A Tale of Two Helices

A study of alpha helix pair conformations in three-dimensional space

by

Robert Malcolm Fraser

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
August 2006

Abstract

This Master's project underwent an evolution, and this thesis paper reflects that. The crux of the work involved a thorough investigation of how pairs of alpha helices are configured in three-dimensional space. The project began with a review of protein structure, and specifically the relationship between the structure of a protein and the corresponding contact map. There did not exist visualization software for alpha helices and their contact maps; Hippy was created to address this deficiency. The package allows the user to explore alpha helices by manipulating them in space while toggling features such as the side chains, van der Waal's shells, and rendering of the contacts on and off. A greater understanding of the significance of patterns in contact maps may be gained by manipulating the contact map through adjustment of the contact threshold and locating the contacts in the helix pair corresponding to the map. To further the scope and usability of Hippy, it was implemented in OpenGL with platform independence and open source code as primary goals.

There were two related properties of the configuration of helix pairs that were chosen for study: the interhelical angle and the packing attribute. Further complicating this study was a lack of a standard algorithm for the calculation of these attributes. A thorough review of the past methods of finding the axis of a helix and the interhelical angles is presented. The ideal algorithm selected for this study is a quaternion-based axis determination method followed by an angle calculation using a modified dot product. For the determination of the packing attribute, a novel geometrical approach is introduced. Using these methods, a correlation was sought between contact maps and the packing of the corresponding helices. The first method was a blind clustering of the contact maps to determine whether maps corresponding to similar packing values are placed into the same clusters, which yielded promising results. Finally, an approach searching for congruency between the packing values of the nearest neighbour of every map was successful for 99% of the maps.

Acknowledgments

First and foremost, I would like to thank my wife Kelly for her constant support through my studies. She has always been there for me. She has helped me when times were tough, we've celebrated the milestones together, and she has believed in me unwaveringly. We have had many wonderful adventures, and helping me to match her Master's degree from Queen's was one of the greatest. I look forward to the rest of our life together.

The support of my supervisors at Queen's has been wonderful. Prof. Janice Glasgow is a pleasure to work with, she provides a good balance of direction and freedom to explore the avenues you desire. She clearly holds the best interests of her students at heart, she encouraged me to publish and even gave me the opportunity to help organize a conference. Janice is a fun person as well, the lab barbecues she hosted were great, and she's guided me through a few hands at euchre at the grad club. Prof. James Stewart inspired me to undertake the visualization part of this research. His passion for his research is contagious, evidenced by the number of grad students who seek his supervision. James also provided guidance on the ice, reminding me of the importance of playing heads-up hockey. There are also many other members of Queen's faculty and staff who have helped me with this work and provided guidance during my time at Queen's, especially Henk Meijer, Hagit Shatkay, Selim Akl, Purang

Abolmaesumi, and Debby Robertson. My time at Queen's has been wonderful, and the people here have reinforced my decision to pursue a career in academics.

Finally, I must thank my family and friends. My whole family has been very supportive of me, even if they don't fully comprehend what it is I am trying to do. My parents Aura-Lee, David, Wes, Bev, Jim and Rollie (that's natural, step, and in-laws) and siblings Brian, Andrew, Nathan, Kim, Sherry and Jamie all have helped me in one way or another. I have made many new friends here at Queen's, and I'm sure many of us will cross paths again. These people helped procrastinate through softball and squash, backgammon and hockey, sudoku and novels, Beirut and Mario, the grad club and the Toucan. Given the help of my friends, I know several faculty thought this thesis would never be completed. I would list you all off, but really, none of you will have read this far into my thesis anyway :-).

Thanks everybody!

Glossary

Ab initio *Ab initio* describes the method of protein structure prediction that is based (conventionally) on purely physical properties of proteins. From Latin, *ab initio* means from the beginning.

Amino Acid Amino acids are the building blocks of proteins.

Angstroms (Å) The Ångstrom is the conventional unit of distance at the atomic scale. One Ångstrom is equivalent to 0.1 nanometers, and is about twice the diameter of a hydrogen atom.

CASP Critical Assessment of Structure Prediction. A biannual event in which researchers are able to test their protein structure prediction algorithms on proteins with coordinate data that is unknown to the community but which has recently been determined by conventional means. This enables a blind test of the algorithms.

cis A *cis* configuration along the backbone of a polymer is one where the previous and following chain bonds are on the same side of a plane aligned with the bond of interest. This term only applies when there is a planar region present. From Latin, *cis* means on the same side.

CMO The Contact Map Overlap problem. CMO addresses the challenge of determining the best algorithm for aligning and comparing contact maps.

Contact Map A contact map is an $N \times N$ matrix, where N is the number of amino acids in the given protein, and entry C_{ij} in the matrix indicates whether amino acid residue i of the protein is within some threshold distance of residue j.

Euler angle Euler angles are conventionally used for performing rotations. The angles are used to transform points from one coordinate system to another.

Force-based Modelling Force-based modelling is a paradigm where a model of a molecule includes the interatomic forces present. These include conventional bonds, van der Waal's forces and electrostatic effects. A change in the position of one atom will affect the positions of many if not all of the other atoms present.

Globular Protein A globular protein is one that forms a (very roughly) spherical shape. They are soluble in water, and thus tend to have hydrophilic surfaces and hydrophobic cores. All proteins used in this thesis work were globular.

GSCA GSCA is the global segment of closest approach. When dealing with two skew lines of infinite length, the GSCA is the line that intersects both lines at a right angle. The GSCA is unique unless the lines are parallel. See also SCA.

Hippy Hippy is a helix pair visualization software package that was developed to allow the user to interact both with the helices and the contact map.

Hole Hole in the context of this thesis refers to the volume of space around that occupied by the side chains of an amino acid residues. It is due to Crick's 'knobs into holes' packing model. See also knobs.

Homology Modelling Homology modelling is an approach to protein structure prediction where databases are searched for regions of sequence similar to that of the unknown structure. The proposed structure can be built using the recovered structural segments.

Hydrophilic Hydrophilic in the context of this thesis refers to parts of molecules that have an affinity for water. It is from Greek, literally meaning 'water friend' or 'water loving'.

Hydrophobic Hydrophobic in the context of this thesis refers to parts of molecules that are repelled by water. It is from Greek, literally meaning 'water fearing'.

Knob Knob in the context of this thesis refers to the volume of space occupied by the side chain of an amino acid residue. It is due to Crick's 'knobs into holes' packing model. See also holes.

Lennard-Jones equation The Lennard-Jones equation approximates the van der Waal's forces between two atoms using empirically derived constants for each atom species.

Native State The native state of a protein is the structure as it is found in nature, usually in the cell.

NMR Nuclear magnetic resonance spectroscopy. It is a traditional means of molecular structure determination.

Omega (Ω) Ω is the conventional symbol for representing the interhelical angle.

Omega (ω) ω is the conventional symbol for representing the bond angle between the N and C_β atoms in the protein backbone.

Packing Classes The determination and subsequent classification of preferential packing angles results in packing classes.

PDB The Protein Data Bank. A repository for protein coordinate files.

Phi (ϕ) ϕ is the conventional symbol for representing the bond angle between the C_α and N atoms in the protein backbone.

Primary Structure Also sometimes referred to as the sequence, this is essentially an ordered listing of the amino acids that compose the protein.

PSI-BLAST PSI-BLAST is a essentially a lookup table that assigns costs to each amino acid species for a substitution for any other.

Psi (ψ) ψ is the conventional symbol for representing the bond angle between the C_α and C_β atoms in the protein backbone.

Quaternary Structure The quaternary structure of a protein describes the three-dimensional structure of the protein when multiple strands come together to form a complex. This does not occur in all proteins.

Quaternion A quaternion is a description of a rotation which requires only four complex values, represented as a scalar and a vector.

Ramachandran plot A graph showing the sterically acceptable configurations of the ψ and ϕ angles on the protein backbone.

Residue An amino acid that has bonded to become part of a peptide chain is referred to as a residue (because a water is produced as a byproduct during polymerization).

rmsd Root mean squared deviation. This is a common tool for measuring the difference between two sets of points, such as two protein structures or two helices. It is usually expressed in Å for proteins.

Rotamer A rotamer is a particular configuration in three-dimensional space taken by the side chain of an amino acid.

SCA SCA is the segment of closest approach. When dealing with two skew line segments, the SCA is the shortest line that intersects both lines. The SCA is equal to the GSCA for a pair of line segments if the SCA intersects the line segments at right angles. See also GSCA.

Secondary Structure The secondary structure of a protein describes the local structures that are adopted, such as alpha helices, beta sheets, and turns. Hydrogen bonding is the predominant factor in establishing secondary structure.

Side Chain The side chain of an amino acid is the atoms that are bonded to the C_α. It essentially defines the species of the amino acid. See also rotamer.

Supersecondary Structure A supersecondary structure describes the configuration of two or more secondary structures.

Tertiary Structure The tertiary structure of a protein is the three-dimensional structure of a single strand of the protein.

TM An abbreviation for Transmembrane proteins.

trans A *trans* configuration along the backbone of a polymer is one where the previous and following chain bonds are on opposite sides of a plane aligned with the bond of interest. This term only applies when there is a planar region present. From Latin, *trans* means across.

van der Waal's The van der Waal's shell of an atom approximates the steric surface of the atom. The shells of two non-bonding atoms are most stable when just touching. At closer distances, there is a repellent van der Waal's force, while at greater distances the van der Waal's forces are attractive. See also the Lennard-Jones equation.

Contents

Abstract i

Acknowledgments iii

Glossary v

Contents ix

List of Tables xi

List of Figures xii

1 Introduction 1

1.1 The interhelical angle 2

1.2 Visualization software 3

1.3 Hypotheses 4

1.4 Organization of Thesis 5

2 Protein Structure 6

2.1 Elements of Protein Structure 7

2.1.1 The Structural Hierarchy 7

2.1.2 The Alpha Helix 16

2.1.3 Globular vs. Transmembrane Proteins 22

2.1.4 The PDB 22

2.2 Contact Maps 25

2.3 Protein Structure Prediction 27

2.3.1 *ab initio* Modelling 30

2.3.2 Homology Modelling 32

2.3.3 Triptych 33

2.3.4 Contact Map Prediction 36

2.3.5 CASP 36

2.4 Protein Visualization Packages 37
2.4.1 Force-based Modelling of Proteins 40

3 The Hippy Visualization Package 45

4 The Interhelical Angle 52
4.1 Helix Packing Models 52
4.1.1 Knobs into Holes 53
4.1.2 Hydrophobic Core Packing 57
4.1.3 Ridges into Grooves 59
4.1.4 Packing of Surface Contact Areas 62
4.1.5 Lattice Superposition Model 64
4.1.6 Other Models 64
4.2 Observed Packing Preferences 65
4.3 Calculating the Interhelical Angle 72
4.3.1 Axis Calculation Approaches 72
4.3.2 Angle Calculation Approaches 83
4.3.3 Other Approaches 93
4.3.4 The Proposed Method 94
4.4 Analysis of the Interhelical Angle Algorithm 100

5 Clustering Contact Maps to Predict Packing 104
5.1 Distance Metrics 105
5.2 Source Data 108
5.3 Clustering Analysis 111
5.4 Clustering Results 114

6 Summary and Conclusions 117
6.1 Conclusions 117
6.2 Future Work 119

Bibliography 121

A Implementing Hippy 134

B Organic Chemistry Primer 138
B.1 Polymerization 138
B.2 *cis-trans* Isomerism 139

List of Tables

2.1 Alpha helix properties 20

4.1 The test proteins 100
4.2 Angle calculation test results 101

5.1 The contact maps used for clustering 111
5.2 The clustering results for packing 115
5.3 The effects of varying parameters on clustering 116

List of Figures

2.1 The structure of an amino acid 9
2.2 The 20 amino acids 10
2.3 Peptide bond angles and lengths 11
2.4 Ramachandran plot of bond angles 13
2.5 Primary structure 14
2.6 Beta sheets 15
2.7 Tertiary and Quaternary Structure 17
2.8 Structure of the alpha helix 19
2.9 Alpha helix cross section 20
2.10 A sample PDB file 24
2.11 Deriving the contact map from the distance map 26
2.12 Contact map with a larger threshold 28
2.13 Finding the contact map interface 29
2.14 The case-based reasoning framework 34
2.15 SCULPT modelling software 39
2.16 Cartoon modelling of the backbone 44

3.1 The helix selection window 47
3.2 Hippy 48
3.3 Hippy's command window 49
3.4 Hippy's contact map window 50

4.1 A parallel helix pair 55
4.2 Knobs and holes on a helix 56
4.3 Side chains are not perpendicular to the axis 57
4.4 An $\alpha - \alpha$ Corner 58
4.5 Unrolling of a helix 60
4.6 Packing of unrolled helices 61
4.7 Angle probabilities 67
4.8 The theoretical distribution 68
4.9 The corrected distribution of interhelical angles 69

4.10 Classes of interhelical configurations 71
4.11 Local helix axis vectors 76
4.12 Orthonormal vectors for an alpha carbon 78
4.13 Incorrectly calculating τ 88
4.15 The closest approach method for interhelical angle 91
4.14 Geometrical approach to τ 92
4.16 Dot product or dihedral angle? 95
4.17 Calculation of the interhelical angle 97

5.1 The different classes of contact map for comparison 108
5.2 Obtaining two contact maps from one 110

A.1 Hippy's UML diagram 135

B.1 *cis*-*trans* isomerism 140

Chapter 1

Introduction

We've discovered the secret of life.

-Francis Crick

The problem of predicting three-dimensional protein structure from the primary amino acid sequence is one that has been the pursuit of researchers for decades [Cri53, CLR81, KVFM05]. The initial sentiment was that the structure could be found *ab initio*, meaning that given the primary sequence of the protein, the secondary and tertiary structure could be assembled incrementally using first principles. The geometry of proteins is well understood [Anf73, CLR81]; physics is sufficiently advanced to allow us to predict probabilistic low-energy configurations of the protein, and we have a good sized body of knowledge in the Protein Data Bank (PDB) [BWF+00] to provide experiential evidence. Although 25 years ago the problem seemed surmountable, it is still considered a hard problem today.

The primary structure of a protein is the amino acid sequence that defines it. The secondary structure considers local three-dimensional configurations that may appear in the structure, such as alpha helices or beta sheets. The tertiary structure is the three-dimensional shape that the entire protein string assumes; the quaternary structure includes the configuration assumed when multiple protein strands come together (this does not occur in all proteins). Furthermore, there are other structures known as supersecondary structures, which contain multiple secondary structures, which describe the configuration adopted by the secondary structures in 3D space.

1.1 The interhelical angle

The present study is concerned with only one type of supersecondary structure, specifically alpha helix pairs. The ability to predict the configuration of pairs of alpha helices would be an asset for protein structure prediction. It is highly likely that the 3D structure of a protein is primarily determined by the supersecondary structures present [VKD97, HSS$^+$02]. Therefore, given the ability to predict the configurations of alpha helix pairs, similar studies could be done for the other supersecondary structures (alpha helix - beta sheet, beta sheet - beta sheet, etc., see Sun et al. [SRPX97] for a review), and ultimately the dream of producing the tertiary structure from the primary sequence may become possible by assembling these results. Perhaps the most significant single characteristic of the alpha helix pair is the interhelical angle (or interaxial angle, for the purposes of this thesis, these are considered to be synonymous).

This thesis proposes a straightforward approach to the determination of the interhelical angle for a pair of alpha helices. The proposed algorithm uses the contact map for the protein as part of the algorithm, thus the algorithm does not work for helices that are not in contact with each other. This is considered appropriate for most applications, as those alpha helices which are in contact are the ones that form supersecondary structures and thus form the basis of the tertiary structure of the protein.

1.2 Visualization software

There exists a myriad of software for modelling proteins, most of which accept the PDB file for the protein as input and extract the relevant information from the file so the user can customize the view for the properties relevant to their studies. A few of the more popular packages are Chimera [PGH^{+}04], Swiss-PdbViewer [GP97] and Protein Explorer [Mar02]. The limitation of these packages with regards to the applications required for this research is that the displays are tailored for the entire protein and do not provide much insight into the interaction between the alpha helices. There was a clear need for software which allows a clear view of helix pairs not only for this study, but for anyone investigating alpha helices in general. The interhelical angle is an example of a relatively simple property that is often difficult for new researchers to fully understand. For some people, it may seem intuitive (although it is erroneous) that the angle could be reduced from a range of -180° to 180° to just 0° to 180° or something similar. None of these modelling packages facilitate the investigation of the interhelical angle, as they are primarily used to model the entire protein. It is possible to isolate the supersecondary structures, but it requires some effort since it is not what the packages were designed for.

None of the modelling systems that were found incorporated contact maps. This was deemed to be a major deficiency in the area, and forms the niche where the Hippy software package is required. This thesis is focused on the creation of a specialized graphics application for a particular problem domain. This is not an unusual problem.

Another example is the Ptuba software package [LSD05], developed specifically to model the surfaces of alpha helices. There are so many facets to the structure of a protein that it is impossible for one graphics package to address all possible needs that a researcher may have, which necessitates specialized packages such as Hippy.

1.3 Hypotheses

The creation of a specialized visualization software package dedicated to the investigation of alpha helices in proteins and the contact maps associated with them will advance the ability to derive fundamental properties of proteins. This project entails a number of subgoals:

- the investigation of the abilities and limitations of existing software packages;
- the review of the packing of alpha helices;
- the development of an algorithm for calculating the interhelical angle of a pair of alpha helices;
- the evaluation of this algorithm with respect to existing approaches;
- the development of the modelling package such that:
    - PDB files can be opened and a pair of alpha helices is isolated immediately;
    - the contact map for the alpha helix pair is displayed along with the graphical model;
    - the correlation between a contact in the map and the corresponding atoms in the model is clear;

Finally, it is proposed that properties of the three-dimensional configuration of a pair of alpha helices may be predicted from the contact map corresponding to the pair.

1.4 Organization of Thesis

We proceed by discussing protein structure and contact maps in the next chapter. Chapter 3 describes the Hippy visualization package in detail. Details regarding the implementation of Hippy are presented afterward in Appendix A. We discuss interhelical angles and the algorithm to solve interhelical angles in Chapter 4. We test the ability to predict a property of the configuration of helix pairs from contact maps in Chapter 5. In Chapter 6, the conclusions of this thesis work are presented, and possibilities for future work are outlined.

Chapter 2

Protein Structure

In all things of nature there is something of the marvelous.

-Aristotle

Proteins are organic molecules which reside at the centre of most of the functions of life. They form the majority of the constituents of the cell, and participate in operations such as metabolism and immunological defense as well as forming the basis of cellular structural elements [Lig74, BD05]. The structure that the protein adopts in the cell is known as the native state of the protein. The structures of proteins are complex, and the standard approach to the determination of the three-dimensional structure is tedious and time consuming, often requiring years for a single protein. We begin our study with an examination of the structure of protein in section 2.1. We discuss contact maps in section 2.2, followed by the history of protein structure prediction in section 2.3.

2.1 Elements of Protein Structure

Petsko and Ringe [PR04] sum up the basis of protein structure prediction succinctly: sequence determines structure, which in turn determines function. The work in this thesis is not involved with the latter portion of this dogma, but the desire to determine structure from sequence is at the very heart. This section of the thesis expands on the concepts used in this work. We explore the different levels of the protein structural hierarchy, followed by an in-depth look at the alpha helix. A summary of the PDB, the standard repository of protein structures, concludes our examination of protein structure.

2.1.1 The Structural Hierarchy

Proteins have several discrete levels of structural complexity, termed primary, secondary, tertiary and quaternary. Each level adds another layer of complexity to the structure of the protein. We examine these in turn, beginning with the primary structure, but it would be prudent to first discuss amino acids, the building blocks of proteins.

Amino Acids

The structure of an amino acid is shown in Figure 2.1. The backbone of the protein is formed when the amino end of one amino acid reacts with the carboxyl end of another, which results in a covalent bond between the nitrogen atom of the first and the beta carbon of the second. A water molecule is produced as a by-product of the reaction. Since the amino acid loses the water molecule in the reaction, the amino acid in the polymeric state is referred to as an amino acid residue, or often just as a residue for short.

The side chain portion of the amino acid can take many different forms, as shown in Figure 2.2, and is the characteristic which defines the species of amino acid. There is an entire study dedicated to the conformations adopted by the side chains in space, known as rotamers. For any bond in a molecule, and in this case in proteins in particular, there are bond angles and lengths that are typically adopted (as shown in Figure 2.3). For the side chain of an amino acid this can be expressed as a class of configurations, and a probability of occurrence in native proteins can be associated with each configuration (see for example [LWRR00]). The nature of the molecules forming the primary sequence determine the higher order structures in the molecule [PR04], so understanding the different rotamer classes is potentially of critical importance to protein structure prediction.

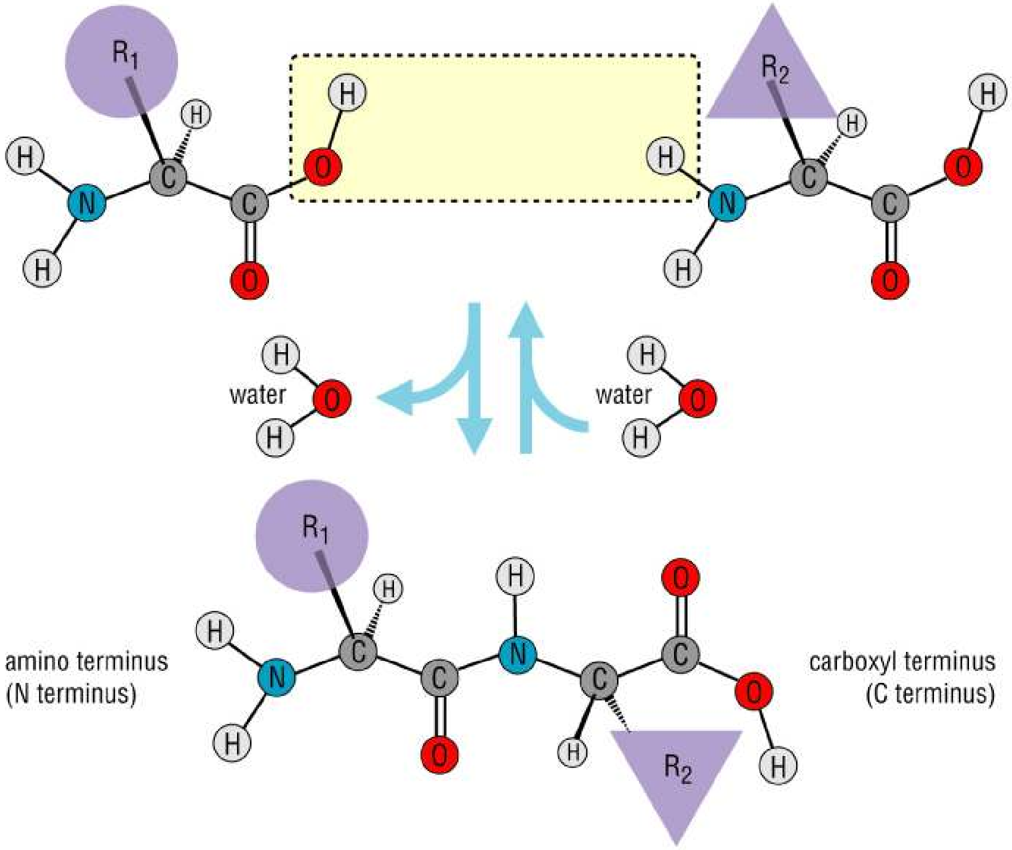


Figure 2.1: The images at the top represent two amino acid monomers. The amino group of one monomer reacts with the carboxyl group of another monomer (or peptide fragment), and a water molecule is a by-product of the reaction. R_1 and R_2 correspond to the side chains of each amino acid. The figure is reproduced with permission from [PR04].

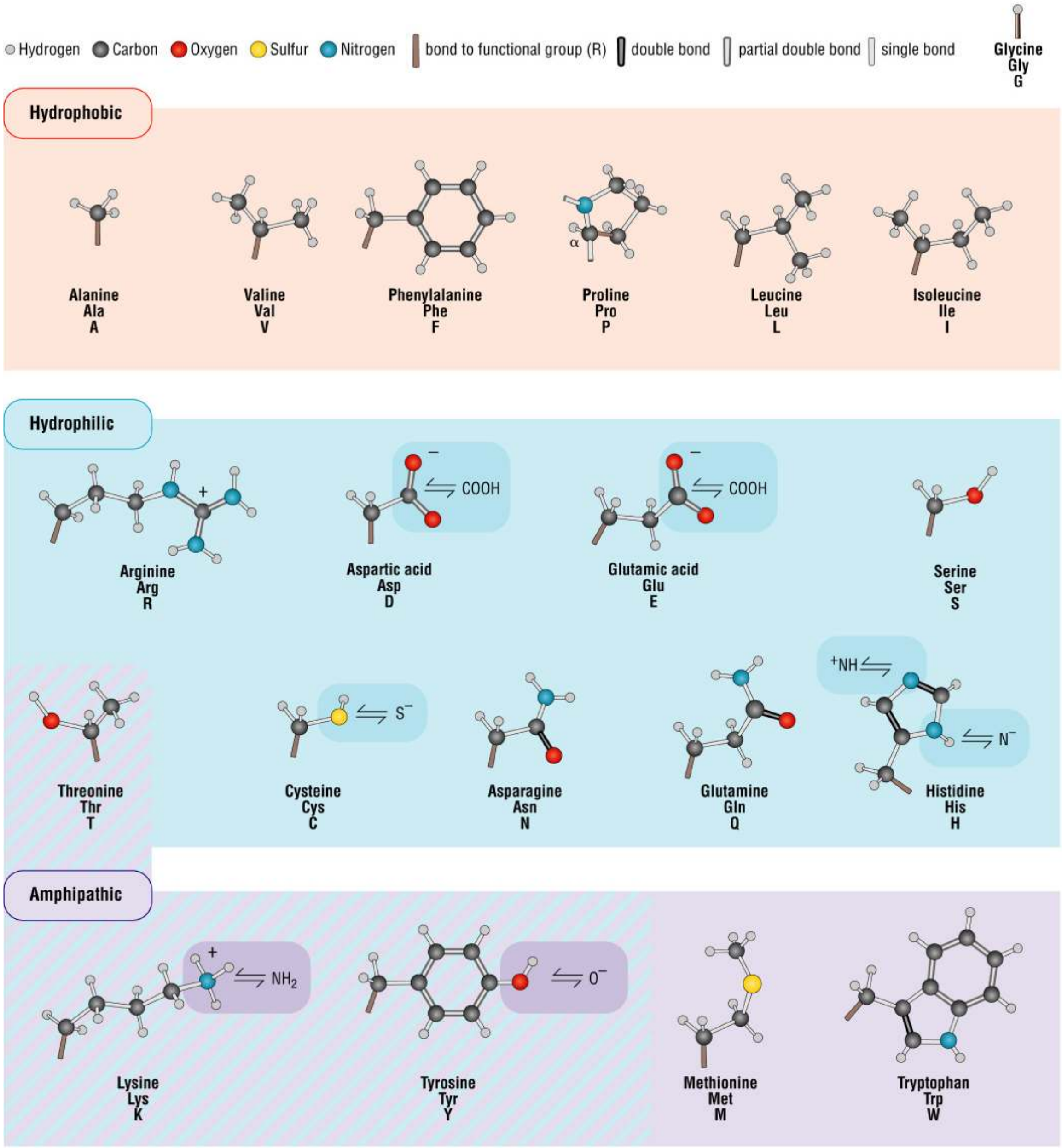


Figure 2.2: These are the 20 amino acid species found in nature. They are separated in this figure according to hydrophobicity. Notice in particular some special cases: Glycine, in the top right, is just a hydrogen atom and not really a side chain at all; Proline actually bonds to both the C_β and C_α atoms. This figure has been modified with permission from [PR04].

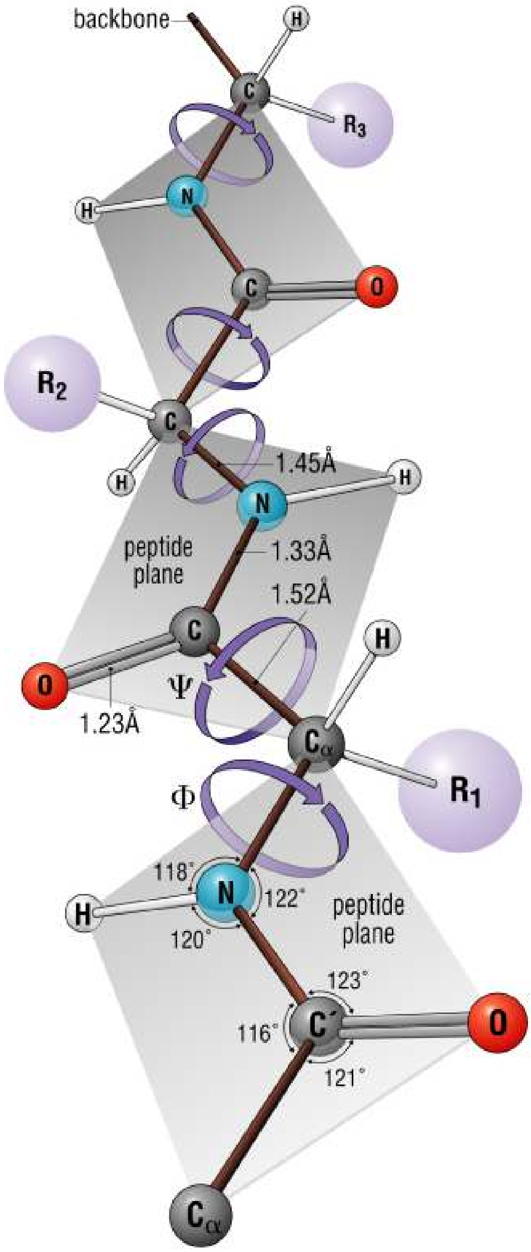


Figure 2.3: The polypeptide chain is illustrated to give the impression of the three dimensional configuration of the chain. Typical bond lengths and angles are shown (except ω, which is the dihedral angle in the peptide plane). The side chains are represented by spheres, where the sphere labelled R_i is meant to represent the side chain for residue i. This image is reproduced with permission from [PR04].

Along the backbone of the protein, the bond angles that are found are very regular and classifiable. This is most clearly illustrated with a Ramachandran plot, as shown in Figure 2.4. The angles in the backbone are referred to as the Psi (ψ), Phi (ϕ) and Omega (ω) angles. There is a planar region in the backbone, which is known to have regular properties for the bond lengths and angles, as shown in Figure 2.3. The ω angle is at the heart of the planar region, which describes the angle of rotation around the amide bond between the nitrogen atom and the alpha carbon atom in the backbone. The angle is rarely found at anything other than $0°$ or $180°$, corresponding to the *cis* and *trans* configurations respectively, and a heavy preference for the *trans* configuration is observed (except in the case of proline) [Lea01]. Refer to Appendix B for a brief review of organic chemistry concepts if desired.

