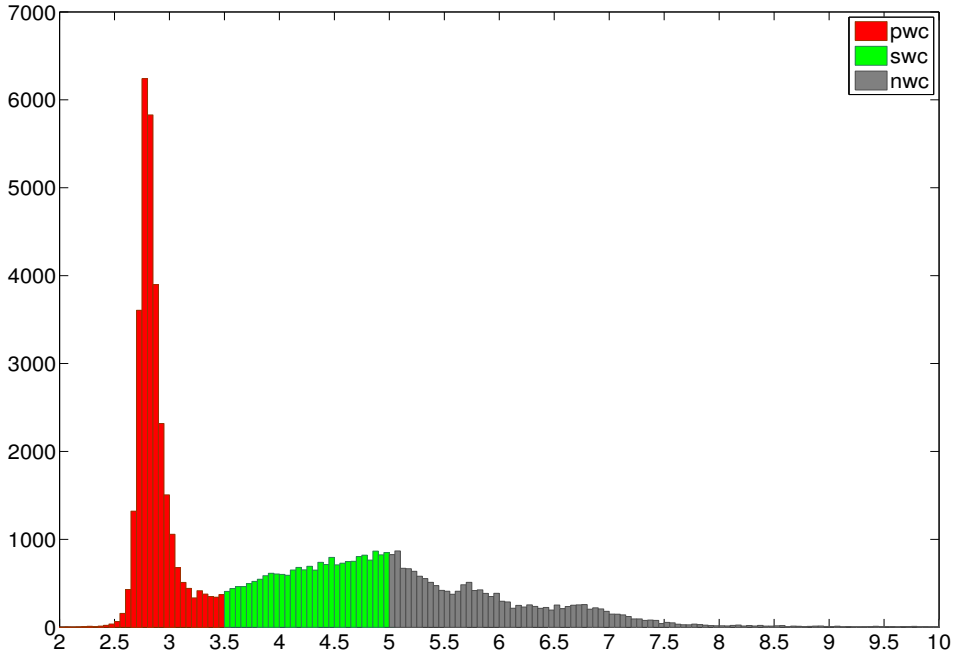


Figure 3.5: Histogram of distances d_i for middle residues in α-helices in high resolution protein dataset colored by the intervals corresponding to: primary water contacts [0,3.5) (red), secondary water contacts [3.5,5) (green) and no water contacts [5,10) (grey). Here only the unique shortest contact with any specific water is maintained. The artificial peaks $\approx$ 5.1Å, $\approx$ 5.8 Å, and $\approx$ 6.7Å, seen in Figure 3.4, are gone.

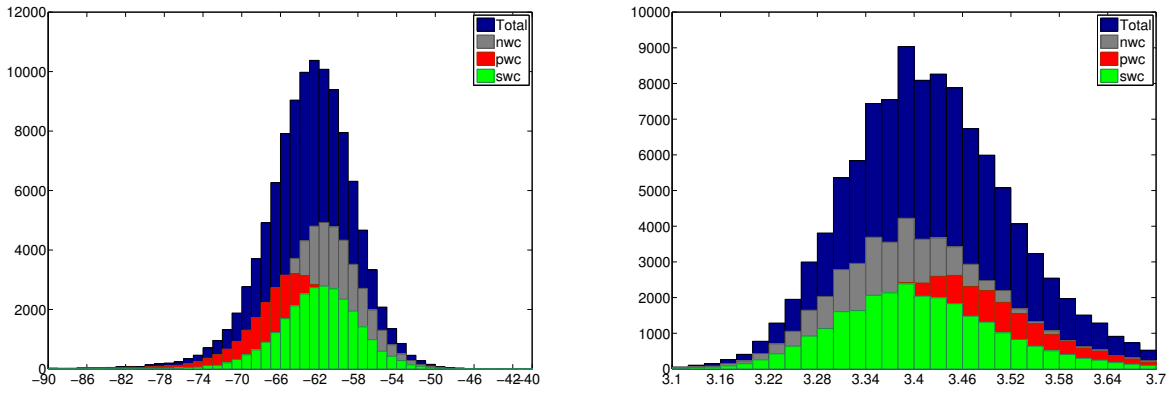


Figure 3.6: Histograms of φ_{i+1} (left) and $d_{O_iO_{i+1}}$ (right) organized by d_i for primary water contacts [0,3.5) (red), secondary water contacts [3.5,5) (green) and no water contacts [5,10) (grey) and the total (blue). In both figures a clear shift is seen for dipeptide parameters in the case of the primary water contact with respect to the cases of secondary and no water contact. We also see that the canonical peak $\varphi = -63^\circ$ in the total dataset results from the combination of two different peaks: $\varphi = -65^\circ$ corresponding to the case of the primary water contact, and $\varphi = -62^\circ$ for other cases.

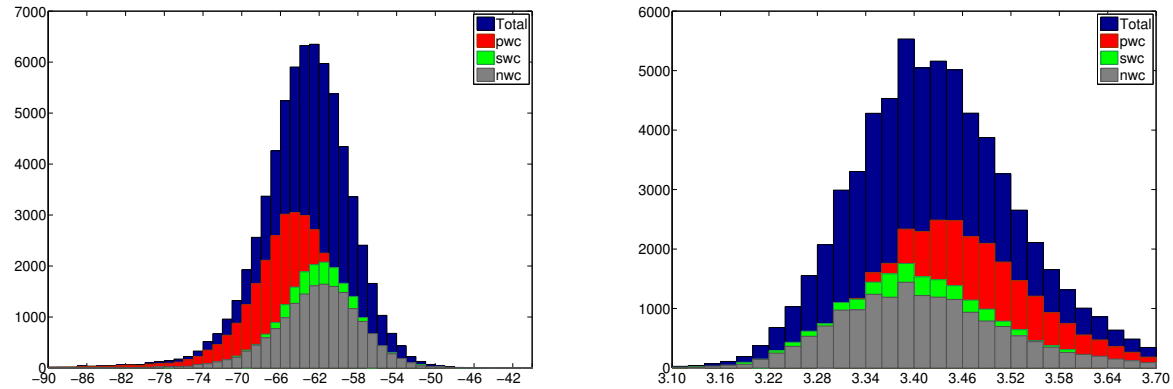


Figure 3.7: Histograms of φ_{i+1} (left) and $d_{O_iO_{i+1}}$ (right) organized by d_i for primary water contacts [0,3.5) (red), secondary water contacts [3.5,5) (green) and no water contacts [5,10) (grey) and the total (blue). Here only the shortest contact with any specific water is maintained. In both figures the same shift is seen for dipeptide parameters in the case of the primary water contact with respect to the cases of secondary and no water contact. We also see that the canonical peak $\varphi = -63^\circ$ in the total dataset results from the combination of two different peaks: $\varphi = -65^\circ$ corresponding to the case of the primary water contact, and $\varphi = -62^\circ$ for other cases.

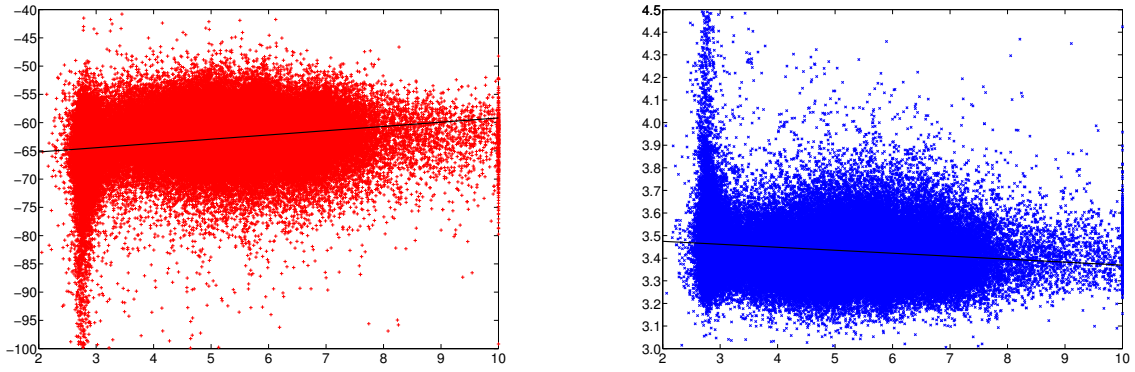


Figure 3.8: Scatter plots of d_i (X-axis) and $\varphi_{(i+1)}$ (left) and $d_{O_iO_{i+1}}$ (right) (Y-axis) is shown for high resolution protein dataset with the least squares fit lines (black) shown to better indicate the changes in $\varphi_{(i+1)}$ and $d_{O_iO_{i+1}}$ parametrized by d_i.

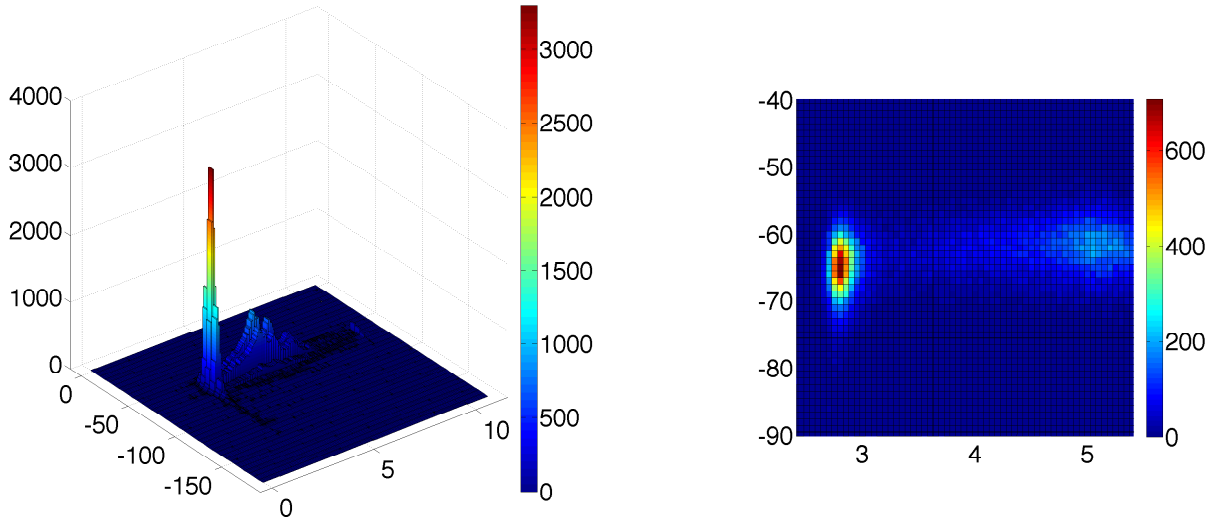


Figure 3.9: 3D histogram (left) and heatmap (right) of dependence of $\varphi_{(i+1)}$ (Y-axis) on distance d_i (X-axis) is shown for high resolution protein dataset.

