Computational Studies of Protein Helix Kinks

Henry Wilman

Systems Approaches to Biomedical Sciences Industrial Doctoral Centre

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2014

Acknowledgements

This thesis would not have been possible without the help and encouragement of many people. I would like to thank my supervisors: Charlotte, for all of the advice and encouragement she has given me in the last four years, and Jiye, for his helpful guidance throughout. The readability of this thesis is a testament to Charlotte's advice and attention to developing my skills in the writing process. This research was funded by the EPSRC and UCB, with travel grants from Hertford College.

I would like to thank everyone in the Oxford Protein Informatics Group (OPIG) for their contributions, big and small, to my work. In particular, thanks to Seb (for his guidance in all things programming and proteins at the start of my project), JP (for his wizardry in producing AHAH), Bernhard (for all his input in making the AHAH paper excellent), and Jamie (for his frequent interesting discussions of our work). I also have to thank everyone who participated in AHAH. I was pleasantly surprised by how many people responded to my pleas for help! Thank you all.

The last four years have been made enjoyable by so many people, and long may it continue (in spite of any geographical constraints). Especially James and Hannah, who with whom I have shared this adventure. They and the other members of OPIG have created to some brilliant experiences. I have enjoyed so many evenings of food, games, and wine with a group of special people. I would like to thank Erika, for her unstinting support, hugs, and endless cups of morning tea. Of course, none of this would have been possible without the support and encouragement from my family. I have come along way since being taught to count smarties at the kitchen table!

Abstract

Kinks are functionally important structural features found in the α-helices of many proteins, particularly membrane proteins. Structurally, they are points at which a helix abruptly changes direction. Previous kink definition and identification methods often disagree with one another.

Here I describe three novel methods to characterise kinks, which improve on existing approaches. First, Kink Finder, a computational method that consistently locates kinks and estimates the error in the kink angle. Second the B statistic, a statistically robust method for identifying kinks. Third, Alpha Helices Assessed by Humans, a crowdsourcing approach that provided a gold-standard data set on which to train and compare existing kink identification methods.

In this thesis, I show that kinks are a feature of long α-helices in both soluble and membrane proteins, rather than just transmembrane α-helices. Characteristics of kinks in the two types of proteins are similar, with Proline being the dominant feature in both types of protein. In soluble proteins, kinked helices also have a clear structural preference in that they typically point into the solvent.

I also explored the conservation of kinks in homologous proteins. I found examples of conserved and non-conserved kinks in both the helix pairs and the helix families. Helix pairs with non-conserved kinks generally have less similar sequences than helix pairs with conserved kinks. I identified helix families that show highly conserved kinks, and families that contain non-conserved kinks, suggesting that some kinks may be flexible points in protein structures.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Overview

Proteins are biological macromolecules that, along with RNA and DNA, are integral to the majority of the processes in an organism (Alberts *et al.*, 2008). The function of a protein depends on its three dimensional structure. Knowing the structure of proteins can help us to understand how they work, what goes wrong with them, and how to interfere with the processes they take part in. Proteins which reside in the membranes of cells (integral membrane proteins) are involved in many processes in the cell. The specific environment of the membrane, which is crucial to membrane protein structure, is difficult to experimentally replicate. Consequently,

our structural and functional understanding of membrane proteins is limited.

This thesis describes my studies of protein helix kinks. Kinks are important functional and structural features that occur particularly in integral membrane proteins. They are important in the function of many important proteins, such as G protein coupled receptors (GPCRs) and ion channels.

The initial vision of my research project was to identify kinks in known protein structures, and characterise them with a view to using this knowledge to predict where and how kinks would appear in protein structures. However, it soon became apparent that the identification of kinks is not a trivial task. Although there were numerous papers that described kinks, the definition of 'kink' changed from study to study, and there is little agreement as to the causes and characteristics of kinks. Thus, I developed my own computational methods to identify kinks (Chapter 2). In Chapter 3, I describe how they, along with three other published methods, compare to my crowd-sourcing approach, Alpha Helices Assessed by Humans. In Chapter 4, I investigate the characteristics of kinks in a non-redundant set of protein structures, and show that the characteristics and frequency of kinks in soluble protein helices are similar to those of kinks in integral membrane proteins. In Chapter 5, I show that some helix kinks are conserved among related proteins, while others are not. The patterns of conservation are similar in membrane and soluble proteins.

In this Chapter, I describe the basics of protein structure, and how it can be determined by experiment, before I explain the methods for the identification of helices and kinks. The final section discusses the reasons for my specific interest in kinks.

1.2 Helix kinks

Helix kinks are a poorly understood feature of protein structure. They are a local distortion in helices, and are an important functional feature of many membrane proteins. Kinks are regions in α-helices where the helix direction changes abruptly, however the specific definition

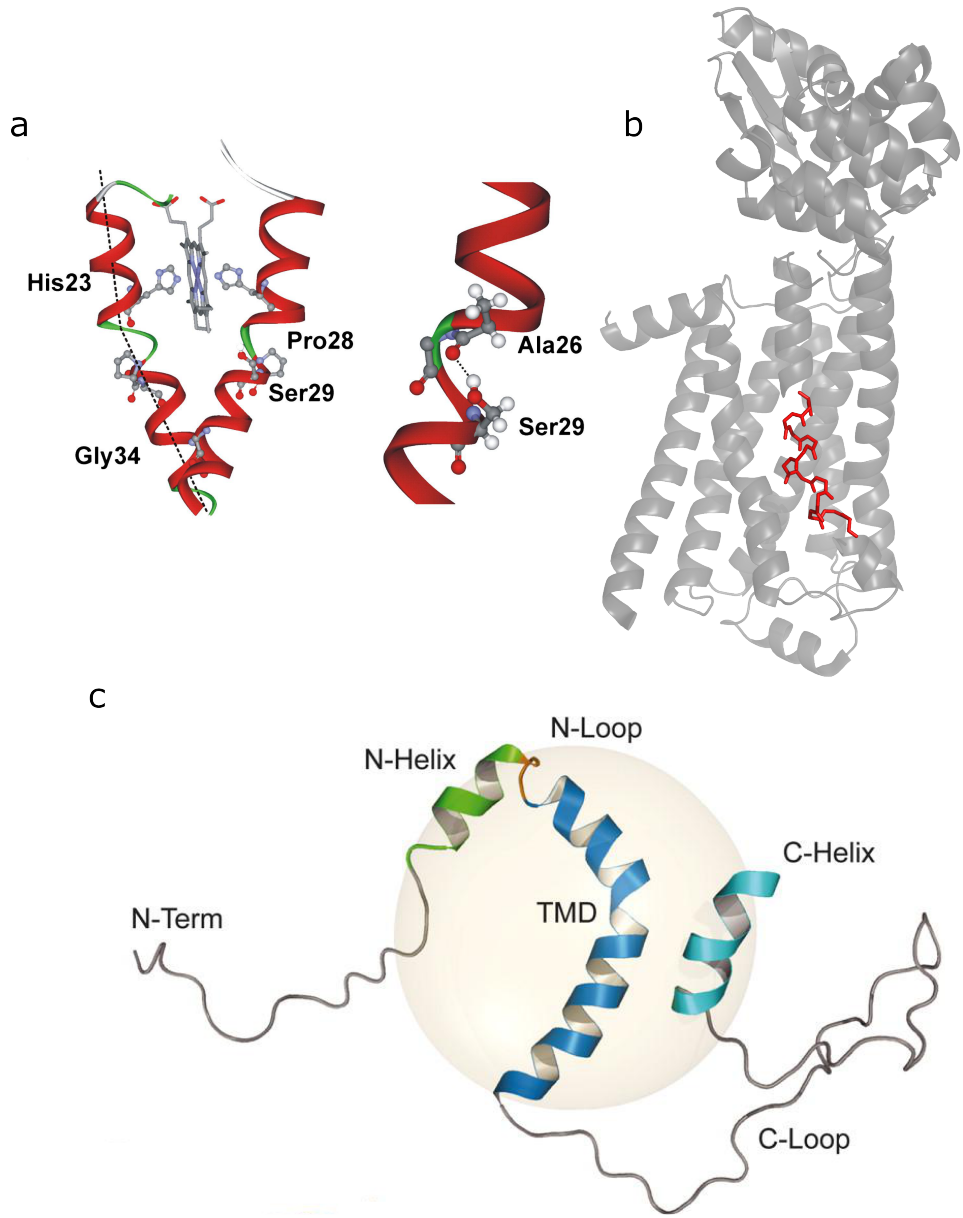


Figure 1.1: **Three example kinks.** (a) Kink in cytochrome b_{559}. Reproduced from Biochimica et Biophysica Acta (BBA) Biomembranes, 1818, Weber, Tome, Otzen & Schneider, A Ser residue influences the structure and stability of a Pro-kinked transmembrane helix dimer, 2103-7, Copyright (2012) with permission from Elsevier. (b) The kink in the sixth transmembrane helix of a GPCR (PDB code 2rh1). Residues around the kink are highlighted in red. (c) Kink in the amyloid precursor protein. Reproduced from Barrett *et al.* (2012). Reprinted with permission from AAAS.

varies from study to study. There are many terms used to describe this type of helix distortion with kink (Bansal *et al.*, 2000; Devillé *et al.*, 2008; Kneissl *et al.*, 2011; Kumar & Bansal, 2012; Meruelo *et al.*, 2011; Seifert *et al.*, 2014; Werner & Church, 2013; Wilman *et al.*, 2014b) being the most popular, but bend (Langelaan *et al.*, 2010), hinge (Sansom & Weinstein, 2000), alteration (Hischenhuber *et al.*, 2012, 2013), and cusp (de Almeida & Holoshitz, 2011) have also been used. Such helix kinks are a common feature of α-helical membrane proteins (Hall *et al.*, 2009; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Nugent & Jones, 2011; Rigoutsos *et al.*, 2003; Werner & Church, 2013; Wilman *et al.*, 2014b) as well as long helices in soluble proteins (Devillé *et al.*, 2008; Rey *et al.*, 2010; Wilman *et al.*, 2014b). They have been implicated in the function of G protein-coupled receptors (Bettinelli *et al.*, 2011; Schwartz *et al.*, 2006; van der Kant & Vriend, 2014; Yohannan *et al.*, 2004b), in the conformational change of ion channels (Fowler & Sansom, 2013; Suchyna *et al.*, 1993; Tieleman *et al.*, 2001), in heme binding in Cytochrome b_{559} (Weber *et al.*, 2012), and in the function of many other proteins (Barrett *et al.*, 2012; Ni *et al.*, 2011; Sansom & Weinstein, 2000). For examples, see Figure 1.1. Accurate determination of kinks is important to the modelling of membrane proteins (Deflorian & Jacobson, 2011; Werner & Church, 2013), computational docking studies of compounds to membrane proteins (Kufareva *et al.*, 2011), and general understanding of protein structure.

1.3 Proteins

1.3.1 Protein structure

Proteins are formed from a polymer chain of amino acids. There are 20 main naturally occurring amino acids (Figure 1.2). This polymer chain can fold up into a three dimensional structure. In natural systems, most proteins fold to a single 3D shape. Protein structure is often broken into four scales - primary (the order of the amino acids), secondary (the local structure of the amino acid chain), tertiary (the structure of a single chain), and quaternary (the structure of chains relative to one another) (Klotz *et al.*, 1970; Linderstrø m Lang, 1952).

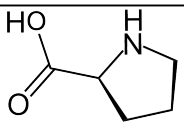
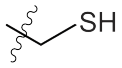
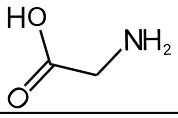
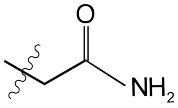
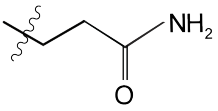
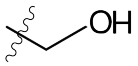
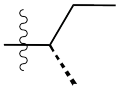
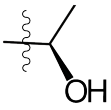
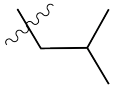
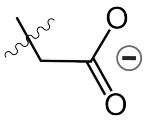
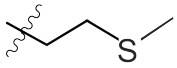
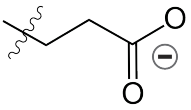
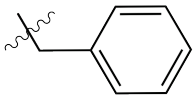
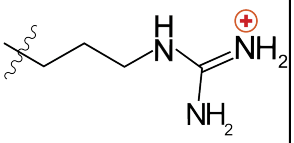
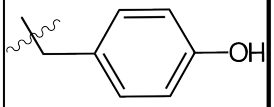
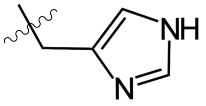
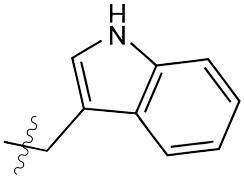
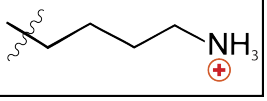
Name			Sidechain	Name			Sidechain
P	Pro	Proline		C	Cys	Cysteine	
G	Gly	Glycine		N	Asn	Asparagine	
A	Ala	Alanine		Q	Glu	Glutamine	
V	Val	Valine		S	Ser	Serine	
I	Ile	Isoleucine		T	Thr	Threonine	
L	Leu	Leucine		D	Asp	Aspartic acid	
M	Met	Methionine		E	Glu	Glutamic acid	
F	Phe	Phenyalanine		R	Arg	Arginine	
Y	Tyr	Tyrosine		H	His	Histidine	
W	Trp	Tryptophan		K	Lys	Lysine	

Figure 1.2: **The side chains of the 20 naturally occurring amino acids.** Top left: special cases, P and G. Middle left: Hydrophobic. Bottom left: aromatic and hydrophobic. Top right: polar, uncharged. Bottom left: electrically charged. Proline and glycine are shown with their backbone residues. The side chain is the R group in Figure 1.3.

1.3.1.1 Primary structure

Protein primary structure describes the order of amino acids (peptides) in the protein polymer (polypeptide). Figure 1.3 shows a two dimensional representation of an amino acid. It is made up of the sidechain, R, which varies from amino acid to amino acid (Figure 1.2), and the main chain (including the heavy atoms N, C^{α},C',and O), which is common to all amino acids. Figure 1.3 also shows how the peptide can form bonds to other peptides, and in doing so, polymerise.

There are 20 different naturally occurring amino acids, each with its own characteristics (Figure 1.2). The side chains differ in size, charge, and acidity. The amino acids are often grouped based on the characteristics of their side chains. The C^{α} atom is chiral for all natural amino acids except glycine, i.e. there are two non-superimposable ways (enantiomers) to arrange the four groups covalently bonded to it (Figure 1.3a). All natural amino acids are the L-enantiomer except glycine. Figure 1.3b shows how two amino acids can bond to form a dipeptide. Further amino acids can be added to form a polypeptide (e.g. Figure 1.3c). Proteins are polypeptides, made up of chains of tens, hundreds, or thousands of amino acids.

1.3.1.2 Ramachandran angles

The three dimensional structure of proteins can be characterised by the dihedral angles along the backbone. Dihedral angles describe the relative positions of four consecutive atoms in the backbone. These are the φ (describing the relative positions of the C', N, C^{α}, and C' atoms), ψ (describing the relative positions of the N, C^{α}, C' and N atoms), and ω (describing the relative positions of the C^{α}, C', N, and C^{α} atoms) angles. These angles are shown in Figure 1.4 as Newman projections, where the atoms are arranged so that the viewer is looking down the central bond, with the groups bonded to the closer atom extending from the centre of the diagram, and the groups bonded to the further atom extending from the edge of the circle (Newman, 1955). The dihedral angles are in the range $(-180°,180°)$. Due to amide bonding, ω is restricted to angles close to either $0°$ or $180°$. The two smaller groups (O and H) are normally not adjacent to one another, hence ω is much more frequently close to $180°$ than it is to $0°$.

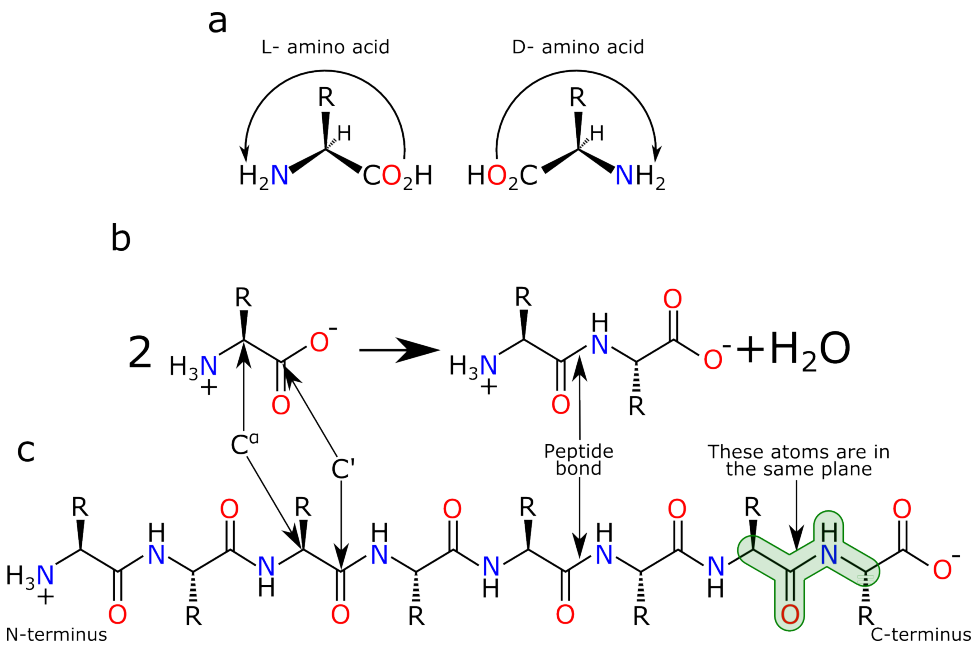


Figure 1.3: **The structure of amino acids**, using stick representation. Covalent bonds are shown as sticks, with atoms at the vertices. Vertices without labels are carbon atoms, and hydrogen atoms attached to carbon are generally not shown. Each carbon has four bonds, and where fewer than this are shown, hydrogen atoms make up the remainder. (a) The two enantiomers of amino acids. The four bonds around the C^{α} are arranged tetrahedrally, so there are two possible, non-superimposable, arrangements. When viewed with the hydrogen pointing away from the viewer, if the CO_2H, R, and NH_2 groups are arranged anticlockwise, it is given the L- designation, otherwise the D- designation. (b) How two amino acids can react to form a dipeptide, where the nitrogen in one amino acid reacts with the carboxyl group of the other, to form the peptide bond. Solid triangles indicate the bond is pointing out from the page, whereas dashed triangles indicate groups pointing into the page. (c) A polypeptide, made of 8 amino acids condensed together. The backbone is formed of the repeating N, C^{α}, C', atoms, running horizontally across the page. The C'-N peptide bond is typically planar. Consequentially, groups of successive C^{α}, C', N, and C^{α} atoms lie in the same plane.

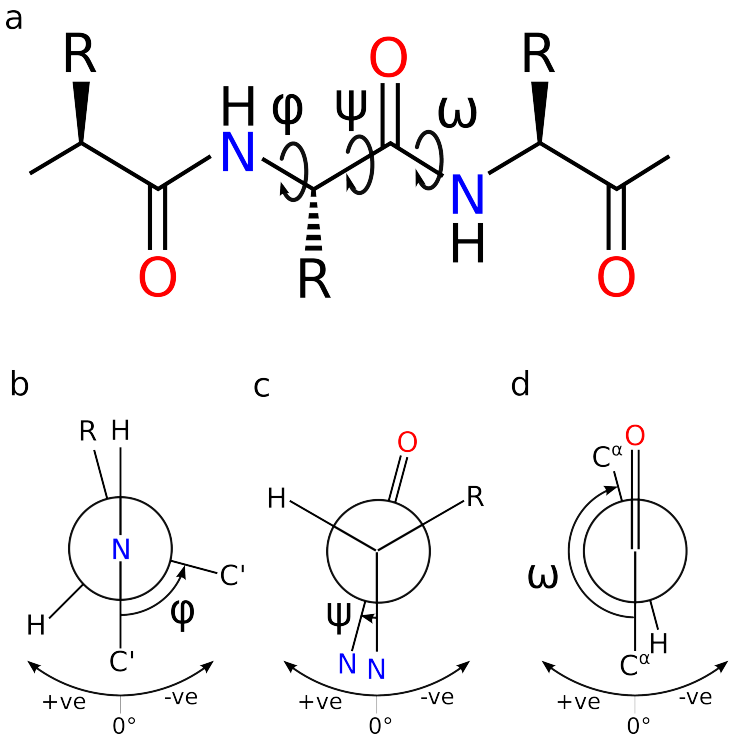


Figure 1.4: Definitions of protein dihedral angles. (a) The dihedral angles for an amino acid within a protein chain. (b), (c), and (d) Newman projections of the three angles. Each projection looks down a bond, the first atom in the bond is shown in the centre of the diagram. The long lines show bonds from this atom to other atoms, whilst the lines that start at the edge of the circle indicate the bonds from the atom at the far end of the central bond. Hence, in (b), the Nitrogen is bonded to the H and C', while the C^{α} (not shown) is bonded to the H, R and C' atoms.

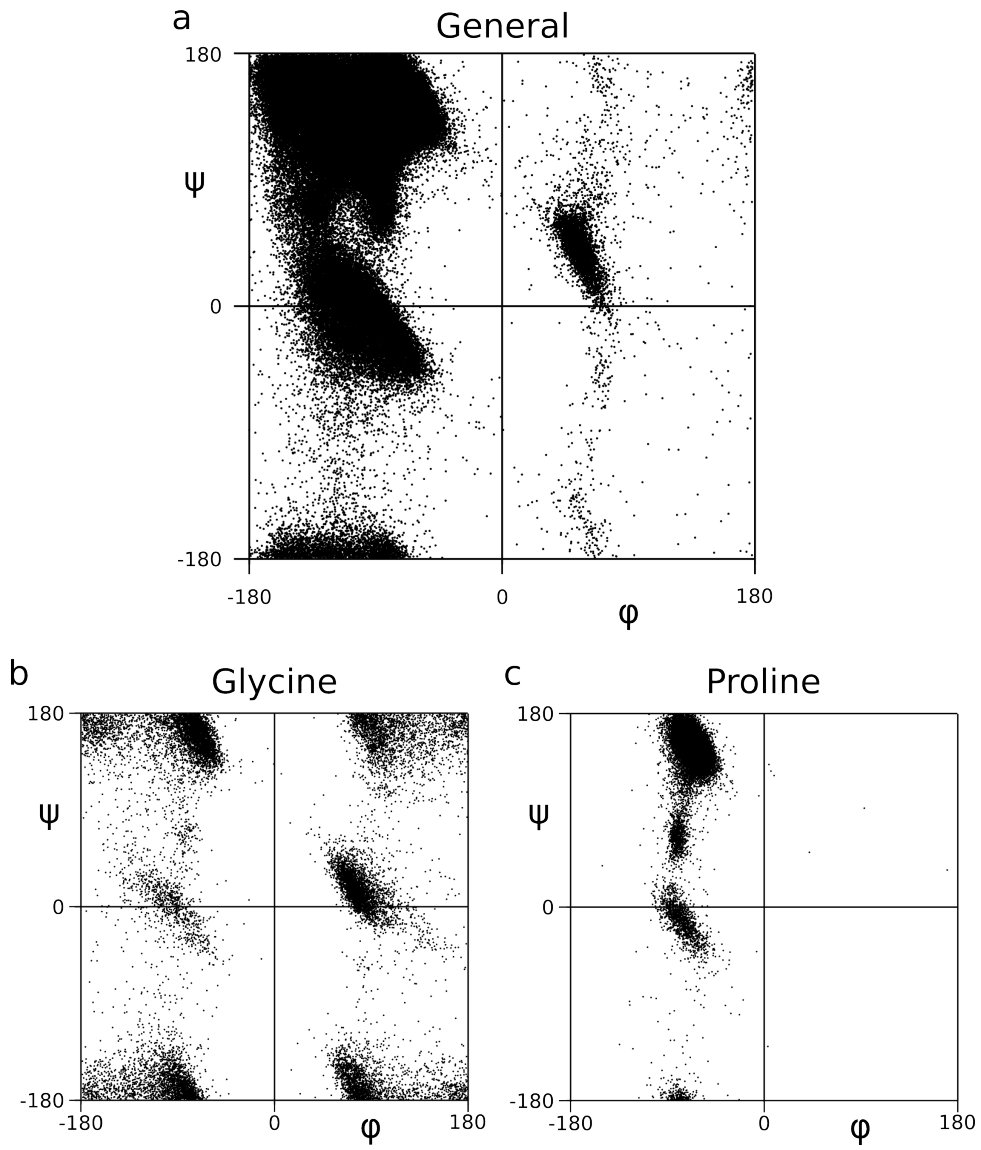


Figure 1.5: **Ramachandran plots.** (a) For all amino acids except proline and glycine, (b) for glycine, and (c) for proline. φ and ψ are defined in Figure 1.4. Data taken from PDB structures culled to 30% sequence identity, excluding α-helices and β-sheets.

The dihedral angles φ and ψ can take on a much larger range of angles. Like with the ω angle, it is more favourable for large groups to be opposite one another. As a result, residues are more often found with φ and ψ angles within certain ranges. Figure 1.5a shows the φ and ψ angle ranges for protein residues. Residues typically favour conformations with negative φ angles, due to their L- chirality.

The preferred φ and ψ angles differ for the different amino acids, due to the different sizes and characteristics of their R groups (Figure 1.5a). The preferences are generally quite similar, but glycine (Figure 1.5b) and proline (Figure 1.5c) are the most different. Glycine's plot is much more symmetric than the general plot. This is because its R group is H, making it a non-chiral molecule, and making negative and positive φ conformations equally favoured. Proline has a cyclic structure, with both the C^{α} and N atoms from the amino acid backbone contained within the ring (Figure 1.2). This severely limits its available φ angles, as shown in Figure 1.5c.

There are two densely populated areas in the Ramachandran plots. These are characteristic of the local three-dimensional structures that protein chains form, called protein secondary structure.

1.3.1.3 Secondary structure

In globular proteins, there are two frequently observed local structures, α-helices and β-sheets. These make up around 35% and 20% of protein structure respectively (although these values vary depending on the set of proteins and method used to identify secondary structure within them) (Martin *et al.*, 2005). They are stabilised by characteristic hydrogen bonding patterns between the amide groups on the backbone of the protein, from the carbonyl oxygen to the hydrogen bonded to the nitrogen atom. Hydrogen bonds are 'an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X-H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation' (Arunan *et al.*, 2011). They are polar, electrostatic, interactions that also show some covalent nature.

1.3.1.4 Helices

The right handed α-helix is a regular repeating 3D structure in proteins (Figure 1.6). Residues within α-helices have (φ,ψ) angles in the region of $(-63°,-40°)$. This arrangement facilitates hydrogen bonds between the i^{th} (CO) and $i+4^{th}$ (NH) amide groups $(i+4 \rightarrow i)$[1], and for the amino acid side chains to point out of the helix. It is characterised by the parameters shown in Figure 1.6, which are the number of residues per turn, the rise (the distance between successive C^{α} atoms along the helix axis), their Ramachandran angles, the hydrogen bonds between the backbone atoms, and the radius.

As well as the α-helix, there are two other helix structures that are observed in proteins. These correspond to helices with hydrogen bonds between the i^{th} (CO) and $i+3^{th}$ amide groups (3_{10}-helix) and bonds between the i^{th} and $i+5^{th}$ amide groups (π-helix). Examples of the three helices are shown in Figure 1.7. Both of these helix types are much less regular than α-helices, and their (φ,ψ) angles vary depending on their position in the helix (Enkhbayar *et al.*, 2006; Fodje & Al-Karadaghi, 2002).

There are known 'ideal' parameters for each of these helices (Pauling *et al.*, 1951), but helices in protein structures deviate from these, lying within allowed regions. The ideal parameters, and average observed parameters are shown in Table 1.1

π-helices are particularly unstable, as there is space in the centre of the helix, that is not large enough to admit small molecules, e.g. water. In addition, the $i+5 \rightarrow i$ hydrogen bond has a greater entropic cost than either the $i+4 \rightarrow i$ or $i+3 \rightarrow i$ bond. For this same reason, the 3_{10} helix is often suggested as a folding intermediate on the path towards an α-helix (Enkhbayar *et al.*, 2006). Although helices are characterised by hydrogen bonds over three, four, or five residues, this does not preclude other hydrogen bonds being present. Indeed, 24% of hydrogen bonds in proteins are bifurcated, i.e. they involve three heavy atoms (either two donors and one acceptor, or two acceptors and one donor) (Preissner *et al.*, 1991).

[1] In this notation, the arrow points from the residue with the hydrogen-bond donor (NH) to the residue with the hydrogen-bond acceptor (CO)

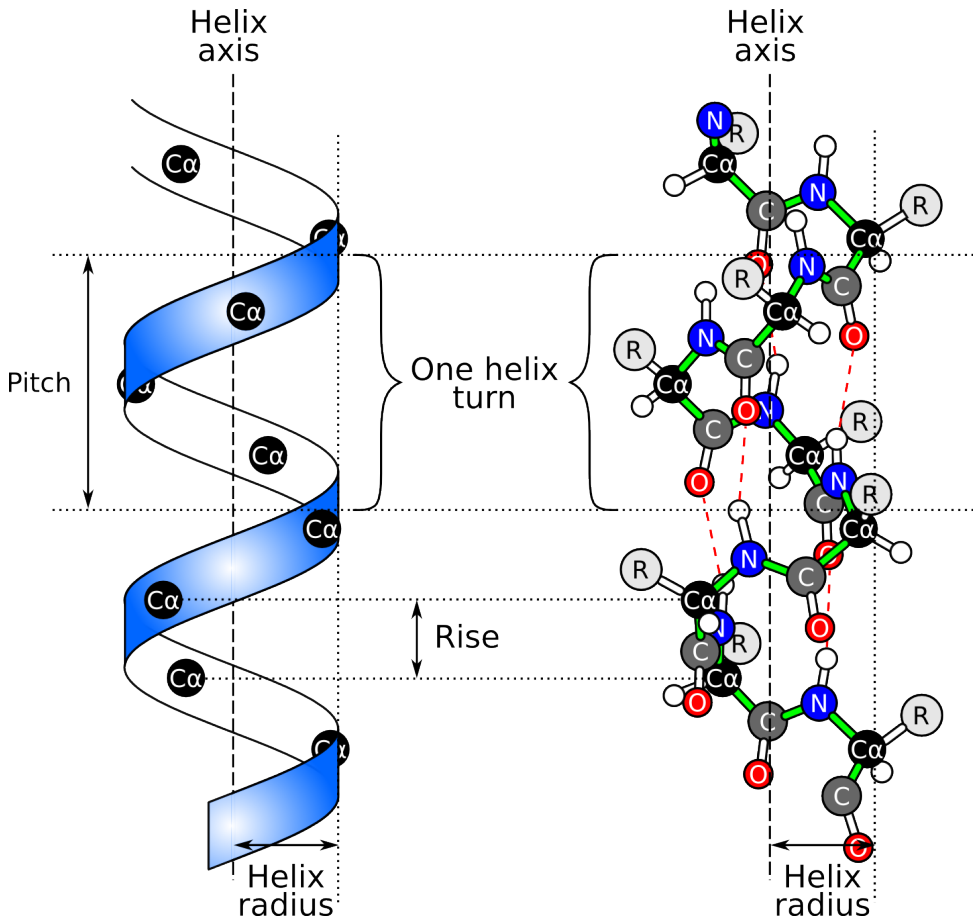


Figure 1.6: **Helix geometry.** Diagrams showing the atomic structure of an alpha helix. A helix turn is one full rotation of the helix, while the rise per residue is the distance along the helix axis between two adjacent C^{α}s. The pitch is the rise over a whole helix turn.

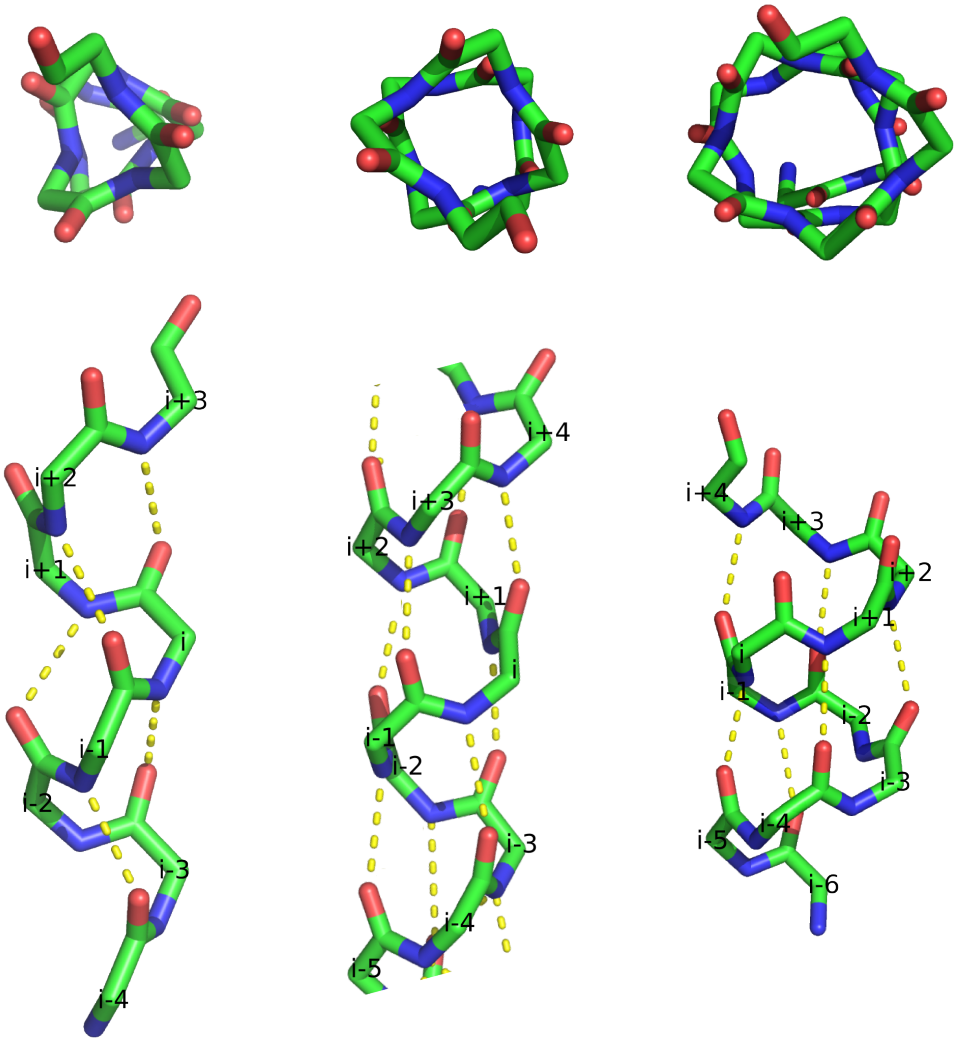


Figure 1.7: **The three types of common protein helices**: 3_{10} (left), α (centre), and π (right). Taken from proteins: 2R9R, chain H, residues 297 to 305, 3BMX4 426-433, and 3BMXA 298-309 respectively. The hydrogen bonds characteristic to each helix are shown by dashed yellow lines. Figures produced using PyMOL (Schrödinger LLC, 2014) using stick representations, with cbag colouring (Oxygen red, Nitrogen blue, Carbon green). Sidechains and hydrogen atoms not shown.

Table 1.1: **Helix parameters.**. The top half shows the observed average parameters in protein structures. The bottom half shows the parameters for ideal helices. Data taken from: [a]Enkhbayar *et al.* (2006), [b]Barlow & Thornton (1988), [c]Pauling *et al.* (1951), [d]Fodje & Al-Karadaghi (2002).

Observed averages

Helix type	Hydrogen bonds	Rise (Å)	Residues per turn	$\varphi(^\circ)$	$\psi(^\circ)$	Backbone radius (Å)	Occurrence
3_{10}	$i \leftarrow i+3$	2.0^a	3.2^a	-74.0^a	-4.0^a	2.0^b	$4\%^a$
α	$i \leftarrow i+4$	1.5^b	3.54^b	-62^b	-41^b	2.3^b	$32-42\%^b$
π	$i \leftarrow i+5$	1.2^d	4.4^d	-76^d	-41^d	2.8^d	$0.3\%^d$

Ideal

Helix type	Hydrogen bonds	Rise (Å)	Residues per turn	$\varphi(^\circ)$	$\psi(^\circ)$	Backbone radius (Å)	Typical lengths
3_{10}	$i \leftarrow i+3$	2.0^c	3.0^c	-74^c	-4^c	1.8^c	5-15 [a]
α	$i \leftarrow i+4$	1.5^c	3.65^c	-48^c	-57^c	2.3^c	5-40
π	$i \leftarrow i+5$	1.14^d	4.4^d	-57^d	-70^d	2.76^d	7-13[d]

1.3.1.5 β-sheets

β-sheets are made up of extended sections of the protein chain, called β-strands. These residues have (φ,ψ) angles in the region of $(-135^\circ,135^\circ)$. Where two adjacent β-strands hydrogen bond to one another, they form a β-sheet. The strands can run parallel (Figure 1.8a) or anti-parallel (Figure 1.8a). $18-25\%$ of protein structures are made up of β-strands (Martin *et al.*, 2005). The amino acid side chains alternate pointing up and down along the β-strand. A single pair of hydrogen bonds between two residues is called a β-bridge (Figure 1.9).

1.3.1.6 Tertiary structure

The three dimensional structure of a protein chain is known as its tertiary structure (for an example, see Figure 1.10a). Most proteins have a well defined three dimensional structure, although 20% of eukaryotic proteins and 4% of archaebacterial and prokaryotic proteins contain disordered regions of more than 80 residues (Dunker *et al.*, 2000; Liu *et al.*, 2002; Schlessinger *et al.*, 2011). The tertiary structure is crucial for the function of proteins. It is responsible for creating active sites that allow the proteins to interact with small biological molecules (Kessel

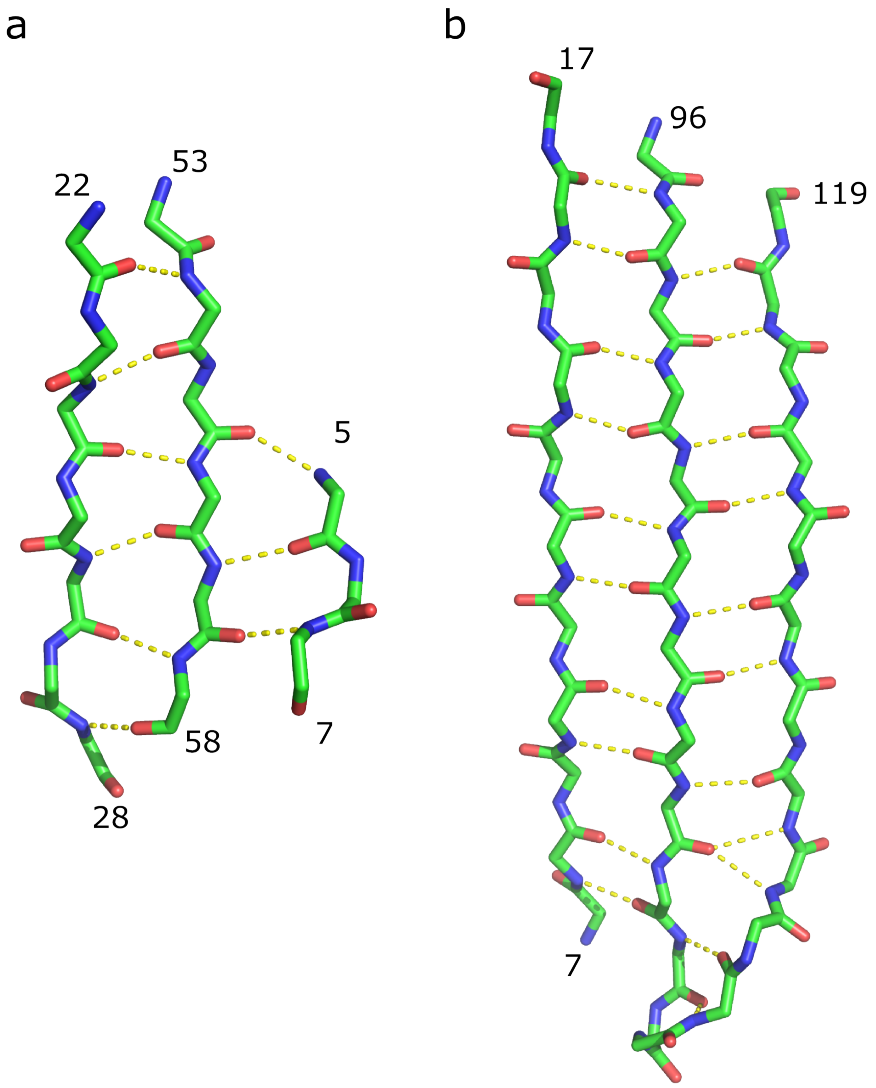


Figure 1.8: **β-sheet example structures.** (a) Parallel β-sheet, from Thioredoxin protein (PDB code 2TRX, chain A). (b) Anti-parallel β-sheet, from zinc α_2-glycoprotein (3ES6, chain A). The numbers in the figure are the PDB residue numbers of the first and last residue in each strand. The hydrogen bonds that stabilise the structure are shown as yellow dashed lines.

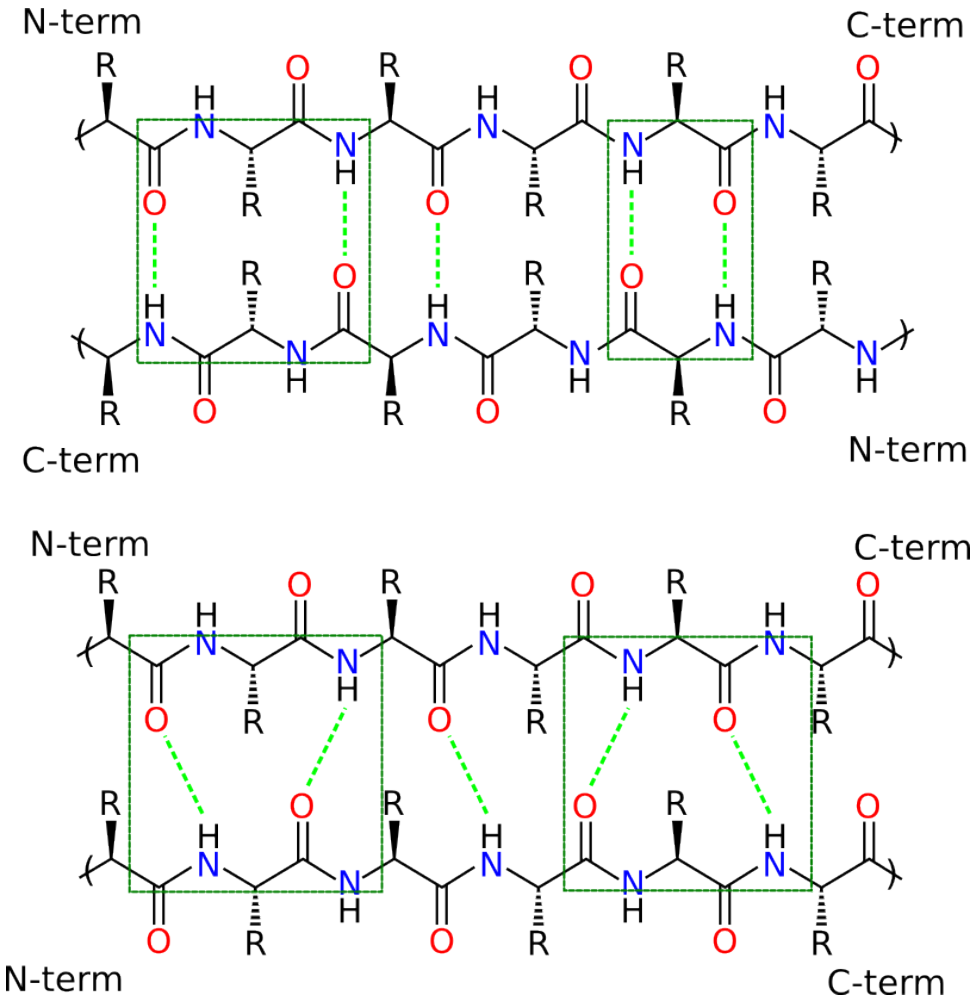


Figure 1.9: **Schematic diagrams of antiparallel (top) and parallel (bottom) β-sheets.** Hydrogen bonds that make up a β-bridge are outlined in green

& Ben-Tal, 2011). Many functions involve conformational change, where signals are sent across the protein with large scale movements.

