THE UNIVERSITY OF CHICAGO

SOLVATION SIGNATURE IN HYDROGEN BOND GEOMETRY OF PROTEIN HELICES - DRAFT

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, 2016
Copyright © 2016 by Tatiana Orlova
All Rights Reserved
Epigraph Text
CONTENTS

LIST OF FIGURES ......................................................... vii
LIST OF TABLES ......................................................... ix
ACKNOWLEDGMENTS ...................................................... x
ABSTRACT ................................................................. xi

1 INTRODUCTION ........................................................ 1

2 HYDROGEN BOND GEOMETRY IN α-HELICES .............................. 2
  2.1 Hydrogen bonds ................................................... 2
    2.1.1 Hydrogen bond models .................................... 2
    2.1.2 Distance and angle optimization ............................ 4
    2.1.3 Solvation optimization .................................... 5
    2.1.4 Solvation signature ........................................ 10
  2.2 Helical structure ................................................ 11
  2.3 Protein backbone local regular structures .......................... 13
    2.3.1 History of the topic ....................................... 13
    2.3.2 Computations .............................................. 17
    2.3.3 Helical conclusions ....................................... 23
  2.4 Feasible protein backbone local regular structures - Allowed (ϕ,ψ) orientations 23
    2.4.1 Motivation .................................................. 24
    2.4.2 Allowed (ϕ,ψ) regions .................................... 25
    2.4.3 Feasible conclusions ....................................... 26
  2.5 Stable protein backbone structures and hydrogen bond geometry .......... 26
    2.5.1 A hydrogen bond and its geometrical parameters .......... 28
    2.5.2 Most common α- and 310-helices (significance of canonical peaks) .. 29
    2.5.3 Computational models and transformation matrices for α- and 310-helix 30
    2.5.4 Optimizing with respect to linearity of a hydrogen bond .......... 34
    2.5.5 Minimizing the length of a hydrogen bond .................. 35
    2.5.6 Computing other geometrical parameters of a hydrogen bond .......... 36
    2.5.7 Optimization results and balance between two criteria .......... 40
    2.5.8 Hydrogen bond geometry in small molecules .................. 42
    2.5.9 Relationship between geometrical parameters .................. 42
    2.5.10 Other geometrical restrictions ............................ 44
    2.5.11 Hydrogen bond conclusions ................................ 44
  2.6 Overall conclusions ............................................. 47

3 SOLVATION SIGNATURE IN α-HELICES .................................. 53
  3.1 Proteins interact with water .................................... 53
  3.2 Measuring backbone solvation in α-helices .......................... 55
    3.2.1 Helix nomenclature ....................................... 55
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.2 Computing geometrical parameters</td>
<td>55</td>
</tr>
<tr>
<td>3.2.3 Dataset</td>
<td>55</td>
</tr>
<tr>
<td>3.3 Solvation effects on backbone geometry in (\alpha)-helices</td>
<td>58</td>
</tr>
<tr>
<td>3.3.1 Results for high resolution protein dataset.</td>
<td>58</td>
</tr>
<tr>
<td>3.3.2 Comparison of our results to the results from [50].</td>
<td>66</td>
</tr>
<tr>
<td>3.4 Solvation patterns in (\alpha)-helices</td>
<td>66</td>
</tr>
<tr>
<td>3.4.1 Solvation dependent amino acid composition in (\alpha)-helices</td>
<td>66</td>
</tr>
<tr>
<td>3.5 Conclusion and further plans</td>
<td>79</td>
</tr>
</tbody>
</table>

REFERENCES                                                                  81
LIST OF FIGURES

2.1 Contours of $|H_5 \cdots O_1|$ as a function of $(\varphi, \psi)$ .......................... 3
2.2 Contours of $\nu = N_5 - H_5 \cdots O_1$ as a function of $(\varphi, \psi)$ .......................... 3
2.3 H-O distance and the NOH angle in $\alpha$-helices ..................................................... 3
2.4 Hydrogen bond geometry .......................................................... 5
2.5 Contours of $d_{O_iO_{i+1}}$ as a function of $(\varphi_{i+1}, \psi_{i+1})$ .......................... 7
2.6 Distribution of distances $O^{i-1} \cdots C_i$ and $O^{i-1} \cdots C_{\beta i}$ parametrized by $\varphi$ .......................... 9
2.7 Ideal geometry of the peptide plane according to Pauling .......................... 12
2.8 Helical parameters: rotation angle, radius, and rise ..................................................... 17
2.9 $(\varphi, \psi)$ allowed regions ..................................................... 25
2.10 $(\varphi, \psi)$ pairs corresponding to $\alpha$-helix ..................................................... 36
2.11 $(\varphi, \psi)$ pairs for $3\alpha$-helix ..................................................... 36
2.12 $(\varphi, \psi)$ pairs corresponding to extremal cases ..................................................... 36
2.13 Relationship diagram between $H_5 \cdots O_1$, $C_1 - \hat{O}_1 \cdots H_5$, and $C_1 - \hat{O}_1 \cdots H_5$ .......................... 43
2.14 Optimization results in connection with the canonical peak for $\alpha$-helix .......................... 45
2.15 Optimization results in connection with the canonical peak for $3\alpha$-helix .......................... 46
2.16 Contours for $C_1 - \hat{O}_1 \cdots H_5$ parametrized by $(\varphi, \psi)$ .......................... 47
2.17 Contours for $(N_5 - \hat{H}_5 \cdots O_1 - C_1 - \hat{O}_1 \cdots H_5)$ parametrized by $(\varphi, \psi)$ .......................... 47
2.18 Contours for different angles as a function of $(\varphi, \psi)$ for $\alpha$-helix .......................... 47
2.19 Contours for $N_4 - \hat{H}_4 \cdots O_1$ parametrized by $(\varphi, \psi)$ .......................... 48
2.20 Contours for $C_1 - \hat{O}_1 \cdots H_4$ parametrized by $(\varphi, \psi)$ .......................... 48
2.21 Contours for different angles as a function of $(\varphi, \psi)$ for $3\alpha$-helix .......................... 48
2.22 Contours for $(N_4 - \hat{H}_4 \cdots O_1 - C_1 - \hat{O}_1 \cdots H_4)$ parametrized by $(\varphi, \psi)$ .......................... 48
2.23 Contours for $|H_4 \cdots O_1|$ parametrized by $\varphi, \psi$ ..................................................... 48
2.24 Contours for angle differences and distances as function of $\varphi, \psi$ for $3\alpha$ helices .......................... 48
2.25 Contours for $N_1 - \hat{H}_1 \cdots O_5$ and $|H_5 \cdots O_1|$ parametrized by $(\varphi, \psi)$ .......................... 49
2.26 Contours for $N_1 - \hat{H}_1 \cdots O_4$ and $|H_4 \cdots O_1|$ parametrized by $(\varphi, \psi)$ .......................... 49
2.27 Contours for $N_1 - \hat{H}_1 \cdots O_k$, where $k = 3, 4$ .......................... 49
2.28 Contours for $C_1 - \hat{O}_1 \cdots H_5$ and $|H_5 \cdots O_1|$ parametrized by $(\varphi, \psi)$ .......................... 50
2.29 Contours for $C_1 - \hat{O}_1 \cdots H_4$ and $|H_4 \cdots O_1|$ parametrized by $(\varphi, \psi)$ .......................... 50
2.30 Contours for $C_1 - \hat{O}_1 \cdots H_k$ and $|H_k \cdots O_1|$, where $k = 4, 5$ .......................... 50

3.1 Backbone hydration .......................................................... 54
3.2 Geometrical parameters of helical backbone hydration ..................................................... 56
3.3 Distances $d_{O_iC_{i+1}}$ and $d_{O_iO_{i+1}}$ parametrized by $\varphi_{i+1}, \psi_{i+1}$ ..................................................... 57
3.4 Histogram of distances $d_i$ ..................................................... 59
3.5 Histogram of distances $d_i$ with unique water contacts only ..................................................... 60
3.6 Histograms of $\varphi_{i+1}$ and $d_{O_iO_{i+1}}$ ..................................................... 61
3.7 Histograms of $\varphi_{i+1}$ and $d_{O_iO_{i+1}}$ with unique water contacts only ..................................................... 61
3.8 Scatter plots $(d_i, \varphi_{i+1})$ and $(d_i, d_{O_iO_{i+1}})$ ..................................................... 62
3.9 Histogram and heatmap for $(d_i, \varphi_{i+1})$ ..................................................... 62
3.10 Histogram and heatmap for $(d_i, d_{O_iO_{i+1}})$ ..................................................... 63
3.11 Histogram and heatmap for $(\varphi_{i+1}, d_{O_iO_{i+1}})$ ........................................... 63
3.12 Least squares fit lines to $(\varphi_{i+1}, d_{O_iO_{i+1}})$ ......................................................... 64
3.13 Amino acid composition for middle residues in $\alpha$-helices ........................................... 64
3.14 Amino acid propensities .................................................................................................. 65
3.15 Least squares fit lines to $(\varphi_{i+1}, d_{O_iO_{i+1}})$ organized by the residue type and position 66
3.16 Least squares fit lines to $(\varphi_{i+1}, d_{O_iO_{i+1}})$ organized by the residue type and position 67
3.17 Least squares fit lines to $(\varphi_{i+1}, d_{O_iO_{i+1}})$ organized by the residue type and position 67
3.18 Backbone solvation pattern in $\alpha$-helix ............................................................................. 69
3.19 Backbone solvation pattern in $\alpha$-helix for middle residues ................................................ 70
3.20 Backbone solvation patterns in $\alpha$-helix at the termini ........................................................ 70
3.21 Backbone solvation at the helix termini .................................................................................. 71
3.22 Backbone solvation in helices of length 6 .............................................................................. 71
3.23 Backbone solvation in helices of length 12 ............................................................................. 72
3.24 Backbone solvation in helices of length 15 ............................................................................. 73
3.25 Backbone solvation in helices of length 16 ............................................................................. 74
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Minimum contact distances between atoms</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>$\phi, \psi$ ranges allowed by the contact distances between atoms</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>$\phi, \psi$ values of globular protein 2VXN</td>
<td>29</td>
</tr>
<tr>
<td>2.4</td>
<td>$\phi, \psi$ values of transmembrane protein 2K9J</td>
<td>31</td>
</tr>
<tr>
<td>2.5</td>
<td>Optimization results in comparison with data for most common helices</td>
<td>41</td>
</tr>
<tr>
<td>2.6</td>
<td>Helical parameters of optimized helices in comparison to most common helices</td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>High resolution protein dataset statistics</td>
<td>57</td>
</tr>
<tr>
<td>3.2</td>
<td>Results for high resolution protein dataset for middle residues.</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Relevant results from [50]</td>
<td>66</td>
</tr>
<tr>
<td>3.4</td>
<td>Helix length statistics for High Resolution Protein Dataset.</td>
<td>76</td>
</tr>
<tr>
<td>3.5</td>
<td>Helix solvation patterns for middle residues in High Resolution Protein Dataset. &lt;br&gt;Part1.</td>
<td>77</td>
</tr>
<tr>
<td>3.6</td>
<td>Helix solvation patterns for middle residues in High Resolution Protein Dataset. &lt;br&gt;Part2.</td>
<td>78</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS
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 influence the geometry of helices and help 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, $\alpha$- and $\beta_{10}$-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 $\alpha$- and $\beta_{10}$-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 $(\varphi, \psi)$ 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 $\varphi$ angles and larger distances between neighboring main-chain carbonyl oxygens. Moreover, we observe a distinct periodic backbone solvation pattern in $\alpha$-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
CHAPTER 2
HYDROGEN BOND GEOMETRY IN α-HELICES

2.1 Hydrogen bonds

The concept of the hydrogen bond was established by 1920 [45], and possibly earlier [41, 15]. Hydrogen bonds are the most important bond 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, 44], they are not yet fully understood and are still actively studied [37, 38]. In particular, it is difficult to model hydrogen bonds quantitatively with any generality and precision.

2.1.1 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” [43]. 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” [43]. 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 [12, 13]. One model of hydrogen bonds in water incorporates polarization terms [11, 52]. 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
Figure 2.1: Contours for different values of the distance $|H_5 \cdots O_1|$ as a function of $\varphi, \psi$ are plotted for $\alpha$-helices. The dot in the smallest contour indicates the $(\varphi, \psi)$ pair that minimizes the H-O distance in $\alpha$-helices.