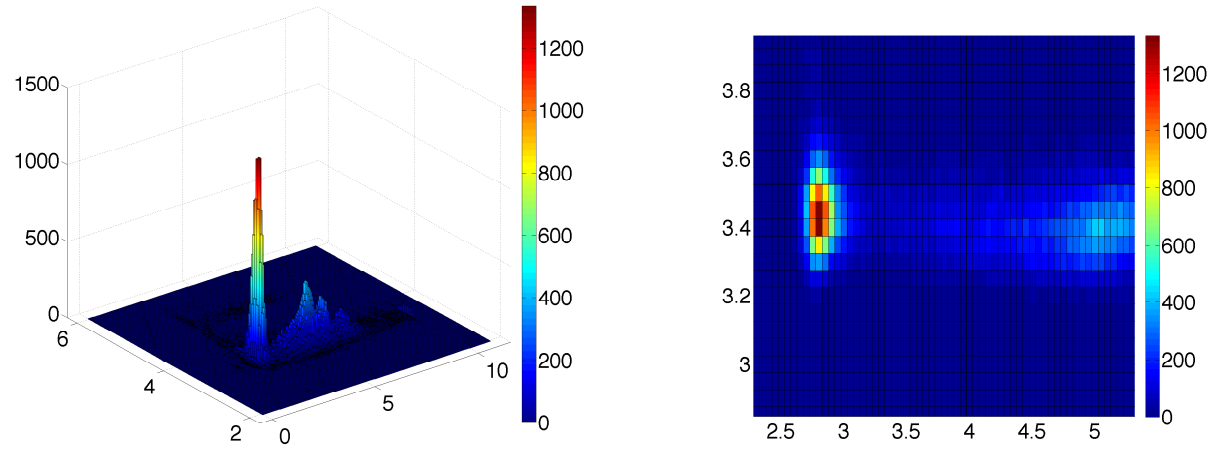


Figure 3.10: 3D histogram (left) and heatmap (right) of dependence of $d_{O_iO_{i+1}}$ (Y-axis) on distance d_i (X-axis) is shown for high resolution protein dataset.

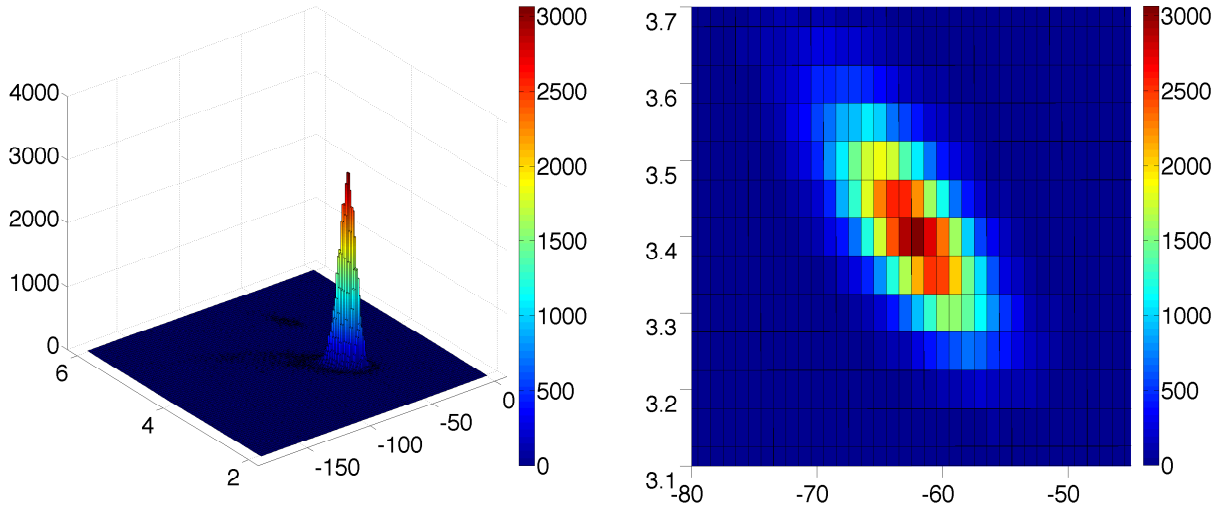
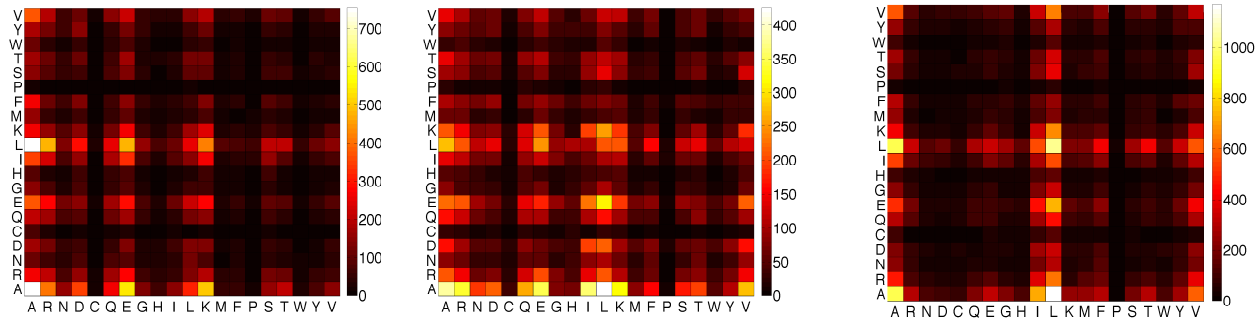


Figure 3.11: 3D histogram (left) and heatmap (right) of dependence of $\varphi_{(i+1)}$ (X-axis) on $d_{O_iO_{i+1}}$ (Y-axis) for high resolution protein dataset. Peak: $(\varphi_{(i+1)}, d_{O_iO_{i+1}}) = (-62°, 3.40\text{Å})$



(a) **(b)** **(c)**

Figure 3.12: Amino acid composition for middle (i. e. excluding N- and C-ter) residues i (X-axis) and $i+1$ (Y-axis) in α-helices: (a) when ith residue carbonyl oxygen has water oxygen within [0,3.5)Å; (b) when ith residue carbonyl oxygen has water oxygen within [3.5,5)Å; (c) when ith residue carbonyl oxygen has water oxygen within [5,10)Å.

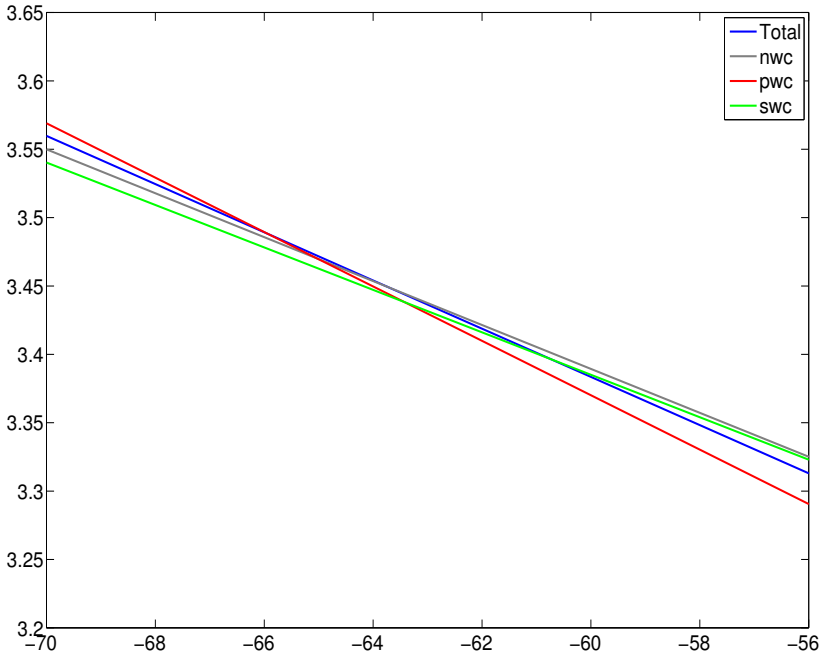


Figure 3.13: Least squares fit lines to scatter plots (the actual points not shown) of φ_{i+1} (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by d_i for primary water contacts [0,3.5)Å(red), secondary water contacts [3.5,5)Å(green) and no water contacts [5,10)Å(grey) and the total (blue)

Group	Amino Acid	Propensities	
		β-strand	α-helix
I	**S**	0.8	0.7
	P	0.4	0.4
	G	0.7	0.4
	N	0.6	0.7
	D	0.6	0.8
II	**A**	0.8	1.4
	C	1.4	0.9
	L	1.1	1.3
	Q	0.7	1.3
	K	0.8	1.2
	R	0.9	1.2
	E	0.7	1.4
III	**W**	1.2	1.1
	M	1.0	1.3
	Y	1.4	1.0
	H	1.0	0.9
	F	1.4	1.0
	T	1.2	0.8
	V	2.0	0.9
	I	1.8	1.0

Table 3.3: Propensity values correspond to total propensities in [29, Table 2 on p.5]. Amino acids are listed in the order of the corresponding lines in Figure 3.14. In that specific order there are three groups: (I) - amino acids that have low propensity for both α-helix and β-sheet, (II) - amino acids that have a higher propensity for α-helix, (III) - amino acids that have a higher propensity for β-sheet. Highlighted are the outliers, Cys and Met. Both amino acids do not fit into the propensity preference of their respective group. This effect is due to the significant underrepresentation of Cys and Met in α-helices, and thus in the full dataset. Overall, it is interesting that as the ith backbone carbonyl oxygen has a primary interaction with water the way how $d_{O_iO_{i+1}}$ changes as φ_{i+1} increases provides a nice amino acid classification that connects backbone solvation to secondary structure propensities

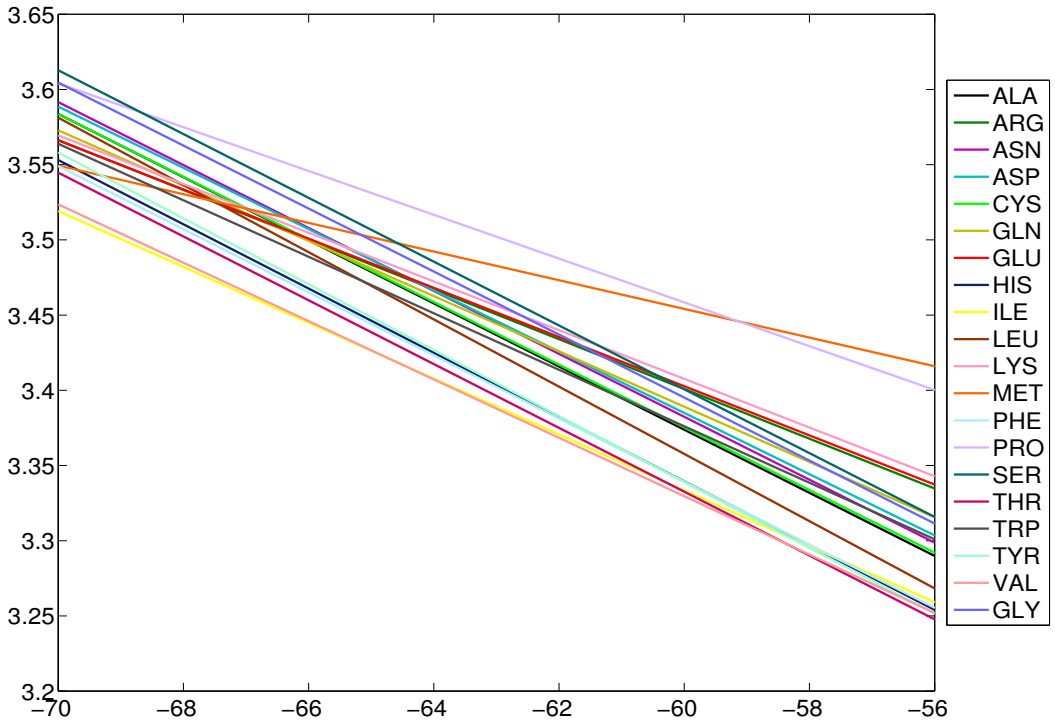


Figure 3.14: Least squares fit lines to scatter plots (the actual points not shown) of φ_{i+1} (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by the residue type located at the $(i+1)$th position, when the ith backbone carbonyl oxygen has a primary contact with water, i. e. $d_i \in [0, 3.5)$Å. We see that the lines are well ordered. This order provides a nice classification of amino acids connecting secondary structure propensities with geometrical possibilities for backbone solvation, as demonstrated in Table 3.3.