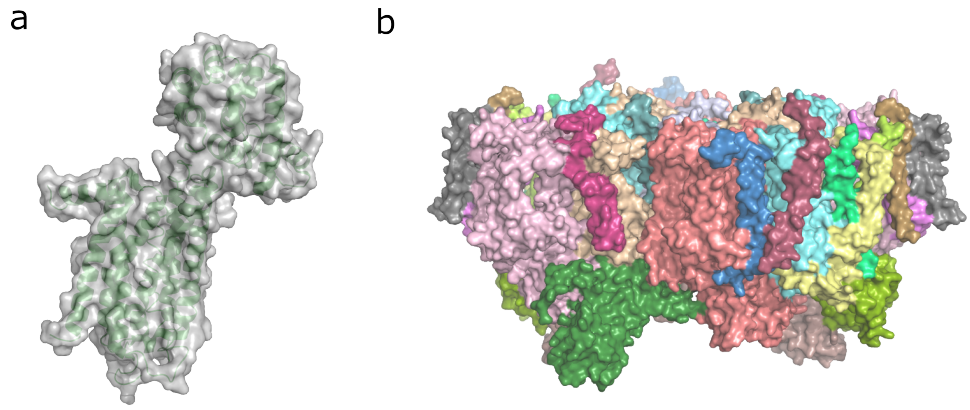


Figure 1.10: **Tertiary and quaternary protein structure** (a) The tertiary structure of protein with PDB code 3eml. The surface is shown in grey, with cartoon secondary structure in green. (b) Quaternary structure of photosystem II (PDB code 1s5l). The surface of each chain is shown, with each chain coloured differently.

Proteins can be grouped into a hierarchy by their similarity, into classes, folds, superfamilies, families, and domains, e.g. the SCOP database (Murzin *et al.*, 1995). The class of a protein is defined by the types of secondary structure it contains. A fold is defined by the topology of its secondary structure elements. A superfamily contains proteins that are thought to have a similar function and structure, indicating a probable common evolutionary origin. Family and domain level classification are based on increasing levels of sequence similarity. The CATH (Sillitoe *et al.*, 2013) database provides a similar protein classification.

Generally proteins with more than 30% of their amino acids in common have the same structure (and so are in the same family), but proteins can be grouped based on their structure even when they share no common sequence elements.

Many proteins are complexes made up of a number of protein chains in contact with one another. The relative positions of the different chains is called the quaternary structure.

1.3.1.7 Quaternary structure

Protein quaternary structure describes how multiple protein chains combine together to form a single complex (Klotz *et al.*, 1970). Many protein functions rely on the interaction of many protein chains, such as transport proteins (Veenhoff *et al.*, 2002), and the light-harvesting system in plants (Rochaix, 2014). Photosystem II, one of the two complexes in the light-harvesting system, is made up of more than twenty protein chains that fit together to form one large membrane bound complex, as shown in Figure 1.10b (Hankamer *et al.*, 1997).

1.3.2 Proteins in membranes

Many of the proteins in the body reside in the cell membrane. The membrane environment is very different to the aqueous environment within the cell.

1.3.2.1 Biological membranes

Cells are surrounded by a membrane made up of lipids, a physical barrier between the inside (cytoplasm) and outside (exoplasm) of the cell. It provides a non-polar barrier that is all but impermeable to the important, water soluble, molecules in the cell, such as ions, small molecules, RNA, and proteins (Kessel & Ben-Tal, 2011). Similarly cell organelles, such as the mitochondria, the endoplastic reticulum, the chloroplasts, and the nucleus are enclosed by membranes.

These biological membranes consist of a lipid bilayer (Figure 1.11a). Amphipathic glycero-phospolipids, with polar heads and non polar tails, are the most common type of molecule found in membranes (Figure 1.11b). Many other molecules, such as sphingolipids, sterols, and ethers, can also make up membranes (Kessel & Ben-Tal, 2011). Although glycero-phospholipids make up the majority of membranes, the concentrations of these differ between different membranes. This includes between different organisms (e.g. the most numerous phospholipid in eukaryotes is phosphatidyl-choline (Morein *et al.*, 1996), while in bacteria phosphatidyl-ethanolamine and phosphatidyl-glycerol are most numerous), and different tissues within the same organism. Similarly, the inner membranes in eukaryotes (such as the mitochondrial and endoplasmic reticulum

membranes) have a different composition than the plasma (cell) membrane (Keenan & Morre, 1970; Yeagle, 2004).

The composition of a single membrane is not uniform. The two leaves of the membrane may have different compositions - in particular, the cytoplastic[1] leaf contains many negatively charged molecules (Bergelson & Barsukov, 1977; Op den Kamp, 1979), and so is negatively charged with respect to the exoplastic[2] leaf. The membrane bound flippase and floppase proteins maintain this asymmetry by transferring lipids from one leaf to the other, in an energy dependent process (Daleke, 2003; Holthuis & Levine, 2005).

Recent work has shown evidence for short range order in the membrane. 'Lipid rafts' are areas dense with proteins, which are responsible for many cell functions (Lingwood & Simons, 2010). The membrane within these lipid rafts is thicker than the surrounding membrane. This can occur because the lipid tails have two possible phases of matter - liquid ordered and liquid disordered. Specific proteins have preferences to be in one or other of these phases, leading to the partitioning of these proteins in the membrane (Schroeder *et al.*, 1998; Wang *et al.*, 2001). Further, proteins can deform the shape of the membrane, proteins are known to both alter the thickness (Mitra *et al.*, 2004) and promote membrane curving (Siegel *et al.*, 2006).

Despite the variation in the membrane constituents, all of the membranes share the common characteristic of a hydrophobic core sandwiched between two hydrophilic layers. The exact properties, including the width and chemical make up, differ between organisms, cells, and even regions within the same membrane. The average width of a membrane is $\sim 50-60$Å, with a ~ 30Å core (tail layer) and two $\sim 10-15$Å polar lipid-water interfaces (head layer) (Kessel & Ben-Tal, 2011).

1.3.2.2 Membrane proteins

The proteins that reside in the membrane are different to the soluble proteins in the rest of the cell. Membrane proteins are very important for biological processes in the body. It is estimated

[1]The membrane leaf on the inside, in contact with the cell cytoplasm
[2]The leaf on the outside of the cell, in contact with the extracellular matrix

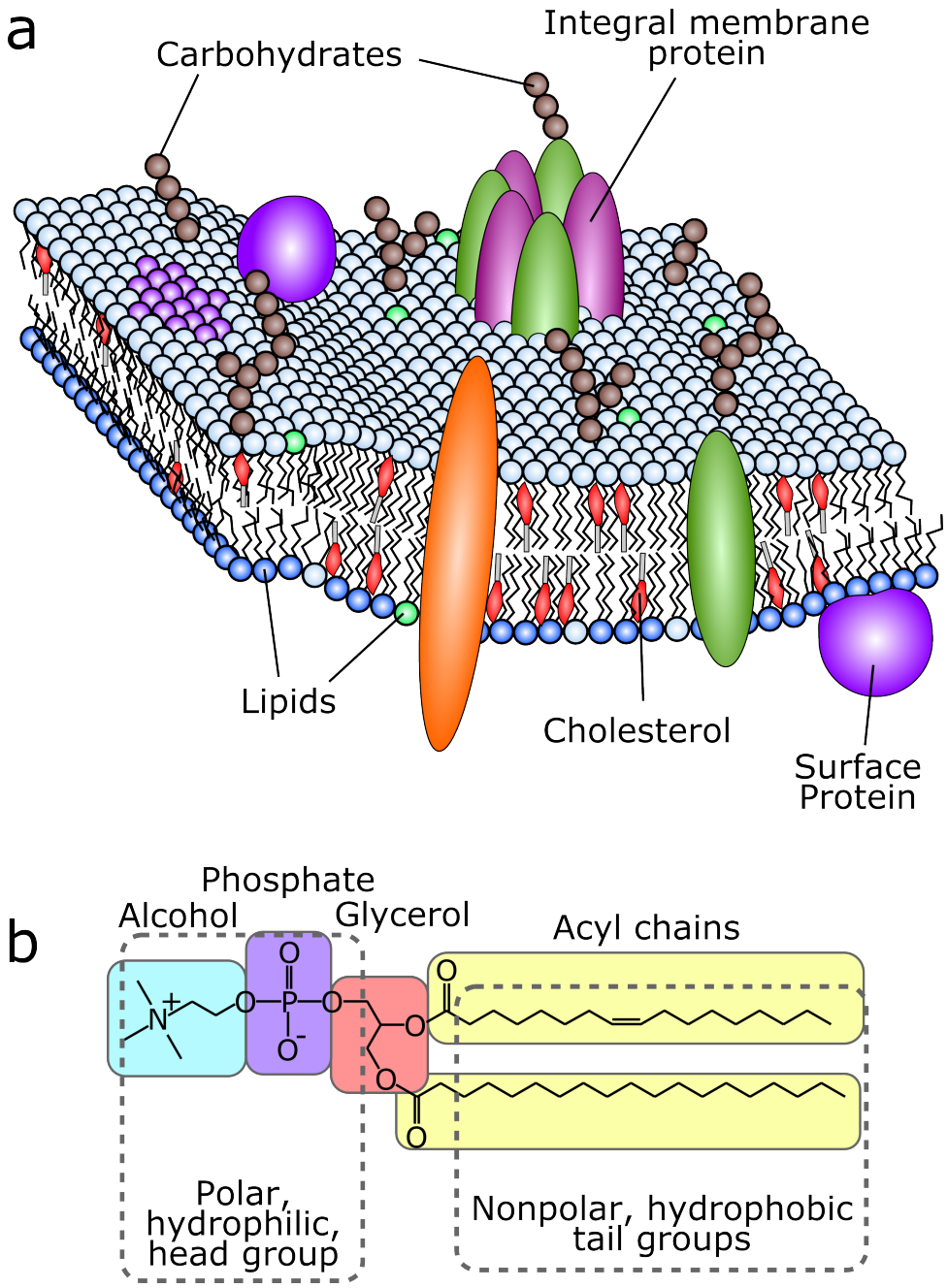


Figure 1.11: **The cell membrane.** (a) Representation of a cell membrane. Integral membrane proteins span the membrane, while membrane surface, or peripheral, proteins are in contact with the membrane. (b) The chemical structure of a phosphatidylcholine, an example phospolipid. Different phospholipids have different alcohol and alkyl groups, but the general structure of a polar head and nonpolar tail is the same.

that 30% of the human genome codes for membrane proteins (Almén *et al.*, 2009). Membrane proteins make up 50% of drug targets (Drews, 2000), with Rhodopsin-like GPCRs alone making up 27% of small molecule drug targets Overington *et al.* (2006).

Proteins that are in contact with the membrane can be split into two types:

- Integral membrane proteins, that span the membrane, and sit inside it.
- Peripheral or surface proteins, which are attached loosely to one or other side of the membrane

Peripheral membrane proteins have many similarities to other soluble proteins, however integral membrane proteins are quite different, as they are exposed to the hydrophobic core of the membrane. There are two dominant structure types of integral membrane proteins, α-helical and β-barrels (Figure 1.12). α-helical proteins have between one and 14 helices that cross the membrane (transmembrane helices), with one, two, and seven transmembrane helix proteins being the most common (Almén *et al.*, 2009). β-barrels proteins are less common, and occur primarily in the outer membrane of Gram-negative bacteria.

It is energetically unfavourable to have polar carbonyl and amide groups with unsatisfied hydrogen bonds in contact with the hydrophobic core of the membrane. α-helices allow the backbone carbonyl groups to satisfy all hydrogen bonds, minimising these unfavourable interactions, as do β-sheets. Thus, the secondary structure elements (α-helices or β-strands) tend to span the entire membrane, creating long elements, whereas in soluble proteins a more globular structure with shorter secondary structure elements is generally observed.

One difference between membrane and soluble proteins is in the core (inside) of the proteins. Studies indicate that the core of membrane proteins is similar to that of soluble proteins (Gimpelev *et al.*, 2004; Hildebrand *et al.*, 2004; Rees *et al.*, 1989), but earlier studies suggest that membrane helices are more tightly packed than soluble helices (Eilers *et al.*, 2000). Interfaces between helices in membrane proteins are also different from those in soluble proteins. They contain more small residues, as opposed to soluble proteins (Eilers *et al.*, 2002). For example,

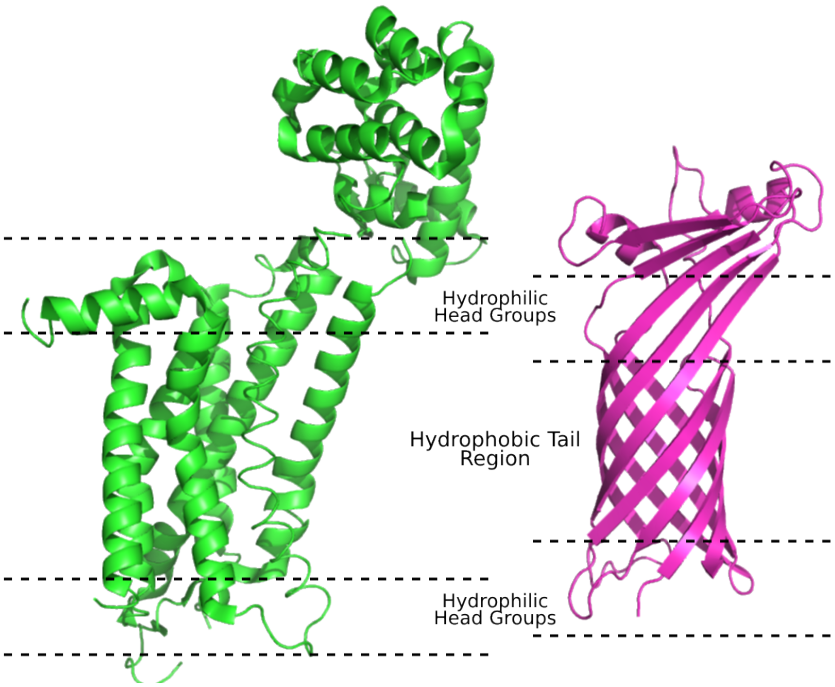


Figure 1.12: **Examples of α-helical bundle and β-barrel membrane proteins.**

the GxxxG motif (where G is glycine, serine, alanine, or threonine) is common in membrane helices, but not frequently observed in soluble protein helices (Eilers *et al.*, 2002).

There has been much debate about the strengths of hydrogen bonds in membrane proteins. Most agree that the strength of classical hydrogen bonds is less in membrane proteins than in soluble proteins (Faham *et al.*, 2004; Joh *et al.*, 2008). NMR studies have shown that carbonyl groups in membrane helices are aligned differently to those in soluble helices, with potential effects on hydrogen bonding, both intra- and inter- helical (Page *et al.*, 2008). A number of studies have concentrated on the Cα–H - - - O bond, with contrasting views of their importance in stabilising membrane protein structure (Bowie, 2011; Yohannan *et al.*, 2004a). There is no definitive evidence demonstrating the difference (or similarity) of the hydrogen-bonding patterns in the core of membrane and soluble proteins. Experimental studies are difficult and consequently limited in scope. This makes is difficult to discern which effects are due to the different secondary structure makeup of membrane and soluble proteins, and which are due to intrinsic differences between the two types of protein. During synthesis, α-helical membrane proteins are extruded into a membrane one helix at a time, before the helices assemble to create the final structure. It is plausible that the hydrogen bonds that provide helical stability throughout this process are different to those found in soluble proteins, but there is no definitive evidence to support this hypothesis.

Residue conservation patterns in membrane proteins are different from those in soluble proteins, with hydrophobic residues being less well conserved in the parts of membrane proteins that are exposed to the centre of the membrane than they are in soluble proteins, whilst hydrophilic residues are generally much better conserved in membrane proteins than soluble proteins (Mokrab *et al.*, 2010). There are fewer hydrophilic residues in membrane proteins, but these are often either structurally or functionally important - e.g. in the lining of the pore in ion channels. Hence in membrane proteins a higher proportion of the hydrophilic residues are important than in soluble proteins. Since residues that are functionally or structurally important are typically more conserved than those that are less important, hydrophilic residues are

generally better conserved in membrane proteins than soluble proteins. Glycine and proline are highly conserved in the middle of the membrane (Mokrab *et al.*, 2010), as are coil residues[1] (Kauko *et al.*, 2008).

The helices within membrane proteins are significantly different from those in soluble proteins. Membrane helices are typically far longer, which is a result of the membrane environment.

1.3.3 Protein structure determination

The three-dimensional structure of proteins defines how it interacts with molecules, natural or otherwise, in the body. There is no direct way in which to observe the structure of proteins. X-ray diffraction is the most popular method used to elucidate protein structures. Elucidated protein structures can be found in the Protein Data Bank (PDB) (Berman *et al.*, 2007), in which, as of August 2014, there are 102,550 structures.

1.3.3.1 X-ray crystallography

X-ray crystallography uses the principle of X-ray diffraction. In a regular solid with a repeating three-dimensional pattern, there are many sets of regularly spaced planes of atoms, which scatter X-rays. Proteins can be crystallised such that they form a regular repeating array, where each protein is in exactly the same orientation.

For data collection, the protein is placed in a goniometer, and irradiated with an X-ray beam. A detector is used to collect the diffracted X-rays, which yields a pattern of dots. The X-rays are diffracted by the electrons within the protein, so the pattern of dots is related to the protein structure. Specifically, the diffraction pattern is a Fourier transform of the electron density.

However, when the Fourier transform is inverted, not all of the information about the electron density is recovered. Because the detector can only detect the amplitude (which provides information on the size of the atom which has diffracted the X-rays), and not the phase (which

[1] Coil residues are those that do not adopt a secondary structure type, i.e. they are not in an α-helix or β-sheet.

provides the information about the position of the atom that diffracted the X-rays), the diffraction pattern provides insufficient data to allow the protein structure to be determined. A common solution to this is molecular replacement, where an initial estimate of some or all of the structure is used in conjunction with the diffraction data to estimate the phases. These phase estimates are used to refine the initial structure estimate, and the process is repeated until the agreement is satisfactory.

The agreement with the diffraction data is often measured by the R factor, and the free R factor. The R factor is a measure of how the measured diffraction pattern differs from the ideal pattern based on the proposed structure. Values in the range of $0.15 - 0.25$ are considered satisfactory (Kessel & Ben-Tal, 2011). The R factor is liable to overfitting, as it is used in the refinement as well as a measure of quality (Brändén & Alwyn Jones, 1990). The free R value uses a cross-validation method to provide a less biased measure of quality, where the quality is measured using experimental data that was not used in the refinement steps (Brünger, 1997). The value of R free is typically slightly larger than that of the R factor, with values in the range of $0.20 - 0.30$ considered satisfactory. Structures are also refined by considering features such as the bond lengths, the dihedral angles, and the packing of the protein. These are compared to both empirical values, and the values found in already known protein structures (Laskowski *et al.*, 1993; Read *et al.*, 2011).

The smallest separation of atom planes that can be identified in a structure is known as the resolution. This typically varies in the range of 5Å, where secondary structure elements can be discerned, to 0.8Å, where the hydrogen atoms can be resolved. Amino acid side chains can be resolved at resolutions of 2.5Å or better (Kessel & Ben-Tal, 2011).

Limitations A crystal is not the native environment of the cell. X-ray structures are of static proteins, and where parts of the protein are not static, the structure cannot be resolved. Further, the proteins are densely packed, which can influence the structure of the protein. That said, the average solvent content of a protein crystal is 48% (Lee & Kim, 2009), in comparison to a

cell, which has a water content of around 72%[1] (Savitz *et al.*, 1964).

The structure is normally calculated iteratively, and rarely has exact agreement with the experimental data. The resolution of the eventual structure is generally insufficient to identify the position of hydrogen atoms. A major stumbling block is the crystallisation of proteins. The conditions necessary for crystallisation are often strict, and change from protein to protein. This is particularly the case for membrane proteins, as their natural environment is lipids, which do not readily crystallise. There are a number of approaches to crystallising integral membrane proteins which are being developed, and the number of available structures is increasing exponentially (Bill *et al.*, 2011; White & Wimley, 1999). These approaches include co-crystalisation with antibodies (Hunte & Michel, 2002), use of detergents, and crystallisation in protein lattices (Sinclair & Noble, 2004). Membrane protein structures are typically of lower resolution than soluble protein structures, with resolution in the region of 2.8—4.0Å.

1.3.3.2 Nuclear magnetic resonance

Nuclear magnetic resonance, or NMR, is a method that uses the nuclear spin of an atom to infer information about its surroundings. Nuclear spin arises from the motion of charges in the nuclei. Only nuclei with odd numbers of protons and/or neutrons (such as ^{1}H, ^{13}C, ^{15}N), have spin. When nuclei with non-zero spin are exposed to a magnetic field, their possible spin states have different energies. If the nuclei are irradiated with photons that have the same energy as the difference between these states, they can absorb the energy to change spin states. This absorption, or the subsequent emission resulting from the relaxation can be measured.

The magnitude of the energy difference depends on the local magnetic field, which is affected by the local electron density, and by nearby nuclei with spin. The rate of relaxation after irradiation is affected by the states of nearby nuclei. These two factors are used to determine constraints within the protein structure, by a variety of techniques. These constraints are then used to deduce the protein structure.

[1] This value is for a red blood cell

NMR structures are reported in the form of an ensemble of models, which provides dynamic information about the protein (Baker & Baldus, 2014). Protein NMR can be done in both the solution and solid states. The solid state is of particular interest to membrane protein structure determination (Middleton, 2007).

Limitations As the structure is built from a series of short range constraints, and because NMR spectra become increasingly complex as the number of atoms within a system increases, this method is limited to small proteins (typically smaller than 35 kDa) (Maslennikov & Choe, 2013). While NMR can be done in the solution state, allowing dynamic information about the protein to be found, the protein is at a much higher concentration than in the cell. Like for X-ray crystallography, that means that the protein is not necessarily in its native conformation when its structure is being determined.

All protein structure determination methods are time consuming, and whilst the techniques are improving rapidly, they are not able to keep pace with the increase in protein sequences that are being identified.

1.3.4 Sequence structure gap

Although there are currently over 100,000 publicly available protein structures in the Protein Data Bank (PDB) (Berman *et al.*, 2007), there are many more known protein sequences. The UniProt database contains over 80 million sequences[1] (The UniProt Consortium, 2013). Despite the number of known structures increasing exponentially, the gap between known structures and sequences is increasing rapidly, with nearly 11 million new UniProt sequences during June and July of 2014, compared to 10,000 new structures deposited in the PDB in all of 2013.

The scarcity of protein structures is most apparent for membrane proteins. As discussed above, determining the structure of membrane proteins is difficult. Although membrane proteins make up *c.* $26-30\%$ of the human proteome (Almén *et al.*, 2009; Fagerberg *et al.*, 2010), of the

[1] as of July 2014

103,015 structures in the PDB (Berman *et al.*, 2007), only 1,521 are of membrane proteins, and just 93 are of human membrane proteins[1] (White & Wimley, 1999).

The relationship between protein sequence and protein structure can be harnessed to predict computationally the structure of proteins without experimentally determined structures. Proteins with similar sequences typically have similar structures. Where a structure exists with a similar amino acid sequence (a homologue) to that of a protein of interest, the structure of the protein of interest can be modelled using the homologous protein structure as a starting point. MODELLER (Sali & Blundell, 1993) is just one example of a homology modelling program, but there are many more. Homology modelling methods work well when the homologue used as a starting point has $\geq 60\%$ sequence identity with the modelled protein (typically giving a model correct to within an RMSD of ≤ 2Å). It is less reliable when the sequence identity of the homologue and modelled protein is $\leq 60\%$, and models are very unreliable when the sequence identity is $\leq 25\%$. Due to the limited number of membrane protein structures, and their relative importance, the modelling of membrane proteins is particularly important. MEDELLER (Kelm *et al.*, 2010) is a membrane specific tool, which outperforms MODELLER on membrane protein structure prediction.

Where no homologous protein structures exist, alternative fragment based approaches can be used, e.g. Rosetta (Simons *et al.*, 1997). Rather than needing a whole homologous protein, these methods find short fragments (three to nine amino acids) that have homologous sequences. The protein structure can be predicted by building up many combinations of fragments (decoys).

1.4 Helices

Kinks, the focus of this thesis, are a feature of protein helices. This section discusses helices in more detail.

[1] as of 08/09/2014

1.4.1 Helix geometry

α-helices make up c. $32-42\%$ of protein structures, and make up a far higher proportion of integral membrane proteins. However, it is only very recently that structure prediction methods have considered possible helix deformations in structure prediction (Chen *et al.*, 2014).

Helices rarely conform exactly to the ideal structure suggested by Pauling *et al.* (1951), indeed they differ quite considerably, with (ϕ,ψ) being in the range of (-63,-40) compared to (-57,-45) for the predicted structures (Table 1.1). Most protein helices curve gently away from the solvent (Blundell *et al.*, 1983).

Even allowing for a range of structures within the definition of α-helix, there are further distortions to helices, which are often identified as important functional sites. There are a number of common distortions to α-helices. First, short sections of π- and 3_{10}- helix are incorporated into helices. These are known as π- and 3_{10}- turns, and tend to occur at the end of helices (particularly π-turns at the C-termini of helices) (Dasgupta & Chakrabarti, 2008). They also occur in the middle of helices, and there is some suggestion that they are stabilised by hydrogen bonds to side-chains and water molecules (Cartailler & Luecke, 2004)

The hydrogen bond patterns within helices are often disrupted, with 10% of backbone hydrogen bonds missing relative to an ideal helix. The backbone atoms are often involved in three centre (bifurcated) hydrogen bonds.

1.4.2 Helix and secondary structure assignment

Although structures submitted to the PDB include secondary structure annotation, it is useful to have a fast, unbiased, computational method that annotates secondary structure. Annotation involves identifying the start and end point of helical structures (and other types of secondary structure) in proteins. Levitt & Greer (1977) provided the first automatic method for secondary structure annotation. In this thesis we primarily use DSSP (Joosten *et al.*, 2011; Kabsch & Sander, 1983).

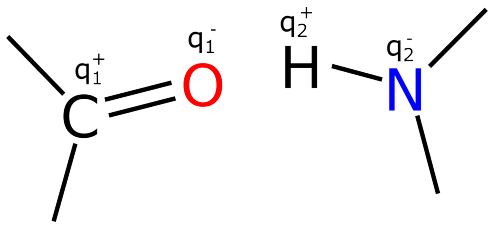


Figure 1.13: **Hydrogen bond energy calculation in DSSP.** Each of the four atoms is modelled with a partial point charge. The hydrogen bond energy is calculated as the sum of the the electrostatic interaction between the four atoms (Equation 1.1).

1.4.2.1 DSSP

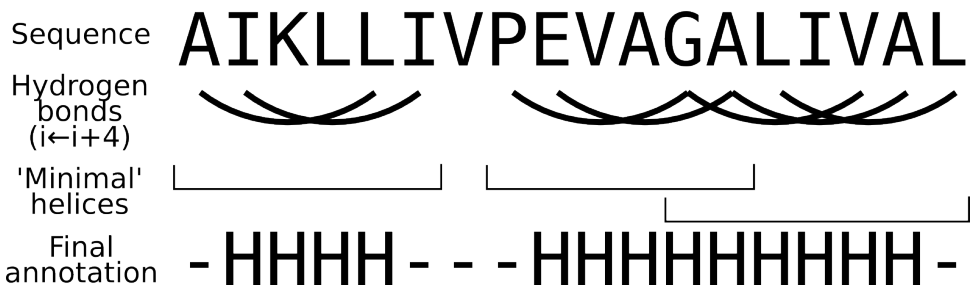


Figure 1.14: **Example DSSP helix annotation.** Sections where with two or more $i \leftarrow i+4$ main chain hydrogen bonds are identified as 'minimal' helices. Where these overlap by two or three residues, they are combined into a single helix, to give the final annotation. The first and last residues in the minimal helices are not given the helical annotation.

DSSP (Define Secondary Structure of Proteins) identifies protein secondary structure based on hydrogen bonding patterns. It uses a simple electrostatic model of bonding (Figure 1.13) to evaluate if a hydrogen bond exists. The energy, E, in kcal/mole, of the hydrogen bond is shown in Equation 1.1:

$$E = q_1 q_2 \left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right) f \tag{1.1}$$

Where q_1 and q_2 are the two partial charges over the CO and NH bonds (Figure 1.13), which are $0.42e$ and $0.20e$ (where e is the unit electron charge). r_{NO} is the distance between the N

and O atoms, in Å. f is a factor with units $\frac{\mathring{A}kcal}{e^2mol}$. In the original work, it has a value of 332, but in subsequent implementations it has been varied to give the best agreement with training sets. A hydrogen bond is considered present when $E < -0.5$ kcal/mole (compared to approximately -3 kcal/mol for a strong hydrogen bond) (Kabsch & Sander, 1983). This means that frequently a single atom can be involved in multiple ('bifurcated') hydrogen bonds.

Elementary patterns - bridges, turns and bends - of backbone hydrogen bonds within a protein structure are then identified. The patterns for beta bridges are shown in Figure 1.9, while $n-$turns are identified at single instances of $i \leftarrow i+n$[1] $(n = 3, 4, 5)$ hydrogen bonds, as shown in Figure 1.7.

Adjacent β-bridges are combined to form β-sheets. Adjacent turns are combined into helices of the relevant type. Where these overlap, by two or three residues, they are combined to form a single helix (Figure 1.14).

There are a total of eight different possible annotations. G, H, and I for the three helices (3_{10}, α, and π respectively), E for an extended strand, B for an isolated β-bridge, T for an isolated turn, S for a residue in a region of high curvature, and loop (–) for everything else. Where there are multiple assignments for the same residue, the preference is for H (α-helix) over B and E (β-bridge and β-sheet) over G (3_{10} helix) over I (π-helix). This, combined with the relatively relaxed definition of the hydrogen bonds, has lead to a criticism that DSSP under annotates π-helices (Zacharias & Knapp, 2014).

DSSP is often found inside analysis software, such as Promotif (Hutchinson & Thornton, 1996), and JOY (Mizuguchi *et al.*, 1998). This, along with its simplicity (thereby allowing reimplementation), has resulted in DSSP being the most popular secondary structure assignment method. However, there are many other available secondary structure assignment methods, that use more information to annotate the structure.

[1]The arrow points from the hydrogen bond donor (typically NH) to the hydrogen bond acceptor (typically CO)

1.4.2.2 Other secondary structure assignment methods

Figure 1.15: **Examples of different secondary structure assignments for protein 5ULL**. SST, STRIDE, DSSP, PDB Author annotation, KAKSI, Palasse, P-Sea, Sticks, Xtlsstr. See text for citations. The nomenclature has been changed to be consistent across all methods, so that: H = α-helix, E = β-sheet, T = turn, B = isolated β-bridge, G = 3_{10} helix, I = π-helix, 5 - 5-turn, 4 - 4-turn, S - Bend, P - poly(L-proline) type 3_1-helix. Some methods use lower case letters at the start and end of secondary structure elements.

DSSP utilises the regular hydrogen bond pattern of secondary structures to identify them. There are many other methods that identify secondary structure, which use both improvements to DSSP's hydrogen bond method, and a number of other repeating features of secondary structure:

- STRIDE (Frishman & Argos, 1995)
- Manual annotation, from the depositor of the structure in the PDB.
- DEFINE-S (Richards & Kundrot, 1988)
- KAKSI (Martin *et al.*, 2005)
- SST (Konagurthu *et al.*, 2012)

- PSSC (Zacharias & Knapp, 2014)
- P-SEA (Labesse *et al.*, 1997)
- P-Curve (Sklenar *et al.*, 1989)
- Xtlsstr (King & Johnson, 1999)
- Stick (Taylor, 2001)

There are a number of differences in the annotation provided by each method. Figure 1.15 shows the annotation for a protein provided by a number of different methods. The notation has been standardised to that of DSSP (H = helix, E = strand, – = none). It is common for them to disagree on the start and end of a helix by a few residues.

On X-ray crystal structures, when the methods are reduced to three states (H, E, and –), pairwise agreements between any two methods are in the range of 95%-75% (Cuff & Barton, 1999; Konagurthu *et al.*, 2012; Martin *et al.*, 2005). However, Konagurthu *et al.* (2012) reported much lower agreement in NMR structures, with an agreement of 69% between DSSP and STRIDE for NMR structures.

The structure of helices is most distorted towards the ends. This is where the majority of annotation disagreements occur (Konagurthu *et al.*, 2012). Indeed, some methods use more relaxed thresholds for helix inclusion at the termini of helices (such as the Cα-Cα distance criteria in KAKSI (Martin *et al.*, 2005). However, methods also disagree on the annotation of helix distortions, with some methods identifying long helices including a distortion, and other methods identifying two separate helices with a distortion between (Martin *et al.*, 2005). KAKSI is an example of a method that explicitly excludes kinks from helices (Martin *et al.*, 2005).

1.4.3 Characteristics of α-helices

The methods described above provide a way to identify the α-helices within known protein structures. Using these methods, the general characteristics of α-helices have been identified.

These characteristics of α-helices distinguish them from other parts of the protein.

1.4.3.1 Amphipathic helices

Soluble protein helices are typically amphipathic, that is, with polar and non-polar faces (Segrest *et al.*, 1990). This phenomenon arises because helices are typically exposed to the polar solvent on one side, and to the non-polar core of the protein on the other side. Water molecules disrupt the hydrogen bonding patterns on the face of the helix exposed to the solvent, which results in longer hydrogen bonds on this side, and the helix curving away from the solvent (Blundell *et al.*, 1983).

1.4.3.2 Residue types

Some amino acids are more frequently found in α-helices than others (Pace & Scholtz, 1998). Typically amino acids with β-branched sidechains are disfavoured, as there is not room for these side chains to pack around the helix. Leucine, arginine, methionine, and lysine all have high helix propensities, while threonine, cysteine and aspartic acid have low helix propensities (i.e. are less favoured in helices). The hydrophobic effect is a major driver for the propensities, specifically residues that can bury more of their side chains against the helix favour the α-helix conformation more.

Glycine has a low helix propensity (Muñoz & Serrano, 1994; Pace & Scholtz, 1998), because of its small size, and the greater entropic cost of restricting its movements by incorporating it into the helix. Its small size makes it difficult for other parts of the protein to pack up against it. Proline has no amide hydrogen, precluding it from forming backbone hydrogen bonds, the major stabilising force for helix formation. In addition, the ring in proline cannot be easily accommodated within the helix structure. For these reasons, proline and glycine often act as helix terminators in proteins, and are rarely found in the middle of α-helices

Helices have a tendency to start and end on the face of the helix in contact with the rest of the protein. Consequently, hydrophobic residues are most favoured at the start and end

of helices, and so the amino acid propensities vary periodically with the position in the helix (Engel & DeGrado, 2004).

1.4.3.3 Helix caps

The ends of soluble helices are often well conserved motifs - called 'caps', which have common hydrophobicity and hydrogen bond patterns (Aurora & Rose, 1998). These helix caps are stabilised by backbone to backbone (*c.* $^2/_3$ of caps) and backbone to side chain (*c.* $^1/_3$ of caps) hydrogen bonds. N-caps frequently have a polar residue as the second residue in the helix, and C-caps normally have a hydrophobic residue four from the end, and a polar residue two from the end.

1.4.3.4 Differences between membrane and soluble α-helices

Helices in integral membrane proteins that cross the membrane (transmembrane helices) are unlike soluble protein helices, as they are exposed to a hydrophobic environment, and are typically longer. Soluble helices are generally much shorter than transmembrane helices, with an average of around 15 residues (Baeza-Delgado *et al.*, 2013; Engel & DeGrado, 2004; Pal *et al.*, 2003), compared to 25 for transmembrane helices (Baeza-Delgado *et al.*, 2013; Ulmschneider & Sansom, 2001). The propensities of amino acids for transmembrane helices are different, due to the environment in which they sit (Jha *et al.*, 2011; Ulmschneider & Sansom, 2001).

Charged residues are much less frequently found in membrane helices, while non-polar and aromatic residues are more frequently found. Glycine is also much more common in transmembrane helices (8.7%) than soluble helices (4.3%) (Baeza-Delgado *et al.*, 2013), as is proline. The amino acid propensities vary with position in the membrane (von Heijne, 2006). Charged residues are disfavoured in the centre of the membrane, but favoured at the interface, while hydrophobic residues are favoured in the centre, but disfavoured on the outside of the membrane (Baeza-Delgado *et al.*, 2013). Charged residues at the interface between the non-polar core of the membrane and the polar head groups often have their side chains pointing into the head

groups, a feature known as 'snorkelling' (Chamberlain *et al.*, 2004). The membrane environment promotes the formation of helices, meaning that sections of protein that would be disordered in solution form helices inside the membrane (Leman *et al.*, 2013).

There are suggestions that ϕ and ψ angles have slightly different distributions for membrane and soluble helices (Hildebrand *et al.*, 2004; Page *et al.*, 2008), and that $i \leftarrow i+3$ hydrogen bonds are more prevalent in membrane helices (Bowie, 2011; Cao & Bowie, 2012). This was suggested by Bowie (2011) as a mechanism to allow transmembrane helices to be more flexible than soluble helices. In contrast, in soluble and membrane protein helices, the proportion of hydrogen bond donor and acceptor groups that are unsatisfied is similar (Joh *et al.*, 2008).

Work by Wong *et al.* (2012) has suggested that some membrane helices act simply as anchors to the membrane, whilst others play a role in structure and function. These can be classed as 'simple' and 'complex' membrane helices. Simple helices have generic, lipophilic, sequences which are similar to other simple helices, regardless of their evolutionary relationship. This highlights a possible difficulty in identifying homologous protein sequences, as proteins with many simple helices may be very similar to other proteins, despite not being evolutionarily related.

Helix structure is affected by the rest of the protein, as well as its own sequence and the membrane environment. The packing of helix pairs falls into a small number of clusters (Walters & DeGrado, 2006), and this packing is largely stabilised by small residue side chains in the interface (with small residues every 4 or 7 residues along each helix). The packing is mostly down to apolar side chains, but polar interactions do have an effect in some proteins (DeGrado *et al.*, 2003). Soluble proteins are held together by virtue of the hydrophobic effect - however this is less the case with membrane proteins, as they are exposed to the hydrophobic core of the membrane.

One particular common feature of transmembrane helices is kinks (Barlow & Thornton, 1988). These are rare in soluble proteins (Nugent & Jones, 2011). Kinks are important for function in a variety of important proteins.

1.5 Kinks in α-helices

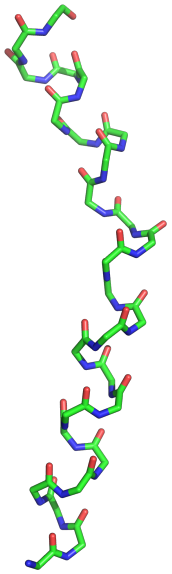


Figure 1.16: **A kinked helix.** Taken from the transmembrane domain of a human A2A adenosine receptor (pdb code: 3eml, chain A). Showing only backbone atoms; Carbon green, Nitrogen blue, and Oxygen red.

Kinks are a type of distortion of α-helices, where there is a sharp change in the helix axis (see Figure 1.16 for an example) (Kneissl *et al.*, 2011). Although there have been many studies into kinks, and other helix distortions, there is no universally accepted definition of helix distortions (Bansal *et al.*, 2000; Hall *et al.*, 2009; Kauko *et al.*, 2008; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Rigoutsos *et al.*, 2003; Visiers *et al.*, 2000; Werner & Church, 2013). They are however, often identified as important in the function of proteins.

1.5.1 Why are kinks important?

Kinks are a structural and functional feature of membrane proteins.

There are many studies that have identified important functional kinks in membrane proteins. Katritch *et al.* (2009) indicated that a flexible kink in the fifth transmembrane helix (TM5) of GPCRs is crucial to the activation mechanism of GPCRs. Highly conserved pro-

line residues cause hinges in the sixth and seventh transmembrane helices (TM6 and TM7) of rhodopsin-like GPCRs that modulate the width of the proteins, and so activate the receptors (Bettinelli *et al.*, 2011; Schwartz *et al.*, 2006; Shi *et al.*, 2002). Three proline residues within transmembrane helices are important to the activation of melatonin receptors (Mazna *et al.*, 2008). Similarly, kink causing prolines are key to Calcitonin binding and signalling (Conner *et al.*, 2005), although it is likely that proline is not necessary for the function of these kinks (Bettinelli *et al.*, 2011). Kinks are not necessarily conserved within the GPCR family - TM1 in CXCR4 and rhodopsin contains a kink, while solved structures of other GPCRs do not (Langelaan *et al.*, 2013).

Kinks are also important in the function of ion channels, where they perform a function in gate-opening (Duclohier *et al.*, 1992; Forrest *et al.*, 1999; Fowler & Sansom, 2013; Johansson & Lindahl, 2006; Kaduk *et al.*, 1997; Mihajlovic & Lazaridis, 2012; Ri *et al.*, 1999; Taylor & Sanders, 1999; Tieleman *et al.*, 1999, 2001; Woolfson *et al.*, 1991).

Kinks in soluble proteins Studies of α-helix kinks have mostly concentrated on membrane proteins. However, there is some work on kinks in soluble protein α-helices (Devillé *et al.*, 2008; Rey *et al.*, 2010), and there are examples of functional kinks in soluble proteins, for example in MHC proteins (de Almeida & Holoshitz, 2011; Rudolph *et al.*, 2006).

1.5.2 How kinks are identified

As with secondary structure assignment, there are many methods available for the computational identification of kinks, however there is no standard method, and most authors have designed their own.

Authors frequently use different terminology, so for clarity I have used consistent terms when describing the methods (but also included the authors' terminology in parentheses at the first mention, if it is different). The methods described below differ in a number of areas, but most in the way they fit helix axes, the length of helix they fit the axes to. Importantly, most of

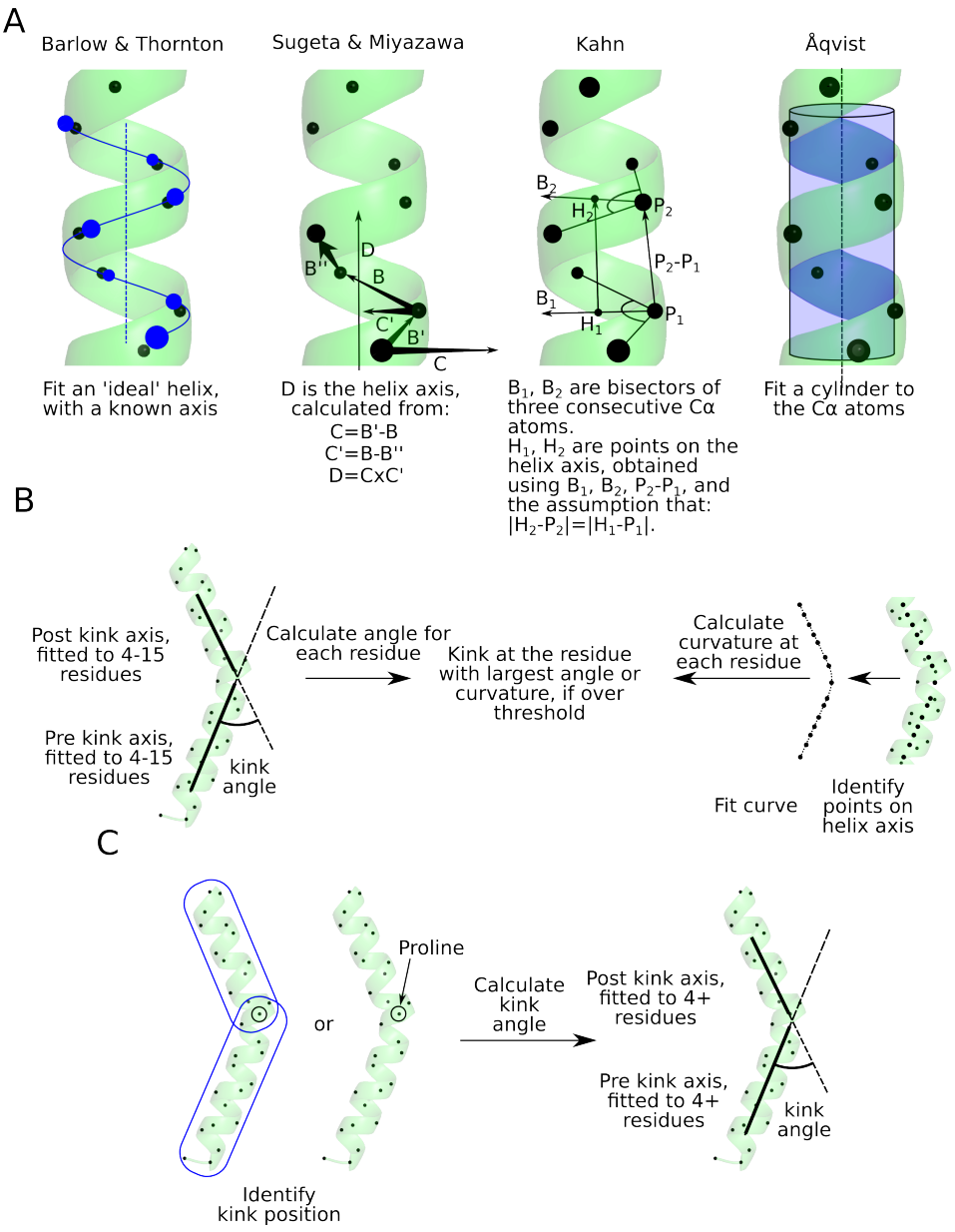


Figure 1.17: **Methods of kink identification.** (a) Four commonly used methods for the calculation of local helix axes and/or points on the helix axis. (b) General approach to using axes or points to identify kinks. (c) Alternative approach of identifying kink position first, before using fitted axes to calculate the kink angle.

them identify the position of the kink (the 'kink residue') based on the residue with the largest angle.