Figure 2.2: Contours for different values of the angle $\nu = N_5 \overrightarrow{H_5} \cdots O_1$ as a function of $(\varphi, \psi)$ are plotted for $\alpha$-helices. The dot in the smallest contour indicates the $(\varphi, \psi)$ pair that minimizes the NOH angle defined subsequently in Figure 2.4.

Figure 2.3: The dots in the smallest contours in the respective figures indicate the $(\varphi, \psi)$ pair that minimizes the (a) H-O distance and (b) the NOH angle in $\alpha$-helices. The remaining dot, to the upper-left, indicates the observed $(\varphi, \psi) = (-63^\circ, -43^\circ)$ pair that occurs most commonly in $\alpha$-helices. Solid black lines indicated the “allowed regions” of $(\varphi, \psi)$ space that do not incur any backbone steric clashes. Dashed black lines indicated the “outer limit” of allowed regions of $(\varphi, \psi)$ 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 [14, 16] 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 $\phi$ and $\psi$ which can vary from unit to unit but in the ideal case can be uniquely determined by a single pair $(\varphi, \psi)$. We focus our attention here on two common types of protein secondary structure: $\alpha$- and $310$-helices which are stabilized by hydrogen bonds between residues $i \rightarrow i + 4$ and $i \rightarrow i + 3$, respectively. Even though both $\alpha$- and $310$-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.

2.1.2 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 2.1 shows that the contours related to different possible hydrogen bond lengths center around a distinctly different $(\varphi, \psi)$ pair. It appears that it would be possible to form much closer associations between backbone
Figure 2.4: Hydrogen bond geometry, styled after [3, Figure 11, page 751] which was reproduced in [8, Figure 22, page 143] (a) Angles $\mu = \angle COH$ and $\nu = \angle NHO$ used to define hydrogen bond quality. (b) Angles $\beta$ (out of plane) and $\gamma$ (in plane) were defined in [3, Figure 11, page 751] and later used in [8, Figure 22, page 143].

There are many angles of interest in defining and analyzing hydrogen bonds, as shown in Figure 2.4. One angle in particular, the $\angle NHO$ angle denoted by $\nu$ in Figure 2.4, relates to the linearity of the hydrogen bond. Unfortunately, this angle is also not decisive in determining $\alpha$-helical structure. Figure 2.2 shows the contours related to different values of this angle, and the “ideal” value of $\nu = 180$ degrees leads us to a $(\varphi, \psi)$ 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 $(\varphi, \psi)$ space, not near the empirically observed values. On the other hand, Figure 2.2 confirms that the empirical $(\varphi, \psi)$ values have a reasonable NHO angle, about 165 degrees. That is, the NHO angle is not very sensitive to choice of $(\varphi, \psi)$ in this region. Thus we are forced to look for another explanation.

### 2.1.3 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 [39]. Soon afterward [32, 42], 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 [42] 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 [42]: “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 [42] 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 [32] 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 interest [9, 10], it now seems clear that the variation in hydrophobicity in proteins has a
Figure 2.5: Contours for various values of distance $d_{OiO_{i+1}}$ as function of angles $\psi_{i+1}$ (Y-axis) and $\varphi_{i+1}$ (X-axis) is shown in black for the case of ideal (Pauling) geometry. It is clear that the distance $d_{OiO_{i+1}}$ is increasing with increasing $\varphi_{i+1}$. Points for $(\varphi, \psi)$ 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.

significant and quantifiable effect on the behavior of proteins [17]. According to [69], “The prevailing view holds that the hydrophobic effect has a dominant role in stabilizing protein structures.” Thus we conclude that formation of $\alpha$-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 $\alpha$-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 $\alpha$-helices in solution. Paradoxically, poly-valine does not so readily form $\alpha$-helices in solution [48, 6, 7]. More precisely, the order of propensity for helix formation is [48, Fig.1, page 4931] as follows:
Ala ≈ Leu > Ile >> Val >>> 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 carboxyl 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 $(\varphi, \psi)$ in typical helices can be characterized by a balance between two competing criteria: maximizing Hbond linearity and maximizing the distance between neighboring main-chain oxygens $d_{O_iO_{i+1}}$, as indicated in Figure 2.5. The latter criteria we think is important for helix solvation in water. As shown in Figure 2.5, and developed in full detail subsequently in this report, the contours related to Hbond linearity and distance are similar. Thus we could equivalently characterize the optimal $(\varphi, \psi)$ angles by a competition between maximizing the distance between neighboring main-chain oxygens $d_{O_iO_{i+1}}$ and minimizing the Hbond distance.

What is striking in Figure 2.5 is that the contours for $d_{O_iO_{i+1}}$ are orthogonal to all other contours, making it the leading candidate for explaining the dominant values of $(\varphi, \psi)$. Many other contours related to other distances and angles will be presented subsequently. Maximizing $d_{O_iO_{i+1}}$ 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 $\varphi$. Then the choice of $\psi$ is largely determined by optimizing the angle $\overline{NHO}$ as shown in Figure 2.2. That is, if we primarily were optimizing the hydrogen bond distance shown in Figure 2.1, then a smaller (more negative) value of $\psi$ would have been obtained.

The choice of $(\varphi, \psi)$ angles also dictates the distance between the oxygen on the backbone carboxyl (C-O) group and the C$_\beta$ carbon on the succeeding residue (except for glycine). We
Figure 2.6: 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 $\varphi$. The distance $O^{i-1} \cdots C^i$ is one of the parameters related to a potential clash in a classic view [34], 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 2.6 that the empirical values for \((\varphi, \psi)\) angles largely maximize this distance, but variation in \(\varphi\) has little effect on the distance since nearby values of \(\varphi\) yield 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 \(C^i_\beta\) 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 2.6. This can be interpreted by saying that the optimal angles maximize the opportunity for backbone hydration, since the hydrophobic \(C^i_\beta\) has been moved as far as possible from the hydrogen bond constituent \(O^{i-1}\).

Figure 2.6 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.

The role of the distance \(O^{i-1} \cdots C^i_\beta\) was considered in [34] 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 \(\overline{NHO}\) 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 \((\varphi, \psi)\) angles.

### 2.1.4 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 \((\varphi, \psi)\) angles. This has been stud-
ied and will be the subject of a separate paper. 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, intercalation by water may occur on a very fast time scale compared to the time scale for helix restructuring. One way to examine that would be by molecular dynamics simulation, which we intend to do. It is likely that, say, the structure of polyalanine in vacuum has a different distribution of \((\varphi, \psi)\) angles from the same system in water. We intend to investigate this in separate work.

2.2 Helical structure

The polypeptide chain has the flexibility to assume many local configurations, as explained in Figure 2.7. The peptide bond typically does not allow rotation, but two other backbone bonds, \(N-C_\alpha\) and \(C_\alpha-C\), do rotate freely. Angles \(\varphi\) and \(\psi\) measure rotation along bonds \(N-C_\alpha\) and \(C_\alpha-C\) 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 \((\varphi_i, \psi_i)\) 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 2.4). 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 \((\varphi, \psi)\) are repeated from unit to unit. On one hand, we can fully specify a helical configuration by a single pair of values \((\varphi, \psi)\). 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: \(\alpha\)-helix and \(3_{10}\)-helix characterized by regular hydrogen bonding.
pattern of type \( i \rightarrow i + 4 \) and \( i \rightarrow i + 3 \), respectively. Various studies show that dihedral angle pairs \((-63^\circ, -43^\circ)\) and \((-60^\circ, -25^\circ)\), respectively, result in the most common helices within each group [35].

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 \((\phi, \psi)\). So, it is natural to study the relationship 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 \( \alpha \)- 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 [47, 63, 57, 19]. 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 \( \alpha \)-helices [19].

In Section 2.3 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 2.7), and allow angles \( \phi \)
and $\psi$ to vary. Any helix is uniquely determined by rotation angle $\theta$, radius $r$, and rise $d$. We derive formulas for these parameters as functions of $\varphi$ and $\psi$, thus showing that any repeated pair of angles $\varphi$ and $\psi$ defines a helical structure of a backbone. In Section 2.4 we give an overview of feasible helical configurations. In Section 2.5 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 $\alpha$-helix respectively. First, we find $(\varphi, \psi)$ such that the angle

$$N_j - H_j \cdots O_i = 180^\circ.$$ 

This property is called *linearity* of a hydrogen bond. We then minimize the distance $|O_i \cdots H_j|$ as a function of $(\varphi, \psi)$. We then show that canonical peaks in the $(\varphi, \psi)$ angle distributions at $(-63^\circ, -43^\circ)$ for $\alpha$-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 paper were performed by using MAPLE$^{\text{TM}}[1]$.

### 2.3 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.

#### 2.3.1 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 [36]

“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 [36]

“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 [36] 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, 21].
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 [47]. The first rigorous mathematical analysis with torsion angles as parameters was published in 1954 by Mizushima and Simanouchi [63]. They applied an idea of Eyring [30].

In [63] 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 $\xi$ and internal rotation angles $\zeta$ are the same from unit to unit. They obtained the following formulas for helical rotation angle $\theta$, rise $d$, and radius $r$:

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

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

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

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 $\xi_1, \xi_2, \xi_3$ respectively. Again applying the method of Eyring [30] they define

$$A = A_3 A_1 A_2 \quad (2.4)$$

and

$$B = A_3 A_1 B_2 + A_3 B_1 + B_3, \quad (2.5)$$

where

$$A_i = \begin{pmatrix}
- \cos \xi_i & - \sin \xi_i & 0 \\
\sin \xi_i \cos \zeta_j & - \cos \xi_i \cos \zeta_i & - \sin \zeta_i \\
\sin \xi_i \sin \zeta_i & - \cos \xi_i \sin \zeta_i & \cos \zeta_i
\end{pmatrix} \quad (2.6)$$
and

\[
B_i = \begin{pmatrix}
L_i \\
0 \\
0
\end{pmatrix}
\]

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

\[
\cos \theta = (a_{11} + a_{22} + a_{33} - 1)/2, \quad (2.8)
\]

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

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

Later in the reference written by Dickerson and Geis [18] published in 1969, we find the following formulas:

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

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

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

\[
d = \vec{b} \cdot \vec{a}. \quad (2.13)
\]

The helical radius was computed in vector form:

\[
r = \frac{1}{2} \cot \frac{\theta}{2} \vec{a} \times \vec{b} + \vec{b} - (\vec{b} \cdot \vec{a})\vec{a}. \quad (2.14)
\]
Figure 2.8: Helical parameters: rotation angle $\theta$, radius $r$, and rise $d$ are shown. Point $P$ is a projection of $C_i^\alpha$ onto the plane containing $C_{i+1}^\alpha$ and orthogonal to the rotation axis.

We will use a method introduced by Dix in 2002 [19] and compute $\theta$, $\bar{a}$, $d$, and $r$ symbolically as functions of torsion angles $\varphi$, $\psi$, and a bond length $L$.

### 2.3.2 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 $E_{(a_1,a_2,a_3)}(\mathcal{R}) = (e_0, e_1, e_2, e_3)$ in the following way (as introduced by Dix in [19]):

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

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

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

$$e_3 = \mathcal{R}_a - e_0 - e_1 - e_2.$$
Let a $4 \times 4$ matrix $A$ 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.

$$E(a'_1,a'_2,a'_3) = E(a_1,a_2,a_3)A.$$ (2.19)

It is of the form

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \bar{0} & A \end{pmatrix},$$ (2.20)

where $A$ is a $3 \times 3$ rotation matrix, $\bar{0} = (0,0,0)$, and $\bar{b}$ denotes the vector $\overrightarrow{a_1a'_1}$ in $\mathbb{R}^3$.