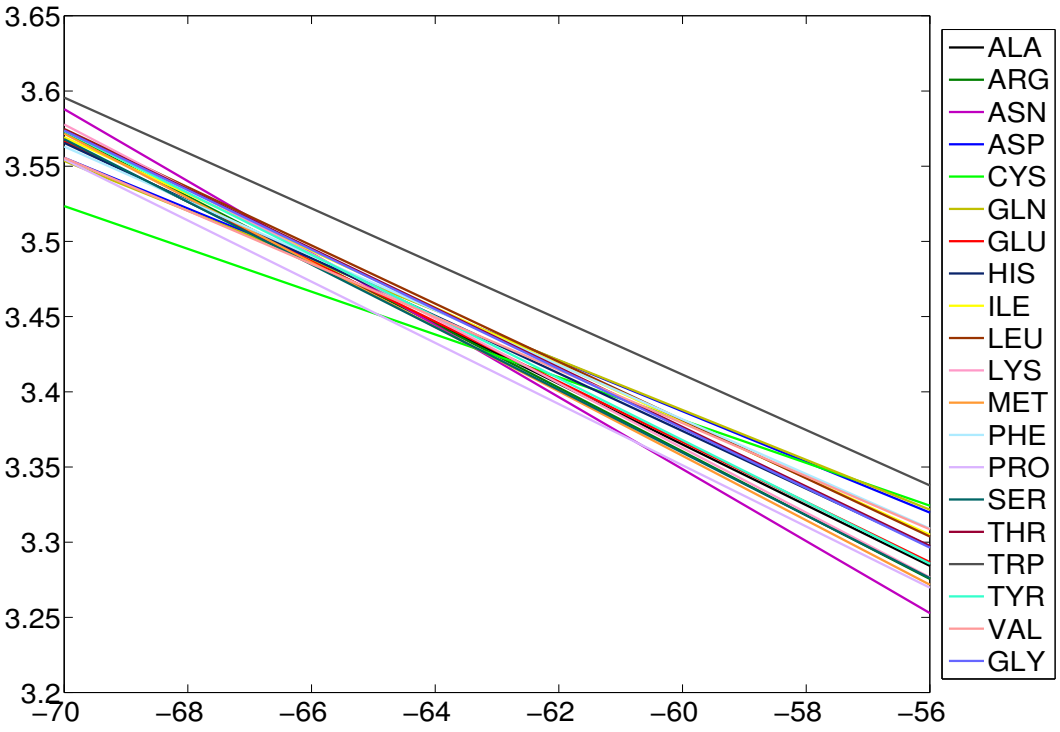


Figure 3.15: Least squares fit lines to scatter plots (the actual points not shown) of φ_{i+1} (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by the residue type located at the ith position, when the ith backbone carbonyl oxygen has a primary contact with water, i. e. $d_i \in [0, 3.5)$Å.

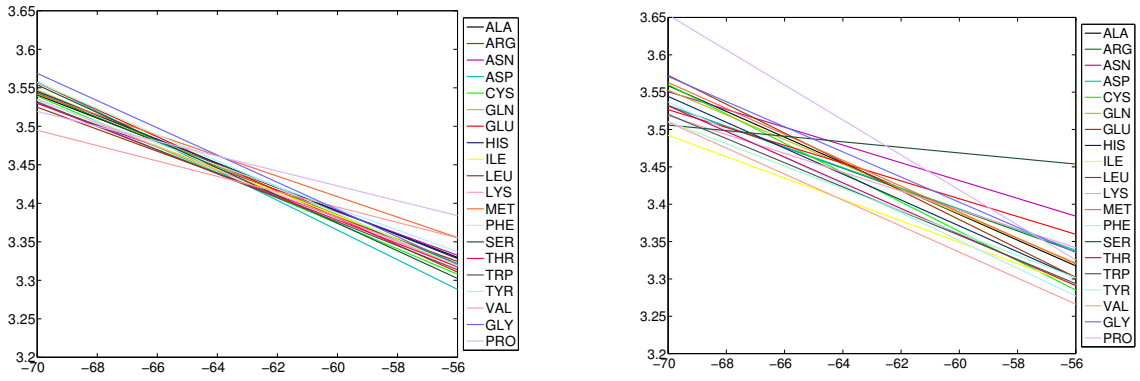


Figure 3.16: Least squares fit lines to scatter plots (the actual points not shown) of φ_{i+1} (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by the residue type located at the ith position (left) and $(i+1)$ (right) when the ith backbone carbonyl oxygen has a secondary contact with water, i. e. $d_i \in [3.5, 5)$Å.

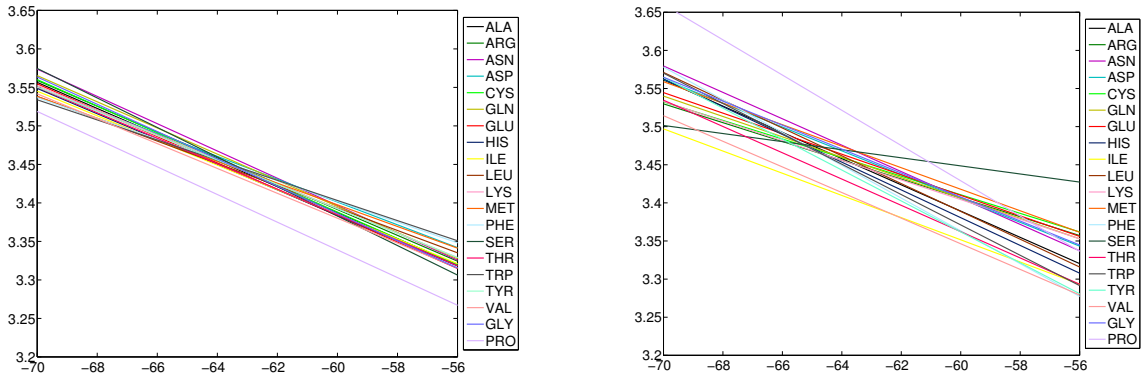


Figure 3.17: Least squares fit lines to scatter plots (the actual points not shown) of φ_{i+1} (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by the residue type located at the ith position (left) and $(i+1)$ (right) when the ith backbone carbonyl oxygen has no contact with water, i. e. $d_i \in [5, 10)$Å.

Data	mean φ, (°)	mean ψ, (°)	mean $d_{O_iO_{i+1}}$, (Å)	PWC	SWC	AA Seq.
CPCM	-61.3 ± 0.4	-44.5 ± 0.8	3.45 ± 0.02	0	0	aaaaaaa
4C3N	-61.0 ± 0.4	-44.7 ± 0.6	3.45 ± 0.04	0	0	aaaaaaa
4C3N11	-65.3 ± 0.9	-42.0 ± 0.7	3.50 ± 0.05	5	2	AAA**a**AAa

Table 3.4: We looked at three models of polyalanine α-helix placed in: **(i) CPCM** – continuum CPCM solvent, **(ii) 4C3N** – continuum CPCM solvent with 4 explicit water molecules added at the C-terminus, and 3 – at the N-terminus, and **(iii) 4C3N11B** – continuum CPCM solvent with 4 explicit water molecules at the C-terminus, 3 – at the N-terminus, and 11 – at the backbone between termini in [49]. Using data provided in the paper we computed mean φ, ψ and $d_{O_iO_{i+1}}$ values for the middle residues in all three cases. Columns **PWC** and **SWC** have number of backbone carbonyl oxygens that have a water molecule within [0,3.5)Å (primary interaction) and $\geq$3.5Å respectively. The rightmost column contains a sequence representation, where "a" is used for Alanine, with capital letters standing for the case when carbonyl oxygen at that position has a primary interaction with a water molecule.

3.3.2 Comparison of our results to the known results obtained via density functional theory computations in [49].

In the conclusion of this section we compare our data mining results with the results obtained by Marianski et. al. in 2012 [49]. This most recent to our knowledge study presents various models of polyalanine α-helices solvated in water obtained via density functional theory calculations. We used atom locations data provided by the authors in a supplementary file. In particular, we looked at three models of polyalanine α-helix placed in: **(i) CPCM** – continuum CPCM solvent, **(ii) 4C3N** – continuum CPCM solvent with 4 explicit water molecules added at the C-terminus, and 3 at the N-terminus, and **(iii) 4C3N11B** – continuum CPCM solvent with 4 explicit water molecules at the C-terminus, 3 at the N-terminus, and 11 at the backbone between termini.

In Figure 3.18 we see that changes in φ angle from cases (i) and (ii) to the case (iii) are comparable with our results. Moreover mean values for φ_{i+1} and $d_{O_iO_{i+1}}$ in density functional theory computations for the backbone between termini also show a similar increase of 4° and 0.05Å respectively from the unsolvated states, i.e. cases (i) and (ii), to the solvated state, i.e. case (iii), see Table 3.4 for details.