1.5.2.1 Barlow & Thornton (1988)

In this study, ideal 'probe' helices are fitted to a sliding window of the helix of interest (Figure 1.17). The ϕ, ψ, and ω angles for each probe helix are based on the average for the helix of interest. The root mean square deviation (RMSD) of the atoms in the helix to those in the probe helix ($C'_{i-1}, N_i, C^{\alpha}_i, C'_i, N_{i+1}$) is calculated, and where this is more than 2σ from the mean for the helix, the helix is kinked. This study identified ten proline induced kinks, 9 of which had very similar kink angles ($26 \pm 5°$).

1.5.2.2 Prokink (Visiers *et al.,* 2000)

Prokink is a method developed to evaluate helix kinks (distortions) caused by proline. The axes of the of the pre-kink (pre-proline) and post-kink (post-proline) helix sections (the proline is the first residue in the post-kink helix sections) are calculated from the Cα coordinates (Figure 1.17C). The helix axes are calculated by the CHARMM program (Brooks *et al.*, 1983), which uses the cylinder method of Åqvist (1986) (Karplus, 2014) (Figure 1.17A). These helix sections can be any length, depending on the length of the helix and position of the proline, but the authors suggest that they should be at least 7 residues long. The kink is characterised by three angles:

- Kink (bend) angle: The angle between the pre-kink and post-kink axes
- Wobble angle: An angle describing the position of the proline Cα atom relative to the pre- and post- kink axes. This is 0° when the helix kinks towards the proline, and 180° or −180° when the helix kinks away from the proline.
- Face shift: An angle describing the tightness of the helix turn before the kink (i.e. the proline atom). An α-turn has a face shift of 0°, a tight, 3_{10}-turn has a negative face shift,

and a wide, π-turn has a positive face shift. It is measured in the range $(180°, -180°)$.

1.5.2.3 Simulaid (Mezei, 2010)

The analysis program, Simulaid, incorporates a version of Prokink (Figure 1.17C). This uses axes calculated using the method within Kahn (1989) (Figure 1.17A).

1.5.2.4 TMKink (Meruelo *et al.*, 2011)

TMKink uses the Prokink version contained within Simulaid. However, it evaluates the kink angle at every single residue, and annotates kinks where the angle is $\geq 24°$, or the average over consecutive residues is $\geq 13°$ (Figure 1.17B). Kinks are removed if there is a larger kink within nine residues. The residue with the largest angle is the kink residue.

1.5.2.5 Hall *et al.* (2009)

This study used the Prokink method in Simulaid. Again, they calculated a kink angle for every residue. Kinks were identified where kink angles were $\geq 13.4°$, providing they were at least seven residues away from any larger kink.

1.5.2.6 Helanal (Bansal *et al.*, 2000), and Helanal-Plus (Kumar & Bansal, 2012)

Helanal and Helanal-Plus classify helices as kinked, curved, or straight (linear). Kink angles are calculated as in Prokink, but with axes calculated using the method described in Sugeta & Miyazawa (1967) (Figure 1.17A). The Sugeta & Miyazawa (1967) method also calculates the midpoint of each four residue segment of the helix. These mid-points are fitted to a straight line, and a circle (Helanal) or sphere (Helanal-Plus). A heuristic algorithm is used to classify the helix, using a combination of the largest kink angle in the helix, and the quality of the line and circle/sphere fits. In Helanal, a helix with a kink angle $\geq 20°$ is classified as kinked. In the updated version, Helanal-Plus, helices with kink angles $\geq 30°$ are classified as kinked, and those with angles between 20° and 30° are classified as kinked if their fits to the line and sphere

are sufficiently poor. These two algorithms are technically quite different, however they give similar results. In both, the residue with the largest angle is the kink residue.

1.5.2.7 Werner & Church (2013)

This study uses the method of Kahn (1989) to calculate local helix axes (Figure 1.17A). An angle for each residue is calculated from the axis for the four residues before it, and the axis of the four residues after it. Kinks are identified where the angle is $\geq 13°$, and the residue with the largest angle is the kink residue.

1.5.2.8 Manual annotation (Kneissl *et al.,* 2011)

In this study, two researchers annotated helices as kinked, curved, or straight. Their criteria are described in the paper: 'as a rule of thumb, a kink can be an abrupt change of the helix axis'. The kink residue is chosen by the human annotators, and no kink angle is ascribed to the kink.

1.5.2.9 MC-Helan (Langelaan *et al.,* 2010)

The MC-Helan algorithm identifies straight helix sections within a protein using strict helix criteria (Figure 1.17C). It identifies five residue long helix seeds, and iteratively grows them by adding the next residue, providing it adheres to the criteria. The criteria are that the ϕ and ψ angles are within the expected ranges (Lovell *et al.*, 2003), and that the backbone atoms are within 3Å of the fitted helix axis. This helix axis is a cylinder fit, using a Monte-Carlo method, rather than the approach in Åqvist (1986).

Kinked (bent) or distorted helices are then identified where a helix (as defined by e.g. DSSP) is made up of more than a single straight helix section. The kink angle is calculated, as for the other methods, from the axes of the two helix sections.

The kink residue is selected using a voting procedure, based on the inter Cα distances, inter Cα angles, and ϕ and ψ angles of the residues close to the boundary between helix sections.

1.5.2.10 Other methods

Many other authors have used their own methods and there is one example which is pertinent here, if only to describe fully the possible approaches to this problem. Hischenhuber *et al.* (2012) used both polynomial fits and splines to characterise the helices in the MH^2c protein. From these fits, the local curvature of the helix was calculated. The curvature was then used to identify kinks in these helices.

The authors discuss the available options in the polynomial and spline fits (i.e. the order of the polynomial, or the smoothing of the spline), and conclude that polynomials with order of between three and seven give the best helix descriptions. It should be noted, however, that this approach may not scale over the range of helix lengths that is required for kink analysis. Indeed, even though the MH^2c helices have fairly similar lengths, different orders of polynomial are recommended by the authors for different helices.

These methods are hard to group. One possible division is between sliding window methods and whole helix methods. However, some methods use approaches from both. For example, Helanal calculates the kink angle for a sliding window of seven residues, but also uses a measures of the straightness and curvedness of the whole helix for classification. Indeed, any method that produces a vector of values that relate to the straightness of the helix can be used to identify kinked helices.

The kink identification methods all have one thing in common - they reduce the three dimensional coordinates of a protein helix to a handful of statistics, and ultimately to a single classification (i.e. kinked, curved, or straight). There are many other methods that reduce three dimensional coordinates of a protein to a more simple representation. Using quaternions (Hanson & Thakur, 2012; Quine, 1999), splines (Geetha & Munson, 1996), and Ramachandran angles are examples of such approaches, and there are many more (Kandiraju *et al.*, 2005; Murray *et al.*, 2011; Reyes, 2011). Some of these methods have been used to characterise the straightness of helices (Kneller & Calligari, 2006; Murray *et al.*, 2011; Quine, 1999), and these could all be used to identify kinks, although this would require their parameters to be optimised

for this purpose.

Approaches to representing proteins are important in both protein structure comparison (and thus structure alignment) and protein visualisation. Comparison and alignment of proteins using their atomic coordinates is a computationally intensive process, and simplifying the protein representation can significantly reduce the computational resources required. Comparison and alignment of proteins using the RMSD (root mean square deviation) between the atoms has several problems, some of which can be addressed using alternative representations (Betancourt & Skolnick, 2001; Can & Wang, 2013).

Reduced representations are also useful for displaying proteins. Protein visualisation software (e.g. PyMol (Schrödinger LLC, 2014) and VMD (William Humphrey *et al.*, 1996)) provide simple ribbon or cartoon representations of proteins in three dimensions. In PyMOL and VMD these use a spline fit to the C^{α} atom coordinates, with a user variable smoothing parameter. PyMOL and VMD also provide simple helix representations - cylinders based on least-squares fits to the C^{α} atom coordinates. Bendix, a VMD plugin, uses a variation of the Sugeta & Miyazawa (1967) method to produce a more accurate helix representation than those in PyMOL and VMD (Dahl *et al.*, 2012).

1.5.3 The causes of kinks

There is a variety of methods that have been used to characterise kinks and explain their causes. Analysis of single proteins in mutation experiments (e.g. Weber *et al.* (2012)) and molecular dynamics studies (e.g. Bettinelli *et al.* (2011); Fowler & Sansom (2013)), provide evidence for the causes of kinks. Evidence also comes from the effects of naturally occurring mutations (Kuechler *et al.*, 2010), the kinks in a protein family (Bettinelli *et al.*, 2011; Devillé *et al.*, 2009), and surveys of kinks in many proteins (Hall *et al.*, 2009; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011).

There have been many suggestions as to the causes of kinks in membrane proteins, with proline being the most commonly suggested factor.

1.5.3.1 Role of Proline

The presence of proline in a helix is strongly associated with that helix being kinked (Barlow & Thornton, 1988; Langelaan *et al.*, 2010; Yohannan *et al.*, 2004b). Proline cannot be fully incorporated into an α-helix, due to its lack of an amide proton, and the ring formed by its backbone and sidechain (Cordes *et al.*, 2002; Hall *et al.*, 2009; Yohannan *et al.*, 2004c). This precludes proline from forming the $i+4 \rightarrow i$ hydrogen bond to a backbone carbonyl, which is the principle stabilising force of the α-helix structure. Although earlier work suggested that all kinks were associated with prolines, or so-called vestigial prolines[1] (Yohannan *et al.*, 2004b), more recent work has shown that there are many kinks that are not associated with proline (Cao & Bowie, 2012; Hall *et al.*, 2009; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010).

1.5.3.2 Other causes of kinks

Many of the studies have investigated the presence of other amino acids at specific points around the kink residue, but no patterns, other than proline, have been consistently found (Hall *et al.*, 2009; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Marsico *et al.*, 2010a; Meruelo *et al.*, 2011; Rigoutsos *et al.*, 2003; Weber *et al.*, 2012; Werner & Church, 2013). For example, glycine has been found to be prevalent around kinks in some studies (Devillé *et al.*, 2008; Hall *et al.*, 2009; Kneissl *et al.*, 2011), but not in others (Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Werner & Church, 2013). It has also been suggested that residues with sidechains that can form hydrogen bonds to the backbone, such as serine and threonine, are important in kinks (Deupi *et al.*, 2004; Weber *et al.*, 2012), although these side chains are rarely seen bonding to the backbone in the known structures of kinks (Deupi *et al.*, 2004; Hall *et al.*, 2009). A number of papers have suggested more complex sequence motifs that may be important in kinks (for example Del Val *et al.* (2012); Hall *et al.* (2009); Marsico *et al.* (2010a,b)), but once again none of these motifs are consistently observed across studies.

The dominance of proline suggests that hydrogen bonds are very important to the forma-

[1] Where proline is not seen in the kinked helix structure, but is observed in a homologous sequence

tion of kinks. There is evidence that membrane protein helices have different hydrogen bonds compared to soluble proteins (Cao & Bowie, 2012). Water molecules have been identified as stabilising some kinks in crystal structures (Tate *et al.*, 2010). The tertiary structure of proteins is also important to the structure of helix kinks, for example helix-helix packing (DeGrado *et al.*, 2003; Eilers *et al.*, 2000; Hildebrand *et al.*, 2008)

These studies provide contrasting explanations for the appearance of kinks within structures. This is largely due to the variety of ways in which kinks are identified and located in different studies (i.e. manual inspection or one of many computational methods). Identifying the method-independent features of kinks is one of the key aims of this thesis.

Hydrogen bonds between the amide group of the $i+4^{th}$ residue and the carbonyl group of the i^{th} residue in the backbone of the protein are the primary feature of α-helices. A missing $i+4 \rightarrow i$ hydrogen bond is frequently observed close to the kink. Kinks may be stabilized by the presence of non-canonical hydrogen bonds, such as $i+3 \rightarrow i$ or $i+5 \rightarrow i$ backbone connections, sidechain to backbone (Deupi *et al.*, 2004; Hall *et al.*, 2009), or other types of hydrogen bonds (Bowie, 2011; Kauko *et al.*, 2008; Rey *et al.*, 2010). These non-canonical hydrogen bonds have also been implicated in the flexibility of membrane helix kinks, by residues shifting their backbone hydrogen bond partners (Cao & Bowie, 2012).

1.5.3.3 Causes in soluble proteins

Kinks in soluble proteins have not been studied to the same degree as those in membrane proteins. There are examples of functional kinks in soluble protein helices (de Almeida & Holoshitz, 2011; Rudolph *et al.*, 2006), and I have found two examples of research on general analysis of helix distortions in soluble proteins. In one analysis of soluble Helix-X-Helix motifs, kinks were usually found when the linker residue (X) was glycine, serine, aspartic acid, or asparagine, or when proline was found after the linker residue (Devillé *et al.*, 2008). Additionally, the linker residue was frequently buried (i.e. not exposed to solvent. A study of proline distortions in soluble protein helices focused on classifying the distortions based on Ramachandran angles and

hydrogen bonding patterns (Rey *et al.*, 2010). However, unlike the studies of membrane helices, it did not consider the residue preferences around these distortions.

1.5.4 Modelling and prediction of kinks

There are specific methods tailored to membrane protein structure predictions (e.g. Kelm *et al.* 2010). These exploit the important structural role that the specific membrane environment has on proteins to improve on methods developed for all types of protein structure. Differences in kinks is one way in which the protein conformation can change. This makes template selection for homology modelling difficult (Worth *et al.*, 2009), so understanding kinks has a role to play in homology modelling. For example, in GPCRs, the quality of docking is affected by 'substantial variations in kinks and helical structure in the binding pocket region' (Kufareva *et al.*, 2011). Understanding kinks is key to building models to guide drug development.

Despite recent experimental structures of GCPRs (Cherezov *et al.*, 2007; Hanson *et al.*, 2012; Rasmussen *et al.*, 2011; Rosenbaum *et al.*, 2011), these structures are closely related and only represent a small subset of GPCRs (Rosenbaum *et al.*, 2009). Therefore, theoretical models are still important for structure based drug discovery (Tang *et al.*, 2012).

There are some existing methods that predict kinks in helices (e.g. Kneissl *et al.* 2011; Meruelo *et al.* 2011; Seifert *et al.* 2014). These are primarily machine learning methods based on the helix sequence. Despite reportedly good performance, the authors do not provide any biochemical explanation of the performance, over and above that of the presence of proline. The quality of results also depends heavily on the method chosen to identify kinks, and it is not clear if these methods perform well enough for use in structure prediction. Other methods have used molecular dynamics to predict the occurrence of kinks (Hall *et al.*, 2009; Mai & Chen, 2014).

There are three recent examples of kinks being used in structure prediction. One study used kink characterisation to score membrane protein model decoys during modelling (Werner & Church, 2013). Two studies have attempted to incorporate kinks into structure prediction

(Chen *et al.*, 2014; Mai & Chen, 2014).

1.6 Identification and characterisation of kinks in membrane and soluble proteins

There are a variety of definitions and means of identification for kinks. I have designed two computational methods to identify kinks (Chapter 2). These are compared to a set of kinks identified by humans (obtained through a crowd sourcing approach), and other kink identification methods (Chapter 3). Using two computational kink identification methods, I show that helix kinks are a feature of long helices in all proteins (soluble and membrane), not just in helices in integral membrane proteins (Chapter 4). Some kinked helices are known to adopt multiple conformations in homologous proteins, and even in the same proteins. The similarities of helices in homologous proteins, and the flexibility of kinks within individual proteins will be explored in Chapter 5.

CHAPTER 2

Computational kink identification

2.1 Introduction

In this chapter I will describe the two computational methods I developed for the identification of helix kinks in protein structures: the B statistic, a statistically robust method that discriminates between kinked and not kinked helices, and Kink Finder, which identifies, locates, and measures kinks in helices, and is used for the majority of my thesis. Kink Finder employs novel methods to consistently locate kinks.

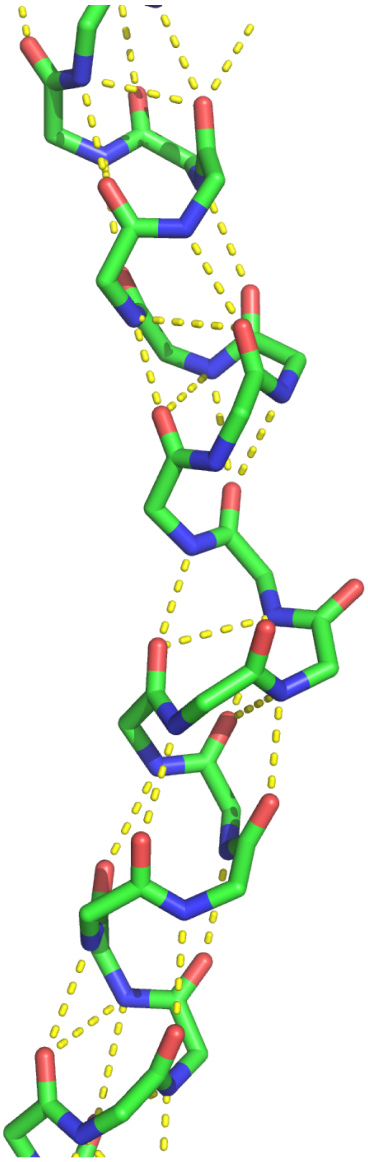


Figure 2.1: **A section of a kinked helix,** showing the hydrogen bonds (dashed yellow). Note the hydrogen bonds between the oxygen and nitrogen atoms. These are not all present around the kink, where the helix changes direction. The helix runs from the N-terminus (bottom) to C-terminus (top). No side chains or hydrogen atoms are shown. Atoms are coloured as follows: carbon - green, nitrogen - blue, oxygen - red.

2.1.1 The significance and role of helix kinks

Kinks are distorted regions of helix, where the helix changes direction abruptly (e.g. Figure 2.1). There are three principal reasons why I am interested in kinks. These are fully described in Chapter 1, but are summarised here. First, kinks are an important but poorly understood feature of protein structures. Many factors that modulate the structure and function of kinks have been suggested (see Section 1.5.3), but there is no clear understanding of the characteristics and causes of helix kinks.

Second, kinks occur frequently in the transmembrane helices of membrane proteins, which are very important in cellular processes, and as drug targets. For example, Rhodopsin-like GPCRs, a subset of membrane proteins, make up $> 26\%$ of small molecule drug targets (Overington *et al.*, 2006). There is a comparatively small number of known membrane protein structures, making the modelling and prediction of membrane proteins important. Differences in kink conformation can change the structure of proteins, so identifying the correct helix kink conformation is key to good structural modelling of proteins (Kufareva *et al.*, 2011; Worth *et al.*, 2009; Yohannan *et al.*, 2004b).

Third, although there have been many studies into kinks and other helix distortions, there is no universally accepted definition of kinks (Bansal *et al.*, 2000; Hall *et al.*, 2009; Kauko *et al.*, 2008; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Rigoutsos *et al.*, 2003; Visiers *et al.*, 2000; Werner & Church, 2013). Methods frequently disagree on the identification, location and size of kinks. The broad definition of 'kink', as a sharp change in direction of the helix is well established. Other synonymous terms have been used (e.g. bend (Langelaan *et al.*, 2010) and hinge (Sansom & Weinstein, 2000)), but the general concept remains the same. However, the specific definition varies from study to study, sometimes significantly, so while the different methods agree on the extreme examples, many helices between the extremes of kinked and not kinked are classified differently by the different methods.

A better characterisation and knowledge of kinks could provide an improvement to methods to predict and understand biologically and pharmaceutically important proteins. As kinks are

a feature of α-helices, the first step towards kink identification in a protein is to identify the α-helices. The next section briefly describes α-helices (for a fuller description, see Section 1.4.3), and the methods that I use to identify them.

2.1.2 The challenge of identifying α-helix kinks

An α-helix is a regular repeating structural unit of a protein, where the amide group of each residue is hydrogen bonded to the carbonyl group of the residue four towards the N-terminus of the protein chain (Pauling *et al.*, 1951). An example of a kinked helix is shown in Figure 2.1. The top and bottom sections of the figure show the regular pattern of the alpha helix, with hydrogen bonds (dashed yellow) between the carbonyl oxygen (red) of one residue, and the amide nitrogen (blue) of the residue 4 above it. When a series of adjacent residues adopt this hydrogen bond, a regular helical shape results, as at the top and bottom of Figure 2.1.

There are a number of available methods to annotate the residues in protein structures which are helical, including DSSP (Kabsch & Sander, 1983), STRIDE (Frishman & Argos, 1995), and KAKSI (Martin *et al.*, 2005) (for further details, see Section 1.4.2). These methods give generally similar annotations, but differ around the ends of helices and in the annotation of distorted helical regions (Martin *et al.*, 2005). The method for helix annotation used here is DSSP (Kabsch & Sander, 1983). This works by using distance criteria to identify the residues that have the hydrogen bonds characteristic of α-helices, as described in Section 1.4.2.1. Where groups of consecutive residues have sufficient numbers of these backbone hydrogen bonds, DSSP annotates these residues as helical (Kabsch & Sander, 1983).

DSSP was chosen as my principal secondary structure assignment method for its simplicity. This allowed me to see clearly how any manipulations that I made to the method affected its results. Additionally, DSSP was the existing method used by the Deane group. It is an important part of the tools iMembrane (Kelm *et al.*, 2009), Medeller (Kelm *et al.*, 2010), and MP-T (Hill & Deane, 2013), which I use in this thesis. All of these rely on the DSSP method as implemented in JOY (Mizuguchi *et al.*, 1998). It is very difficult argue that any one secondary

structure assignment method is better than another, as there is no definitive benchmark set on which to test the methods. There have been many subsequent methods which are arguably better, particularly in specific circumstances. However, there is no method which is clearly and consistently better than DSSP. Its simplicity and the decision of the authors to distribute the method freely has ensured that it remains as one of the most popular methods.

There are two specific problems when using any helix annotation method for creating a dataset of helices with which to identify kinks. First, some residues in kinked regions tend not to be annotated as helical, even though they are between two helical segments. Some methods, e.g. KAKSI (Martin *et al.*, 2005), even go as far as explicitly removing kinks from helices. Second, the irregular helix structure in the residues at the termini of some helices causes kink identification methods to identify kinks where they are not present. The methods I employed to combat these problems are described below.

2.1.2.1 Kinks not included in helices

I searched for kinks in protein helices, so it was important that the helices used include kinks. In regions of proteins where there are kinks, some of the backbone hydrogen bonds that would otherwise be present in an α-helix are not present (e.g. in Figure 2.1). Although DSSP does not require all of the expected hydrogen bonds to be present for a residue to be annotated as helical, if some are missing a residue can be annotated as non-helical. This results in two shorter helices separated by one or two non-helical residues, rather than a single long helix. These regions of the protein need to be incorporated into a single helix when I attempt to identify kinks in protein structures. Therefore, in the methods presented here, where sections of the protein that are annotated as helical were separated by only one or two residues, I combined these sections into a single helix.

2.1.2.2 Helix termini

Kink identification methods can be very sensitive to residues that are incorrectly annotated as helical. Often residues at helix termini have a less regular structure than other parts of helices. Secondary structure annotation methods frequently disagree on the annotation of residues at, or close to, the ends of helices. My initial analysis identified large numbers of kinks at the starts and ends of helices. To ensure that this was not an artefact of incorrect helix annotation, I checked the ends of the helices for their helical nature.

I ensured the helical nature of the termini of the helices using a method based on a routine that identifies 'helix seeds' in MC-Helan (Langelaan *et al.*, 2010). Where the first or last five residues of helices that I identified were not a helix seed, I iteratively removed residues from the end of the helix until this condition was met. The requirements for a helix seed are threefold. First, the first residue must have dihedral angles in the alpha-helical region (Lovell *et al.*, 2003). Second, the angles ($C^{\alpha}_{i+x}\widehat{C^{\alpha}_{i}C^{\alpha}_{i+1}}$, $x = 2,3,4$) must lie within the expected ranges for an alpha helix ($35-50°$ for $x = 2$, $60-80°$ for $x = 3$, and $45-65°$ for $x = 4$). Third, the $C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+x}$ ($x = 2,3,4$) distances must be within 0.5 Å of the values for an ideal α-helix.

Once a set of helices has been identified in a protein, there are a number of possible approaches to identifying kinks. The next section summarises the problems with existing kink identification methods.

2.1.3 Existing kink identification methods

There are a number of available methods to identify, locate, and measure kinks in helices (e.g. Kumar & Bansal (2012); Langelaan *et al.* (2010); Meruelo *et al.* (2011), see Section 1.5.2). These classify helices into one of a number of groups. The terms used to describe these groups vary from study to study. Some algorithms are binary classifiers (e.g. kinked/straight in TMKink (Meruelo *et al.*, 2011)), some are ternary classifiers (e.g. kinked/curved/linear in Helanal-Plus Kumar & Bansal (2012)). The terms used are not always the same, but the terms used by each method broadly correspond to kinked, straight, and, in ternary classifiers, curved. Unfortunately, the

Helix	3DDLA 90-112	1RHZA 365-394	3DDLA 13-36	3DDLA 47-72	1RHZA 138-162	1RHZA 314-335	1RHZA 102-128	3DDLA 151-173
Kneissl	Kinked	Kinked	Curved	Curved	Straight	Straight	Curved	Curved
Helanal	Curved	Kinked	Curved	Curved	Linear	Linear	Curved	Kinked
MCH	Kinked	Kinked	Kinked	Kinked	Straight	Straight	Kinked	Kinked
KF	Kinked	Kinked	Straight	Straight	Straight	Straight	Kinked	Kinked

Table 2.1: **Classifications of the helices in Figure 2.2 by four methods.** A manual classification by Kneissl *et al.* (2011), Helanal-Plus (Kumar & Bansal, 2012), MC-Helan (MCH) (Langelaan *et al.*, 2010), and Kink Finder (KF). The classification terms are defined differently by each of the methods.

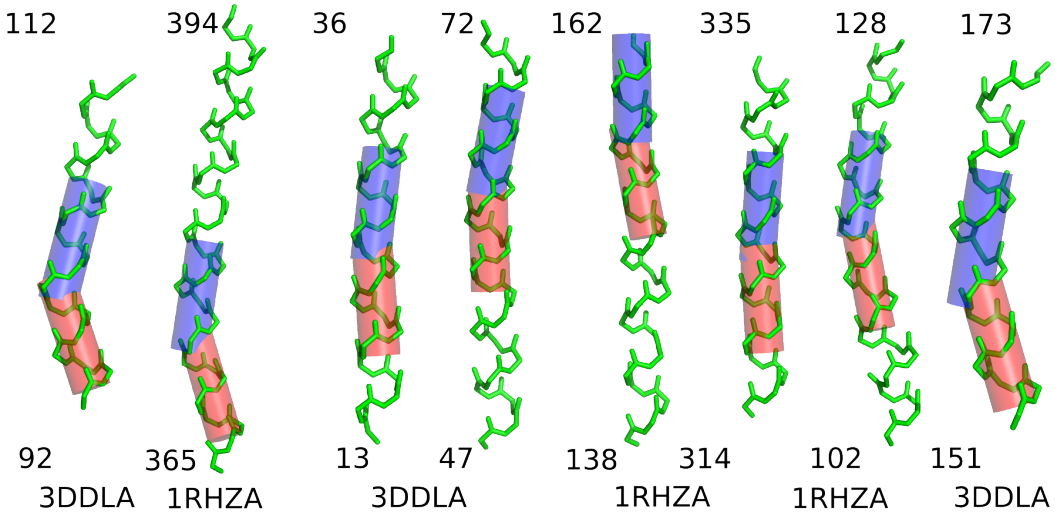


Figure 2.2: **Eight example helices.** The sections of the helix before and after the largest kink angle in each helix are fitted to cylinders as calculated by Kink Finder, shown here. The classification of these helices by four methods is shown in Table 2.1. Helices are labelled with their 4 character PDB code and chain identifier, and number of first and last residue.

terms used by each study do not exactly correspond, nor do the measurements (e.g. a kinked helix identified by Helanal-Plus is not necessarily the same as a kinked helix by TMKink, and a measurement of 25° by Helanal-Plus does not necessarily correspond to a 25° measurement by TMKink). The algorithms use different methods to identify kinks, and frequently disagree on the classification of helices (see Figure 2.2 and Table 2.1). The differences between the classifications are further explored in Chapter 3.

There are three primary deficiencies in these algorithms; (1) they all require many parameters to be set in their algorithms to identify kinks, (2) they locate kinks inconsistently relative to the helix structure (which means that methods rarely agree on the specific location of the kink in a given helix), and (3) they make no estimation of the uncertainty of the measured angle. The first two are discussed in further detail below, while the third is a focus of Chapter 5.

2.1.3.1 Decisions and parameters

In each of these kink identification methods, there are a large number of decisions that have been made, and parameters set to provide the classification of helices and identification of helix kinks. Examples of these decisions include the method to fit axes to helix sections, the length of the sections, and the threshold(s) to discriminate between different groups. The decisions are often interdependent and their complex interaction makes it difficult to infer any biological meaning from each decision through to the eventual output. Indeed, there is rarely any strong biological or statistical rationale to help make these decisions. Similarly, the output of the methods does not provide any good rationale with which to select an appropriate threshold. Neither is there an available set of kinks on which to train a classifier. Helices do not fall into two clear subsets (i.e. kinked and not kinked); instead they form a continuum from kinked to not kinked (Figure 2.2). In all methods, the threshold between classifications is a sharp division based on a statistic (e.g. angle) with a unimodal distribution. As a result, I am aware of no method that has a strong statistical or biological justification for its choice of threshold.

2.1.3.2 Location of kinks

Often the conclusions drawn by studies of helix kinks rely on the choice of kink location. For example, many studies characterise the residues that occur one position after the kink residue (see Section 1.5.2). Current methods each use a different method to locate the kink residue and no method does this consistently relative to the helix shape (see Figure 2.3). The kink

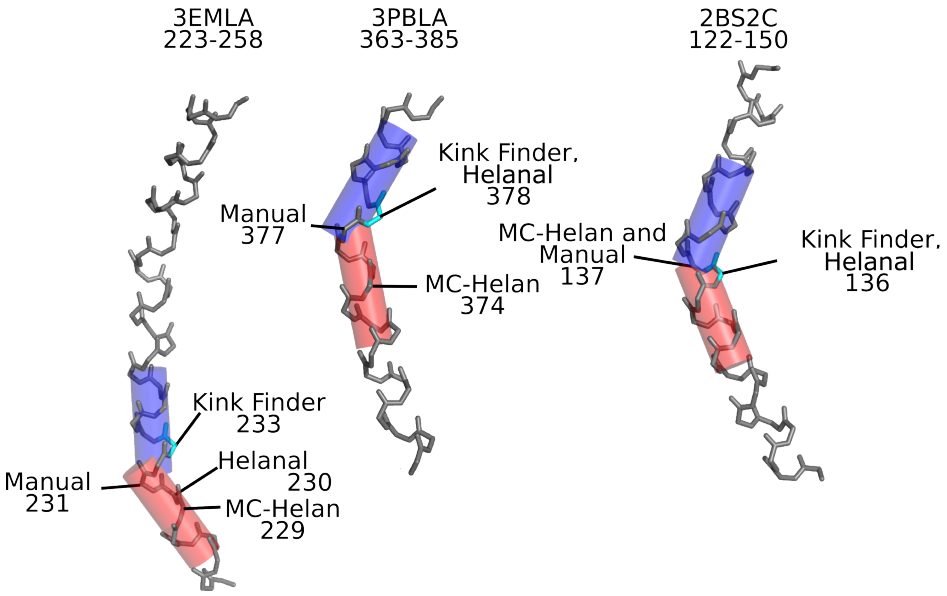


Figure 2.3: **Kink position disagreement between existing methods.** Three example helices, showing disagreements between the kink position as annotated by four methods, MC-Helan (Langelaan *et al.*, 2010), Manual (Kneissl *et al.*, 2011), Helanal (Kumar & Bansal, 2012), and Kink Finder (described in this chapter). The existing methods are described in Section 1.5.2. Helices are from the PDB structures indicated, with only the backbone atoms shown. The cylinders pictures were fitted by Kink Finder.

residue (the location of the kink) is often chosen as the residue with the largest angle, which is sometimes on one side of the kinked helix, and sometimes on the other. I believe that this inconsistent location of the kink has led to unclear and contradictory results in these studies of the characteristics of kinks.

2.1.4 My methods

I created two methods to identify kinks, which resolve some of the problems identified above: the B statistic, and Kink Finder.

2.1.4.1 B statistic

The B statistic provides a statistically robust method to classify helices into one of two groups. This method requires no decisions, and the classification threshold is derived from a Gaussian mixture model fitted to a set of helices. The threshold gives a classification different from that of previous methods, with a much smaller group of 'kinked' helices than other methods. This method, however, does not provide the location of the kink in the helix. I developed this method after Prof. K. V. Mardia approached us with the initial concept of applying this statistic to protein helices.

2.1.4.2 Kink Finder

Kink Finder is the primary method that we use in this thesis to identify, locate, and measure helix kinks. It is similar to TMKink (Meruelo *et al.*, 2011) and Prokink (Visiers *et al.*, 2000), using local axes to calculate kink angles in helices. However, unlike these and other methods, it locates the kink residue consistently relative to the shape of the helix, and it provides an estimate of the uncertainty in measured kink angles (see Chapter 5).

2.2 Data set

The data set used in this chapter is the helix data set created by Kneissl *et al.* (2011). It contains helices from from 132 α-helical membrane proteins. These membrane proteins were taken from the MPtopo database, which contains all membrane proteins with experimentally verified transmembrane regions, and has no sequence identity threshold (Jayasinghe *et al.*, 2001). The resolution of these structures ranges from 1.60Å to 4.50Å, and 60% of the chains have resolution of less than 3.00Å.

The helices were filtered to have $\leq 95\%$ sequence identity. Helices that are not in the membrane were removed. This gave a set of 1014 helices, the annotation of which was manually curated. The length of these helices ranges from 12 to 43 residues. Of these, 357 are annotated as kinked, 461 straight, and the remaining 196 curved. This set provides a high quality helix set on which to evaluate my methods. For the initial B statistic analysis, I remove helices that have a π-turn from this set. In this case, π-turns were defined as where the SST program, which identifies hydrogen bonds in the same way as DSSP (Kabsch & Sander, 1983), identified a $i+5 \rightarrow i$ main chain to main chain hydrogen bond. This subset contains 597 helices.

2.3 B statistic

In this section I describe the B statistic method, which provides a very simple and statistically robust method to categorise helices as kinked or not kinked.

2.3.1 Method

This method uses a principal component analysis of C^{α} atom coordinates to classify helices. In all helices, the principal component will approximate the helix axis. The second and third components are perpendicular to this axis. In a straight helix, the second and third eigenvalues are approximately equal. Conversely, a kinked helix has non-equal second and third eigenvalues (Figure 2.4). This observation forms the basis of the method.

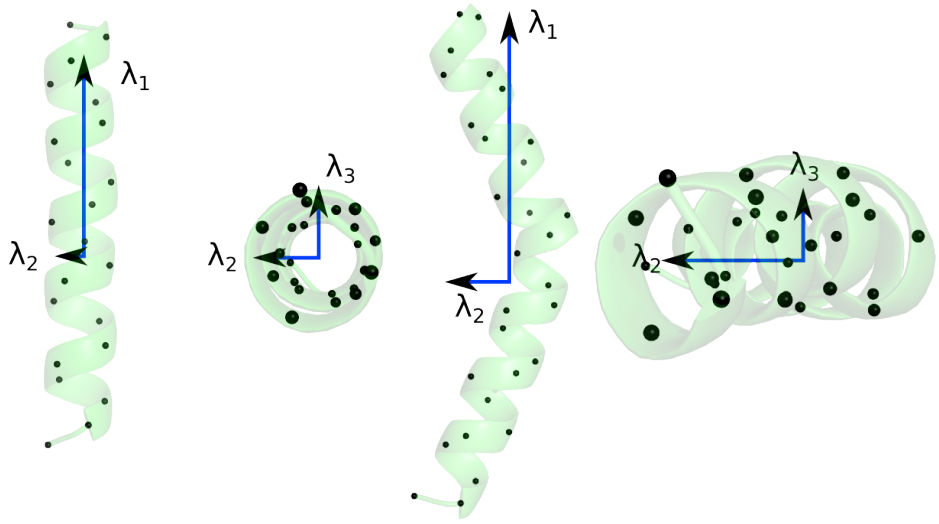


Figure 2.4: **B statistic method for identifying kinked helices,** using the principal components of the C^{α} atom coordinates. The principal components are labelled in descending order, with λ_1 the largest. The second (λ_2) and third (λ_3) eigenvalues of straight helices are very similar. However, these eigenvalues are very different in kinked (or otherwise distorted) helices.

The similarity of the eigenvalues can be evaluated using the following statistic, B:

$$B = 2\left(n - \frac{17}{6}\right)\log\left(\frac{a}{g}\right) \tag{2.1}$$

where a and g are the arithmetic and geometric means of the second and third eigenvalues, and n is the number of atoms used in the principal component analysis. This is a specific form of this equation:

$$B = \left(n - \frac{2p+11}{6}\right)(p-k)\log\left(\frac{a}{g}\right) \tag{2.2}$$

with $k = 1$ and $p = 3$. This general form is used to test for isotropy of eigenvalues, and evaluate if the $k+1^{th}, k+2^{th}, \ldots, p^{th}$ eigenvectors contain the same amount of information (Mardia *et al.*, 1979).

I implemented the B statistic method in Python 2.7 (www.python.org). The principle components of the C^{α} atom coordinates were calculated using the singular value decomposition function in the linear algebra module (linalg.svd) of the NumPy package (van der Walt *et al.*, 2011).

The distribution of $log(B)$ for the 597 transmembrane helices without π turns is used to determine the threshold between kinked and not kinked helices (Figure 2.5a). The distribution can be approximated as a mix of two normal distributions, which are fitted with a Gaussian mixture model. $Log(B) = 0$ is the threshold for discriminating between the kinked and not kinked groups.

2.3.2 Results

The distribution of the B statistic for the 597 helices without π-turns in the Kneissl *et al.* (2011) data set is bimodal (Figure 2.5a). The fitted Gaussian mixture model corroborates this, and indicates that the ideal classification threshold is log(B) = 0. The mixture model indicates that the helices divide into a 'kinked' group (18% of helices) and a 'not kinked' group (82% of helices). Most other kink identification methods classify a higher proportion of helices as

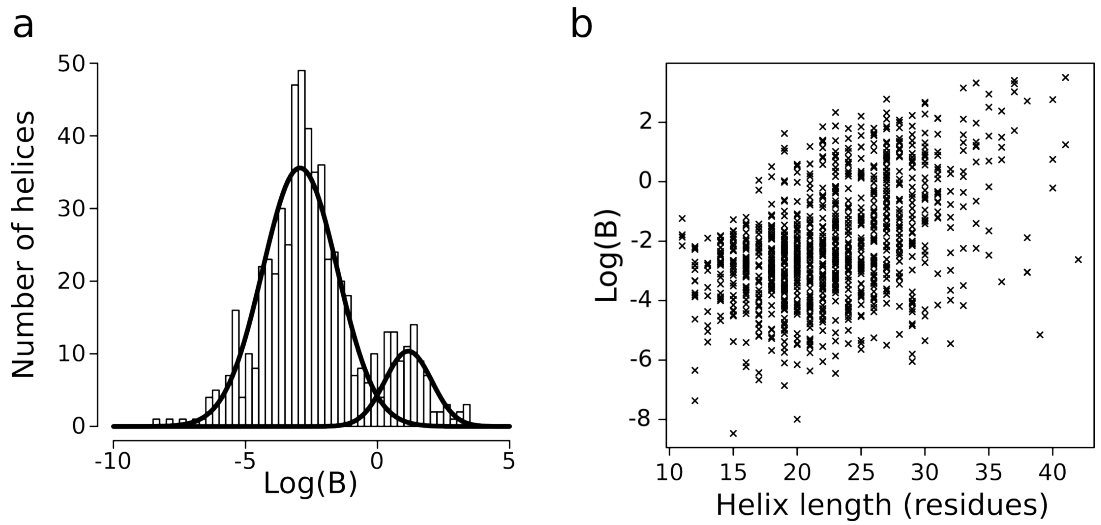


Figure 2.5: (a) **Distribution of log(B)** for 597 membrane helices which contain no π-turns. A Gaussian mixture model is fitted to the distribution. (b) Variation of log(B) with helix length for all 1014 helices.

kinked. In comparison, Kniessl et al. annotate 35% of these helices as kinked, and a further 19% as curved. If the helices with π-turns are included, the separation between the 'kinked' and 'not kinked' groups is less pronounced, but there is still a dip between the two peaks. The value of log(B) is generally more positive for longer helices (Figure 2.5b, $r^2 = 0.21$), indicating longer helices are more often kinked (and kinked to a greater degree) than shorter helices.

Unlike most other methods, this does not provide a location of the kink in the helix. In addition, this method only works for helices with at least 10 residues. For shorter helices, the principal component analysis does not accurately identify the helix axis.

2.4 Kink Finder

This section describes Kink Finder, the principal method I developed to identify, locate and measure kinks in helices. An outline of the method is shown in Figures 2.6 and 2.7. The basis of this method is the cylinder fits, which are described below, followed by the method used to consistently locate kinks. A method used to estimate the angle measurement error is described

in Chapter 5.

2.4.1 Methods

There have been many approaches used to identify kinks in protein helices (see Section 1.5.2). As discussed in Section 2.1.3.1, these contain many subjective decisions and parameters. The effect that changing these would have on the identification of kinks is unclear from the algorithms. Developing Kink Finder provided both a way to understand the intricacies of finding kinks, and an algorithm that was easy to modify for my research goals.

For many of the approaches described in Section 1.5.2, the classification (as kinked or otherwise) of a section of a helix depends on the helix that it is a part of. For example, kink classifications in MC-Helan (Langelaan *et al.*, 2010) rely on the helix annotation in the protein chain and on the helix axes, which are estimated from the coordinates of a variable number of residues. Helanal (Kumar & Bansal, 2012) relies on fitting points on the helix axis to a line and a sphere for kink identification. This means it is possible for a 12-residue section of helix to be identified as kinked, while an identical section is not identified as kinked if the helices that surrounded the two are different (either due to different secondary structure annotation or a different helix structure). This is also the case for methods that use spine or polynomial fits to characterise kinks. I did not want this to be the case in Kink Finder - identical helix sections should be given the same annotation, regardless of their surrounding. Therefore, I chose to use a method in Kink Finder that treated helix sections in isolation from their surroundings.

To do this, Kink Finder calculates angles for a residue based on estimated axes for the section before and the section after that residue in the helix. Therefore, a crucial part of Kink Finder (as for many other kink identification methods) is the method used to estimate the helix axis. Kink Finder uses cylinder fits to do this.

There are many approaches that have been used to fits axes to helix sections (see Section 1.5.2). These all work well on ideal, or close to ideal, helix sections, but work less well on distorted helix sections. Cylinder fits to all backbone atoms of six-residue sections of the helix

provide the best compromise between accuracy and sensitivity.

Generally, fits to longer sections give more accurate estimation of the helix axis than fits to shorter sections. Conversely, fits to shorter sections are more sensitive to helix distortions than fits to longer sections. Six residue helix sections provide a good compromise between these two considerations, and equates to the shortest section of a helix for which an axis can be visually estimated. Fitting axes to shorter sections, for example using the approach of Sugeta & Miyazawa (1967) or Kahn (1989), can give very poor estimates of the axes in distorted regions - e.g. π-turns. Therefore, angles calculated from estimated axes of short sections, must be filtered (to remove large angles that arise from distortions other than kinks) before they can be used to identify kinks. Indeed, the Helanal (Kumar & Bansal, 2012) method contains an algorithm to do exactly this. Further, to describe kinks accurately good estimates of the helix axis in kinked regions are necessary. However, fits to four-residue helix sections are not robust in kinked regions and so, even when these axes can be used to identify kinks, they often give a poor estimate of the kink angle.