As explained in [19], any coordinate transformation matrix can be expressed as a product of coordinate transformation matrices of 3 types. Namely,

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

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

and

$$T_3(\zeta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\zeta & -\sin\zeta \\ 0 & 0 & \sin\zeta & \cos\zeta \end{pmatrix}. $$ (2.23)
These three matrices represent the following basic types of coordinate system transformations:

\[ E_{(a_2,a_1,a_3)} = E_{(a_1,a_2,a_3)}T_1(L), \]  

(2.24)

where \( L = L_{a_1,a_2} > 0 \) is a length of the \( a_1 \rightarrow a_2 \) bond,

\[ E_{(a_1,a_3,a_2)} = E_{(a_1,a_2,a_3)}T_2(\xi), \]  

(2.25)

where \( \xi = \xi_{a_1,a_2,a_3} > 0 \) is the angle \( \xi = \overline{a_1a_2a_3} \), and

\[ E_{(a_1,a_2,a')} = E_{(a_1,a_2,a)}T_3(\pm \zeta), \]  

(2.26)

where \( \zeta = \zeta_{a_1,a_2} \) is the dihedral connecting angle \([a_1a_2a]_a\). We limit the subscripts for \( \zeta \) 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

\[ E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})} = E_{(C^\alpha_{i},C'_{i},N_{i+1})}\mathcal{A}, \]  

(2.27)

or we can also write

\[ E_{(C^\alpha_{i},C'_{i},N_{i+1})} \xrightarrow{\mathcal{A}} E_{(C^\alpha_{i+1},C'_{i+1},N_{i+2})}. \]  

(2.28)

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

$$E_{i}^{(α,α',N_{i+1})} \xrightarrow{A} E_{i+1}^{(α,α',N_{i+2})} = E_{i}^{(α,α',N_{i+1})} \xrightarrow{T_{1}(L_{i}^{α,α'})}$$

$$E_{i}^{(α',α,N_{i+1})} \xrightarrow{T_{2}(ξ_{i}^{α,α',N_{i+1}})} E_{i+1}^{(α',α,N_{i+1})} \xrightarrow{T_{3}(±ξ_{i}^{α',α,N_{i+1}})}$$

$$E_{i}^{(α,α',N_{i+1})} \xrightarrow{T_{1}(L_{i}^{α,α'})} E_{i+1}^{(α,α',N_{i+1})} \xrightarrow{T_{2}(ξ_{i}^{α',α,N_{i+1}})} E_{i+1}^{(α,α',N_{i+1})} \xrightarrow{T_{3}(±ξ_{i}^{α',α,N_{i+1}})}$$

$$E_{i}^{(N_{i+1},α',α,N_{i+1})} \xrightarrow{T_{2}(ξ_{i}^{α',α,N_{i+1}})} E_{i+1}^{(N_{i+1},α',α,N_{i+1})} \xrightarrow{T_{1}(L_{i}^{α,α'})} E_{i+1}^{(N_{i+1},α',α,N_{i+1})} \xrightarrow{T_{3}(±ξ_{i}^{α',α,N_{i+1}})}$$

$$E_{i}^{(α',α,N_{i+1})} \xrightarrow{T_{1}(L_{i}^{α,α'})} E_{i+1}^{(α',α,N_{i+1})} \xrightarrow{T_{2}(ξ_{i}^{α',α,N_{i+1}})} E_{i+1}^{(α',α,N_{i+1})} \xrightarrow{T_{3}(±ξ_{i}^{α',α,N_{i+1}})}$$

We can write $A$ as a product of nine matrices:

$$A = T_{1}(L_{i}^{α,α'})T_{2}(ξ_{i}^{α,α',N_{i+1}})T_{3}(±ξ_{i}^{α',α,N_{i+1}})T_{1}(L_{i}^{α,α'})T_{2}(ξ_{i}^{α',α,N_{i+1}})T_{3}(±ξ_{i}^{α',α,N_{i+1}})$$

$$T_{3}(±ξ_{i}^{N_{i+1},α',α,N_{i+1}})T_{1}(L_{i}^{α,α'})T_{2}(ξ_{i}^{α',α,N_{i+1}})T_{3}(±ξ_{i}^{α',α,N_{i+1}})$$

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

$$L_{i}^{α,α'} = L_{i+1}^{α,α'} = L_{i+1}^{α,α'} = L,$$

$$ξ_{i}^{α,α',N_{i+1}} = 120^\circ, \quad ξ_{i+1}^{α,α',N_{i+1}} = 109.47^\circ,$$

$$ξ_{i}^{α',N_{i+1}} = 180^\circ, \quad ξ_{i+1}^{α',N_{i+1}} = ϕ, \quad ξ_{i+1}^{α',N_{i+1}} = ψ$$

(see Figure 2.7). Thus,

$$A = T_{1}(L)T_{2}(120^\circ)T_{3}(180^\circ)T_{1}(L)T_{2}(120^\circ)T_{3}(±ϕ)T_{1}(L)T_{2}(109.47^\circ)T_{3}(±ψ)$$

We performed all symbolic computations in MAPLE 16 [1] and obtained the following results:
\[ \mathcal{A} = \begin{pmatrix} 1 & 0^T \\ \frac{5L}{2} & \sqrt{3L/2} \\ \sqrt{3L/2} & 0 \end{pmatrix} \]  

(2.32)

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

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

(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 \( A \)

\[
\text{tr}(A) = \frac{1}{3} - \frac{4}{3} \cos(\varphi) \cos(\psi) + \frac{4}{3} \sin(\varphi) \sin(\psi).
\]

(2.34)

The eigenvalues of \( A \) are

\[
\lambda_1 = 1, \lambda_{2,3} = \cos(\theta) \pm i \sin(\theta),
\]

(2.35)

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

\[
\text{tr}(A) = 1 + 2 \cos(\theta).
\]

(2.36)
Then (2.34) and (2.36) give
\[
\cos(\theta) = -\frac{1}{3} - \frac{2}{3} \cos(\varphi) \cos(\psi) + \frac{2}{3} \sin(\varphi) \sin(\psi) = \frac{-1 - 2 \cos(\varphi + \psi)}{3} = \frac{1 - 4 \cos^2\left(\frac{\varphi + \psi}{2}\right)}{3}.
\]  

(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:
\[
\bar{u} = (A_{32} - A_{23}, A_{13} - A_{31}, A_{21} - A_{12})^T \\
= \left(-\frac{4}{3} \sin(\varphi - \psi), \frac{4\sqrt{2}}{3} \sin\left(\frac{\varphi + \psi}{2}\right) \cos\left(\frac{\varphi + \psi}{2}\right), \frac{4\sqrt{2}}{3} \sin\left(\frac{\varphi + \psi}{2}\right) \sin\left(\frac{\varphi - \psi}{2}\right)\right)^T.
\]

(2.38)

Then the normalized axis vector is
\[
\bar{a} = \frac{\bar{u}}{2 \sin(\theta)}.
\]

(2.39)

The distance \(d\) traveled from \(C^\alpha_i\) to \(C^\alpha_{i+1}\) parallel to the axis of rotation (also known as a rise of a helix) can be found by the formula
\[
d = |\bar{b} \cdot \bar{a}| \\
= \left|\frac{\sqrt{2L}}{2 \sin(\theta)} \left(-\frac{4}{3} \sin(\varphi - \psi), \frac{4\sqrt{2}}{3} \sin\left(\frac{\varphi + \psi}{2}\right) \cos\left(\frac{\varphi + \psi}{2}\right), \frac{4\sqrt{2}}{3} \sin\left(\frac{\varphi + \psi}{2}\right) \sin\left(\frac{\varphi - \psi}{2}\right)\right)\right| \\
= \left|\frac{-2L \sin\left(\frac{\varphi + \psi}{2}\right) \left(10 \cos\left(\frac{\varphi + \psi}{2}\right) - \sqrt{6} \cos\left(\frac{\varphi - \psi}{2}\right)\right)}{2 \sin(\theta)}\right|^2 \\
= \left|\frac{-L \sin\left(\frac{\varphi + \psi}{2}\right) \left(10 \cos\left(\frac{\varphi + \psi}{2}\right) - \sqrt{6} \cos\left(\frac{\varphi - \psi}{2}\right)\right)}{2 \sqrt{2 - 2 \cos^4\left(\frac{\varphi + \psi}{2}\right) + 2 \cos^2\left(\frac{\varphi + \psi}{2}\right)}}\right|^2.
\]

(2.40)

Let point \(P\) be a projection of \(C^\alpha_i\) onto the plane containing \(C^\alpha_{i+1}\) and orthogonal to the rotation axis. Let point \(D\) lie on the rotation axis such that \(\overline{C^\alpha_{i+1}D}\) is orthogonal to the axis. To find the radius \(r\) we consider triangles \(\triangle C^\alpha_i C^\alpha_{i+1}P\) and \(\triangle C^\alpha_{i+1}PD\) (see Figure 2.8).
Clearly,
\[ \frac{|C_{i+1}^\alpha P|^2}{\sin^2(\theta)} = \frac{\vec{b} \cdot \vec{b} - d^2}{\sin^2(\theta)} = \frac{r^2}{(\sin(\frac{180^\circ - \theta}{2}))^2}. \]  

Thus we have
\[
\begin{align*}
    r^2 &= \frac{(\vec{b} \cdot \vec{b} - d^2) \cos^2(\frac{\theta}{2})}{(1 - \cos^2(\theta))} = \frac{(\vec{b} \cdot \vec{b} - d^2)(\cos(\theta) + 1)}{2(1 - \cos(\theta))(1 + \cos(\theta))} \\
    &= \frac{(\vec{b} \cdot \vec{b} - d^2)}{2(1 - \cos(\theta))} = \frac{(\vec{b} \cdot \vec{b} - d^2)}{2(1 - \frac{\text{tr}(A) - 1}{2})} = \frac{(\vec{b} \cdot \vec{b} - d^2)}{3 - \text{tr}(A)} \\
    &= \frac{7L^2 - \frac{L^2 \sin^2(\frac{\phi + \psi}{2})}{8 - 8 \cos^4(\frac{\phi + \psi}{2}) + 8 \cos^2(\frac{\phi + \psi}{2})}}{\frac{8}{3} + \frac{4}{3} \cos(\phi + \psi)}. \tag{2.42}
\end{align*}
\]

2.3.3 **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 [18]. 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.

2.4 **Feasible protein backbone local regular structures - Allowed**

\( (\varphi, \psi) \) orientations

It is important to note that not all possible \((\varphi, \psi)\) 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.

2.4.1 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 [62]. 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 [25, Figure 10, page 4627]. However, for us the main conclusion is that the location of the preferred \((\phi, \psi)\) angles does not lie on the boundary of the normally allowed region. Thus the empirically observed \((\phi, \psi)\) values are not determined by steric hindrance. They could be perturbed substantially and still lie within the normally allowed region.
Sasisekharan treated atoms as simple impenetrable spheres and found three major normally allowed \((\phi, \psi)\) regions for standard radii of peptide atoms and outer limit regions for the smallest radii that can be still considered plausible. The \(\phi, \psi\) 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 2.9. 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 \(\alpha\)-helix and \(\beta\)-sheet and on small peptides (see [58, Figure 6, page 54]).
2.4.3 Feasible conclusions

Here we gave an overview of results from [62] 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 \((\phi, \psi)\). We use these results in Section 2.5 where we perform the optimization in the domain outlined in Figure 2.9.

<table>
<thead>
<tr>
<th>Contact</th>
<th>Normally allowed, A</th>
<th>Outer limit, A</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_\alpha \cdots C_\alpha)</td>
<td>3.20</td>
<td>3.00</td>
</tr>
<tr>
<td>(C \cdots C)</td>
<td>2.95</td>
<td>2.90</td>
</tr>
<tr>
<td>(C \cdots O)</td>
<td>2.80</td>
<td>2.70</td>
</tr>
<tr>
<td>(C \cdots N)</td>
<td>2.90</td>
<td>2.80</td>
</tr>
<tr>
<td>(C \cdots H)</td>
<td>2.40</td>
<td>2.20</td>
</tr>
<tr>
<td>(O \cdots O)</td>
<td>2.70</td>
<td>2.60</td>
</tr>
<tr>
<td>(O \cdots N)</td>
<td>2.70</td>
<td>2.60</td>
</tr>
<tr>
<td>(O \cdots H)</td>
<td>2.40</td>
<td>2.20</td>
</tr>
<tr>
<td>(N \cdots N)</td>
<td>2.70</td>
<td>2.60</td>
</tr>
<tr>
<td>(N \cdots H)</td>
<td>2.40</td>
<td>2.20</td>
</tr>
<tr>
<td>(H \cdots H)</td>
<td>2.00</td>
<td>1.90</td>
</tr>
</tbody>
</table>

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

2.5 Stable protein backbone structures and hydrogen bond geometry

Recall that all backbone local regular structures (i.e., helices) can be defined by a pair \((\phi, \psi)\) for \(-180^\circ \leq \phi, \psi \leq 180^\circ\), 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 [19].
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: $\alpha$-helix and $3_{10}$-helix characterized by regular hydrogen bonding pattern of type $i \rightarrow i + 4$ and $i \rightarrow i + 3$, respectively. Even though both $\alpha$- 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 [35]. Using mathematical models of protein $\alpha$- and $3_{10}$-helices, incorporating empirical angles and bond lengths, we compute various geometrical parameters of a main chain – main chain hydrogen bond in each structure as functions of $(\varphi, \psi)$ and study their influence on helical stability. We optimize with respect to hydrogen bond linearity and length in $\alpha$- and $3_{10}$-helix and show that canonical $(\varphi, \psi)$ 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 $(\varphi, \psi)$ angles.