Finally, it is worth pointing out that in the case (iii) the distances between backbone car-

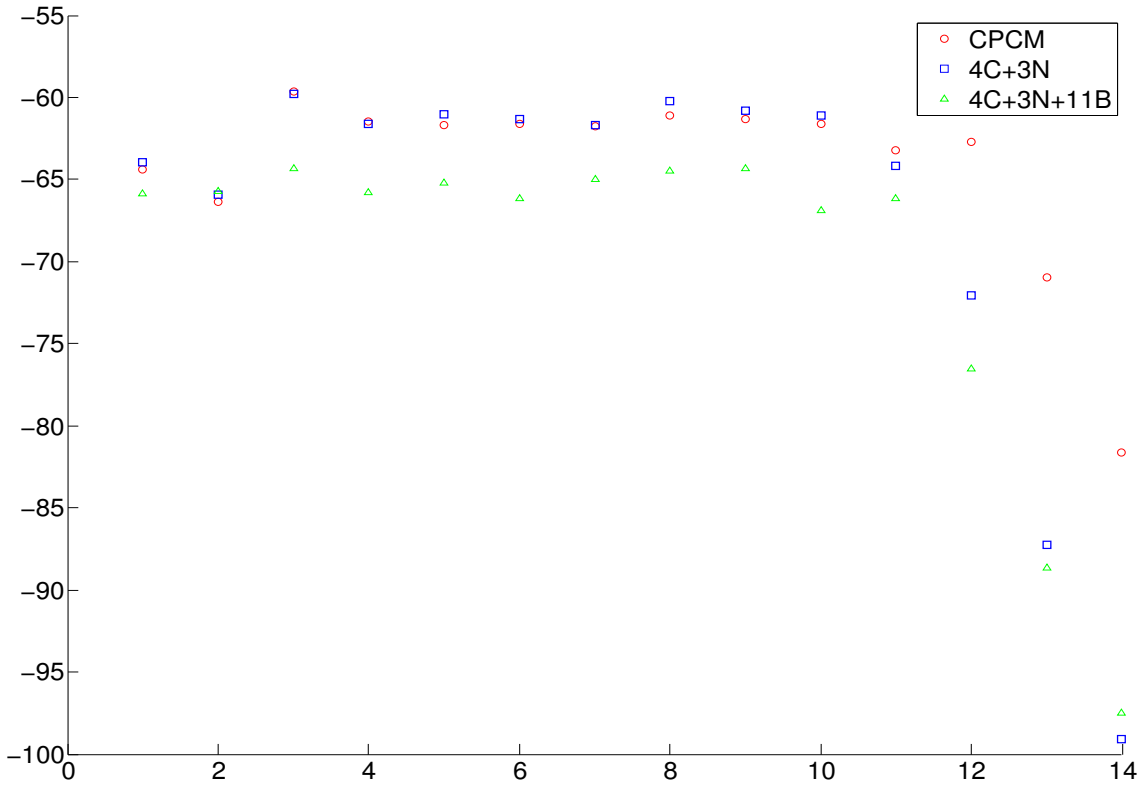


Figure 3.18: φ values plotted for three cases corresponding to the models of polyalanine α-helix placed in: **(i) CPCM** – continuum CPCM solvent, **(ii) 4C3N** – continuum CPCM solvent with 4 explicit water molecules added at the C-terminus and 3 at the N-terminus, and **(iii) 4C3N11B** – continuum CPCM solvent with 4 explicit water molecules at the C-terminus, 3 at the N-terminus, and 11 at the backbone between termini. Atom locations were taken from [49].

i	d_i, (Å)	$d_{O_iO_{i+1}}$, (Å)
5	3.443	3.493
6	3.363	3.499
7	3.488	3.478
8	3.536	3.457
9	3.389	3.456
10	3.341	3.548
11	5.010	3.596

Table 3.5: Computed d_i and $d_{O_iO_{i+1}}$, where residue position in the helix $i \in [5, 11]$, for the case (iii) of polyalanine α-helix solvated at the termini and backbone using density functional theory computations. Atom locations were taken from [49]. Here d_i values vary significantly from the mean 2.8Å observed by us in the experimental dataset.

Length	Quantity	Length	Quantity
3	3	21	371
4	22	22	274
5	242	23	219
6	**1997**	24	164
7	1266	25	145
8	1368	26	95
9	1544	27	90
10	1467	28	63
11	1617	29	55
12	**1639**	30	70
13	**1533**	31	45
14	**1304**	32	37
15	**1484**	33	40
16	**1118**	34	19
17	**853**	35	9
18	**741**	36	8
19	**651**	≥37	79
20	**479**		

Table 3.6: Helix length statistics for high resolution protein dataset. Helix lengths in bold were considered for backbone solvation pattern, see Figures 3.20, 3.21, and 3.22. For higlighted lengths see full histograms of solvated and unsolvated amino acids for every helical position in Figures 3.24-3.27.

bonyl oxygens and water molecules are much larger than the distance of 2.8Å most commonly observed in the experimental data, see Table 3.5.

3.4 Solvation patterns in α-helices

3.4.1 Solvation dependent amino acid composition in α-helices

Backbone hydrogen bonding in helices is highly directional. Residues at N-terminus and C-terminus have unsatisfied hydrogen-bond donors and acceptors respectively. Moreover, such directionality should impose restrictions in the helical interior (i.e. middle residues) as well. Various cases of amino-acid positional propensities at helical termini have been noticed and studied. The middle residues however most often are either characterized together, or amino acid propensities for pairs $(i, i+k)$, where $k = 1, ..., 4$ are considered in connection with

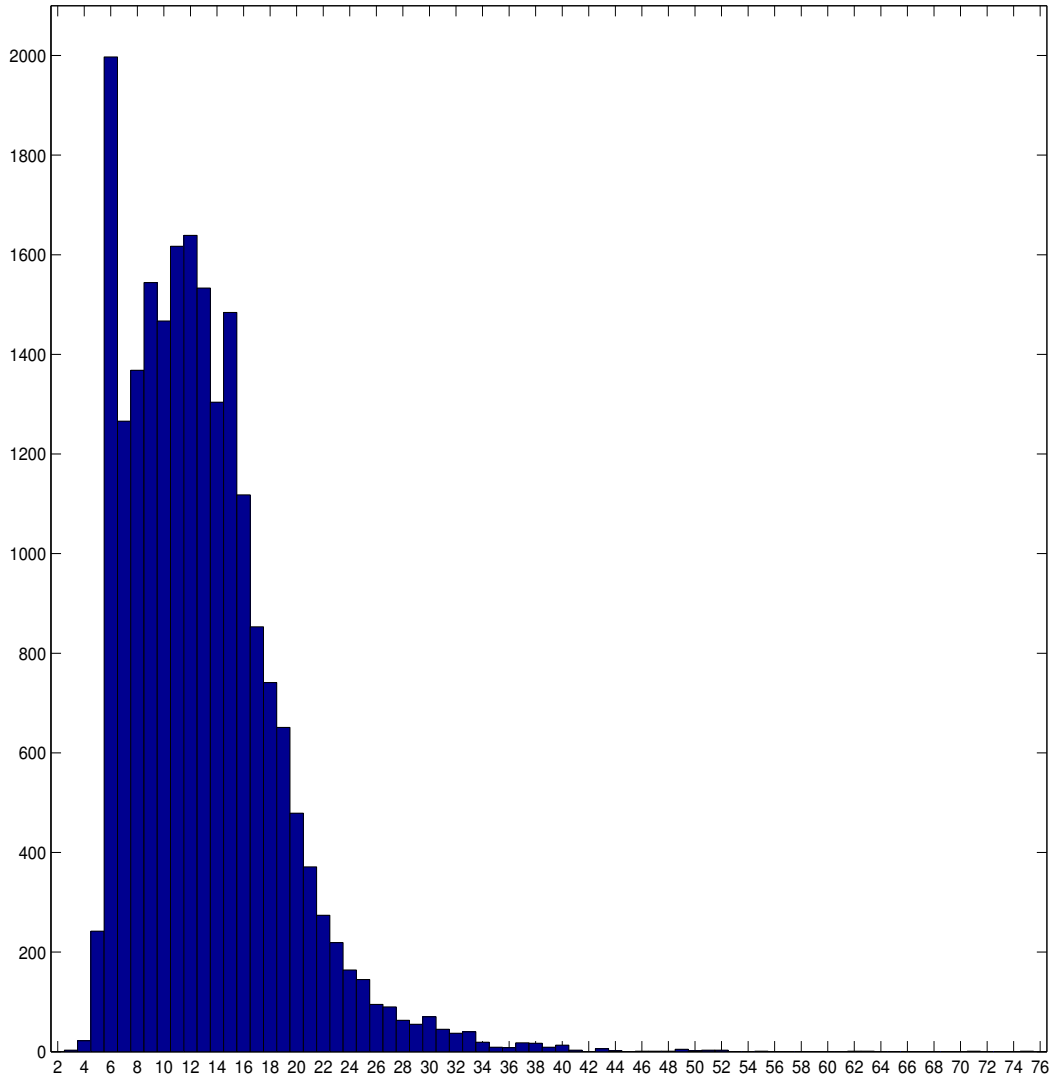


Figure 3.19: Histogram of α-helix lengths in high resolution protein dataset.

main-chain main-chain hydrogen bond, salt bridge, or other commonly classified interaction irrespective of their exact position within a helix interior.

Here we connect positional amino acid propensities with backbone solvation patterns and backbone geometry. We group helices by length and compute amino acid composition and proportional solvation at every helical position including middle residues. We observe a distinct periodic backbone solvation pattern in α-helices, suggesting that, as a result of optimizing for solvation, most helices have a very specific orientation and position specific residue preferences.

For solvation pattern in α-helix see Figures 3.20-3.22. For statistics of hydrated helices at the termini see Figure 3.23. For solvation of full helices of various lengths see Figures 3.24-3.27.

We also observe that at the least solvated helical positions, such as Residue 5, which is the first residue in a helix interior, (also Residue 8, Residue 9 etc. for longer helices) amino acids Ala, Leu, Ile, Val are overrepresented. Such pattern is conserved irrespective of helix length. This effect could also be seen in the amino acid compositions of dipeptides with no water contact, see Figure 3.12 (c) for details, where dipeptides combinations consisting of these four residues types are the most common. We found two recent studies that mention this particular set of amino acids. In one study the effects of side chains in helix nucleation versus helix propagation were investigated, and Ala, Leu, Ile, Val were found to have the highest rate in helix nucleation, see [52] for more details. The second study is the main subject of the next subsection.