In comparison to the methods of Sugeta & Miyazawa (1967) and Kahn (1989), cylinder fits provide a robust method to calculate axes. Fitting cylinders can accommodate wide and tight turns, as well as other helix distortions, while still giving good axis estimates. Although least-squares fits are the most common method of estimating the axis of a set of points, they are unsuitable for fitting axes to short helix sections. The quality of least-squares fits varies periodically with the length of the section fitted to. Least-squares fits to $n + \frac{1}{2}$-turn-long helix sections (where n is an integer, and one turn of an α-helix is 3.6 residues) typically give good axis estimates. However, fits to n-turn-long sections give axis estimates that are systematically up to 10° from the true axis. This occurs even when fitting to sections of an ideal helix. In comparison, cylinder fits provide good axes regardless of the section lengths. Cylinder fits cannot provide good estimates for helix sections that are shorter than six residues. Although cylinder fits to sections longer than six residues give better axis estimations, in regular sections of helices this improvement is $< 2°$ (even for fits to 10 residue or longer sections). This does

not compensate for the loss of sensitivity resulting from using longer helix sections. Using all backbone atoms rather than just the C^{α} atoms improves the accuracy of axis estimations in difficult cases.

Thus, cylinder fitting to all backbone atoms of six-residue helix sections provides a good balance between axis estimation accuracy and sensitivity to kinks. The cylinder fit method provides good axis estimates, and provides axis estimates are visually wrong by more than 5° in $< 1\%$ of cases. Cylinder fits are robust to distortions in the helix, so they provide good axis and angle estimates even around kinks, and these estimates are improved by using all backbone atoms, rather than just the C^{α} atoms. Finally, the cylinder fits provide quality of fit parameters that can be used to quantitatively assess the quality of each estimated axis.

2.4.1.1 Cylinder fits

The Åqvist (1986) study provides a method to fit cylinders to a set of points, by minimising the distance of the points from a cylinder surface, i.e. minimising:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \hat{d}\right)^2} \tag{2.3}$$

where d_i is the shortest distance from point i to the fitted axis, and $\hat{d}$ is the mean of these distances. Unlike standard least-squares regression fits, this has no algebraic solution. This Åqvist (1986) formulation contains two undesirable features. First, the derivatives are fractions, with d_{ik} as the denominators. Second the objective function (Equation 2.3) and its derivatives contain square roots. As a result, the derivatives rapidly approach infinity when d_i is small, while the objective function would be better behaved if the problem could be reformulated so that it did not require square roots.

In conjunction with Prof. Mardia, I developed a formulation that avoided these two problems. Prof. Mardia re-formulated the method after we had had many discussions about the problem in general and the specific approach in Åqvist (1986). I then implemented the method

in C. We use the same notation as in Åqvist (1986), so the observed points (x_i, with x, y, and z components x_{i1}, x_{i2}, and x_{i3}), orthogonal distance of the axis from origin (r_0, with components r_{01}, r_{02}, and x_{03}), and the axis direction (a, , with components a_1, a_2, and a_3) are:

$$x_i^T = (x_{i1}, x_{i2}, x_{i3}) \;\; i = 1, .., n; \;\; r_0^T = (r_{01}, r_{02}, r_{03}), a^T = (a_1, a_2, a_3), \tag{2.4}$$

with the conditions that:

$$|a| = 1, \;\; a^T r_0 = 0. \tag{2.5}$$

We use the objective function:

$$\Delta = \sum \left(d_i^2 - \overline{d^2} \right)^2, \;\; \overline{d^2} = \frac{1}{n} \sum d_i^2 \tag{2.6}$$

where the orthogonal distance squared of x_i to a_i is d_i^2:

$$d_i^2 = x_i^T \left(I - aa^T \right) x_i - 2 x_i^T r_0 + r_0^T r_0, \tag{2.7}$$

and I is the 3x3 identity matrix. Let us use the notation:

$$\xi_k^T = (r_{01}, r_{02}, r_{03}, a_1, a_2, a_3). \tag{2.8}$$

The derivatives of the function can be expressed as:

$$\frac{\partial \Delta}{\partial \xi_k} = 2 \sum (d_i^2 - \overline{d^2}) \left(\frac{\partial d_i^2}{\partial \xi_k} - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial d_i^2}{\partial \xi_k} \right), \tag{2.9}$$

where

$$\frac{\partial d_i^2}{\partial \xi_k} = -2 x_{ik} + 2 r_{0,k}, \;\; k = 1, 2, 3, \tag{2.10}$$

and then

$$\frac{\partial d_i^2}{\partial \xi_k} = -2 \left(x_i^T a \right) x_{i,k-3}, \;\; k = 4, 5, 6. \tag{2.11}$$

Note that in Equation 2.11 the square root of d_i disappears, and in Equations 2.10 and 2.11 the denominator disappears (compared to Equations 11, 12, and 13 in Åqvist (1986)).

To fit a cylinder, we optimized Δ (Equation 2.6) under the conditions in Equation 2.5. We used the gradients (which in our case are given by Equations 2.9-2.11) as part of the optimisation routine.

Kink Finder uses the above formulation, and a nonlinear conjugate gradient method adapted from Numerical Recipes in C (Press *et al.*, 1992) for minimizing the function (which I implemented in C), to fit cylinders to all backbone atoms from six residues of the helix. Given a starting vector of six parameters (the components of a and r_0) that describe an estimate of the helix axis, this method calculates the gradient of the function in the vicinity of this starting point. Initially, the direction with the steepest descent is chosen. A line minimisation is performed in this direction. From this new point, a new direction is chosen, using the Polak-Ribiere function, which combines the previous direction and the steepest descent from this new point. This process of line minimisations and choosing a new direction is repeated until a putative global minimum is reached. This fit relies on a starting point being chosen which is reasonably close to the correct answer. Without a good starting estimate, the fit is likely to be stuck in a local minimum, some distance from the global minimum.

2.4.1.2 Fit starting point

Initially, the starting points are taken from least-squares regression fits to 18, 21, 24, and 36 backbone atoms. These correspond to 1.5 turns of a 3_{10}, α-, and π-helix, and 2.5 turns of an α-helix. Least-squares fits to backbone atoms of a helix only give good helix axis approximation when they use $n + \frac{1}{2}$ turns of the helix (where n is an integer, and one turn of an αhelix is 3.6 residues). Using, for example, two turns of an α-helix (i.e. 7 residues) gives an axis estimate *c.* 10° from the true axis. Cylinders are fitted with each of these least-squares fits as starting points, and the fit with the smallest r is chosen. Although I optimise the function given in Equation 2.6, I use the original Åqvist (1986) formulation to evaluate the goodness of fit. As

such, r is defined as:

$$r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \hat{d}\right)^2} \tag{2.12}$$

where d_i and $\hat{d}$ are as defined in Equation 2.3. This gives the global function minimum for all but the most difficult fits. Where these provide poor fits, the axes of nearby helix sections are used as additional starting points. When calculating angles for each residue in the helix Kink Finder extends the length of the fit (up to 10 residues), where this results in a fit with a lower r.

The result of the cylinder fitting section of the algorithm is that each residue (other than the first 5 and last 6) has an axis for the 6-10 residues N-terminal to it, and an axis for the 6-10 residues C-terminal to it (Figure 2.6b). When running Kink Finder on the Kneissl et al. data set (which results in 50,000 fits), the fit is extended in 35% of cases. In 20% of cases the fit is extended to seven residues, in 8% to eight residues, in 4% to nine, and in 4% of cases to ten residues. Extending to more than 10 residues has an effect on very few calculated angles. Additionally, it can occur only when there is sufficient helix either side of the residue for which an angle is being calculated.

2.4.1.3 Identifying kinks

A local angle for each residue is calculated from the angle between the axes immediately N- and C-terminal to it (Figure 2.6b). Kink Finder identifies kinks by ordering the residues in a helix by their angle (largest first), and taking each residue in turn (Figure 2.6d-g). A kink is identified where a residue:

1. has an angle greater than 10°,

2. is at least four residues from any already identified kinks, and

3. has a residue with an angle less than 10° between it and any already identified kinks.

This residue is the 'initial kink residue' (or the 'uncorrected kink residue'). The method used to locate the 'kink residue' (or the 'corrected kink residue') in each kink is described in the next section. Where a kink definition larger than 10° is used, the initial set of kinks identified here are reduced to only those with an angle greater than the desired threshold.

2.4.1.4 Locating the kink

Kink Finder locates the specific site of the kink (the 'kink residue'), based not only on the local helix angle but also the local structure of the helix. The kink residue is designated as the residue with the wobble angle nearest 0° of the following four residues: the residue before the initial kink residue, the initial kink residue, and the two residues after the initial kink residue (Figure 2.7). Residues with wobble angles close to 0° are on the inside of kinks.

The wobble angle for a residue is calculated from the position of its C^{α} atom relative to the local helix axes. Two vectors are calculated: first, the C-terminal local axis is projected onto a plane perpendicular to the N-terminal local axis and containing the C^{α} atom; second, the vector between the C^{α} atom and the point where the N-terminal axis intersects the plane. The 'wobble angle' is calculated, which is the angle between these two vectors (Figure 2.8).

2.4.2 Results

For the residue with the largest angle in the kink (i.e. the initial kink residue), the wobble angle distribution is shown in Figure 2.9a. There is a bias for these wobble angles to be on the inside - i.e. between 270° and 90° - with a (circular) mean of 308°. Applying the position correction in Kink Finder means that the kink residues have the wobble angle distribution shown in Figure 2.9b. All kink residues have a wobble angle between 306° and 57°, and the (circular) mean is 354°.

Kink Finder's method of choosing kink residues on the inside of the kink results in the positions around a kink being consistently placed relative to the helix shape. Figure 2.10 shows the positions around the kink residue, with negative numbers towards the N-terminus, and

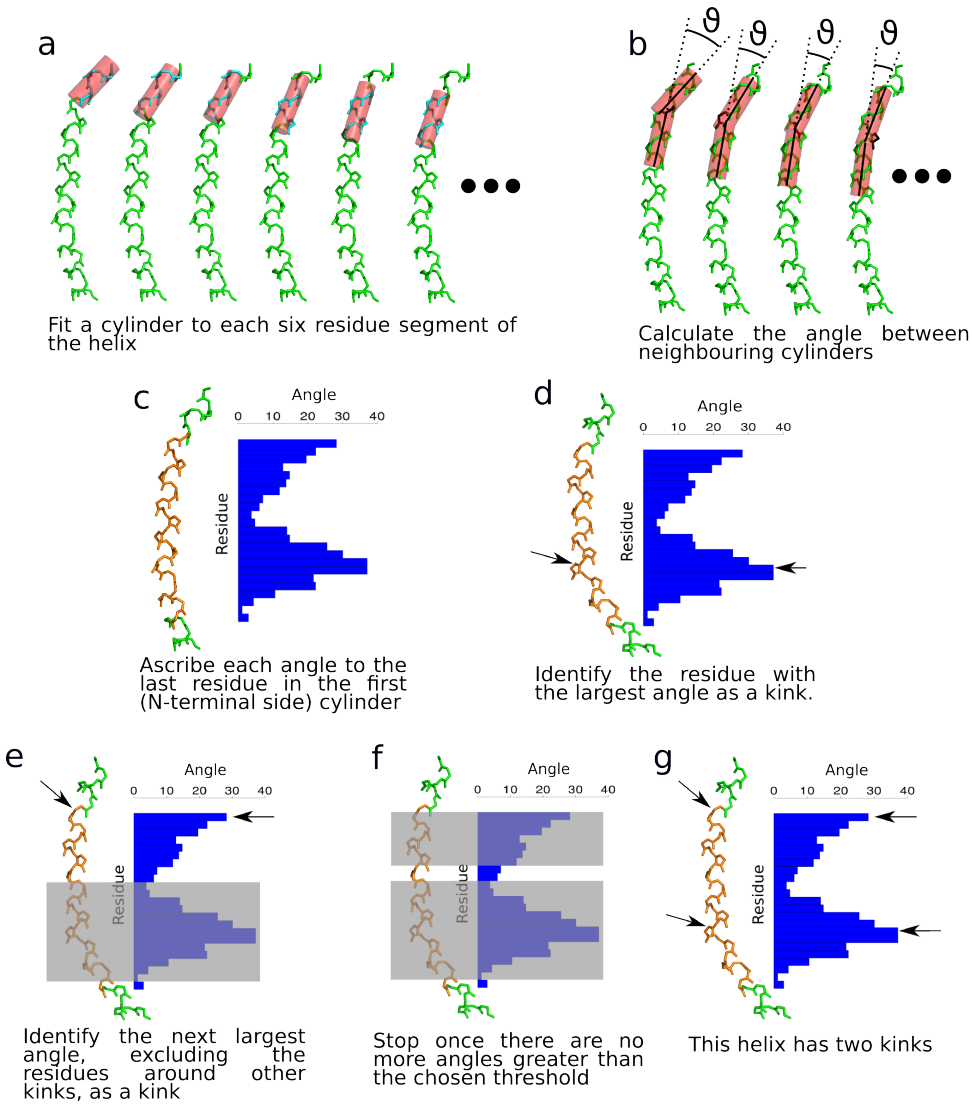


Figure 2.6: **Kink Finder algorithm.** (a) Axes are fitted to the backbone atoms of a sliding window of six residues. (b) and (c) These axes are used to calculate a local angle for each residue. (d) The residue with the largest angle is annotated as kinked. (e) Additional kinks are identified at residues where the angle is $\geq 10°$, the residue is at least four residues from another kink, and there is at least one residue with a local angle $\leq 10°$ between the residue and all other kinks in the helix.

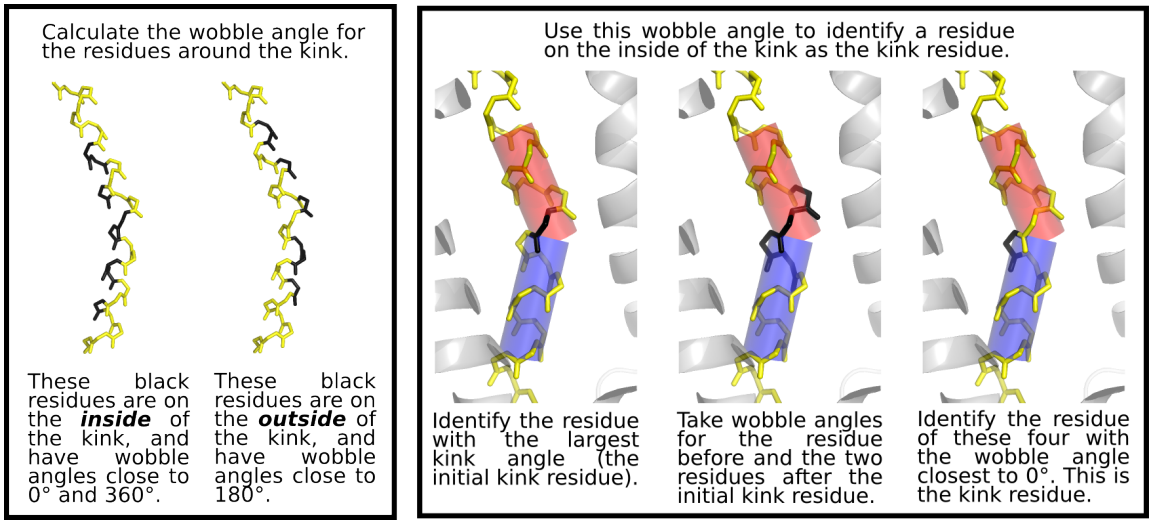


Figure 2.7: **Kink Finder's method to locate the kink residue.**

positive numbers towards the C-terminus. Some residues are consistently on the outside of the kink (e.g. residues -2 and $+2$), and some are consistently on the inside of the kink (e.g. residues -4, 0, and $+4$). This helps to clarify the residue patterns around kinks, and ensures that residue positions are consistently annotated with respect to the structure of the helix around kinks (for an example of the effect, see Figure 2.11).

Figure 2.12 shows the change in the amino acid percentage occupancies on correcting the kink position. The bias described above is hinted at in the uncorrected occupancies, but the patterns are smeared across several positions. For example, proline is seen at residues $+1$ to $+5$ in the uncorrected occupancies. Correcting the position concentrates the percentage occupancies - for example isoleucine (I) is more frequent at position $+4$, but less frequent at positions $+3$ and $+5$ in the corrected occupancies compared to the uncorrected occupancies. Similarly, the range of valine (V) occupancies increases from $8-12\%$ in the uncorrected kinks to $9-16\%$ in the corrected kinks, suggesting that this approach captures more information.

The distribution of maximum angle for each helix in the Kneissl et al. data set is shown in Figure 2.13a. This does not have the clear bimodality of the distribution of the B statistic. Instead it is a smooth distribution, with a shoulder around 20°, and another one around 35°.

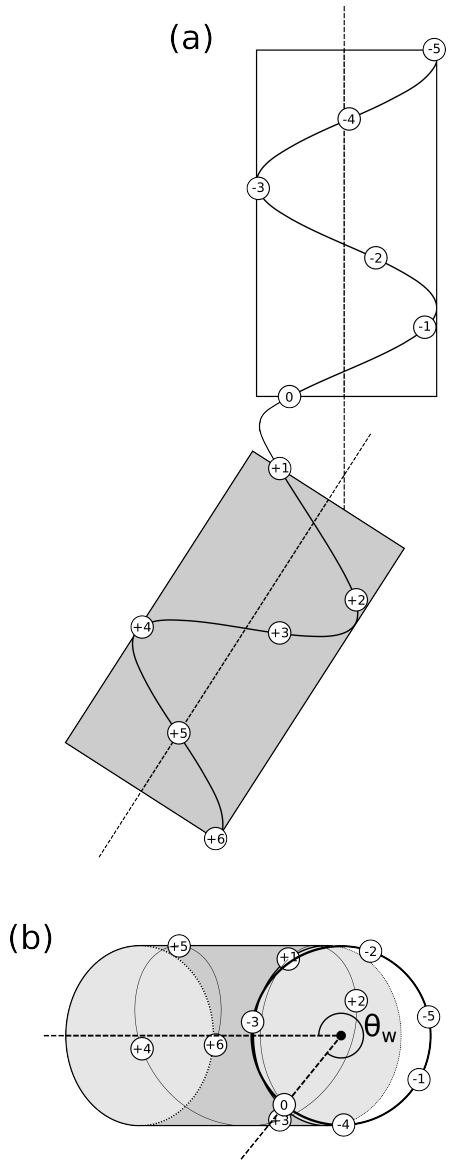


Figure 2.8: **Calculating the wobble angle**. (a) Representation of a kinked helix with fitted cylinders. Only C^{α} atoms are shown, numbered sequentially -5 (N-terminus) to +6 (C-terminus). The fitted axes are shown by dotted lines, and are in the plane of the page. The N-terminal cylinder is white, and the C-terminal cylinder is grey. (b) The same representation, rotated to look down the N-terminal axis. Atom 0 is in the plane of the page, and the C-terminal axis has been projected onto this plane. The wobble angle for each C^{α} atom is the angle between this vector, and the vector perpendicular to the N-terminal axis which passed through the C^{α} atom. θ_w is the wobble angle for atom 0. The wobble angle is used to select the kink residue, as shown in Figure 2.7.

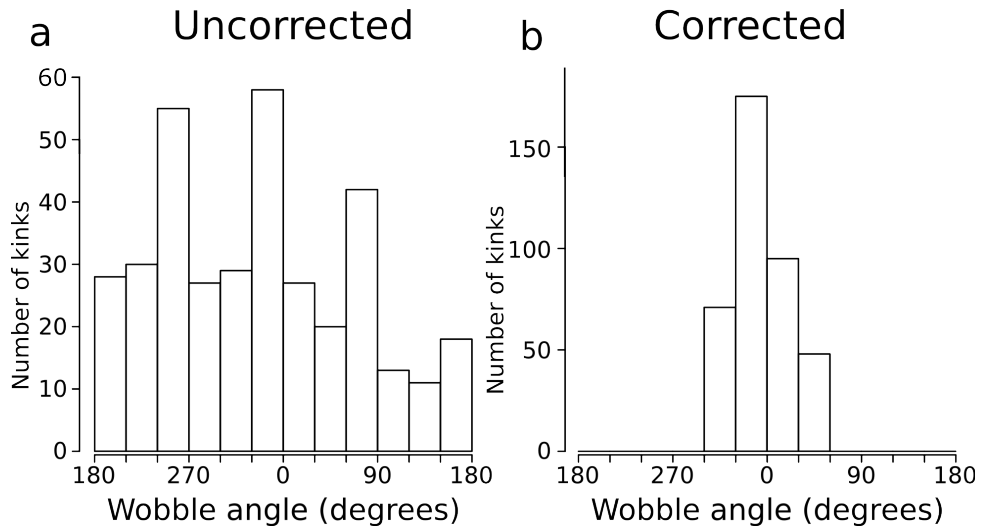


Figure 2.9: **Distribution of wobble angles for the uncorrected and corrected kink residues.** (a) In a kink, the uncorrected kink residue is the residue in the helix with the largest angle. (b) In a kink, the corrected kink residue is the residue close to the uncorrected kink residue that is most on the inside of the kink. The selection method for the corrected kink residue is fully described in Figure 2.7.

There is no clear location for a classification threshold between kinked and straight helices. Like the B statistic, the maximum kink angle is typically larger for longer helices (Figure 2.13b, $r^2 = 0.15$).

Figure 2.14 shows the effect of changing the kink angle threshold on the percentage occupancies. In a similar manner to the angle distributions, there is a smooth change as the angle threshold is changed - there is no sharp point where the occupancies change dramatically. As the angle increases, the occupancies become more noisy, due to the reducing number of kinks. Interestingly, glycine is most clearly present at position 0 at the lower angles, indicating that it may be more important in smaller kinks than larger kinks.

2.5 Discussion

In this chapter I described two methods that we developed to identify, locate and measure kinks in helices. I aimed to deal with two deficiencies in existing kink identification methods. I have

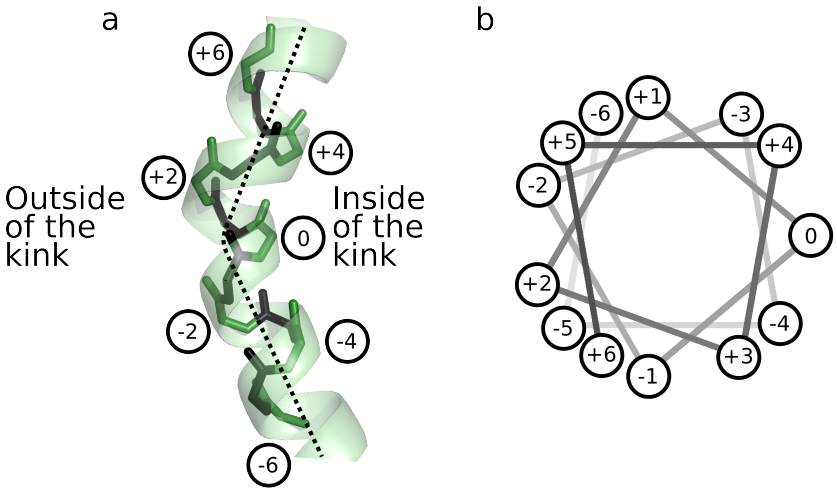


Figure 2.10: **Numbering of residues in kinks.** Position 0 is the kink residue, selected as the residue on the inside of the kink. This shows an ideal kink, where the wobble angle is 0°. The wobble angle of residue 0 can vary between −50° (between +3 and -4 in (b)) and +50° (between -3 and +4 in (b)) (Figure 2.9b). Consequently, the exact location of each position can vary from kink to kink. For example, while position 0 is always on the inside of the kink and position -2 is always on the outside, position +1 may be towards the inside of one kink and towards the outside of another.

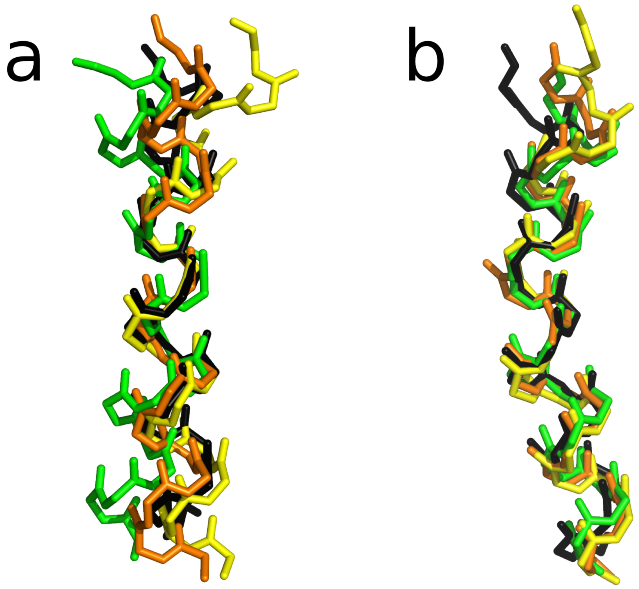


Figure 2.11: **Four example kinks identified by Kink Finder** aligned by: (a) the residue with the largest angle (the uncorrected kink residue), and (b) the kink residue (on the inside of the kink). For each figure, the alignment is of the four backbone atoms in a single residue (either with the largest angle or the kink residue). The alignment was obtained using a custom algorithm written in Python 2.7, which utilises the linear algebra module of the NumPy package (van der Walt *et al.*, 2011). This algorithm produces an alignment where the RMSD between the four backbone atoms in a pair of structures is minimised. The kinks were individually pairwise aligned to the kink from 2gfp. Only the backbone atoms are shown, and each helix is shown in a different colour. All atoms in a given helix are the same colour. The helices are coloured as follows: green is a helix from the protein with PDB identifier 2gfp, chain A, residues 17-36; yellow 2vpz, chain C, residues 120-139; orange 3ehz A 284-303; black 3k07 A 923-942.

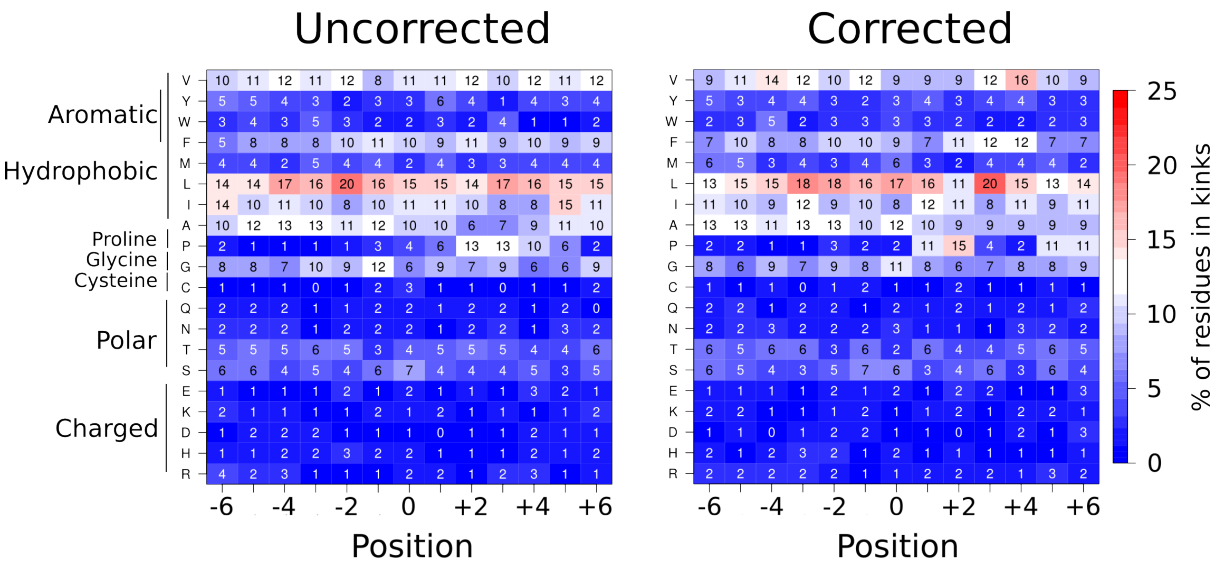


Figure 2.12: **Amino acid percentage occupancy for corrected and uncorrected kink positions.** Each box shows the percentage of kinks that contain that amino acid at that position relative to the kink residue. The uncorrected kink position is the residue with the largest angle in the helix, while the corrected kink position is a residue on the inside of the kink, close to the uncorrected kink position. The method used to identify the corrected kink position is described fully in Figure 2.7.

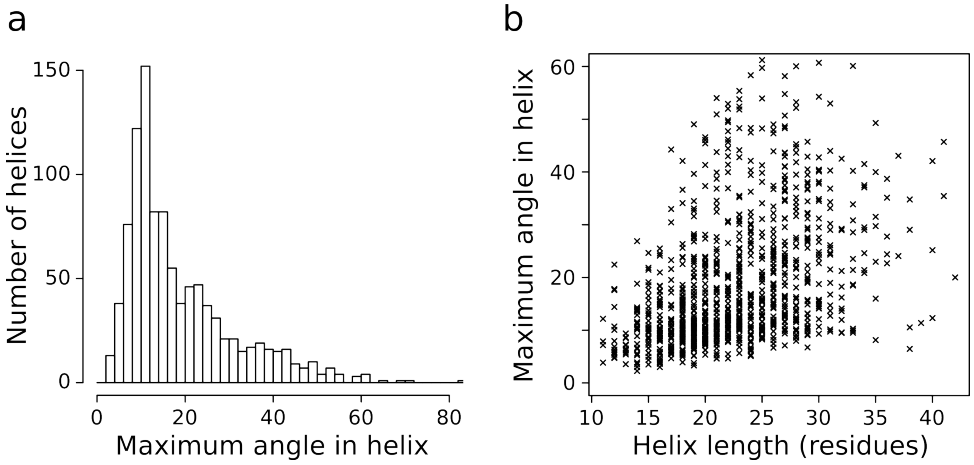


Figure 2.13: **Kink Finder maximum angles.**(a) Distribution of maximum kink angles in 1014 membrane helices, calculated by Kink Finder. (b) Variation of maximum kink angle with helix length, for helices analysed by Kink Finder

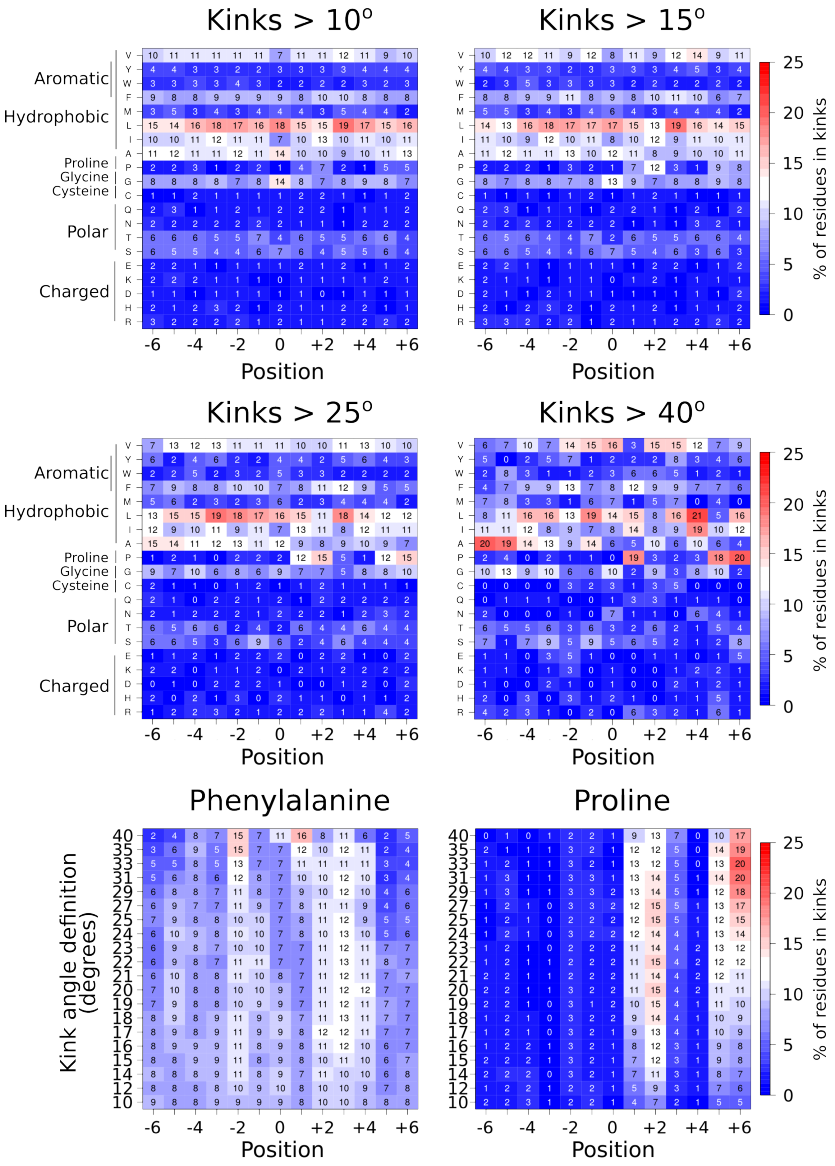


Figure 2.14: **Percentage occupancy of amino acids around kinks.** Each square represents the percentage of residues at that position relative to the kink that are of the type indicated. The threshold for inclusion as a kink is varied between 10° and 40°. The top four plots show how these vary for all twenty amino acid types when using 10°, 15°, 25°, and 40° cut offs, which have, respectively, 979, 552, 246, and 86 kinks. The bottom two plots show the variation for the amino acids phenylalanine (F) and proline (P) at the full range of kink angles.

achieved this by creating a statistically robust helix classifier (B statistic), and by designing a method to locate kinks consistently (Kink Finder). A module of Kink Finder that can estimate the error in measured kink angles is described in Chapter 5.

2.5.1 Decisions and parameters

The B statistic method provides a statistically robust method to classify helices, without the large number of parameters used in other methods. However, it does not locate kinks. I developed Kink Finder to locate kinks, but the disadvantage is that, like other methods, it has parameters that have to be subjectively set by the user. These include: the length of each helix section, the types of backbone atoms used in the cylinder fits, the generation of starting points for the cylinder fits, the number of residues from which the kink residue is chosen, and the angle threshold. Throughout my work, the parameters have been selected to provide best agreement with the expected annotations from visual inspection. Although changing these parameters had an effect on the results, it did not affect the conclusions that have been drawn from them. There are many methods, such as parameter sweeps and unsupervised learning, which can be applied to problems where the optimal parameters are unknown. However, these are not suitable in this case because there is no definitive benchmark set of kinks or kinked helices on which to evaluate a such an approach. Chapter 3 describes my crowdsourcing approach that yielded an objective human annotated training set of helices.

2.5.2 Using a binary classifier

In both the B statistic method and Kink Finder, we have only considered a binary classifier between kinked and not kinked helices. Although a binary classifier is an over-simplification of helix kinks, it provides the simplest and clearest classification.

2.5.2.1 B statistic classification threshold

The B statistic method, unlike all other kink identification methods, uses no pre-set parameters, providing a statistically robust classification method. The classification threshold is statistically derived from the data. We are not aware of any other method that does this. The B statistic discriminates clearly between two subsets of helices, particularly when helices that contain π-turns are removed. No other kink identification method produces a statistic with a bimodality. However, the kinked group that this method identifies is made up of fewer than 20% of the helices. Most other previous methods, including human annotation, have indicated that at least double this number of helices are kinked. So though perhaps statistically robust, the B statistic as it stands may not be biologically relevant.

2.5.2.2 Kink Finder classification threshold

The Kink Finder method does satisfy our aim for consistent kink location. However, it still requires researcher-led decisions to aid in the identification of kinks. The most important of these decisions are the angle threshold for classification between kinks and non-kinks, and the decisions that govern the fitting of local helix axes.

Unlike the B statistic, the maximum kink angle in each helix computed by Kink Finder has no clear boundary between two (or more) subgroups. Similarly, there is no clear step change in the amino acid patterns around kinks that indicates an obvious threshold. Studies have suggested that helices with angles as low as 13° should be classified as kinked (Hall *et al.*, 2009; Werner & Church, 2013), but a threshold in the region of 20° appears from Figure 2.13a to be more suitable. This classifies *c.* 40% of membrane helices as kinked, which is in good agreement with some other methods (e.g. Kneissl *et al.* (2011)). However, throughout this thesis I have varied the threshold to check that it does not affect the results presented.

2.5.2.3 Length of helix sections in Kink Finder

Kink Finder uses six-residue sections of helices to fit local axes. The length of the helix section is very important in kink identification methods. Short sections (3-4 residues), such as those used in Helanal, lead to a more sensitive method, but require a method to distinguish between kinks and other types of distortions, such as π-turns. Using cylinder fits the minimum length required for a reasonable approximation of the helix axis is six residues. Although longer sections provide more confident fits, they give the method less sensitivity. Unlike most methods (see Section 1.5.2), coordinates of all the heavy backbone atoms are used, rather than just those of the C^{α} atoms.

2.5.2.4 Cylinder fitting method

In conjunction with Prof. Mardia, I developed a more robust approach to the cylinder fitting algorithm of Åqvist (1986). This method is still a numerical approach, using a non-linear conjugate gradient method to optimise the function. Like many other minimisation algorithms, this relies on the function being minimised to be continuous and well behaved, and a suitable starting point being provided for the parameter values. Our improved approach resulted in a 5-fold reduction in the number of fits that failed to reach the global minimum, indicating that this is a better behaved formulation. The starting estimate is still crucial to calculated a good fit.

Initially, I used starting points from least-squares regression to 21 backbone atoms, which correspond to $1\frac{1}{2}$ helix turns. My experimentation with least-squares regression showed that good axis approximations (within $1-2^{\circ}$ of the true axis) are generally found when fitting to $n+\frac{1}{2}$ turn sections of helices. The true axis is not calculable, it is perhaps best described as the axis that would be calculated from a least-squares fit to the backbone atoms of an infinitely long, perfectly straight, and regular α-helix. Fitting to lengths between this number of turns (e.g. 2 turns, or seven residues) gives poor axis estimates (within $6-10^{\circ}$ of the true axis). Around helix distortions, 21 backbone atoms do not always correspond to $1\frac{1}{2}$ helix turns, so

the initial estimate can be poor. Nine-residue (36-atom, $2\frac{1}{2}$ helix turns) fits also give a good estimate of the helix axis, which we utilise as an alternative starting point, as well as 24- and 18-atom fits (6- and $4\frac{1}{2}$-residue), which correspond to $1\frac{1}{2}$ turns of a π and 3_{10} helix respectively.

Unlike the majority of other methods, Kink Finder uses all heavy backbone atom (N, C^{α}, C', O) coordinates to fit axes. This provides more information than just the C^{α} atom coordinates. However, the four heavy atoms in the backbone of each residue do not all lie the same distance from the helix axis. This means that the function landscape for a fit to all these atoms is less smooth than the landscape for just the C^{α}s, but the multiple starting points provide the global minimum in almost all cases. For the same reason, the goodness of fit, r, is rarely ≤ 0.3Å.

2.5.2.5 Kink location

The kink residue is chosen to be on the inside (as opposed to the outside) of the kink, as that was where the initial kink residue was most frequently found (Figure 2.9). As will be described in Chapter 3, most other methods also identify the kink residue more frequently on the inside of the kink than on the outside of the kink.

Many studies describe the effects of particular residues on kinks, but in reality, it is only meaningful to study these patterns if the kink is positioned consistently with respect to the shape of the helix. Figure 2.11 shows four kinks, aligned by their kink residues, both the initial kink residues (those residues with the largest angles) and the kink residues on the inside of the kink. It is clear that selecting the kink residue in this way means that the residues at a specific position (see Figure 2.10) are consistently in the same position relative to the shape of the kinked helix. This means that, unlike in other kink identification methods, the positions relative to the kink residue have meaning for Kink Finder.

As shown in Chapter 4, the outside of kinks (i.e. positions -2, +2 etc.) tends to point into the solvent (or membrane). Some of the patterns around kinks identified using Kink Finder's method of kink residue selection could be due to the effect of being in contact with the solvent or the rest of the protein. However, the bias already exists in the uncorrected initial kink position,

and so needs to be accounted for. This feature has, to my knowledge, not been reported by any other researchers.

2.5.3 Conclusion

In this chapter I have described the novel methods used in this research to identify, locate, and measure helix kinks in protein structures. The B statistic provides a statistically robust method to classify helices that have kinks, whilst Kink Finder provides a consistent method to locate kinks relative the the local helix structure. My method to estimate error in kink angles is described in Chapter 5. These features are applied to the characterisation of kinks in the following chapters. In the next chapter, these two methods are compared to annotations from other methods and from humans.

CHAPTER 3

Comparison of kink finding methods and the development of kink identification by crowdsourcing

The majority of the work in this chapter was published as an article in 2014. Much of the text and the figures in the article are used here. These are reprinted (adapted) with permission from 'Crowdsourcing yields an new standard for protein helix kinks' (Wilman, Ebejer, Shi, Deane & Knapp, 2014a). Copyright 2014 American Chemical Society. I wrote the article and undertook

the analysis for the paper with input from the co-authors. Bernard Knapp, Jean-Paul Ebejer, and I worked together on the visual development and testing of the crowdsourcing application, AHAH (Alpha-Helices Assessed by Humans), while Jean-Paul was responsible for the technical implementation of the web application. I, and my co-authors, are indebted to all those who participated in the study.

3.1 Introduction

Current kink definition and identification methods often disagree with one another. In this chapter, I show how these differences remain between methods even when the variation in data set, thresholds and choice of kink residue are removed. Using the same data set, with equivalent thresholds, and a consistent method to identify the kink residue improves helix annotation agreement between the methods from 50% to 65%, and the number of helices where there is exact agreement increases from 0 to 33.

To see if a more consistent kink set could be built, we developed a crowdsourcing approach to the problem. Using an online interface, we collected more than 10,000 classifications of 300 helices into straight, curved, or kinked categories. My analysis of the results found that participants were better at discriminating between straight and not-straight helices than between kinked and curved helices. Surprisingly, more obvious kinks were not necessarily identified as more localised within the helix. We published a set of 252 helices where more than 50% of the participants agree on a classification. This set can be used as a reliable gold standard to develop, train and compare computational methods. An interactive visualisation of the results was made available online at `http://opig.stats.ox.ac.uk/webapps/ahah/php/experiment_results.php`.

3.1.1 Existing kink finding methods

There are a large number of possible computational approaches to identifying kinks (Section 1.5.2, and Chapter 2). Several algorithms that identify kinks in proteins have been published, in addition to my method, Kink Finder (Bansal *et al.*, 2000; Kneissl *et al.*, 2011; Kumar & Bansal, 2012; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Visiers *et al.*, 2000). The conclusions of studies of kinks have often been contradictory. There is no consistent method to identify the kink residue among the methods.

In this chapter I describe a human rather than computational approach to the problem. Kneissl *et al.* (2011) took a human approach to the problem. They relied on two researchers to annotate kinks in a set of 1014 membrane helices. Crowdsourcing provides a better approach for annotation than a single researcher, as it is not biased by the opinions of a small number of people. Many crowdsourcing approaches have recently been used to solve computationally difficult problems in a wide variety of scientific fields (Good & Su, 2013).

3.1.2 Crowdsourcing science

Crowdsourcing is a technique that employs the human intelligence of a large number of participants to solve computationally challenging problems (Good & Su, 2013; Parvanta *et al.*, 2013; Ranard *et al.*, 2014). Galaxy Zoo, Foldit and Phylo are three relevant examples of recent of crowdsourcing approaches.

The Galaxy Zoo project (Land *et al.*, 2008; Skibba *et al.*, 2012) uses volunteers to classify images that would otherwise take individual researchers years. This has expanded, in the form of the Zooniverse project (Zooniverse Team, 2014), to include projects in cosmology, climate, humanities, nature and biology, with a current combined output of 69 papers[1]. Many of the Zooniverse projects use large numbers of volunteer participants to provide classifications of images that normally deviate from the ideal definitions, and often have characteristics of more than one ideal definition. This makes it very similar to the kink identification problem.

[1] As of September 2014

In the field of bioinformatics, Foldit (Cooper *et al.*, 2010; Khatib *et al.*, 2011) invites volunteers to fold protein structures via an interactive game. This sees human intelligence exploring the conformational space of proteins using the same scoring function as in the Rosetta protein structure prediction algorithm (Simons *et al.*, 1997). Phylo also uses an online game, but in this case to solve difficult alignment problems (Kawrykow *et al.*, 2012). In both of these examples, crowdsourcing harnesses the human ability to make large, helpful, moves that computational methods are unable to identify.

A common characteristic of these, and other examples, is the simplicity of their interfaces. The crowdsourcing approach does not rely on every user providing a 'correct' or nuanced response. Instead, many simple classifications provide as much, if not more, information as a detailed classification by an expert. For instance, classification certainty can be calculated from the degree of consensus between respondents. Consequently, the aim is for simple, fast, interfaces, to allow as many responses as possible.