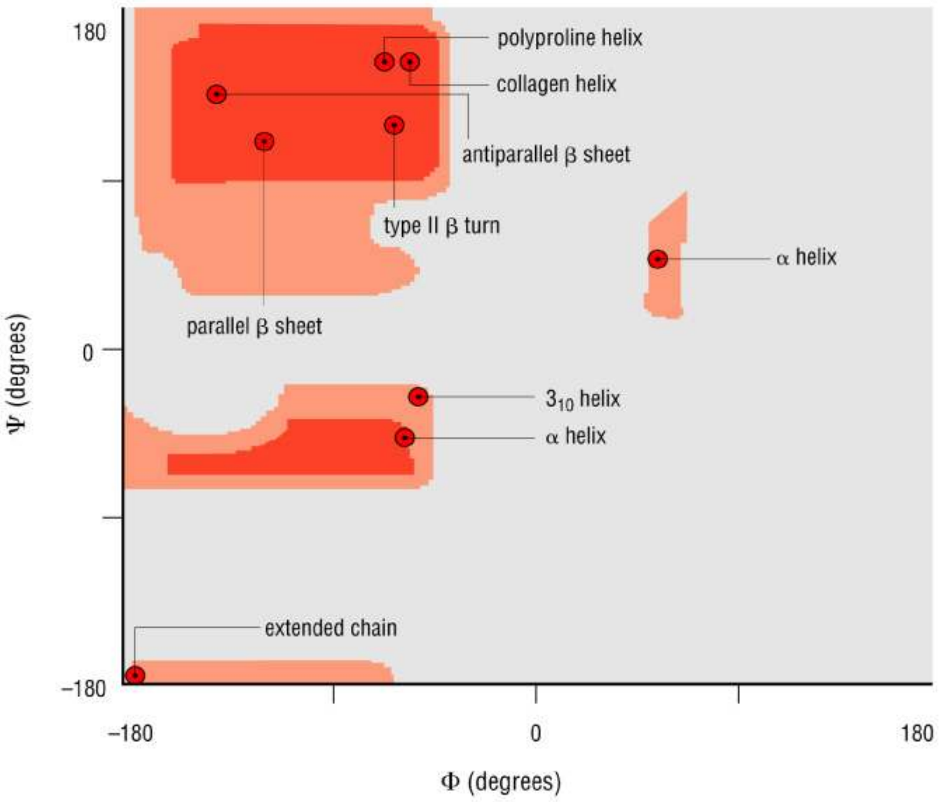


Figure 2.4: The ψ and ϕ angles found in proteins are plotted on this graph. The majority of the angles are not often seen in nature [RS68], and the areas of the plot which are characteristic of different structures are indicated. Darker regions of the figure correspond to preferred configurations. This image is reproduced with permission from [PR04].

Primary Structure

A protein is a polymer molecule constructed by assembling a sequence of amino acid monomers (also referred to as a polypeptide) [GGML99]. Proteins are constructed from roughly 100 - 1000 amino acids [EJT04], although numbers outside this range are not uncommon. The primary structure is often simply expressed as a sequence of characters, each corresponding to an amino acid as shown in Figure 2.2. In terms of organic chemistry, the primary sequence is a complete description of the molecule, as it describes the entire atomic constituency of the molecule [Lig74]. Two different proteins will have two different primary sequences, as a given primary sequence corresponds to a particular protein structure. There is a defined order to the primary sequence, as adjacent amino acids are bonded covalently along the protein backbone, and the order is given from the amino end of the molecule to the carboxyl end, which corresponds to the 5' to the 3' in the encoding genes[1], as shown in Figure 2.5.

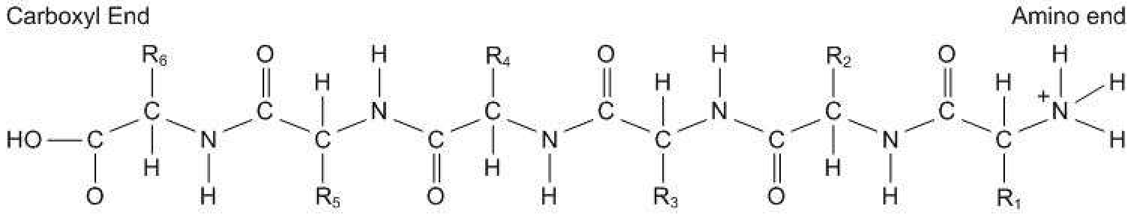


Figure 2.5: This illustrates the primary structure of a protein schematically. R_k simply represents the k^{th} side chain. This figure has been adapted from [GGML99].

Understanding that the primary sequence fundamentally entails an ordering is vital, because vectors representing a portion of the protein will have a defined directionality that is derived from the primary sequence.

Secondary Structure

The secondary structure defines local three-dimensional configurations that may appear in the structure, such as alpha helices or beta sheets. Secondary structures arise in large part due to interaction between non-adjacent amino acids of the peptide, predominantly in the form of hydrogen bonding between the polar groups of each amino

[1] For DNA and RNA molecules, one end of the molecule is referred to as the 3' and the other the 5'. Nucleic acids are synthesized in the 5' to 3' direction, so they are conventionally written in this direction [GGML99].

acid residue [PR04, EJT04]. The methods for determining these structures from primary structure, contact maps, or other means are robust, see [VKD97] and [Jon99] for details. The structure of beta sheets is shown in Figure 2.6, and alpha helices are discussed in much greater detail in section 2.1.2.

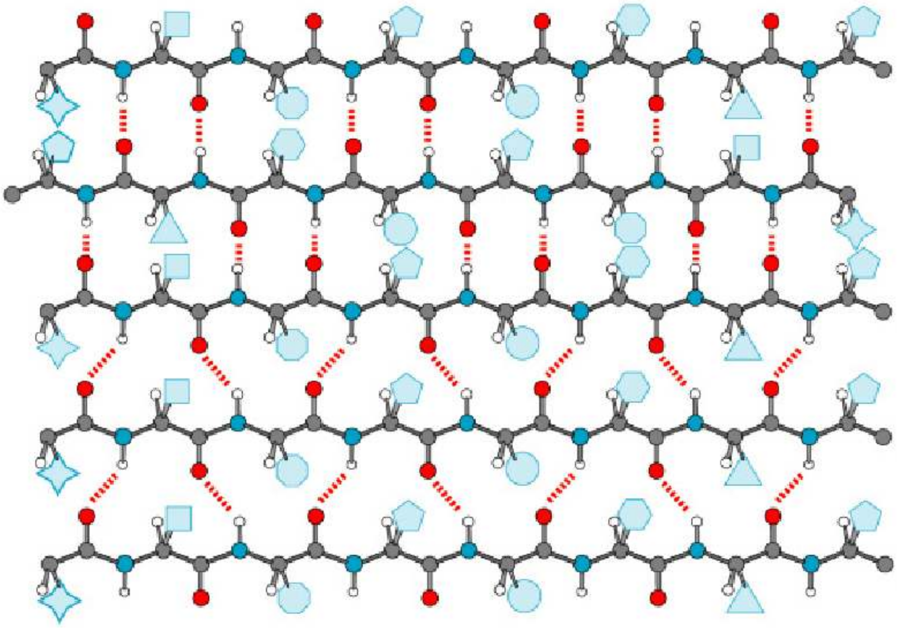


Figure 2.6: This figure illustrates the typical configurations adopted by beta sheets, one of the most common types of secondary structure. This figure has been reproduced with permission from [PR04].

Tertiary and Quaternary Structure

The tertiary structure is the three-dimensional shape that the entire protein string assumes. The structure is a result of cumulatively connecting secondary structural elements [GGML99]. The quaternary structure includes the configuration assumed when multiple protein strands come together (which does not occur in all proteins).

There are several kinds of quaternary structures. If the structure is composed of multiple identical strands, it is referred to as a homo-multimer, while hetero-multimers are composed of different protein strands bonded together. Both tertiary and quaternary structure are shown in Figure 2.7.

Supersecondary Structures

Supersecondary structures are considered to lie between the secondary and tertiary levels; my research concerns itself with this type of structure exclusively. A supersecondary structure contains multiple secondary structures, and describes the configuration adopted by the pair (or triplet, etc.) in three-dimensional space. This subject is further restricted in this study, as we will be examining only pairs of alpha helices, and investigating means of gathering information to assist the prediction of this supersecondary structure given primary structure only. It is highly likely that the 3D structure of a protein is primarily determined by its supersecondary structures [VKD97, HSS+02]. Our hypothesis is that given the capability to predict the configuration of alpha helices, similar studies can be done for the other supersecondary structures, and ultimately the dream of producing the tertiary structure from the primary sequence may become possible by assembling these results.

2.1.2 The Alpha Helix

Secondary structures are significant for the stability of a protein due to the extensive networks of the hydrogen bonding [PR04]. As mentioned above, the 3D structure of a protein is primarily determined by the supersecondary structures present [VKD97, HSS+02]. The alpha helix is the most common type of secondary structure [PR04],

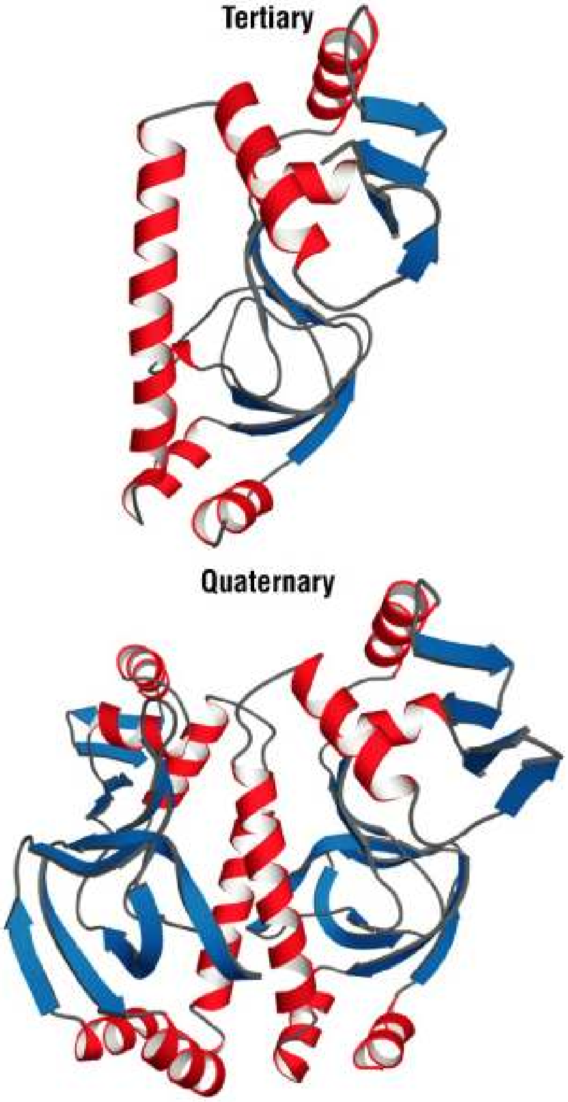


Figure 2.7: The top image shows a protein strand in cartoon format that has been folded; this three dimensional configuration is the tertiary structure. The red helices are alpha helices, and the blue arrows are beta sheets. The lower image shows two of the structures from the top image coming together to form a complex. This is the quaternary structure of the protein. This figure has been adapted with permission from [PR04].

as over a third of residues in globular proteins are found in helices [BT88], and this thesis focusses on the alpha helix for this reason. The structure of the alpha helix was first published by Pauling, Corey and Branson [PCB51], and their insights were remarkable. They identified the planar regions of the backbone. All of their bond lengths were within 0.02Å and all the bond angles were within 2° [PCB51, Eis03] of those values published here, an incredible accomplishment given that these values were derived from crude models of only four proteins and considering the natural variations possible for these values. The alpha helix is formed when the carbonyl atom in an amino acid residue forms a hydrogen bond with the amide nitrogen atom from another residue four residues further along the chain. This structure is illustrated in Figure 2.8.

The orientation of the amino acid residues in the alpha helix results in a configuration where all of the side chains are pointing away from the axis of the helix (not precisely normal however, see Figure 2.9), so that it is the side chains of the residues which form the surface of the alpha helix. The structure of the alpha helix is very regular, which facilitates modelling. The conformational parameters are summarized in Table 2.1. The planar regions of the backbone, shown earlier in Figure 2.3 on page 11, are usually nearly parallel with the helix axis [GP98]. There are other properties that arise in helices that are remarkable. For example, they tend to form dipoles and frequently one side of the helix is hydrophillic while the opposite face is hydrophobic, but such phenomena are peripheral to the task at hand. Suffice it to say, alpha helices are wholly remarkable and complex structures, worthy of significant study.

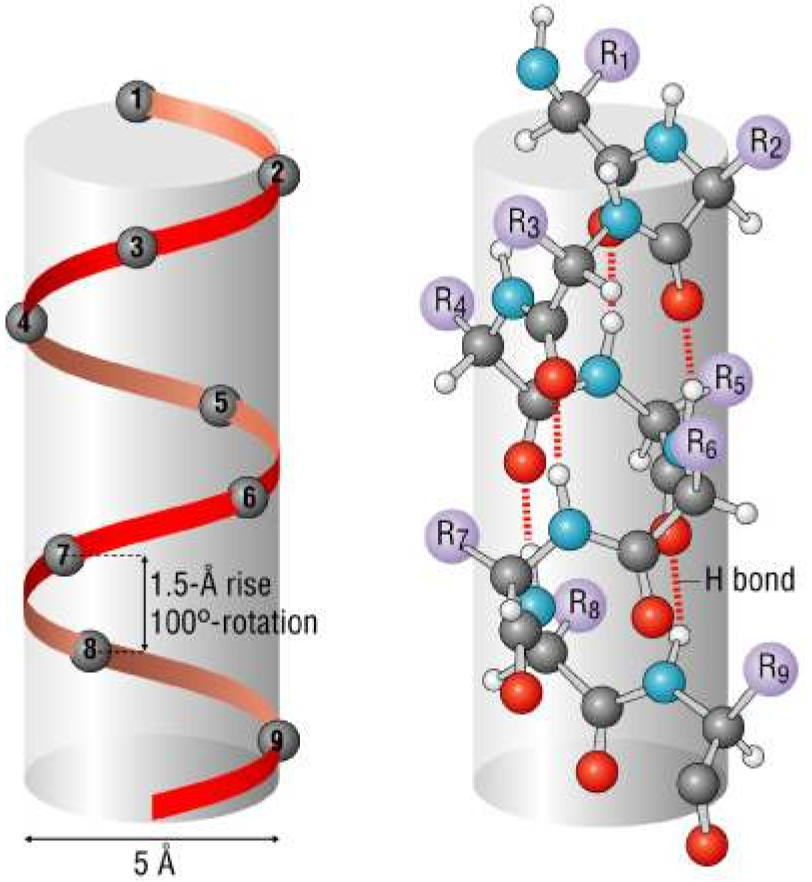


Figure 2.8: The helix on the left is a simplified helix showing the C_α atoms only. This is the model used predominantly throughout the thesis. The helix on the right shows the full structure (with side chains represented as usual), and hydrogen bonds represented as dashed red lines. This figure has been reproduced with permission from [PR04].

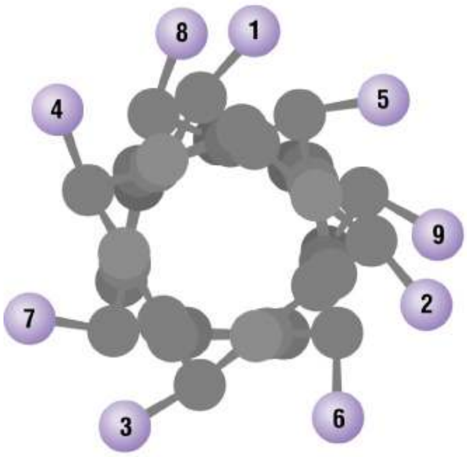


Figure 2.9: This shows the cross-section of the alpha helix. Notice in particular the angle that the side chains assume to the axis. In addition, they are slightly oriented towards the amino-terminal end of the helix [GP98]. The periodicity of the residues is also apparent. This figure has been reproduced with permission from [PR04].

Property	**Value**
Phi (ϕ)	$-57°$
Psi (ψ)	$-47°$
Omega (ω)	180°
Rotation about axis per residue	100°
Residues per turn	3.6
Translation per residue	1.5Å

Table 2.1: The properties of the ideal alpha helix. Phi, psi, and omega are the torsion angles, refer back to Figure 2.3 on page 11 for reference if necessary. The translation is the distance in Ångströms along the helical axis moved per residue [PR04].

Alpha helices are easily distorted, a fact that makes their study non-trivial. The axis of a helix is never a straight line, as helices are always bent to some degree. This is often attributable to steric interactions with other regions of the protein or between side chains within the helix, forcing the adoption of non-optimal configurations [BT88]. The different amounts that the helix is bent along the backbone can be classified. If the axis of the helix for one residue is less than 20° different from the next, the helix is classified as being linear or smoothly curved. If the angle exceeds 20° however, a hydrogen must be broken and the helix is kinked. If the angle exceeds 60°, then the kink is so severe that the helix structure on either side of the kink is treated as two different helices, with one or no residues in the loop region [BKV00].

The region where two helices pack together often experiences some degree of distortion, something that will be discussed in greater detail later. This variation, however, has a minimal impact on the energy of the system, as packed helices have energies that are close to the energies of each helix when isolated. This is because the torsion angles remain in the areas normally permitted as shown in the Ramachandran plot (see Figure 2.4 on page 13) [CLR77]. Proline affects the structure of an alpha helix because of the cyclic side chain (refer to Figure 2.2 on page 10). The nitrogen atom is bonded to the side chain, so it is clearly unavailable for the hydrogen bonding characteristic of alpha helices. Also, helices on the surface of the folded protein are often curved away from the solvent to allow the backbone oxygen atoms to increase their hydrogen bonding capabilities with the solvent [GP98]. Surprisingly, however, the length of a helix is inversely correlated with the curvature, so that shorter helices tend to be more curved than longer ones. It is believed that this is a fundamental property of helices; the stability of a curved helix decreases when it becomes lengthy [KB98].

2.1.3 Globular vs. Transmembrane Proteins

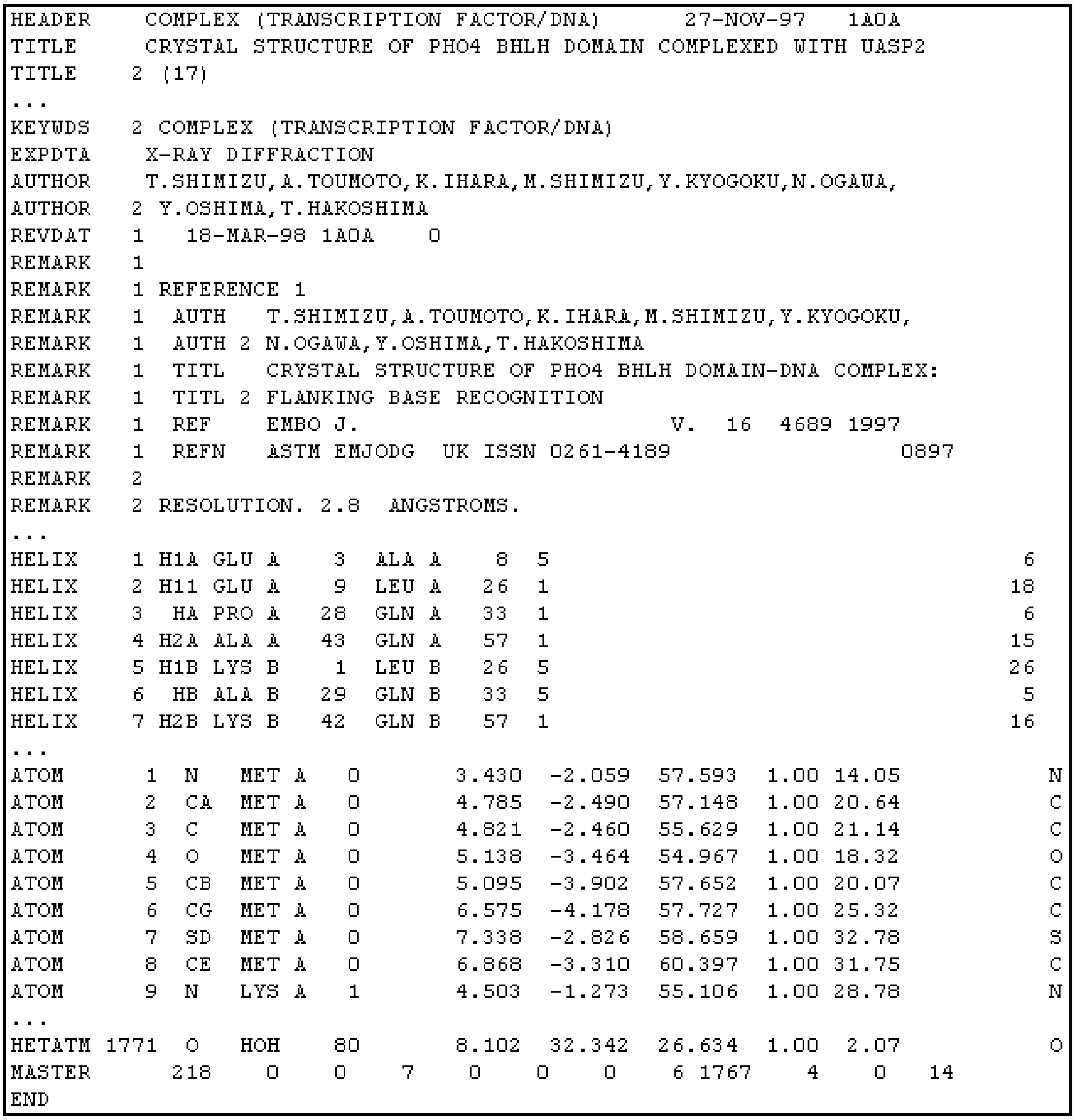
The two major classes of proteins that are studied at present are globular and transmembrane proteins. Globular proteins are densely packed and are more easily crystallized generally, which results in a relative abundance of proteins of this type that have been determined experimentally. Transmembrane (TM) proteins are associated with the phospho-lipid bilayer which forms the walls of cells and organelles found in cells. Transmembrane proteins may be proteins that are nearly globular yet have a tail which is embedded in the membrane, the protein may have components that are on each side of the bilayer, or the protein could be entirely embedded in the membrane [TMBA01]. The fundamental difference between these two classes of proteins, as far as alpha helices are concerned, is that transmembrane proteins tend to pack together at angles that are closer to parallel and anti-parallel than their globular counterparts. This is attributed in part to the constraint of crossing the membrane, so the helices tend to be oriented normal to the membrane [Bow05]. Also, the helices found in transmembrane proteins tend to be more kinked than those found in globular proteins; kinks in alpha helices may provide more flexibility (they may act as hinges), allowing the protein to perform some movements necessary for its function [TSUS01].

2.1.4 The PDB

The Protein Data Bank (PDB) is the standard repository for all protein structures once they have been determined. The standard source is online at the Research Collaboratory for Structural Bioinformatics (RSCB) PDB database at `http://www.rscb.org`. There is also the Worldwide Protein Data Bank (wwPDB) at `http://www.wwpdb.org/`, which is a consortium of RSCB, the Macromolecular Structure

Database at the European Bioinformatics Institute (MSD-EBI) and the Protein Data Bank Japan (PDBj). All three institutions mirror the same database, and the RSCB has control over the archives [BHN03], so the RSCB site is still the most direct point of access to PDB files (wwPDB simply provides links to the three member institutions). The coordinates contained in the files stored in the PDB have been determined by either X-ray diffraction or NMR spectroscopy. The PDB has a set of standards to which the files must conform, although in practice there are a number of files with idiosyncratic characteristics, such as the index of the first amino acid, or the labelling of the peptide chains.

A portion of a sample PDB file is shown in Figure 2.10. There is a large variety of information stored in the file. The helix information gives an index value for the helix, the species of the first residue in the helix, the chain ID and the position of the helix on that chain, the corresponding information for the ending residue, and the length of the alpha helix. The information for every atom is also listed (except perhaps for hydrogen, depending on the resolution of the model). Along the ATOM line is the atom index, the atom ID for that residue type, the residue type, the chain, the index of the residue, the coordinates, the occupancy value, the B-value, and the species of the atom. The occupancy value (where the values are all 1.00 in this example) is used if there are alternate conformations. A value of 1 indicates that there are no alternative conformations in the file (meaning none are known). The column after occupancy is called the B-value, a temperature value indicating the amount of disorder for the position of this atom due to thermal vibrations or other such effects [EJT04]. The columns where information is stored in the file for a particular line type is always the same, and the sequence that the atoms are listed in for a particular amino acid is



```
HEADER    COMPLEX (TRANSCRIPTION FACTOR/DNA)      27-NOV-97   1A0A
TITLE     CRYSTAL STRUCTURE OF PHO4 BHLH DOMAIN COMPLEXED WITH UASP2
TITLE    2 (17)
...
KEYWDS   2 COMPLEX (TRANSCRIPTION FACTOR/DNA)
EXPDTA    X-RAY DIFFRACTION
AUTHOR    T.SHIMIZU,A.TOUMOTO,K.IHARA,M.SHIMIZU,Y.KYOGOKU,N.OGAWA,
AUTHOR   2 Y.OSHIMA,T.HAKOSHIMA
REVDAT   1   18-MAR-98 1A0A    0
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   T.SHIMIZU,A.TOUMOTO,K.IHARA,M.SHIMIZU,Y.KYOGOKU,
REMARK   1  AUTH 2 N.OGAWA,Y.OSHIMA,T.HAKOSHIMA
REMARK   1  TITL   CRYSTAL STRUCTURE OF PHO4 BHLH DOMAIN-DNA COMPLEX:
REMARK   1  TITL 2 FLANKING BASE RECOGNITION
REMARK   1  REF    EMBO J.                      V.  16  4689 1997
REMARK   1  REFN   ASTM EMJODG  UK ISSN 0261-4189                 0897
REMARK   2
REMARK   2 RESOLUTION. 2.8  ANGSTROMS.
...
HELIX    1 H1A GLU A    3  ALA A    8  5                                   6
HELIX    2 H11 GLU A    9  LEU A   26  1                                  18
HELIX    3  HA PRO A   28  GLN A   33  1                                   6
HELIX    4 H2A ALA A   43  GLN A   57  1                                  15
HELIX    5 H1B LYS B    1  LEU B   26  5                                  26
HELIX    6  HB ALA B   29  GLN B   33  5                                   5
HELIX    7 H2B LYS B   42  GLN B   57  1                                  16
...
ATOM      1  N   MET A   0       3.430  -2.059  57.593  1.00 14.05           N
ATOM      2  CA  MET A   0       4.785  -2.490  57.148  1.00 20.64           C
ATOM      3  C   MET A   0       4.821  -2.460  55.629  1.00 21.14           C
ATOM      4  O   MET A   0       5.138  -3.464  54.967  1.00 18.32           O
ATOM      5  CB  MET A   0       5.095  -3.902  57.652  1.00 20.07           C
ATOM      6  CG  MET A   0       6.575  -4.178  57.727  1.00 25.32           C
ATOM      7  SD  MET A   0       7.338  -2.826  58.659  1.00 32.78           S
ATOM      8  CE  MET A   0       6.868  -3.310  60.397  1.00 31.75           C
ATOM      9  N   LYS A   1       4.503  -1.273  55.106  1.00 28.78           N
...
HETATM 1771  O   HOH    80       8.102  32.342  26.634  1.00  2.07           O
MASTER      218    0    0    7    0    0    0    6 1767    4    0   14
END
```

Figure 2.10: This is a condensed version of the PDB file for 1a0a.pdb, the protein used in other examples in this thesis. The format of the file is rigidly specified so that programs can parse the files to extract the wanted information. The first six characters in every line describe the contents of the line. Most of the REMARKs have been removed from the file as displayed here, as have been other types not used in this work. This simply gives the impression of the file format, and a sample of what the header information looks like, and how the HELIX and ATOM lines are structured.

consistent.

The file names in the PDB always begin with a number, followed by a series of characters. This initial number serves as a versioning system for the files, so that if a protein corresponding to a file in the PDB is determined again, perhaps at a higher resolution or after the clarification of the species of an amino acid, the initial number of the file is incremented. Some files are no longer available in the RSCB PDB database, such as PDB ID 1cpa. It has been deemed obsolete, and now only the files 3cpa-8cpa are available. For more information on PDB file formats, see the official PDB guide [CCD+96].

2.2 Contact Maps

This research pursues a slightly different tack from the conventional, by using contact maps as a fundamental tool. One way to represent the three-dimensional structure of a protein is a distance map. A distance map is an $N \times N$ matrix, where N is the number of amino acids in the given protein, and entry D_{ij} in the matrix is the distance from amino acid i to amino acid j in 3D space, typically measured in Ångströms (Å). Of course, it is not possible at present to predict distance maps from an amino acid sequence[2], but the prediction of contact maps is being performed, as discussed in Section 2.3.

A contact map can be thought of in a conventional sense as a binary version of the distance map, where a threshold has been applied to yield Boolean values, as shown

[2] Solving the problem of predicting a distance map is tantamount to predicting three-dimensional protein structure. Given all the distances from every point to every other point in three-dimensional space, the configuration can be found exactly. This is known as the molecular distance geometry problem [YGW00].

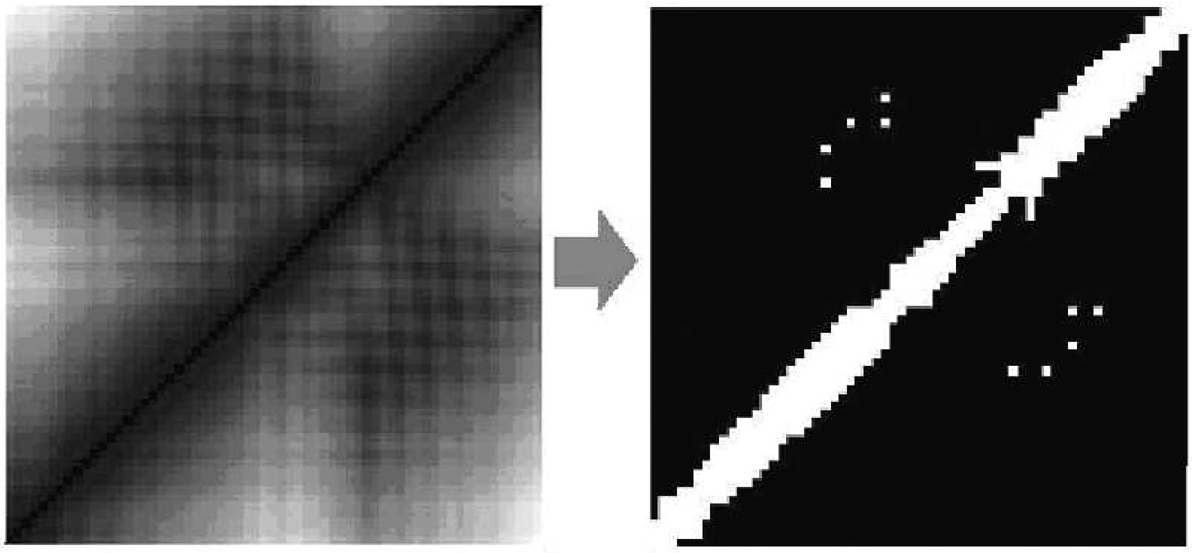


Figure 2.11: The image on the left is a distance map derived from the PDB file for one chain of the protein with PDB ID a1o1, a transcription factor (DNA). The darker the region, the closer those two amino acids are in space. The diagonal axis is black, as an amino acid is distance 0 from itself. The distance map has been converted to the contact map on the right by applying a threshold of 7Å to the distance map. In the contact map a white square indicates a true value, meaning that those amino acids are within the threshold distance of each other. Notice the smattering of white squares distant from the central axis in the contact map. These represent contact between regions of the protein which are distant from each other on the backbone. These are the regions where we will find the contacts for supersecondary structures.

in Figure 2.11. The threshold value is the definition of a contact.

The means of determining an ideal value to use for the contact map threshold distance is by no means an objective process at present. A literature review found many varieties of values; indeed the choice of what to measure for distance is by no means standardized at present. For example, Fariselli et al. [FOVC01] are using 8Å between C_β as their model, Vendruscolo et al. [VKD97] use 9Å between C_α.

Additionally, they add the constraint of eliminating all contacts that are between amino acids located within seven positions of each other in the sequence to avoid getting contacts which are associated with turns. Källblad and Dean [KD04] use 5Å between C_β with a five position gap instead of the seven used by the former group, and Hu at al. [HSS$^+$02] use a double threshold of 4 and 7Å between C_α, the thought being that this will eliminate contacts not associated with supersecondary structures. Our group uses 10Å between C_α as the threshold, as for our purposes it provides the most useful information, as illustrated in Figure 2.12 on the next page.

Work by previous students has produced software which has the ability to isolate the contact maps associated with a particular pair of proteins. This software provides the means of testing the hypothesis that properties of alpha helix pairs may be predicted from contact maps. This involves isolating the so-called interface region of the contact map for the pair as a first step. This concept is shown in Figure 2.13.