Previously, mathematical optimization was only applied to the $\alpha$-helix, and only the linearity of the hydrogen bond was considered [19]. We successfully reproduce his result and use this as a verification that our computational model is correct. Our original contribution to the topic starts in Sections 2.5.3 and 2.5.4 where we compute a similar linearity result for the $3_{10}$-helix. It then continues in Sections 2.5.5, 2.5.6, and 2.5.7 where we minimize the length of a hydrogen bond in both $\alpha$- and $3_{10}$-helices, compute other relevant parameters, and give numerical results respectively. In particular, our results in Section 2.5.7 show that canonical peaks at $(-63^\circ, -43^\circ)$ in $\alpha$-helix and at $(-60^\circ, -25^\circ)$ 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 2.5.9. First, we notice a clustering of the in-plane component $\gamma_{C_i-O_i...H_j}$
Table 2.2: Ranges of $\phi$ and $\psi$ allowed by the contact distances between the atoms $N_1, C_2,$ and $C_\beta$ and the remaining backbone atoms in the two residues for non-glycyl residues and between the atoms $N_1, C_\alpha,$ 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.

around $12-13^\circ$ for canonical peak values in both $\alpha$- and $3_{10}$-helices (see Figure 2.4(b) for the definition of the $\gamma$ angle), and then show how the relationship between angles $C_i \rightarrow O_i \cdots H_j$, $N_j \rightarrow H_j \cdots O_i$, and the distance $|H_j \cdots O_i|$ play a very important role in formation of most common $\alpha$- and $3_{10}$-helices.

2.5.1 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 strong enough to hold molecules XH and YZ together at normal temperatures and directional enough so that this association is orientationally specific [33]. Here we consider hydrogen bonds formed within a chain of amino acids. Namely, we study backbone-backbone hydrogen bonds $N_j \rightarrow H_j \cdots O_i = C_i$, formed between groups $C_i = O_i$ of the $i$-th amino acid and $N_j \rightarrow H_j$ 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 2.4):

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°, −43°). Structure was solved in 2010 via X-Ray chrystallography with resolution 0.82Å. Data is taken from Protein Geometry Database [22].

<table>
<thead>
<tr>
<th>ID</th>
<th>Chain ID</th>
<th>AA</th>
<th>ϕ(°)</th>
<th>ψ(°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>A</td>
<td>Gln</td>
<td>-64.15</td>
<td>-43.26</td>
</tr>
<tr>
<td>112</td>
<td>A</td>
<td>Lys</td>
<td>-61.02</td>
<td>-44.53</td>
</tr>
<tr>
<td>113</td>
<td>A</td>
<td>Val</td>
<td>-60.18</td>
<td>-42.60</td>
</tr>
<tr>
<td>114</td>
<td>A</td>
<td>Ser</td>
<td>-63.65</td>
<td>-43.56</td>
</tr>
<tr>
<td>115</td>
<td>A</td>
<td>Glu</td>
<td>-64.30</td>
<td>-40.59</td>
</tr>
<tr>
<td>116</td>
<td>A</td>
<td>Ala</td>
<td>-64.76</td>
<td>-41.70</td>
</tr>
<tr>
<td>117</td>
<td>A</td>
<td>Cys</td>
<td>-63.19</td>
<td>-41.83</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

Dihedral angle pairs (−63°, −43°) and (−60°, −25°) 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 [35, 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°, −43°) [40]. An example of such helices in membrane...
proteins is found in the fragment of a transmembrane protein integrin depicted in PDB file 2K9J. The $\langle \varphi, \psi \rangle$ values for an $\alpha$-helix fragment 2K9J.A.12-30 are shown in Table 2.4.

2.5.3 Computational models and transformation matrices for $\alpha$- and $3_{10}$-helix

Here we are interested in two coordinate transformations:

$$E(H_j,N_j,C'_{j-1}) \xrightarrow{A_{j-i}} E(O_i,C'_i,C''_i) \quad (2.43)$$

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 $\alpha$-helices for $j-i = 4$.

The transformation matrix $A_4$ was computed by Dix in [19] 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 [19] a proper graph representation of three-dimensional molecular structure of an $\alpha$-helix was created in IMIMOL and a Maple procedure to compute $A_4$ 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 $A_4$ exactly as it was done in Section 5.4 [19]. However, when we used the procedure exported by IMIMOL for computing $A_4$ to compute the optimal $\langle \varphi, \psi \rangle$ values for the linearity of a hydrogen bond in $\alpha$-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
Table 2.4: INTEGRIN ALPHAIIb-BETA3 TRANSMEMBRANE COMPLEX 2K9J [29].

An example of a membrane protein α-helix with ϕ and ψ values tightly clustered around canonical peak (−63°,−43°). Structure was solved in 2009 via solution NMR. Data is taken from Protein Data Bank [22].

A₄ 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 A₄ 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 A₄ and compute A₃ in a similar manner. Recall, that the ideal parameters of the peptide chain are ξ_{C'_{i+1},N_{i+1}} = 120°, and ξ_{N_{i+1},C'_{i+1},C_{i+1}} = 120°, ξ_{C'_{i+1},N_{i+1},C_{i+1}} = 109.47°, ξ_{C'_{i+1},C_{i+1}} = 180°. As before, we denote the torsion angles by ϕ and ψ: ξ_{N_{i+1},C'_{i+1}} = ϕ, ξ_{C'_{i+1},C_{i+1}} = ψ. To compute A₃ and A₄ 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]

$$L_{C_i^\alpha,C_i^\alpha} = 1.52, \ L_{C_{i+1}^\alpha,N_{i+1}} = 1.33, \ L_{N_{i+1},C_{i+1}^\alpha} = 1.45$$

(see Figure 2.7), corresponding to the ideal $\alpha$-helical geometry. We will also use two additional bond lengths $L_{O_1,C'_i} = 1.23, L_{H_{J_i},N_J} = 1$ also described in [56]. To simplify notation we can set $i = 1$ without any loss of generality. Then from (2.43) for an $\alpha$-helix

$$E_{(H_5,N_5,C'_4)} \xrightarrow{\ T_3(0)\ } E_{(N_5,C'_4,C'_4^\alpha)} \xrightarrow{\ T_1(1.33)\ } E_{(C'_4^\alpha,N_5,C'_4^\alpha)} \xrightarrow{\ T_2(120)\ } E_{(N_5,H_5,C'_4)} \xrightarrow{\ T_2(120)\ } E_{(N_5,C'_4,H_5)}$$

$$E_{(C'_4^\alpha,N_4,C'_4^\alpha)} \xrightarrow{\ T_3(1.52)\ } E_{(C'_4^\alpha,C'_4^\alpha,N_4)} \xrightarrow{\ T_2(109.27)\ } E_{(C'_4^\alpha,N_4,C'_4^\alpha)} \xrightarrow{\ T_3(\varphi)\ } E_{(C'_4^\alpha,N_4,C'_4^\alpha)} \xrightarrow{\ T_1(1.45)\ } E_{(N_5,C'_4,C'_4^\alpha)}$$

$$E_{(N_4,C'_3^\alpha,C'_3^\alpha)} \xrightarrow{\ T_3(\varphi)\ } E_{(C'_3,C'_3^\alpha,N_3)} \xrightarrow{\ T_1(1.52)\ } E_{(C'_3^\alpha,C'_3^\alpha,N_3)} \xrightarrow{\ T_2(109.27)\ } E_{(C'_3^\alpha,N_3,C'_3^\alpha)} \xrightarrow{\ T_3(\varphi)\ } E_{(C'_3^\alpha,N_3,C'_3^\alpha)}$$

$$E_{(N_3,C'_3^\alpha,C'_2^\alpha)} \xrightarrow{\ T_2(120)\ } E_{(N_3,C'_2^\alpha,C'_3^\alpha)} \xrightarrow{\ T_3(1.52)\ } E_{(C'_3^\alpha,C'_2^\alpha,N_3)} \xrightarrow{\ T_2(109.27)\ } E_{(C'_3^\alpha,N_2,C'_2^\alpha)}$$

$$E_{(C'_2^\alpha,N_2,C'_1^\alpha)} \xrightarrow{\ T_3(\varphi)\ } E_{(C'_2^\alpha,C'_2^\alpha,N_3)} \xrightarrow{\ T_1(1.45)\ } E_{(N_2,C'_2^\alpha,C'_1^\alpha)} \xrightarrow{\ T_2(120)\ } E_{(N_2,C'_1^\alpha,C'_1^\alpha)} \xrightarrow{\ T_3(180)\ } E_{(N_2,C'_1^\alpha,C'_1^\alpha)}$$

$$E_{(C'_1,N_2,C'_1^\alpha)} \xrightarrow{\ T_1(1.33)\ } E_{(C'_1,C'_1^\alpha,N_2)} \xrightarrow{\ T_2(120)\ } E_{(C'_1,C'_1^\alpha,O_1)} \xrightarrow{\ T_3(180)\ } E_{(C'_1,O_1,C'_1^\alpha)} \xrightarrow{\ T_2(120)\ } E_{(C'_1,O_1,C'_1^\alpha)}$$

$$\xrightarrow{\ T_1(1.23)\ } E_{(O_1,C'_1^\alpha,C'_1^\alpha)}.$$

(2.44)
and for $3_{10}$ helix

$$
\begin{align*}
E(H_{4,N_4,C'_4}) \xrightarrow{A_3(\varphi,\psi)} E(O_1,C'_1,C''_1) &= E(H_{4,N_4,C'_4}) \xrightarrow{T_1(1)} E(N_4,H_{4,C'_4}) \xrightarrow{T_2(120)} E(N_4,C'_4,H_{4}) \\
T_3(0) &\xrightarrow{T_1(1.33)} E(N_4,C'_4,C''_3) \xrightarrow{T_2(120)} E(C'_3,N_4,C''_3) \xrightarrow{T_3(\varphi)} E(C''_3,C'_3,N_3) \\
T_1(1.52) &\xrightarrow{T_2(109.27)} E(C''_3,N_3,C'_3) \xrightarrow{T_3(\varphi)} E(C''_3,N_3,C'_2) \xrightarrow{T_1(1.45)} E(N_3,C''_3,C'_2) \\
T_2(120) &\xrightarrow{T_3(180)} E(N_3,C''_2,C'_2) \xrightarrow{T_2(109.27)} E(C''_2,N_3,C'_2) \xrightarrow{T_3(\varphi)} E(C''_2,N_2,C'_2) \\
T_3(\psi) &\xrightarrow{T_1(1.52)} E(C''_2,C'_2,N_2) \xrightarrow{T_2(120)} E(C''_2,C'_2,N_2) \xrightarrow{T_3(180)} E(N_2,C''_2,C'_2) \xrightarrow{T_3(\varphi)} E(N_2,C''_2,C'_1) \xrightarrow{T_1(1.33)} E(C''_2,N_2,C'_1) \\
T_1(1.45) &\xrightarrow{T_2(120)} E(N_2,C''_2,C'_1) \xrightarrow{T_3(180)} E(N_2,C''_2,C'_1) \xrightarrow{T_2(120)} E(C''_1,C'_1,O_1) \xrightarrow{T_1(1.23)} E(O_1,C''_1,C'_1)
\end{align*}
$$

(2.45)

We can then express $A_4$ and $A_3$ as products of $T_1, T_2, T_3$ matrices:

$$
A_4(\varphi, \psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45) \\
T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)T_2(120)T_3(180) \\
T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120) \\
T_3(180)T_2(120)T_1(1.23),
$$

(2.46)

and

$$
A_3(\varphi, \psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45) \\
T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45)T_2(120)T_3(180) \\
T_1(1.33)T_2(120)T_3(180)T_2(120)T_1(1.23).
$$

(2.47)
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 [19]. We will explain his result and compute a similar result for the case of 310-helix.