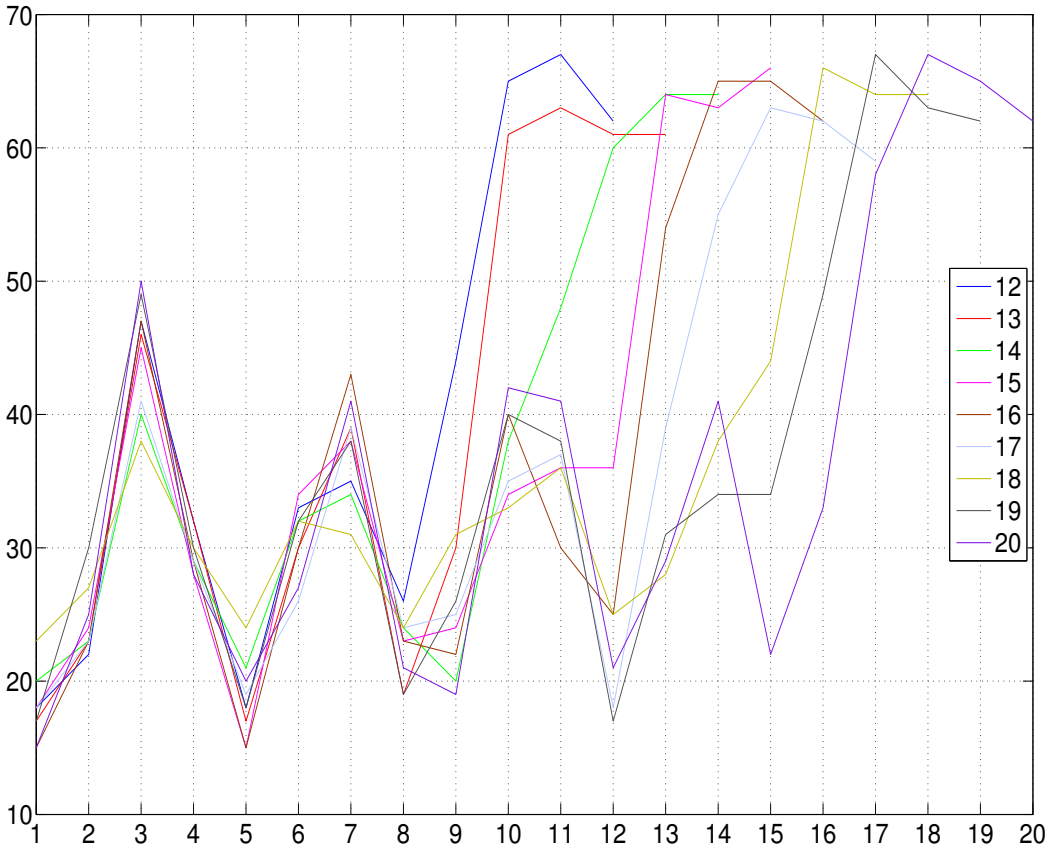


Figure 3.20: Backbone solvation pattern in α-helices of lengths $12, 13, ..., 20$ in high resolution protein data set. (X-axis) - position in the helix, i.e. 1-Ncap, 2-N1, 3-N2 etc., last-Ccap (note, that last position varies in the plot depending on the helix length), (Y-axis) percentage of solvated residues at this position in the helix. Here we observe a distinct periodic pattern. This pattern also implies that most helices assume specific orientation.

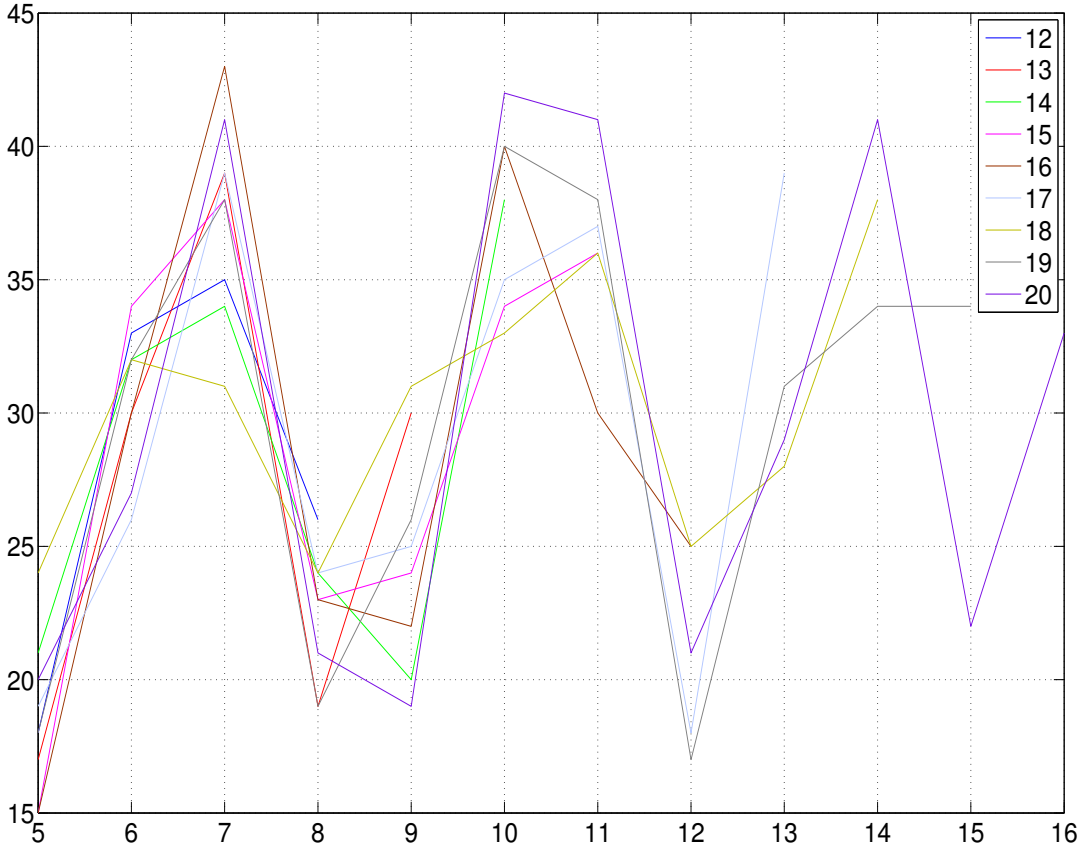


Figure 3.21: Backbone solvation pattern in α-helix for middle residues. (X-axis) - position in the middle of a helix, i.e. 1-N4, 2-N5, etc., last-C4, (Y-axis) percentage of solvated residues at this position. The solvation pattern aligns very well irrespective of helix length.

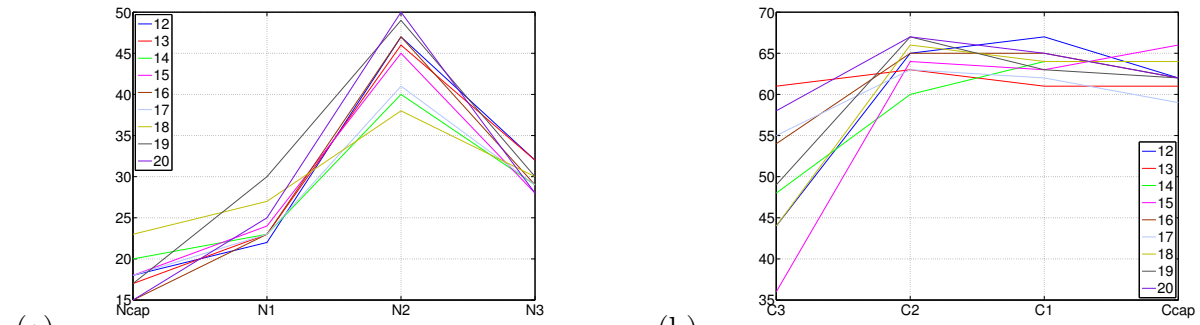


Figure 3.22: Backbone solvation patterns in α-helix at the **(a)** N-terminus, and **(b)** C-terminus. (X-axis) - position at the termini, (Y-axis) percentage of solvated residues at this position. The solvation patern aligns very well at N-terminus. However at C-terminus the percentage of solvated residues vary significantly at position C1. Such effect is due to the C1 residue falling on a different position in the periodic pattern depending on helix length. This further confirms the periodicity of solvation pattern in α-helix.

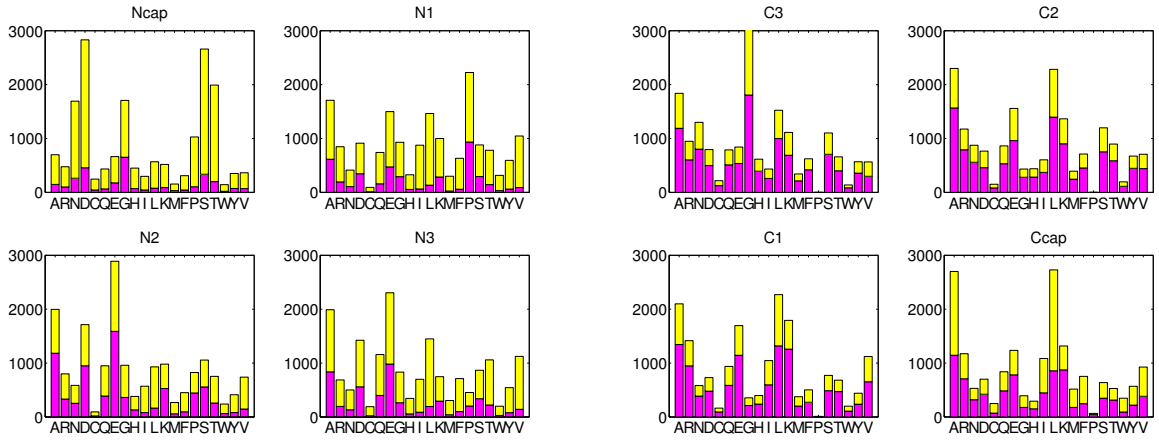


Figure 3.23: Amino acid compositions of backbone solvation at the helix termini. Histogram of solvated (magenta) and unsolvated (yellow) amino acids is plotted for all 20 types of amino acids for each position at the termini.

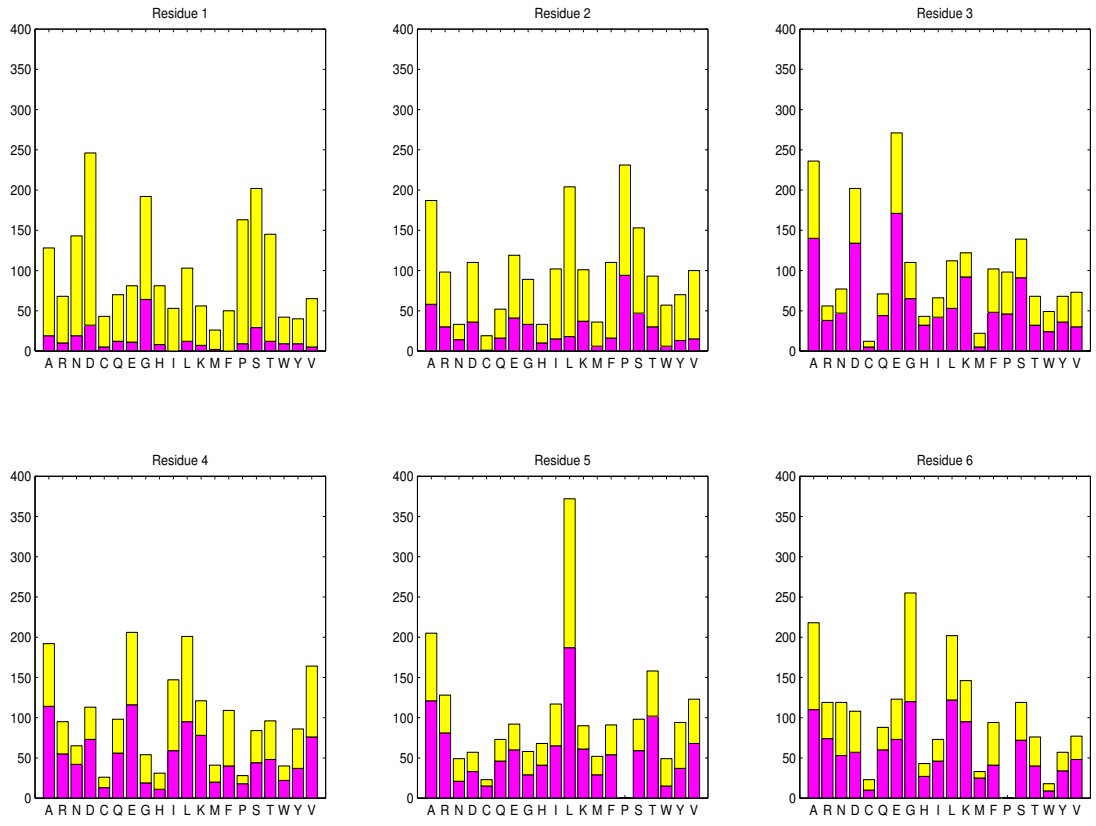


Figure 3.24: Amino acid composition of backbone solvation in helices of length 6. Histogram of solvated (magenta) and unsolvated (yellow) amino acids is plotted for all 20 types of amino acids for each position at the α-helix of length 6.