Once we have the contact map, we would like to extract as much useful information as possible. Clearly, simply using the contact map provides some constraints: for any given pair of amino acids, it is known whether they are closer than the threshold value or not to each other in space. Other than this obvious information, there is nothing very apparent in the information represented by the contact maps.

2.3 Protein Structure Prediction

The ultimate objective of predictive approaches to structure prediction is that since function is determined by structure, the structure of an individual's proteins could

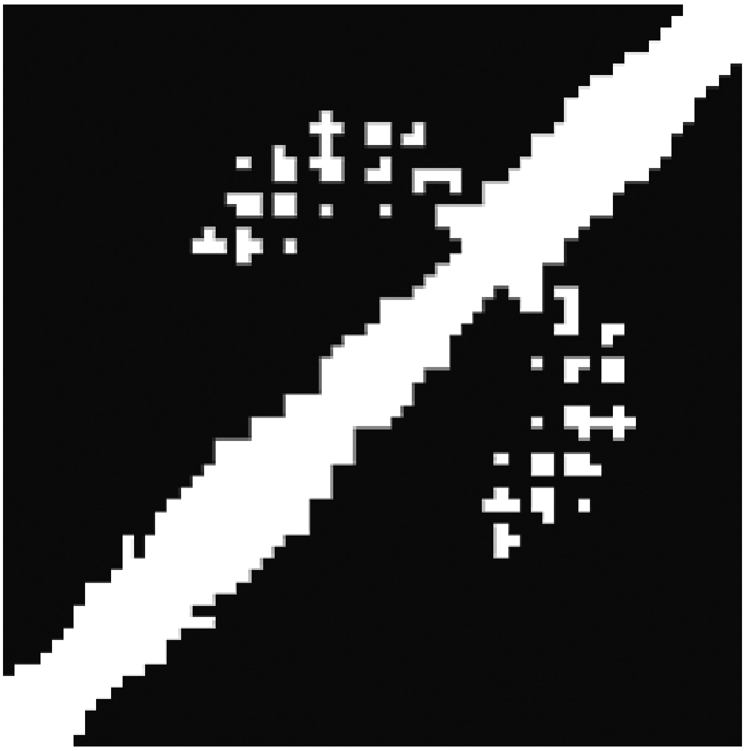


Figure 2.12: This is a contact map for the same protein (a1o1) that was used for Figure 2.11 on page 26. Notice that in this case, there is much more information corresponding to the supersecondary structures. With a threshold of 7Å, we would often have a contact interface with only one point of contact. By increasing the threshold to 10Å, we are increasing the amount of information available to describe the interface.

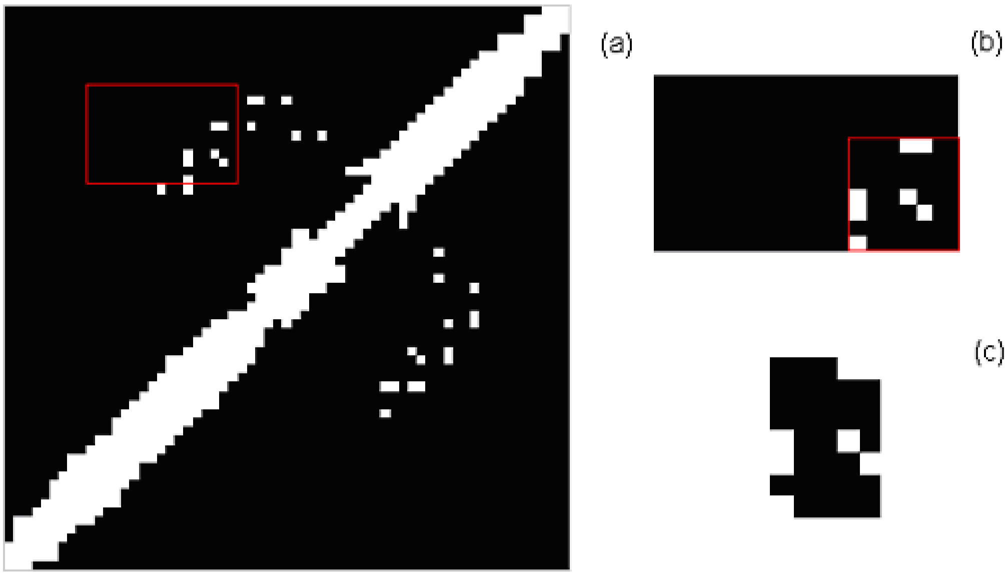


Figure 2.13: (a) The contact map for protein 1a0a is used again. The contact map interface is found by isolating the smallest rectangle containing all of the contact points from the contact map for the helix pair. The red rectangle indicates the area occupied by two helices, shown in (b). The contact map represents all of the amino acids for one alpha helix along the vertical axis and the other along the horizontal. This has been further refined to the interface area, shown in (c).

be predicted from their primary sequence (see Fetrow et al. [FGS98] for an example of an application of the "sequence-to-structure-to-function paradigm"). Now that the human genome has been sequenced [LLB$^+$01, VAM$^+$01], the primary sequence is accessible. A major long term of goal of this line of research is custom drug design. The effect of drugs could be predicted from their structure, circumventing the lengthy process of synthesis and trials which is the status quo. If an individual were ill due to a genetic defect, the determination of the defects in their proteins would facilitate treatment. The modest goal of the present research is to assist in the prediction of supersecondary structures from the primary sequence.

The conventional means for protein structure determination is by either X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. The former technique produces an electron density map, while the latter produces a graph with peaks corresponding to the shift due to each nucleus in the molecule. The researchers interpret this data to produce the coordinates for the protein. These methods are susceptible to minor inaccuracies, and are limited by the resolution of the tools used (for example, it is often difficult to determine the location of hydrogen atoms) [PR04].

2.3.1 *ab initio* Modelling

The basic premise behind the *ab initio* approach to protein structure prediction is that the three-dimensional structure of a protein corresponds to the global minimum energy configuration of the molecule (termed the "thermodynamic hypothesis" [Anf73]). Proteins do not have completely static structures; the core of a protein packs as densely as organic crystals, as opposed to liquid hydrocarbons [RW93], but this does not hold for the surface. It is possible that the native structure is not precisely the global minimum, but it is likely close (within 1.5 Å rmsd[3]) [BB01, EJT04, SFWB$^+$05], although this view is not held universally [ACSR97].

In order for such an approach to be successful, the appropriate energy function must be determined to model the energy of the molecule. An example of such an energy function is discussed in detail in section 2.4. As mentioned earlier, this approach to protein structure prediction is intractable because of the literally astronomical number of possible conformations that can be adopted by the protein, corresponding to local minima. For example, consider that the upper bound on the number of elementary particles in the known universe is less than 10^{89}[Bry03]. Each amino acid has at least three known rotamer classes [LWRR00], so given a small protein of 187 amino acids we already have 3^{187} (which is greater than 10^{89}) possible conformational states due to the side chains alone. Clearly some heuristics must be applied if there is any hope of using this approach. There are several methods of simplifying the problem, such as solving for different types of interactions incrementally, or beginning with a coarse structure and progressively refining it. Blundell et al. [BSST87] claim that due to the numerous local minima, energy minimization is only effective if the structure is already within 1Å of the correct structure. Another tactic is to begin by solving for local structures in the protein that are well packed, such as supersecondary structures, and then progress by assembling the results [SFWB$^+$05], although this approach can also run into trouble because of the numbers of local minima [Nag89]. Of course, another approach is to wait for technology to catch up (Moore's Law, [Moo65]), this tactic has borne fruit recently, as several approaches that were not possible 15 years ago are now tractable [SFWB$^+$05]. Similarly, the number of files in the PDB is

[3] rmsd is short for root mean square deviation or distance, a conventional method for comparing two sets of data points. See [MC94] for a thorough discussion of aligning and comparing protein structures.

experiencing accelerating growth due to improvements in high-throughput structure determination technology [WFC$^+$03], and so approaches using the PDB as a resource are experiencing a dual benefit.

2.3.2 Homology Modelling

Homology-based modelling relies heavily on the information stored in the PDB. These approaches search for structures in the PDB with sequences and properties similar to those of the structure that is being predicted, and then assemble the results of the searches into a final possible structure. This tactic (along with various refinements) has produced the best results to date for protein structure prediction [BB01]. A caveat, however, is that there exists a bias in the PDB towards protein structures that are more easily sequenced. There is a need for many structures to be determined within some protein families where none exist now. There are many proteins where a search of the PDB will yield no results for proteins with 30-35% sequence similarity [Bur00].

The cornerstone of homology-based modelling is sequence alignment. Given a primary sequence for which you intend to predict the tertiary structure, you search the PDB files in your database for portions of proteins with primary sequences similar to your target. The premise for this approach is that similar primary sequences yield similar tertiary structures. Differences in primary structure are acceptable because there are different rates of structure conservation for each level of the structural hierarchy (specifically, tertiary structure is conserved better than primary structure, which in turn is conserved better than DNA sequences [BSST87]). Thus, similar but distinct primary structures can conceivably yield the same tertiary structures. Since identical sequences are not going to be found (except perhaps for highly conserved segments), algorithms have been developed to provide a metric for the degree of similarity between two sequences. This is not a simple problem, as the properties of the side chains for the amino acid residues result in situations where a change from one amino acid to another may have no effect, or conversely could significantly affect the structure of the protein.

Threading (also known as inverse folding or fold recognition) was developed as a more intelligent form of homology [GGGR05]. Threading works by matching sequences to structures, in the hopes of finding similar structures with no evolutionary similarity. With the development of more sophisticated alignment tools, such as PSI-BLAST, hybrid approaches between threading and homology permitted greater accuracy in predictions. PSI-BLAST uses a look-up table to assign a cost for substitutions, insertions and deletions [AMS$^+$97]. Once the configurations of the homologous portions have been determined, they must be assembled to create the overall tertiary structure (just as with the approach where alpha helix pairs would be predicted). The two major domains of this problem are the prediction of the loop structures and the side chain conformations [ALJXH01]. There are several approaches for each, and the accuracy is encouraging. For loop regions, several approaches report prediction errors of less than 2Å rmsd for eight residue loops [FDS00, ALJXH01] and similar results are reported for side chains [SM98, MBCS00].

2.3.3 Triptych

Triptych is a novel case-based reasoning framework that is being developed in an attempt on the protein structure prediction problem. Case-based reasoning is well

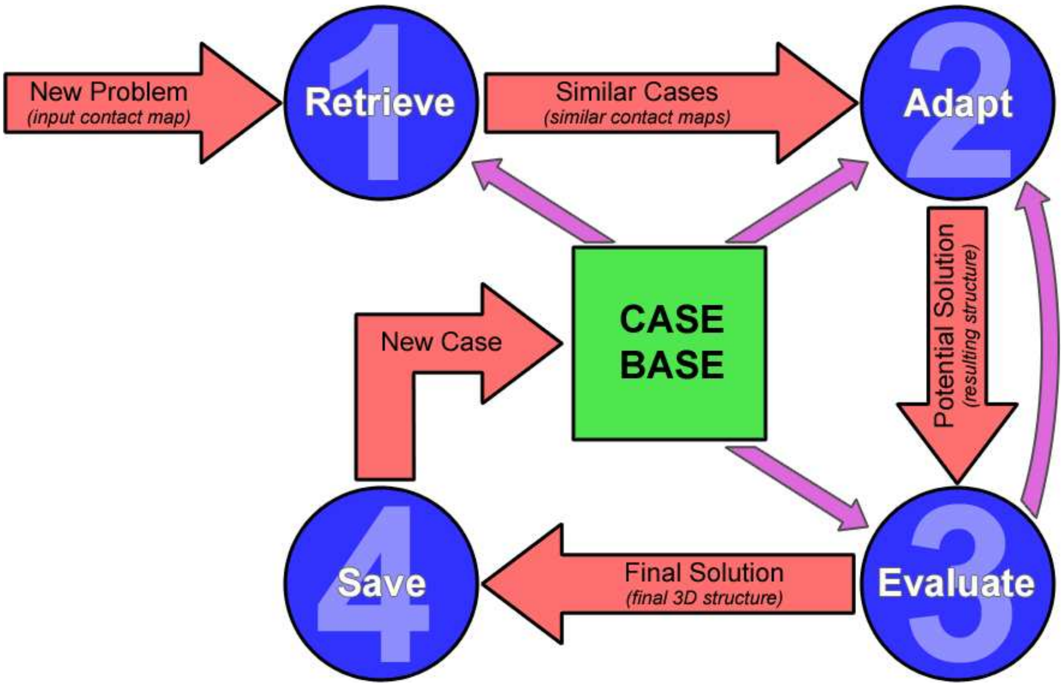


Figure 2.14: A schematic of the case-based reasoning system.

suited to this application because a database of protein structures inherently provides a vast stock of experiences to draw upon. Case-based reasoning has been used successfully in molecular biology applications in the past (see [JG04] for a review). The case-based reasoning framework that is used in Triptych is outlined in Figure 2.14.

Triptych employs a hierarchical bottom-up approach, first constructing super-secondary structures which are then assembled to create the entire protein structure using higher level contact maps. The process begins by taking a contact map as input [DGK06, GKD06]. For the protein structure prediction task, the contact map has been predicted from the primary sequence. Accuracy for predicted contact maps is less than 65% presently [PB02], but several groups are making progress on the problem[4]. Areas of the contact map corresponding to super-secondary structures can be identified and isolated. Now the case-based reasoning system is applied, composing of four steps:

- *retrieve* - In the retrieval step, the database is searched to find regions of contact maps that are similar to those from the predicted contact map.

- *adapt* - In the adaptation step, the structures corresponding to the contact maps obtained in the retrieval step are adapted to suit the predicted contact map. This provides a set of potential solutions to the prediction problem.

- *evaluate* - The evaluation step is the heart of the case-base reasoning framework. The proposed structures obtained from the adaptation phase are evaluated and ranked by a committee of advisors which are also using cases from the database. The advisors are implemented in a neural network so that the most useful advisors have the greatest influence on the final score associated with the proposed structure. Triptych uses many advisors, such as evaluating the steric stability of the structure, the hydrophobicity, the possible rotamer classes of the side chains, and so on. The final system will likely have 20 to 30 advisors [DGK06]. The more advisors that are present in the evaluation phase, the more robust the case-based reasoning system will be. If the proposed structures are poor, they are not accepted and the system returns to the adaptation phase to produce another set of candidate structures.

- *save* - Once the evaluation step accepts a structure, the prediction is complete. This structure can now be exported, and it can be saved into the case base to

[4] see Fariselli et al. [FOVC01], Pollastri and Baldi [PB02] and Punta and Rost [PR05] for details about contact map prediction.

provide more cases for future predictions.

The work being performed in this thesis is related to the Triptych project. Hippy provides a tool for researchers developing advisors, as they are able to explore pairs of alpha helices and gain intuition regarding the properties of their advisors. The algorithm for the interhelical angles can be used as an advisor. The advisor would compare the angle between pairs of helices produced in the adaptation phase to those of pairs of alpha helices with similar contact maps in the case base.

2.3.4 Contact Map Prediction

The groups working to predict contact maps ([FOVC01, PB02, PR05]) do not have the luxury of a distance map from which the contact map may be created, thus their methods involve predicting the contact map. Farieselli at al. [FOVC01] use neural networks to perform their predictions with evolutionary pathways, conserved regions of sequence, and predicted secondary structures among the information that is used as inputs. Regardless, this task is peripheral to the present study; for the purposes of this research, it is being assumed that the results of contact map prediction will be highly accurate at some later stage. A challenge at present would be to use an empirically determined contact map to determine the three-dimensional configuration of a pair of alpha helices in the contact map.

2.3.5 CASP

The metric in the field for comparing protein structure prediction algorithms is performance at the biannual Critical Assessment of Structure Prediction (CASP), of which six have been held. More recently, another similar program called the Critical Assessment of Fully Automated Structure Prediction (CAFASP) has been held where no human intervention is permitted in the prediction [GGGR05]. The idea behind CASP and CAFASP is that there exists a great opportunity to test prediction algorithms blindly given a structure that has been determined experimentally by conventional means, but is not yet published. The concept of CASP originated in the early 1990's. Researchers suggested this challenge to ensure that prediction tools, particularly those employing homology modelling, are unbiased in their predictions [BR93]. In effect, it was to prove to the community that the predictions are not a matter of soothsaying [BCG92]. In the last iteration of CASP, over 200 prediction teams representing 24 countries participated.

The methods used at CASP are classified based upon the algorithms used, but many of those classified as *ab initio* unfortunately used some information from databases to augment their methods [Osg00]. Pure *ab initio* methods use only a representation of the geometry of the protein, a force field, and an algorithm for searching for the global minimum. Since the first CASP, the accuracy of the predicted models has been steadily improving, with the results at the last session producing results twice as accurate as those in the first. If such progress continues, the results at CASP 11 (to be held a decade from now) should be of excellent quality [KVFM05].

2.4 Protein Visualization Packages

There is a significant number of software packages designed for modelling proteins, most of which accept the PDB file for the protein as input and extract the relevant information from the file so the user can customize the view for the properties relevant to their studies. Among the most popular are Rasmol [SMW95], Chimera [PGH$^+$04], Swiss-PDB viewer [GP97], and Protein Explorer [Mar02].

The limitation of these packages with regards to the applications required for this research is that the displays are tailored for the entire protein and do not allow for much insight into the interaction between the alpha helices at the level of intermolecular forces. The structures of proteins, as read from the PDB files, are treated as static in these packages, so there is no ability for the user to attempt to move the helices to witness the effects. The most relevant package that has been found for this application is called SCULPT, a package which displays a protein and allows the user to interactively move individual atoms to gain an understanding of the effects on the structure and interatomic forces involved [Sur92]. This package is proprietary, and the features are limited in scope as related to the study at hand, as there is no ability to examine the effects of manipulating the structure on the contact map interface. The package models interatomic forces very well. Van der Waals forces are modelled by rendering translucent spheres around atoms at the van der Waals radii of the respective atoms, and the colour of the sphere reveals the nature of the interaction (blue for attraction, red for repulsion), as shown in Figure 2.15 on the next page [SRRJ94].

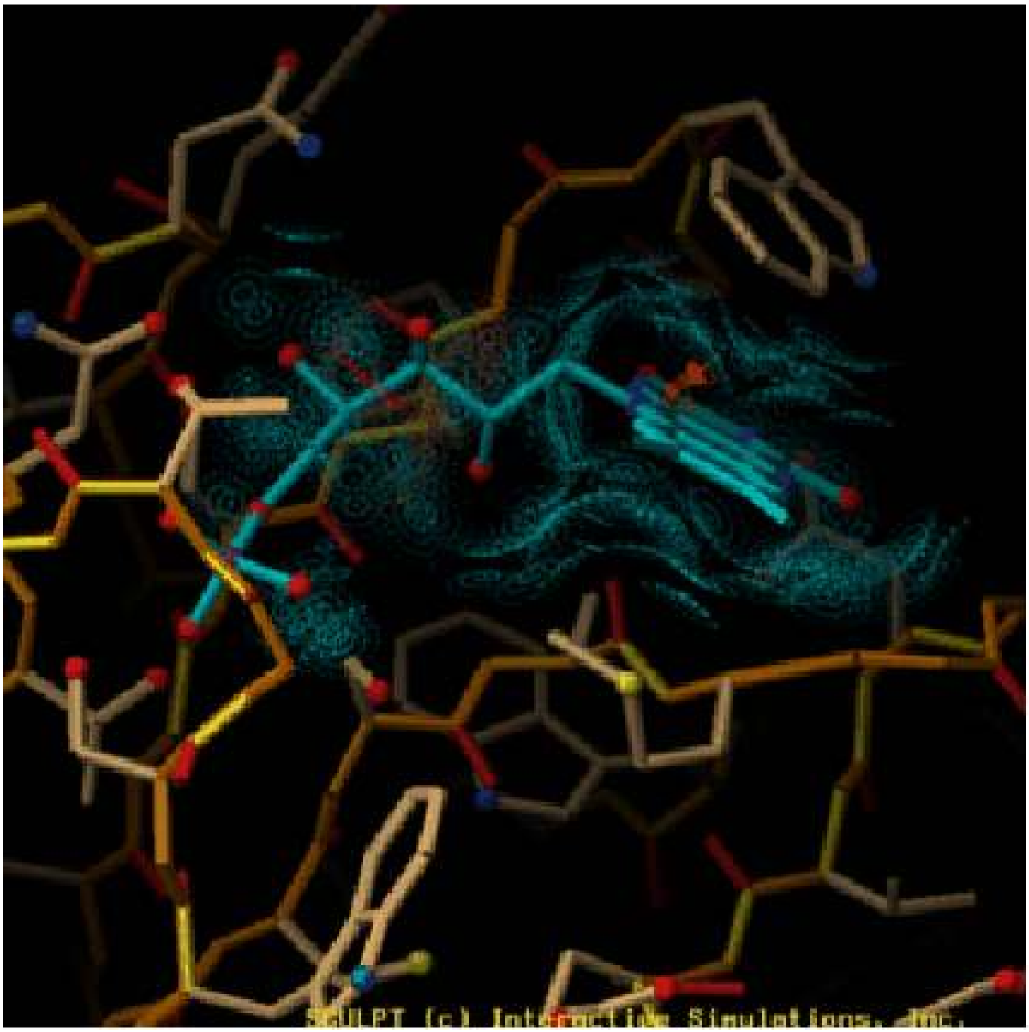


Figure 2.15: A sample image created with the SCULPT modelling package [SRRJ94]. The atoms and covalent bonds are represented in a traditional ball and stick manner, but the rendering of the van der Waals radii as translucent partial spheres is innovative and useful for understanding their effects on the geometry of the protein.

The aim of this portion of the thesis is to create a system that expands the understanding of the interface of helix pairs and how changes to the configuration of the protein (both at the primary and supersecondary levels) affect the contact map for the interface. The system will allow the user to select any pair of the helices from the protein, and the pair will be shown along with whichever relevant properties and forces are wanted, and the contact map will be shown so the effects of the user interaction can be viewed in real time. Initially, it was thought that a modelling system which was based upon the energy of the system would prove most useful, as the protein native state is commonly thought to be a low energy configuration (Anfinsen's thermodynamic hypothesis [Anf73]). The following section discusses the modelling of proteins based on the energy of the system.

2.4.1 Force-based Modelling of Proteins

The forces and properties that are relevant to a protein model are:

- bond length
- bond angle
- single dihedral angle
- multi-value dihedral angle
- hydrogen bond
- van der Waals interaction
- electrostatic charge
- solvent interaction

The SCULPT package discussed earlier constrains the first three items to their ideal values [SRRJ94], which is not invalid because the strength of those forces is at least an order of magnitude greater than the others [Str88]. This is implemented primarily for reasons of efficiency, as with increasing numbers of atoms the number of interatomic interactions increases quadratically. The SCULPT package uses springs to model the other forces, except solvent interaction, which is ignored. This means that for a given van der Waals interaction between two atoms, for example, there is an ideal distance that separates the two atoms. Using the spring model, the energy of the system increases exponentially as the distance from this ideal is changed. It was thought that since the proposed system involved a limited number of atoms (those found in the two helices), it should be possible to model all of the interatomic forces and properties accurately, except for solvent interaction. The forces existing between two atoms can be described with the following equation [Lea01]:

$$\begin{aligned}\nu\left(r^{N}\right) = \sum_{bonds}\frac{k_i}{2}\left(l_i - l_{i,0}\right)^2 + \sum_{angles}\frac{k_i}{2}\left(\theta_i - \theta_{i,0}\right)^2 + \sum_{torsions}\frac{v_n}{2}\left(1 + cos\left(n\omega - \gamma\right)\right)\\ + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \qquad (2.1)\end{aligned}$$

where

$\nu\left(r^{N}\right)$ is the potential energy of the system
r is the position of the atoms

N is the total number of atoms

k is the stretching constant (experimentally determined)

l_i is the length of the bond

$l_{i,0}$ is the ideal length of the bond

θ_i is the bond angle

$\theta_{i,0}$ is the ideal bond angle

V_n is the barrier height (experimentally determined)

n is the multiplicity (the number of minimum energy angles)

γ is the phase factor (at which angle the minimum energy exists)

ω is the torsion angle

ε is the well depth (experimentally determined)

σ is the collision diameter (experimentally determined)

q is the charge of the atom

ε_0 is the permittivity of space

Notice that the first two terms, modelling bond length and bond angle respectively, are represented using Hooke's Law [Lea01]. The third term models torsion angles, and the fourth term combines the two non-bonded interactions. The first part is the van der Waals forces, modelled using the Lennard-Jones equation; the remainder uses Coulomb forces to model electrostatic interactions [Lea01]. This equation does not need to be evaluated for all pairs of atoms in practice. Obviously, the bonded interaction terms only need to be solved if the two atoms are actually bonded. The non-bonded terms only need to be considered for interatomic distances of under 10Å, and only the electrostatic interaction needs to be evaluated if the distance is greater than 6Å and less than 10Å [SRRJ94]. Thus a binary matrix could be maintained which stores which atoms are bonded to which others, and another distance matrix can be maintained to determine which interactions need to be calculated. When there is user interaction with the system, the distance matrix will need to be modified accordingly, but the bond matrix is static. From the distance matrix, it is a very simple matter to create the contact map for the helix pair dynamically. The contact map, as described earlier in section 2.2, is concerned with only the alpha carbons in our work. The contact map is just a binary (by an applied threshold) sub-matrix of the distance matrix for the entire protein. The relation for each pair of atoms can be entered into a matrix storing all of the equations as a linear system. The solution can be found in a reasonable amount of time because the matrices are very sparse, and there are tools available to exploit this [WK88].

There are a number of static protein visualization packages available that are similar in spirit to the proposed system. One particularly nice package is WebMol [Wal97], which is Java-based and embedded in a web browser. The package has the same limitations as most of the other packages already discussed, specifically the emphasis on the structure of the entire protein and the lack of the contact map display. WebMol does have the option of displaying the distance map, a nice feature and only a handful of lines of code away from including contact map functionality. A program named Dotter [SW98] was created which allowed the user to dynamically adjust the threshold value that is used to create the contact map. Dotter is a useful tool for displaying the contact map for a whole protein, but we are interested in only pairs of alpha helices.

There were several reasons why the force-based model was not used in this thesis. Firstly, the values in equation 2.1 that were described as being experimentally determined are not universal. Several sets of data that can be used have been found. These include the MM2/MM3/MM4 packages [AYL89, ACL96] and AMBER [PCC$^+$95]. In addition, the relevance of such a system is questionable. Several extant modelling systems have implemented the features (such as Chimera), and so for the labour involved with the implementation, it was decided that other features would be more useful for Hippy. Finally, another feature found in other packages that will not be used in Hippy is the cartoon representations of the backbone, as shown in the Chimera screen shot in Figure 2.16. This feature makes for very nice graphics, but they do not

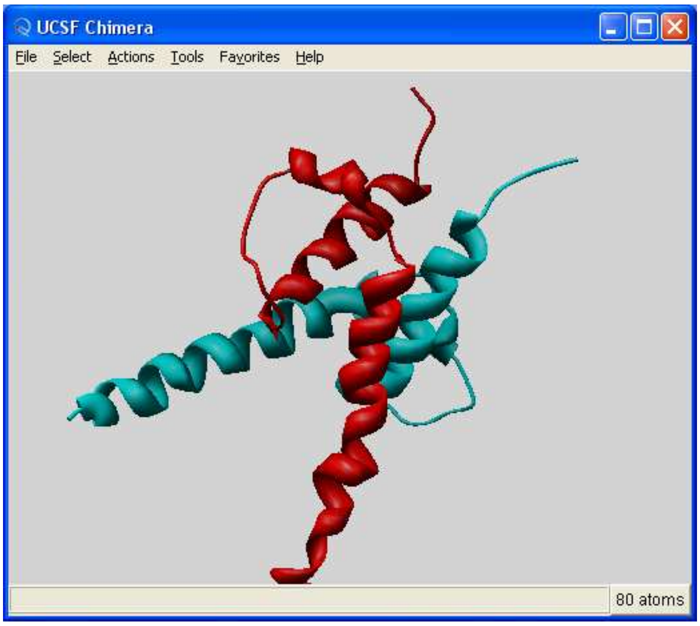


Figure 2.16: This is an image of protein 1a0a created with the Chimera modelling package [PGH+04]. The backbone and helices are rending with an attractive ribbon shape. Such features are more aesthetic than functional.

contribute much to the understanding of the configuration of two helices. It could be argued that such a feature clearly indicates where the secondary structures are in the protein, but since Hippy only displays pairs of helices this is not a necessity.

Chapter 3

The Hippy Visualization Package

If we knew what we were doing, it wouldn't be called research, would it?

-Albert Einstein

Hippy, the helix pair viewing software, was developed to efficiently configure the view according to the needs of the user [FG06]. Since Hippy is designed specifically for the visualization of pairs of alpha helices, many assumptions can be made. The initial helix pair selection tool (see Figure 3.1) shows the entire protein, and the user can move through the pairs of alpha helices and see the pair and the corresponding contact map at various thresholds. Once the user has selected a pair for viewing, the properties of the display may be tailored in the main window, as shown in Figure 3.2. The rendering of multiple properties of the helices may be switched on or off so that only relevant information is being displayed. Hippy's features are manipulated using the keyboard and mouse; see Figure 3.3 for the command window. The implemented properties include:

- The alpha carbons only can be shown or all other backbone atoms can be shown as well;

- The sidechains can be displayed or hidden;
- The opacity of the sidechains can be varied on a discrete linear scale;
- The van der Waals shells of the side chains can be shown or hidden;
- The contacts (derived from the contact map) between alpha carbons can be shown or hidden;
- The threshhold distance being used to calculate the contact map can be varied.

Figure 3.2 illustrates a screenshot from the main window of the software, showing most of these features in action. The window shows the helix pair viewer rendering the second and fourth alpha helices (as indexed in the PDB file) from the protein 1a0a. In this example, all backbone atoms are being drawn and the sidechains are being rendered with very low opacity. The contacts between alpha carbons are shown as the translucent bars, and the van der Waals shells are shown with a different colour used for each helix. Where the van der Waals shells from one helix are in contact with those of the other is an indication of areas with significant interaction between the alpha helices.

There are numerous conventions that were used in the design of this package:

- the secondary structure information was taken from the PDB file.
- the colouring scheme for the atoms is the standard C-P-K scheme (after Corey, Pauling and Koltun) [Kol65].

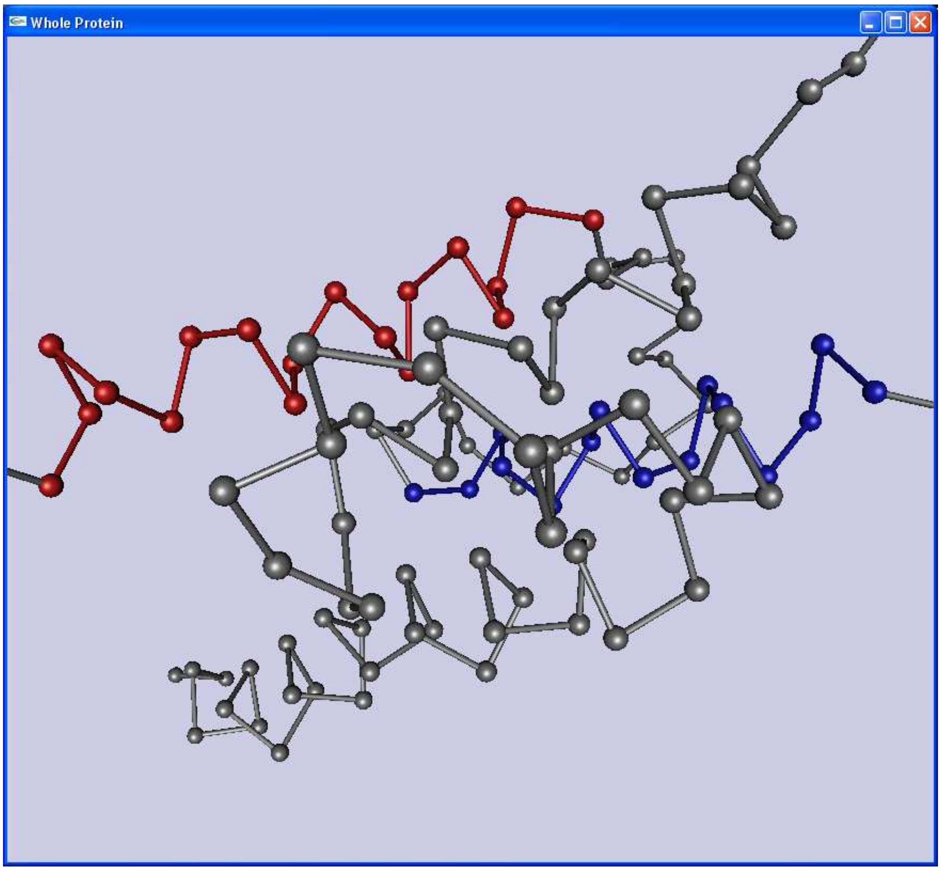


Figure 3.1: The helix selection window from Hippy, with the helices of index 2 and 4 from protein 1a0a highlighted. Scrolling through the helices here changes what is being rendered in the main Hippy window and the corresponding contact map.