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

$$E(O_1,C'_1,C''_1)(\mathbf{R}) = A_4(\varphi,\psi).$$

We want to find optimal $\varphi, \psi$ that will result in angle $N_5 \hat{H}_1 \cdots O_1$ being $180^\circ$. 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 $E(O_1,C'_1,C''_1)(\mathbf{R})$ we have

$$\mathbf{R}_{O_1} = E(O_1,C'_1,C''_1)(\mathbf{R})(1,0,0,0)^T = E(H_5,N_5,C'_4)(\mathbf{R})A_4(\varphi,\psi)(1,0,0,0)^T.$$  

On the other hand

$$\mathbf{R}_{O_1} = E(H_5,N_5,C'_4)(\mathbf{R})(1,x_{O_1},y_{O_1},z_{O_1})^T,$$

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

$$(1,x_{O_1},y_{O_1},z_{O_1})^T = A_4(\varphi,\psi)(1,0,0,0)^T$$

from which we can deduce $y_{O_1} = A_4(\varphi,\psi)_{3,1}$ and $z_{O_1} = A_4(\varphi,\psi)_{4,1}$. 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 $\varphi$ and $\psi$

$$A_4(\varphi,\psi)_{3,1} = 0, \text{ and } A_4(\varphi,\psi)_{4,1} = 0.$$  

Solving (2.52) simultaneously will give us the optimal $(\varphi, \psi)$ values in α-helix favoring lin-
earity of the hydrogen bond $N_5-H_5\cdots O_1$. By the exact same argument solving
\[ A_3(\varphi, \psi)_{3,1} = 0, \text{ and } A_3(\varphi, \psi)_{4,1} = 0 \] (2.53)
will give us optimal $(\varphi, \psi)$ values for a hydrogen bond $N_4-H_4\cdots O_1$ in a $3_{10}$-helix with angle $\overline{N_4H_1O_1} = 180^\circ$. Details regarding both solutions are summarized in Table 2.5.

2.5.5 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
\[ \ell_{j-i}(\varphi, \psi) = |H_j\cdots O_i| = \|H_jO_i\| = \|(A_{j-i}(\varphi, \psi)_{2,1}, A_{j-i}(\varphi, \psi)_{3,1}, A_{j-i}(\varphi, \psi)_{4,1})\|. \] (2.54)

Then solving
\[ \frac{\partial \ell_{j-i}(\varphi, \psi)}{\partial \varphi} = 0, \text{ and } \frac{\partial \ell_{j-i}(\varphi, \psi)}{\partial \psi} = 0 \] (2.55)
simultaneously will give us a pair $(\varphi, \psi)$ that minimizes the length of a hydrogen bond $N_j-H_i\cdots O_i$.

It follows from (2.55), that solving
\[ \frac{\partial \ell_4(\varphi, \psi)}{\partial \varphi} = 0, \text{ and } \frac{\partial \ell_4(\varphi, \psi)}{\partial \psi} = 0 \] (2.56)
simultaneously will give us a pair $(\varphi, \psi)$ that minimizes the length of a hydrogen bond $N_5-H_1\cdots O_1$ in $\alpha$-helix. And solving the system of equations
\[ \frac{\partial \ell_3(\varphi, \psi)}{\partial \varphi} = 0, \text{ and } \frac{\partial \ell_3(\varphi, \psi)}{\partial \psi} = 0 \] (2.57)
will give us a pair $(\varphi, \psi)$ 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.
2.5.6 Computing other geometrical parameters of a hydrogen bond

Clearly, for an \( \alpha \)-helix

\[
\cos(N_5H_5\cdots O_1) = \frac{H_5N_5 \cdot H_5O_1}{\|H_5N_5\| \|H_5O_1\|} = \frac{(1, 0, 0)(A_4(\varphi, \psi)_{2,1}, A_4(\varphi, \psi)_{3,1}, A_4(\varphi, \psi)_{4,1})^T}{\ell_4(\varphi, \psi)} = \frac{A_4(\varphi, \psi)_{2,1}}{\ell_4(\varphi, \psi)}
\]  

and for a \( 3_{10} \)-helix

\[
\cos(N_4H_4\cdots O_1) = \frac{H_4N_4 \cdot H_4O_1}{\|H_4N_4\| \|H_4O_1\|} = \frac{(1, 0, 0)(A_3(\varphi, \psi)_{2,1}, A_3(\varphi, \psi)_{3,1}, A_3(\varphi, \psi)_{4,1})^T}{\ell_3(\varphi, \psi)} = \frac{A_3(\varphi, \psi)_{2,1}}{\ell_3(\varphi, \psi)}
\]

We will be also interested in computing angle \( C_1O_1\cdots H_5 \) along with its two components: in-plane \( \gamma_{C_1O_1\cdots H_5} \), and out-of-plane \( \beta_{C_1O_1\cdots H_5} \), for an \( \alpha \)-helix. Similarly, for
a $3_{10}$-helix we will compute $C_1 - O_1 \cdots H_4$ together with in-plane $\gamma_{C_1-O_1\cdots H_4}$, and out-of-plane $\beta_{C_1-O_1\cdots H_4}$ components, respectively. Let

$$C_4(\varphi, \psi) = T_1 (1)T_2 (120)T_3 (0)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)$$

$$T_1 (1.45)T_2 (120)T_3 (180)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)T_1 (1.45)$$

$$T_2 (120)T_3 (180)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)T_1 (1.45)T_2 (120)$$

$$T_3 (180)T_1 (1.33).$$

Clearly

$$\overline{H_5C_1} = (C_4(\varphi, \psi)_{2,1}, C_4(\varphi, \psi)_{3,1}, C_4(\varphi, \psi)_{4,1}). \quad (2.61)$$

Then

$$\overline{C_1O_1} = \overline{H_5O_1 \cdots H_4C_1} = \quad (2.62)$$

$$= (A_4(\varphi, \psi)_{2,1} - C_4(\varphi, \psi)_{2,1}, A_4(\varphi, \psi)_{3,1} - C_4(\varphi, \psi)_{3,1}, A_4(\varphi, \psi)_{4,1} - C_4(\varphi, \psi)_{4,1}),$$

and

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

where $||O_1C_1|| = 1.23$.

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

$$N_4(\varphi, \psi) = T_1 (1)T_2 (120)T_3 (0)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)$$

$$T_1 (1.45)T_2 (120)T_3 (180)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)T_1 (1.45) \quad (2.64)$$

$$T_2 (120)T_3 (180)T_1 (1.33)T_2 (120)T_3 (\psi)T_1 (1.52)T_2 (109.27)T_3 (\varphi)T_1 (1.45).$$

Then

$$\overline{H_5N_1} = (N_4(\varphi, \psi)_{2,1}, N_4(\varphi, \psi)_{3,1}, N_4(\varphi, \psi)_{4,1}), \quad (2.65)$$
and

\[ N_1 C_1 = H_5 C_1 - H_5 N_1 = \]
\[ = (C_4(\varphi, \psi)_{2,1} - N_4(\varphi, \psi)_{2,1}, C_4(\varphi, \psi)_{3,1} - N_4(\varphi, \psi)_{3,1}, C_4(\varphi, \psi)_{4,1} - N_4(\varphi, \psi)_{4,1}). \]  \hspace{1cm} (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

\[ \overline{N} = \frac{N_1 C_1 \times C_1 O_1}{\| N_1 C_1 \times C_1 O_1 \|}, \]  \hspace{1cm} (2.67)

and let \( \overline{O_1 H_5}^{Proj} \) be a projection of \( \overline{O_1 H_5} \) onto that plane

\[ \overline{O_1 H_5}^{Proj} = \overline{O_1 H_5} - (\overline{N} \cdot \overline{O_1 H_5}) \overline{N}. \]  \hspace{1cm} (2.68)

Then

\[ \cos(\gamma_{C_1 - O_1 \cdots H_5}) = \frac{C_1 O_1 \cdot \overline{O_1 H_5}^{Proj}}{1.23 \| \overline{O_1 H_5}^{Proj} \|}, \]  \hspace{1cm} (2.69)

and

\[ \cos(\beta_{C_1 - O_1 \cdots H_5}) = \frac{\overline{O_1 H_5} \cdot \overline{O_1 H_5}^{Proj}}{\ell_4(\varphi, \psi) \| \overline{O_1 H_5}^{Proj} \|}. \]  \hspace{1cm} (2.70)

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

\[ C_3(\varphi, \psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi) \]
\[ T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45) \]  \hspace{1cm} (2.71)
\[ T_2(120)T_3(180)T_1(1.33), \]

and thus

\[ C_1 O_1 = H_4 O_1 - H_1 C_1 = \]
\[ = (A_3(\varphi, \psi)_{2,1} - C_3(\varphi, \psi)_{2,1}, A_3(\varphi, \psi)_{3,1} - C_3(\varphi, \psi)_{3,1}, A_3(\varphi, \psi)_{4,1} - C_3(\varphi, \psi)_{4,1}), \]  \hspace{1cm} (2.72)
\[
\cos(C_1 - O_1 \cdots H_4) = \frac{O_1 H_4 \cdot O_1 C_1}{\|O_1 H_4\| \|O_1 C_1\|} = \frac{O_1 H_4 \cdot O_1 C_1}{1.23 \|O_1 H_4\|},
\] (2.73)

and
\[
\mathcal{N}_3(\varphi, \psi) = T_1(1)T_2(120)T_3(0)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)
\]
\[
T_1(1.45)T_2(120)T_3(180)T_1(1.33)T_2(120)T_3(\psi)T_1(1.52)T_2(109.27)T_3(\varphi)T_1(1.45).
\] (2.74)

Then
\[
\overline{H_4 N_1} = (\mathcal{N}_3(\varphi, \psi)_{2,1}, \mathcal{N}_3(\varphi, \psi)_{3,1}, \mathcal{N}_3(\varphi, \psi)_{4,1}),
\] (2.75)

and
\[
\mathcal{N}_1 C_1 = \overline{H_4 C_1} - \overline{H_4 N_1} =
\]
\[
= (C_3(\varphi, \psi)_{2,1} - \mathcal{N}_3(\varphi, \psi)_{2,1}, C_3(\varphi, \psi)_{3,1} - \mathcal{N}_3(\varphi, \psi)_{3,1}, C_3(\varphi, \psi)_{4,1} - \mathcal{N}_3(\varphi, \psi)_{4,1}).
\] (2.76)

Let \(\overline{O_1 H_4}^{\text{Proj}}\) be a projection of \(\overline{O_1 H_4}\) onto that plane
\[
\overline{O_1 H_4}^{\text{Proj}} = \overline{O_1 H_4} - (N \cdot \overline{O_1 H_4})N.
\] (2.77)

Then
\[
\cos(\gamma_{C_1-O_1\cdots H_4}) = \frac{C_1 O_1 \cdot \overline{O_1 H_4}^{\text{Proj}}}{1.23 \|\overline{O_1 H_4}^{\text{Proj}}\|},
\] (2.78)

and
\[
\cos(\beta_{C_1-O_1\cdots H_4}) = \frac{\overline{O_1 H_4} \cdot \overline{O_1 H_4}^{\text{Proj}}}{\ell_3(\varphi, \psi) \|\overline{O_1 H_4}^{\text{Proj}}\|}.
\] (2.79)

Details regarding all of these parameters for the optimal solutions are summarized in Table 2.5.
2.5.7 Optimization results and balance between two criteria

We considered the model of a protein helix described above with angles $\phi$ and $\psi$ 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 respectively. Note that values $(\phi,\psi)=(-56.5^\circ,-50.5^\circ)$ and $|O_1\ldots H_5| = 1.76\text{Å}$ are the same as obtained by Dix in [19] which verifies the correctness of our computational model. In the third column we computed hydrogen bond parameters for $(\phi,\psi)$ pairs known as canonical peaks ($(\phi,\psi)$ values that occur most commonly) for $\alpha$- and $3\overline{1}_0$-helix respectively. The significance of these peaks can be clearly seen in [35, 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 [19] (see Sections 5.3-5.4 [19] for details).