All crowdsourcing experiments require participants. Approaches to get participants and annotations include volunteers (Land *et al.*, 2008), gamification (Cooper *et al.*, 2010), payment (Nguyen *et al.*, 2012), and forced labour (e.g. reCAPTCHA (von Ahn *et al.*, 2008)). Gamification is the use of game-like mechanics, such as competition through scoring, in order to encourage participation. Our approach, Alpha Helices Assessed by Humans (AHAH), relied on volunteers, and contained a small level of gamification.

3.2 Materials and methods

3.2.1 Kink identification methods

I compared the results of our crowdsourcing approach against four previous approaches. These were Kink Finder (Wilman *et al.*, 2014b), MC-Helan (Langelaan *et al.*, 2010), Helanal-Plus (Kumar & Bansal, 2012), and the Kneissl annotation (Kneissl *et al.*, 2011), which are described in detail in Sections 1.5.2 and 2.4. These methods use different terminology, i.e. 'kink', 'bend',

and 'distortion', and 'straight', 'linear', and 'good'. In this chapter, I treat bent and distorted (MC-Helan) as synonymous with kinked. Linear (Helanal-plus) and good (MC-Helan) are synonymous with straight.

These methods have different thresholds for identifying helices as kinked, and the kink residue (the precise position of the kink in the helix). Kink Finder uses a maximum angle of 20° as the threshold to classify a helix as kinked. MC-Helan classifies any helix that does not satisfy its 'ideal helix' definition as kinked, with no angle threshold. Helanal-Plus classifies all helices with maximum angles $\geq 30°$ as kinked, and some helices with maximum angles $\geq 20°$. The Kneissl annotation relied on the researchers' own judgement, both for the classification of the helices, and the choice of kink residue.

Kink Finder identifies a residue on the inside of the kink as the kink residue. MC-Helan selects a kink residue using a voting procedure, based on the inter Cα distances, inter Cα angles, and ϕ and ψ angles of the residues close to the boundary between helix sections. Helanal identifies the residue with the largest angle in the helix as the kink residue.

3.2.2 Comparison between methods

The number of helices annotated as kinked by the methods was equalised by changing their parameters, so that the number of kinked helices identified by each method was 357, the same as for the Kneissl *et al.* (2011) annotation. After equalising the number of helices identified as kinked by each method, helices were kinked if Kink Finder identified a maximum angle $\geq 19.9°$. For MC-Helan, the helices were kinked if MC-Helan had identified them as bent or distorted, and their bend angle was $\geq 16.0°$. Helanal-Plus identified kinked helices if it had identified them as kinked, or their maximum kink angle was $\geq 22.2°$.[1]

The choice of the kink residue has an effect on the patterns observed around a kink. Each of the methods use a different approach, but only Kink Finder uses a consistent method that means that the positions relative to the kink residue have meaning (Section 2.5.2.5). For each

[1]This yielded 356 kinked helices for Helanal-Plus, as there was a three-way tie in the angle for the 357^{th} smallest kink.

of the four methods, the kink position was corrected using the same method as for Kink Finder, taking the initial kink position as identified by the method, and using axes fitted by Kink Finder (see Figure 2.7 in Chapter 2).

3.2.3 Amino acid propensities

Amino acid propensities were calculated from the number of residues at a specific position in kinks compared to the number of each amino acid type in all of the helix set. The equation for propensity is given in Equation 3.1;

$$\text{Propensity} = \text{P}_i^a = log_e \left(\frac{N_i^a / N_i^{\text{all}}}{N_{\text{helices}}^a / N_{\text{helices}}^{\text{all}}} \right) \tag{3.1}$$

where N_i^a is the number of residues (N) of a particular amino acid type (a), e.g. glycine, observed at a particular position relative to the kink (i). N_i^{all} is the total number of residues observed at a particular position relative to the kink. $N_{\text{helices}}^{\text{all}}$ is the total number of amino acids (all) observed in the helices within the data set. The background distribution ($N_{\text{helices}}^a / N_{\text{helices}}^{\text{all}}$) comes from the set of 1014 helices.

3.2.4 Helix data set

For the comparison of four existing annotation methods, I used the set of 1014 helices manually annotated by Kneissl *et al.* (2011). For the AHAH crowdsourcing, we randomly selected 300 of these helices, while maintaining the same proportion of kinked, curved, and straight helices as annotated by Kneissl et al. These helices are taken from the MPtopo database (Jayasinghe *et al.*, 2001). There is no redundancy threshold on the proteins, but no pair of helices has more than 95% sequence identity. The resolution of the protein structures varies between 1.60Å and 4.50Å, and 60% have resolution less than 3.00Å.

Table 3.1: **Participant backgrounds.** Each value shows the proportion of participants with the given background. Participants were able to indicate multiple backgrounds. The 'None' category includes 77 (24.8%) school pupils.

Background	% of participants
Structural Biology	14.8
Chemistry/Biochemistry	22.9
Physics/Maths	18.7
Computer Science/I.T.	24.8
Other Science	20.0
Other non-science	11.0
None	27.4

3.2.5 AHAH technical implementation

The crowdsourcing application, AHAH, was implemented as a web application using PHP (version 5.4.4) for middleware and MySQL (version 9.1.13) for the database backend. Results were analysed using Python (2.7.5) (van Rossum & de Boer, 1991) and R (3.0.2) (R Core Team, 2014). A schematic diagram of the participants interaction with the web application is shown in Figure 3.1.

3.2.6 Helix data representation

The helices were displayed to participants using the JSmol viewer[1], in a Cα only ribbon representation (Figure 3.2a and b). Residues were coloured by their type, and participants were able to rotate them freely and zoom in. The helices rotated slowly by default.

3.2.7 Participants

A total of 310 people with diverse backgrounds (Table 3.1), and education levels (Table 3.2) registered and took part. Each participant annotated a minimum of 30 randomly selected helices.

[1] JSmol: an open source HTML5 viewer for chemical structures in 3D. http://wiki.jmol.org/index.php/JSmol#JSmol.

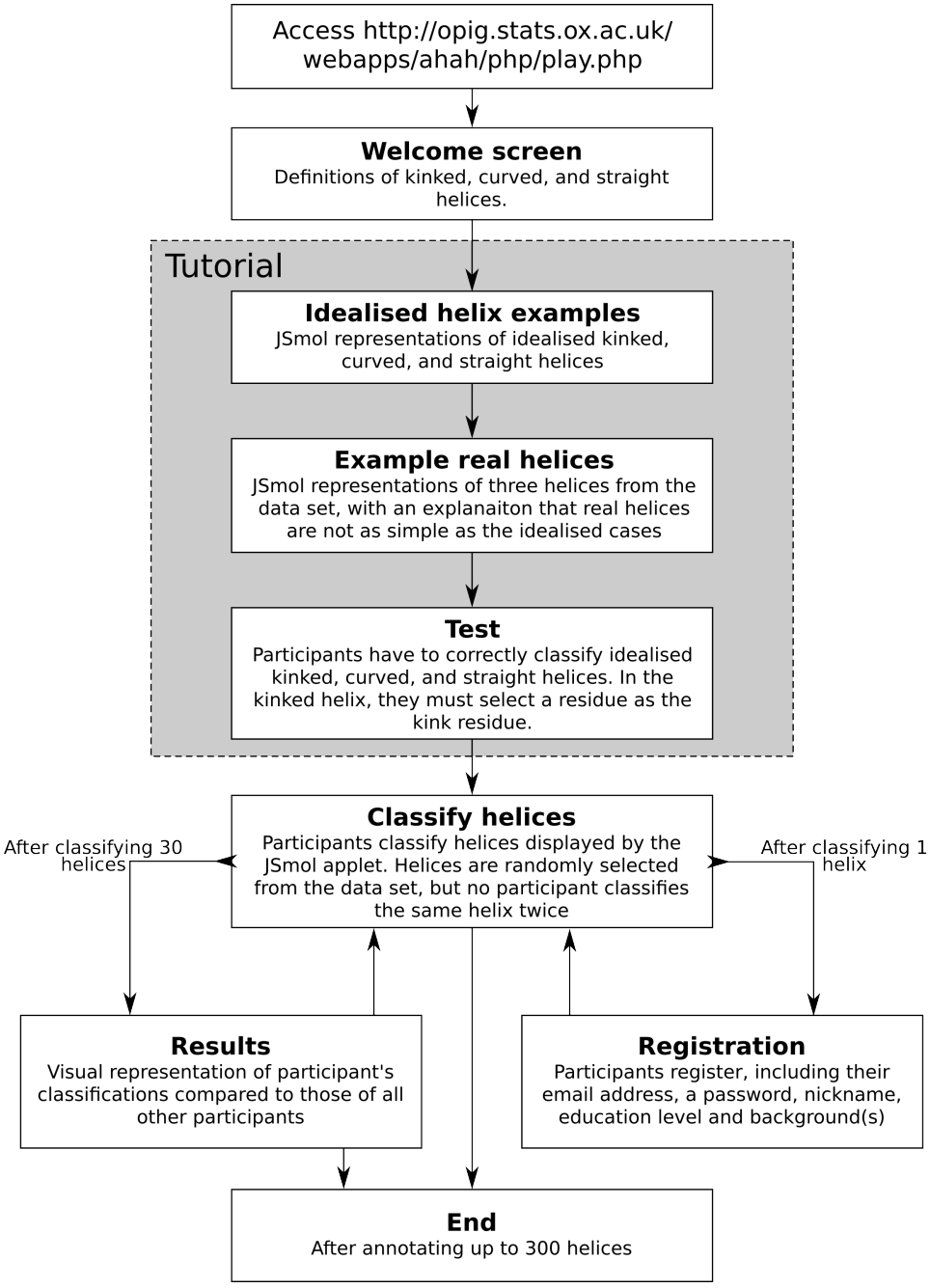


Figure 3.1: **AHAH web server schematic.**

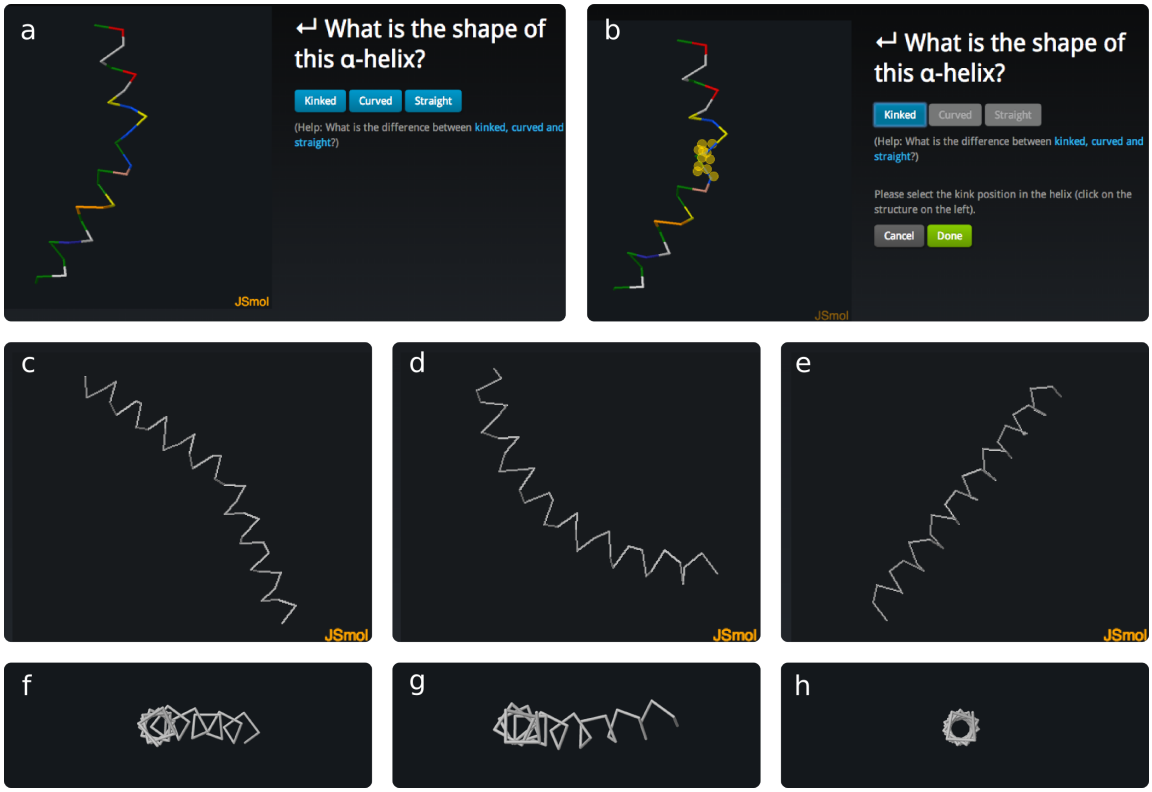


Figure 3.2: **Representation of helices in AHAH.** (a) AHAH user interface. Participants identify the helix as kinked, curved, or straight. (b) Where they identify a helix as kinked, they select a residue at the kink position. (c)-(h) Idealised helices shown to participants in the tutorial. (c) and (f) are two representations of a kinked helix, (d) and (g) are two representations of a curved helix, and (e) and (h) show a straight helix.

Table 3.2: **Participant education levels.** The 'Other' category includes 77 (24.8%) school pupils.

Education level (Achieved to date)	% of participants
Post-doctoral and above	17.7
Undergraduate and above	48.4
Secondary	5.2
Other	26.1
None indicated	2.6

3.2.8 Training of participants

Participants were trained in three stages. First, written descriptions of the classification definitions were given, along with idealised examples of each helix type (Figure 3.2c-h). The definitions were as follows.

- Kinked - 'There is a clear location where the direction of the helix changes. Only a small part of the helix is involved in this.'
- Curved - 'There is a slow but steady change of the direction of the helix. This can happen over a large part or even all of the helix.'
- Straight - 'There is no change in the overall direction of the helix.'

Second, examples of three real helices were illustrated. In addition, we stated that such real helices might be more ambiguous than the idealised examples shown previously. Finally, participants had to annotate three idealised helices correctly before they could continue. For the idealised kinked helix, this includes identifying a residue close to the site of the kink. This step attempted to ensure that only those participants who understood the concept of kinked, curved and straight could participate in the project.

3.2.9 Crowdsourcing survey

After successfully completing the tutorial, participants could annotate one helix before being required to register. Participants were then asked to classify 30 randomly selected helices. After 30 annotations, participants were thanked for participating, and shown a comparison between their responses and those of other participants. The participants were then given an option to either stop annotating helices or continue until a maximum of 300 helices. Each helix was shown only once to the respective participant, and corrections after the initial assessment were not possible. Participants were not able to skip a helix.

For each helix, participants were shown the helix, displayed in JSmol as described above, and had to indicate if it was kinked, curved, or straight by clicking the appropriate button

(see Figure 3.2a). If the participant indicated that the helix was kinked, then he/she had to select a residue as the kink point, by clicking on it in the JSmol viewer. This residue was then highlighted, and they could change their choice of kink residue, or confirm it by clicking a button (see Figure 3.2b).

3.2.10 Response consistency

In order to assess the annotations I considered response consistency. The consistency of a subset of results was calculated by taking the number of annotations that agreed with the majority view. Annotations were divided into subcategories based on features such as the time taken, the backgrounds of the participant, and education level of the participant. The time taken for each annotation was calculated as the difference between the time at which the annotation was made, and the time at which the previous annotation was made. There were 138 annotations that took less than 2 seconds which I removed from the analysis. These were outliers with respect to time taken, and were disproportionately provided by a small number of participants (79 from the same five participants).

3.2.11 AHAH kink positions

For the AHAH analysis, the kink residue for a kinked helix was the residue most frequently identified as the site of the kink by the participants. Where there was a tie, out of the tied residues the closest residue to the mean position was chosen. If this still failed to separate them, the residue closer to the N-terminus of the helix was chosen.

3.3 Results

The agreement between the four existing kink identification methods, Kink Finder, Kneissl *et al.* (2011), MC-Helan, and Helanal-Plus is relatively low. Figure 3.3 shows a Venn diagram of the helices that are identified as kinked by the four methods. The methods agree on only half of

the helices (517 out of 1014). The percentage of kinked helices varies from 24.1% to 67.7%. Of the 201 helices that are classified as kinked by all four methods, there are none where all four approaches identify the same kink residue within the helix.

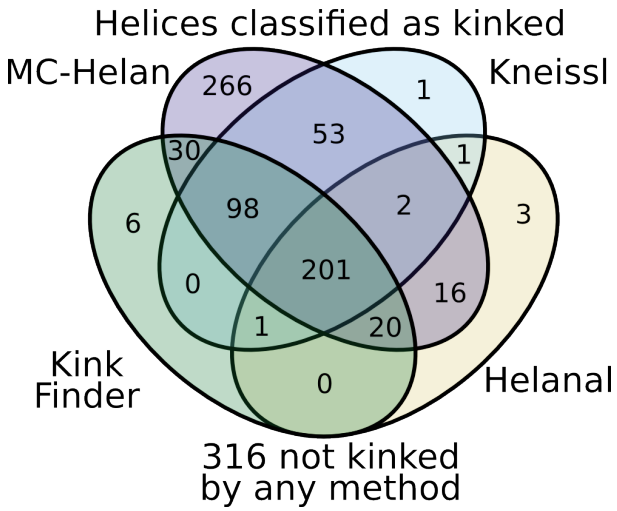


Figure 3.3: **Number of helices classed as kinked by four helix classification methods.** All the methods agree in only 517 (201 kinked, 316 not-kinked) of 1014 cases.

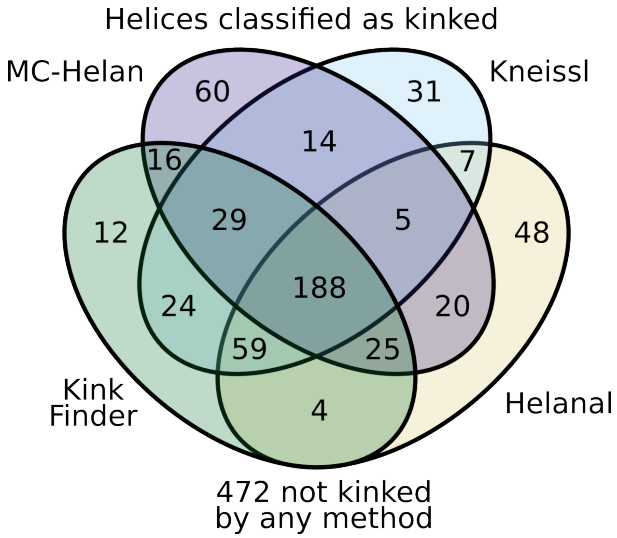


Figure 3.4: **Number of helices classed as kinked by four helix classification methods, when their thresholds are changed to each annotate the same number of kinked helices (357).** All the methods agree in 660 (188 kinked, 472 not-kinked) of 1014 cases.

As the methods each identify a different proportion of helices as kinked, they cannot agree on every helix. The thresholds for each of the methods were changed such that all methods annotated the same number of helices as kinked[1] (see Section 3.2.2). This improves the agreement, the methods now agree in 660 (65%) cases (Figure 3.4), as opposed to 517 (51%). However, there are now fewer kinked helix agreements and a large increase in the number considered unkinked by all methods.

In the 188 conserved kinked helices, there are none where all the methods agree on the kink position. Correcting the kink position, to a nearby residue on the inside of the kink, using the method in Kink Finder (Figure 2.7), further improves the consistency of the methods. When the kinks of each method are repositioned, there are 33 kinked helices which have the same kink position by all four methods.

3.3.1 Features of kinks identified by all methods

The existing kink identification methods have identified contradictory patterns around kinks (see Section 1.5.3). When I adjusted the thresholds in the methods so that they identify the same number of kinks, and used the same method to identify the kink residue, the four methods show much more similar amino acid patterns.

The effect of these adjustments is shown in Figure 3.5. The uncorrected propensities (left) show some similarities, however, other than proline being favoured in the positions after the kink, very few residues are over-represented at the same positions in the propensities of all four methods. However, the corrected propensities are much more consistent. For example, in the corrected propensities for all four methods, proline is favoured at positions +1 and +2, and disfavoured at +4. Glycine is favoured at position +1, valine at +3 and +4, phenylalanine at +2 and +3, serine at -1, and tryptophan at -4.

[1] 357 and 356 in the case of Helanal

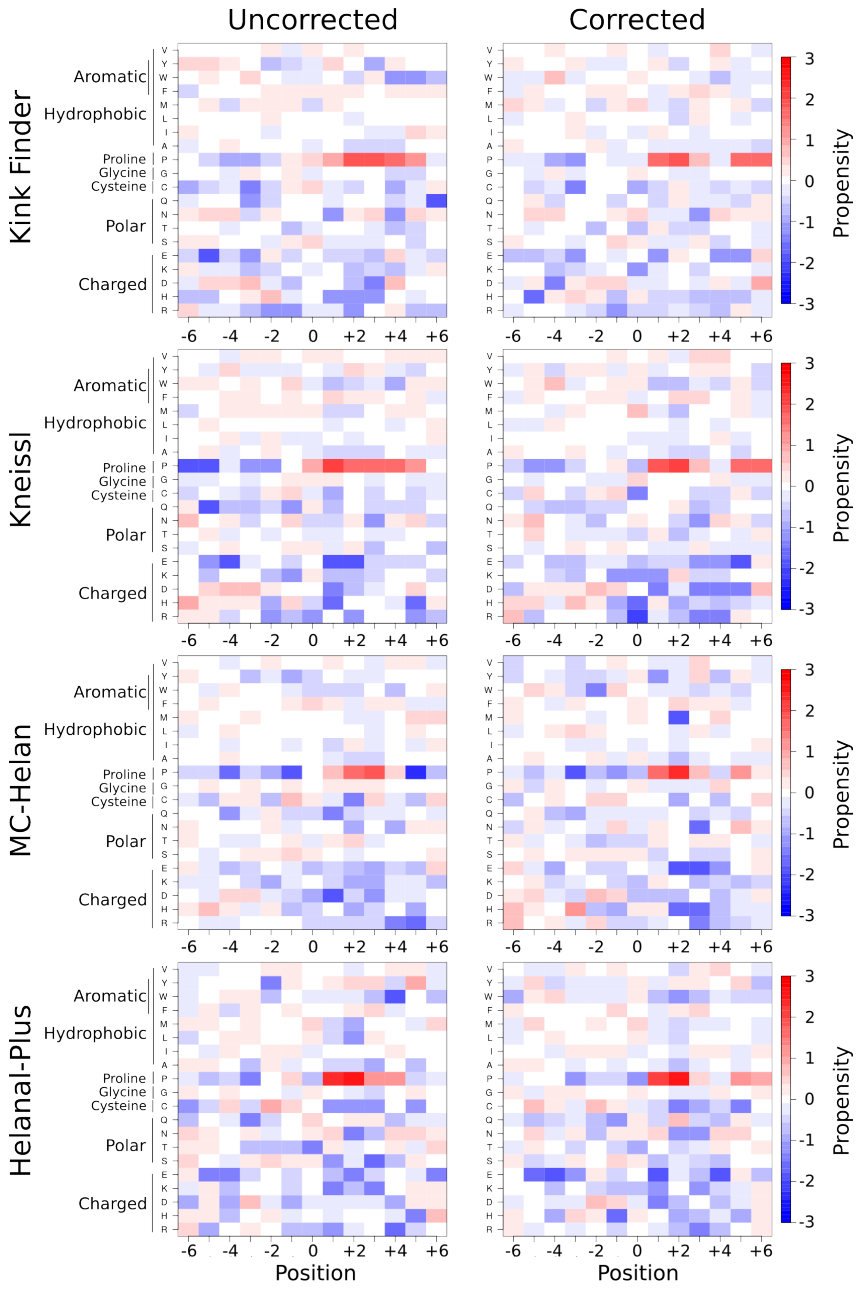


Figure 3.5: **Amino acid propensities around kinks as identified by the four algorithms.** Left are propensities of the amino acids around the kinks as identified by each method. Right are the propensities for the corrected kinks, where the thresholds have been changed to give 357 kinks, and the kink position is corrected (see Section 3.2.2). Red (positive propensity) indicates that the residue is observed more frequently than in helices on average, and blue (negative propensity) indicates that the residue is found less frequently at that position than in helices on average.

3.3.2 Crowdsourcing

As shown above, the problem of classifying helices is computationally difficult and ambiguous and there is no objective basis to identify the correct or ideal threshold(s). In order to address this problem we built AHAH (Alpha Helices Assessed by Humans), a crowdsourcing experiment to identify helix kinks.

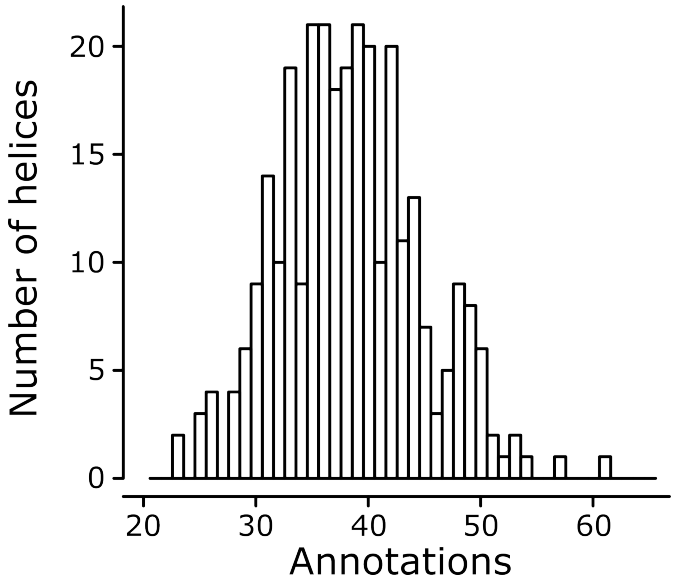


Figure 3.6: **Number of annotations for each helix in AHAH.** The histogram shows the number of helices that have been annotated by the corresponding number of participants. For example, there were nine helices annotated by 30 participants.

3.3.3 AHAH participants and annotations

A total of 310 participants registered to take part in AHAH (Tables 3.1 and 3.2), of which 290 annotated at least thirty helices. In addition, there were a further 928 annotations by unregistered participants. This yielded a total of 10,665 helix annotations. The average number of annotations was 35.6 per helix and 34.4 per registered participant. Figure 3.6 shows the distribution of the number of annotations for each of the helices.

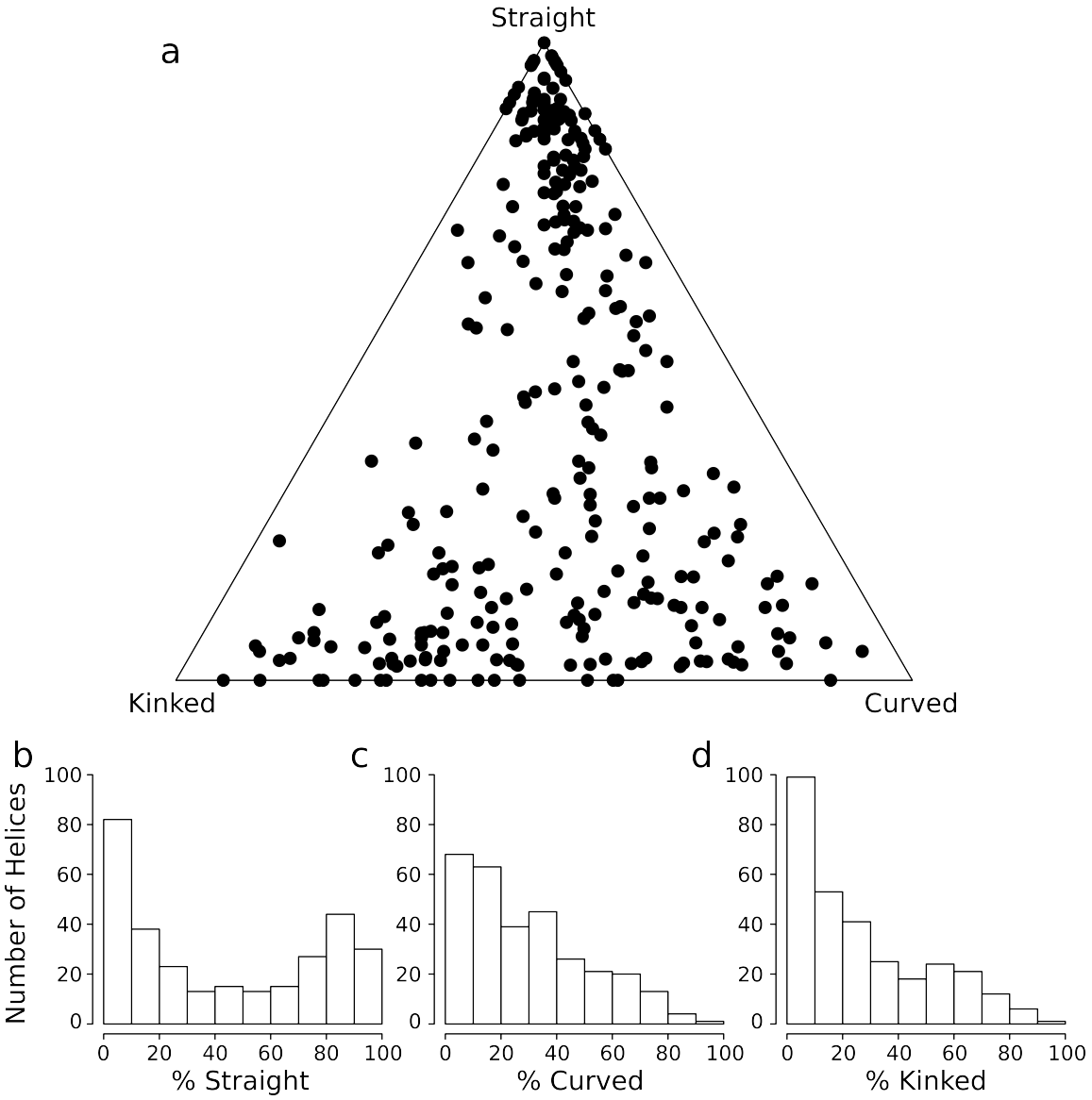


Figure 3.7: **AHAH helix classifications.** (a) Ternary diagram showing the AHAH responses for each helix. Each point represents a helix. The perpendicular distance from each side is the proportion of responses of each type. So those points on the bottom line represent helices that were not annotated as straight by any participants, and those points in the top corner represent helices that were annotated as straight by all participants. (b-d) The frequency of helices with a given percentage of straight (b), curved (c), and kinked (d) responses.

3.3.4 Helix classification by our participants

Figure 3.7 illustrates the responses of our participants for each helix. It shows that many helices do not fall obviously into one of the three possible classifications. For only 86 (28.7%) of the 300 helices is the majority view held by more than 80% of the participants who classified it, for 194 (64.7%) the majority view is held by $\geq 60\%$ of participants, and for 252 (83.7%) the majority view is held by $\geq 50\%$ of participants.

It is not clear from previous studies if a binary or tertiary classification of helices is most appropriate (Kneissl *et al.*, 2011; Kumar & Bansal, 2012; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011). Of the possible binary classifications (straight/not-straight, curved/not-curved or kinked/not-kinked), straight/not-straight provides the clearest separation. There are 191 (63.7%) helices where more than 80% of participants agree on a straight/not-straight classification (Figure 3.7b), compared to 134 (44.7%) for curved/not-curved (Figure 3.7c), and 155 (51.7%) for kinked/not-kinked (Figure 3.7d). This indicates that it is easier for people to identify if helices are straight or not, but more difficult to differentiate between kinked and curved.

3.3.5 Consistency of AHAH responses

The ambiguity of the results could be caused by poor quality responses from subgroups within our participants. I tested this hypothesis in two ways. First, I investigated if the consistency of a response is related to the time taken by the participant to make that classification. Second, I compared the consistency of participant groups, to ensure no group is significantly worse than any other.

Divergence from the majority view did not vary with time taken to classify a helix (Figure 3.8). There were a small number of annotations that took under two seconds, which were less consistent. As described in the methods, these were removed from the analysis. The annotations from unregistered participants were only slightly less consistent than those of the registered participants.

I divided the participants into background and education level groups based on their re-

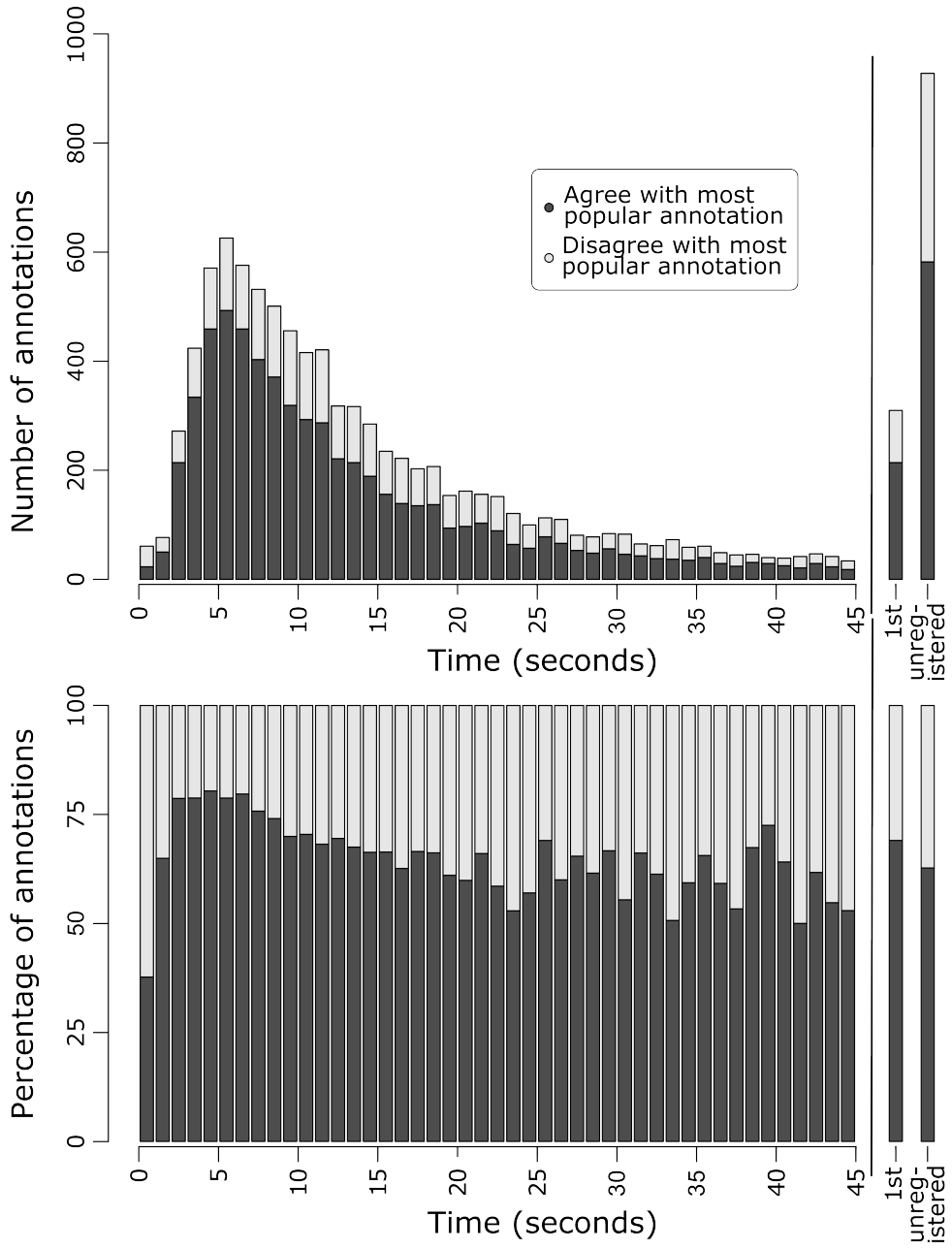


Figure 3.8: **Helix annotations grouped by time taken**, expressed as counts (top) and percentages (bottom). The right-hand-most bars show the first annotation by each participant, and those annotations by unregistered participants. Dark grey indicates annotations that agree with the majority view, and the pale grey bars indicate annotations that disagree with the majority view. The time taken for a participants n^{th} annotation was taken as the difference between the timestamp of their n^{th} and $(n-1)^{th}$ annotation. No time is calculated for the first annotation by each participant.

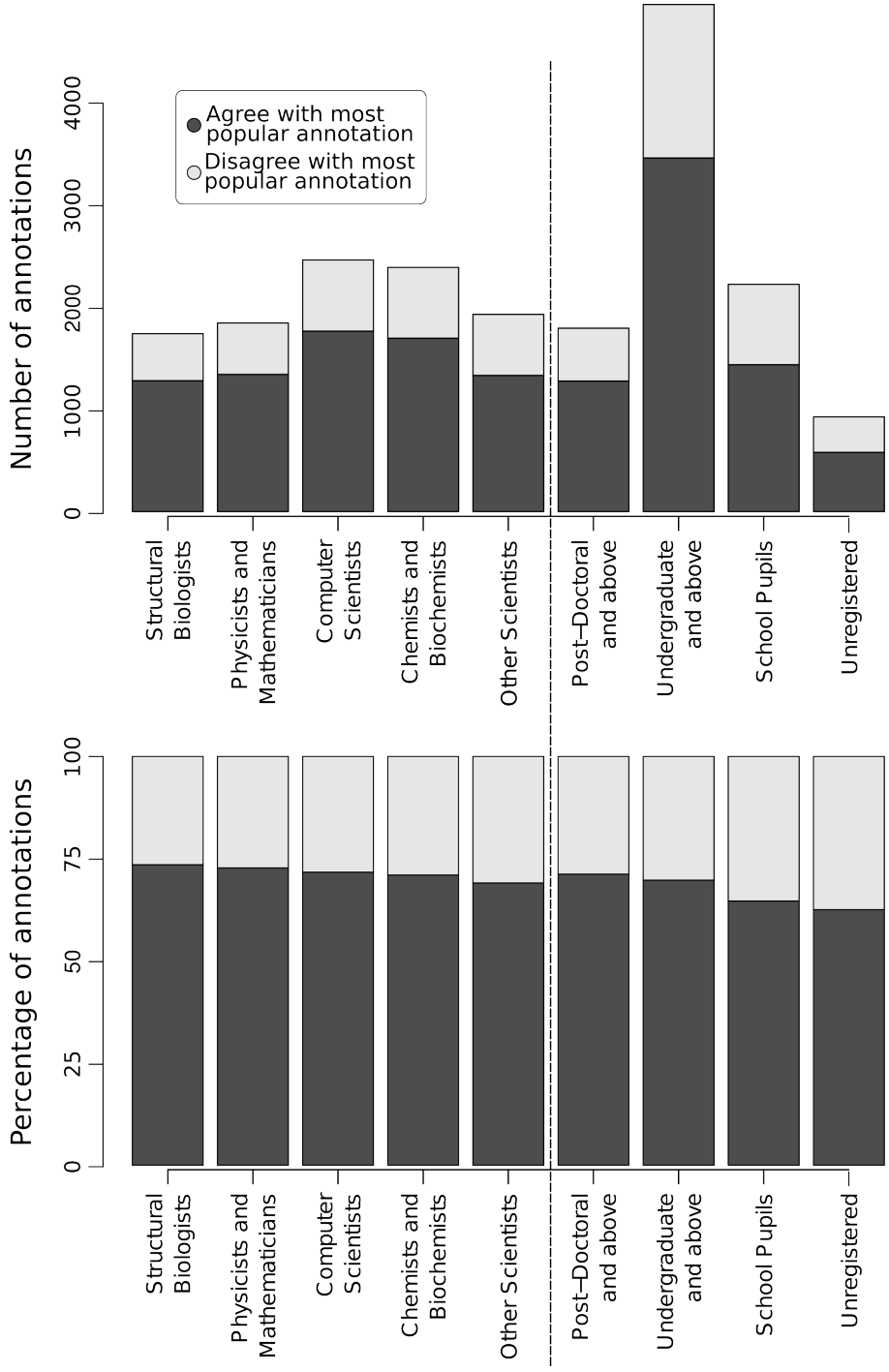


Figure 3.9: **The consistency of annotations from selected groups of participants.** Groups by background are on the left, and groups by education level are on the right. Each bar shows the number (top) or percentage (bottom) of annotations by the participants within the group that agree with the majority view.

sponses during registration. Although there is some difference between the consistency of the groups, no group is sizeably worse than any other. Structural Biologists appear to be the best group (73.5% agreement with the majority view) and Other Scientists the worst group (69% agreement) in terms of backgrounds (Figures 3.9). Dividing the data by education level, the Post-Doctoral and above group has the best agreement with the majority view (71.2% agreement), whereas the School Pupils have the worst agreement (64.6% agreement), except for the unregistered annotations (62.5% agreement).

The differences between the groups of participants are small. Therefore I excluded only those annotations that took less than two seconds, but kept all other data.

3.3.6 Comparison of AHAH with other methods

In Figure 3.10 I compare the results of our crowdsourcing approach with other methods from the literature. Two of these methods (Kink Finder and MC-Helan) only split helices into straight and kinked groups, while Kneissl and Helanal-Plus divide helices into kinked, curved and straight groups. None of the individual methods have strong agreement with the AHAH participants.

The Kink Finder and the Kneissl classifications are more consistent with the crowd-sourced classifications than the other two methods (Figures 3.10a and 3.10b). Kink Finder correctly annotated all but one of the helices which $\geq 60\%$ of participants classified as kinked. There are, however, a number of helices towards the curved corner that Kink Finder classified as straight (Figure 3.10a). The Kneissl annotation broadly agrees with the crowd-sourced data, although helices classified by Kneissl as curved are distributed across most of the ternary diagram (Figure 3.10b). There are two helices where more than 90% of participants disagree with the Kneissl annotation, and 14 where more than 80% disagree.

MC-Helan appears to over classify helices into the kinked category, classifying the majority of helices as kinked, except those that nearly all participants annotate as straight (Figure 3.10c). Helanal-Plus classified far fewer helices as kinked than any of the other methods (Figure 3.10d).

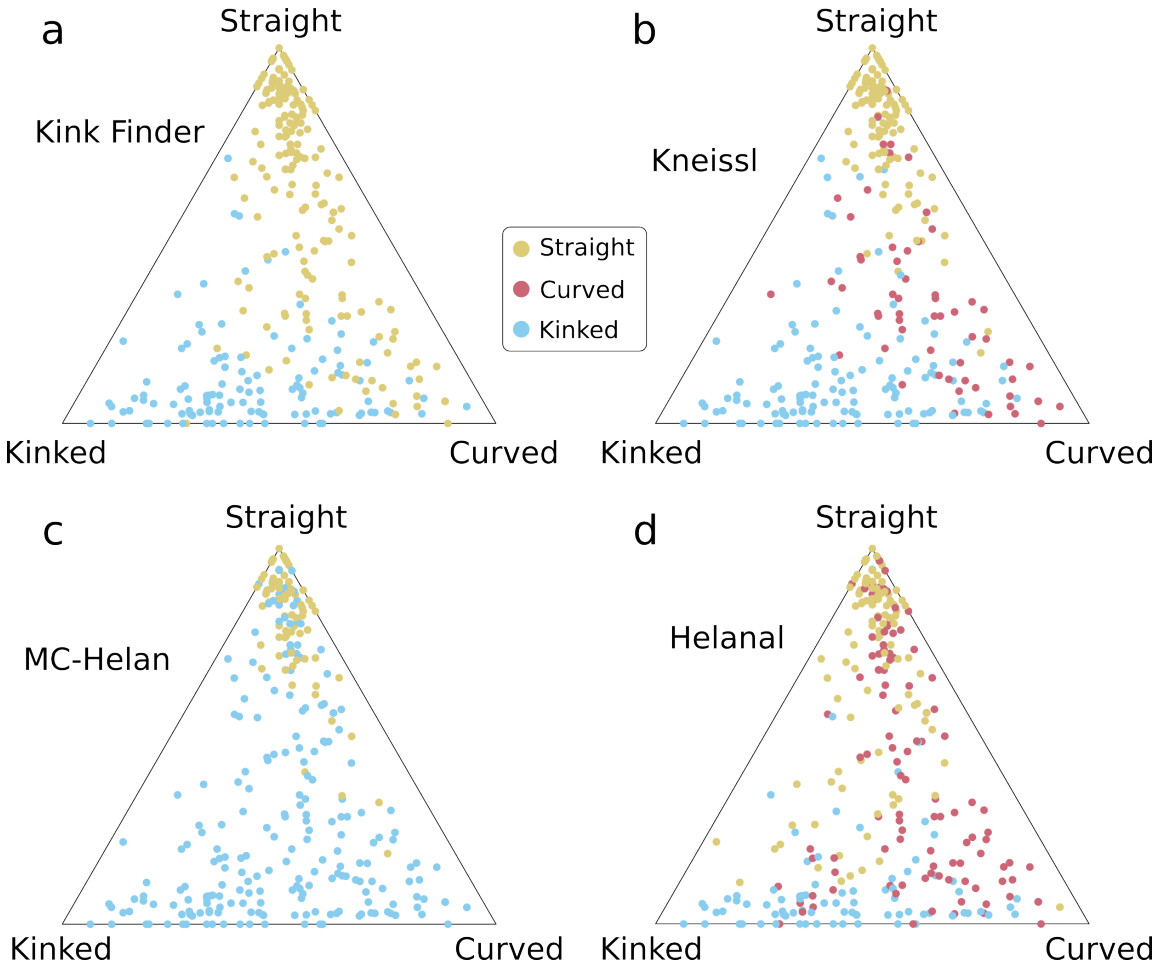


Figure 3.10: **Ternary diagrams showing the responses for each helix, coloured by the annotation of four helix classification methods.** (a) Kink Finder, (b) Kneissl manual annotation, (c) MC-Helan, and (d) Helanal-Plus.