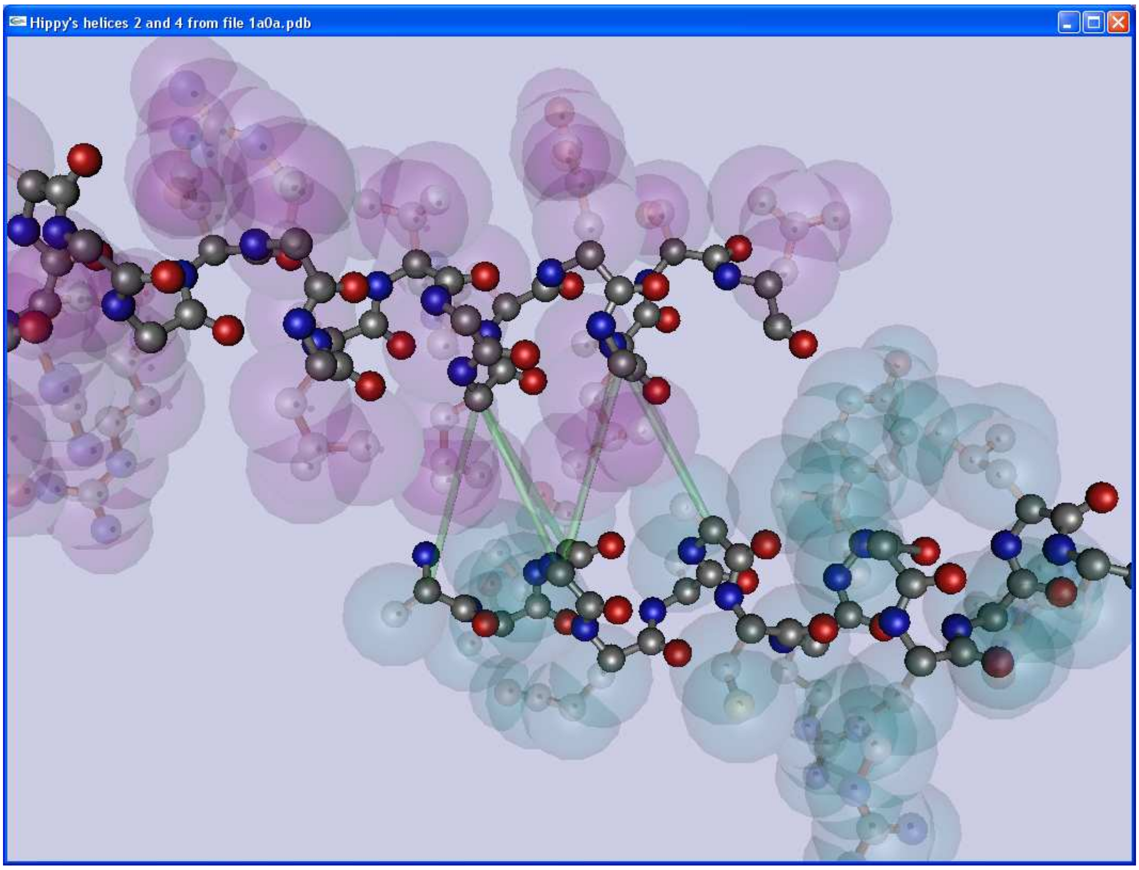
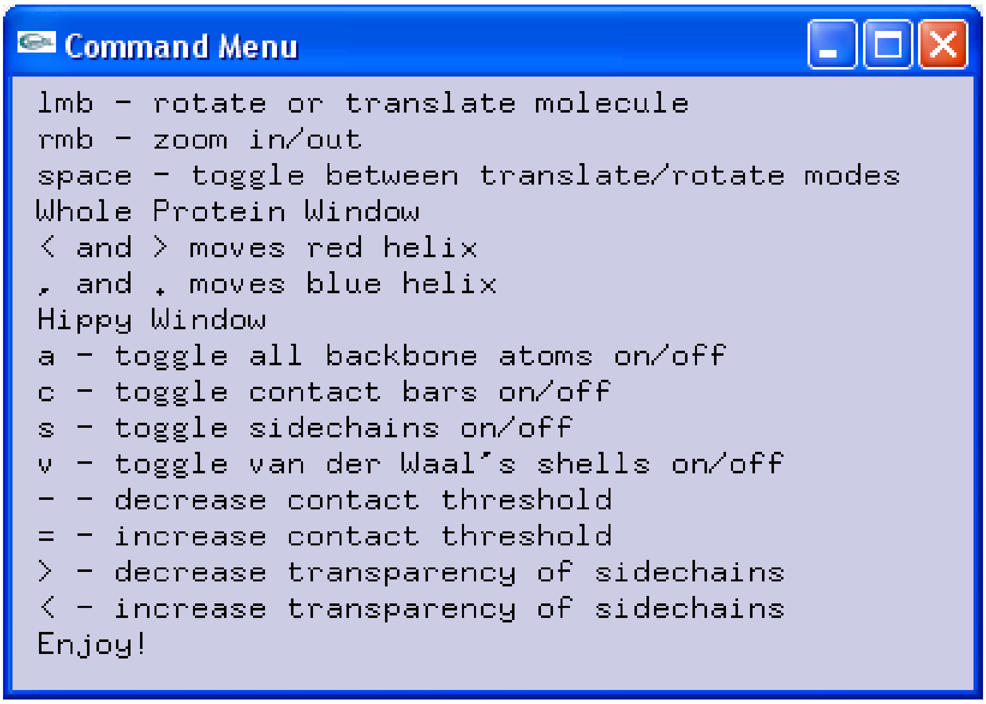


Figure 3.2: Hippy rendering the helix pair from Figure 3.1. The purple and green spheres illustrate the van der Waals shells of the atoms on each helix. The green bars are connecting alpha carbons that are in contact. All atoms of the backbone are being rendered, and the side chain atoms are rendered with low opacity.



```
lmb - rotate or translate molecule
rmb - zoom in/out
space - toggle between translate/rotate modes
Whole Protein Window
< and > moves red helix
, and . moves blue helix
Hippy Window
a - toggle all backbone atoms on/off
c - toggle contact bars on/off
s - toggle sidechains on/off
v - toggle van der Waal's shells on/off
- - decrease contact threshold
= - increase contact threshold
> - decrease transparency of sidechains
< - increase transparency of sidechains
Enjoy!
```

Figure 3.3: This is Hippy's command window. The user can refer to this window for the keystrokes required for Hippy's functionality.

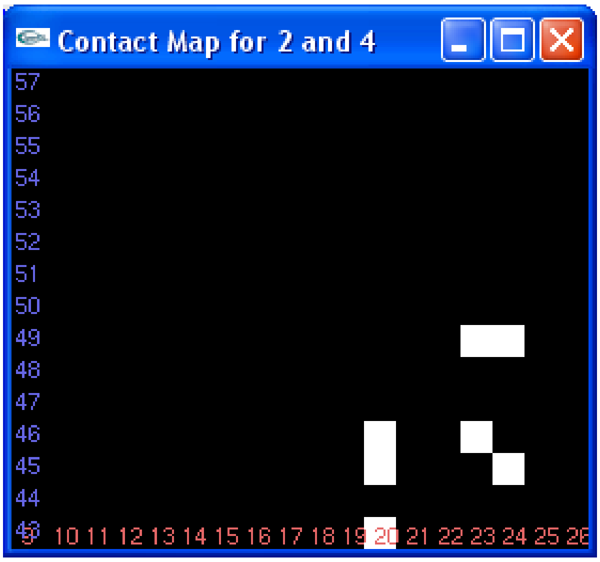


Figure 3.4: The contact map window from the helix pair viewer, corresponding to the helices of index 2 and 4 from protein 1A0A. Note the correspondence with Figure 1b. Clicking on a contact here causes the bar associated with the contact in the main window to flash.

- the radii for the van der Waals shells uses the United Atom Radius, which is the convention used by Rasmol-based viewers [MS00]. This model creates a sphere which approximates the radii for the heavy atom and the hydrogen atoms bonded to it as one. This is needed because most PDB files do not contain hydrogen atoms (due to the inability of X-ray crystallography techniques to resolve them).

To open a file, the user simply specifies the name of the file containing the helices that are desired[1], and then selects the indices of the helices themselves. The file is then parsed to extract all of the coordinate data corresponding to those helices. The contact map for the pair of helices is calculated from the coordinate data using

[1] The standard format for input files is that of PDB files; users can create their own files that contain the atom and helix data for a molecule in PDB format if so desired.

the desired contact threshhold (the default is 8Å), and the contact map window is rendered, as shown in Figure 3.4. This figure shows Hippy's contact map window corresponding to the pair of alpha helices shown in the previous figure. Notice that there is a clear pattern that becomes obvious from this contact map, as there is some symmetry. The bottom-most contact (corresponding to the left-most contact bar in Figure 3.2) may or may not be significant depending on the packing of the helices. Hippy facilitates determining the significance by allowing a researcher to examine the helix shapes and the van der Waals shells of the side chain atoms in the main window and by adjusting the contact threshold.

The program, source code, and documentation may be obtained from the author. For details regarding the implementation of Hippy, refer to Appendix A.

Chapter 4

The Interhelical Angle

We may, I believe, anticipate that the chemist of the future who is interested in the structure of proteins, nucleic acids, polysaccharides, and other complex substances with high molecular weight will come to rely upon a new structural chemistry, involving precise geometrical relationships among the atoms in the molecules and the rigorous application of the new structural principles, and that great progress will be made, through this technique, in the attack, by chemical methods, on the problems of biology and medicine.

-Linus Pauling

4.1 Helix Packing Models

Perhaps the most significant single characteristic of the alpha helix pair is the interhelical angle. This was recognized in the earliest papers, and was indeed considered the first stepping stone towards tackling the problem of predicting the configuration of alpha helices. Intuition may speak to some that helices should tend towards parallel or anti-parallel configurations, due to steric effects between the side chains of each helix. A parallel configuration would minimize the energy of the pair. This same line of reasoning has been used to explain why this parallel configuration is actually rarely observed in nature. Once the possible configurations allowing interdigitation of side chains are considered, the result is a high energy system if the helices were parallel. A more preferable configuration is to have the helices at an angle of about 20° to each other and slowly coiling around one another [WW03].

In this section of the thesis, we will examine different packing models in turn, serving to illustrate the evolution of the alpha helix pair packing model. First, we define what we mean by the packing of two helices. An appropriate definition would be that the distance between two helices is less than some threshold value, in order to ensure that some steric contact is actually occurring. The other consideration is that the line representing the line of closest approach between the two helices should be normal to both axes, this ensures that the helices are actually packing and not just contacting incidentally [TS04]. These guidelines are followed (or at least assumed implicity) in all of the packing models described here.

4.1.1 Knobs into Holes

Francis Crick [Cri53] was a pioneer in the study of alpha helices; he was the first to attempt to determine whether there were preferential angles in which helices pack. His model established that the surface of the helix is essentially composed of bulges and pockets, and that these properties should dictate packing configurations. The bulges are the steric srufaces of extruding side chains, which can be viewed as knobs on the helix. The pockets left between the knobs are the holes. He began with the assertion that parallel helices are unlikely, as a lower energy configuration would be to have coiled antiparallel coils, a configuration where the two helices are slightly intertwining. This is shown in Figure 4.1.

The filling of holes with the knobs of another helix (both idealized) led Crick to predict that the two helices should have an interhelical angle in the neighbourhood of 20° (as well as a suboptimal packing angle of -70°), and that the two helices could wrap around one another indefinitely at this angle [Cri53]. The model makes intuitive sense even 50 years later: consider that the side chains of the amino acid residues are ‘knobs,’ which correspond to the van der Waal’s shells of the atom composing the chain. These can fit into the ‘holes’ left between the ‘knobs’ forming the steric surface of the helix, as shown in Figure 4.2.

Even at this early stage in the field, Crick recognized that the directionality of the helix axes is significant, although for the wrong reasons. He indicated that if we are to have interdigitation[1], the situation where the helix axes are running anti-parallel would be much preferable to one where the helices were running parallel, since the side chains are not perpendicular to the axis. This insight is still valuable today. When dealing with simple lines as a representation of the backbones of helices, the direction of the helix axes is irrelevant because they are symmetrical. This is not the case with the actual helices however, since the side chains are not normal to the axis and there is a clear directionality to the sequence of the atoms along the backbone, as shown in Figure 4.3. Due the intuitiveness of the knobs-into-holes model, it has

[1] Interdigitation refers to the regular filling of holes on one helix by the knobs on the other.

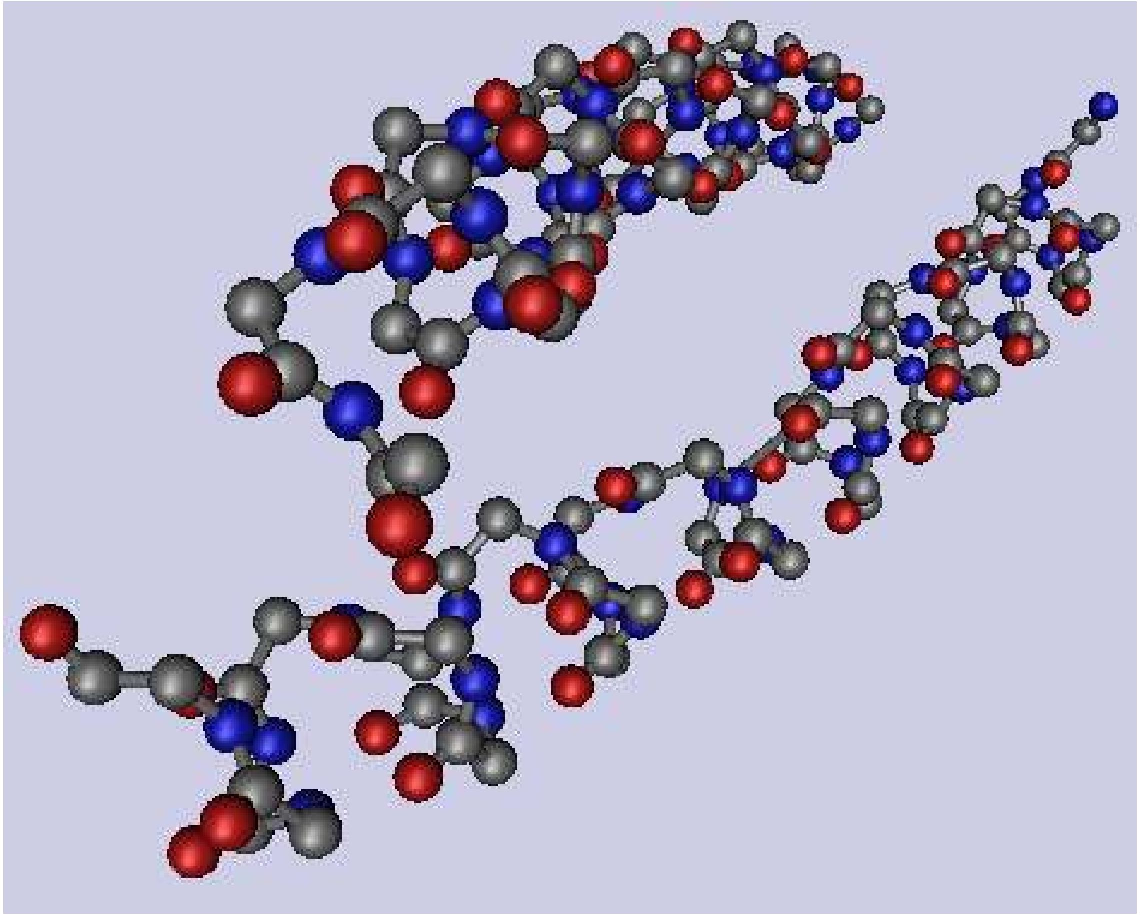


Figure 4.1: This is a pair of alpha helices (indices 1 and 2) from protein 1zii in the PDB. Notice how the two helices intertwine. The side chain atoms are hidden for clarity. Contrary to what Crick predicted, however, this helix pair is actually a stable parallel configuration with respect to the primary sequence. For more on this class of protein known as Leucine zippers, see [OKKA91].

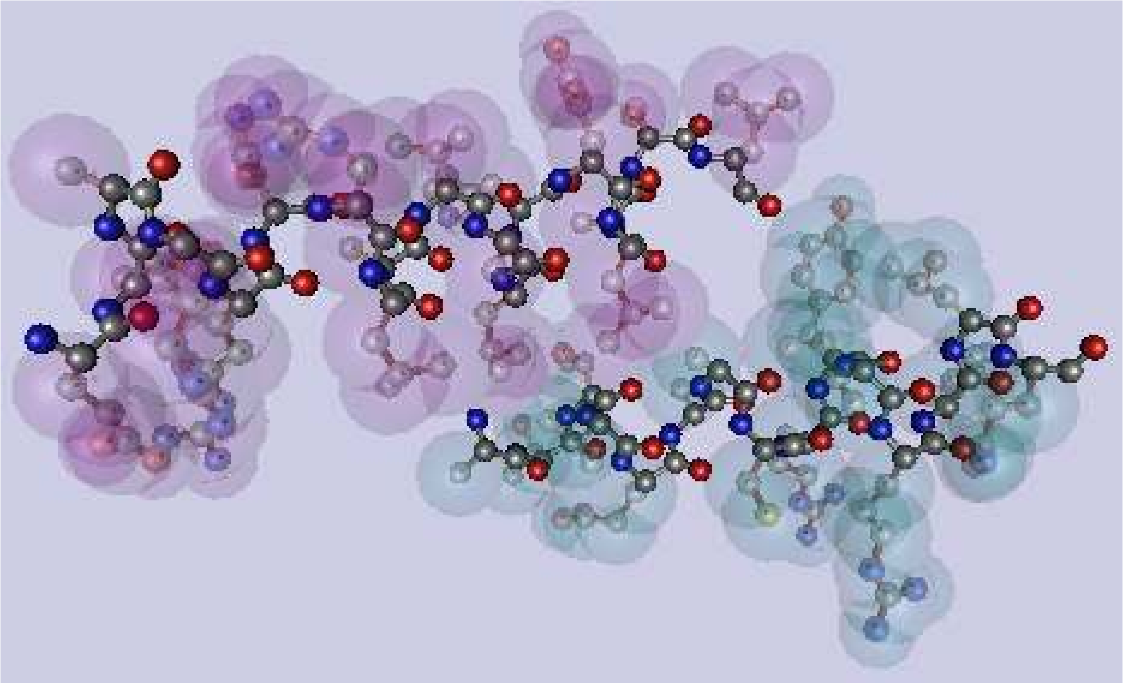


Figure 4.2: This is a pair of alpha helices (indices 2 and 4) from protein 1a0a. The sidechains are rendered with moderate opacity for clarity, and the van der Waal's shells of the side chain atoms are shown (none of the backbone shells are rendered) in a different colour for each helix. Notice the 'knobs' and 'holes' of each helix.

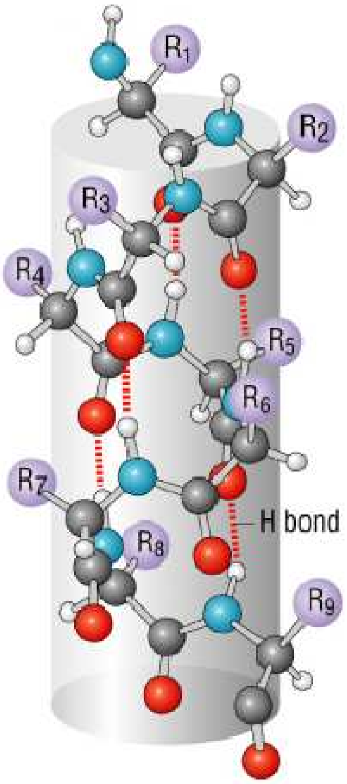


Figure 4.3: It is evident from this diagram that the side chains of an alpha helix are not perpendicular to the axis of the helix. Thus, the directionality of the helix axis is significant when considering packing. This figure has been adapted with permission from [PR04].

persisted and it continues to be used to explain the packing of alpha helices [WW03].

4.1.2 Hydrophobic Core Packing

The hydrophobic core packing model, first published by Efimov in 1979 [Efi79, Efi99], was a sort of marriage between the knobs into holes model with knowledge of the preferential folding of proteins with respect to the hydrophobicity of the alpha helices. Proteins tend to have a hydrophillic surface (as the solvent is water), and the cores

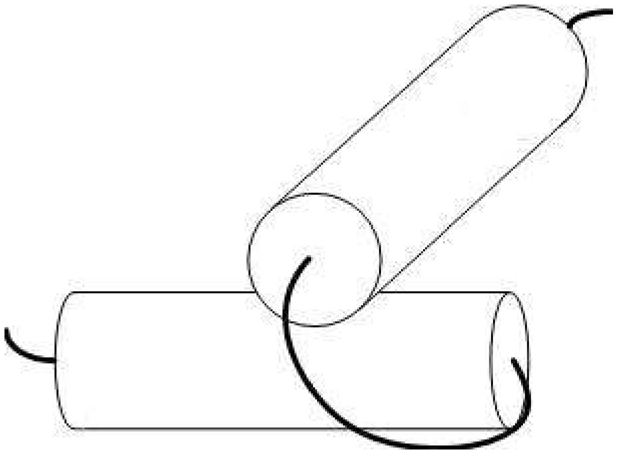


Figure 4.4: This Figure illustrates the $\alpha - \alpha$ corner. Most configurations of alpha helix pairs with near orthogonal interhelical angles are of this type. The figure has been adapted from [Efi93].

of globular proteins are hydrophobic. Using this knowledge, the rotamers of the side chains were predicted based upon the nature of the environment of the amino acid residues and the polarity of the residues themselves. This model was used to reinforce the results of Chothia et al. discussed earlier [Efi79].

In later work, Efimov [Efi93] studied the configurations of bundles of alpha helices to support his hydrophobic packing model. Among some of his other results was one that is quite relevant to the discussion at hand: pairs of alpha helices that pack at near orthogonal angles tend to be separated by only a turn in the primary structure, a structure termed an $\alpha - \alpha$ corner (shown in Figure **??**). This result is surprising, intuitively one might have expected that near orthogonal configurations would have been adopted predominantly by helices that were distant from one another in the primary sequence.

4.1.3 Ridges into Grooves

The work of Chothia, Levitt and Richardson [CLR81, CF90] expanded the knobs into holes model, extending the model to consist of ridges and grooves along the helices. Their model was developed empirically, using graphics software to observe the contacts occurring between residues on different secondary structures. The critical underlying assumption to their work is the following: the secondary structures pack in such a way that the van der Waal's energy is minimized and there is minimal steric strain.

This model is illustrated using a method created by Crick, where each helix is treated as a tube which is slit down the side and then unrolled so that it can be treated as a flat surface (notice that this method is still finding contemporary use, as in the Ptuba helix modelling package [LSD05]). The two surfaces representing each helix can be overlaid to illustrate the packing of the helix, as the interdigitation is clear, as elucidated in Figure 4.5.

It is clear from this model how the concept of ridges arose. Lines can be traced along the points in the sheet (representing amino acid residues) to connect them in straight lines. For example, point i can be connected to point $i \pm 3$, point $i \pm 6$, and so on. Now we can overlay two of these sheets, each representing a helix to see the different packing classes that may arise. This concept is illustrated in Figure 4.6. This was a strong departure from previous thinking, which held that most alpha helix pairs were aligned nearly parallel or anti-parallel, as described by Crick [Cri53] and Chothia and Levitt only a year prior to the publishing of this new model [LC76].

This model is highly dependent upon the regularity of the idealized structure. The

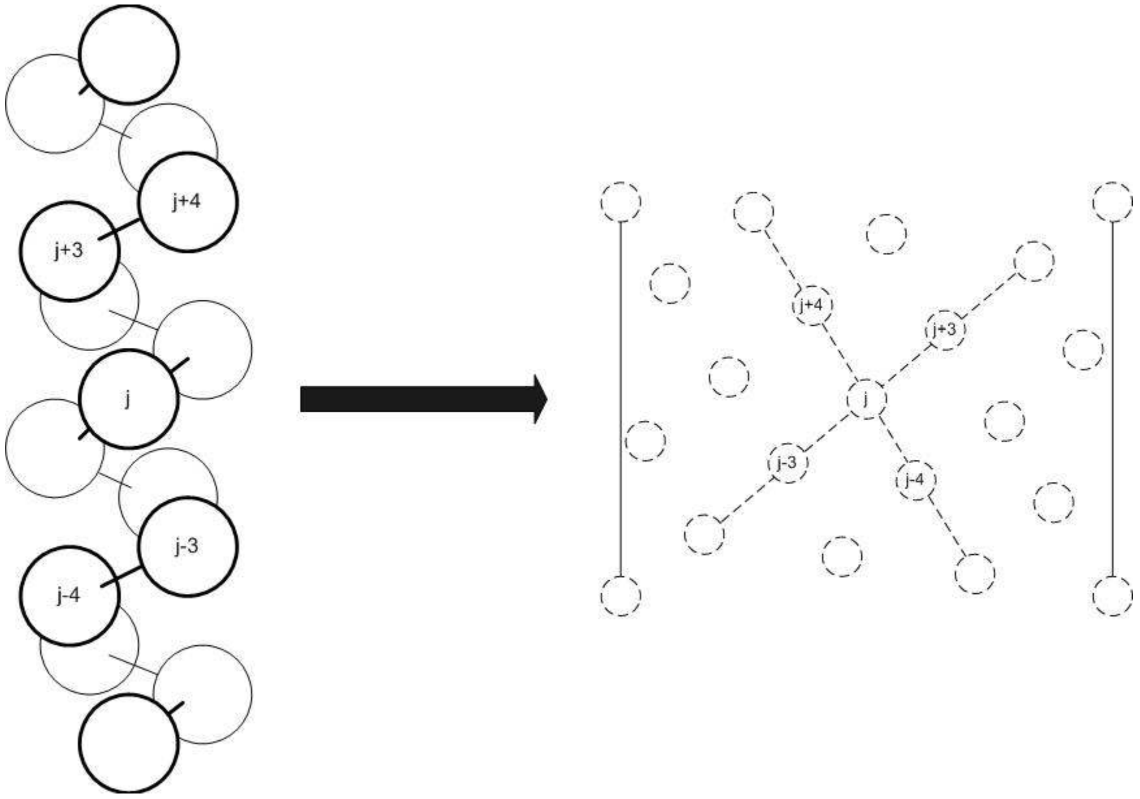


Figure 4.5: This shows how the alpha helix is sliced and each residue becomes a point on a flat sheet. Two possible ridges are illustrated, one through point j and then $j \pm 3$, and the other through $j \pm 4$. The figure has been adapted from [CLR81].

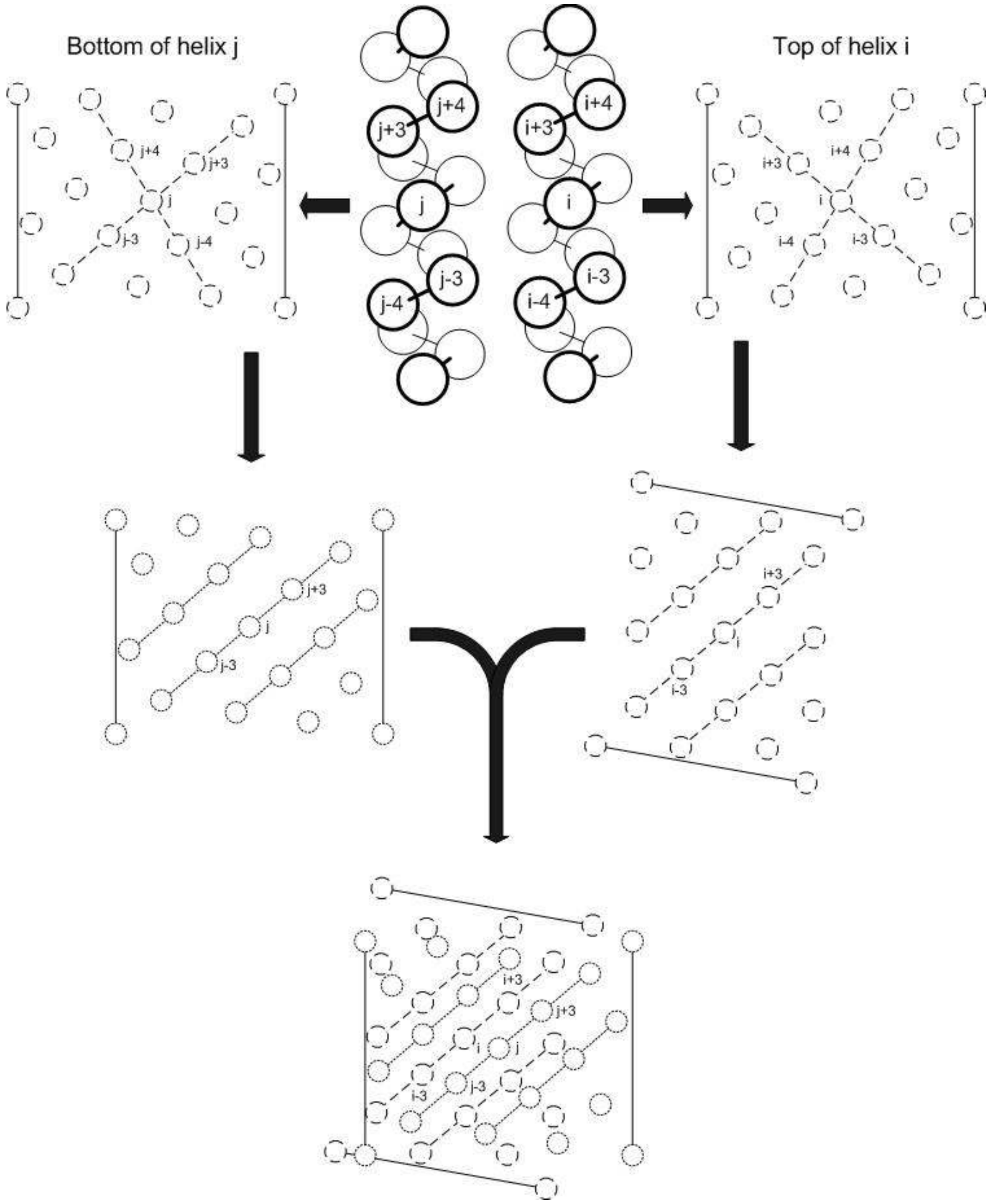


Figure 4.6: This figure illustrates the packing of unrolled helices. The different packing classes are evident from how the ridges (lines) and grooves of the helices come together. The figure has been adapted from [CLR81].

three primary classes identified for interhelical angles, occurring at -105°, -81° and -3° occur when there are 3.4 residues for each turn in the helix. If the helix is expanded slightly such that there are 3.8 residues per turn, the classes shift to -66°, -32° and +40°, so the classes are highly dependent on the helix structure. Notice that these are the three primary classes, and that others are possible. Also, the directionality of the helices is ignored, effectively creating a range of only 180° for the possible interhelical angles. A clear flaw with this model is that all side chains are considered equal since all amino acid residues are treated as simple points. This is a gross oversimplification, as can be seen from the variation in the volume of space occupied by the different side chains shown in Figure 4.2.