In Figures 2.10 and 2.11 we plotted $(\phi,\psi)$ pairs corresponding to the shortest hydrogen bond, linear hydrogen bond and canonical peaks for $\alpha$-helix and for $3\overline{1}_0$-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 $\alpha$-helix by considering the contours shown in Figures 2.1 and 2.2. 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 $3\overline{1}_0$-helix is the same but using Figures 2.19 and 2.23.
Table 2.5: With \( \varphi \) and \( \psi \) as free parameters hydrogen bonds \( N_5\cdots H_5\cdots O_1 \) in \( \alpha \)-helix and \( N_4\cdots H_4\cdots O_1 \) 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 \( (\varphi, \psi) \) pairs known as canonical peaks \( ((\varphi, \psi) \) values that occur most commonly) for \( \alpha \)- and \( 3_{10} \)-helix respectively (values were taken from [35]). 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.

| \( \alpha \)-helix | \( N_5\cdots H_5\cdots O_1 = 180^\circ \) | min \( |O_1\cdots H_5| \) | canonical peak | data from [8] |
|-------------------|---------------------------------|-----------------|----------------|----------------|
| \( (\varphi, \psi) \) \( (^\circ, ^\circ) \) | \((-56.5, -50.5)\) [19] | \((-56.5, -57.5)\) | \((-63, -43)\) [35] | N/A |
| \( \varphi + \psi \) \( (^\circ) \) | 107 | 114 | 106 | N/A |
| \( |O_1\cdots H_5| \) \( (\text{Å}) \) | 1.76 [19] | 1.54 | 1.92 | 2.05 \( \pm \) 0.15 |
| \( N_5\cdots H_5\cdots O_1 \) \( (^\circ) \) | 180 | 146 | 166 | 157 \( \pm \) 9 |
| \( C_1\cdots O_1\cdots H_5 \) \( (^\circ) \) | 168 | 147 | 156 | 147 \( \pm \) 7 |
| \( \gamma_{C_1\cdots O_1\cdots H_5} \) \( (^\circ) \) | -3.8 | -25 | -11.8 | -18 \( \pm \) 9 |
| \( \beta_{C_1\cdots O_1\cdots H_5} \) \( (^\circ) \) | 10.6 | 22 | 20.5 | 28 \( \pm \) 8 |

| \( 3_{10} \)-helix | \( N_4\cdots H_4\cdots O_1 = 180^\circ \) | min \( |O_1\cdots H_4| \) | canonical peak | data from [8] |
|-------------------|---------------------------------|-----------------|----------------|----------------|
| \( (\varphi, \psi) \) \( (^\circ, ^\circ) \) | \((-50.3, -23.3)\) | \((-36.6, -39)\) | \((-60, -25)\) [35] | N/A |
| \( \varphi + \psi \) \( (^\circ) \) | 73.6 | 75.6 | 85 | N/A |
| \( |O_1\cdots H_4| \) \( (\text{Å}) \) | 1.76 | 1.62 | 1.93 | 2.17 \( \pm \) 0.16 |
| \( N_4\cdots H_4\cdots O_1 \) \( (^\circ) \) | 180 | 158 | 155 | 153 \( \pm \) 10 |
| \( C_1\cdots O_1\cdots H_4 \) \( (^\circ) \) | 144 | 159 | 122 | 114 \( \pm \) 10 |
| \( \gamma_{C_1\cdots O_1\cdots H_4} \) \( (^\circ) \) | -20.8 | -5.9 | -13 | -30 |
| \( \beta_{C_1\cdots O_1\cdots H_4} \) \( (^\circ) \) | 30 | 19.2 | 56 | 60 |

Table 2.6: Rotation angle \( \theta \), rise \( d \), and radius \( r \) computed for the following \( (\varphi, \psi) \)-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 \( (\varphi, \psi) \) pairs known as canonical peaks \( ((\varphi, \psi) \) pairs that occur most commonly) for \( \alpha \)- and \( 3_{10} \)-helix respectively (peak values \( (\varphi, \psi) \) values were taken from [35]).
2.5.8 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···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···N and between C–O and O···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),

\[ 40° \leq (N - \overrightarrow{H} \cdots O) - (C - \overrightarrow{O} \cdots H) \leq 70°. \]

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

Olovsson and Jonsson in 1976 [55] studied geometry of bonds N–H···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 \( C - \overrightarrow{O} \cdots H \) is clustered around 125° with some large deviations for longer bonds.

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

2.5.9 Relationship between geometrical parameters

We created detailed contour plots for geometrical parameters of hydrogen bonds \( N_5 - \overrightarrow{H_5} \cdots O_1 \) (see Figures 2.2, 2.16, 2.17, and 2.1) and \( N_4 - \overrightarrow{H_4} \cdots O_1 \) (see Figures 2.19, 2.20, 2.22, and 2.23) as functions of \( (\varphi, \psi) \). We then combined some contours in one plot for \( \alpha \)- and one for 3_10-helix, to show graphically a possible role that the difference \( (N_j - \overrightarrow{H_j} \cdots O_i) - \)
Figure 2.13: Relationship between $\alpha$-helix hydrogen bond parameters $H_5 \cdots O_1$, $C_1 - O_1 \cdots H_5$ and $C_1 - O_1 \cdots H_5$ is shown in the diagram. It suggests that minimizing the bond distance $H_j \cdots O_i$, maximizing linearity and minimizing the angle $C_i - O_i \cdots H_j$ simultaneously moves us closer to the canonical peaks for $\alpha$-helix.

$(C_i - O_i \cdots H_j)$, for $(j - i) = 3, 4$, might play in formation of peaks (see Figures 2.14 and 2.15). In particular, plots suggest that minimizing the bond distance $H_j \cdots O_i$, maximizing linearity and minimizing the angle $C_i - O_i \cdots H_j$ simultaneously moves us closer to the canonical peaks for both $\alpha$-helix and $3_{10}$-helix. We show this relationship graphically in the diagram (see Figure 2.13).

It has been noted in [8] that the hydrogen bond energy is very sensitive to the angle $N - \overset{\sim}{H} \cdots O$, but is hardly affected by the angle $C - \overset{\sim}{O} \cdots H$. Correlation between linearity and length of a hydrogen bond can be seen in Figures 2.25 and 2.26, where we plotted contours for $N_j - H_j \cdots O_i$ and $H_j \cdots O_i$ for $j - i = 3, 4$ simultaneously. We also plotted contours for $C_i - O_i \cdots H_j$ and $H_j \cdots O_i$ for $j - i = 3, 4$ (see Figures 2.28 and 2.29). It is clear that within allowed helical region a shorter hydrogen bond distance is favored for larger values of $N - \overset{\sim}{H} \cdots O$. On the other hand, a much broader range of $C - \overset{\sim}{O} \cdots H$ values is allowed when the hydrogen bond distance is restricted to smaller values.
2.5.10 Other geometrical restrictions

Reference [40] defines $\delta$ as the “peptide plane tilt angle” with respect to the helical axis, with $\delta = 0$ being the parallel case [40, Figure 1, page 2086]. The preferred range is $4^\circ \leq \delta \leq 12^\circ$. The figure [40, Figure 2, page 2087] presents peptide plane tilt angles with respect to the helical axis ($\delta$) diagramed as a function of $\varphi, \psi$ torsion angles from uniform helical models leading to the development of this Ramachandran-delta diagram. Superimposed on this diagram are $\varphi, \psi$ values from 500 high-resolution ($\geq 1.8\text{Å}$ resolution) crystal structures. The figure [40, Figure 3, page 2088] shows $\delta$ contours $\varphi, \psi$ values for a membrane protein.

2.5.11 Hydrogen bond conclusions

In this section we reviewed the definition of a hydrogen bond and related geometrical parameters. We explained the significance of ($\varphi, \psi$) values known as canonical peaks: ($-63^\circ, -43^\circ$) in $\alpha$-helix and ($-60^\circ, -25^\circ$) in $3_{10}$-helix [35]. Following the method of Dix [19] we expressed main chain – main chain hydrogen bond parameters in $\alpha$- and $3_{10}$-helix as functions of dihedral angles ($\varphi, \psi$). We successfully reproduced the optimization result from [19], where optimization was done with respect to the linearity of a hydrogen bond in $\alpha$-helix. We obtained the same values ($\varphi, \psi$)=($-56.5^\circ, -50.5^\circ$) as in [19]. 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 ($\varphi, \psi$)=($-50.3^\circ, -23.3^\circ$). We then minimized the length of a hydrogen bond in both $\alpha$- and $3_{10}$-helices obtaining values ($\varphi, \psi$)=($-56.5^\circ, -57.5^\circ$) and ($-36.3^\circ, -39^\circ$) respectively. For the four optimal pairs mentioned above and for the canonical peaks ($-63^\circ, -43^\circ$) in $\alpha$-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 2.10 and 2.11, we showed that canonical peaks ($-63^\circ, -43^\circ$) in $\alpha$-helix and ($-60^\circ, -25^\circ$) in $3_{10}$-helix do not result from optimization of hydrogen bonding with re-
Figure 2.14: This plot suggests that minimizing the bond distance $H_5 \cdots O_1$, maximizing linearity and minimizing the angle $C_1 \cdots O_1 \cdots H_5$ simultaneously moves us closer to the canonical peak for $\alpha$-helix: $\gamma_{C_1 \cdots O_1 \cdots H_5} = -15$ is shown in red and $= -10$ in blue dots.
Figure 2.15: 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: $\gamma_{C_1 \widehat{O}_1 \cdots H_4} = -15$ is shown in red and $= -10$ in blue dots.

spect to any linear combination of the two major hydrogen bond requirements: linearity and length constraints. We noticed a clustering of the in-plane component $\gamma_{C_i \widehat{O}_i \cdots H_j}$ around $12 - 13^\circ$ for canonical peak values in both $\alpha$- and $3_{10}$-helix (see Figures 2.14 and 2.15). And we showed how the relationship between angles $C_i \widehat{O}_i \cdots H_j$, $N_j \widehat{H}_j \cdots O_i$, and the distance $|H_j \cdots O_i|$ play a very important role in formation of most common $\alpha$- and $3_{10}$-helices (see the diagram in Figure 2.13, and contour plots in Figures 2.1 - 2.29).

Our results show the clustering of the in-plane component $\gamma_{C_i \widehat{O}_i \cdots H_j}$ at $12^\circ - 13^\circ$ for canonical peak values in both $\alpha$- and $3_{10}$-helices which we believe to be significant (see
2.6 Overall conclusions

In this paper 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 \(\theta\), rotation axis \(\vec{u}\), radius \(r\), and rise \(d\) in terms of \(\varphi, \psi\) and \(L\). Thus we showed that any repeated pair of angles \(\varphi\) and \(\psi\) defines a helical structure of a backbone. Formula (2.37) for rotation angle was previously published in [18]. To the best of our knowledge explicit formulas for rotation axis (2.38), rise (2.40), and radius (2.42) as functions of \(\varphi, \psi\) and \(L\) can not be found in the current literature.

Since not all helical backbone configurations are feasible, we continued by reviewing results on feasible backbone helical configurations from [62] and [60] obtained by Sasisekharan.
Figure 2.19: Contours for different values of $N_4\cdots\hat{H}_4\cdots O_1$ as a function of $(\phi, \psi)$ are plotted.

Figure 2.20: Contours for different values of $C_1\cdots\hat{O}_1\cdots H_4$ as a function of $(\phi, \psi)$ are plotted.

Figure 2.21: Contours for different angles as a function of $(\phi, \psi)$ for $3_{10}$-helix.

Figure 2.22: Contours for different values of the difference $(N_4\cdots\hat{H}_4\cdots O_1 - C_1\cdots\hat{O}_1\cdots H_4)$ as a function of $\phi, \psi$.

Figure 2.23: Contours for different values of the distance $|H_4\cdots\hat{O}_1|$ as function of $\phi, \psi$.

Figure 2.24: Contours for angle differences and distances as function of $\phi, \psi$ for $3_{10}$ helices.

48
Figure 2.25: Contours for different values of $N_1$–$\hat{H}_1$···$O_5$ and $|H_5$···$O_1|$ as functions of $(\phi, \psi)$ are plotted simultaneously.

Figure 2.26: Contours for different values of $N_1$–$\hat{H}_1$···$O_4$ and $|H_4$···$O_1|$ as functions of $(\phi, \psi)$ are plotted.

Figure 2.27: Contours for different values of $N_1$–$\hat{H}_1$···$O_k$ where $k = 4$ for $\alpha$-helix and $k = 3$ for $3_{10}$-helix. Clearly, a shorter hydrogen bond distance is favored for larger values of $N$–$\hat{H}$···$O$.  

49
Figure 2.28: α-helix: contours for different values of $C_1\cdots\tilde{O}_1\cdots H_5$ and $|H_5\cdots O_1|$ as functions of $(\phi, \psi)$.

Figure 2.29: $3_{10}$-helix: contours for different values of $C_1\cdots\tilde{O}_1\cdots H_4$ and $|H_4\cdots O_1|$ as functions of $(\phi, \psi)$.