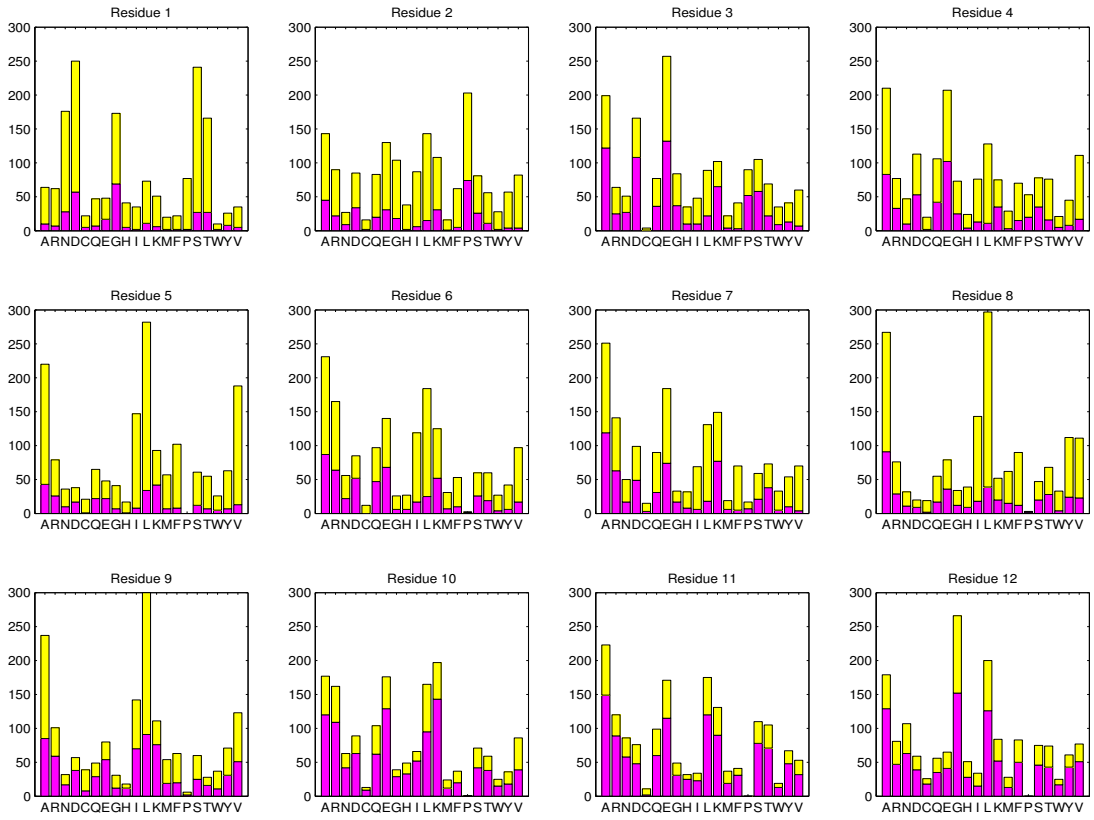


Figure 3.25: Amino acid composition of backbone solvation in helices of length 12. Histogram of solvated (magenta) and unsolvated (yellow) amino acids is plotted for all 20 types of amino acids for each position at the α-helix of length 12.

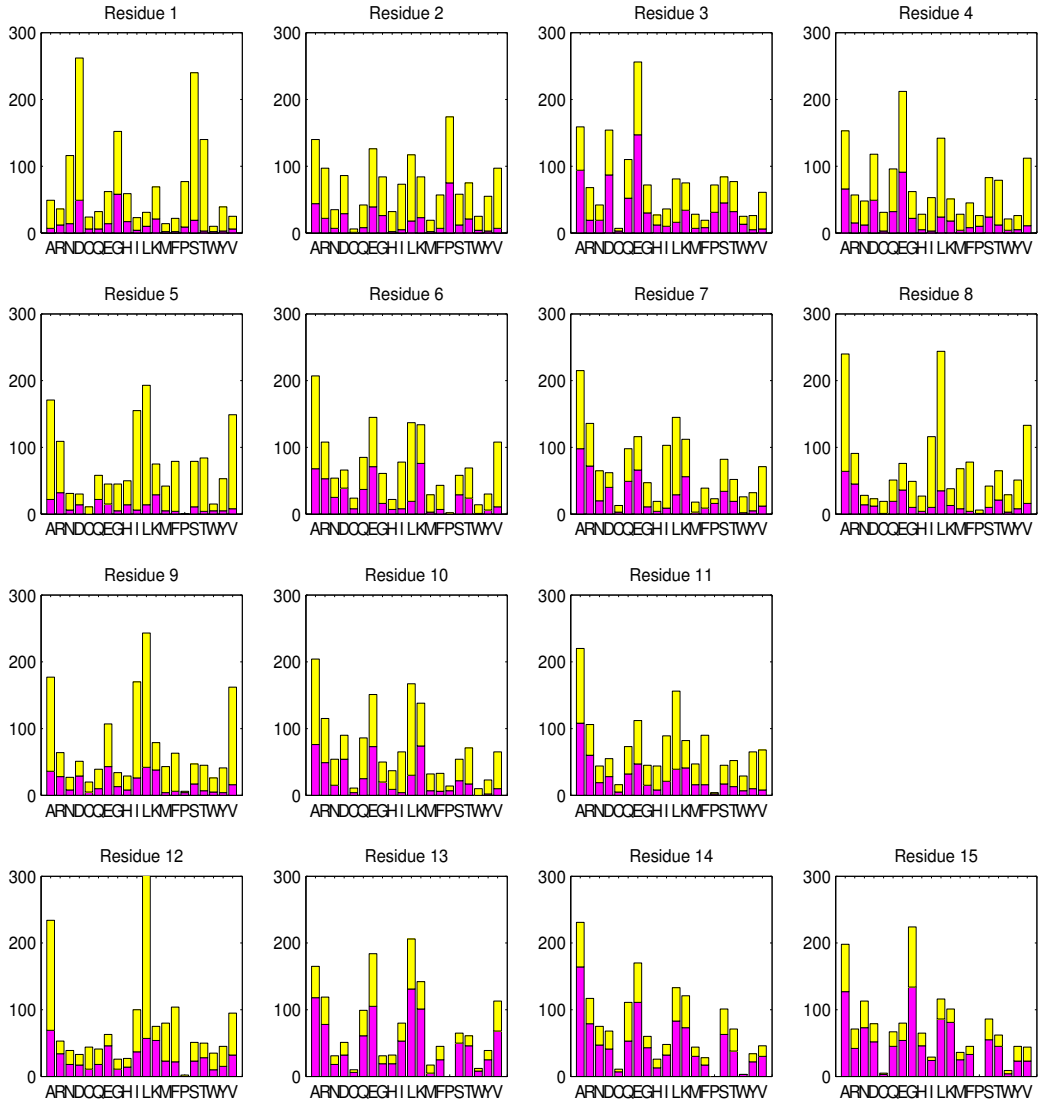


Figure 3.26: Amino acid composition of backbone solvation in helices of length 15. Histogram of solvated (magenta) and unsolvated (yellow) amino acids is plotted for all 20 types of amino acids for each position at the α-helix of length 15.

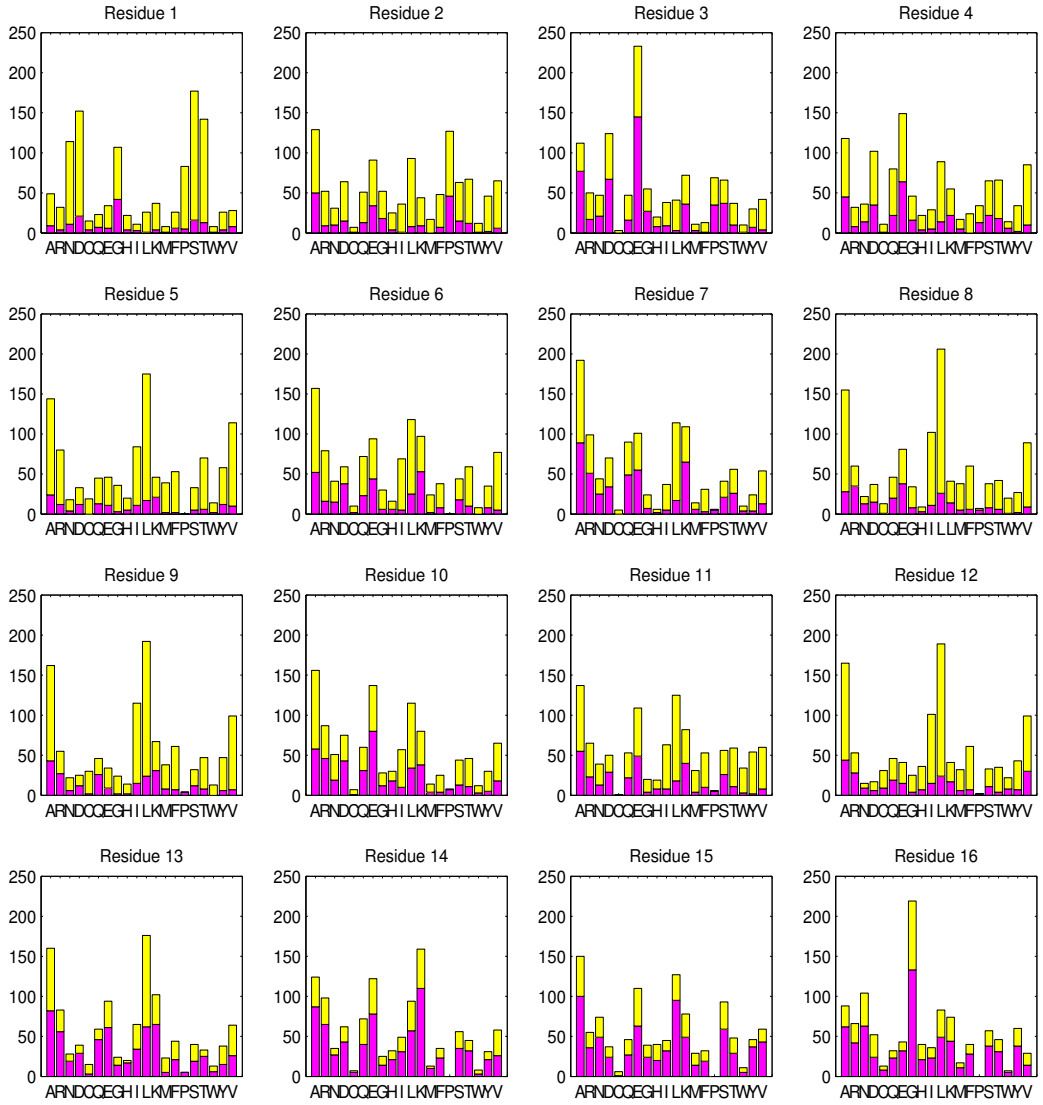


Figure 3.27: Amino acid composition of backbone solvation in helices of length 16. Histogram of solvated (magenta) and unsolvated (yellow) amino acids is plotted for all 20 types of amino acids for each position at the α-helix of length 16.