Chothia et al. [CLR77, CLR81] met with some success when verifying preferential packing angles empirically based on this model. 50 pairs of alpha helices were examined and fit to these classes. The precise methods of calculating the interhelical angles and the results of their study will be discussed in greater detail in section 4.3 on page 72.

4.1.4 Packing of Surface Contact Areas

Richmond and Richards [RR78] took a geometrical approach to explain the packing of alpha helices. The first step to creating their model is to determine the surface of the alpha helices. Using a spherical probe with a radius varying from 0 to 1.4Å (where 1.4Å was the previous convention, representing the radius of a water molecule), the surface could be mapped. At 0Å, the probe is exploring the true surface, essentially the van der Waal's shells of all the atoms composing the helix. At 1.4Å, the result is the solvent accessible surface of the helix. The helices were then packed in order to minimize the solvent accessible surface of the pair.

Another approach used in the paper is to unroll the helices, as Crick [Cri53] and Chothia et al. [CLR81] did, and then describe the amino acid residues each as a large sphere, which are then packed to create the model which they call the close-packed spheres model. The use of this model is very similar to the approach of Chothia et al. They emphasize the importance of the residues in the $i \pm 3$ and $i \pm 4$ positions in the packing of helices. The packed spheres model is used to explain that there are narrow ranges of interhelical angles at which the helices should pack, although this claim is tempered by stating that large side chains on the central residues could affect the packing significantly. Based on this, they derived a set of guidelines for amino acid residue species-specific packing. Only the residue central to the contact area and the surrounding six residues are considered. For each packing class identified by Chothia et al., they list the residue species that are likely to be found in the central position (Recent work has expanded on this line of study, and found that residues in the interface region tend to have shorter side chains [JV00, JV04]). These models were significant, because they were the first to examine the effects of different species on the packing of the helices, rather than using purely idealized structures [RR78]. The packed spheres model was subsequently tested to determine how well it could predict the tertiary structure of protein structure composed of only helical secondary structures (Myoglobin in this case) [CRR79]. The results were promising, but required extensive tweaking of multiple parameters to achieve, which is why this is not a solved problem.

4.1.5 Lattice Superposition Model

The lattice superposition model, introduced by Walther et al. [WEA96], is a further refinement of the "knobs into holes" model, which incorporates the hydrophobic core packing model as well as other physical principles of protein chemistry. The result is that three optimal packing angles are found, which are very similar to those found in the "ridges into grooves" model, while attempting to explain variations away from the ideal packing angles. Given hydrophobic side chains, the packing should be such that the surface area of these side chains is minimized.

Three classes of packing are identified, characterized by angles of -37.1° (which is the same as 142.9° if directionality is ignored), -97.4° (82.6°), and 22° (-158°). The essence of the algorithm was the relaxation of the ridges into grooves model so that there can be the crossing of ridges. Given properties such as the hydrophobic packing mentioned earlier, hydrogen bonding between the helices, disulphide bonds and salt bridges, they determine that there may need to be some shifting of the lattices so that there is no longer an ideal packing as published by Chothia et al. [CLR81] The result is the concept of a local packing angle rather than a global one, and using this technique, much cleaner results were achieved than had been seen earlier [WEA96].

4.1.6 Other Models

There are numerous other models that have surfaced over the years that are worth mentioning, but without going into great detail. Reddy and Blundell [RB93] showed that there are preferred packing angles for helix pairs, and they created a function to predict the interhelical distances for pairs of alpha helices in each class. This was accomplished using the volume of the residues forming the interface region between the pair. Chou, Némethy and Scheraga [CNS83] identified ten packing classes by identifying the low energy configurations of pairs of helices, with a clear preference for nearly parallel configurations. The method was to minimize the energy of the system, and in particular the non-bonded interactions in the form of van der Waal's forces, with a lesser emphasis on electrostatic interactions. This energy minimization approach has been applied more recently by applying docking approaches to helices (docking is typically used to determine how two complete proteins interact) [JTV03], and purely geometrical docking approaches have also been applied [ACHC97]. Murzin and Finkelstein (and later Chothia) [MF88, Cho89, CF90, CHB^{+}97] demonstrated a model where helices were treated as rigid cylinders (ignoring all physical properties of the helices) and attempted to form polyhedra composed of multiple helices (cylinders) which were quasi-spherical. They applied this method to known protein structures, and found that their predictions were within 20° of the actual interhelical angles.

4.2 Observed Packing Preferences

There is also a field of research dedicated to normalizing the distributions that are observed in nature, as it is thought that there are inherent biases in the databases collected to date because of small sample sizes. This line of thought was pioneered by Bowie [Bow97a], who contended that all observed preferential packings were were due to statistical biases in the observed helices, and that none of the described models accounts for the true nature of preferential packing angles. The basic crux of the argument is that there is a greater statistical likelihood of there being a large (closer to perpendicular) interhelical angle than otherwise (see Figure 4.7). Therefore, the distribution must be normalized to remove the bias. In two dimensions, all angles have an equal probability of occurrence, but in three dimensions there are much more possibilities for a helix of a greater interhelical angle. The circumference of a circle transcribed by the rotation of an axis around another will increase with the sine of the interhelical angle.

This, naturally, is a purely geometrical analysis, and does not take into account any of the physical properties of the helices themselves. To elaborate on this point, a graph showing the theoretical distribution can be created using the following equation:

$$p_i = \frac{cos(\Omega'_i) - cos(\Omega''_i)}{\sum_{j=1}^{N} cos(\Omega'_j) - cos(\Omega''_j)} \tag{4.1}$$

where

p_i is the probability of having the interhelical angle in bin i
Ω'_i is the minimum angle for bin i
Ω''_i is the maximum angle for bin i
N is the number of bins in the distribution; 18 is conventional

This distribution is shown graphically in Figure 4.8. To validate the theoretical distribution, the angles of pairs of alpha helices *not* in contact were found and plotted against the theoretical distribution. The rationale is that helices not in contact should have no influence on each other, and the interhelical angle should be randomized.

This indicates that there is an inherent bias present in the distributions of interhelical angles. Figure 4.8 indicates that helices that are not in contact tend to be found at angles that are nearer to normal than parallel. This is not due to any sort of preference to do so, because these helices are not in contact, but due to the probability

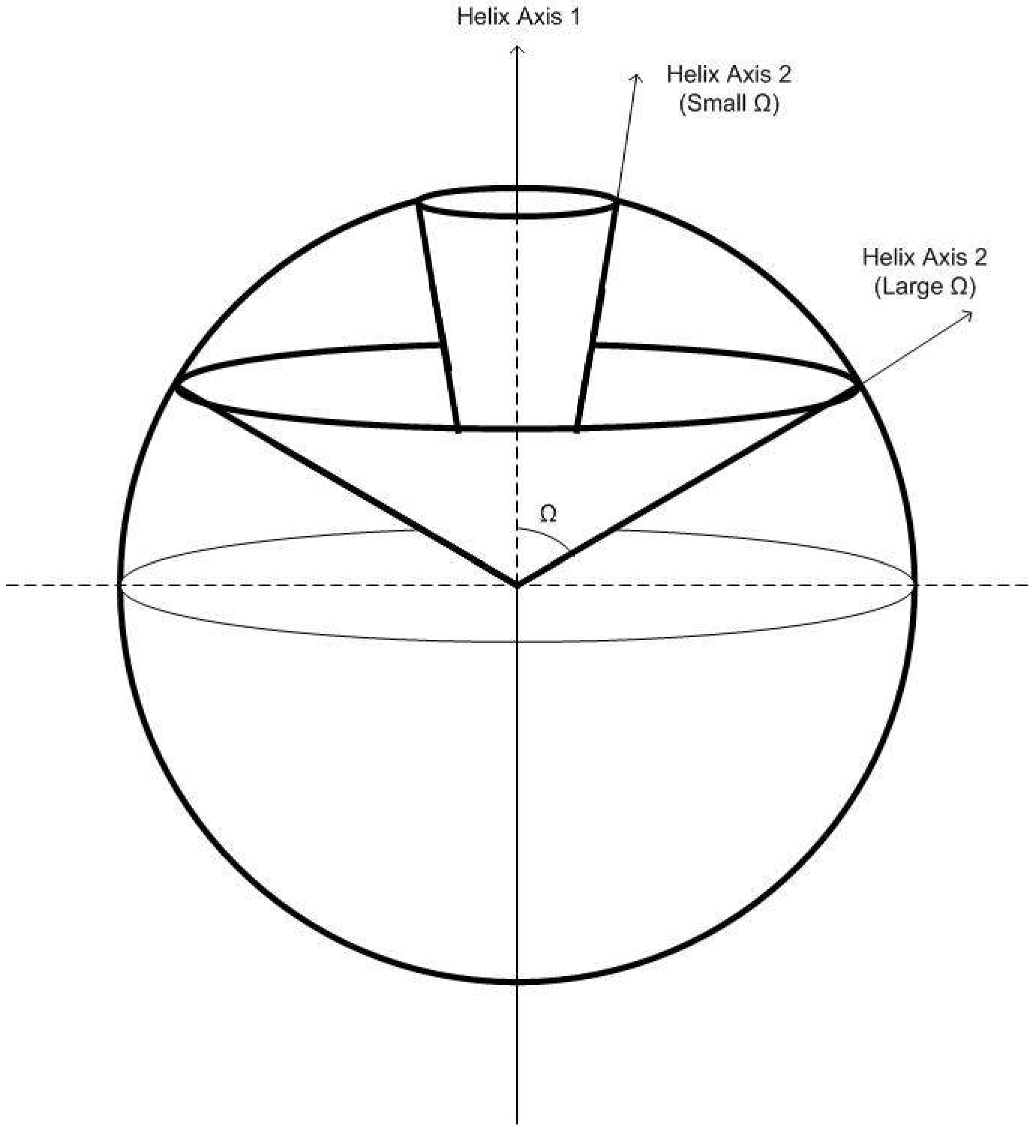


Figure 4.7: Each cone represents the circle that a helix axis at a given interhelical angle transcribes on the surface of a sphere. The figure has been adapted from [Bow97a].

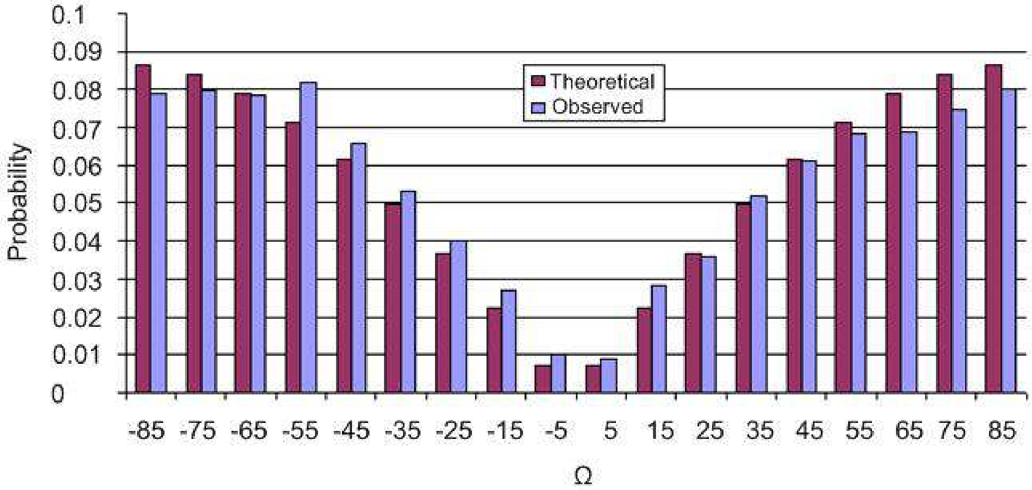


Figure 4.8: The theoretical distribution for bins of 10° is shown along with the values observed from a database of 12,605 alpha helix pairs that were *at least* 15Å apart, and at least 20 amino acid residues apart along the backbone. The figure has been adapted from [Bow97a].

of these angles occurring. Now, given the observation of helices that are in contact, Bowie contends that it is necessary to divide by the distribution shown in Figure 4.8 to normalize the data, removing the inherent statistical bias. The application of this correction is shown in Figure 4.9.

This corrected distribution is used to explain that none of the models explained earlier sufficiently account for this “true” distribution. The essence of the argument is that unrolling two helices into flat surfaces in order to see how they pack, particularly a lattice model where a residue is represented only by a backbone carbon atom, is fallacious. The actual interface between the helices is only a small part of the lattice. Due to this, it is claimed that there are not really any preferences that will be found for packing, because local steric effects will be the dominant effect on interhelical angle [Bow97a].

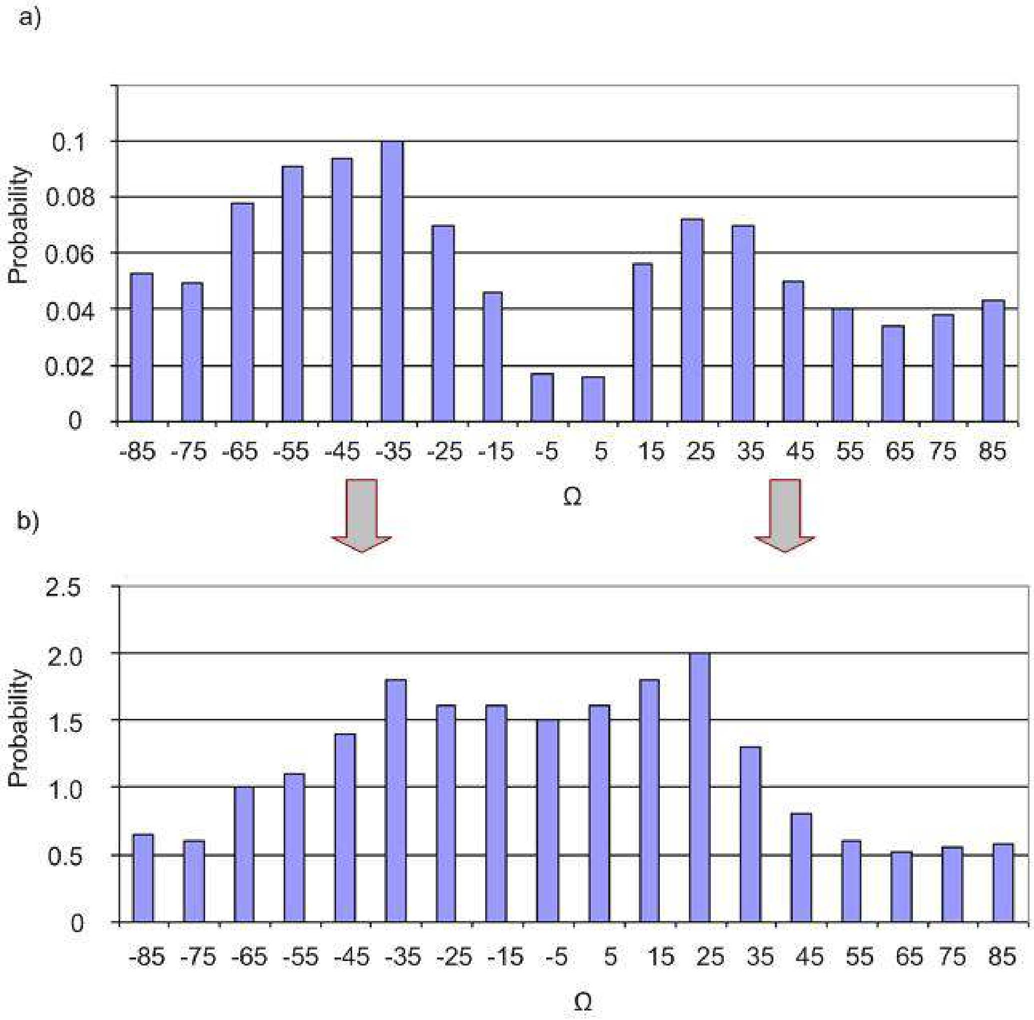


Figure 4.9: a) This is the observed distribution of alpha helix pairs which are in contact. 2145 pairs were used, and the definitions of Ω and contact were the same as those used by Chothia et al. As with the previous distribution, pairs were rejected if there were less than 20 residues along the backbone between the two helices. b) This "corrected" distribution was obtained by dividing the distribution in part a) by the distribution observed in Figure 4.8. Notice that there are not any clear packing preferences anymore, just a general preference for near parallel configurations. A probability value of 1 means the observed probability of the angle is what would be expected by chance, while values greater than 1 correspond to angles that are observed more often than is expected. The figure has been adapted from [Bow97a].

This statistical correction has been adopted by many in the field, although several have modified it in order to prove that statistical biases do in fact exist. Walther et al. [WSC98] responded that the helices used by Bowie were treated as essentially being of infinite length, and that near perpendicular packing preferences return once the axes are treated as finite. Hespenheide and Kuhn [HK03] used Bowie's tenet to correct the interhelical angle distribution that they observed, and concluded that there is a preference for parallel and anti-parallel configurations. Trovato and Seno [TS04] achieved similar results, performed by dividing the types of helix pairs into classes based upon their orientation[2]. The standard class is one where the line of closest approach intersects both helix axes between their endpoints. The other three classes are ones where this occurs only with one or neither of the two helices, as shown in Figure 4.10 on the next page. A massive database of random helix pairs was built, and the distribution of interhelical angles was used as a reference distribution. By separating the reference data into these classes and eliminating pairs which, if not belonging to the first class are within a very small threshold, the corrected unbiased histogram maintains packing preferences that are close to those predicted by packing models.

The recent work of Lee and Chirikjian [LC04] illustrates a correlation between the interhelical angles of alpha helix pairs and the distance between the pairs. The limitations of this work are an overly complex method of determining the interhelical angle using rigid body transformations and the lack of an evaluation of their algorithm. The nature of their study may have necessitated their approach, however, for

[2] Note that these are classes of configurations in space, something distinct from the classes of interhelical angles defined by Chothia et al., where they are referring to preferential packing angles.

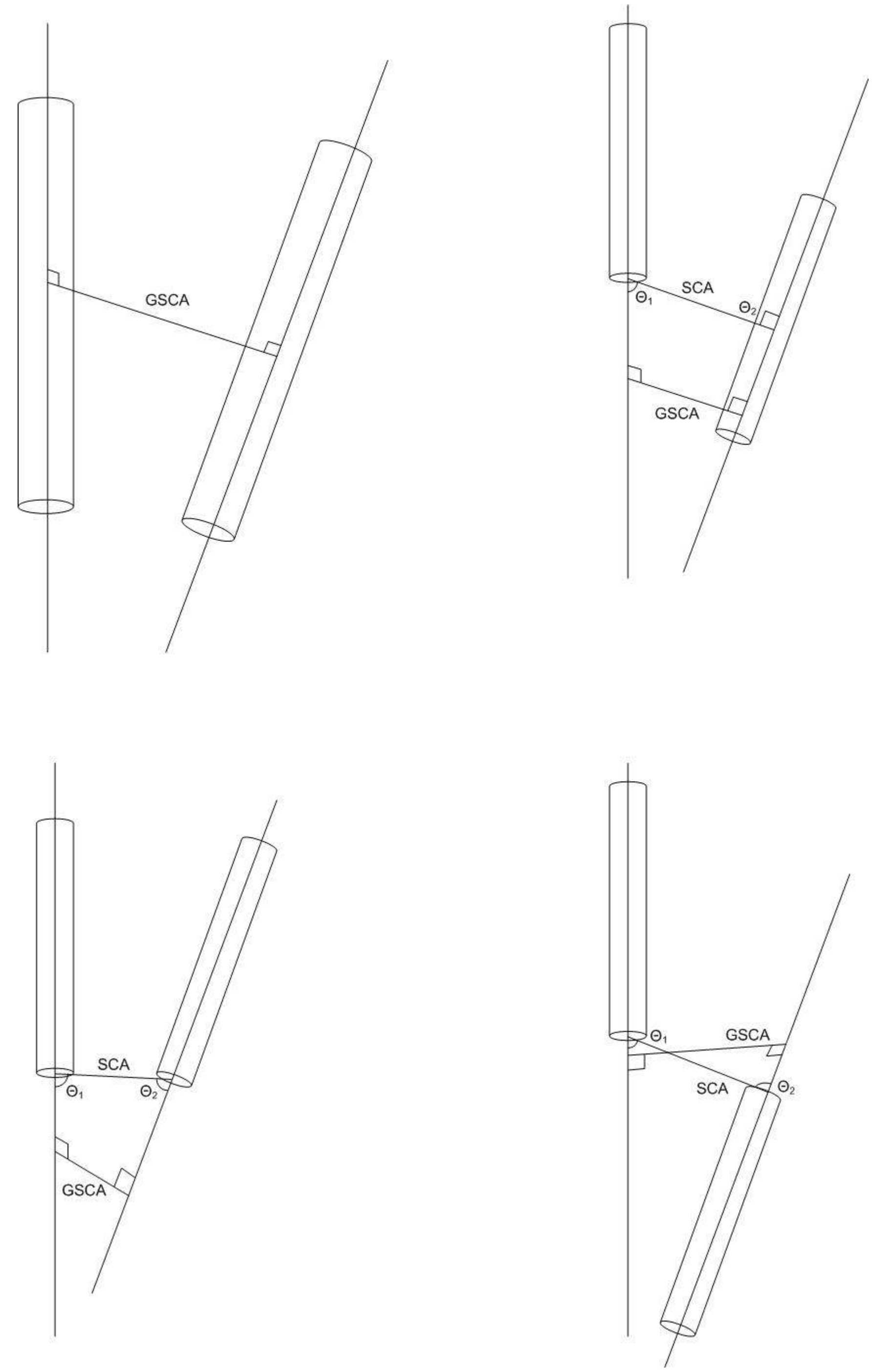


Figure 4.10: The four classes of interhelical configurations. The top left is the typical class for interacting helix pairs, where the global segment of closest approach (GSCA) is within the endpoints of both helices. The top right pair has the GSCA outside the endpoints of one helix, and the bottom figures have the GSCA intersecting both axes outside of the helices. The segment of closest approach (SCA) intersects the axes at the angles θ_1 and θ_2. In the top left pair, the GSCA and SCA are the same thing, and θ_1 and θ_2 are both normal. The figure has been adapted from [TS04].

most applications a simplified method would suffice.

4.3 Calculating the Interhelical Angle

For every packing model, there is usually an attempt to use it to explain observed distributions of interhelical angle distributions. There is a critical flaw with this field of research however, in that there is no standard algorithm that is used for determining the interhelical angle, and so the distributions observed in each paper often differ. Many studies do not even state the method used for calculating the angle. We will now study why there is no conventional means for calculating the interhelical angle at present. The calculation of interhelical angle can be divided into two essential steps. The first step is to determine an axis that represents the helix well, and the second step is to determine the angle from this axis. We will examine the methods that have been used for calculating each of these in the past, followed by a thorough explanation of the proposed method. For all methods, we will use only the coordinates of the alpha carbons of each amino acid as the source data for the calculations.

4.3.1 Axis Calculation Approaches

Axis Determination - Chothia et al.

We will begin this discussion with a return to the work of Chothia et al. [CLR81]. The first step in determining the axis using their method is to find vectors that are normal to the axis (yet to be determined, obviously). These vectors are defined as follows (for a graphical illustration, refer ahead to Figure 4.11 on page 76, where the vector is referred to as B_i rather than the P_i used here):

$$P_i = r\left(C^{\alpha}_{i,k}\right) + r\left(C^{\alpha}_{i+2,k}\right) - 2r\left(C^{\alpha}_{i+1,k}\right)$$

where

P_i is a vector approximately normal to the helix axis at residue i
$C^{\alpha}_{i,k}$ is the i^{th} alpha carbon on helix k

$r\left(C^{\alpha}_{i,k}\right)$ are the coordinates of the atom $C^{\alpha}_{i,k}$; in the future this will be simplified to $r_{i,k}$

The next step in the process is to determine an inertia matrix M from the distribution of these vectors, defined as

$$M = \begin{pmatrix} M_{x^2} & M_{xy} & M_{xz} \\ M_{xy} & M_{y^2} & M_{yz} \\ M_{xz} & M_{yz} & M_{z^2} \end{pmatrix}$$

where entry M_{xy} in the matrix is given by

$$M_{xy} = \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

where

x_i is the x coordinate of the i^{th} alpha carbon of the helix

$\bar{x}$ is the mean x coordinate of the helix, given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Now the axis vector a_k of the helix can be found by calculating the eigenvectors of M, and choosing the eigenvector with the smallest eigenvalue, which corresponds to the direction with the thinnest distribution of P_i values. Now that we have the axis vector, it must be assigned a position in the axis. This is accomplished by creating a line segment in the helix along the helix axis. The beginning point b_k and end point e_k of the axis for helix k are calculated using the following equations:

$$b_k = \bar{r}_k + [a_k \cdot (r_{1,k} - \bar{r}_k)] \times a_k$$

$$e_k = \bar{r}_k + [a_k \cdot (r_{n,k} - \bar{r}_k)] \times a_k$$

where

$\bar{r}_k$ is the centroid of helix k

We now have a straight line segment the length of the helix which describes the axis.

Axis Determination - Walther et al.

As alluded to earlier, Walther et al. [WEA96] begin the determination of the helix axis in a method similar to Chothia et al. The latter calculated vectors normal to the axis of the helix referred to as P_i; Walther et al. calculate the same vector in the same manner, but refer to it as B_i. The use of the vectors fundamentally differs however, as the aim is to determine a local helix axis for each amino acid residue rather than a single axis for the entire helix. This is found using an iterative process, and the first iteration estimates the local helix axis u_i as the cross product of two consecutive normals, as shown in Figure 4.11. Thus,

$$B_i = r_i + r_{i+2} - 2r_{i+1}$$

and

$$u_i = B_i \times B_{i+1}$$

For the helix axis for residue r_n (the one at the C-terminus of the helix), the axis vector of r_{n-1} is used again. In order to assign positions for the axis vectors, the geometric center of four consecutive residues around the present is calculated:

$$A_i = \frac{r_{i-1} + r_i + r_{i+1} + r_{i+2}}{4}$$

The formula needs to be adjusted for the ends of the helices. The axis vectors (u_i) are adjusted so that their lengths are all 1.5Å, which is the average rise per residue along the axis for an ideal alpha helix. The axis of the helix is now described by a series of line segments. The endpoints, b_i and e_i, of the local helix axis for residue r_i are given by:

$b_i = A_i$

$e_i = A_i + u_i$

The axes are smoothed using an iterative approach. The first step is to take the average of three consecutive helix axis vectors (two at the ends):

$$u_{i,smoothed} = \frac{u_{i-1} + u_i + u_{i+1}}{3}.$$

Now the average point coordinates A_i are adjusted by finding the midpoint between the beginning point of the current helix and the endpoint of the previous one:

$$A_{i,smoothed} = \frac{b_i + e_{i-1}}{2}$$

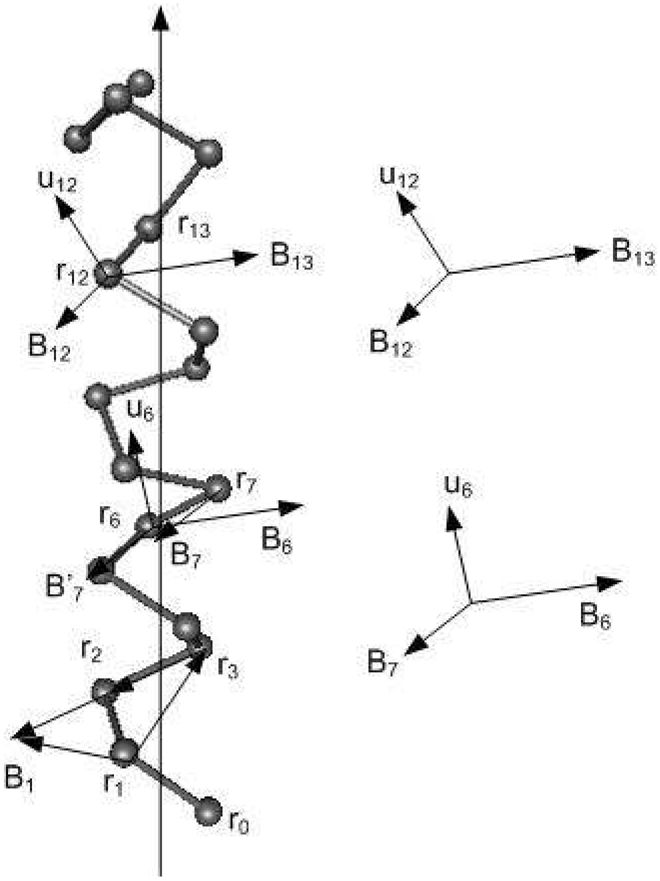


Figure 4.11: The method described by Walther et al. for determining local helix axis vectors. The vectors labelled B_i are the same as those referred to by Chothia et al. as P_i previously. The local axis vectors u_i are calculated in the first iteration by taking the cross product of two of these normals consecutively.

This smoothing process is repeated three times; the result is a series of local helix axis line segments which approximate the curve of the helix.

Axis Determination - Trovato and Seno

Trovato and Seno [TS04] use a modified version of the Walther et al. [WEA96] method of local helix axes. Diverging from the Chothia et al. [CLR81] method of finding the normal vectors to the axis, Trovato and Seno introduce a new method based upon the "bond" vectors b_i between successive alpha carbon positions:

$$b_i = \frac{r_{i+1} - r_i}{|r_{i+1} - r_i|}$$

Figure 4.12 illustrates these points and vectors. Now the axis vector a_i and a set of three orthonormal vectors t_i, v_i and u_i associated with each alpha carbon are defined as follows (also note that all vectors are normalized):

$$t_i = \frac{b_i + b_{i+1}}{|b_i + b_{i+1}|}$$

$$v_i = \frac{b_i \times b_{i-1}}{|b_i \times b_{i-1}|}$$

$$u_i = t_i \times v_i$$

$$a_i = u_i \times u_{i+1}$$

The algorithm proceeds in much the same manner as Walther et al., except that the vectors are initially set to a length of 1.45Å, and only two iterations of the smoothing procedure are performed. The coordinates of the local helix axis

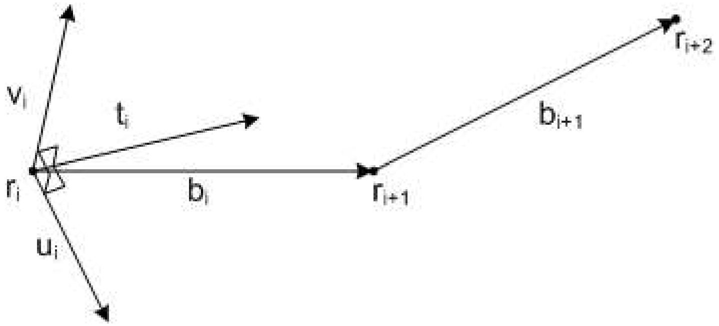


Figure 4.12: The three orthonormal vectors (t_i,v_i, and u_i) associated with an alpha carbon r_i in the Travato and Seno [TS04] axis determination method are illustrated here. The axis vector a_i (not shown) is given by the cross product of u_i and u_{i+1}.

are calculated in an identical manner as Walther et al. Notice that the nomenclature has changed[3], and that u_i here corresponds to what Walther et al. referred to as B_i, and the local axis vector here is defined as a_i, while Walther et al. defined this vector as u_i. This approach of using local axes has been adopted by many, as it permits the calculation of the angle between the helices at the point of closest approach; a local axis is essentially a tangent at that point. Bansal et al. [BKV00] created a program called HELANAL which uses local axes to characterize the shapes of helices as either straight, curved or kinked, based upon the difference between successive axis vectors. The HELANAL program is available online at `http://www-lecb.ncifcrf.gov/~kumarsan/HELANAL/helanal.html`, and has found use as a helix axis calculation tool since its release [DMW03, ED05a].

[3] It is clearly confusing to have the same names for different vectors depending on the method, but it was assumed that it would be preferred to maintain the original nomenclature used by the authors so that referring to their literature would be more clear.