Table 3.3: **Percentage overlap between our gold standard and other kink identification methods.** Bold numbers indicate where the methods agree with our gold standard. K - Kinked, C - Curved, S - Straight, U - Unassigned.

% Gold	Kink Finder		Kneissl			MC-Helan		Helanal-Plus		
Standard	K	S	K	C	S	K	S	K	C	S
K	**20.7**	0.7	**21.0**	0.3	0.0	**21.3**	0.0	**15.7**	3.3	2.3
C	10.3	9.3	10.3	**8.7**	0.7	19.0	0.7	7.3	**12.0**	0.3
S	1.0	**42.0**	2.0	4.0	**37.0**	16.7	**26.3**	0.7	14.0	**28.3**
U	5.7	10.3	6.7	7.0	2.3	15.0	1.0	3.0	7.0	6.0

The helices classified as curved by Helanal-Plus are distributed across the diagram, like the curved helices of Kneissl. The helices annotated as straight by Helanal-Plus are distributed across the straight and kinked regions.

All of the methods annotate a number of the helices in the curved corner of the ternary plot as kinked, particularly helices that are annotated as straight by very few, if any, participants.

3.3.7 Gold standard

A major aim of the AHAH study was to provide a reliable gold standard set to be used in the development of future computational annotation approaches. Figure 3.11 shows the cutoffs we used to produce our gold standard set. A helix falls into a class if more than 50% of the manual annotations agree. Otherwise the helix is not assigned to any class. The gold standard set is shown in full in Table A1 in Appendix A.

A total of 64 (21.3%) helices are classified as kinked, 59 (19.7%) as curved, and 129 (43.0%) as straight. The remaining 48 (16.0%) helices were not assigned a classification. Table 3.3 shows the overlap between this gold standard set and the other kink identification methods.

Previous methods typically overestimate the number of kinked helices, annotating 37.7% (Kink Finder), 40% (Kneissl), 67.3% (MC-Helan), and 26.7% (Helanal-Plus) of the helix set as kinked (Table 3.3). For each method, there are between 22 (7.3%) and 57 (19.0%) helices that are annotated as kinked, but are classified as curved in our gold standard.

The Kink Finder annotation agrees well with the straight and kinked classifications, with

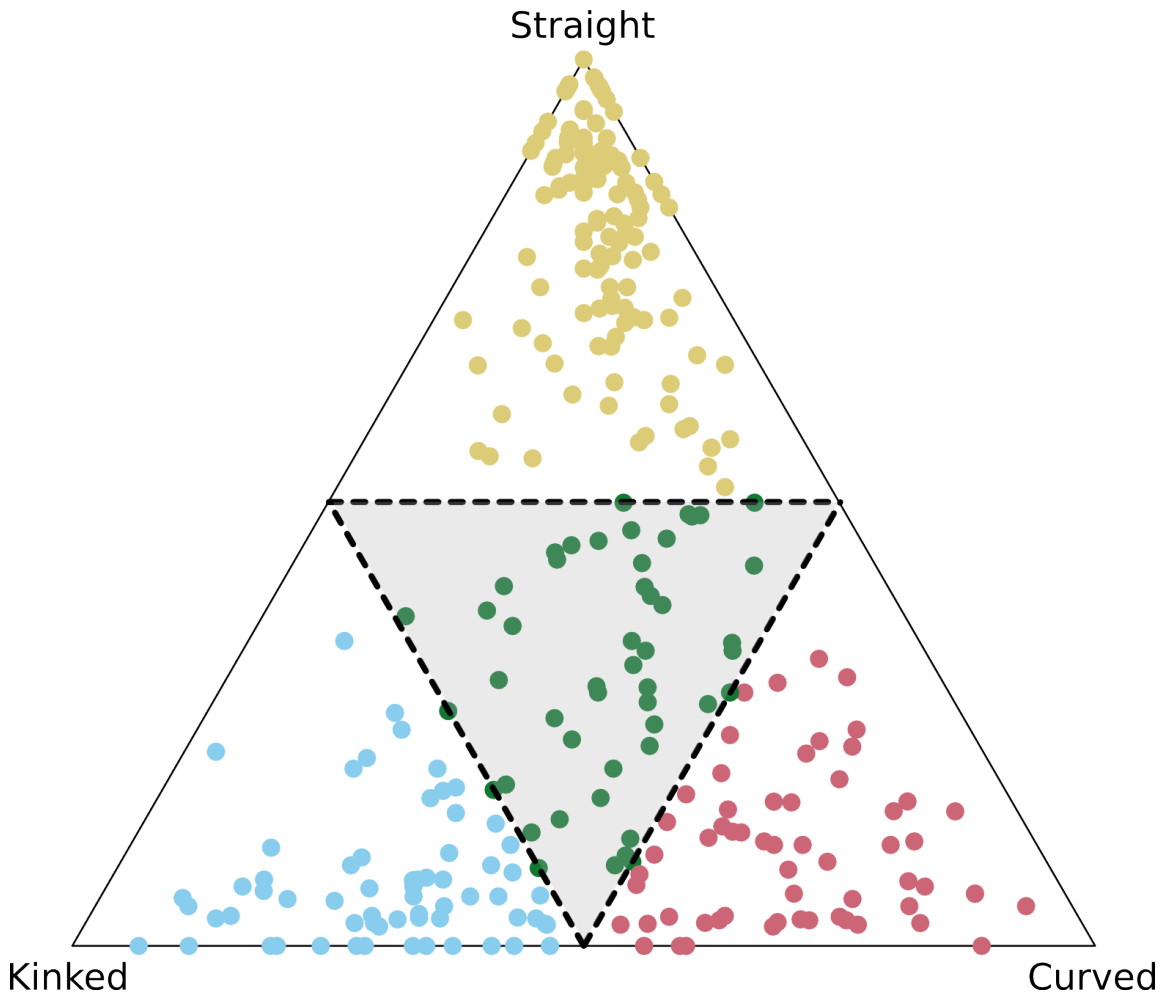


Figure 3.11: **Ternary diagram showing the thresholds used to define the gold standard.** Helices within the grey area were not classified into any group. Blue - kinked, red - curved, yellow - straight, green - unassigned. An interactive version of this figure is available online at `http://opig.stats.ox.ac.uk/webapps/ahah/php/experiment_results.php`.

only 5 (1.7%) helices annotated as kinked by one of the gold standard or Kink Finder, and straight by the other. The helices classified as curved in our gold standard are roughly evenly split between kinked and straight by Kink Finder. Similarly, for the Kneissl annotation, only 6 (2.0%) helices are classified as straight by either Kneissl or our gold standard, but kinked by the other (Table 3.3). The difference between straight and the other two classifications is much clearer than the difference between kinked and curved.

All but 5 of the 88 helices annotated as straight by MC-Helan are in the straight group in the gold standard. However, the helices that MC-Helan identified as kinked are split roughly evenly between the kinked, curved, and straight groups in the gold standard (Table 3.3). Many helices annotated as curved by Helanal-Plus are annotated as straight in the gold standard set (Table 3.3).

In terms of individual participant annotations, the Kneissl annotation agrees with 61.3% (6537 out of 10665) of our participants responses, Kink Finder 58.1%, Helanal-Plus 53.1%, and MC-Helan 48.3%. Of the 48 (16.0%) helices that AHAH is unable to classify, there are only four cases where the other methods give the same classification, and in over half of the cases (25) they give all three classifications (kinked, curved, and straight). Of the 64 kinked helices, 39 are annotated as kinked by all four existing methods. Of the 62 helices in the 300 AHAH helices that were annotated as kinked by all four existing methods, 39 were annotated as kinked, 17 curved, 1 straight, and the remaining 5 were unclassified in the gold standard data set.

3.3.8 Effect of helix length on classification

In our gold standard set, shorter helices are far more likely to be annotated as straight - of the 63 helices with fewer than 20 residues, 50 (79.3%) are annotated as straight (Figure 3.12). From 20 to 27 residues, the proportion of straight helices falls to nearly zero. Only 4 (6.3%) of the 64 helices longer than 27 residues are classified as straight. This relationship is also seen in the classifications of the 4 methods - Kink Finder, Kneissl, MC-Helan, and Helanal-Plus (Figure 3.13).

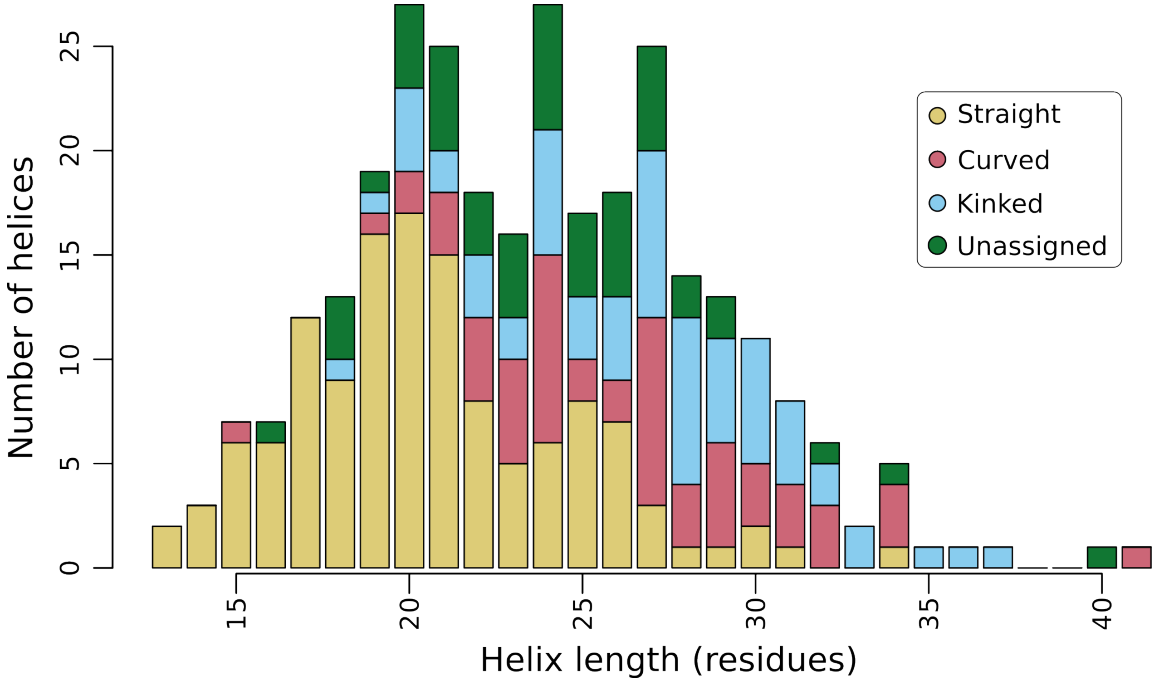


Figure 3.12: **Helix annotations in the gold standard set, grouped by helix length.**

The agreement with the majority view is also higher for shorter helices (Figure 3.14). In particular helices shorter than 20 residues are classified with a very high agreement among participants. This suggests that straight helices are easier to classify.

3.3.9 Kink positions

It is unclear from previous studies if a single kink residue is appropriate, or if kinks are localised over several residues, or even several turns of the helix. Among the 64 helices classified as kinked in the gold standard set, the selected kink residue varies from participant to participant. The most popular residue is selected by more than half of the participants for only three helices. The size of the variation of the chosen position differs between the helices. Looking at the standard deviation of the kink positions identified by the participants gives a rough indication of the degree of localisation of the kink.

The standard deviation of the kink position selected by the participants varies between 0.6

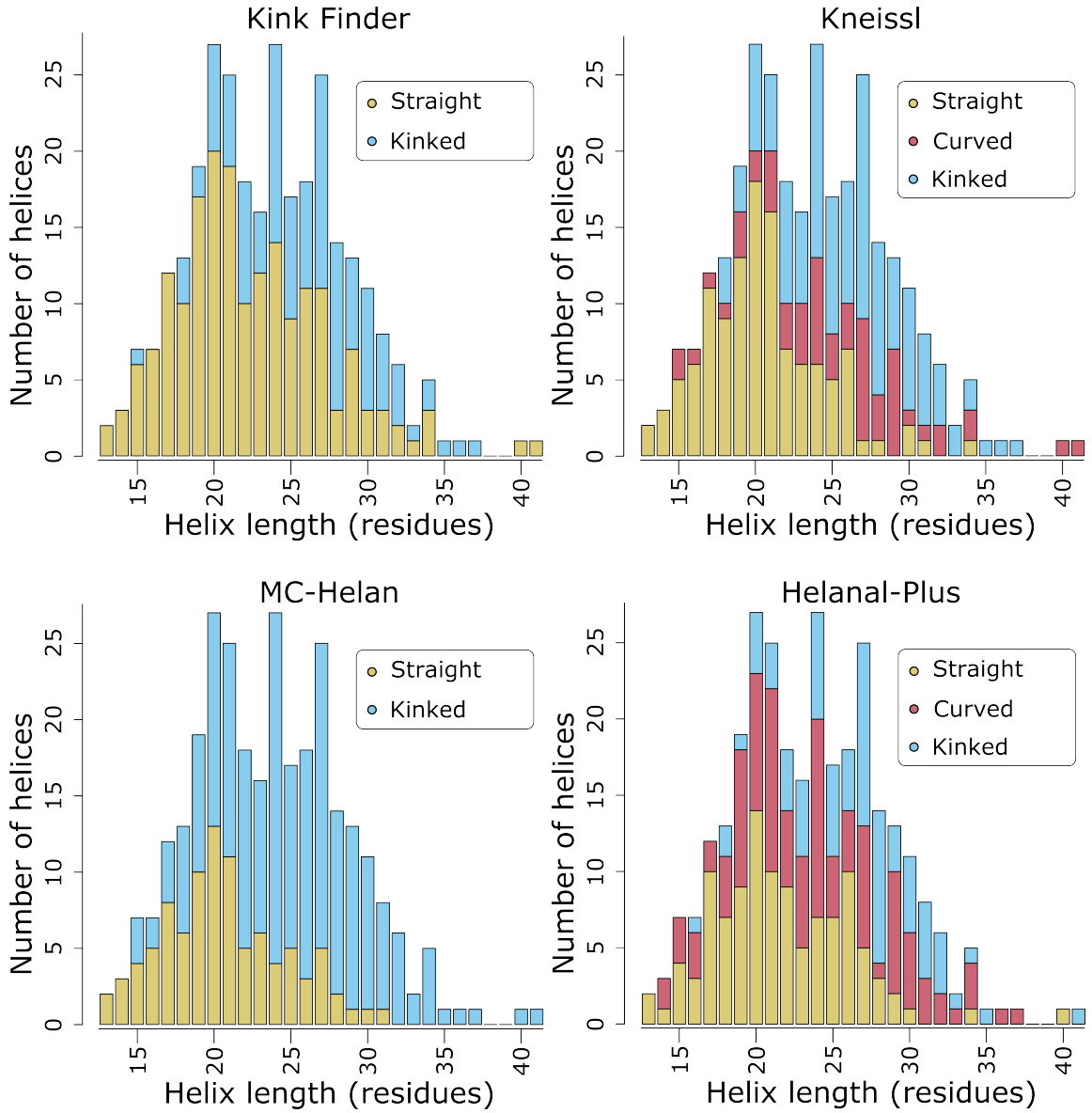


Figure 3.13: **Classification of 300 helix set grouped by helix length, according to the four methods**, Kink Finder (top left), Kneissl (top right), MC-Helan (bottom left), and Helanal-Plus (bottom right).

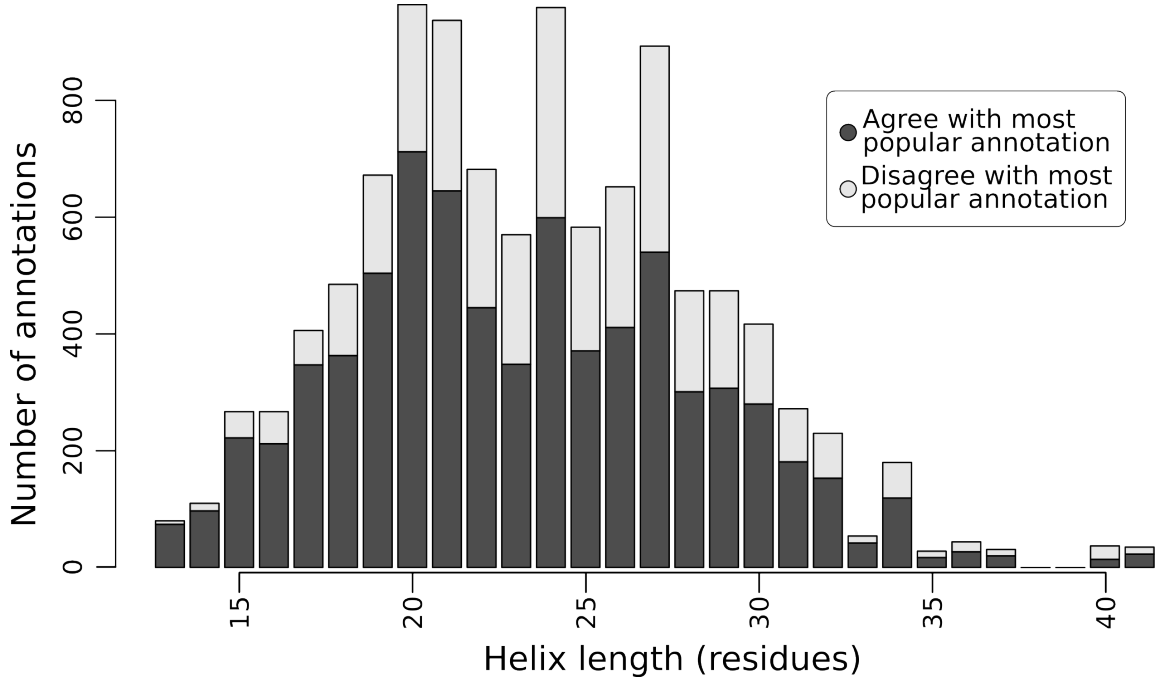


Figure 3.14: **Helix annotations grouped by helix length.** Dark grey indicates that the annotation agree with the consensus, the pale dark bars indicate annotations that disagree with the majority view.

and 5.9 residues (Figure 3.15a). The distribution of standard deviations is unimodal, with a median of 1.9, but is skewed towards 0, and has a tail to the right. This indicates that there is no archetypal kink, and that the degree of localisation is not constant. The 'average' kink is localised over a turn (3.6 residues) or thereabouts, and most are localised over no more than two turns.

There is no correlation ($r^2 = 0.08$) between the percentage of participants annotating the helix as kinked and the standard deviation (Figure 3.15b). This indicates that more obvious kinks are not necessarily localised over a smaller region of the helix.

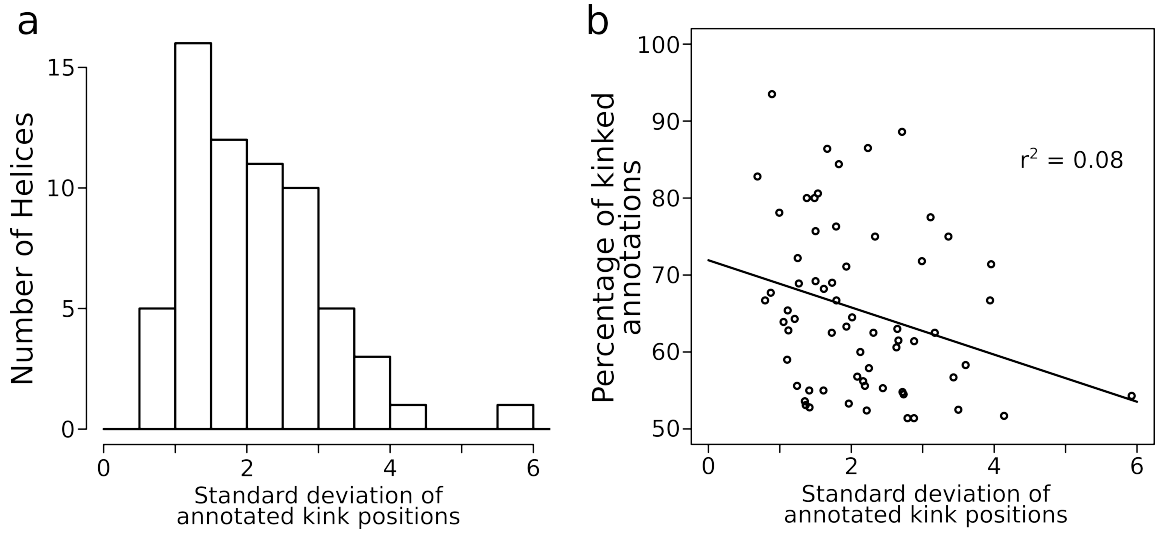


Figure 3.15: **Variation of the standard deviation of kink positions in helices identified by participants.** (a) The distribution of standard deviations. (b) The relationship between standard deviation and the percentage of participants that classified the helix as kinked.

3.3.10 Amino acids in gold standard kinks

The counts of the residues in the 64 kinks at each of the positions around the kink are shown in Figure 3.16. Like for previous studies of residue preferences around kinks (Hall *et al.*, 2009; Kneissl *et al.*, 2011; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011), there is a clear preference for proline to be in the turn after the kink. Proline is concentrated around positions +2 (twelve kinks), +3 (eight), and +4 (nine) indicating that users tend to identify the kink position at

residues two or three before it, i.e. on the inside of the kink. There are seven kinks where proline is at the kink position, suggesting that the proline may have been the most obvious feature in some kinks.

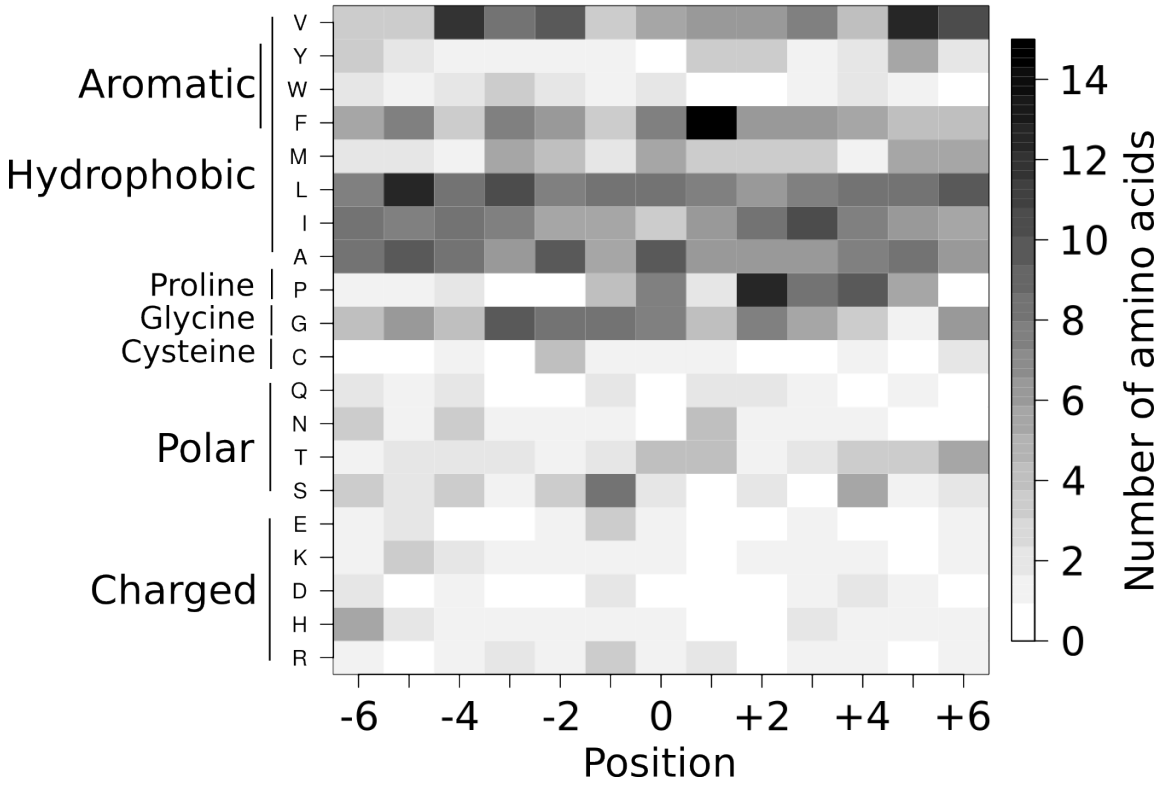


Figure 3.16: **Residue counts for the 64 kinks in the gold standard data set.** Positions calculated from the modal value, with the closest to the mean as a tiebreaker, then the N-terminal one if still tied.

3.4 Discussion

Previous studies of kinks have identified conflicting characteristics of kinks, each study used a different data set, and a different method to identify kinks. In this chapter, I have compared the results of five distinct methods (three from previous studies, and two of our own design) on the same data set. By changing the angle thresholds, the existing methods were modified

so that they identified the same number of kinks. By applying the same position correction as I developed for Kink Finder (see Figure 2.7), with the initial (uncorrected) kink position defined as the kink position identified by the existing method, I modified the existing methods to consistently identify the kink position with respect to the geometry of the helix.

3.4.1 Disagreement among current helix characterisation tools

The specific definition of kinks is subjective, and changes from study to study. The four helix classifiers, Kink Finder (Wilman *et al.*, 2014b), Kneissl et al. (Kneissl *et al.*, 2011), MC-Helan (Langelaan *et al.*, 2010), and Helanal-Plus (Kumar & Bansal, 2012) considered in this chapter provide a range of approaches to kink identification. They agree on the classification of only half of the helices in our tests.

However, when I altered the methods to identify the same number of kinks, and to consistently identify the kink residue relative to the shape of the helix, the results of each method, were much more similar to each other than the results reported in the corresponding papers. This indicates that the different conclusions from previous studies was due to a combination of the different data sets, the different kink identification algorithms, the choice of parameters within these algorithms, and the method used to identify the kink residue.

The disagreements between the methods is likely to be caused by a lack of a gold standard, which makes a performance comparison impossible. In this chapter, we provide a crowdsourcing based gold standard aimed to be used in the training and comparison of computational methods.

3.4.2 Crowdsourcing can tackle difficult problems

Crowdsourcing has been used by a number of researchers to provide classifications for computationally difficult tasks that require a large amount of time. Researchers found for micro (Lintott *et al.*, 2010, 2008) and mega (Cooper *et al.*, 2010; Kawrykow *et al.*, 2012; Khatib *et al.*, 2011) tasks, that a large number of citizens provide responses that are as good as, if not better than, expert classifications. Studies rarely need to apply acceptance criteria for the crowd

data (Luengo-Oroz *et al.*, 2012; Nguyen *et al.*, 2012), and weighting user responses has little effect on the classifications (Lintott *et al.*, 2010, 2008). In agreement with this, we found that specialists (i.e. Structural Biologists) are only slightly more consistent at classifying the helices than untrained individuals. The school pupils were slightly worse than the average participant (although not statistically significantly so). However most of the poor quality responses (which we removed from the analysis) came from three of the school pupils.

3.4.3 A discrimination between straight and not-straight helices

The classifications provided by our participants indicate that many helices do not fall into the clear classes of kinked, curved, and straight. There are a number of helices that are consistently classified as straight. There are also many helices which are very rarely classified as straight. However, there is no clear dividing line between a set of 'kinked' and a set of 'curved' helices. This could be due to the participants, or our chosen definitions, but given the disagreements in the published computational methods, it is perhaps more likely to be due to the inherent nature of the helices.

Helix length has a strong influence on helix classification. In our gold standard set, short helices (≤ 20 residues) are generally straight, while longer helices (≥ 27 residues) are generally either kinked or curved. This is a common feature of all of the classification methods discussed here, and replicates one of the findings of Chapter 4, indicating that this is not an artefact of the method used to identify kinks.

3.4.4 Little agreement in exact kink position

The difficulty in identifying kinks is also demonstrated by the kink positions annotated by the participants in AHAH. For some helices, the position is regularly annotated within a few residues, while in others it is identified over a larger range of residues. It is very rare for a single residue to be consistently identified as the site of a kink - either by our participants, or the existing methods. It is, perhaps, surprising that the variation in the kink position does not

correlate ($r^2 = 0.08$) with the proportion of participants that annotate the helix as kinked, i.e. a more obviously kinked helix is no indication of a more localised kink. Users tend to annotate the kink position on the inside, in agreement with the Kink Finder repositioning method.

Modifying the four methods so that the kink residue is consistently chosen with respect to the local geometry of the helix increases the number of helices in which all methods identify the same kink residue, and in part eliminates the different findings of the methods. For example, some studies have identified an enrichment in glycine residues at specific positions relative to the kink position (Devillé *et al.*, 2008; Hall *et al.*, 2009; Kneissl *et al.*, 2011; Wilman *et al.*, 2014b), but others have not (Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Werner & Church, 2013). However, my method results in the four methods all identifying glycine as more frequent at the kink position, while there are also specific patterns with serine, valine, phenylalanine, and tryptophan common to all of the methods.

3.4.5 A new gold standard training set

Our participants classified a total of 300 helices. In 252, the agreement was above 50%. We suggest this subset as a high quality gold standard set for the testing of future computational approaches. The raw data from our survey was made available online, to allow researchers to produce their own gold standard, with different criteria, if they so wish.

This set has a lower proportion of kinked helices than previous methods, which suggests that previous methods have been overly sensitive to helix distortions. It is noticeable that the proportion of kinked helices in the gold standard set is much lower than the computational methods. Only 66 (22%) of the 300 helices are annotated as kinked, compared to 37% in the Kneissl *et al.* (2011) analysis. This could be changed, with a relaxing of the 50% threshold. Removal of the 'unclassified' group, and reclassifying these helices based on the majority decision would simplify the results, but not result in any more agreement with the computational methods. It would add ambiguous helices to our gold standard set, which is likely to be detrimental to its purpose as a training set. More complex divisions of the helices are possible, for example

by expanding the kinked region to include areas towards the curved corner. This would improve the agreement with existing classifiers, but there is no justification other than that.

The annotation is dependent on the instructions to the participants, where we described kinked helices as 'There is a clear location where the direction of the helix changes. Only a small part of the helix is involved in this.' This is very similar to the rule of thumb described in Kneissl *et al.* (2011), suggesting that their two researchers provided annotations that were not consistent with this. Indeed, there are two cases where their annotation disagrees with more than 90% of our respondents, and 14 cases where they disagree with more than 80% of the participants. Providing a more relaxed definition of kinks to the participants may have resulted in more helices being annotated as kinked.

3.4.6 Experiment parameters and number of annotations

For ease of use, the helices were displayed using the amino acid colouring scheme in JSmol, which meant that each of the twenty amino acids types was coloured differently. It is possible that this could have biased the participants - particularly if they knew the colouring scheme, and had prior knowledge of the link between proline and kinks. However, this would not be the case for the majority of users. Alternatively, it might be that brighter residues were more likely to be selected by participants, but there is insufficient data for such a bias to be detected.

We attempted to ensure that we had many annotations from a lot of people, rather than just relying on a few users. Our use of a simple scoring function provided a modest level of 'gamification', with participants from the group and the DTC comparing and competing with their results. The use of gamification, as in Foldit, and in the Zooniverse project, does seem to encourage participation (Eveleigh *et al.*, 2013).

3.5 Conclusions

Existing kink identification methods disagree on the identification of kinks, with no objective standard. The differences can be reduced, but not eradicated, by using similar thresholds, by using a consistent method to choose the kink residue, and by considering the same data set. We used a crowdsourcing approach to create a gold standard data set that can be used for training, testing and evaluation of kink identifiers and predictors. The gold standard annotation identifies fewer kinks than existing kink identification methods, suggesting that many of the kinks identified by these methods are subtle, rather than obvious, changes in the helix axis. The crowd sourcing experiment also revealed that the difference between kinked and curved helices is less clear than between straight and not-straight helices. Further, the degree to which kinks are localised in the helix varies from kink to kink, but kinks are generally localised to a single helix turn.

CHAPTER 4

Helix kinks in soluble and membrane proteins.

The majority of the work in this chapter was published as an article entitled 'Helix kinks are equally prevalent in soluble and membrane proteins' in the journal Proteins: Structure, Function, and Bioinformatics (Wilman, Shi & Deane, 2014b).[1]

[1]The article is available under the terms of the Creative Commons Attribution License (CC BY), Copyright 2014 The Authors. Much of the text and all of the figures from the article are reproduced with adaptions in this chapter.

4.1 Introduction

Kinks are known to occur in many proteins (Section 1.3.4) (Barlow & Thornton, 1988; Bowie, 2013; Nugent & Jones, 2011), and they are known to be functionally important (Barrett *et al.*, 2012; Bettinelli *et al.*, 2011; Ni *et al.*, 2011; Sansom & Weinstein, 2000; Schwartz *et al.*, 2006; Suchyna *et al.*, 1993; Tieleman *et al.*, 2001; Weber *et al.*, 2012; Yohannan *et al.*, 2004b). It has been claimed that kinks are far less frequent in soluble protein helices than membrane proteins helices (Cao & Bowie, 2012; Devillé *et al.*, 2008; Langelaan *et al.*, 2010; Mai & Chen, 2014). This separation of membrane and soluble proteins reflects research on membrane protein generally. It is well understood that the membrane environment has an effect of protein structure (Stevens & Arkin, 1999; Ulmschneider & Sansom, 2001), and that membrane specific tools for e.g. modelling (Ebejer *et al.*, 2013) and alignment (Chang *et al.*, 2012; Hill & Deane, 2013; Hill *et al.*, 2011) outperform similar generic tools. The majority of previous research on kinks has concentrated on transmembrane helices, but there is little comparable research into soluble helix kinks, and to my knowledge there is no previous work that makes a direct comparison between the frequency and features of kinks in helices of membrane and soluble helices.

The relative lack of membrane protein structural data (Section 1.3.4) has hampered previous research on kinks. As shown in this chapter, using reasonably relaxed redundancy and quality thresholds, the PDB (Berman *et al.*, 2007) contains c. 1200 helices with more than 12 residues from membrane proteins. A higher redundancy threshold gives a larger data set, which allows stronger statistical conclusions to be drawn. However, increasing the redundancy raises the chances of bias.

There are far more solved structures of soluble proteins, with much higher diversity. In order to see if I can take advantage of this data, I compared helix kinks in soluble and membrane proteins. My aim was to identify the similarities and differences of kinks in membrane and soluble helices.

In this chapter, I investigate kinks in both membrane and soluble protein α-helices. I

demonstrate that kinks are equally prevalent in the two types of proteins. Specifically I find that kinks occur in long helices (≥ 20 residues) in *all* proteins, and kinks occur equally frequently in long helices regardless of their environment. Kinks in the two types of proteins are similar in the residue patterns, and frequency of broken hydrogen bonds. Proline is a dominant feature of kinks in both types of protein, while other residue patterns are similar, allowing for the expected differences in the amino acid distributions. One notable feature that occurs only in soluble protein helices is that kinks have a structural preference to point into the solvent, revealed by patterns in both the hydrophobicity and solvent accessible surface area of residues around the kink.

4.2 Methods

To compare membrane and soluble protein kinks, I required two analogous sets of protein helices. These protein helix structures were collected from the PDB (Berman *et al.*, 2007), one containing soluble protein helices, and the other transmembrane helices.

4.2.1 Soluble data set

To compile a soluble protein data set, the PDB was filtered with PISCES (Wang & Dunbrack, 2003), using the following settings: less than 80% sequence identity, $40 <$ chain length < 1000 residues, resolution < 5Å, R-factor < 0.4, include non X-ray structures, exclude Cα-only structures. I use the first conformer in each NMR protein structure. To remove any transmembrane structures from this set, any protein chains that were in the PDBTM (Kozma *et al.*, 2013; Tusnády *et al.*, 2004) and/or in the Membrane Proteins of Known Structure Database (White & Wimley, 1999) were excluded. The JOY program (Mizuguchi *et al.*, 1998) was used to annotate the helix structures (see Section 4.2.8). Initially, regions of contiguous helix were identified, using the DSSP algorithm (Kabsch & Sander, 1983). Those helical regions separated by only one or two coil residues were combined. These helices were split if they contained a kink angle,

as defined by Kink Finder (see Chapter 2), greater than 60°.

Correctly annotating the end of helices is important in kink identification, as errors in assignment can lead to false positives in kink finding methods (see Chapter 3 for further discussion). The ends of the helices were checked for helical nature. Where the first five and last five residues of the helix were not a helical seed, as defined in MC-Helan (Langelaan *et al.*, 2010), residues were iteratively removed from the end of the helix, until this condition was met. The requirements for a helical seed, as defined by MC-Helan, are threefold. First, the first residue must have dihedral angles in the alpha-helical region (Lovell *et al.*, 2003). Second, the angles ($C^{\alpha}_{i+x}C^{\alpha}_{i}C^{\alpha}_{i+1}$) must lie within the expected ranges for an alpha helix ($35-50°$ for $x=2$, $60-80°$ for $x=3$, and $45-65°$ for $x=4$). Third, the $C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+x}$ ($x=2,3,4$) distances must be within 0.5 Å of the values for an ideal α-helix. Helices with 12 or fewer residues were removed, as this is the shortest length for which kinks can be identified by Kink Finder.

The final data set contained 9742 protein chains, with a total of 29699 helices. I also compiled data sets with three other sets of thresholds:

- less than 80% sequence identity, resolution < 5Å, R-factor < 0.9, gave a set with 31633 helices from 10822 chains.
- less than 80% sequence identity, resolution < 3Å, R-factor < 0.9, gave a set with 29064 helices from 9528 chains.
- less than 50% sequence identity, resolution < 5Å, R-factor < 0.9, gave a set with 24245 helices from 7992 chains.

These sets gave very similar results to each other (see Section 4.3).

4.2.2 Membrane data set

PDB codes of membrane proteins were initially identified from the Membrane Proteins of Known Structure Database (White & Wimley, 1999) and from the PDBTM (Kozma *et al.*, 2013; Tusnády *et al.*, 2004). Structures derived from electron microscopy experiments, and

those containing only Cα atom coordinates were removed. These proteins were split into their constituent chains, and filtered using the same criteria as for the soluble data set (less than 80% sequence identity, resolution < 5Å, R-factor < 0.4). The membrane position for each residue was annotated by iMembrane (Kelm *et al.*, 2009), with only chains annotated with ≥ 1 residue in the tail region of the membrane retained. The tail region of the membrane is the hydrophobic core, in contact with the hydrophobic tail groups of the phospholipids that make up the membrane (Figure 1.11). All remaining non membrane proteins were removed by visual inspection. Although the resolutions were generally better for the soluble data set, only 20 chains in this membrane set had resolution ≥ 4Å. Helices were identified in the same manner as for the soluble data set. Helices that had no residues annotated as being located in the tail region of the membrane by iMembrane were removed. Helices were trimmed so that no more than five residues at either end were outside the membrane, by iteratively removing residues from the end of a helix until this was the case. In this way, the membrane annotation provided by iMembrane allows us to exclude non-membrane helices, such as the so called H8 helix in GPCRs, from this dataset. This allows Kink Finder to calculate an angle for each of the residues in the head or tail region of the membrane. The final data set contained 268 protein chains, with a total of 1208 helices (see Table B4 in Appendix B). I also compiled data sets with three other sets of thresholds:

- less than 80% sequence identity, resolution < 5Å, R-factor < 0.9, gave a set with 1326 helices from 281 chains.

- less than 80% sequence identity, resolution < 3Å, R-factor < 0.9, gave a set with 861 helices from 196 chains.

- less than 50% sequence identity, resolution < 5Å, R-factor < 0.9, gave a set with 1002 helices from 220 chains.

These sets gave very similar results to each other (see Section 4.3).

4.2.3 Kink identification

Kinks in helices were identified by the Kink Finder algorithm (Chapter 2). For comparison, identical analysis was undertaken using kinks identified by the MC-Helan algorithm (Langelaan *et al.*, 2010). As described in Chapter 3, these two methods identify a different proportion of helices as kinked, and identify the kink position in a different way. I applied an angle cut off to MC-Helan that classified a similar proportion of helices as kinked, and modified the algorithm to select the kink residue by the same method as Kink Finder (as in Figure 2.7). This provided results that had good agreement with those from the Kink Finder analysis. These results are shown in Appendix B, in Figures B1, B2, B3, and Table B2.

4.2.4 Length matching

Sets of soluble helices with the same length distribution as the membrane helices were selected by sampling from the full soluble helix set. For each sample, soluble helices were chosen, without replacement, to match the length distribution of the membrane helices. This was possible as currently there are far more structures of soluble helices than membrane helices. For this work, 50 samples were taken from the soluble data set.

4.2.5 Sequence homologue collection

Homologous sequences for each protein chain in the two sets were collected. Sequences were obtained by using PSI-BLAST (Altschul, 1997) to search the UniProt90 database (The UniProt Consortium, 2013), with the following settings: 2 iterations, $1e^{-5}$ as the E value threshold (to include a sequence in the model used by PSI-BLAST to create the Position Specific Substitution Matrix, or 'inclusion E thresh'), and $1e^{-3}$ as the E value threshold (to retain a sequence). Sequences with a large length difference to the structure were removed (those with length greater than 3/2 or less than 2/3 of the structure). The remaining sequences were aligned using MAFFT 7.015b (Katoh & Standley, 2013), with options: –maxiteration 1000 –localpair –treeout.

4.2.6 Sequence profiles

Protein sequence profiles were built using the sequence alignments, with each sequence weighted according to its similarity to every other sequence, with more dissimilar sequences having higher weights (Shi *et al.*, 2001; Vingron & Argos, 1989; Vingron & Sibbald, 1993). Helices had on average 52 homologous sequences. The homologous sequences were used only in the calculation of amino acid propensities and hydrophobicities.

4.2.7 Hydrophobicity

Amino acid hydrophobicities were taken from experimentally-derived interface-octanol data (White & Wimley, 1999). The sequence profiles derived from sequence homologues were used to calculate the hydrophobicity of each position near the kink.

4.2.8 Hydrogen bonds and solvent accessible surface area

The JOY program, version 5.10 (Mizuguchi *et al.*, 1998), was used with its default parameters to annotate both backbone and sidechain hydrogen bonds in the protein structures, and the solvent accessible surface area of each residue.

4.2.9 Propensities

Amino acid sequence profiles can be used to examine whether residues are favoured at specific positions around a kink, using their propensity to be at a given position relative to a kink, as shown in Equation 4.1;

$$\text{Propensity} = \mathrm{P}_i^a = log_e \left(\frac{N_i^a / N_i^{\text{all}}}{N_{\text{helices}}^a / N_{\text{helices}}^{\text{all}}} \right) \tag{4.1}$$

where N_i^a is the number of residues (N) of a particular amino acid type (a), e.g. glycine, observed at a particular position relative to the kink (i). N_i^{all} is the total number of residues observed at a particular position relative to the kink. $N_{\text{helices}}^{\text{all}}$ is the total number of amino acids

(all) observed in the helices within the data set. $N^{\mathrm{a}}_{\mathrm{helices}}$ is the total number of a given amino acid type in all helices within the data set. The background distribution ($N^{a}_{\mathrm{helices}}/N^{\mathrm{all}}_{\mathrm{helices}}$) comes from the set of helices from which the kinks are identified (hence the membrane and soluble sets each have their own background distribution). The data used to calculate the propensities come from the sequence profiles, which are compiled using homologous sequences (Section 4.2.6).

4.2.10 Motifs

Sequence motifs were identified by searching 13 residue segments around each kink, and the proportion of kinks containing one or more instances of the motif was recorded. This value was compared to a background, calculated by selecting a random 13 residue segment from each of the straight helices (those with maximum angle $\leq 14°$) in the set, and searching these for the motif. This random sampling was repeated 50 times, and the result averaged.

The motif notation is based on regular expressions. Each character represents a single amino acid, the single letter amino acid identifiers are used, x represents any amino acid, and letters in square brackets (e.g. [ACD]) represent any one of the letters in the square brackets (e.g. A, C or D). Ar represents any of the amino acids with aromatic side chains (i.e [FYW]). The 'G_0 or G_{+1}' motif is a special case, which represents all kinks where glycine is at position 0 or +1.