3.4.2 Solvation signature in chameleon sequences

Chameleon sequences are amino acid sequences that can form different structures in different proteins [45]. Such sequences are relevant to amyloid diseases [18], thus understanding the mechanisms that force these sequences to assume different secondary structures is very important [45, 36]. Here we connect a type of secondary structure with a backbone solvation signature for some chameleon sequences.

We demonstrate this concept on two examples: (a) non-homologous proteins, see Figure 3.28 for details, (b) homologous proteins, see Figure 3.29 for details. It is remarkable that in both examples patterns of backbone carbonyl oxygen interactions with water for α-helix and β-strand are reciprocals of each other.

Another argument that supports the relevance of backbone-water interactions to the resulting secondary structure a particular chameleon sequence is most likely to adopt is that Ala, Leu, Ile, Val are highly overrepresented in chameleon sequences comparing to the amino acid composition across all proteins [45]. As we saw earlier Ala, Leu, Ile, Val amino acids dominate in statistically least solvated positions in α-helix. So, the changes in solvation could possibly make the helix less stable and contribute to formation of a different structure, such as insoluble fibrils via β-sheet formation and aggregation, or even aggregation of helices into insoluble fibrils [36].

The above arguments suggest that more systematic analysis of this correlation is needed. The main challenge is that currently the database of chameleon sequences is rather small, where many structures are only available at low resolution, and/or with no water information for both or one of the structures. We hope that as more instances of chameleon sequences are identified, and more structures are solved with better quality these questions can be investigated further using data mining techniques. We also believe that molecular dynamic simulations applied to chameleon sequences interacting with water could be a very useful tool in this direction of research.

Figure 3.28: Here two fragments of protein sequences for structures 3S30 chain A and 2Q0Y chain A are pictured. The identical fragment, chameleon sequence **iddlelvc**, is highlighted in magenta in each structure. Amino acid is denoted by a capital letter if a backbone carbonyl oxygen at that position has a primary interaction with water molecule. If such interaction occurs within the chameleon sequence the amino acid is also highlighted. For easier comparison we also provide a binary representation of a solvation pattern. We see that the solvation pattern in the case of a β-strand conformation is almost reciprocal to the case when the chameleon sequence is a part of an α-helix.

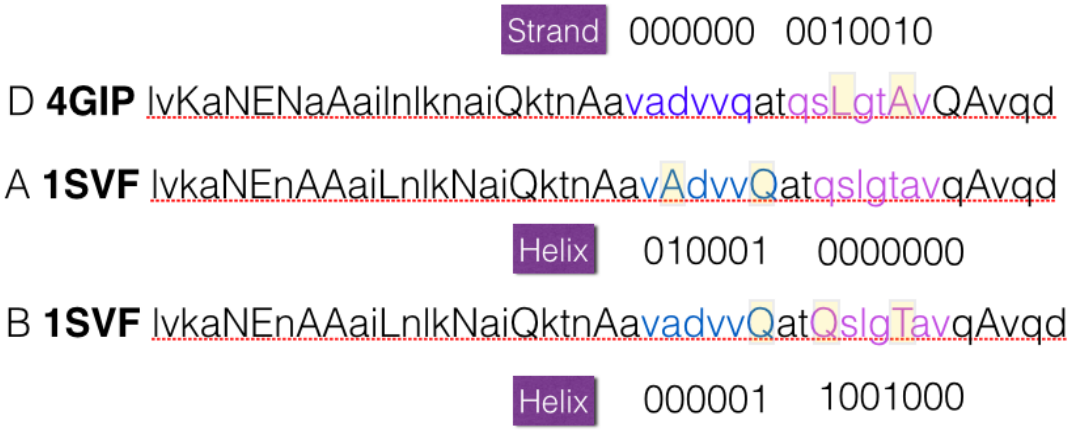


Figure 3.29: Here three fragments of protein sequences for structures 4GIP chain D, 1SVF chains A and B are pictured. Two identical fragments, chameleon sequences **vadvvq** and **qslgtav**, are highlighted in blue and magenta respectively in each structure. Amino acid is denoted by a capital letter if a backbone carbonyl oxygen at that position has a primary interaction with water molecule. If such interaction occurs within the chameleon sequence the amino acid is also highlighted. For easier comparison we also provide a binary representation of a solvation pattern. We see that the solvation pattern in the case of a 0 and -1 β-strands conformation is almost reciprocal to the other two cases when both fragments are a part of an α-helix.

3.5 Summary and conclusions

In this chapter we studied how backbone geometry of α-helix changes due to interactions with water.

First, we identified a strong dependence between water proximity to the backbone carbonyl oxygen O_i and an increase in magnitude of $|\varphi_{i+1}|$ and $d_{O_iO_{i+1}}$, as shown in Figures 3.8–3.11. As expected, our results confirmed that the formation of a canonical peak [31] $(\varphi, \psi) = (-63^\circ, 43^\circ)$ in experimental α-helices is influenced by the need to maximize the distance between neighboring main-chain oxygens $d_{O_iO_{i+1}}$ in order to allow a better interaction with water. However, as seen in Table 3.2, and in Figure 3.6, in high resolution protein dataset the total peak at $\varphi = -63^\circ$ results as an average of two different peaks at -65° and -62°, corresponding to the case when main-chain oxygen has a primary interaction with water and other cases respectively. We were able to validate our findings by comparing to the existing data found in [49] obtained via density functional theory calculations.

Moreover, we found that the rate with which the distance $d_{O_iO_{i+1}}$ grows as a function of φ_{i+1} strongly depends on the kind of amino acid at the $(i+1)$th position in $(i, i+1)$ dipeptide. In particular, when backbone carbonyl oxygen at the ith position has a primary contact with water the least squares fit lines for $(\varphi_{i+1}, d_{O_iO_{i+1}})$ corresponding to different kinds of amino acids at $(i+1)$th position are well ordered, as shown in Figure 3.14. This order provides a nice classification of amino acids connecting secondary structure propensities with geometrical possibilities for backbone solvation, as demonstrated in Table 3.3.

We also discovered a distinct periodic backbone solvation pattern in α-helices. This pattern is position dependent, see Figures 3.20–3.22, and shows position specific amino acid preferences, see Figures 3.23–3.27. It is interesting to note, that in most helices, amino acid composition at Residue 5 (and other similar least solvated positions in longer helices such as Residue 8, Residue 9, Residue 12 etc.) show preference for Ala, Leu, Ile, and Val. This group of amino acids appeared in recent research as having the highest propensity for helix nucleation [52], and as statistically overrepresented in the database of chameleon

sequences [45]. Both topics are important for helix formation and stability and need further investigation in the context of solvation in water.

Finally, in Figures 3.28 and 3.29 we gave two examples of chameleon sequences having reciprocal backbone solvation patterns while adopting different types of secondary structure. This important connection between secondary structure formation and backbone carbonyl oxygen solvation pattern in identical protein sequences requires further analysis which we plan to do in the future.

CHAPTER 4

CONCLUSIONS

Water is often referred to as a "lubricant of life" because of its unique ability to quickly form hydrogen bonds with peptides and other water molecules. In this work we focused our attention on the role of solvation in hydrogen bond geometry of protein helices. On one hand the regular structure of protein helices makes it a very appealing object for mathematical modeling and optimization, on the other hand, the recent growth of the database of protein structures and improvement in the quality of data, in particular, the availability of water location information, makes "data science" techniques very useful as well. Here we were able to show how combining these methods could lead to the discovery of important results in protein geometry that are relevant to the understanding of some challenging diseases.

First, we used computer algebra systems to study the geometry of hydrogen bonds in α-helix. We obtained optimization data, suggesting that the most common α-helices are optimized for main-chain main-chain hydrogen bonds linearity, and for maximizing backbone solvation in water. We used these results to perform a further investigation on experimental protein structures.

As a next step, we explored the effects of solvation on α-helices by analyzing high quality protein data set. We observed a solvation signature in the experimental α-helices. It altered helical backbone geometry by requiring larger distances between neighboring backbone carbonyl oxygens. We found that solvation effects on helical backbone are universal irrespective whether a helix is located at the surface or inside the protein. Also, we obtained a classification of amino acids connecting secondary structure propensities with geometrical possibilities for backbone solvation, and showed how backbone solvation causes the preferred orientation and position specific amino acid preferences in α-helices. We applied our findings to a current research problem in connection with chameleon sequences relevant to amyloid diseases.

Finally, some research directions and possible applications of our results that we plan

to explore in the future include: water location prediction in experimental protein helices based on backbone geometry, backbone flexibility model improvement that would incorporate solvation in protein design algorithms, deeper understanding of the role of backbone solvation in chameleon sequences, and application of similar data analysis to membrane and ice-structuring proteins.

BIBLIOGRAPHY

[1] MAPLE 16. *Maplesoft, a division of Waterloo Maple Inc.* Waterloo, Ontario.

[2] M. Alahuhta and R.K. Wierenga et al. Atomic resolution crystallography of a complex of triosephosphate isomerase with a reaction-intermediate analog: new insight in the proton transfer reaction mechanism. *Proteins*, 78(8):1878–1888, 2010.

[3] P.J. Artymiuk and C.C. Blake. Refinement of human lysozyme at 1.5 Å resolution analysis of non-bonded and hydrogen-bond interactions. *Journal of Molecular Biology*, 152(4):737–762, 1981.

[4] W.T. Astbury, van Iterson, R.D. Preston, E.G. Cox, and J.M. Preston. General discussion. *Transactions of the Faraday Society*, 29:71–77, 1933.

[5] R. Aurora and G.D. Rose. Helix capping. *Protein Science*, 7(1):21–38, 1998.

[6] F. Avbelj. Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins. *Journal of Molecular Biology*, 300(5):1335–1359, 2000.

[7] F. Avbelj and R.L. Baldwin. Role of backbone solvation in determining thermodynamic β propensities of the amino acids. *Proceedings of the National Academy of Sciences, USA*, 99(3):1309–1313, 2002.

[8] E.N. Baker and R.E. Hubbard. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44(2):97–179, 1984.