Axis Calculation - Rotational Least Squares

The rotational least squares approach to calculating the axis of a helix involves first constructing an idealized helix with the same number of residues as the target. The ideal helix provides a perfect reference axis, since the positions of the ideal helix are constructed around the axis that defines it! This method was introduced by McLachlan [McL79]. He first calculates the center of gravity of the two helices and aligns them. Now the target helix is aligned with the idealized one, minimizing the rmsd between the points by performing a series of Euler rotations about the center of gravity. This approach was later refined by MacKay [Mac84], who replaced the Euler angle approach (which involved the simultaneous solution of nine parameters) with a cleaner algorithm using a quaternion transformation.

The application of quaternions for these rotations will follow the discussion of MacKay [Mac84]. A quaternion is a description of a rotation which requires only four values, represented as a scalar and a vector. Given a unit quaternion Q with the scalar p_s and the vector p such that $p_s^2 + |p|^2 = 1$, a unit vector r is related to r' by a rotation of θ about an axis with the direction cosines l, m, and n. p can be described in terms of the orthonormal unit vectors i, j, and k, where

$$i^2 = j^2 = k^2 = ijk = -1.$$

Note that these values are associative, such that $ij \neq ji$. In the application of quaternions to alpha helices, each alpha carbon atom is an instance of r, and the corresponding atom on the idealized helix aligned with the x-axis is the instance of r'. Rather than r corresponding to the coordinates of the alpha carbon atom, each r is a unit vector with the origin at the centre of mass. Since there are many pairs of vectors to solve for, a unique set of values for Q can be derived. We find r' from r by

$$r' = Q^{-1} r Q$$

where

$$Q = \cos\frac{\theta}{2} + l \sin\frac{\theta}{2i} + m \sin\frac{\theta}{2j} + n \sin\frac{\theta}{2k}$$

and

$$Q^{-1} = \cos\frac{\theta}{2} - l \sin\frac{\theta}{2i} - m \sin\frac{\theta}{2j} - n \sin\frac{\theta}{2k}$$

Note that $Q^{-1}Q = 1$ and $Q = \cos\frac{\theta}{2} + \sin\frac{\theta}{2\mathbf{n}}$, where $\mathbf{n}$ is the unit vector along the axis of rotation. Multiplying the quaternions Q and Q^{-1} with the vector r yields

$$r' = \left(\cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2}\right) r - \sin\theta\,(\mathbf{n} \times r) + 2\sin^2\frac{\theta}{2\,(\mathbf{n} \cdot r)\,\mathbf{n}}$$

Now we are able to find r' from r by plugging in the values of θ and l, m, and n. We find these values by performing least squares on a series of equations. For each pair of points r and r', it can be shown that

$$m \tan\frac{\theta}{2}(z + z') - n \tan\frac{\theta}{2}(y + y') = (x' - x)$$

$$-l \tan\frac{\theta}{2}(z + z') + n \tan\frac{\theta}{2}(x + x') = (y' - y)$$

$$l \tan\frac{\theta}{2}(y + y') - m \tan\frac{\theta}{2}(x + x') = (z' - z)$$

where

$$r = (x, y, z)$$

and

$$r' = (x', y', z')$$

Given N atoms in the helix, there are $3N$ equations to solve for. In order to perform the least squares operation, MacKay [Mac84] uses the matrix operation $AX = H$ to produce $X = \left[A^TA\right]^{-1} A^TH$; note that $\left[A^TA\right]^{-1} A^T$ is the pseudo inverse of A. This equation can be solved for the three unknowns $l \tan\frac{\theta}{2}$, $m \tan\frac{\theta}{2}$, and $n \tan\frac{\theta}{2}$. Taking the squared sum of these values gives $\tan^2\frac{\theta}{2}$, and given this l, m, and n can be found by division. The only contingency to this algorithm is that there exists a degeneracy at $\theta = 180°$, as at this value $\tan\frac{\theta}{2}$ is infinite. The addition of a check for this instance can account for it by setting Q to (1,0,0,0) in the event. One advantage of this approach is the inherent simplicity once the algorithm is understood. MacKay implemented the algorithm in Basic in 150 lines of code[4].

Axis Calculation - Comparing the Methods

Conveniently, an exhaustive comparative study of axis calculation methods has been performed by Christopher et al. [CSB96]. In this study, they compare the effectiveness of methods including those discussed here, in addition to several others:

- a parametric least squares method that is simple and easy to implement.
- a moment matrix method, which is solved by finding the eigenvectors that describe a plane of best fit. As with the previous method, this is simply a three-dimensional linear regression approach.
- a method where the points on the helix (in our case the alpha carbons) are fit to a cylinder. This method was introduced by Åqvist [Åqv86].

[4] This is also a testament to his programming skill, or my lack thereof, since I used about 160 lines of Matlab.

- a cross product of triad bisectors method. In this method, three consecutive points are used to find a vector normal to the as yet unknown helix axis. The cross product of consecutive such normals gives a local helix axis. To obtain a single helix axis, three-dimensional linear regression is applied to these local axes. This is the type of approach used by Chothia et al. [CLR81] and Walther et al. [WEA96]. In [CSB96], the work of Kahn [Kah89] is cited. The primary difference between Kahn and the others is that Kahn uses the triad vectors to determine the position of the axis, whereas the others simply compute the average positions of the residues.
- a rotational least squares method that was described by MacKay [Mac84]. This method requires that a set of rotations are performed such that the helix is aligned with an axis-aligned standard helix, for which the helix axis is easily defined. Christopher et al. [CSB96] used a variation of this approach, using a mapping of the helix to itself, one atom out of register. In other words, r_i is mapped to r_{i+1} to obtain a screw transform. They contended that this method would be more accurate for less ideal helices.

The objective of the work by Christopher et al. was to determine which of these methods would most accurately describe the axis of a helix, with the aim being to use the algorithm on alpha helices and A-form DNA helices. Gaussian noise was introduced to the positions of the atoms of idealized helices, and varying amounts of curvature were added to determine the effects on the accuracy of each algorithm. The first two methods, understandably, were subject to greater error than the other methods since they are simple linear regressions. The result is a line of best fit rather than a true helix axis. If there is an unbalanced number of atoms on opposite sides of the helix, the axis can be affected. The accuracy of the regression methods became comparable to the accuracy of the other methods for alpha helices with greater than 25 residues, but helices of this length are more the exception than the norm. For short helices, the error for the parametric least squares method and the moment matrix method were both significant, even on ideal helices. Their use should be discouraged in applications where accuracy is an important issue.

The method introduced by Åqvist performed well relative to the other methods, however it is a costly algorithm. First, an initial estimate of the helix axis is made, followed by the minimization of six parameters associated with each point on the helix. Both this method and the product of triad bisectors approach are accurate for idealized helices, but they begin to suffer when the position of the residues vary or when the helices are curved. This leaves the rotational least squares method. Christopher et al. concluded that this algorithm was better than the others in nearly every instance where one of the approaches was significantly better than the others. They qualify the results with a warning that the study was performed using modelled data since there is no definite answer for what the true axis for a natural helix is, but they recommend using the rotational least squares method nonetheless on the basis of their findings. Paradoxically, no papers were found in this survey which used this approach.

4.3.2 Angle Calculation Approaches

Angle Calculation - Chothia et al.

Recall that the axes produced by the Chothia et al. method were straight line segments. We have the beginning and end points of the line, b_k and e_k, and so using a scalar coefficient S_k we can define any point along the line defined by the axis as:

$$t_k = b_k + S_k(e_k - b_k),$$

and the point t_k will be on the axis within the helix if S_k is between 0 and 1 (which correspond to b_k and e_k respectively. There exist terms S_k and S_j can be found for helices k and j such that the length of the line $(t_k - t_j)$ is minimized. This corresponds to the global segment of closest approach. In order to determine the coefficients, we must first find an expression for the distance between the two points of closest approach. The distance d between the points t_k and t_j can be expressed as

$$d^2 = (t_k - t_j) \cdot (t_k - t_j) = (t_k - t_j)^2$$

Now we can substitute the line equation derived earlier to express the distance as

$$d^2 = (b_k + S_k\,(e_k - b_k) - b_j - S_j\,(e_j - b_j))^2$$

By minimizing the value of d^2 with respect to the coefficients S_k and S_j, their value can be found.

$$\frac{\partial(d^2)}{\partial S_k} = 0 = (b_k + S_k\,(e_k - b_k) - b_j - S_j\,(e_j - b_j)) \cdot (e_k - b_k)$$

$$\frac{\partial(d^2)}{\partial S_j} = 0 = (b_k + S_k\,(e_k - b_k) - b_j - S_j\,(e_j - b_j)) \cdot (e_j - b_j)$$

Using this pair of equations, we can create expressions solving for one coefficient in terms of the other:

$$S_k = 0 = -\frac{(b_k - b_j - S_j\,(e_j - b_j)) \cdot (e_k - b_k)}{(e_k - b_k)^2} \tag{4.2}$$

$$S_j = 0 = \frac{(b_k + S_k\,(e_k - b_k) - b_j) \cdot (e_j - b_j)}{(e_j - b_j)^2} \tag{4.3}$$

Solving both equations simultaneously yields:

$$S_k = \frac{-W_1U_{22} + W_2U_{12}}{Det}$$

$$S_j = \frac{W_2U_{11} - W_1U_{12}}{Det}$$

where

$$W_1 = (b_k - b_j) \cdot (e_k - b_k)$$

$$W_2 = (b_k - b_j) \cdot (e_j - b_j)$$

$$U_{11} = (e_k - b_k) \cdot (e_k - b_k)$$

$$U_{12} = (e_k - b_k) \cdot (e_j - b_j)$$

$$U_{22} = (e_j - b_j) \cdot (e_j - b_j)$$

$$Det = U_{11}U_{22} - U_{12}^2$$

Now we substitute the coefficients back into the line equations to find t_k and t_j, the points of closest approach on each line[5]:

$$t_k = b_k + S_k(e_k - b_k)$$

$$t_j = b_j + S_j(e_j - b_j)$$

Now we have points t_k and t_j, the endpoints of the line of closest approach. Finally, the interhelical angle is calculated by finding the dihedral angle of the points $[b_k \quad t_k \quad t_j \quad b_j]$[6]. This approach was used by Bowie in his work [Bow97a], among others [SSS+95, PMP99].

[5] Note that at this point this line is the global segment of closest approach, and thus t_k will be outside the line segment described by b_k and e_j if S_k is greater than 1 or less than 0.

[6] For the moment, we will assume that the global segment of closest approach and the segment of closest approach are one and the same.

Angle Calculation - Walther et al.

Recall that Walther et al. [WEA96] produced a series of local axes corresponding to each residue in each helix. In order to determine the interhelical angle for the pair, they first determine the point of closest contact on each local helix axis with all those from the other helix, and take the global minimum. Once these are found, the algorithm proceeds in much the same way as with Chothia et al. [CLR81], where the angle can be calculated as a dihedral angle. All pairs where the line of closest approach intersected the endpoint of a helix segment and the angle τ (to be discussed momentarily) was greater than 5° were discarded, as these did not represent packing pairs. This method of calculating the angle has been used by most others since [DMW03].

The Angle τ

The angle τ becomes relevant when the pair of helices of interest do not have the intersection of the global segment of closest approach within the endpoints of the line segment representing the helix axis. Obviously, this angle only occurs with finite helix axes[7]. When the segment of closest approach differs from the global segment of closest approach, as shown in Figure 4.10 on page 71, the angle θ between the segment of closest approach and the helix axis differs from the perpendicular. The amount of this difference is expressed as τ. An extensive survey of the literature revealed no satisfactory method for calculating τ.

The first requirement is to find the points of closest approach on each helix axis such that the points are within each helix. We will continue from where we left off

[7] This is not to be confused with the application of τ in transmembrane proteins, which is vexingly also referring to a tilt angle, but in the application we are not interested in it refers to the tilt of the axis of an alpha helix to the normal of the membrane. A clear discussion of this property is given by Bowie [Bow97b]. τ will refer to the prior definition in the context of this thesis.

with the Chothia et al. [CLR81] method. It is suggested that in order to find these points, if S_k is greater than 1, set it to be 1, and if it is less than 0, set it to be 0. Intuitively, this approach may make sense, but there is an inherent problem as illustrated in Figure 4.13.

A correction to this method to calculate the points of closest approach is to use this step as an initial iteration, but then to use equations 4.2 and 4.3 to recompute the other value. There are several scenarios that could occur, as outlined in Algorithm 1. The algorithm will rarely require going to the third conditional statement, but it is conceivable. This algorithm could be improved.

Algorithm 1 Iteratively finding the closest points

```
if $t_k$ is outside helix then
  move $t_k$ to the nearest endpoint of helix $k$
  recompute $t_j$
  if $t_j$ is outside helix then
    move $t_j$ to the nearest endpoint of helix $j$
    recompute $t_k$
    if $t_k$ is outside helix then
      move $t_k$ to the nearest endpoint of helix $k$
    end if
  end if
end if
```

A purely geometric and intuitive approach to finding the points of closest approach and interhelical angle is recommended. It is simple to implement, particularly if the helix axes were derived using the rotational least squares method. Assuming that the necessary transformations for each helix have been stored, it is simple to apply them to both in turn to find the angle directly. This algorithm is outlined formally in Algorithm 2.

This approach begins by selecting one helix, say it has axis j. In this first iteration,

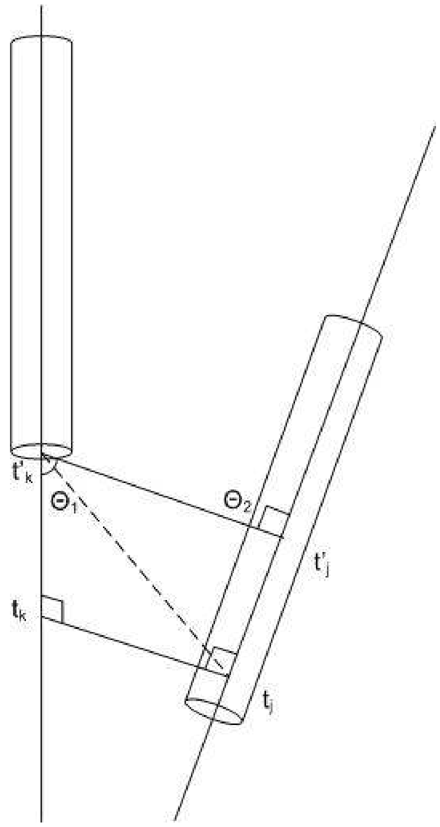


Figure 4.13: The global segment of closest approach $(t_k - t_j)$ is perpendicular to both helix axes, but lies outside of one of the helices. Simply moving t_k to the endpoint of the helix at t'_k is incorrect, as this would create a line of closest approach from t'_k to t_j, shown as a dashed line in the figure, and the τ values would be higher than actual. Thus, t'_j must be found, and in the case of this example, it would remain in the helix and the line of closest approach remains perpendicular to this axis.

Algorithm 2 Geometrical approach to finding the points of closest approach

```
for the helix pair, define the axes $j$ and $k$
in the first iteration $i$=$j$ and $\neg i$=$k$; vice versa in the second
for each helix axis $i$ do
    translate both axes with the translation used to align helix $i$ with the axis-aligned helix in the axis determination step earlier
    rotate both axes with the quaternion for $i$ ($i$ is now aligned with the x-axis)
    rotate $\neg i$ around $i$ until $\neg i$ lies parallel to the x-y plane
    if $\neg i$ crosses the x-z plane then
        find point the of intersection of $\neg i$ and x-z plane, call it $int_{\neg i} = (x_{\neg i}, 0, z_{\neg i})$
    else
        find point of $\neg i$ closest to x-z, either $b_{\neg i}$ or $e_{\neg i}$, call it $int_{\neg i} = (x_{\neg i}, y_{\neg i}, z_{\neg i})$
    end if
    if $x_{b_i} \leq x_{\neg i} \leq x_{e_i}$, where $b_i = (x_{b_i}, y_{b_i}, z_{b_i})$ then
        $t_i = (x_{\neg i}, 0, 0)$
    else
        if $x_{b_i} > x_{\neg i}$ then
            $t_i = b_i$
        else
            $t_i = e_i$
        end if
    end if
end for
```

we seek the point on axis j which is closest to any point on helix axis k. If we translate and rotate helix j so that it is aligned with the x-axis and the beginning point is at the origin. By applying the same transformations to helix axis k, their spatial configuration is conserved. Next, everything is rotated around the x-axis until axis k lies in a plane parallel with the x-y plane. This transformation yields the property that if axis k crosses the x-z plane (ie. a point on the axis where y=0), then this crossing point is the nearest point to the x-axis. Divide space into three regions, the divisors being planes parallel to the y-z plane. One will be placed at the beginning point of axis j at $x = 0$ (which is the y-z plane); we define this plane as x_b. The other plane is placed at the end point of axis j, $x_e = e_j$. This notation is continuing with that used by Chothia et al. [CLR81], where b_j and e_j are the endpoints of helix axis j.

Now we divide the problem into two cases: Either axis k crosses the x-z plane or it does not. If it crosses the plane then the solution is simple. Take the x-coordinate of the point where y=0 in axis k, and find this point on axis j. If it is between b_j and e_j, then it is the point of nearest approach. If not, simply take the endpoint of j that is nearest. If axis k does not cross the x-z plane, take the endpoint of k that is closest to the x-axis. Now using the x coordinate of this point, repeat the above steps.

Note that this procedure finds the point of nearest approach on axis j, but this does not necessarily find the point for axis k. To find the corresponding point on the other axis, this procedure must be repeated by aligning axis k with the x-axis and finding where axis j lies. This procedure is illustrated in Figure 4.14. The advantage of using this method is that it reduces the problem to a simple and intuitive geometric relation.

Angle Calculation - Hespenheide and Kuhn

Hespenheide and Kuhn [HK03] use a model that is similar to the proposed. They find the axis for each helix by determining the line of best fit through the alpha carbons of the helix (I am assuming by linear regression). The line of closest approach is then found, and Ω is defined as shown in Figure 4.15. All pairs of helices that do not pack face to face (where the global segment of closest approach does not lie within the length of both helices) are eliminated in their method, and only helices that are within 13Å of each other are used to ensure interaction.

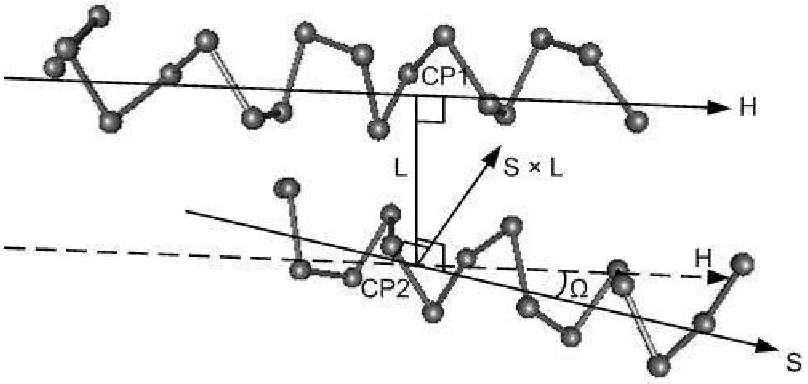


Figure 4.15: The closest approach method for calculating Ω. CP1 and CP2 are the points of closest approach for helix axes H and S respectively. This line of closest approach is designated as L. S×L defines a vector normal to both S and L, and then Ω is the angle between S and H when H is projected into the plane defined by S and S×L. The figure has been adapted from [HK03].

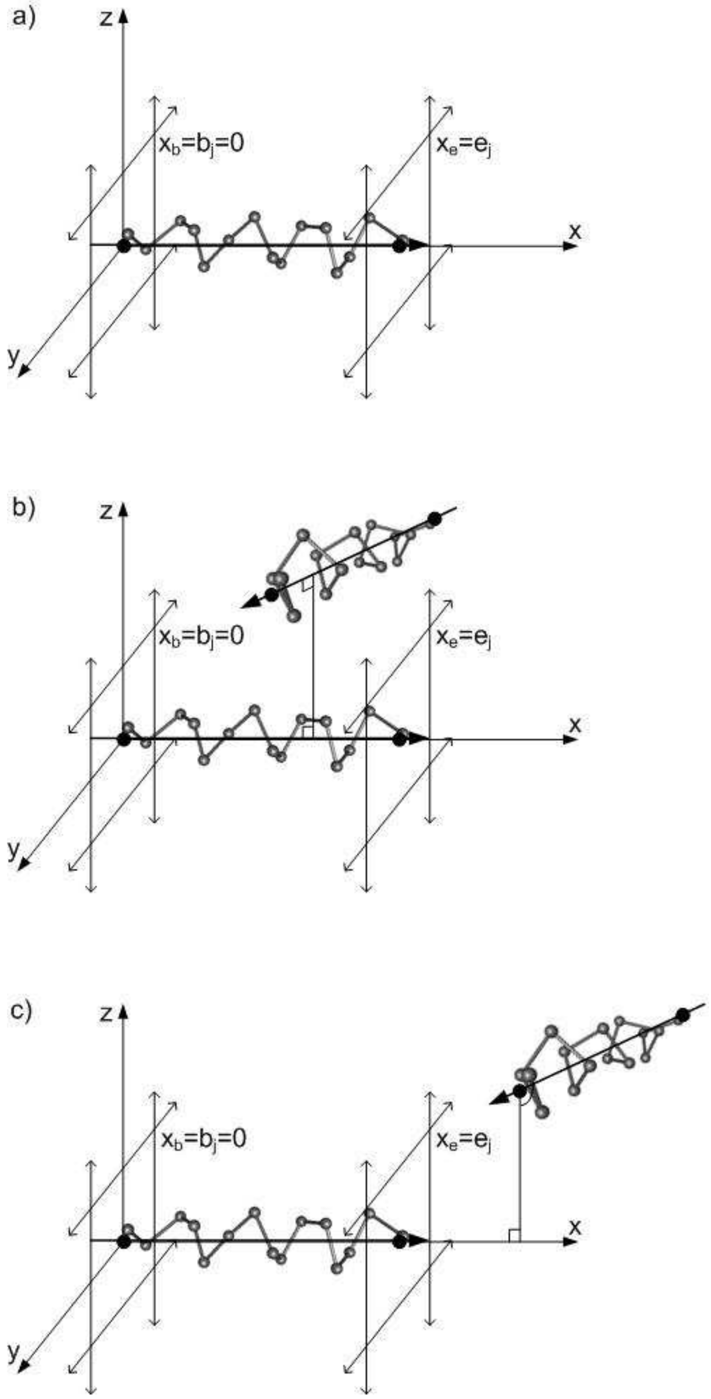


Figure 4.14: a) The axis of helix j is aligned with the x axis and the point b_j is positioned at the origin. Planes are defined parallel to the y-z plane at $x_b = 0$ and $x_e = e_j$, the end points of the helix. b) In this case axis k crosses the x-axis. Since the crossing point is between the planes ($x_b \leq x \leq x_e$), then we need to find x where $y = 0$. The point on axis j at x is the endpoint of the segment of closest approach. c) Axis k does not cross the x axis, so we find x at $min(|y|)$. In this case $x > x_e$, so the endpoint x_e is nearest on axis j.

4.3.3 Other Approaches

Barlow and Thornton [BT88] outline a method for calculating the helix axis which uses all of the backbone atoms rather than just the alpha carbons. The product is a series of points through the helix corresponding to each atom, through which a curve may be fit. This approach is similar in spirit to that of Walther et al. [WEA96], in that if one were to use it for calculating interhelical angles (which the authors do not, their aim is to classify the curvature of helices), the most likely candidate would be the tangent to the axis curve at the point of closest approach on each helix. Another approach was found, available in a software package called PROMOTIF, but the precise method used for calculating the interhelical angle was unspecified [HT96, SR02].

Bywater et al. [BTV01] use straight lines connecting the center of each end of a cylinder representing the helix. This approach is ideal for straight helices, and the authors claim that it is also sufficient for continuously curved helices as well. In helices where there is a bend or a kink, identified by the determination of atypical torsion angles, the axis is calculated for each part of the helix individually. The authors are examining the properties of transmembrane proteins, and they found that the global segment of closest approach (GCSA, see Figure 4.10 on page 71) is usually contained within the length of both helices [BTV01]. This conclusion would permit the direct application of the proposed method for calculating the interhelical angle. Unfortunately, the same can not be said for globular proteins. Lee and Chirikjian [LC04] surveyed 1290 proteins to obtain a database of 28,365 interacting helix pairs (within 15Å of each other). Of these, fewer than 20% represent pairs that pack.

4.3.4 The Proposed Method

To consider helices as idealized for the purposes of finding the interhelical angle is a gross oversimplification. Helices in nature are usually deformed to some degree [MWK74, BT88]; there may be severe bends, bulges, or other features. The first step towards determining the interhelical angle is therefore to determine an accurate manner of representing the axis of each helix. The method used to find the axis in this study begins by finding the residues on each helix in contact with those on the other. This defines the portion of the helix in the interface region, and the axis is determined for this section only. The interface is the packing portion of the helix, so it is the only part of the helix relevant to the interhelical angle. This mitigates effects on the axis from bends in the helix resulting from interactions with other protein structures. Depending on the size of the interface, it may be extended by several residues in each direction (i.e., if it is less than five residues long), provided that they are still within the alpha helix. This will ensure that enough of the helix is being used to determine the axis correctly. Simply finding the axis locally as a tangent to the point on the axis at the intersection with the segment of closest approach will be subject to the local properties of the nearest residues, although this effect is mitigated by smoothing operations. Nonetheless, the most logical method of describing the interhelical angle for a helix pair would be to describe the configuration over the interacting region between the pair. The helix axis is found by the rotational least squares method on the basis of the Christopher et al. study[CSB96].

If the objective is to find the angle only between helices that pack face to face[8],

[8] There are a few reasons why this might be the case. Dealing with pairs that do not pack face to face is more difficult, as the interhelical angle is highly susceptible to the choice of endpoints of each axis, as illustrated in Figure 4.16. Also, finding the interhelical angles of non-packing pairs is

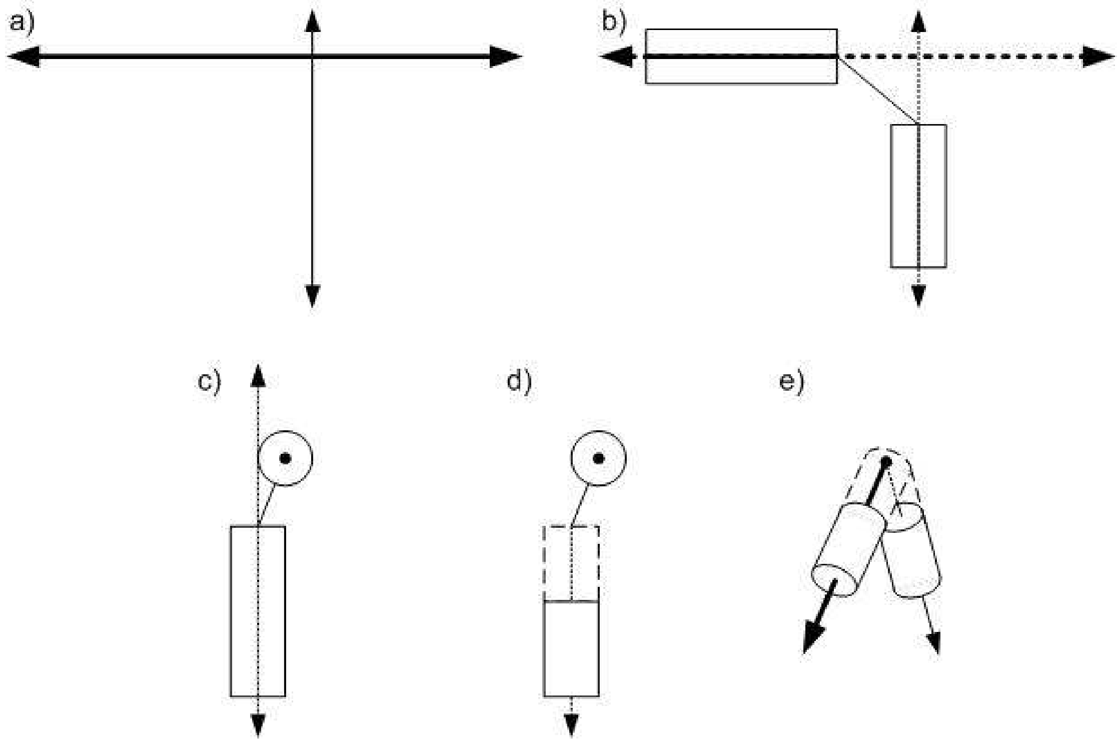


Figure 4.16: a) This is a pair of skew lines of infinite length, both are parallel to the plane of the page. The thicker line is closer to the viewer. Notice the crossing angle of the lines is 90°. b) Now we have two helices represented by the cylinders. The global segment of closest approach intersects neither helix, and the segment of closest approach joining the two end of the helices is shown as a line. c) We have rotated the figure about one helix axis, and now we are looking directly down the axis of what was the closer axis. d) The representation of the cylinders and axes is changed to make e) more clear. e) The dihedral angle is illustrated by rotating the figure so that we are looking down the segment of closest approach. Notice that the angle is nowhere near 90°.

the method is simple. After performing your first rotation, if the point where axis k crosses the x-axis is not between x_b and x_e (refer back to Figure 4.14 on page 92), then you do not have face to face packing, and the pair can be discarded. Otherwise, the interhelical angle can be found easily using a modified dot product method [FSG06]. Simply finding the dot product of the vectors representing each axis yields oversimplified angles. This concept is discussed further in Figure 4.17.

The key to understanding the difference in the angle calculations is that having selected a viewpoint perpendicular to the axis of a helix, the near and far surfaces of the helix will have 'ridges' of different orientations. One of the helices will be considered as being above the other in space when viewed by looking down a vector which is perpendicular to both axis vectors. There can be considered top and bottom surfaces to each helix using this model (Figure 4.17 (a)), and the ridges of one helix can be thought of as fitting into the grooves of the other helix (Figure 4.17 (b)). The concept of packing alpha helices using the ridges and grooves model may be dated and it may not be the true model for how packing occurs, but it suffices well for this illustration. It was the original rationale behind the concept of preferential interhelical angles [CLR81], while recent papers have supported the idea using more sophisticated means. In order to determine what the value of the angle is, the axis vectors must be found using the method explained earlier. There will be a region in space where these two lines, when positioned as passing through the center of masses of their respective helix regions, pass closest to each other. This line of closest approach is necessarily perpendicular to both axis vectors (the global segment of closest approach). The

of questionable value for the purposes of protein structure prediction, since it is packing pairs of helices which form the structure of the protein.