Figure 2.30: Contours for different values of $C_1\cdots\tilde{O}_1\cdots H_k$ and $|H_k\cdots O_1|$ as functions of $(\phi, \psi)$, where $k = 5$ for α-helix and $k = 4$ for $3_{10}$-helix. A much broader range of $C\cdots\tilde{O}\cdots H$ values is allowed when the hydrogen bond distance is restricted to smaller values.
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 2.5 (see Figure 2.9).

Finally, we studied main chain – main chain hydrogen bonds in two types of common helical configurations: $\alpha$- and $3_{10}$-helices. We started by defining a hydrogen bond and its geometrical parameters. We then explained the significance of $(\varphi, \psi)$ values known as canonical peaks: $(-63^\circ, -43^\circ)$ in $\alpha$-helix and $(-60^\circ, -25^\circ)$ in $3_{10}$-helix [35]. We wanted to know how hydrogen bond geometry influenced the formation of canonical peaks. We computed main chain – main chain hydrogen bond parameters in $\alpha$- and $3_{10}$-helix as functions of dihedral angles $(\varphi, \psi)$. Optimization with respect to the linearity of a hydrogen bond in $\alpha$-helix was done by Dix in [19], where the pair $(\varphi, \psi)=(-56.5^\circ, -50.5^\circ)$ was obtained. We reproduced this result to test the correctness of our computation. Then following the method in [19] we computed a linearity result for $3_{10}$-helix and obtained values $(\varphi, \psi)=(-50.3^\circ, -23.3^\circ)$. Another important geometrical parameter we studied was the length of a hydrogen bond. We computed lengths of main chain – main chain hydrogen bonds in $\alpha$- and $3_{10}$-helices in terms of $(\varphi, \psi)$. By minimizing these lengths we obtained values $(\varphi, \psi)=(-56.5^\circ, -57.5^\circ)$ and $(-36.3^\circ, -39^\circ)$ for $\alpha$- 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 $\alpha$-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 $\alpha$-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 2.10 and 2.11). We also showed that the clustering of the in-plane component $\gamma_{C_i-\hat{O}_i...H_j}$ around $12 - 13^\circ$ for canonical peak values in both $\alpha$- and $3_{10}$-helix and the relationship between angles $C_i-\overrightarrow{O_i}...H_j$, $N_j-\overrightarrow{H_j}...O_i$ and the distance $|H_j...O_i|$ play a very important role in formation of most common $\alpha$- and $3_{10}$-helices. However, the most obvious contributor to the canonical peak values in both $\alpha$-helix and $3_{10}$-helix appears to be maximizing solvation of the back bone as

51
indicated in Figure 2.5 and Figure 2.6.
CHAPTER 3
SOLVATION SIGNATURE IN $\alpha$-HELICES

3.1 Proteins interact with water.

Proteins actively interact with hydration water at the surface as well as with water molecules buried inside [46]. Such interactions are essential for their structure and function [61], [27]. Despite the importance, relationship between solvent molecules and peptide structure is still poorly understood and remains an active research topic [46], [26]. $\alpha$-helix is the most abundant type of protein secondary structure. It plays an essential role within a protein and in various protein-protein interactions [64]. Thus it is especially important to understand how solvation influences the geometry of $\alpha$-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 [28]. Later some other studies of $\alpha$-helix interactions with the solvent and side-chains were performed in [20] and [53]. More detailed analysis of hydrated alpha-helices was done by Sundaralingam and Sekharudu in 1989 [49]. In [49] 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 $\alpha$-helix (see Figure 3.1). A nice summary on the topic was written in 2003 by McColl et. al. [23] where the importance of the main-chain carbonyl oxygen hydration was reemphasized as the most common and “especially relevant to aqueous solution studies.” Here we focus our study on the hydration of main-chain carbonyl oxygens, examining changes in a local $\alpha$-helical geometry of a dipeptide as a function of the distance between backbone carbonyl oxygen and the closest water molecule.

Our earlier analysis of $\alpha$-helix geometry in the previous chapter suggests that the values of dihedral angles $(\varphi, \psi) = (-63, 43)$, also known as the canonical peak [35], 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 $d_{O_iO_{i+1}}$, see Figure 2.5. 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 \(\alpha\)-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 \(\alpha\)-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 \(\alpha\)-helix, we measure distances between carbonyl oxygen \( O_i \) and all water oxygens available in the file and record the shortest such distance \( d_i = \min_W d_{O_iO_W} \), with the cut-off value of 10 Å. We also consider relevant \(\alpha\)-helical geometrical parameters, namely, distances \( d_{O_iC_{i+1}^\beta} \) and \( d_{O_iO_{i+1}} \), and dihedral angles \( \psi_i, \varphi_{(i+1)}, \psi_{i+1} \), see Figure 3.2. From ideal geometry computation we expect \( d_{O_iC_{i+1}^\beta} \) not to change significantly, see Figure 3.3 (A). On the other hand, distance \( d_{O_iO_{i+1}} \) is increasing with increasing \( \varphi_{i+1} \), see Figure 3.3 (B). We perform a detailed study of the changes in these angles and distances when external hydration occurs.

3.2.3 Dataset

We considered all .pdb files available in Protein Data Bank [22] (as of July 23, 2016) with resolution < 1.5 Å. We found 8754 such structures. After selecting files containing at least one \(\alpha\)-helix, and information about the location of (at least one) water molecule we were
Figure 3.2: We consider a pair of consecutive \(i\)-th and \((i+1)\)-th residues in an \(\alpha\)-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 \(d_i = \min_{W} d_{O_iO_W} < 10\AA\) (\(d_{O_iO_W}\)), i.e. the distance from the carbonyl oxygen \(O_i\) to the oxygen of the closest water molecule within 10Å range.
Figure 3.3: (A) Dependence of distance $d_{O_iC_{i+1}}$ (Y-axis) on angle $\phi_{i+1}$ (X-axis); (B) Dependence of distance $d_{O_iO_{i+1}}$ (Z-axis) on angles $\psi_{i+1}$ (Y-axis) and $\phi_{i+1}$ (X-axis) are shown for the case of ideal geometry [56].

<table>
<thead>
<tr>
<th>HR dataset (Res &lt; 1.5Å, R≤30)</th>
<th>Protein Structures</th>
<th>Number of $\alpha$-helices</th>
<th>Qualified Dipeptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>2176</td>
<td>20838</td>
<td>Nter 70065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mid 109410</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cter 53482</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Waters 990258</td>
</tr>
</tbody>
</table>

Table 3.1: Number of relevant protein structures, $\alpha$-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, ···, C5-C4, C4-C3, located in the middle of a helix; and C ter corresponds to dipeptides C3-C2, C2-C1, C1-Ncap at the C-terminus.

left with 8117 structures. We then culled these files for identical strutures using PISCES server [68], 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 $\leq 25$. We obtained a dataset of 2176 structures with the total of 20838 $\alpha$-helices, and 990258 waters. Amino-acids with alternative $C_\alpha$ positions were omitted from computations.

For High Resolution Protein Dataset statistics see Table 3.1.
<table>
<thead>
<tr>
<th></th>
<th>(d_i) (Å)</th>
<th>[0,3.5)</th>
<th>[3.5,5)</th>
<th>[5,10)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dipeptides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean (\varphi_{i+1}) (°)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean (d_{OiO_{i+1}}) (Å)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique water contacts only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean (\varphi_{i+1}) (°)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean (d_{OiO_{i+1}}) (Å)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

3.3 Solvation effects on backbone geometry in \(\alpha\)-helices

3.3.1 Results for high resolution protein dataset.

Basic dataset statistics is summarized 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. This explains why helical amino acid propensities computed in [24, Table 2 on p.5], see Figure 3.14, do not differ for exposed versus buried residues. The backboone interaction with water, as expected, highly correlates with larger \(\varphi\) angles and larger distances between neighboring main-chain carbonyl oxygens. For relationships between \(d_i\), \(\varphi_{i+1}\), and \(d_{OiO_{i+1}}\) see Figures 3.4-3.17. For amino-acid composition see Figure 3.13.
Figure 3.4: Histogram of distances $d_i$ for middle residues in $\alpha$-helices in High Resolution Protein Data Set 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 $\approx 5.1\,\text{Å}$, $\approx 5.8\,\text{Å}$, and $\approx 6.7\,\text{Å}$ 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. 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 that 10Å.)
Figure 3.5: Histogram of distances $d_i$ for middle residues in $\alpha$-helices in High Resolution Protein Data Set 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 $\approx 5.1\text{Å}$, $\approx 5.8\text{ Å}$, and $\approx 6.7\text{Å}$, seen in Figure 3.4, are gone.
Figure 3.6: Histograms of $\varphi_{i+1}$ (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 the dipeptide parameters in the case of the primary water contact with respect to the cases of secondary and no water contact.

Figure 3.7: Histograms of $\varphi_{i+1}$ (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 the dipeptide parameters in the case of the primary water contact with respect to the cases of secondary and no water contact.
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: 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: 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: Heatmap of dependence of $\varphi_{(i+1)}$ (X-axis) on $d_{O_iO_{i+1}}$ (Y-axis) for high resolution protein dataset. Peak: $(\varphi_{(i+1)}, d_{O_iO_{i+1}}) = (-62, 3.40)$
Figure 3.12: Least squares fit lines to scatter plots (the actual points not shown) of $\varphi_{i+1}$ (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) 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).

Figure 3.13: Amino acid composition for middle (i.e. excluding N- and C-ter) residues $i$ (X-axis) and $i+1$ (Y-axis) in $\alpha$-helices: (A) when $i$th residue carbonyl oxygen has water oxygen within $[0,3.5)\text{Å}$; (B) when $i$th residue carbonyl oxygen has water oxygen within $[3.5,5)\text{Å}$; (C) when $i$th residue carbonyl oxygen has water oxygen within $[5,10)\text{Å}$.
Figure 3.14: [24, Table 2 on p.5] Helical amino acid propensities do not differ for exposed and buried residues, suggesting that helices are likely to be equally solvated by either hydration or buried water independent of location. The table also shows that this is not true for β-sheets.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>α-helix</th>
<th></th>
<th></th>
<th>β-strand</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Buried</td>
<td></td>
<td>Exposed</td>
<td>Buried</td>
<td>Total</td>
</tr>
<tr>
<td>V</td>
<td>0.83</td>
<td>0.89</td>
<td>0.91</td>
<td>2.31</td>
<td>1.57</td>
<td>2.00</td>
</tr>
<tr>
<td>I</td>
<td>0.96</td>
<td>1.01</td>
<td>1.04</td>
<td>2.02</td>
<td>1.39</td>
<td>1.79</td>
</tr>
<tr>
<td>L</td>
<td>1.16</td>
<td>1.27</td>
<td>1.28</td>
<td>1.18</td>
<td>0.93</td>
<td>1.15</td>
</tr>
<tr>
<td>M</td>
<td>1.03</td>
<td>1.29</td>
<td>1.26</td>
<td>1.01</td>
<td>0.84</td>
<td>1.01</td>
</tr>
<tr>
<td>P</td>
<td>0.48</td>
<td>0.41</td>
<td>0.44</td>
<td>0.49</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td>A</td>
<td>1.43</td>
<td>1.37</td>
<td>1.41</td>
<td>0.48</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>C</td>
<td>0.63</td>
<td>0.85</td>
<td>0.85</td>
<td>1.24</td>
<td>1.07</td>
<td>1.36</td>
</tr>
<tr>
<td>F</td>
<td>0.88</td>
<td>0.99</td>
<td>1.00</td>
<td>1.50</td>
<td>1.10</td>
<td>1.40</td>
</tr>
<tr>
<td>Y</td>
<td>0.91</td>
<td>0.98</td>
<td>0.98</td>
<td>1.71</td>
<td>1.12</td>
<td>1.37</td>
</tr>
<tr>
<td>W</td>
<td>0.87</td>
<td>1.09</td>
<td>1.07</td>
<td>1.90</td>
<td>0.91</td>
<td>1.23</td>
</tr>
<tr>
<td>Q</td>
<td>1.34</td>
<td>1.21</td>
<td>1.26</td>
<td>0.96</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>S</td>
<td>0.74</td>
<td>0.80</td>
<td>0.76</td>
<td>0.86</td>
<td>0.85</td>
<td>0.81</td>
</tr>
<tr>
<td>T</td>
<td>0.72</td>
<td>0.84</td>
<td>0.78</td>
<td>1.58</td>
<td>1.08</td>
<td>1.21</td>
</tr>
<tr>
<td>N</td>
<td>0.74</td>
<td>0.77</td>
<td>0.73</td>
<td>0.71</td>
<td>0.76</td>
<td>0.63</td>
</tr>
<tr>
<td>H</td>
<td>0.90</td>
<td>0.85</td>
<td>0.87</td>
<td>1.15</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>D</td>
<td>0.91</td>
<td>0.73</td>
<td>0.82</td>
<td>0.61</td>
<td>0.76</td>
<td>0.55</td>
</tr>
<tr>
<td>K</td>
<td>1.25</td>
<td>1.13</td>
<td>1.17</td>
<td>1.14</td>
<td>0.98</td>
<td>0.76</td>
</tr>
<tr>
<td>E</td>
<td>1.51</td>
<td>1.25</td>
<td>1.39</td>
<td>0.89</td>
<td>0.86</td>
<td>0.65</td>
</tr>
<tr>
<td>R</td>
<td>1.31</td>
<td>1.13</td>
<td>1.21</td>
<td>1.27</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>G</td>
<td>0.28</td>
<td>0.59</td>
<td>0.44</td>
<td>0.41</td>
<td>0.81</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Figure 3.15: Least squares fit lines to scatter plots (the actual points not shown) of $\varphi_{i+1}$ (X axis) versus $d_{O_iO_{i+1}}$ (Y axis) organized by the residue type located at the $i$th position (left) and $(i+1)$ (right) when the $i$th backbone carbonyl oxygen has a primary contact with water, i.e. $d_i \in [0, 3.5]$Å.