[9] R.L. Baldwin. In search of the energetic role of peptide hydrogen bonds. *Journal of Biological Chemistry*, 278(20):17581–17588, 2003.

[10] R.L. Baldwin. Energetics of protein folding. *Journal of Molecular Biology*, 371(2):283–301, 2007.

[11] D.J. Barlow and J.M. Thornton. Helix geometry in proteins. *Journal of Molecular Biology*, 201(3):601–619, 1988.

[12] E.R. Batista, S.S. Xantheas, and H. Jonsson. Multipole moments of water molecules in clusters and ice Ih from first principles calculations. *Journal of Chemical Physics*, 111(13):6011–6015, 1999.

[13] 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 Research*, 28(1):235–242, 2000.

[14] T. Beuming, Y. Che, R. Abel, B. Kim, V. Shanmugasundaram, and W. Sherman. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins*, 80(3):871–883, 2012.

[15] T. Blundell, D. Barlow, N. Borkakoti, and J. Thornton. Solvent-induced distortions and the curvature of α-helices. *Nature*, 306(5940):281–283, 1983.

[16] A.D. Buckingham and P.W. Fowler. Do electrostatic interactions predict structures of van der Waals molecules? *Journal of Chemical Physics*, 79(12):6426–6428, 1983.

[17] A.D. Buckingham and P.W. Fowler. A model for the geometries of van der Waals complexes. *Canadian Journal of Chemistry*, 63:2018–2025, 1985.

[18] F. Chiti and C.M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75:333–336, 2006.

[19] H. Choi, H. Kang, and H. Park. New angle-dependent potential energy function for backbone-backbone hydrogen bond in protein-protein interactions. *Journal of Computational Chemistry*, 31(5):897–903, 2010.

[20] P.R. Connelly, R.A. Aldape, F.J. Bruzzese, S.P. Chambers, M.J. Fitzgibbon, M.A. Fleming, S. Itoh, D.J. Livingston, M.A. Navia, and J.A. Thomson. Enthalpy of hydrogen bond formation in a protein-ligand binding reaction. *Proceedings of the National Academy of Sciences, USA*, 91(5):1964–1968, 1994.

[21] B.I. Dahiyat, D.B. Gordon, and S.L. Mayo. Automated design of the surface positions of protein helices. *Protein science*, 6(6):1333–1337, 1997.

[22] C. Deremble and R. Lavery. Macromolecular recognition. *Current Opinion in Structural Biology*, 15(2):171–175, 2005.

[23] G.R. Desiraju. A bond by any other name. *Angewandte Chemie International Edition*, 50:52–59, 2011.

[24] R. Dickerson and I. Geis. *The structure and action of proteins*. Harper and Row, Publishers / New York Evanston London, 3rd edition, 1969.

[25] D. Dix. IMIMOL: A computer program for molecular geometry specification and computation. http://www.math.sc.edu/ dix/imimol.pdf, 2004. [Online; accessed 24-Feb-2014].

[26] D. Eisenberg. The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences, USA*, 100(20):11207–11210, 2003.

[27] H. Eyring. The resultant electric moment of complex molecules. *Physical Review*, 39(4):746–748, 1932.

[28] J.S. Franzen and R.E. Stephens. The effect of a dipolar solvent system on interamide hydrogen bonds. *Biochemistry*, 2(6):1321–1327, 1963.

[29] K. Fujiwara, H. Toda, and M. Ikeguchi. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology*, 12:18, 2012.

[30] B.K. Ho, A. Thomas, and R. Brasseur. Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. *Protein Science*, 12(11):2508–2522, 2003.

[31] S.A. Hollingsworth and P.A. Karplus. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *BioMolecular Concepts*, 1(3-4):271–283, 2010.

[32] M.L. Huggins. The structure of fibrous proteins. *Chemical Reviews*, 32(2):195–218, 1943.

[33] G.A. Jeffrey. *An introduction to hydrogen bonds*. Oxford Science Publications, Clarendon Press, Oxford, 1997.

[34] G.A. Jeffrey. Hydrogen-bonding: An update. *Crystallography Reviews*, 9(2-3):135–176, 2003.

[35] W. Kauzmann. Some factors in the interpretation of protein denaturation. In *Advances in Protein Chemistry*, volume 14, pages 1–63. Academic Press, 1959.

[36] B. Kim, T.D. Do, E.Y. Hayden, D.B. Teplow, M.T. Bowers, and J.E. Shea. Aggregation of chameleon peptides: Implications of α-helicity in fibril formation. *The Journal of Physical Chemistry B*, 120(26):5874–5883, 2016.

[37] S. Kim and T.A. Gross. Uniformity, ideality, and hydrogen bonds in transmembrane α-helices. *Biophysical Journal*, 83(4):2084–2095, 2002.

[38] I.M. Klotz. Solvent water and protein behavior: View through a retroscope. *Protein Science*, 2(11):1992–1999, 1993.

[39] I.M. Klotz and J.S. Franzen. Hydrogen bonds between model peptide groups in solution. *Journal of the American Chemical Society*, 84(18):3461–3466, 1962.

[40] P.A. Kollman and L.C. Allen. The theory of the hydrogen bond. *Chemical Reviews*, 72(3):283–303, 1972.

[41] J. Kroon, J.A. Kanters, J.G.C.M. van Duijneveldt-van De Rijdt, F.B. van Duijneveldt, and J.A. Vliegenthart. O-H $\cdot\cdot$ O hydrogen bonds in molecular crystals a statistical and quantum-chemical analysis. *Journal of Molecular Structure*, 24(1):109–129, 1975.

[42] W.M. Latimer and W.H. Rodebush. Polarity and ionization from the standpoint of the Lewis theory of valence. *Journal of the American Chemical Society*, 42(7):1419–1433, 1920.

[43] T.L. Lau, C. Kim, M.H. Ginsberg, and T.S. Ulmer. The structure of the integrin αIIbβ3 transmembrane complex explains integrin transmembrane signalling. *The EMBO Journal*, 28(9):1351–1361, 2009.

[44] Y. Levy and J.N. Onuchic. Water and proteins: A love-hate relationship. *Proceedings of the National Academy of Sciences, USA*, 101(10):3325–3326, 2004.

[45] W. Li, L.N. Kinch, P.A. Karplus, and N.V Grishin. ChSeq: A database of chameleon sequences. *Protein Science*, 24(7):1075–1086, 2015.

[46] B.W. Low and H.J. Grenville-Wells. Generalized mathematical relationships for polypeptide chain helices: The coordinates of the II helix. *Proceedings of the National Academy of Sciences, USA*, 39(8):785–801, 1953.

[47] P. Luo and R.L. Baldwin. Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. *Proceedings of the National Academy of Sciences, USA*, 96(9):4930–4935, 1999.

[48] N. Mandel, G. Mandel, B.L. Trus, J. Rosenberg, G. Carlson, and R.E. Dickerson. Tuna cytochrome c at 2.0 Å resolution. III. Coordinate optimization and comparison of structures. *The Journal of Biological Chemistry*, 252(13):4619–4636, 1977.

[49] M. Marianski and J.J. Dannenberg. Aqueous solvation of polyalanine α-helices with specific water molecules and with the CPCM and SM5.2 aqueous continuum models using density functional theory. *The Journal of Physical Chemistry B*, 116(4):1437–1445, 2012.

[50] I.H. McColl, E.W. Blanch, L. Hecht, and L.D. Barron. A study of α-helix hydration in polypeptides, proteins, and viruses using vibrational Raman optical activity. *Journal of the American Chemical Society*, 126(26):8181–8188, 2004.

[51] K.H. Meyer and L.E.R. Picken. *Natural and Synthetic High Polymers: A Textbook and Reference Book for Chemists and Biologists*, volume IV of *High Polymers*. Interscience Publishers Inc., New York and London, 1942.

[52] S.E. Miller, A.M. Watkins, N.R. Kallenbach, and P.S. Arora. Effects of side chain in helix nucleation differ from helix propagation. *Proceedings of the National Academy of Sciences, USA*, 111(18):6636–6641, 2014.

[53] C. Millot and A.J. Stone. Towards an accurate intermolecular potential for water. *Molecular Physics*, 77(3):439–462, 1992.

[54] K.T. No, O.Y. Kwon, S.Y. Kim, M.S. Jhon, and H.A. Scheraga. A simple functional representation of angular-dependent hydrogen-bonded systems. 1. amide, carboxylic acid, and amide-carboxylic acid pairs. *The Journal of Physical Chemistry*, 99(11):3478–3486, 1995.

[55] I. Olovsson and P.G. Joensson. X-ray and neutron diffraction studies of hydrogen bonded systems. In G. Zundel P. Schurster and C. Sandorfy, editors, *The Hydrogen Bond*, volume 7, pages 393–456. North-Holland, Amsterdam, The Netherlands, 1976.

[56] L. Pauling and R.B. Corey. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences, USA*, 37(5):235–240, 1951.

[57] J.R. Quine. Helix parameters and protein structure using quaternions. *Journal of Molecular Structure: THEOCHEM*, 460:53–66, 1999.

[58] C. Ramakrishnan. Ramachandran and his map. *Resonance*, 6:48–56, 2001.

[59] C. Ramakrishnan and N. Prasad. Study of hydrogen bonds in amino acids and peptides. *Int. J. Prot. Research*, III(1-4):209–231, 1971.

[60] C. Ramakrishnan and G.N. Ramachandran. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*, 5(6):909–933, 1965.

[61] R. Renthal. Buried water molecules in helical transmembrane proteins. *Protein Science*, 17(2):293–298, 2008.

[62] M. Sandaralingam and Y.C. Shekharudu. Water-inserted alpha-helical segments implicate reverse turns as folding intermediates. *Science*, 244(4910):1333–1337, 1989.

[63] V. Sasisekharan. In N. Ramachandran, editor, *Collagen*, page 39. John Wiley and Sons, New York, 1962.

[64] T. Simanouchi and S. Mizushima. On the helical configuration of a polymer chain. *The Journal of Chemical Physics*, 23(4):707–711, 1955.

[65] A.J. Stone. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters*, 83(2):233–239, 1981.

[66] A.J. Stone and M. Alderton. Distributed multipole analysis: methods and applications. *Molecular Physics*, 56(5):1047–1064, 1985.

[67] R. Taylor and O. Kennard. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research*, 17(9):320–326, 1984.

[68] N. Thanki, J.M. Thornton, and J.M. Goodfellow. Distributions of water around amino acid residues in proteins. *Journal of Molecular Biology*, 202(3):637–657, 1988.

[69] G. Wang and R.L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[70] D. Xu, C.-J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering*, 10(9):999–1012, 1997.

[71] S. Yamamoto, T. Furukawa, P. Bour, and Y. Ozaki. Solvated states of poly-L-alanine α-helix explored by Raman optical activity. *The Journal of Physical Chemistry A*, 118(20):3655–3662, 2014.

[72] S.Q. Zhang, D.W. Kulp, C.A. Schramm, M. Mravic, I. Samish, and W.F. DeGrado. The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions. *Structure*, 23(3):527–541, 2015.