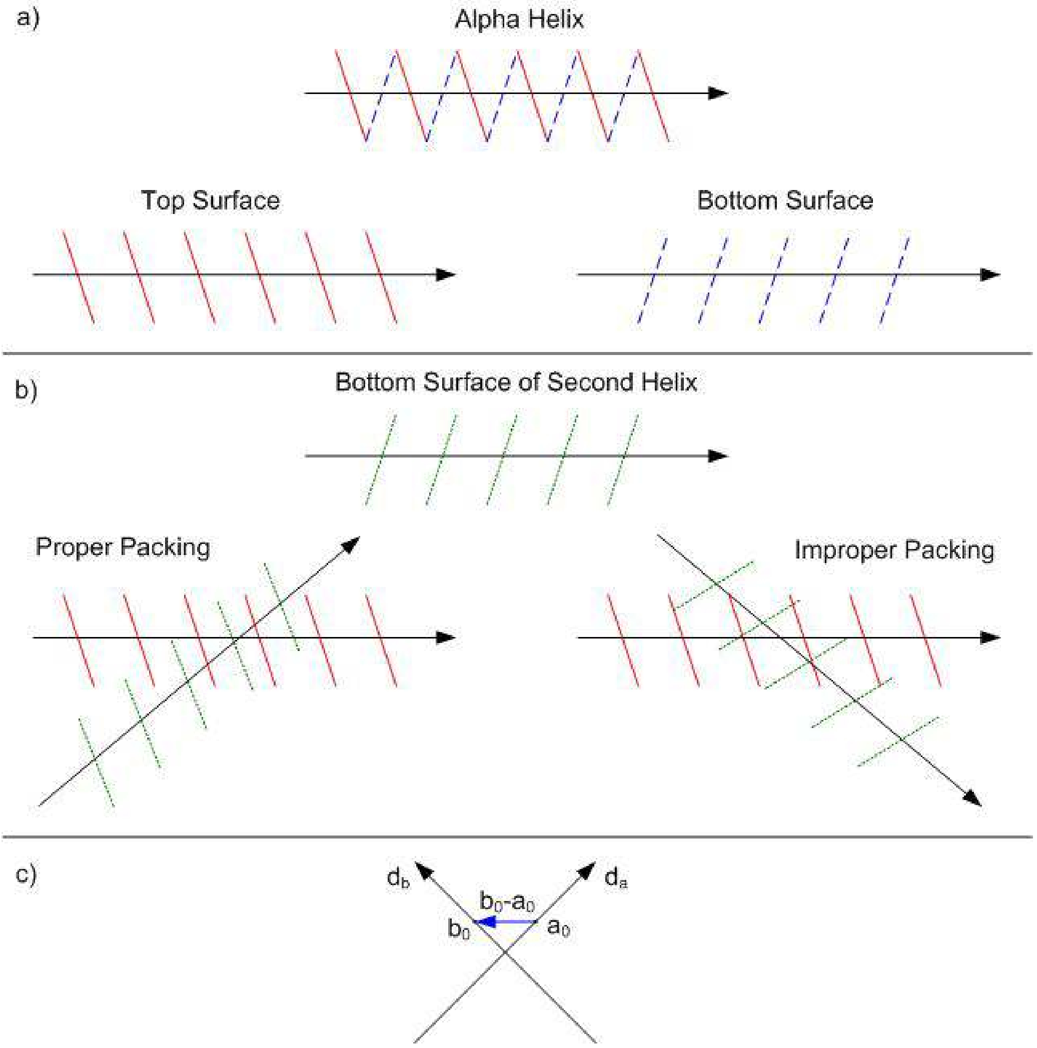


Figure 4.17: a) The top image represents a simplified alpha helix. The black arrow is the helix axis, and is lying in the plane of the page. The red lines are the sections of the helix backbone that may be considered the top surface; they are projecting out of the page. The blue lines are the bottom surface, and are beneath the surface of the page. These two surfaces are shown separately below the top image for clarification. b) The green lines represent the ridges on the bottom surface of a second alpha helix. In the two images below, this helix has been superimposed on the top surface from part a) at an angle of 45°. Notice that there is a significant difference in the resultant packing, therefore a sign must be used: the left image is negative, the right is positive [CLR81] (using the right-hand rule, find the angle from the upper helix to the lower one). c) Two vectors d_a and d_b, each representing one of the helix axes, are shown.

points of intersection can be determined, and will be labelled as a_0 on axis d_a and b_0 on axis d_b, as shown in Figure 4.17 (c). Now the vector formed by subtracting the points can be compared to the cross product of the vectors (they will be either parallel or anti-parallel) to determine the rotation of the top vector relative to the bottom, as outlined in Algorithm 3.

Algorithm 3 Modified dot product method for determining the interhelical angle

```
find axis $d_a$ and $d_b$ and closest points $a_0$ and $b_0$
$\Omega = \arccos\left(\frac{d_a \cdot d_b}{|d_a||d_b|}\right)$
if $(d_a \times d_b) \cdot (b_0 - a_0) < 0$ then
  $\Omega = -\Omega$ (the top vector is rotated clockwise)
end if
```

The dot product of the cross product of the vectors with the difference between the points of closest approach (a_0 on axis d_a and b_0 on axis d_b) can determine which of the vectors is above the other from the current perspective. Note that this approach will always yield the same result regardless of which is chosen as the top vector. Choosing the other vector as the top would change the signs of both the cross product and the vector representing the line of closest approach. Thus, the sign of the dot product will be consistent. This approach to calculating the angle (not the axis however) is similar to that outlined by Engel and DeGrado [ED05b]. Once it has been established which helix is on top, the angle can be found using the right-hand rule, finding the angle from the upper helix to the lower one. Although the reference point is chosen arbitrarily, it is irrelevant as the resultant angle will be the same no matter the perspective. Once the direction of rotation has been established, the sign of the angle can be assigned. If the rotation is clockwise, the angle is negative, as defined by Chothia et al. [CLR81]. This method produces a fully automated system for accurately determining the interhelical angles of any helix pair.

Many of the past approaches have used the dihedral angle formed by the axes and the segment of closest approach, as described earlier. If the pair is a packing pair, this method is not necessary and the dot product method can be applied. If the pair is not packing, then the dihedral angle calculation method should be used, as the difference between the dot product of the axes and the dihedral angle can be large, as illustrated in Figure 4.16.

To summarize, the proposed method for finding the interhelical angle and packing attribute for a pair of alpha helices is as follows:

1. Find the contact map for the pair and isolate the interface region.

2. Find the axis using the quaternion-based rotational least squares method for each helix over the atoms in this interface region. Using an axis local to the interface is the most logical approach, since the axis should describe the helix in the area where interaction between the helices is occurring.

3. Determine whether or not the pair is packing and find the points of closest approach using Algorithm 2. This is a simple method to implement, as the transformations required have already been computed in the previous step.

4. Calculate the interhelical angle using Algorithm 3 if the pair is packing. If the pair is not packing, calculate the interhelical angle using a dihedral method if the angle is desired.

Protein	PDB ID	Reference
Carboxypeptidase A	5CPA	Hartsuck & Lipscomb (1971) [HL71]
Flavodoxin	2FOX	Burnett et al. (1974) [BDK+74]
Haemoglobin, horse met	2MHB	Ladner et al. (1977) [LHP77]
Lysozyme, hen egg	6LYZ	Imoto et al. (1972) [IJN+72]
Lysozyme, bacteriophage T4	2LZM	Remington et al. (1978) [RAO+78]
Nuclease, staphylococcal	2SNS	Arnone et al. (1971) [ABC+71]
Subtilisin	1SBT	Wright et al. (1969) [WAK69]
Thermolysin	1TLX	Matthews et al. (1974) [MWK74]
Tobacco mosaic virus protein	2TMV	Bloomer et al. (1978) [BCB+78]
Triose phosphate isomerase	1TIM	Phillips et al. (1978) [PSTW78]

Table 4.1: The names of the proteins and the references as listed in Chothia et al. [CLR81]. The protein ID is the one found in the PDB that was considered to be the closest match to that in the original paper. The proteins were searched by name and authors, and the oldest matching entry was chosen. It is possible that some of the data used in this study is different from the data used in Chothia et al., as in most cases the PDB file was a revised version of the original data set.

4.4 Analysis of the Interhelical Angle Algorithm

The algorithm introduced in the previous section was validated by comparing the interhelical angles using that algorithm to those found observationally by Chothia et al. [CLR81]. In their study, ten proteins were chosen for analysis based upon the high resolution of their structures (at the time), and the angles were determined by manually manipulating models of the structures and finding the angles based upon direct observation. For this reason, the angles that they calculated are being considered as the standard for this study. The proteins used are shown in Table 4.1. The results of the algorithm are shown in Table 4.2.

There were several special cases that needed to be dealt with in the analysis. Some of the indices for the helices in some pairs used were not accurate. For several, such as 2FOX pair (1,4), the helices were quite distant and not in contact in any conventional

Helix Indices	Calculated Angle	Observed Angle	Difference
1sbt			
3,6	-38.9	-44	5.1
4,5	12.9	9	3.9
6,8	-21.2	-27	5.8
7,8	130.2	132*	1.8
2tmv			
2,3	-167.7	-152*	15.7
3,4	-160.8	-163*	2.2
6lyz			
1,2	126.9	132*	5.1
1,4	-58.2	-62	3.8
1tim			
1,2	-29	-24	5
2,3	150	147*	3
7,8	-47.1	-40	7.1
2lzm			
1,5	-46.2	-42	4.2
3,5	106.1	106	0.1
7,8	-163.9	-164*	0.1
8,10	168.6	163*	5.6
2sns			
1,2	141.1	146*	4.9
5cpa			
1,2	8.2	7	1.2
2,3	-90.5	-86	4.5
2,8	-59.9	-56	3.9
3,8	51	55	4
6,8	142.6	143*	0.4

Table 4.2: The angle calculation test results. The calculated angles are those found using Algorithm3. The observed angles are those presented by Chothia et al. [CLR81], and were determined manually through observation with modelling software. All fields are in degrees. The purpose of this table is simply to indicate that the algorithm performs well. There is inherent subjectivity in this measure, given from the version of the PDB file used, and particularly in the choice of axis.

sense. In this example, there was a pair (1,5) which were in contact and within 3.4° of the value given by Chothia et al. for (1,4). It would be speculative to assume that this would be the pair that was intended, as there were many instances of pairs found not investigated by Chothia et al., and five instances of pairs they found which were not pairs in contact in the PDB files used. All such instances were excluded from the study. For seven pairs, Chothia et al. did not provide an angle as a result of a configuration that they could not use observationally. The problem scenario is one where the point of nearest approach for the two vectors was not in the area where the helices were considered in contact for one of the helices (in other words, not a packing pair). In these cases, these helix pairs were skipped for the purposes of this study as they provide no reliable metric. Also, in the Chothia et al. study, the direction of the vectors were not taken into account, which resulted in angles that could be different from the actual by 180°. Therefore, it was considered acceptable to add or subtract 180° from the angle observed in their study. The result of this work is an average difference of 4.1° and a standard deviation of 3.2°. Further, if we consider the angle for helices 2 and 3 in 2tmv to be an outlier[9], the results improve to an average difference of 3.6° and a standard deviation of 1.9°. For comparison, Chothia et al. had an average difference of 16° and a standard deviation of 12.1° on the same alpha helix pairs used in this study when calculating the interhelical angle based on their packing model. The actual values are not shown here for the sake of brevity; they are listed in complete detail in their paper.

[9] It is possible that the data used in the present study was different than the original data, because different versions of PDB files were used. Also, it is possible that Chothia et al. were mistaken in their observations for this pair. Finally, the use of the helix axis local to the interface rather than over the entire helix accounts for some of the difference.

This study evaluated the effectiveness of a new algorithm for calculating interhelical angles for protein alpha helices using contact maps for a frame of reference. The algorithm is simple and provides results in a full 360° range that are within 4.1° on average of classically derived results (by manual observation of models). This algorithm should prove useful to anyone working in the field of protein structure prediction which involves alpha helix supersecondary structures. Future work will involve attempting to predict the interhelical angle based upon contact information and helix axes found using predicted coordinates, as this would be a key step in the overall goal of predicting tertiary protein structure from the primary sequence.

Chapter 5

Clustering Contact Maps to Predict Packing

> *Here is an ordinary square. But, suppose we extend the square beyond the two dimensions of our Universe along the hypothetical Z-axis, there. This forms a three-dimensional object known as a "cube", or a "Frinkahedron" in honor of its discoverer.*
>
> -Professor Frink, *The Simpsons*

The purpose of this chapter is to demonstrate that it is possible to predict properties of pairs of alpha helices given only the the contact map for the pair. To accomplish this task, we will use the binary attribute of packing, as discussed previously in Chapter 4. The hypothesis for this section of the thesis is that if two pairs of alpha helices have similar contact map interfaces, then they will pack in similar ways.

To test the hypothesis, we must first determine how we are to compare contact maps, as we need a distance metric for the comparisons. This problem is a field of study aside from the task at hand, so only a brief outline of the challenges will be given here. Once we have the comparison method, we can perform clustering and determine whether clusters of similar contact maps correspond to groups of helix pairs with similar packing attributes.

5.1 Distance Metrics

To begin the discussion, assume for the time being that we are comparing apples to apples: the contact maps that we are comparing are all of the same size and of similar orientations. If we have two contact maps C_1 and C_2, then similar values at $C_1(i,j)$ and $C_2(i,j)$ indicate a similarity between the maps. If the maps we are comparing are of size i by j, the maps may also be represented as strings or vectors of length $n = i \times j$. We have the choice of many distance metrics for comparing these strings. The naive choice would be to use Euclidean distance. The Euclidean distance is the 2 norm distance between the strings, given by:

$$d_E = \sqrt{\sum_{i=1}^{n} (C_1(i) - C_2(i))^2}$$

However, there are other metrics that are better suited to data of this type [ESK02]. Euclidean distance has the property that true values and false values for attributes carry the same weight, and this does not work well for contact maps which are generally sparse. Consider that if you had two maps, where each has only one contact, but the contacts are in different locations in each map. The distance between the maps would be $\sqrt{2}$. Now consider two maps where the whole map is filled with contacts save for one location each. You would expect there to be much higher similarity between the two maps that are full of contacts rather than the two nearly devoid maps, but their respective Euclidean distances are the same. Ertöz et al. [ESK02] assert that when dealing with sparse data sets in high dimensions, it is often the case that the presence of an attribute is more significant than its absence. Therefore, the metrics chosen to measure distance should reflect this property. The Jaccard distance (or Jaccard coefficient or Jaccard index) is given by [Jac08]:

$$d_J = \frac{C_{11}}{C_{11} + C_{10} + C_{01}}$$

where

C_{11} is the number of contacts shared by both contact maps
C_{10} is the number of contacts in the first map not found in the second
C_{01} is the number of contacts in the second map not found in the first

Another metric with similar properties to the Jaccard distance is the cosine distance, which is given by the dot product between the normalized vectors corresponding to each contact map:

$$d_{cos} = \frac{C_1 \cdot C_2}{||C_1||||C_2||}$$

Of course, raw contact maps from pairs of alpha helices are not all of the same size, and so we must manipulate the maps so that we can use the distance metrics outlined above (we must make each map an apple). The comparison of contact maps involves first aligning the contact maps, so that a contact in one map has a similar meaning to a contact in another map. The proper alignment of contact maps is an open research problem, known as the Contact Map Overlap (CMO) problem [CL02]. Different visual techniques may be used to measure the similarity of the maps, these are discussed in detail in [Kuo05] and alluded to in [DGK06]. There exists another approach using graph theory and Lagrangian relaxation [CL02]. The nature of the data being used in the present application permits a simplified approach, since we are not dealing with contact maps for entire proteins. The challenge is to properly align the contact maps; this problem yielded a solution which involves dividing the contact maps into sub-classes. The easiest maps to align will be those maps oriented in a corner of the map, as the corner[1] provides a natural alignment point. Therefore, the first class of contact map is any map with a contact in a corner of the map. For this application, the corners were treated as 2×2 areas in each corner. This makes sense, since this corresponds to the end of the helix. At the third residue, a turn around the axis is nearly complete, so it is not really at the end of the helix. However arbitrary, this threshold is effective, as will be demonstrated shortly.

The next class of contact maps is edge contacts, specifically any that have a contact within the outer two rows or columns around the perimeter of the contact map, but not in a corner. These maps can be aligned by the contacts present in the perimeter area, based upon the center of mass of the contacts in the perimeter. Finally, the third class is the central contacts, which would be those maps that have no contacts within the outer perimeter. These maps are fairly difficult to align. The maps could be aligned again by their center of masses, but with two dimensions now factoring into the calculation, there is more room for error. Alternatively, one of the tools discussed earlier for solving the contact map alignment problem could be used. Fortunately, the alignment of these maps is not an issue for the present problem, as will be shown. This classification process is performed greedily in the order presented

[1] For the purposes of this application, we will consider a contact map as being oriented in a corner if there are contacts present in that corner. There is not necessarily any physical meaning behind the classification.

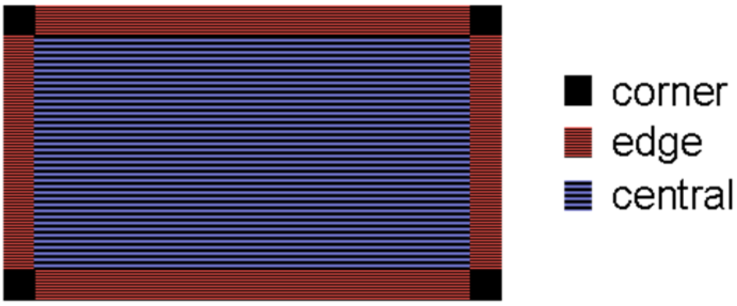


Figure 5.1: The different classes of contact map for comparison. Corner contacts have contacts in a 2×2 corner of the map. Edge contacts have contacts along the perimeter of the map, but not in a corner. Central contacts have no contacts in the outer two rows or columns of the map.

here (corner → edge → central), so that any map belonging to a corner class is placed there first. Instances where maps could potentially belong to both the corner and edge classes are thus all considered instances of the corner class. Once all corner and edge maps have been removed from the source list, the central maps are what remain. Figure 5.1 shows the different classes of contact map schematically.

5.2 Source Data

The data used for this study consists of 1078 pairs of alpha helices, collected from the PDB files of 171 proteins. These helix pairs were selected such that all helices were of length at least 6, and that the contact interface for the pair was at least 2×2. The central class of map contained 112 contact maps, and all maps corresponded to packing pairs of helices. Since all instances have a true value for packing, we can consider this to be one class. We do not have to align the contact maps to compare them, since all instances have the same packing value and no further analysis is necessary. Both of the edge and corner classes contain instances of both values for the packing however, so clustering may be necessary. The treatment of each class will be discussed in turn.

The corner maps, as mentioned earlier, are easy to align as they share a common origin. In addition, each contact map can be used as two data points since they can be reflected about a diagonal line through the origin, as shown in Figure 5.2. We can then transform each corner contact map such that all have a common origin through a series a flips. By these operations, we obtain a data set of 862 contact maps oriented in the corner. 794 of these maps correspond to packing pairs of helices, while 68 correspond to non-packing pairs. Now we have to modify the contact maps so they are all the same size. This has been accomplished in this study by simply taking the 15 × 15 square map[2] rooted in the corner containing the contacts. If the map is smaller than this square in either dimension, the map is simply padded with zeroes. Now we have a collection of aligned maps of the same size that may be compared with a distance metric.

The edge contact maps can be treated similarly, except that there is the challenge mentioned earlier of having no clear point of origin to use for the alignment step. Once we have the alignment method, we can do flips of maps so that all edge maps can compared. However, as with the central maps the task is not useful for this application. Using the same set of helix pairs as previously, we obtain 535 contact maps. Of these maps, only 4 do not correspond to packing helix pairs, so it is fair

[2] This size has been chosen arbitrarily. Other sizes may be used, the effect on the performance of the clustering is quite minimal, as will be demonstrated later in this chapter.

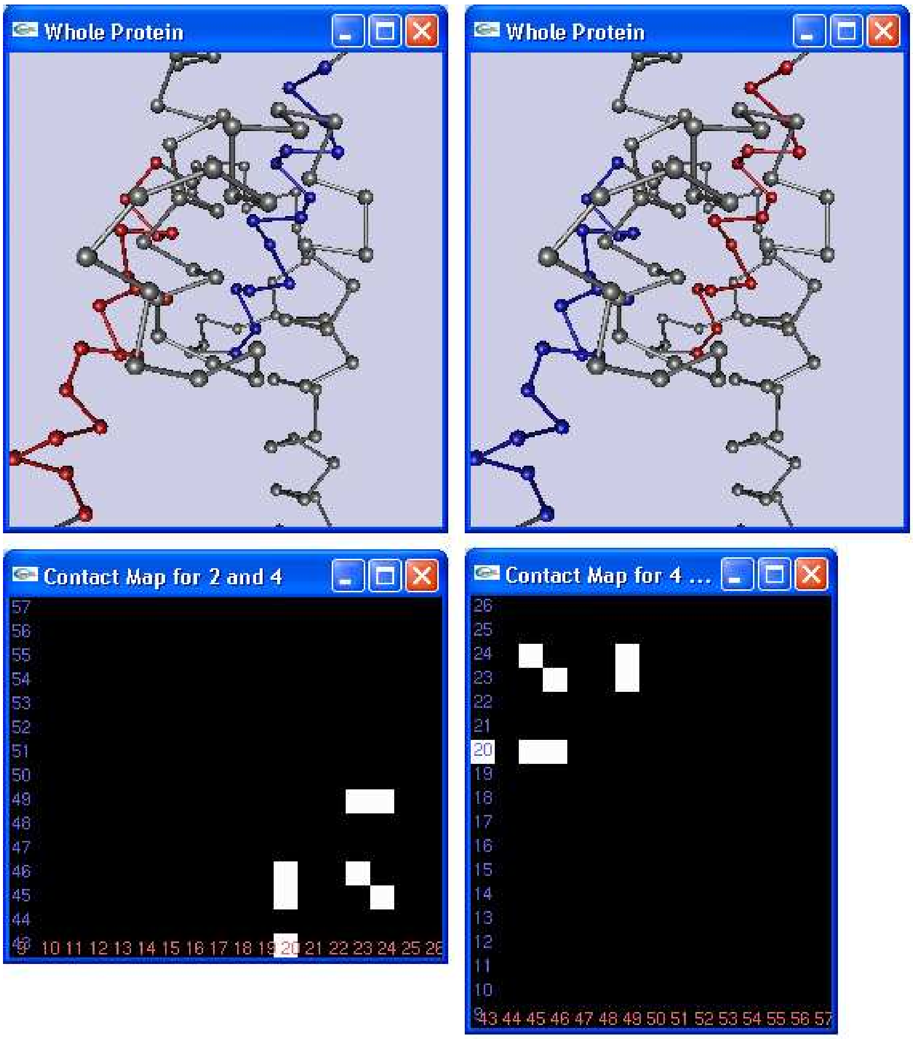


Figure 5.2: A contact map can be reflected to get two data points from one. This corresponds to just switching which helix is represented by each axis. In the figure, helices 2 and 4 from protein 1A0A are used again. The red helix in the top windows corresponds to the horizontal axis. In the left figure, helix 2 is on the horizontal axis versus helix 4 on the vertical. Switching the helices causes a reflection about the origin (lower left in this case). Both maps obviously correspond to the same three-dimensional configuration.

Contact Map Class	**Total Instances**	**Packing**	**Non-packing**
Central	112	112 (100%)	0 (0%)
Edge	535	531 (99.3%)	4 (0.7%)
Corner	431	397 (92.1%)	34 (7.9%)
Doubled Corner	862	794 (92.1%)	68 (7.9%)
All Maps	1078	1040 (96.5%)	38 (3.5%)

Table 5.1: The characteristics of the contact maps used for clustering are summarized here. For each class of contact map, the number of instances in the source data set is given, as well as the number of corresponding pairs of alpha helices that exhibit packing or otherwise.

to decide that these maps are outliers[3]. Thus, contact maps with edge contacts generally correspond to packing pairs of alpha helices. The nature of the test data is summarized in Table 5.1.

5.3 Clustering Analysis

Now we may proceed to cluster the corner contact maps, and determine whether clusters of maps share the same value for the packing attribute. Throughout this discussion, a contact map will often be referred to as a point, since that is what the clustering algorithm is working with. A 15×15 contact map which has been converted to a vector of length 225 is now being treated as a point in 225 dimensional space. The algorithm to be used to perform the clustering is a shared nearest neighbour approach, as outlined by Ertöz et al. [ESK02].

We can review the implementation of the algorithm step by step.

[3] For fun, clustering was performed on the data to see if all four instances corresponding to non-packing would end up in one cluster, but this was not the case. In fact, the clustering algorithm considered three of the instances to be noisy data.

Algorithm 4 Shared Nearest Neighbour Clustering Algorithm

```
1 - Construct the similarity matrix
2 - Sparsify the similarity matrix using k-nn sparsification
3 - Construct the shared nearest neighbour graph from the k-nn sparsified similarity matrix
4 - For every point in the graph, calculate the total strength of links coming out of the point
5 - Identify representative points by choosing the points that have high total link strength
6 - Identify noise points by choosing the points that have low total link strength and remove them
7 - Remove all links that have weight smaller than a threshold
8 - Take connected components of points to form clusters, where every point in a cluster is either a representative point or is connected to a representative point
```

1 - Construction of the similarity matrix

The similarity matrix was constructed using both the Jaccard coefficient and the cosine distance in turn. Both yielded similar results in the final clustering, so the cosine distance was arbitrarily selected as the measure to be used for the remainder of the study. Given n contact maps to be clustered, the similarity matrix S is size $n \times n$, where entry $S(i, j)$ is the cosine distance from contact map i to j.

2 - Sparsification of the similarity matrix

The similarity matrix is sparsified using k nearest neighbours (k-nn) sparsification. k nearest neighbours is a simple concept: it is simply that for some value k, find the k points closest to each point using the chosen distance metric. This is accomplished by first finding the k smallest distances in each row of the similarity matrix; the provides the k-nn for each point. To sparsify, we next check the members of the lists for each point to ensure mutuality. Suppose that point j is a member of the k-nn list for point i. Now we check the k-nn list for j to see if i is a member. If it is not, we remove j

from the k-nn list for point i. Once this procedure is complete, there will be complete mutuality in the k-nn lists for each contact map. The selection of an appropriate value for k is a trial and error process; the effectiveness depends on the size of the data set, the nature of the data being clustered, and the number of desired clusters in the outcome. For the corner contact maps, values between 4 and 15 were tested, and 12 produced the best results.

3 - Construction of the shared nearest neighbour graph

We now will use a weighting scheme introduced by Jarvis and Patrick [JP73] to determine how well connected points are. This is done by finding the number of nearest neighbours two points share and how well connected they are. The strength of a connection between two points i and j is given by:

$$str(i,j) = \sum (k+1-m) \times (k+1-n)$$

where

$i_m = j_n$, ie. some third point p in both lists
m is the position of p in the nearest neighbour list of i

n is the position of p in the nearest neighbour list of j

This sum is over every instance of a shared nearest neighbour in the lists for the points i and j. The shared nearest neighbour graph is an $n \times n$ matrix Snn, where entry $Snn(i,j)$ is defined by $str(i,j)$.

4 through 8 - Formation of the clusters

The next step is to find the connectivity of each point, $con(i)$, which is found by taking the sum of the strength of all of its connections:

$$con(i) = \sum_{j=1}^{n} Snn(i, j)$$

The connectivity of a point is used to determine how the clusters form. Points that have a connectivity higher than some threshold are chosen as representative points, and are used to nucleate clusters. To facilitate the selection of an appropriate threshold, the connectivity values were normalized to values between 0 and 1, and 0.6 was found to produce good clustering in this study.

In the original Ertöz et al. [ESK02] study, they removed all points with connectivity values below a given threshold and removed all links from the connectivity graph with weights below another threshold to eliminate noise. For the purposes of this study, the effects of these noise removal steps were minimized as it was desirable to consider all data as good data. As a result, 0.001 was used as the threshold for both steps.

Finally, the clusters are formed by taking the representative points and all of the points that they are connected to as clusters.

5.4 Clustering Results

The best results, as mentioned previously, were obtained by using a value of 12 for k when using the corner contact maps data set. A summary of the clustering results are presented in Table 5.2. Afterwards, the effects of varying the parameters is shown

Cluster Number	Packing	Non-Packing
Packing		
1	21	0
2	21	0
3	18	0
4	26	0
5	16	0
6	20	0
7	37	0
8	16	0
9	29	0
10	8	0
11	18	0
Mixed		
12	8	4
13	30	2
14	83	4
15	13	4
16	20	4
Non-Packing		
17	0	14
18	0	14

Table 5.2: The clustering results are listed. There were 11 clusters of packing pairs, 2 of non-packing pairs, and 5 were mixed.

in Table 5.3.

Since the results of the clustering analysis showed promise, it could be concluded that the packing configuration of a pair of alpha helices could be predicted from the contact map. However, it is worth asking how effective it would be to simply exhaustively search the database of contact maps to find the contact map closest to a target map. To test this, the data set of corner contact maps used in the clustering analysis was analyzed. For every data point, the closest neighbour by cosine distance was found, and the packing values for the neighbours was compared. Of the 862 contact maps in the data set, only 9 had a packing value different from the nearest

Parameters	**Packing**	**Non-Packing**	**Mixed**	**Total**
Standard	11	2	5	18
$k = 6$	51	2	9	62
$k = 9$	21	2	6	29
$k = 15$	7	0	6	13
$rep.thresh. = 0.8$	4	2	1	7

Table 5.3: The effects of varying the clustering parameters on the clustering results is summarized. The 'normal' parameters are those that provided the best results, specifically $k = 12$, the representative threshold is 0.6, the noise threshold was 0.001, and the contact maps were sized at 15×15. The clustering appears good when the representative threshold of 0.8 is used, but the total population of the clusters is only 88.

neighbour. That is just shy of 99% accuracy, and should be considered reliable for performing predictions. Thus, for prediction of the packing attribute for a pair of alpha helices corresponding to a contact map, an exhaustive search for the packing value of the nearest neighbour to the contact map in the database is recommended.

Chapter 6

Summary and Conclusions

6.1 Conclusions

It is a far, far better thing that I do, than I have ever done; it is a far, far better rest that I go to than I have ever known.

- Sydney Carton, *A Tale of Two Cities*

Hippy is a straightforward and user-friendly program that has the potential to facilitate breakthroughs in the study of alpha helix pairs. It assists in the investigation of the configuration and packing of a pair and the manipulation of the contact map, which no other known software provides. The program is open source and platform independent; it is possible for anyone to modify the package to include additional features that an individual may desire. Researchers working in protein structure prediction using contact maps will benefit from this tool by visualizing the correlation between the three-dimensional structure of a helix pair and the corresponding contact map. Hippy was instrumental in the other sections of this thesis.

The interhelical angle of a pair of alpha helices was studied extensively in this thesis. There are many methods available for calculating the axis of an alpha helix and the angle formed between two axes, but all of the methods reviewed were found wanting. Thus, a new method was introduced. The axis for each helix is found using a rotational least squares method. Once the axes are determined, the angle can be found easily by aligning one helix axis with an axis of the reference coordinate system. This rotation also allows for a check of packing by checking for the overlap of coordinates. These methods were further refined with the use of the contact map for the pair of alpha helices. This allows the determination of the interface region for the pair. Since helices are generally curved to some degree, it is best to determine the interhelical angle based solely on the sections of the helices where they are packing together. Given this, the algorithm presented in this thesis is the most accurate algorithm known for determining interhelical angles.

It was demonstrated that it is possible to predict properties of the configuration of a pair of alpha helices from the contact map. First, it was shown that contact maps cluster well using a k-nearest neighbours algorithm tailored to sparse data. The clustering algorithm necessitated a novel contact map classification scheme, where the location of the contacts in the map is used to distinguish between central, edge, and corner classes of contact map. Finally, the packing attribute of a pair of alpha helices was found to match that of its nearest neighbour 99% of the time. These results provide definitive support for the premise that other attributes of the configuration of proteins may be predicted from contact maps.

6.2 Future Work

> *We can only see a short distance ahead, but we can see plenty there that needs to be done.*
>
> - Alan Turing

The Hippy package has many potential extensions, some of which are in development at present. These include:

- The next version of Hippy will include an option where the full range of possible motions for the helix pair would be demonstrated. The contact map values and the steric surface of the helices will be used as the constraints.

- There is no ability for the user to move the helices or atoms to witness the effects at present. Since the contents of the contact map are linked dynamically to the coordinate data in Hippy, if either the coordinate data or the contents of the contact map window are changed, the contents of the other window will be adjusted accordingly. In addition, the substitution of different species of residues would be interesting.

- A modelling system which was based upon the energy of the system might prove useful, as the protein native state is commonly thought to be a low energy configuration (Anfinsen's thermodynamic hypothesis [Anf73]). This may still be added in some future version.

- Hippy could be extended to include triplets or quadruplets of helices, and it could also be used for beta sheets or other structures.

- One helix pair could be aligned with another helix pair of interest using conventional alignment techniques. This would be illustrative of differences between pairs of helices with respect to their contact maps.

- The secondary structure from Database of Secondary Structures in Proteins (DSSP) [KS83] could be used instead of that directly from the PDB file.

The discovery of the relationship between the packing attribute and the contact map will be useful for researchers working on protein structure prediction. In particular, this will be used as an advisor in the Tryptych case-based reasoning system discussed in section 2.3.3 to evaluate whether the predicted packing configuration of pairs of alpha helices correspond well to what would be expected from the contact map.