<table>
<thead>
<tr>
<th>Data</th>
<th>mean $\varphi$</th>
<th>mean $\psi$</th>
<th>mean $d_{O_iO_{i+1}}$</th>
<th>PWC</th>
<th>SWC</th>
<th>AA Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4C3N11</td>
<td>-65.3 ± 0.9</td>
<td>-42.0 ± 0.7</td>
<td>3.50 ± 0.05</td>
<td>5</td>
<td>1</td>
<td>AAAaAAa</td>
</tr>
<tr>
<td>4C3N</td>
<td>-61.0 ± 0.4</td>
<td>-44.7 ± 0.6</td>
<td>3.45 ± 0.04</td>
<td>0</td>
<td>0</td>
<td>aaaaaaa</td>
</tr>
<tr>
<td>CPCM</td>
<td>-61.3 ± 0.4</td>
<td>-44.5 ± 0.8</td>
<td>3.45 ± 0.02</td>
<td>0</td>
<td>0</td>
<td>aaaaaaa</td>
</tr>
</tbody>
</table>

Table 3.3: Relevant results from [50]

### 3.3.2 Comparison of our results to the results from [50].

Our results are comparable with the results of simulation performed in [50] using density functional theory, see Table 3.3.

### 3.4 Solvation patterns in $\alpha$-helices

#### 3.4.1 Solvation dependent amino acid composition in $\alpha$-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
Figure 3.16: Least squares fit lines to scatter plots (the actual points not shown) of $\varphi_{i+1}$ (X axis) versus $d_{O_i O_{i+1}}$ (Y axis) organized by the residue type located at the $i$th position (left) and $(i + 1)$ (right) when the $i$th backbone carbonyl oxygen has a secondary contact with water, i. e. $d_i \in [3.5, 5)\AA$.

Figure 3.17: Least squares fit lines to scatter plots (the actual points not shown) of $\varphi_{i+1}$ (X axis) versus $d_{O_i O_{i+1}}$ (Y axis) organized by the residue type located at the $i$th position (left) and $(i + 1)$ (right) when the $i$th backbone carbonyl oxygen has no contact with water, i. e. $d_i \in [5, 10)\AA$. 
studied. The middle residues however most often are either characterized together, or amino acid propensities for pairs \((i, i + k)\), where \(k = 1, \ldots, 4\) are considered in connection with main-chain main-chain hydrogen bond, salt bridge, or other commonly classified interaction irrespective of their 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 \(\alpha\)-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 \(\alpha\)-helix see Figures 3.18-3.20. For statistics of hydrated helices at the termini see Figure 3.21. For solvation of full helices of various lengths see Figures 3.22-3.25.
Figure 3.18: Backbone solvation pattern in α-helix. Here we observe a distinct periodic pattern. This pattern also implies that most helices assume specific orientation.
Figure 3.19: Backbone solvation pattern in $\alpha$-helix for middle residues.

Figure 3.20: Backbone solvation patterns in $\alpha$-helix at the (a) N-terminus, and (b) C-terminus.
Figure 3.21: 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 3.22: 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 3.23: 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 12.
Figure 3.24: 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 15.
Figure 3.25: 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 16.
Figure 3.26: Histogram of α-helix lengths in high resolution protein dataset.
<table>
<thead>
<tr>
<th>Length</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>242</td>
</tr>
<tr>
<td>6</td>
<td>1997</td>
</tr>
<tr>
<td>7</td>
<td>1266</td>
</tr>
<tr>
<td>8</td>
<td>1368</td>
</tr>
<tr>
<td>9</td>
<td>1544</td>
</tr>
<tr>
<td>10</td>
<td>1467</td>
</tr>
<tr>
<td>11</td>
<td>1617</td>
</tr>
<tr>
<td>12</td>
<td>1639</td>
</tr>
<tr>
<td>13</td>
<td>1533</td>
</tr>
<tr>
<td>14</td>
<td>1304</td>
</tr>
<tr>
<td>15</td>
<td>1484</td>
</tr>
<tr>
<td>16</td>
<td>1118</td>
</tr>
<tr>
<td>17</td>
<td>853</td>
</tr>
<tr>
<td>18</td>
<td>741</td>
</tr>
<tr>
<td>19</td>
<td>651</td>
</tr>
<tr>
<td>20</td>
<td>479</td>
</tr>
<tr>
<td>21</td>
<td>371</td>
</tr>
<tr>
<td>22</td>
<td>274</td>
</tr>
<tr>
<td>23</td>
<td>219</td>
</tr>
<tr>
<td>24</td>
<td>164</td>
</tr>
<tr>
<td>25</td>
<td>145</td>
</tr>
<tr>
<td>26</td>
<td>95</td>
</tr>
<tr>
<td>27</td>
<td>90</td>
</tr>
<tr>
<td>28</td>
<td>63</td>
</tr>
<tr>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>33</td>
<td>40</td>
</tr>
<tr>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>35</td>
<td>9</td>
</tr>
<tr>
<td>≥36</td>
<td>87</td>
</tr>
</tbody>
</table>

Table 3.4: Helix length statistics for High Resolution Protein Dataset.
<table>
<thead>
<tr>
<th>Helix length</th>
<th>Top binary solvation patterns</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>01</td>
<td>668</td>
</tr>
<tr>
<td></td>
<td>00</td>
<td>525</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>114</td>
</tr>
<tr>
<td>11</td>
<td>000</td>
<td>580</td>
</tr>
<tr>
<td></td>
<td>001</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td>010</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>011</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>41</td>
</tr>
<tr>
<td>12</td>
<td>0000</td>
<td>354</td>
</tr>
<tr>
<td></td>
<td><strong>0010</strong></td>
<td>261</td>
</tr>
<tr>
<td></td>
<td>0001</td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>0101</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>0100</td>
<td>114</td>
</tr>
<tr>
<td>13</td>
<td>00001</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td><strong>00101</strong></td>
<td>199</td>
</tr>
<tr>
<td></td>
<td>00000</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>01001</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>01101</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>00100</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>01000</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>00010</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>10001</td>
<td>48</td>
</tr>
<tr>
<td>14</td>
<td>000001</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>000000</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td><strong>001001</strong></td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>010000</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>001000</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>000101</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>001101</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>000010</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>010010</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>011000</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>010001</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>000100</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>011001</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>100000</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 3.5: Helix solvation patterns for middle residues in High Resolution Protein Dataset. Part 1.
<table>
<thead>
<tr>
<th>Helix length</th>
<th>Top binary solvation patterns</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>00000000 00100000 00100100 00000100 0110000 0000001 0010000 0010001 0100010 0100001 0110010 0011000 0000100 0011000 0000001</td>
<td>180 79 76 67 57 55 51 46 42 39 37 36 35 28</td>
</tr>
<tr>
<td>16</td>
<td>00000000 00000001 00100101 00000101 00100001 00100110 00000010 01100101 00100000 00100011 00000000 00000000 00000010 00000010 00000010</td>
<td>68 67 42 35 29 28 27 25 24 22</td>
</tr>
<tr>
<td>17</td>
<td>0000000001 000000000 001001001 001000001 001001101 001001101 000001001 001000101 000001001 000001001 0000100110 0000110001 0000010010 0010001000 0010001000</td>
<td>43 33 24 22 21 21 19 17 16 15 14</td>
</tr>
<tr>
<td>18</td>
<td>00000000000 0000000001 0000001000 1100000000 0001001101 000001001 0010001001 0010001001 0001000001 0000100010 0010001018</td>
<td>23 21 19 16 16 12 12 12 10 10</td>
</tr>
</tbody>
</table>

Table 3.6: Helix solvation patterns for middle residues in High Resolution Protein Dataset. Part2.
3.5 Conclusion and further plans

Main Conclusions:

1. Strong dependence between hydration of carbonyl oxygen $O_i$ and value of $\varphi_{i+1}$.

2. There is a very clear dependence of $d_{OiO_{i+1}}$ on $\varphi_{i+1}, \psi_{i+1}$ in ideal geometry, see Figure 2.5, and 3.3. Upon examining PDB data, see Figure 3.2 for the definition of parameters, we found that distance $d_{OiO_{i+1}}$ is strongly correlated with the value of $\varphi_{i+1}$, see Figures 3.6, 3.8, 3.11, and the interaction of backbone carbonyl $O_i$ with water, see Figures 3.10, 3.12. Also, this correlation strongly depends on the $(i, i + 1)$-dipeptide’s amino acid composition, see Figures 3.13, 3.15, 3.16, 3.17.

3. It is known that water positions are largely underestimated in PDB. Using the above results we want to predict where backbone carbonyl in a helix interacts with water but no information in PDB is recorded.

3. Order of amino acids in see Figure 3.15 coincides with the propensities in the Table 3.14.

4. See Tables 3.4, 3.5, 3.6, and Figures 3.21, 3.22, 3.23, 3.24, 3.25 for helix solvation statistics. There is a repeating pattern that needs further exploration.

Points to expand and directions to explore:

0. Helix nucleation vs propagation.

0. Chameleon sequences.

1. In [5] authors mention that “solvent (referring to [49]) and other bound molecules (referring to [31]) can participate in capping hydrogen bonds.”

2. We have noticed a discrepancy ($\pm 1$ or 2) between starting and ending positions in helices used in [5] and in the current files in PDB. So far, I found that .pdb files are being revised (continually) on the subject of compliance with PDB Exchange Dictionary.
The current version of PDB Exchange Dictionary is V4. For example, 1CLL.pdb was released in 1993, referred to in 1998 paper with helix start/end positions different from the current published version revised in 2007. My major concern is if a standard could differ from file to file in the current database and in my dataset? I need to find out this before proceeding to N- and C-termi study. I read in some papers that they recompute the helix positions using different software and omit helices that have discrepancies between PDB assignment and the software.

3. So far, we have not considered flanking residues, i.e. \( \cdots N'' - N' - C' - C'' - \cdots \) in the standard helix nomenclature \( \cdots N'' - N' - N_{cap} - N1 - N2 - N3 - \cdots - C3 - C2 - C1 - C_{cap} - C' - C'' - \cdots \). This is not so relevant for the middle residues study we are currently doing but could be useful in the study of N- and C-termi.

4. Investigate the relationship between \((i, i+4)\) and \((i, i+3)\) hydrogen bonds, and backbone carbonyl water contacts for the middle residues, also consider how this relationship varies for different amino acids (choose for example helices of length 15 that give possibilities for 3 \( \alpha \)- and 4 \( 3_{10} \)- helical internal hydrogen bonds and 7 carbonyl oxygens)

5. What is special about helices of length 6?

6. Add data for membrane proteins.

7. Add data for ice structuring proteins.

8. Calcium binding protein (1CLL.pdb)

9. Vitamin D binding protein, actin, prevention of fibril formations.

10. Antimicrobial peptides with tendencies to form amphipathic helices with antibiotic action (change membrane permeability and promote perturbation in the pathogen cell) membrane)
BIBLIOGRAPHY