4.3 Results

This results section shows first that the kinkedness of membrane and soluble protein helices is similar, allowing for the differences in the length distribution of the the helices, and second that the features of kinks in the two data sets are similar.

4.3.1 Angle distributions

A simple comparison of the largest angle of each helix suggests that the proportion of highly kinked helices is lower in soluble proteins than in membrane proteins (Figure 4.1a). However, the

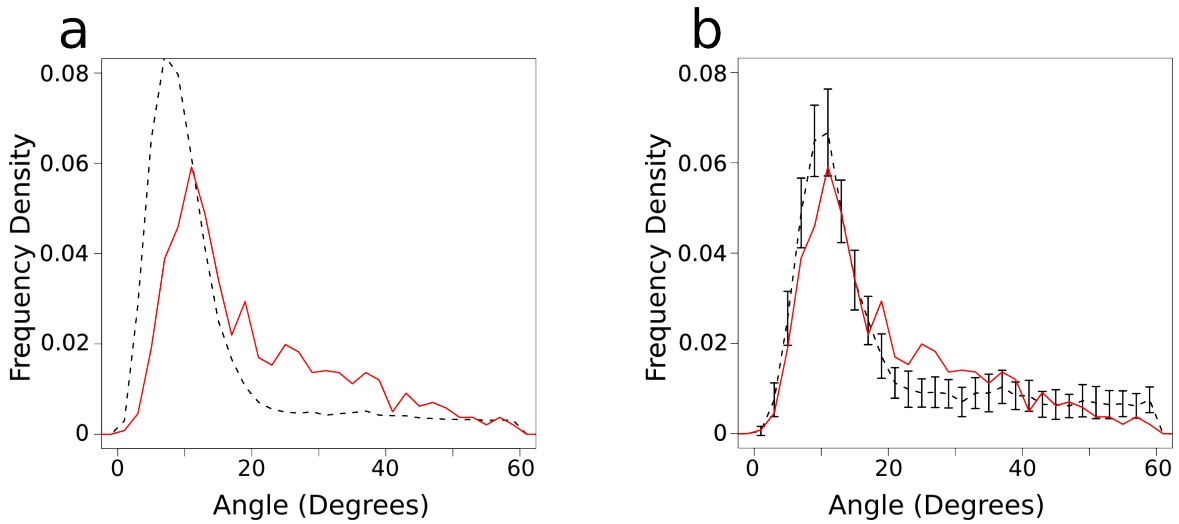


Figure 4.1: **Kink angle distributions.** (a) The distribution of maximum angles in all membrane (solid red) and all soluble (dashed black) helices. (b) The distribution of maximum angles in length matched sets of membrane and soluble helices. The soluble values are means of 50 samples which are matched to the same length distribution as the membrane helices. Bars show 1 s.d.

distribution of the lengths of the two types of helices are markedly different, and the maximum kink angle in the helix is dependent on the length of the helix (see Figure 4.2 and Figure B3 in Appendix B). In both membrane and soluble proteins, the maximum angle in the helix increases as the length of the helix increases. There are many more short soluble helices than short membrane helices. This skews the distribution of angles - in these data sets, the average soluble helix has *c.* 14 residues, and a maximum kink angle of *c.* 10°, whereas the average membrane helix has *c.* 23 residues, and an angle of *c.* 15°.

I compared the maximum angle distributions of soluble and membrane helices of the same length (e.g. all 20 residue long soluble helices compared to all 20 residue long membrane helices), using a two sample Kolmogorov-Smirnov test (Chakravarti *et al.*, 1967; R Core Team, 2014). For the majority of lengths, the distribution of angles for membrane and soluble helices are not different (p value ≥ 0.05, p values for each test are shown in Table B3). For shorter helices (lengths of 19 residues and shorter), the Kolmogorov-Smirnov tests show a difference

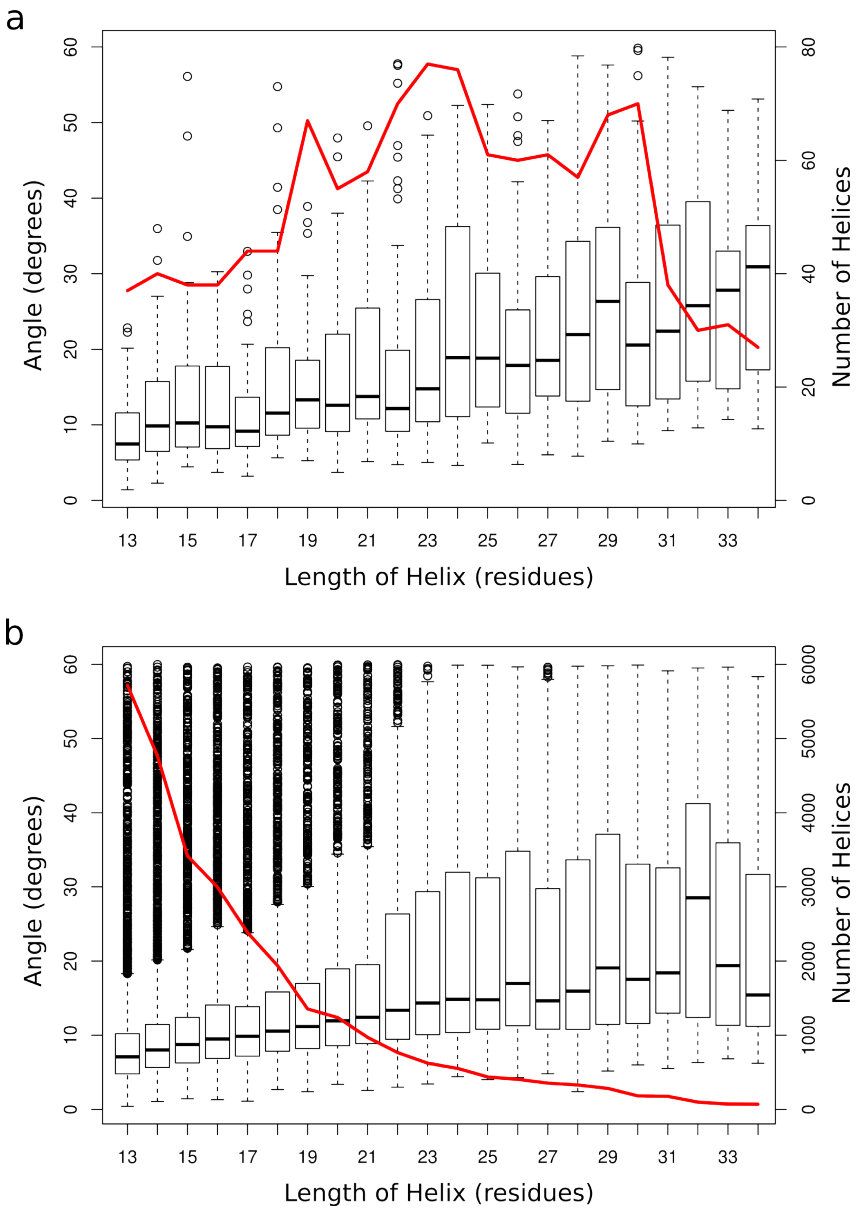


Figure 4.2: **Lengths and angles of helices.** (a) membrane, (b) soluble helices. Plot of maximum kink angles (calculated by Kink Finder) for each different length of helix. Each boxplot shows the distribution of maximum angle in each helix of that length (scale on left axis). The red line shows the total number of helices of each length (scale on right axis). Note how in both plots the maximum angle increases as the helices get longer. In (a) the majority of the helices are between 20 and 30 residues long, however in (b), majority of helices are fewer than 15 residues long.

between the soluble and membrane helix angle distributions. The means of these distributions are similar, however, indicating that the differences measured in the test may be due to the very large number of short soluble helices compared to the small number of short membrane helices (see Figure 4.2).

These results indicate that helices of the same length should be compared when considering helix kinks. Throughout this work, membrane helices are compared to length-matched samples of the soluble helices.

Figure 4.1b shows the distribution of angles for the membrane set, and 50 length-matched samples from the soluble set. These two distributions are very similar. There are a similar number of helices with maximum angles less than 20° in the two datasets, and the two distributions peak at the same angle (10°). A few more membrane helices than soluble helices have angles between 20° and 30°, and a few more soluble helices than membrane helices have maximum angles above 50°. This is likely due to the membrane environment restricting the degree to which transmembrane helices can kink. This result is not affected by the thresholds used to create the datasets, as Figure 4.3 shows.

In order to investigate the properties of kinked helices, it is necessary to annotate helices as kinked or unkinked. However, the angle distribution is continuous (Figure 4.1), which is consistent with the lack of clear agreement in the literature on the ideal angle cut-off (see Chapter 3). In this chapter, I use 20°, as an angle threshold (see Chapters 2 and 3). This means that in the length matched sets, $32 \pm 1\%$ (1 s.d.) of soluble helices are kinked, and 40% of membrane helices are kinked. Altering this angle threshold in the range of $15 - 25°$ has only a small effect on the results discussed in this Chapter.

4.3.2 Amino acid patterns around kinks

To investigate if the residue patterns around kinks in soluble and membrane helices are similar, I examined the amino acid propensities in positions close to the kink, using length-matched sets of membrane and soluble helices. The propensities are calculated using sequence profiles.

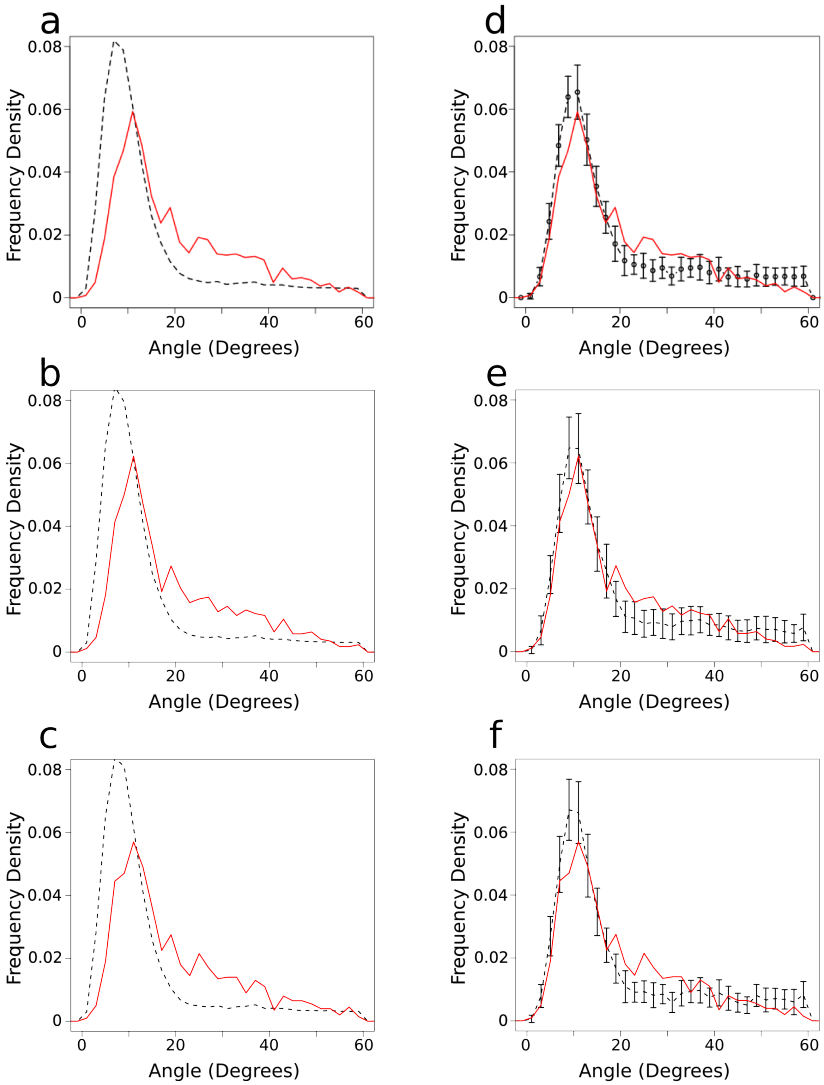


Figure 4.3: **Kink angle distributions.** (a)-(c) The distribution of maximum angles in all membrane (solid red) and all soluble (dashed black) helices. (d)-(f) The distribution of maximum angles in length matched sets of membrane and soluble helices. The soluble values are means of 50 samples which are matched to the same length distribution as the membrane helices. Bars show 1 s.d. (a) and (d) show data from the data set with 80% sequence i.d., resolution ≤ 5Å, R-factor ≤ 0.9. (b) and (e) 80% sequence i.d., resolution ≤ 3Å, R-factor ≤ 0.9. (c) and (f) 50% sequence i.d., resolution ≤ 5Å, R-factor ≤ 0.9.

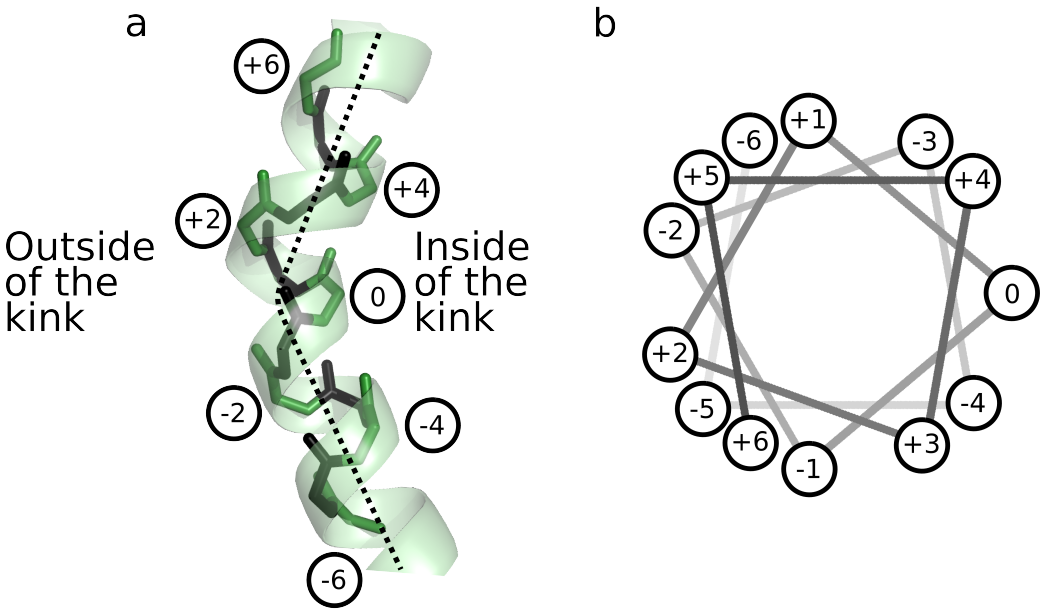


Figure 4.4: **Kink numbering.** (a) Numbering for an example kink, with even numbered residues (green) labelled. A The broken line shows an approximate helix axis and is in the plane of the page. (b) A standard helical wheel diagram, showing the helix in (a). Kinks are numbered from N-terminal to C-terminal, with the kink residue given number 0. This scheme results in residues -4, -3, 0, +3, and +4 being on the inside of the kink, while residues -5, -2, +2, and +5 being on the outside. This shows an ideal kink, where the wobble angle is 0°. The wobble angle of residue 0 can vary between −50° (between +3 and -4 in (b)) and +50° (between -3 and +4 in (b)) (Figure 2.9b). Consequently, the exact location of each position can vary from kink to kink. For example, while position 0 is always on the inside of the kink and position -2 is always on the outside, position +1 may be towards the inside of one kink and towards the outside of another. This figure is identical to Figure 2.10, and is included again here due to its importance in the interpretation of the results in this chapter.

Kink Finder (Chapter 2) chooses a kink residue on the inside of the kink, as shown in Figure 4.4. The residues around each kink are numbered relative to this kink residue, with the kink residue as number 0, those residues toward the N-terminus as negative numbers, and those toward the C-terminus as positive numbers. Thus, position -5 is five residues from the kink residue, toward the N-terminus. A consequence of this numbering scheme is that some positions are consistently on the outside of the kink (e.g. -2 and +2), and some positions are consistently on the inside of the kink (e.g. -4, 0, and +4). This positioning of kink residue means that these results are not comparable to other methods, e.g. Helanal (Kumar & Bansal, 2012) or MC-Helan (Langelaan *et al.*, 2010). As in this study, the region around the kink has a definite shape (Figure 4.4), whereas this is not the case in other methods.

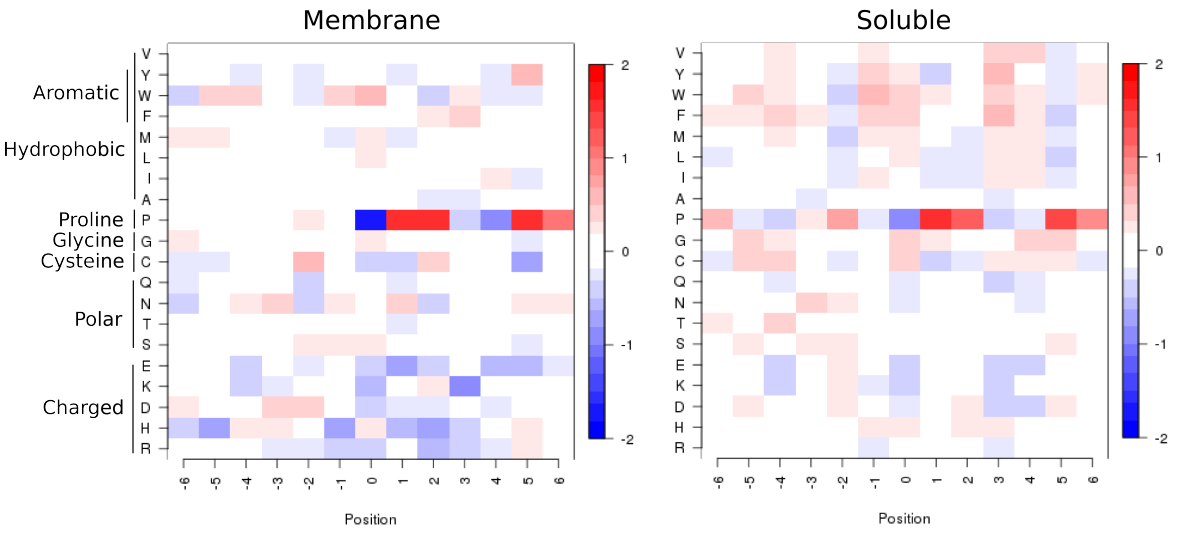


Figure 4.5: **Amino acid propensities for membrane (left) and soluble (right) kinks.** Each value is the propensity for a given residue to be at a given position relative to a kink. These are calculated using sequence profiles. Position 0 is the kink residue, position +1 is one residue towards the C-terminus from the kink residue, as shown in Figure 4.4. Red indicates positive propensities, while blue indicates negative propensities. Residues with positive propensities are found more frequently at that position in kinked helices than in helices in general. See Figure B1 for the analogous figure based on the MC-Helan analysis.

Table 4.1: **Proline in Helices.** Percentage of kinked and non kinked helices containing proline, as identified by Kink Finder. Each 2x2 table shows the percentage of helices in each of 4 groups - kinked with proline in the sequence (profile), kinked without proline in the sequence (profile), straight with proline in the sequence (profile), straight without proline in the sequence (profile). The bottom half of the table ignores the first 4 (N-terminal) residues of helices when determining if there is a proline in the sequence (profile), as these do not have the same effect on the helix structure as those later in the helix.

	Membrane		Soluble length-matched	
	Kinked	Straight	Kinked	Straight
Proline in helix	29.6	18.3	17.0	12.2
No proline in helix	9.9	42.3	15.0	55.7
Proline in profile	33.9	27.8	23.2	26.3
No proline in profile	5.5	32.8	8.9	41.6
Proline in helix (5th and subsequent residues)	26.7	4.8	13.5	0.7
No proline in helix (5th and subsequent residues)	12.7	55.8	18.5	67.3
Proline in profile (5th and subsequent residues)	31.1	10.3	18.9	5.4
No proline in profile (5th and subsequent residues)	8.3	50.2	13.2	62.5

4.3.2.1 Proline

Figure 4.5 shows the amino acid propensities around kinks ($\geq 20°$) in the membrane and length-matched soluble helix sets. For both membrane and soluble helix kinks, proline is present most strongly at position +2 (two residues C terminal to the kink, see Figure 4.4), but also at other positions: +1,+5, and +6. These are all positions on the *outside* of the kink. Conversely it is disfavoured at positions 0 and +4, the *inside* of the kink. In soluble kinks proline is also slightly favoured at positions before the kink (e.g. -2, -3, and -6), which are also on the outside of the kink.

My analysis has revealed that in both types of helices proline typically occurs on the outside of kinks. This is due to both the physical size of its ring, and its lack of amide proton precluding the $i+4 \rightarrow i$ backbone hydrogen bond. Both of these factors tend to lengthen the distance between the i^{th} residue (proline) and the $i-4^{th}$ residue, which causes the helix to kink away

from the side of the helix containing proline.

Proline is very important to kinks. Approximately half of the membrane helices contain a proline, compared to only one third of the length-matched soluble helix sets (Table 4.1). In terms of its ability to disrupt the $i+4 \rightarrow i$ backbone hydrogen bond, proline must be at least four residues from the N-terminus of the helix. Sixteen percent of transmembrane helices, and on average fifteen percent of the length-matched soluble helix sets contain a proline only in the first four residues. In these helices, we do not expect proline to cause a kink. Classifying these helices with prolines only in the first four residues as non proline containing, 95% of soluble helices, and 85% of membrane helices that contain proline are kinked.

These prolines are not necessarily close to the kink position identified above, but this trend does indicate that presence of proline is likely to be linked to helix kinking. For comparison, however, not all kinked helices contain proline: 36% of kinked transmembrane helices contain no proline, and a clear majority (61%) of (length matched) kinked soluble helices do not contain a proline. Additionally, there is a large proportion of kinked helices in both sets that do not contain a proline even in the homologous sequence alignment.

4.3.2.2 Other amino acids

For other residues, the patterns in Figure 4.5 are less obvious, but those discussed below are seen in the results of both the Kink Finder and MC-Helan analyses. In soluble helices, glycine is overrepresented on the inside of kinks (positions -4, 0, +4 (see Figure 4.4)), and slightly overrepresented at position 0 in membrane helix kinks. This may be due to the flexibility of glycine, or the lack of a sidechain allowing the helix to bend towards the glycine.

Serine is overrepresented in positions before the kink (-2 and 0 for membrane kinks, and -3 for soluble kinks), as is asparagine (position -3 for both soluble and membrane). The aromatic residues (phenylalanine, tyrosine, and tryptophan) show some clear periodic patterns in the soluble propensities, being overrepresented at positions -5, -4, -1, 0, +3 and +4. There are similar but weaker patterns in the membrane propensities - for example phenylalanine is overrepre-

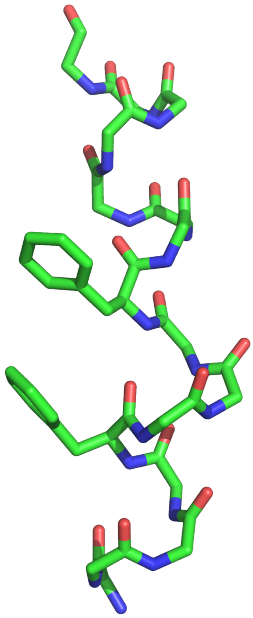


Figure 4.6: **Example of FxxxF kink motif.** Taken from chain A of protein 3eml, residues 75-90. Only the backbone atoms and the two phenylalanine side chains are shown.

Table 4.2: **Frequency of broken backbone hydrogen bonds in kinks and helices.** Each value is the percentage of residues that have no mainchain to mainchain hydrogen bond to their carbonyl group. The second row excludes kinks and helices that contain proline residues.

	Membrane			Soluble		
	Kinks	Kinked Helices	Straight Helices	Kinks	Straight Helices	Kinked Helices
Residues missing backbone hydrogen bond	15.6%	12.8%	6.7%	13.5%	11.1%	5.7%
Residues missing backbone hydrogen bond, excluding prolines	14.6%	12.3%	6.3%	11.8%	9.7%	5.7%

sented at positions +2, +3, and +4. Although it might be expected that large residues would be found on the outside of kinks due to their size, the aromatic residues are more frequently found on the inside of kinks. This suggests that there may be some pi-stacking interactions that stabilise kinks, an example of which is shown in Figure 4.6.

There is a hydrophobicity pattern in the soluble propensities, where we see small positive propensities from many hydrophobic residues at -4, 0, +3 and +4 (positions on the inside of the helix kink), and corresponding negative propensities for charged residues at these positions. This suggests that kinks in soluble protein helices favour having charged residues on the outside of the kinks, and hydrophobic residues on the inside of the kink (Figure 4.4).

4.3.3 Hydrogen bonds

Many of these residue patterns are thought to be due to the role of hydrogen bonds in kinks. While proline cannot form hydrogen bonds from its amide nitrogen, other residues, like serine and threonine, may stabilize kinked helices which lack backbone hydrogen bonds. I found that broken backbone hydrogen bonds are equally common in the membrane and length-matched soluble sets of helices (12.8% and 11.1% of residues in kinked membrane and soluble helices respectively), and equally more common in the 13 residues around kinks (15.6% and 13.5% of residues in membrane and soluble kinks, see Table 4.2).

$i+4 \rightarrow i$ backbone hydrogen bonds are more frequently broken in kinks, and particularly on

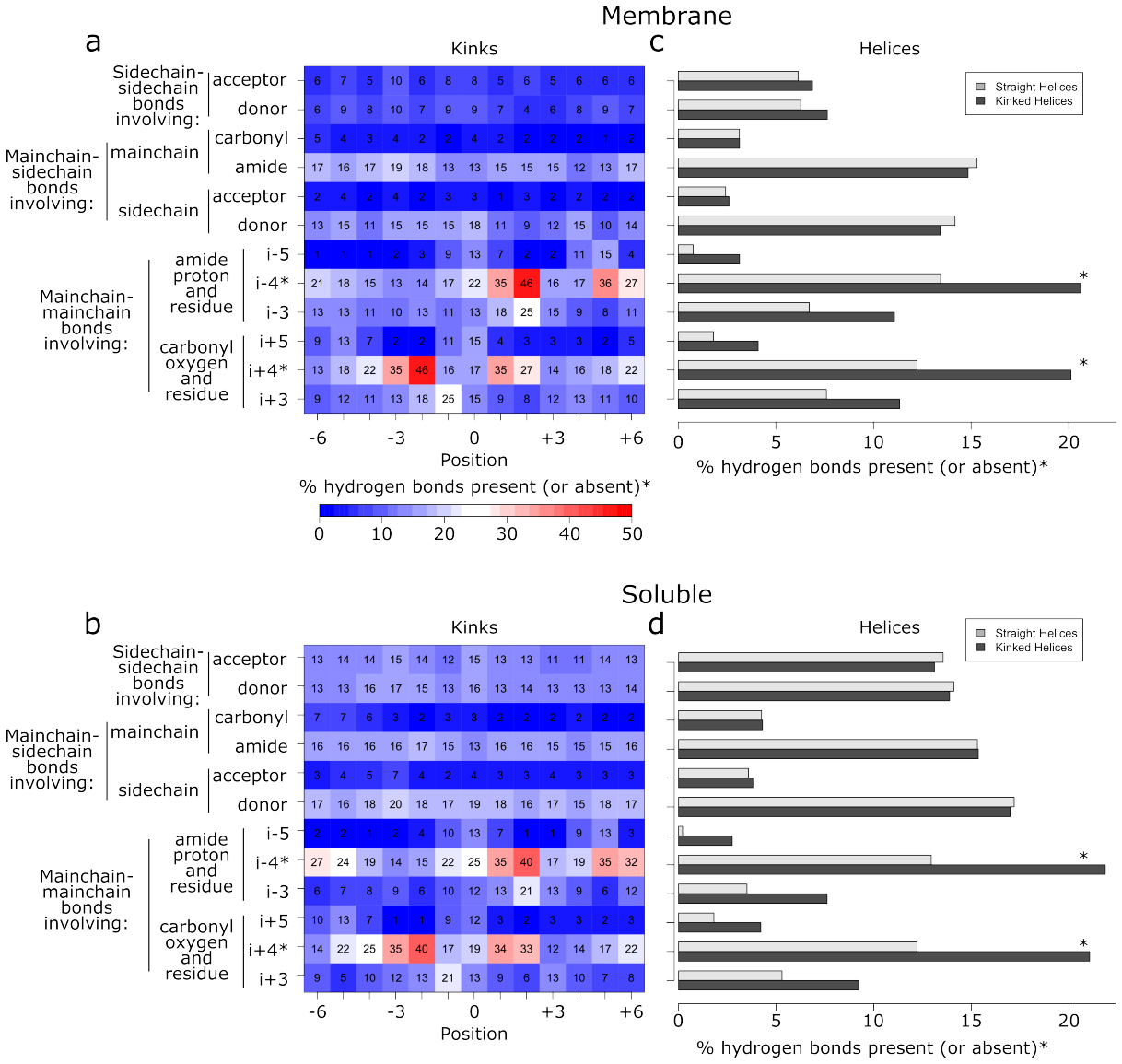


Figure 4.7: **Hydrogen bonds in membrane and soluble kinks and helices.** (a) and (b) show the percentage of residues around kinks (for numbering, see Figure 4.4) that have hydrogen bonds involving the stated atom type. The i to i+4 and i to i-4 rows (starred) show the percentage of residues without that hydrogen bond. E.g. the bottom row shows the proportion of residues that have a hydrogen bond involving that residue's backbone carbonyl oxygen, and the amide hydrogen atom of the residue 3 positions towards the C-terminus. (c) and (d) show the percentage of residues that have hydrogen bonds of the type indicated in straight (pale grey) and kinked (dark grey) helices.

the outside of kinks (Figure 4.7), than in helices on average. Conversely, $i+3 \rightarrow i$ and $i+5 \rightarrow i$ hydrogen bonds are more frequently seen around kinks than they are seen in the average helix. This disruption in the canonical α-helical hydrogen bonding pattern suggests that there is a general disruption of the normal α-helical structure around kinks. In positions around kinks where $i+3 \rightarrow i$ and $i+5 \rightarrow i$ hydrogen bonds are more frequently present, $i+4 \rightarrow i$ backbone hydrogen bonds are more frequently broken. This indicates that in kinks, $i+3 \rightarrow i$ and $i+5 \rightarrow i$ hydrogen bonds act to compensate for the loss of the canonical $i+4 \rightarrow i$ backbone hydrogen bonds.

The frequency of the different types of mainchain-mainchain hydrogen bonds are different for straight and kinked helices. However, there is no increase in frequency of sidechain to backbone or sidechain to sidechain hydrogen bonds around kinks, or in kinked helices compared to straight helices. It is clear that side chains have no systematic effect on the kinking of protein helices, as they are equally frequent in straight and kinked helices. These results are very similar in the two types of protein.

4.3.4 Motifs

Previous studies have suggested a number of sequence motifs which may cause kinks. Some of these have been suggested from observation in structures (e.g. (Deupi *et al.*, 2004; Devillé *et al.*, 2008; Hall *et al.*, 2009)), while others have been identified from conserved sequence patterns (e.g. (Marsico *et al.*, 2010a,b)). The occurrence of these motifs in both kinked and straight helices are shown in Table 4.3. Table 4.3 also contains motifs that I identified as likely to be important based on the amino acid propensities. For example, the propensities for the aromatic residues vary periodically with position (Figure 4.5), with a period of between three and four residues. This suggested that sequence motifs containing aromatic residues separated by two or three other residues may be enriched in kinks. Similarly, glycine propensities vary periodically, suggesting that small residue containing motifs may be over represented in kinks. This analysis uses only the sequences from the helix structures. Relevant motifs should be both frequently

observed in kinks, and also more frequently observed in kinks than comparable segments of straight helices. Motifs that are present in more than 10% of kinks, and motifs that are present more than twice as frequently in kinked helices as compared to straight helices are highlighted in bold in Table 4.3.

Table 4.3: **Table of amino acid motif frequency in kinks and randomly selected parts of straight helices.** Motifs that occur in more than 10% of kinks, or are more than twice as frequent in kinks than straight helices are highlighted in bold. First section: proline containing motifs. Second section: motifs highlighted by other authors. Third section: aromatic and polar motifs. Fourth section: small residue motifs. Ar = aromatic residue (F, Y or W). See Table B2 for MC-Helan data. Notation is described in Section 4.2.10. 'x' means any residue and square brackets indicate any one of the residues contained within. For example, xxxP matches any three residues followed by a proline and xP matches any residue followed by proline.

	Membrane			Soluble		
Motif	% Kinks	% Straight Helices	Ratio	% Kinks	% Straight Helices	Ratio
[AVILMFYW]xxxP (Devillé *et al.*, 2008)	**38.8**	0.9	**41.3**	**12.8**	0.1	**181.0**
xxxxP	**58.7**	1.7	**34.3**	**35.6**	0.2	**154.2**
[ST]P (Deupi *et al.*, 2004; Hall *et al.*, 2009)	5.3	0.0	**255.2**	3.4	0.2	**22.2**
[DR]P (Sansom & Weinstein, 2000)	1.5	0.4	**4.1**	3.4	0.1	**43.4**
xP	**61.1**	6.5	**9.4**	**37.9**	1.8	**21.3**
xxxP	**58.8**	2.3	**26.1**	**36.2**	0.3	**118.7**
[AVILMFYW]xP (Devillé *et al.*, 2008)	**47.0**	2.9	**16.1**	**24.5**	0.5	**48.4**
GxP	3.8	0.2	**16.7**	1.4	0.0	**55.7**
xxP	**60.2**	3.8	**16.0**	**37.1**	0.8	**48.0**
P[ST] (Deupi *et al.*, 2004; Weber *et al.*, 2012)	4.3	0.6	**7.1**	2.7	0.4	**6.1**
Px	**55.6**	9.9	**5.6**	**34.8**	4.7	**7.4**
P	**61.4**	10.0	**6.1**	**39.0**	4.7	**8.3**
[VALT]LWx[AG]YP (Marsico *et al.*, 2010a)	0.2	0.0	NA	0.0	0.0	NA
GHPxVY[FI] (Marsico *et al.*, 2010a)	0.0	0.0	NA	0.0	0.0	NA

Continued on next page

Continued from previous page

[ST][ST] (Del Val *et al.*, 2012; Devillé *et al.*, 2008)	**10.6**	10.8	1.0	9.1	8.2	1.1
[ST]xx[ST] (Del Val *et al.*, 2012)	9.8	10.2	1.0	8.4	7.0	1.2
[ST]xxx[ST] (Del Val *et al.*, 2012)	8.2	9.2	0.9	8.1	6.8	1.2
GxxGxxxG (Hall *et al.*, 2009)	1.4	1.0	1.3	0.0	0.1	0.0
GxxxGxxG (Hall *et al.*, 2009)	0.7	1.1	0.6	0.2	0.0	**5.2**
ArxxxSxxxAr (Hall *et al.*, 2009)	0.2	0.3	0.5	0.3	0.2	1.8
LSAxF (Hall *et al.*, 2009)	0.0	0.3	0.0	0.1	0.0	NA
WLF[ST] (Marsico *et al.*, 2010b)	0.5	0.0	NA	0.0	0.0	NA
ArxxxAr	**21.4**	12.3	1.7	9.6	6.0	1.6
ArxxxArxxAr	6.3	0.7	**9.1**	0.8	0.5	1.5
ArxxxArxxxAr	4.3	1.0	**4.3**	0.8	0.4	**2.1**
ArxxArxxxAr	3.3	1.2	**2.7**	1.1	0.4	**2.9**
[RHDEK]xxx[RHDEK]	7.7	8.9	0.9	**49.1**	57.6	0.9
[ASG]xxx[ASG]	**38.6**	47.0	0.8	**27.7**	32.8	0.8
[STNQ]xxx[STNQ]	**18.7**	17.0	1.1	**25.3**	25.3	1.0
G_0 or G_{+1}	**15.6**	17.7	0.9	**10.5**	6.5	1.6

I restricted the search for motifs to the 13 residues around a kink (positions -6 to +6 in Figure 4.4). This allows a comparison to a background distribution of motifs in 13 residue sections of straight helices. The choice of a window means that longer motifs are less likely to be found than shorter ones (e.g. P is more frequently found that xxxxP). This is true for both the kinked and straight sections, so both the frequency of a motif and its enrichment ratio (i.e. comparing the number of times a motif is observed in kinked and straight helices), are important. Interesting motifs are both frequent (in more than 10% of kinks) and have an enrichment ratio above 1.

Except those that involve proline, very few of the motifs are observed in more than 10% of kinks. Similarly, few motifs are seen more than twice as frequently in kinks compared to straight helices (Table 4.3). Motifs that contain proline are generally no more enriched in kinks than the corresponding sequence length containing just proline. For example, the GxP motif has an enrichment ratio of 16.7 (i.e. it is found 16.7 times more frequently in kinks than straight helices), and the xxP motif has an enrichment ratio of 16.0. The [ST]P motif is an exception (found in 5% of membrane kinks), being very rarely in sections of straight membrane helices (enrichment ratio 255). Unlike the [ST]P motif, motifs containing serine/threonine without proline (e.g. [ST][ST]) are not seen more frequently in kinks than in straight helices.

It has been suggested in some studies that small residues (particularly glycine) allow flexibility in helices (Devillé *et al.*, 2008; Hall *et al.*, 2009; Kneissl *et al.*, 2011). The soluble propensities show glycine as overrepresented at positions 0, +1, +4, and +5, suggesting that there may be motifs such as GxxxG present in kinks. However, I found no small-residue motifs that occur more frequently in kinks than in straight helices. The 'G_0 or G_{+1}' motif, which counts the number of examples of glycine at the kink residue, or the next residue, has an enrichment of 1.6 in soluble helices, suggesting that it could be a kink indicator, although this is not replicated in the membrane helices.

As described in the previous section, aromatic residues had a high propensity to be around helix kinks. We find the [FYW]xxx[FYW] (ArxxxAr in Table 4.3) motif to be slightly enriched for both membrane and soluble kinks. It occurs in 23% of membrane kinks, compared to only 12% of straight membrane helices. Although this motif is less frequent in soluble kinks (in 10% of kinks), it is similarly more frequent than in the straight soluble helices (6%). The longer aromatic motifs are more enriched in membrane protein kinks, but much less frequent than the [FYW]xxx[FYW] motif. Finally, we note that many of the remaining motifs are found in very few structures in our set, which prevents us from assessing if they are associated with kinks.

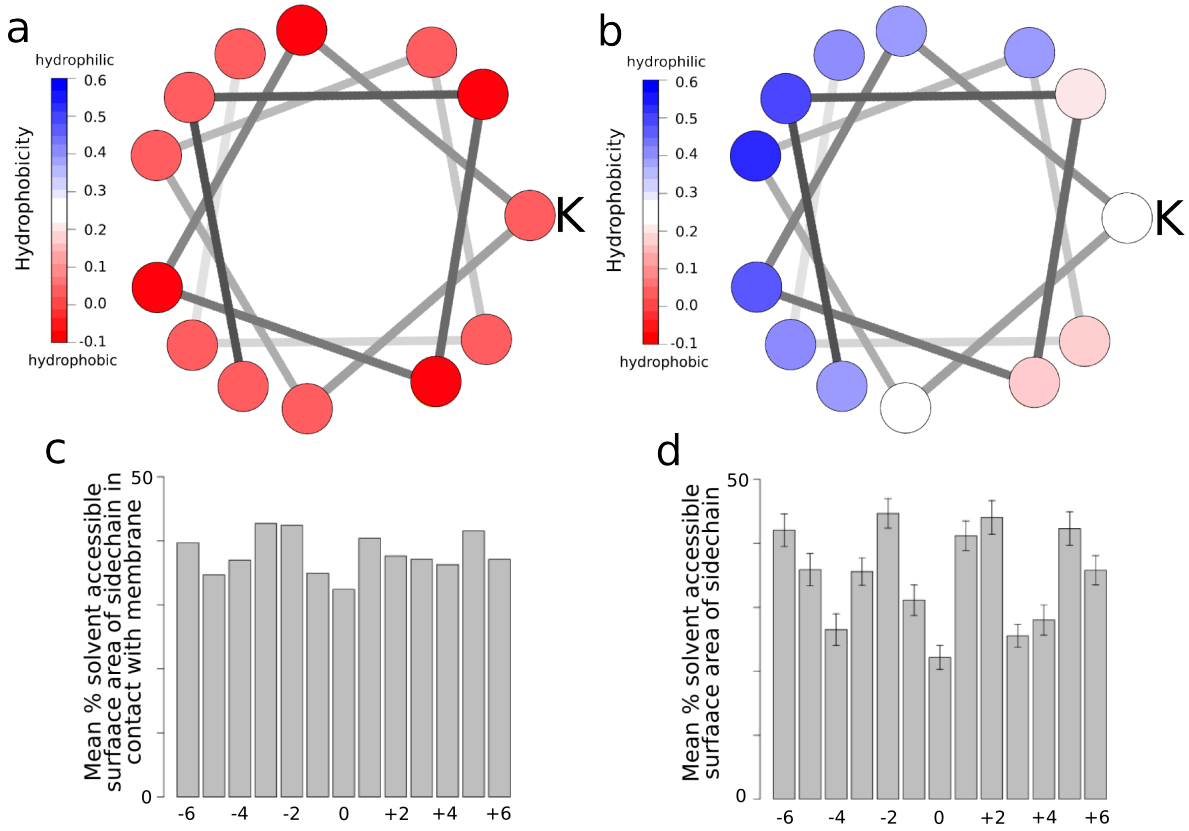


Figure 4.8: **Hydrophobicities, solvent accessible surface areas, and membrane contacts.** (a) Helical wheel diagram showing the average hydrophobicity of residues around membrane kinks. K indicates the kink residue (position 0 in Figure 4.4). (b) Wheel diagram for soluble kinks. K indicates the kink residue. (c) Average percentage of residue in contact with the membrane in kinks. (d) Average solvent accessible percentage of residues in soluble kinks. See Figure B2 for the analogous figure based on the MC-Helan analysis.

4.3.5 Hydrophobicity and solvent accessible surface area

The amino acid propensities around kinks in Figure 4.5 show a periodic pattern. Hydrophobic residues are more frequently observed in positions on the inside of kinks (-4,0,4), while charged and polar residues are less frequent at these positions. The mean hydrophobicity of each position in membrane and soluble kinks are shown in Figure 4.8.

There is a clear pattern in the hydrophobicity of soluble kinks, with more hydrophobic residues on the inside of kinks, and more hydrophilic residues on the outside of kinks. This pattern is repeated in the fraction of residues which are solvent accessible, with residues on the outside of kinks being more solvent accessible than those the inside (Figure 4.8d). This indicates that soluble kinks point into the aqueous environment, meaning that the residues on the outside of the kink will be in the solvent (Figure 4.4).

For the membrane proteins, rather than the solvent accessible surface area, I calculated the percentage of residue sidechains in contact with the membrane. This is the fraction of accessible side chain (taken from the JOY output) if iMembrane indicates that the side chain is in contact with the head or tail region of the membrane, and zero if iMembrane indicates that the side chain is not in contact with the membrane. This does not replicate the pattern seen in the solvent accessible surface area of soluble kinks. Solvent accessible surface area (SASA) calculations do not take the membrane environment into account, so it is not a good measure to use with membrane proteins. Nevertheless, Werner & Church (2013) used SASA as a measure for membrane proteins. The plot of membrane kinks SASA (not shown) is very similar to the soluble kinks SASA (Figure 4.8).

This pattern of hydrophobicity and solvent accessible surface area may be related to the effect of solvent on hydrogen bonds in helices. Backbone hydrogen bonds on the hydrophilic side of amphipathic helices are generally longer than those on the hydrophobic side (Blundell *et al.*, 1983) so that helices curve, or kink, away from the solvent side. This is due to the other available hydrogen bond donors and acceptors on the hydrophilic side of the helix (e.g. sidechains, or solvent), which can form bonds to the backbone groups, withdrawing electron density from the

backbone hydrogen bonds, causing them to lengthen. As expected, this hydrophobicity pattern is not repeated in the membrane kinks, as in this case the helix is surrounded by the rest of the protein or lipid.

4.4 Discussion

The investigations described in this chapter shows that kinked helices are not particular to membrane proteins. If soluble helices with the same length distribution as membrane helices are considered, a similar number of kinks are seen in membrane and soluble helices. Compared to membrane helix kinks, there are more soluble kinks with larger angles, which may be due to the membrane environment restricting the degree to which a helix can kink. This overall similarity of kinks in soluble and membrane helices is independent of the method used to assess their occurrence. Kink residues in this study are identified such that they are in a geometrically similar place with respect to the kink. However, since no other method takes the geometry of the kink into account when selecting the kink residue, the results reported here will not necessarily agree with those in earlier studies.