Finally, it would be useful to explore other properties of alpha helices that could be predicted from contact maps and used as advisors in Tryptych. The next property that will be explored is the interhelical angle, but others such as the interhelical distance should be straightforward to determine as well.

The protein structure prediction problem is still a long way from being solved, but discoveries such as those presented in this thesis are shedding light on the complex nature of the problem piece by piece. They have brought us one small step closer to the solution.

Bibliography

[ABC$^+$71] A. Arnone, C.J. Bier, F.A. Cotton, V.W. Day, E.E. Hazen Jr., D.C. Richardson, J.S. Richardson, and A. Yonath. A high resolution structure of an inhibitor complex of the extracellular nuclease of Staphylococcus aureus. I. Experimental procedures and chain tracing. *Journal of Biological Chemistry*, 246:2302–2316, 1971.

[ACHC97] G. Ausiello, G. Cesareni, and M. Helmer-Citterich. ESCHER: A new docking procedure applied to the reconstruction of protein tertiary structure. *PROTEINS: Structure, Function, and Genetics*, 28:556–567, 1997.

[ACL96] N.L. Allinger, K. Chen, and J.-J. Lii. An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry*, 17:642–668, 1996.

[ACSR97] R. Aurora, T.P. Creamer, R. Srinivasan, and G.D. Rose. Local interactions in protein folding: Lessons from the α helix. *Journal of Biological Chemistry*, 272(3):1413–1416, 1997.

[ALJXH01] B. Al-Lazikani, J. Jung, Z. Xiang, and B. Honig. Protein structure prediction. *Current opinion in chemical biology*, 5:51–56, 2001.

[AMS$^+$97] S.F. Altschul, T.L Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[Anf73] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[Åqv86] J. Åqvist. A simple way to calculate the axis of an α-helix. *Computers & Chemistry*, 10(2):97–99, 1986.

[AYL89] N.L. Allinger, Y.H. Yuh, and J.-J. Lii. Molecular mechanics. The MM3 force field for hydrocarbons I. *Journal of the American Chemical Society*, 111:8551–9556, 1989.

[BB01] R. Bonneau and D. Baker. Ab initio protein structure prediction: Progress and prospects. *Annual Review of Biophysics and Biomolecular Structure*, 30:173–189, 2001.

[BCB$^+$78] A.C. Bloomer, J.N. Champness, G. Bricogne, R. Stader, and A. Klug. Protein disk of tobacco mosaic-virus at 2.8-Å resolution showing interactions within and between subunits. *Nature*, 276:362–368, 1978.

[BCG92] S.A. Benner, M.A. Cohen, and D. Gerloff. Correct structure prediction? *Nature*, 359:781, 1992.

[BD05] E.N. Baker and G.G. Dodson. Proteins - discovery and detail. *Current Opinion in Structural Biology*, 15:652–654, 2005.

[BDK$^+$74] R.M. Burnett, G.D. Darling, D.S. Kendal, M.E. Lesquesne, S.G. Mayhew, W.W. Smith, and M.L. Ludwig. The structure of the oxidized form of clostridial flavodoxin at 1.9-Å resolution. Description of the flavin mononucleotide binding site. *Journal of Biological Chemistry*, 249:4383–4392, 1974.

[BHN03] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, 10:980, 2003.

[BKV00] M. Bansal, S. Kumar, and R. Velavan. HELANAL: A program to characterize helix geometry in proteins. *Journal of Biomolecular Structure & Dynamics*, 17(5):811–819, 2000.

[BKW$^+$77] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: A computer-based archival file for macromolecular structures. *European Journal Of Biochemistry*, 80(2):319–324, 1977.

[BMMT02] J.R. Banavar, A. Maritan, C. Micheletti, and A. Trovato. Geometry and physics of proteins. *Proteins: Structure, Function, and Genetics*, 47:315–322, 2002.

[Bow97a] J.U. Bowie. Helix packing angle preferences. *Nature Structural Biology*, 4:915–917, 1997.

[Bow97b] J.U. Bowie. Helix packing in membrane proteins. *Journal of Molecular Biology*, 272:780–798, 1997.

[Bow05] J.U. Bowie. Solving the membrane protein folding problem. *Nature*, 438:581–589, 2005.

[BR93] G.J. Barton and R.B. Russell. Protein structure prediction. *Nature*, 361:505–506, 1993.

[Bry03] B. Bryson. **A Short History of Nearly Everything.* Doubleday Canada, 2003.

[BSST87] T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg, and J.M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347–352, 1987.

[BT88] D.J. Barlow and J.M. Thornton. Helix geometry in proteins. *Journal of Molecular Biology*, 201:601–619, 1988.

[BTV01] R.P. Bywater, D. Thomas, and G. Vriend. A sequence and structural study of transmembrane helices. *Journal of Computer-Aided Molecular Design*, 15:533–552, 2001.

[Bur00] S.K. Burley. An overview of structural genomics. *Nature Structural Biology*, 7:932–934, 2000.

[BWF$^+$00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[CCD$^+$96] J. Callaway, M. Cummings, B. Deroski, P. Esposito, A. Forman, P. Langdon, M. Libeson, J. McCarthy, J. Sikora, D. Xue, E. Abola, F. Bernstein, N. Manning, R. Shea, D. Stampf, and J. Sussman. Protein Data Bank contents guide: Atomic coordinate entry format description, `http://www.rcsb.org/pdb/file_formats/pdb/pdbguide2.2/guide2.2_frame.html`. 1996.

[CF90] C. Chothia and A.V. Finkelstein. The classification and origin of protein folding patterns. *The Annual Review of Biochemistry*, 59:1007–1039, 1990.

[CHB$^+$97] C. Chothia, T. Hubard, S. Brenner, H. Barns, and A. Murzin. Protein folds in the all-β and all-α classes. *Annual Review of Biophysics and Biomolecular Structure*, 26:597–627, 1997.

[Cho89] C. Chothia. Polyhedra for helical proteins. *Nature*, 337:204–205, 1989.

[CL02] A. Caprara and G. Lancia. Structural alignment of large-size proteins via Lagrangian relaxation. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 100–108, New York, NY, USA, 2002. ACM Press.

[CLR77] C. Chothia, M. Levitt, and D. Richardson. Structure of proteins: Packing of α-helices and pleated sheets. *Proceedings of the National Academy of Science USA, Chemistry*, 74(10):4130–4134, 1977.

[CLR81] C. Chothia, M. Levitt, and D. Richardson. Helix to helix packing in proteins. *Journal of Molecular Biology*, 145:215–250, 1981.

[CNS83] K.C. Chou, G. Némethy, and H.A. Scheraga. Energetic approach to the packing of α-helices. 1. Equivalent helices. *Journal of Physical Chemistry*, 87:2869–2881, 1983.

[Cri53] F. Crick. The packing of α-helices: simple coiled coils. *Acta Crystallographica*, 6:689–697, 1953.

[CRR79] F.E. Cohen, T.J. Richmond, and F.M. Richards. Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structures with Myoglobin as an example. *Journal of Molecular Biology*, 132:275–288, 1979.

[CSB96] J.A. Christopher, R. Swanson, and T.O. Baldwin. Algorithms for finding the axis of a helix: Fast rotational and parametric least-squares methods. *Computers & Chemistry*, 20(3):339–345, 1996.

[DGK06] J. Davies, J. Glasgow, and T. Kuo. Visio-spatial case-based reasoning: A case study in prediction of protein structure. *Computational Intelligence*, in press, 2006.

[DMW03] J.A.R. Dalton, I. Michalopoulos, and D.R. Westhead. Calculation of helix packing angles in protein structures. *Bioinformatics*, 19(10):1298–1299, 2003.

[ED05a] D.E. Engel and W.F. DeGrado. $\alpha - \alpha$ linking motifs and interhelical orientations. *PROTEINS: Structure, Function, and Bioinformatics*, 61(2):325–337, 2005.

[ED05b] D.E. Engel and W.F. DeGrado. $\alpha - \alpha$ linking motifs and interhelical orientations - Supplemental Materials. *PROTEINS: Structure, Function, and Bioinformatics*, 61s(2):38–57, 2005.

[Efi79] A.V. Efimov. Packing of α-helices in globular proteins. Layer-structure of globin hydrophobic cores. *Journal of Molecular Biology*, 134:23–40, 1979.

[Efi93] A.V. Efimov. Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, 60(3):201–239, 1993.

[Efi99] A.V. Efimov. Complementary packing of α-helices in proteins. *Federation of European Biochemical Societies Letters*, 463:3–6, 1999.

[Eis03] D. Eisenberg. The discovery of the α-helix and β sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences of the USA*, 100(20):11207–11210, 2003.

[EJT04] I. Eidhammer, I. Jonassen, and W.R. Taylor. *Protein Bioninformatics: An Algorithmic Approach to Sequence and Structure Analysis.* John Wiley and Sons, Ltd., West Sussex, UK, 2004.

[ESK02] L. Ertöz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining*, 2002.

[FDS00] A. Fiser, R.K.G. Do, and A. Sali. Modeling of loops in protein structures. *Protein Science*, 9:1753–1773, 2000.

[FG06] R. Fraser and J. Glasgow. Introducing Hippy: A visualization tool for understanding the α-helix pair interface. In *Proceeding of the 2006 International Conference on Bioinformatics and Computational Biology (BIOCOMP)*, 2006.

[FGS98] J.S. Fetrow, A. Godzik, and J. Skolnick. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *Journal of Molecular Biology*, 282(4):703–711, 1998.

[For02] M.J. Forster. Molecular modelling in structural biology. *Micron*, 33:365–384, 2002.

[FOVC01] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835–843, 2001.

[FSG06] R. Fraser, J. Stewart, and J. Glasgow. Automated determination of interhelical angles for protein alpha helices from coordinate data. In *First Canadian Student Conference on Biomedical Computing (CSCBC 2006)*, 2006.

[GGGR05] K. Ginalski, N.V. Grishin, A. Godzik, and L. Rychlewski. Practical lessons from protein structure prediction. *Nucleic Acids Research*, 33(6):1874–1891, 2005.

[GGML99] A.J.F. Griffiths, W.M. Gelbart, J.H. Miller, and R.C. Lewontin. *Modern Genetic Analysis*. W.H. Freeman and Company, New York, 1999.

[GKD06] J. Glasgow, T. Kuo, and J. Davies. Protein structure from contact maps: A case-based reasoning approach. *Information Systems Frontiers*, 8:29–36, 2006.

[GM87] P.R. Gerber and K. Müller. Superimposing several sets of atomic coordinates. *Acta Crystallographica*, A43:426–428, 1987.

[GP97] N. Guex and M.C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723, 1997.

[GP98] N. Guex and M.C. Peitsch. Tutorial: Comparative protein modelling. In *The Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 98)*, 1998.

[HHC95] H. Hart, D.J. Hart, and L.E. Crane. *Organic Chemistry: A Short Course*. Houghton Mifflin Company, Toronto, 9th edition, 1995.

[HK03] B.M. Hespenheide and L.A. Kuhn. Discovery of a significant, nontopological preference for antiparallel alignment of helices with parallel regions in sheets. *Protein Science*, 12:1119–1125, 2003.

[HL71] J.A. Hartsuck and W.N. Lipscomb. Carboxypeptidase A. *Enzymes*, 3:1–56, 1971.

[HSS$^+$02] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M. Zaki. Mining protein contact maps. In *2nd BIOKDD Workshop on Data Mining in Bioinformatics*, 2002.

[HT96] E.G. Hutchinson and J.M. Thornton. PROMOTIF–A program to identify and analyze structural motifs in proteins. *Protein Science*, 5:212–220, 1996.

[IJN$^+$72] T. Imoto, L.N. Johnson, A.C.T. North, D.C. Phillips, and J.A. Rupley. Vertebrate lysozymes. *Enzymes*, 7:665–868, 1972.

[Jac08] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270, 1908.

[JG04] I. Jurisica and J. Glasgow. Applications of case-based reasoning in molecular biology. *AI Magazine*, 25(1):85–95, 2004.

[Jon99] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.

[JP73] R.A. Jarvis and E.A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, C22:1025–1034, 1973.

[JTV03] S. Jiang, A. Tovchigrechko, and I.A. Vakser. The role of geometric complementarity in secondary structure packing: A systematic docking study. *Protein Science*, 12:1646–1651, 2003.

[JV00] S. Jiang and I.A. Vakser. Side chains in transmembrane helices are shorter at helix-helix interfaces. *Proteins: Structure, Function, and Genetics*, 40(3):429–435, 2000.

[JV04] S. Jiang and I.A. Vakser. Shorter side chains optimize helix-helix packing. *Protein Science*, 13:1426–1429, 2004.

[Kah89] P.C. Kahn. Defining the axis of a helix. *Computers & Chemistry*, 13(3):185–189, 1989.

[KB98] S. Kumar and M. Bansal. Geometrical and sequence characteristics of α helices in globular proteins. *Biophysical Journal*, 75:1935–1944, 1998.

[KD04] P. Källblad and P.M. Dean. Backbone-backbone geometry of tertiary contacts between α-helices. *PROTEINS: Structure, Function, and Bioinformatics*, 56:693–703, 2004.

[Kea90] S.K. Kearsley. An algorithm for the simultaneous superposition of a structural series. *Journal of Computational Chemistry*, 11(10):1187–1192, 1990.

[Kol65] W.L. Koltun. Precision space-filling atomic models. *Biopolymers*, 3(6):665–679, 1965.

[KS83] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[Kuo05] T. Kuo. A computational approach to contact map similarity. Master's thesis, Queen's University, 2005.

[KVFM05] A. Kryshtafovych, C. Venclovas, K. Fidelis, and J. Moult. Progress over the first decade of CASP experiments. *PROTEINS: Structure, Function, and Bioinformatics Supplement*, 7:225–236, 2005.

[LC76] M. Levitt and C. Chothia. Structural patterns in globular proteins. *Nature*, 261:552–558, 1976.

[LC04] S. Lee and G.S. Chirikjian. Interhelical angle and distance preferences in globular proteins. *Biophysical Journal*, 86:1105–1117, 2004.

[Lea01] A.R. Leach. *Molecular Modelling: Principles and Applications.* Prentice Hall, Essex, England, 2nd edition, 2001.

[LHP77] R.C. Ladner, E.J. Heidner, and M.F. Perutz. The structure of horse methaemoglobin at 2.0 Å resolution. *Journal of Molecular Biology*, 114:385–414, 1977.

[Lig74] A. Light. *Proteins: Structure and Function.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.

[LLB$^+$01] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, and 244 others. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[LSD05] J.A Lopera, J.N. Sturgis, and J.-P. Duneau. Ptuba: a tool for the visualization of helix surfaces in proteins. *Journal of Molecular Graphics and Modelling*, 23:305–315, 2005.

[LWRR00] S.C. Lovell, J.M. Word, J.S. Richardson, and D.C. Richardson. The penultimate rotamer library. *PROTEINS: Structure, Function, and Genetics*, 40:389–408, 2000.

[Mac84] A.L. Mackay. Quaternion transformation of molecular orientation. *Acta Crystallographica*, A40:165–166, 1984.

[Mar02] E. Martz. Protein Explorer: Easy yet powerful macromolecular visualization. *Trends in Biochemical Sciences*, 27:107–109, 2002.

[MBCS00] J. Mendes, A.M. Baptista, M.A. Carrondo, and C.M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Structure, Function, and Genetics*, 37(4):530–543, 2000.

[MC92] V.N. Maiorov and G.M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *Journal of Molecular Biology*, 227:876–888, 1992.

[MC94] V.N. Maiorov and G.M. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235:625–634, 1994.

[McL79] A.D. McLachlan. Gene duplications in the structural evolution of chymotrypsin. *Journal of Molecular Biology*, 128:49–79, 1979.

[MF88] A.G. Murzin and A.V. Finkelstein. General architecture of the α-helical globule. *Journal of Molecular Biology*, 204:749–769, 1988.

[MJ96] S. Miyazawa and R.L. Jernigan. Residueresidue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.

[Moo65] G.E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, 1965.

[MS00] E. Martz and R. Sayle. Bonds in rasmol/chime, `http://www.umass.edu/microbio/rasmol/rasbonds.htm`. 2000.

[MWK74] B.W. Matthews, L.H. Weaver, and W.R. Kester. The conformation of thermolysin. *Journal of Biological Chemistry*, 249:8030–8044, 1974.

[Nag89] K. Nagano. Prediction of packing of secondary structure. In G.D. Fasman, editor, *Prediction of Protein Structure and the Principles of Protein Conformation*, pages 467–548. Plenum Press, NY, 1989.

[NCP$^+$03] M. Nanias, M. Chinchio, J. Pillardy, D.R. Ripoll, and H.A. Scheraga. Packing helices in proteins by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences USA*, 100(4):1706–1710, 2003.

[OKKA91] E.K. O'Shea, J.D. Klemm, P.S. Kim, and T. Alber. X-ray structure of the gcn4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, 254(5031):539–544, 1991.

[Osg00] D.J. Osguthorpe. *Ab initio* protein folding. *Current Opinion in Structural Biology*, 10:146–152, 2000.

[PB02] G. Pollastri and P. Baldi. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18(Supplement 1):S62–S70, 2002.

[PCB51] L. Pauling, R.B. Corey, and H.R. Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Science USA, Chemistry*, 37:205–211, 1951.

[PCC$^+$95] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.R. Ross, T.E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Computational Physics Communications*, 91:1–41, 1995.

[PGF99] R. Preißner, A. Goede, and C. Frömmel. Spare parts for helix-helix interaction. *Protein Engineering*, 12(10):825–831, 1999.

[PGH$^+$04] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605–1612, 2004.

[PGZR03] D.E. Platt, C. Guerra, G. Zanotti, and I. Rigoutsos. Global secondary structure packing angle bias in proteins. *Proteins: Structure, Function, and Genetics*, 53:252–261, 2003.

[PMP99] R.V. Pappu, G.R. Marshall, and J.W. Ponder. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nature Structural Biology*, 6(1):50–55, 1999.

[PR04] G.A. Petsko and D. Ringe. *Protein Structure and Function.* New Science Press Ltd, London, 2004.

[PR05] M. Punta and B. Rost. PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, 2005.

[PSTW78] D.C. Phillips, M.J.E. Sternberg, J.M. Thornton, and I. Wilson. An analysis of the structure of triose phosphate isomerase and its comparison with lactate dehydrogenase. *Journal of Molecular Biology*, 119:329–351, 1978.

[RAO⁺78] S.J. Remington, W.F. Anderson, J. Owen, L.F. Ten Eyck, C.T. Grainger, and B.M. Matthews. Structure of the lysozyme from bacteriophage T4: an electron density map at 2.4 Å resolution. *Journal of Molecular Biology*, 118:81–98, 1978.

[RB93] B.V.B. Reddy and T.L. Blundell. Packing of secondary structural elements in proteins: Analysis and prediction of inter-helix distances. *Journal of Molecular Biology*, 233:464–479, 1993.

[RR78] T.J. Richmond and F.M. Richards. Packing of α-helices: Geometrical constraints and contact areas. *Journal of Molecular Biology*, 119:537–555, 1978.

[RS68] G.N. Ramachandran and V. Sasisekharan. Conformations of polypeptides and proteins. *Advanced Protein Chemistry*, 241:283–438, 1968.

[RW93] G.D. Rose and R. Wolfenden. Hydrogen bonding, hydrophobicity, packing, and protein folding. *The Annual Review of Biophysics and Biomolecular Structures*, 22:381–415, 1993.

[SFWB$^+$05] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker. Progress in modeling of protein structures and interactions. *Science*, 310:638–642, 2005.

[SM98] R. Samudrala and J. Moult. Determinants of side chain conformational preferences in protein structures. *Protein Engineering*, 11(11):991–997, 1998.

[SMW95] R.A. Sayle and E.J. Milner-White. RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, 20(9):374–376, 1995.

[SR02] R.H. Spencer and D.C. Rees. The α-helix and the organization and gating of channels. *Annual Review of Biophysics and Biomolecular Structure*, 31:207–233, 2002.

[SRPX97] Z. Sun, X. Rao, L. Peng, and D. Xu. Prediction of protein supersecondary structures based on the artificial neural network method. *Protein Engineering*, 10(7):763–769, 1997.

[SRRJ94] M.C. Surles, J.S. Richardson, D.C. Richardson, and F.P. Brooks Jr. Sculpting proteins interactively: continual energy minimization embedded in a graphical modeling system. *Protein Science*, 3(2):198–210, 1994.

[SSS$^+$95] M.S. Sansom, H.S. Son, R. Sankararamakrishnan, I.D. Kerr, and J. Breed. Seven-helix bundles: molecular modeling via restrained molecular dynamics. *Biophysical Journal*, 68(4):12951310, 1995.

[STI$^+$97] T. Shimizu, A. Toumoto, K. Ihara, M. Shimizu, Y. Kyogoku, N. Ogawa, Y. Oshima, and T. Hakoshima. Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *The EMBO Journal*, 16(16):4689–4697, 1997.

[Str88] L. Stryer. *Biochemistry*. W.H. Freeman and Company, New York, 3rd edition, 1988.

[Sur92] M.C. Surles. Interactive modeling enhanced with constraints and physics with applications in molecular modeling. In *Proceedings of the 1992 symposium on Interactive 3D graphics*, pages 175–182, 1992.

[SW98] E.L.L. Sonnhammer and J.C. Wootton. Dynamic contact maps of protein structures. *Journal of Molecular Graphics and Modelling*, 16:1–5, 1998.

[TMBA01] W.R. Taylor, A.C.W. May, N.P. Brown, and A. Aszódi. Protein structure: geometry, topology and classification. *Reports on Progress in Physics*, 64:517590, 2001.

[TS04] A. Trovato and F. Seno. A new perspective on analysis of helix-helix packing preferences in globular proteins. *PROTEINS: Structure, Function, and Bioinformatics*, 55:1014–1022, 2004.

[TSUS01] D.P. Tieleman, I.H. Shrivastava, M.R. Ulmschneider, and M.S.P. Sansom. Proline-induced hinges in transmembrane helices: Possible roles in ion channel gating. *PROTEINS: Structure, Function, and Genetics*, 44(2):63–72, 2001.

[UTS05] M.B. Ulmschneider, D.P. Tieleman, and M.S.P. Sansom. The role of extra-membranous inter-helical loops in helix-helix interactions. *Protein Engineering, Design & Selection*, 18(12):563–570, 2005.

[VAM$^+$01] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, and 275 others. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[VKD97] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

[WAK69] C.S. Wright, R.A. Alden, and J. Kraut. Structure of subtilisin BPN' at 2.5 Ångström resolution. *Nature*, 221:235–242, 1969.

[Wal97] D. Walther. WebMol - a Java-based PDB viewer. *Trends in Biochemical Sciences*, 22(7):274–275, 1997.

[WEA96] D. Walther, F. Eisenhaber, and P. Argos. Principles of helix-helix packing in proteins: the helical lattice superimposition model. *Journal of Molecular Biology*, 255:536–553, 1996.

[WFC$^+$03] J. Westbrook, Z. Feng, L. Chen, H. Yang, and H.M. Berman. The Protein Data Bank and structural genomics. *Nucleic Acids Research*, 31(1):489–491, 2003.

[WK88] A. Witkin and M. Kass. Spacetime constraints. *Computer Graphics*, 22(4):159–168, 1988.

[WSC98] D. Walther, C. Springer, and F.E. Cohen. Helix-helix packing angle preferences for finite helix axes. *PROTEINS: Structure, Function, and Genetics*, 33:457–459, 1998.

[WW03] J. Walshaw and D.N. Woolfson. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *Journal of Strucural Biology*, 144:349–361, 2003.

[YGW00] J.-M. Yoon, Y. Gad, and Z. Wu. Mathematical modeling of protein structure using distance geometry. Technical Report TR00-24, Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA, 2000.

Appendix A

Implementing Hippy

Programming is one of the most difficult branches of applied mathematics; the poorer mathematicians had better remain pure mathematicians.

-Edsger Dijkstra

Hippy was implemented in OpenGL with the aim of platform independence with a powerful graphics engine. The program has been compiled and tested on Windows XP, although all of the code is present so that it should compile on Linux and Apple operating systems as well. The core of the program is a window class which included the functionality of for rotating, zooming and panning, and all of the window classes in Hippy extend this class. To clarify the structure of the Hippy package, a rudimentary class diagram is shown in Figure A.1.

The basic classes used in the package are `seq`, `vector`, `header` and `font`. `seq` is an implementation of a wrapper for arrays, enabling basic operations such as adding items, removing select items and checking for membership. `vector` is simply a vector class, where member attributes `x`,`y`, and `z` correspond to indices 0,1, and 2 in an

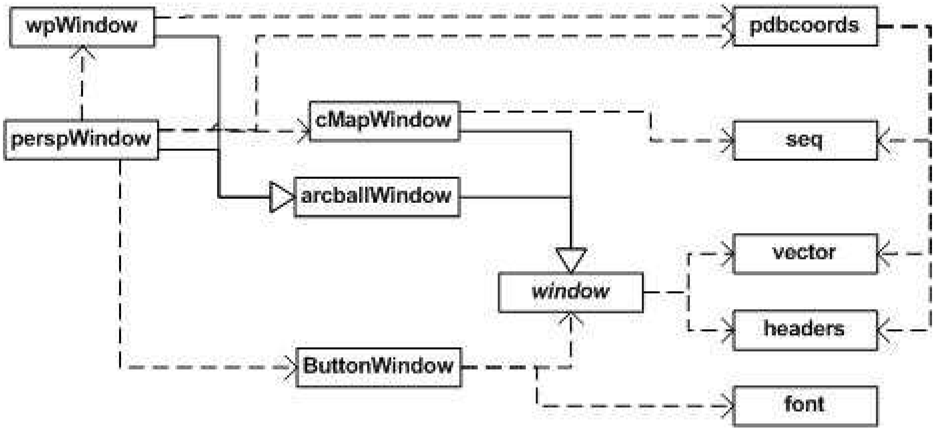


Figure A.1: This is a basic UML diagram of the Hippy package to illustrate the interaction of the different classes.

array respectively. `header` is just an include file, where constants are defined and the different includes for platform independence are found. `font` is a class for creating fonts in the windows.

`pdbcoords` implements functions for parsing the PDB files. Depending on the type of data that is required, it finds the amino acid indices for the two desired helices, and extracts the corresponding data for either the entire helix or just the alpha carbons. The data is stored in `seq` data structures, and provides accessor functions for retrieving the data.

The `window` class is abstract and provides most of the basic functionality required for a window. It includes basic functions such as assigning identifiers to the window, correct repositioning and reshaping, and the virtual action functions.

The `arcballWindow` class is abstract and extends the `window` class. This class provides the functions for the three-dimensional windows. Some of the action functions of the `window` class are implemented so that the user is able to rotate, translate, and zoom the contents of the window.

cMapWindow implements the window class to provide the contact map window for Hippy. The only significant action performed by this class is that upon detecting a mouse click, the corresponding contact in the map is determined and stored and a flag is set. This way, when Hippy discovers the flag it may cause the corresponding contact bar in the main window to flash.

buttonWindow is a misnomer; originally it was thought that this window would provide an array of buttons that the user could toggle to activate or deactivate the various features of Hippy. It was decided afterwards that this is not necessary, as it is quite clear which features are in effect. Instead, the window provides a listing of the features available and the keystrokes that are used to activate and deactivate them. buttonWindow implements the window class.

The wpWindow class uses the pdbcoords class to obtain the coordinates of every alpha carbon in the protein from the PDB file. It implements the arcballWindow class so that the user can manipulate the protein in three dimensions. The default display for the window is to have the helices of index 1 and 2 highlighted. It is linked dynamically to the main window class so that if the user changes which helices should be displayed, the contents of all other windows are modified accordingly.

Finally, perspWindow is the heart of Hippy. The basic structure is similar to wpWindow in that it implements arcballWindow and uses pdbcoords to obtain the coordinate data for the helices. When the user changes which helices should be displayed, the current coordinate data is discarded and the data for the new helices is retrieved using pdbcoords. If the side chains are requested, they are displayed along with the appropriate bonds. There are two different ways of drawing the bonds. One would be that every pair of atoms at are within a distance equivalent to the length of a covalent bond are connected with a bond. This method was not used because it seems expensive, and there is a better method available. The atoms in a side chain are always placed in the PDB file in the same order. It is much more code that the previous method, but we can have a separate condition in a draw function for each species of side chain. This way, we connect the bonds in the proper way each time efficiently.

The platform was developed using the Eclipse IDE with the CDT plug-in for C++. Hippy has been tested and runs well on a variety of Windows XP machines. The program has been developed to be platform independent (the code is in place so that is should compile and run on both Linux and Apple operating systems), but it has not been tested on any other platforms as this time.

Appendix B

Organic Chemistry Primer

A physicist is the atoms' way of thinking about atoms.

-Anonymous

Organic chemistry is not as scary as many people believe it to be. It is simply chemistry based on molecules that are carbon-based. The molecules often polymerize, fortunately for us; examples of such molecules are fossil fuels, plastics, DNA, and proteins. Obviously, the latter are of great concern to this thesis.

B.1 Polymerization

Polymerization is the process where two smaller molecules (referred to as monomers if the molecule is a single building block, or an oligomer if the molecule is composed of multiple monomers already) come together to bond and form a single larger molecule. In the example of proteins, this polymerization process is detailed in Figure 2.1 on page 9. The polymerization process in this case involves the reaction of the amino end of one amino acid with the carboxyl end of another. An amino group on a molecule is simply the presence of a nitrogen with a hydrogen bonded to it. A carboxyl group is more complex, it involves a carbon atom having two oxygens bonded to it. One of the oxygens is in a hydroxyl group, where the oxygen is bonded to both the carbon atom and a hydrogen atom, and the other oxygen atom is bonded to the carbon in a double bond. A covalent bond (one where there is an equal sharing of electrons) is formed between the nitrogen atom of the first amino acid and the beta carbon of the second. A water molecule produced as a by-product of the reaction, as the hydroxyl group bonds with the hydrogen atom that was bonded to the nitrogen atom. This is why polymerized amino acids are referred to as amino acids residues.

B.2 *cis-trans* Isomerism

The back bone of a protein, or any polymer for that matter, has preferred bonding angles, as discussed in the body of the thesis. Planar configurations are common in polymers, and in this situation there are two possible configurations that molecules can adopt if there is a planar chain with an angle in the backbone, a phenomenon referred to as *cis*-*trans* isomerism (or sometimes geometric isomerism) [HHC95]. A *cis* configuration is one where the previous and following chain bonds are on the same side of a plane aligned with the bond of interest. A *trans* configuration would be one with a protein having the exact same molecular composition as the other (thus the isomerism), but the previous and following bonds along the backbone to the one of interest are on opposite sides of a plane aligned with it. This is illustrated in Figure B.1. From latin, *cis* means on the same side, and *trans* means across. There is no rotation possible about the bond in this planar region, so once the molecule is formed it will remain in one configuration or the other.

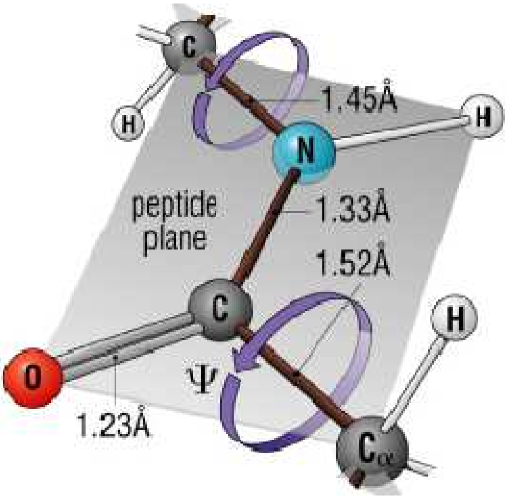


Figure B.1: This section of a protein backbone (the brown bonds) is a planar region. There is no rotation possible around the central bond. This is an example of the *trans* configuration, as the chain bonds before and after this central one are on opposite sides from each other of the central bond. If the bottom two atoms (the oxygen and the alpha carbon) were switched, the bonds would be on the same side, and it would be the *cis* isomer of our example. This image is used with permission from [PR04].