Proline is dominant in both membrane and soluble helix kinks, although there are more prolines incorporated into long membrane helices than long soluble helices. Excluding the first four residues of the helix makes proline a much better indicator of kinks than considering the whole helix. The vast majority of proline containing helices are kinked, but there are many kinked helices that do not contain proline. The residue patterns around kinks are dominated by proline, both for membrane and soluble kinks. The consistent choice of kink residue relative to the helix shape (the kink residue is on the inside of the kink), reveals that proline occurs on the outside of the helix kinks. This consistent positioning reveals a number of other patterns - glycine is favoured on the inside of the kink, serine is favoured before and on the outside of the kink, and aromatic residues are favoured on the inside of kinks. These patterns are observed in both the membrane and soluble helix sets.

Motifs containing aromatic residues are more frequently observed in kinks than straight helices, both in soluble and membrane proteins. Particularly, the [FYW]xxx[FYW] motif (an example of which can be seen in Figure 4.6) is a factor of 1.9 more frequent in membrane kinks than in straight membrane helices, and is found in 23% of membrane helix kinks. Many other motifs highlighted in previous research are either seen very infrequently or no more frequently than in straight helices. Machine learning approaches (for example, inductive logic programming) may be able to identify novel motifs, although the lack of membrane data makes such approaches prone to false positives. Indeed, many of the results presented here are contrary to those presented in one such study (Marsico *et al.*, 2010a). The increase in available membrane protein structural data will make machine learning approaches more applicable in the future. Allowing for the expected differences in residue frequencies in the two types of helix, the amino acid patterns around helix kinks are very similar in membrane and soluble proteins.

A strong hydrophobicity pattern is observed in soluble kinks - where solvent accessible and hydrophilic residues are seen on the outside of kinks, indicating that soluble kinks point into the solvent. While there is no hydrophobicity pattern in the membrane kinks, there is a tendency for aromatic residues to be observed on the inside of membrane kinks. A recent study (Werner & Church, 2013) indicated that kinks were more buried than helices, specifically that, on average, amino acids in nine residue kink sections had a lower solvent accessible surface area and more residue contacts than the average residue in a helix. The effect was small but significant. The nine residues around the kinks (i.e. residues -4 to +4, see Figure 4.4) identified in this chapter have, on average, much lower solvent accessible surface areas than the average (Figure 4.8), although the thirteen residues, on average, have a higher SASA. The (unpublished) kink finding method used in that study (Werner & Church, 2013) is likely to have a bias towards annotating kink residues on the inside of the kink, which would explain the observation reported in their study. This highlights the importance of consistently selecting the kink residue with respect to the shape of the helix.

4.5 Conclusion

The results in this chapter show that soluble helices are equally likely to be kinked as their membrane counterparts, and that these soluble kinks point into the solvent. This suggests that kinks are an important structural feature of soluble proteins and may, like their membrane counterparts, be functionally important. Functional and structural importance of kinks could be assessed by using online databases of function (e.g. Prosite (Sigrist *et al.*, 2013)), or by manually inspecting the protein structure, associated literature, and conservation patterns. Further, we can apply knowledge gained from the much larger set of soluble protein structures to the modelling of membrane protein kinks. Information about sequence patterns and solvent accessibility can be built into predictors in many ways. The most simple method is to use this to score sequences for their similarity to kinks. There are many more complex methods that involve machine learning (e.g. Kneissl *et al.* (2011); Meruelo *et al.* (2011); Werner & Church (2013)), however these are prone to overfitting on the relatively small amount of membrane protein data. However, this chapter has shown that soluble kinks are similar to membrane protein kinks, and therefore membrane kink prediction methods could be validated against the larger soluble data set. Predicting when helices are kinked could be used in the modelling process at the template selection stage, and for scoring the quality of generated models. The results in this chapter suggest that this approach, i.e. using data from all known kinks to predict kinks, is unlikely to be successful. The current methods do not outperform simple metrics based on the length of helices and the presence of proline. More specific approaches, that use information from proteins that are homologous to the protein of interest, are more likely to prove useful for kink prediction. The next chapter describes my study of kinks in related proteins.

CHAPTER 5

How well are kinks structurally conserved?

This work follows on from work carried out by another student, Eleanor Law. She undertook a series of preliminary studies during an undergraduate project that I helped to supervise.

5.1 Introduction

This chapter describes my study on the structural conservation of α-helix kinks within protein families. The structural conservation of kinks is important in the prediction of membrane protein structures, and as an indication of the function and importance of kinks in protein structure.

In terms of membrane protein structure prediction, correctly predicting changes in the core (here the core is defined as the part of the protein buried in the tail region of the membrane) between target and template is a known problem (Chen *et al.*, 2014; Kelm *et al.*, 2010; Kufareva *et al.*, 2011, 2014). Very frequently during homology modelling, the 'best' structure prediction of the target sequence (as measured by RMSD or GDT-TS (Zemla *et al.*, 1999)), uses a protein core copied from the template. Indeed, this is one of the central principles in our group's membrane protein modelling pipeline, Memoir (Ebejer *et al.*, 2013), and specifically MEDELLER (Kelm *et al.*, 2010), the coordinate generating algorithm contained within Memoir.

Kinks are functional structural motifs in proteins (see Section 1.5.1). They are known to be flexible in some proteins, including GPCRs (Barrett *et al.*, 2012; Fowler & Sansom, 2013; Katritch *et al.*, 2013). Knowledge of their structure is important to protein structure prediction. Similarly, understanding how flexible they are, and what makes them flexible is important in understanding the function and modelling of membrane proteins. Kinks play a part in biologically crucial functions, in pharmaceutically relevant membrane proteins.

In this chapter, I aim to answer the following questions:

1. Are kinks conserved in protein families?
2. If so, how conserved are they? Are some kinks more conserved than others?
3. What are the features indicative of conserved and/or non-conserved kinks?

Although previous research has focused on membrane proteins, and the primary interest is in these, soluble proteins provide a much greater structural diversity. The small number of similar membrane protein structures could result in a bias in any results. Hence, as in the previous chapter, I use both membrane and soluble protein structure data sets to investigate these questions.

As described in Chapter 3, kink classification is difficult. However, characterising the differences between the kinkedness of two helices is more difficult and has not been attempted previously.

In Chapter 4, I used a simple binary classification of helices (kinked/straight), however this work requires a different approach. A binary classifier would conclude that helix A (just the kinked side of a threshold) and helix B (just the straight side of a threshold), are differently kinked. The simplest solution to this problem would be to require the angles of the helices to differ by a given amount for them to be classified as different. However, there is no good guidance on what this difference should be.

Although most kink identification methods give an absolute measurement of the angle of each kink, none provide an estimate of the uncertainty of the angle. The measured kink angle depends on the way in which axes have been fitted to the helix. If the structure around the kink is not completely regular, these fits may be poor approximations of the true axis. In this chapter, I use the quality of these fits to estimate the error in the measured angle, thus allowing me to identify when two helices are differently kinked.

In this chapter, I show that kink angles are estimated within $\pm 6^\circ$ 95% of the time when they are based on good cylinder fits, and within $\pm 12^\circ$ 95% of the time when based on the worst cylinder fits. The helices identified as kinked by more AHAH participants (see Chapter 3) tend to have lower error than those kinked helices that were identified as kinked by fewer participants.

Two data sets were used to consider the conservation of kinks. From a set of soluble proteins, 226,072 pairs of homologous helices were identified, and from a set of membrane proteins 1864 homologous helices pairs. Among the membrane (soluble) aligned homologous helix pairs, 200 (24,784) helix pairs were found to be differently kinked, using the error method described below. Dissimilarly kinked aligned helix pairs have less similar sequences - both at the chain and helix level - than the similarly kinked helix pairs. However, there is no more difference in the helix sequences than in the chain sequences. Loss of proline correlates with the loss of a kink in 60% of membrane and 80% of soluble aligned homologous helix pairs. The study was then extended from pairs to families, where I found that there are families of proteins that contain conserved kinks, and families that contain non-conserved kinks.

5.2 Methods

5.2.1 Data sets

The two data sets used for the pairwise helix comparisons are identical to those in Chapter 4. One set is a collection of membrane protein structures, the other soluble protein structures. Both are culled to less than 80% sequence identity, and all structures have R value ≤ 0.4 and resolution better than 5Å. The initial list of membrane proteins are derived from Steve White's database (White & Wimley, 1999) and the PDBTM (Kozma *et al.*, 2013). Helices were annotated using JOY (Mizuguchi *et al.*, 1998), and trimmed using a method described in Chapter 2. The structure annotations, such as hydrogen bonds and solvent accessible surface area, were generated using JOY and iMembrane (Kelm *et al.*, 2009), as described in Section 4.2. Similarly, the sequence profiles were generated in the same way as in Section 4.2.

The comparison to the AHAH results uses the same 300 helices taken from the Kneissl *et al.* (2011) data set as in the AHAH study (Chapter 3).

5.2.2 Error in the measured kink angle

The cylinder fitting method provides a goodness of fit parameter - similar to a root mean square deviation - for each fitted axis. In the next section, I describe the method I developed to relate this to the error in the measured angle.

5.2.2.1 Relating RMSD to error

The error in the measured kink angle is estimated from the goodness of fit, r, of the two axes used to determine that angle. The fitting function for the axis (cylinder) fit is shown in Equation 5.1;

$$r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \hat{d}\right)^2} \tag{5.1}$$

where n is the number of points (atoms) that the axis is fitted to, d_i is the shortest distance from atom i to the fitted axis, and $\hat{d}$ is the mean of these distances.

The module of Kink Finder that estimates the error in the angle is a heuristic method, based on observed errors in 'ideal' kinks with good cylinder fits. The angle measured by Kink Finder can be expressed as:

$$\text{measured angle} = \text{true angle} + \epsilon \tag{5.2}$$

Note that this means that the error, ϵ has the same shape of distribution as the measured angle. Each cylinder fit has a value of r (Equation 5.1), which measures the distance of the atoms from the cylinder surface. The calculation of a kink angle requires two cylinder fits, one for the set of six residues N-terminal of the kink position, and one for the set of six residues C-terminal to the kink (see Figure 5.1). I assumed that the goodness of fit parameters of these two fits (r_n and r_c, i.e. the r for the N- and C-terminal cylinder fits) have an equal effect on the error, ϵ.

For each calculated angle, I approximated the 'goodness of fit' with the sum of the r of the two cylinder fits. Although there is no way to directly measure the 'true' angle, I used 18 ideal kinks to estimate the effect, assuming that the fitted axes for these provided the 'true' angle. These 18 kinked helices (with the lowest $r_n + r_c$, of the kinks in the membrane protein set) have a range of true angles between 0° and 50°. The $r_n + r_c$ for all of these 18 is below 0.6Å. As I use all backbone atoms, rather than just the Cα atoms, to fit the cylinders, the $r_n + r_c$ cannot reach 0Å. In the protein the atoms in and directly connected to the amide (peptide) bond are in the same plane (i.e. the backbone Cα, C, O, and N atoms of one residue, and the Cα atom of the next residue, see Figure 1.3), so there is no way to fit a cylinder that passes through every backbone atom. Consequently, even for an ideal helix, the r cannot be less than 0.27Å.

Taking these 'ideal' kinks, I simulated the relationship between $r_n + r_c$, the true angle, and ϵ. For each measurement in each kink, both cylinders were rotated about their midpoint by

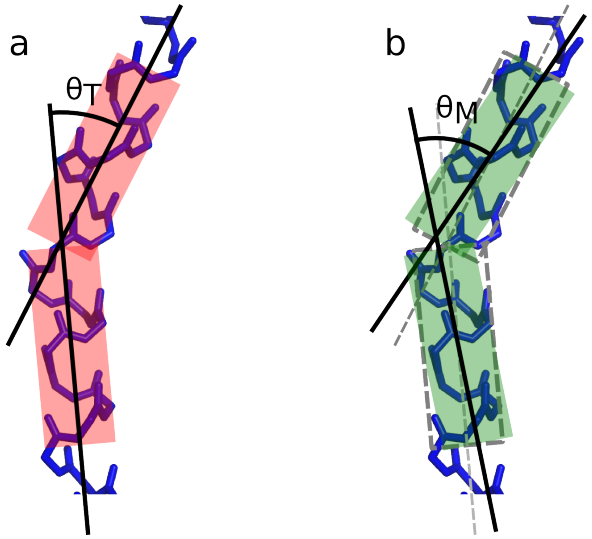


Figure 5.1: **Relating angle error to goodness of fit.** (a) Example 'ideal' kink, with low r_n and r_c. The true angle (θ_T) is the angle between the two fitted axes. (b) Cylinders are rotated (in green) from their fitted positions (dashed lines), and a measured angle (θ_M) is calculated. r_n and r_c are calculated from the rotated cylinders (green). Carrying out this rotation many times provides the data for Figure 5.2.

an angle and direction, using a randomly generated rotation matrix. This provides a series of measured angles based on non-optimised cylinder fits. The $r_n + r_c$ and measured angle are recorded for each (see Figure 5.1), and used to characterise the relationship between the two. Example results for two kinks are shown in Figure 5.2. The other kinks show similar behaviour.

This process is repeated for each of the 18 ideal kinks. For a given range of $r_n + r_c$, the error (and similarly, the measured angle), has a distribution that is close to normal. In order to estimate error, I assume that it is normally distributed, i.e:

$$\epsilon \sim N\left(0, \sigma_\epsilon^2\right). \tag{5.3}$$

where $N(a, b)$ indicates a normal distribution with mean a, and variance b, and σ_ϵ^2 is the variance of this distribution.

The data is binned based on $r_n + r_c$, and for each bin I assume that the measured angle

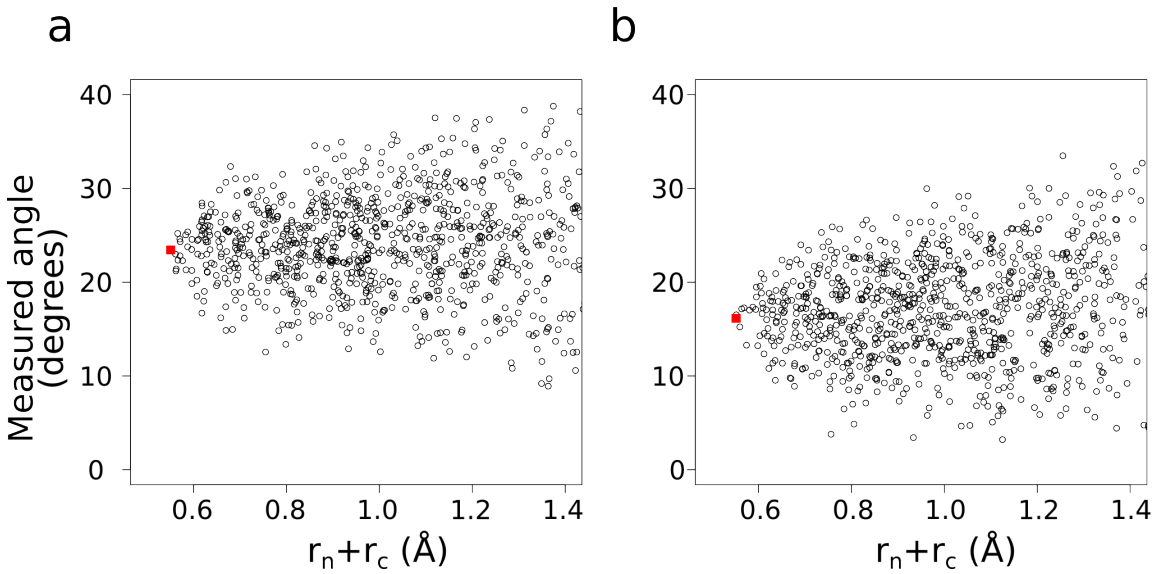


Figure 5.2: Measured angle against goodness of fit ($r_n + r_c$) for two kinks. (a) At residue 255 in chain A of protein 1PB2. (b) At residue 259 in protein 1Y2LA. The red squares indicate the angle and $r_n + r_c$ for the optimum cylinder fits.

is normally distributed. For each bin in each kink, the error is summarised with the standard deviation of the angle error (which is the same as the standard deviation of the measured angle - see Equation 5.2). Figure 5.3 shows how, when excluding kinks under 10°, the size of the error is not affected by the true angle - it only depends on the value of $r_n + r_c$. A 20° kink and a 50° kink, with the same $r_n + r_c$ values have the same magnitude of uncertainty in the kink angle. Therefore, I combined the errors for all of the kinks with angles above 10° to calculate the relationship between $r_n + r_c$ and angle error. I used a statistical confidence interval, rather than using the standard deviation, as this does not rely on the assumption of normality.

The error for all 12 kinks over 10° was binned into ranges of $r_n + r_c$. The modulus of the error values was taken, and the value at the 95% percentile of each bin was taken. The resulting relationship is discussed in the results, Section 5.3.1

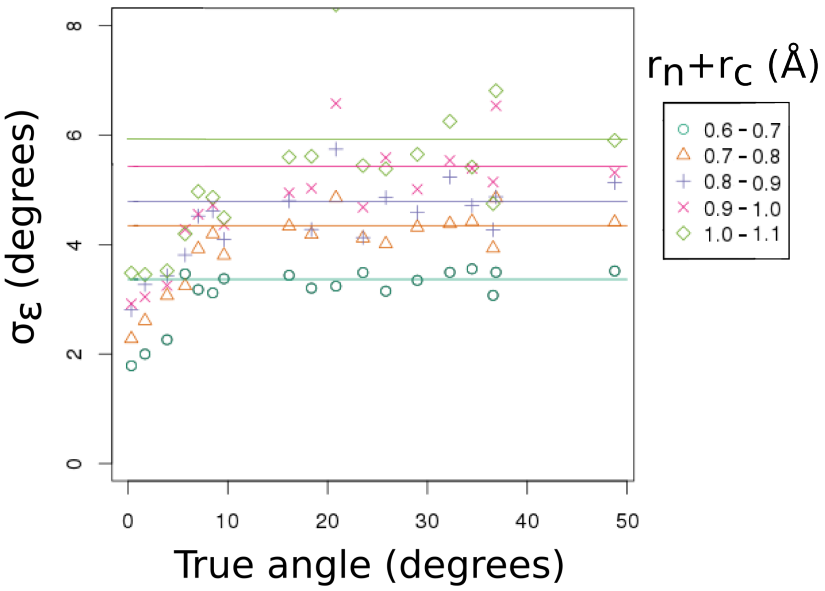


Figure 5.3: **Standard deviation of error for kinks**. The standard deviation of the angle error (measured angle - true angle) for bins of $r_n + r_c$ (y-axis) are shown for 18 ideal kinks, plotted against their true angle as determined by the optimised cylinder fits. The standard deviation of the angle error for a given range of $r_n + r_c$ is constant for angles above 10°. Horizontal lines are fitted to the points where true angle $> 10°$ for each range of $r_n + r_c$.

5.2.3 Identifying homologous aligned helix pairs in the data sets

Separately for each data set (soluble and membrane), I identified homologous aligned helix pairs (for an overview, see Figure 5.4). In order to identify aligned helix pairs within a set first I identified all homologous proteins in the set, and then I extracted aligned helices in these homologues. I performed an all-against-all structural comparison of the protein chains in the two sets described in Chapter 4, using TMAlign (Zhang & Skolnick, 2005). Pairs of chains are not considered as homologues if either:

- the number of residues in the longer chain is more than 50% greater than the number in the shorter chain; or,
- the TM-score of the alignment was less than 0.5.

Both of these requirements are somewhat empirical. The first is a fast way to remove pairs of chains that are unlikely to be evolutionarily related, and the second is the threshold of similarity considered to show two structures have the same fold (Zhang & Skolnick, 2005). As in Chapter 4, helices and kinks were annotated using Kink Finder, but this time output included a confidence interval for each output angle, as described in the previous section. In a homologous pair of chains, aligned helix pairs were extracted if their ends were offset by no more than four residues in the TM-align alignment. For an aligned helix pair, the largest angle in one of the helices, and the largest angle in a 9 residue window around the corresponding position in the other helix were compared (Figure 5.5). This was repeated for the other helix in the pair, meaning that for each helix pair, two comparisons resulted. These were only different where the maximum angles in each of the helices were not within 4 residues of one another.

5.2.3.1 Helix pair properties

Proline was identified as present in a helix if it occurred in the helix, with the exception of the first turn (see Chapter 4). A gap was identified in a helix pair where there was a gap in between the bounds of the helix alignment calculated by TM-Align.

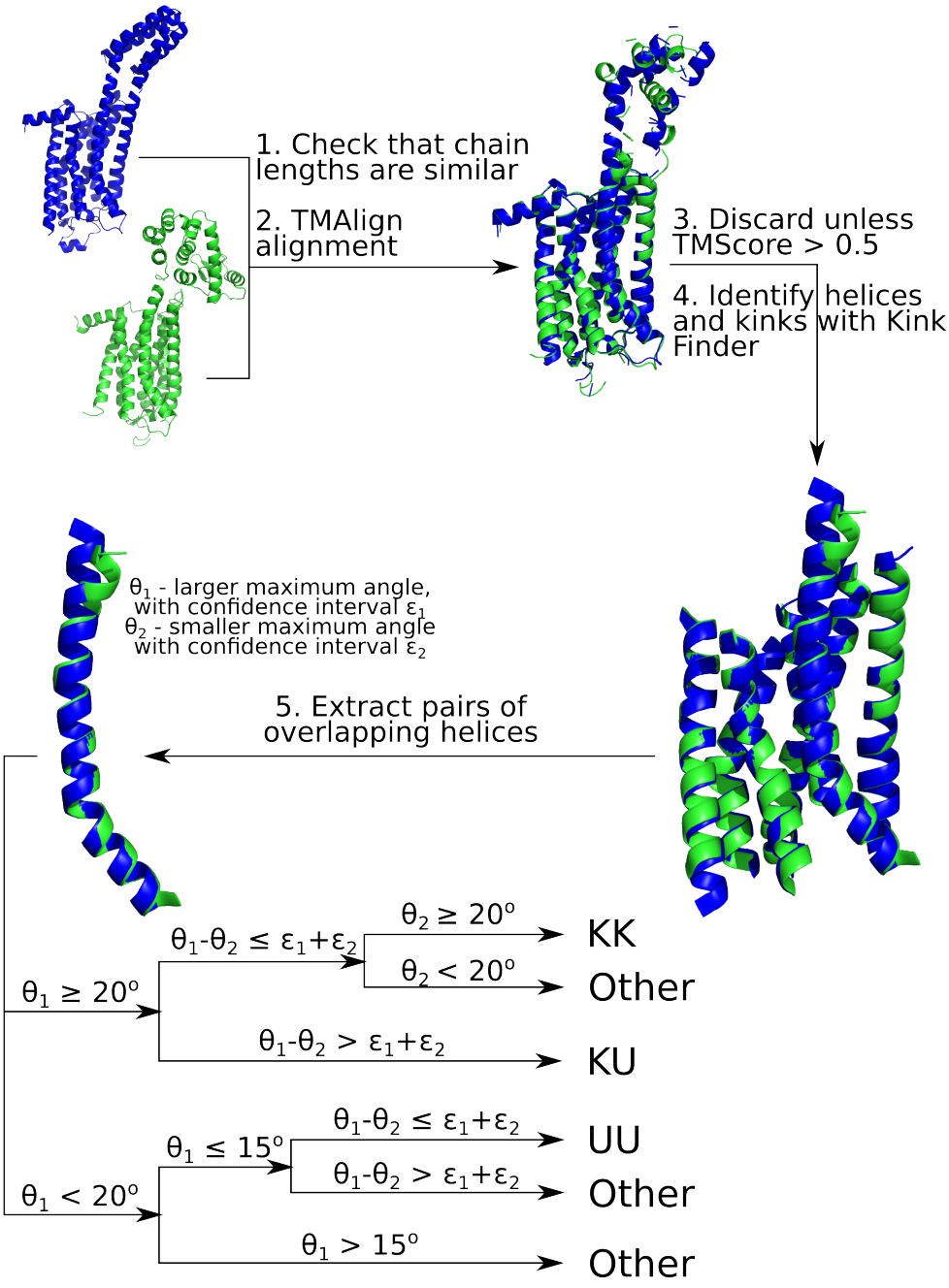


Figure 5.4: **Identifying and classifying homologous helix pairs.** Pairs of protein chains are aligned with TM-Align, providing they have similar lengths - i.e. the longer chain is no more than 50% longer than the shorter chain. Those pairs with TM-scores greater than 0.5 are retained, and aligned helix pairs are identified using Kink Finder. These aligned helix pairs are then classified based on their maximum angles, and the errors associated with those errors, into one of four categories - KK, KU, UU, Other. θ_1 is larger than θ_2.

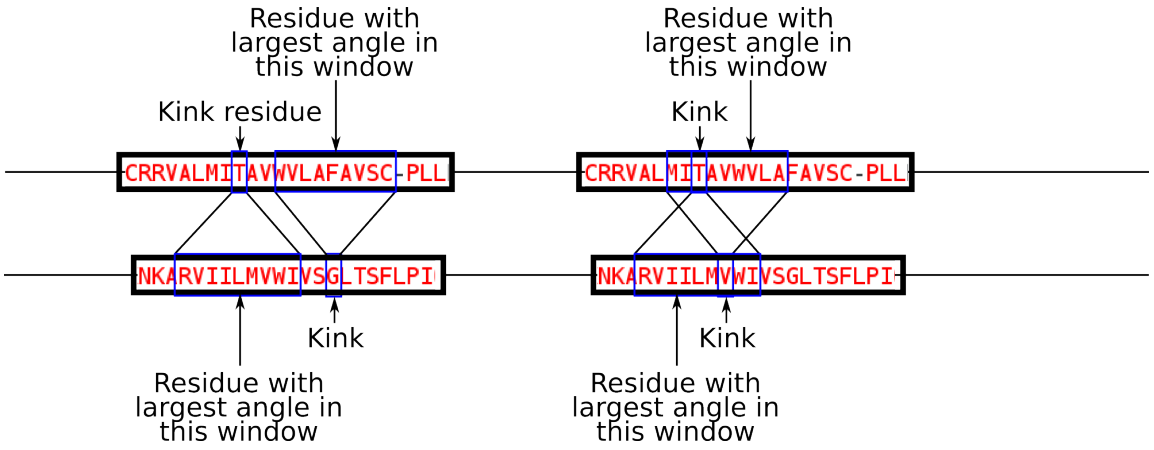


Figure 5.5: **Comparing helices.** The largest angle in each helix (the kink) is compared to a 9 residue window of angles in the other helix. Where the kinks are within 4 residues of each other (right), the two comparisons are the same. If they are not within 4 residues of one another (left), two different comparisons result.

Helix and chain sequence similarities were calculated using the TM-Align structural alignments and environment specific substitution tables, taken from Hill *et al.* (2011). Environment annotation was provided by the JOY and iMembrane programs. The use of environment specific tables leads to asymmetric similarity scores - they differ where the environment annotations of the two aligned residues differ. The scores for the helix with the larger angle in each pair is used. Gaps were ignored for scoring purposes.

5.2.4 Comparison with AHAH

The angle errors in the AHAH data set were compared to the user responses. Although for the same values of $r_n + r_c$, different sized kinks have the same error, kinks with bigger angles tend to have larger $r_n + r_c$ values, and thus larger errors. Helices with larger kinks are also more likely to be annotated as kinked by participants. Therefore, to compare different helices fairly, the angle errors were linearly scaled to remove this correlation, as below:

$$\epsilon_{norm} = \epsilon - 0.048 * \theta \tag{5.4}$$

Where ϵ is the error in the angle (see Equation 5.2), θ is the measured angle of the kink, and ϵ_{norm} is the normalised error. ϵ_{norm} was then linearly scaled such that the minimum value was 0, and the maximum value was 1. This relationship was calculated using only helices with angles between 10° and 60°.

5.2.5 Families

In order to identify homologous families of helices rather than just pairs, a network was constructed using the helix alignments. In this network, each node represented a helix, and two nodes (helices) were joined by an unweighted edge where the two helices were a part of an aligned helix pair. Helix families were extracted using a community detection algorithm, with resolution parameter 25 (Blondel *et al.*, 2008; Traag *et al.*, 2011). Such algorithms identify groups of nodes that are more densely connected to each other than they are to the rest of the network. Within each extracted community, the least connected helices were removed iteratively if they were connected to fewer than 90% of the other members of the group. This approach and these thresholds were trained on the membrane helices to give good agreement with visual inspection, and the same thresholds were applied to the soluble helices.

Once helix families had been constructed, the protein chains that they were a part of were aligned using MAMMOTH-mult (Lupyan *et al.*, 2005), a multiple structural alignment tool. This protein alignment was challenging at times. The two structural alignment tools used here (TM-Align and MAMMOTH-mult) do not give the same alignment. This is in part because TM-Align is a pairwise alignment tool, while MAMMOTH is a multiple structural aligner, but also because any two alignment methods will have differences between an alignment of the same pair of proteins. Often the alignments differ by a few residues, but occasionally the differences are much larger.

In some cases MAMMOTH-mult and TM-Align gave very different alignments for a pair of proteins. MAMMOTH-mult occasionally aligned a helix in one protein chain to a different helix in the other protein than TM-Align did. Consequently, agreement in the both the MAMMOTH-

mult and TM-Align alignments was required for two helices to be a helix pair in the family analysis. The same overlap requirement was used as with the TM-Align alignment, i.e. where the helix ends were not within four residues of one another, they were not an aligned helix pair, and the corresponding edge was removed from the network.

Comparison of helices was difficult where their secondary structure had been annotated differently - i.e. where a position in the alignment had both helical and non-helical annotation. This results in positions in the alignment where some helices have an angle while others do not, which is particularly problematic when in one helix a position has an angle sufficient for it to be kinked, while in another helix in the family, no angle can be calculated. We consequently harmonised the helix annotation for the whole alignment, and smoothed the angles, by taking the largest angle in a three residue window around each residue. This has the possibility of causing problems for Kink Finder where family members had non-helical sections, although not in the examples shown in this chapter. The error estimation method in Kink Finder could provide a method to identify and remove these sections.

5.3 Results

5.3.1 Angle error

Simulations of the 18 kinks demonstrated the relationship between the standard deviation of the error (σ_ϵ), the true angle, and $r_n + r_c$ (Figure 5.3). For angles over 10°, the true angle had little effect on the size of the error. However, the error did increase as $r_n + r_c$ increased. Combining the error distribution for the 12 ideal kinks with angles larger than 10°, a statistical confidence interval (CI) was derived for the error for different values of $r_n + r_c$ from the simulated data.

Figure 5.6 shows the variation of the 95% CI with $r_n + r_c$. The error distribution is symmetric, and is assumed to be so in the rest of this work. I use a log fit to quantify the relationship

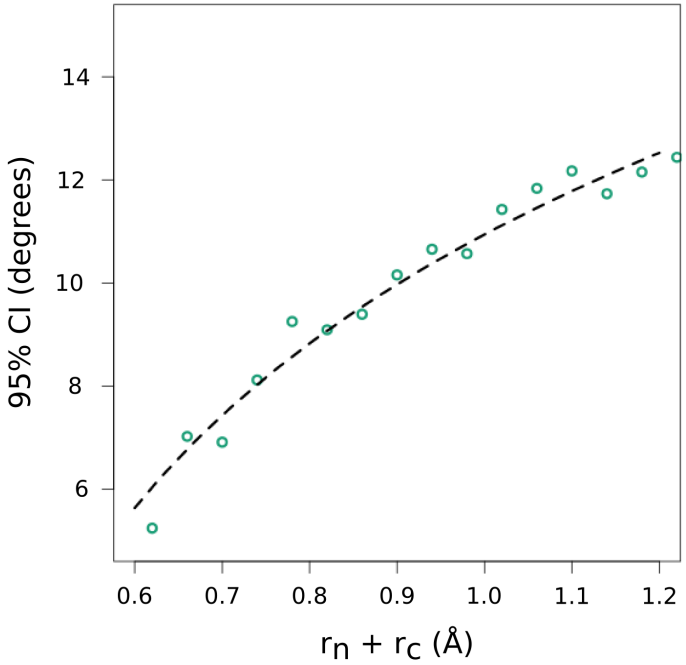


Figure 5.6: **The 95% confidence interval of angle error for a range of values of $r_n + r_c$.** For the combined data from all 12 kinks with angles $\geq 12°$, the angle errors are binned by their $r_n + r_c$ values. The modulus of the angle errors is taken, and the value at the 95^{th} percentile is taken for each $r_n + r_c$ bin (green points). A log plot is fitted to the values between 0.6 and 1.0 (dashed black line).

between $r_n + r_c$ and the size of the 95% CI interval, which gives:

$$CI = \pm(6.349 \times \log(r_n + r_c - 0.2937) + 13.15) \tag{5.5}$$

This method provides an estimate of the uncertainty in the kink angles measured by Kink Finder. It provides a simple way to discriminate between high and low confidence instances of kinks, and to compare and contrast the angles in two helices, from either homologous helices, or different structures of the same helix (such as from experiments or simulations, e.g. NMR or Molecular Dynamics).

Using the derived relationship between goodness of fit and angle confidence, Kink Finder can calculate the confidence interval for every angle it measures. The confidence intervals for the largest angle in each of the membrane helices in the Chapter 4 data set are shown in Figure 5.7. There is a weak correlation between the CI and the angle - larger angles have larger confidence intervals ($r^2 = 0.3$). Although most confidence intervals (errors) in the membrane set are between five and eight degrees, there is one greater than 12 degrees, and 267 larger than eight degrees. In the soluble set, there are four helices with error greater than 12 degrees, and 2097 with error larger than eight degrees.

The histograms suggest that the error is generally smaller for the soluble helices than for the membrane helices, but this is due to the greater proportion of helices with small angles in the soluble data set (Figures 5.7a and b).

5.3.2 Agreement with AHAH

The AHAH results were compared to the error estimation of Kink Finder. In this comparison I found that helices that were more difficult to classify were likely to also have less certain angles. As shown in the previous section, the error is correlated with the angle. Kink Finder is less able to assign angles to more kinked helices compared to less kinked helices.

Figure 5.8 shows the variation of error with the classifications provided by AHAH users.

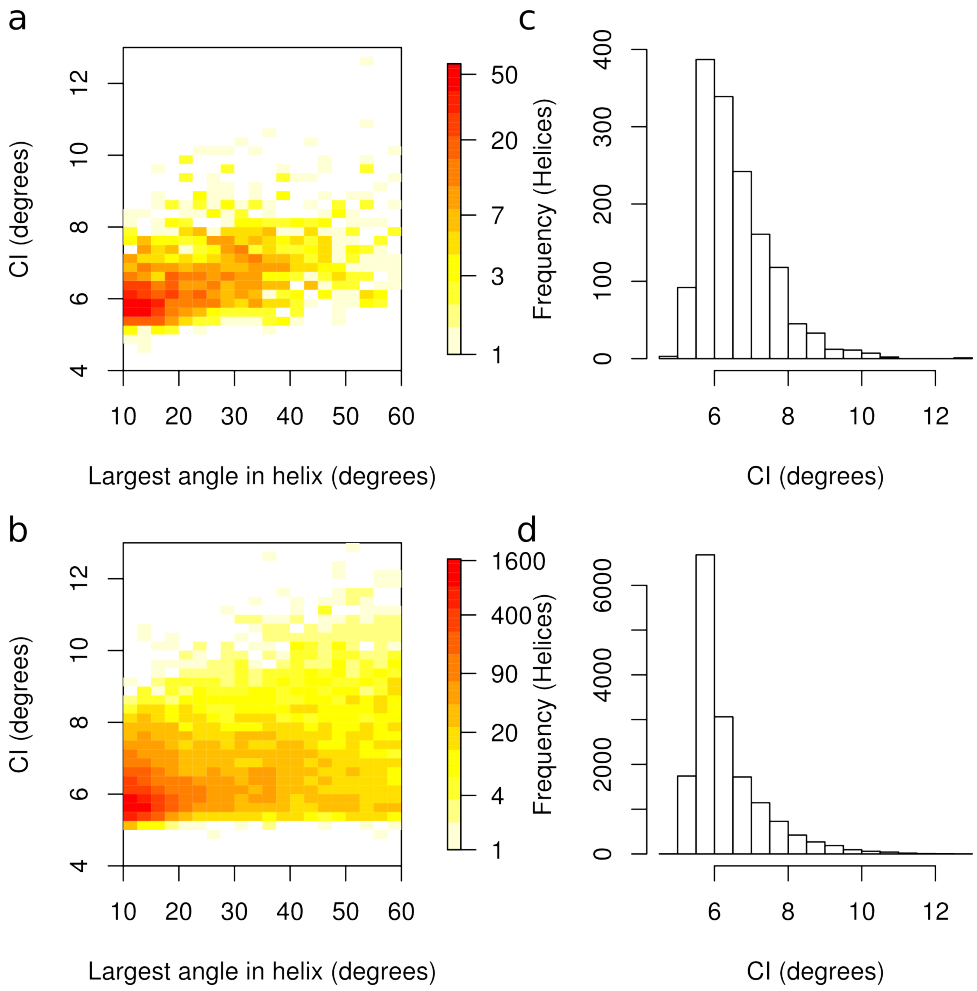


Figure 5.7: **The confidence intervals (CIs) for the maximum kink angles in the membrane (a and c) and soluble (b and d) helices**. Helices with maximum angle $\leq 10°$ are not included. (a) and (b) Heat map showing the variation of CI with angle. Coloured using a log scale. (c) and (d) Histogram of the CIs.

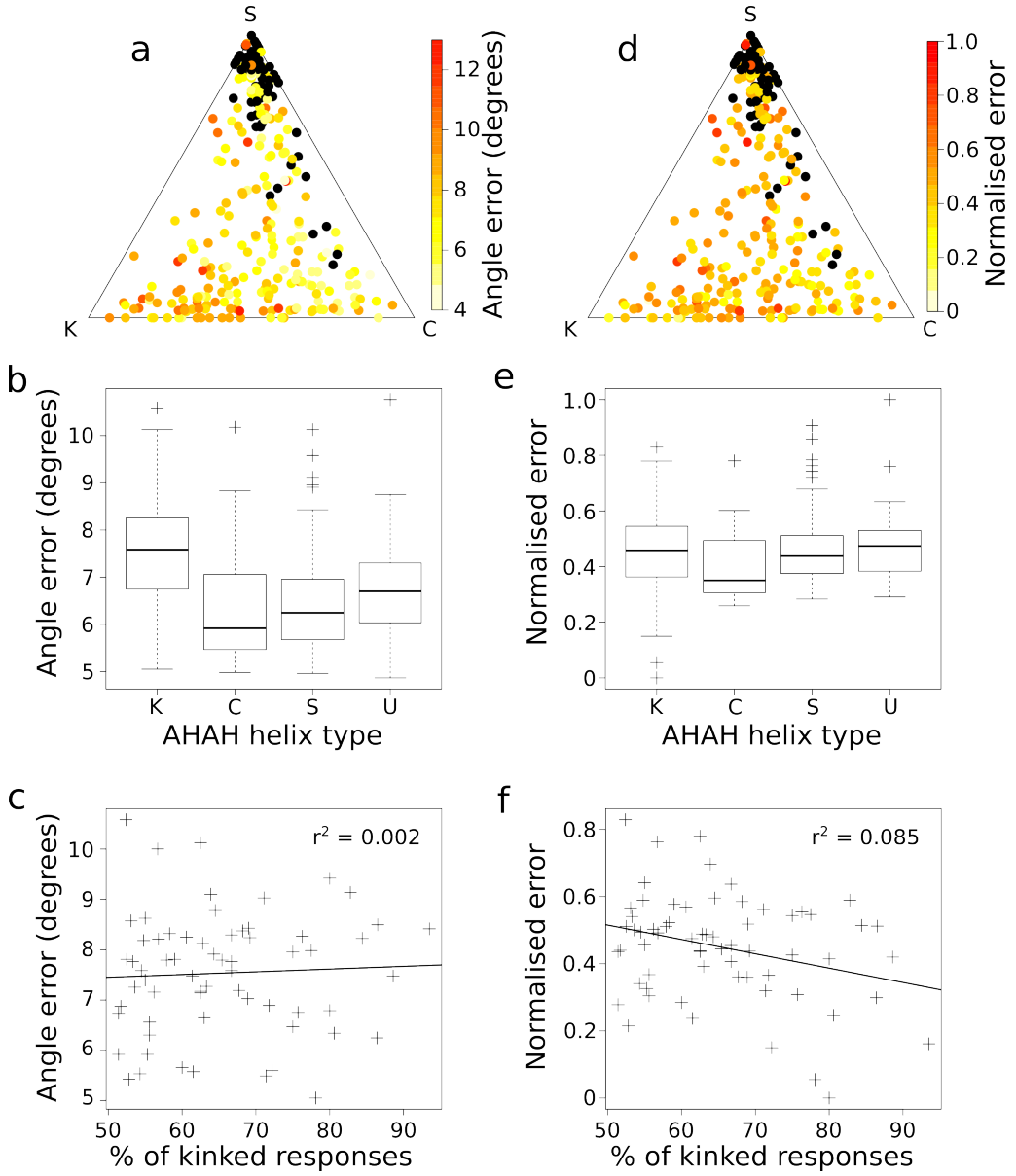


Figure 5.8: **Kink angle error correlation with AHAH results**. (a)-(c) use errors, and (d)-(f) use normalised errors. (a) and (d) show the ternary plot of responses, with points coloured by the error (black points are helices with angles under 10°). (b) and (e) show the distribution of errors for the different classifications of the helices. (c) and (f) show the correlation in the kinked groups between the error and the proportion of users that annotated the helix as kinked. Helices are grouped by their AHAH classification. K - kinked, C - curved, S - straight, U - unassigned.

Figures 5.8(d)-(f) show the normalised errors. The curved and straight helix groups have lower error than the unassigned groups. The kinked group has much higher mean error than even the unassigned group (Figure 5.8b). However, the unassigned group has slightly smaller mean normalised error than the kinked group. Considering the kinked helices alone, Figure 5.8f shows that helices which a higher proportion of participants annotate as kinked typically have less error (when normalised for the effect of angle size). However, the correlation coefficient is small ($r^2 = 0.085$).

5.3.3 Helix pair grouping

Pairs of homologous helices were extracted as described in the Section 5.2.3 and Figure 5.4. The initial membrane set (a total of 268 chains, containing 1,211 helices with 13 or more residues) has a total of 1,864 aligned helix pairs. The soluble set (9,742 chains, with 30,184 helices) has a total of 226,072 aligned helix pairs.

These helix pairs are divided into four groups, based on the sizes and uncertainty of the kinks in the two helices (Figure 5.4):

1. Kinked-kinked (KK) - both helices have angles $\geq 20°$, and overlapping confidence intervals
2. Kinked-unkinked (KU) - one helix has angle $\geq 20°$, and the other has a smaller angle, and the confidence intervals do not overlap.
3. Unkinked-unkinked (UU) - both helices have angles $\leq 15°$, and their confidence intervals overlap.
4. Other (O) - helix pairs falling into none of the above groups.

The number of helix pairs in each of these groups is shown in Table 5.1.

5.3.4 Similarity of helix pairs

Figure 5.9 shows the sequence similarity of helix pairs broken down by the type of the pair. Sequence similarity is a score of how likely the mutations required to change one sequence to

	Membrane				Soluble			
	KK	KU	UU	Other	KK	KU	UU	Other
Totals	596	200	583	509	17229	24784	123997	70315
PP	398	23	12	85	5214	562	131	5541
P-	65	98	30	88	2890	11665	570	6992
-P	91	4	16	25	2702	334	455	6025
- -	42	75	487	349	6423	12223	122841	51757
G	103	48	55	17	2140	7500	12767	9767
N	488	152	528	492	15089	17284	111230	60548

Table 5.1: **Number of aligned helix pairs, and presence of proline and gaps in them.** PP: both helices contain proline, P-: the helix with the larger kink angle contains proline, -P: the helix with the smaller kink angle contains proline, - -: neither helix contains proline. As with my previous analyses, prolines in the four N-terminal residues of the helices are not counted. G indicates that there is a gap in the structural alignment of the two helices, while N indicates that there is no gap. KK, KU, UU, and Other are defined in Section 5.3.3.

another are. Higher scores indicate the mutations are more likely, and hence that the sequences are more closely related. Similarity correlates well with sequence identity. A similarity of 6 equates roughly to an identity of 0.9, a similarity of 3 to an identity of 0.4, and a negative similarity indicates a sequence identity of less than 0.2.

The sequence similarity is lower for the KU pairs than for the KK and UU pairs, indicating that less similar proteins are more likely to have differently kinked helices. Similar behaviour is seen for the sequence identity of the helices and the chains. Furthermore, the helix and the chain similarity correlate well. The distributions of sequence similarity for the helices in each of the KK, KU, and UU sets are very similar to each other (Figure 5.9c and f). There is no obvious general sequence effect in helices that is distinct from the whole chain that results in differently kinked helices. This suggests that the cause of changes to kinks is either more global than the helices sequence (e.g. the packing of the whole protein chain), or a specific residue effect.

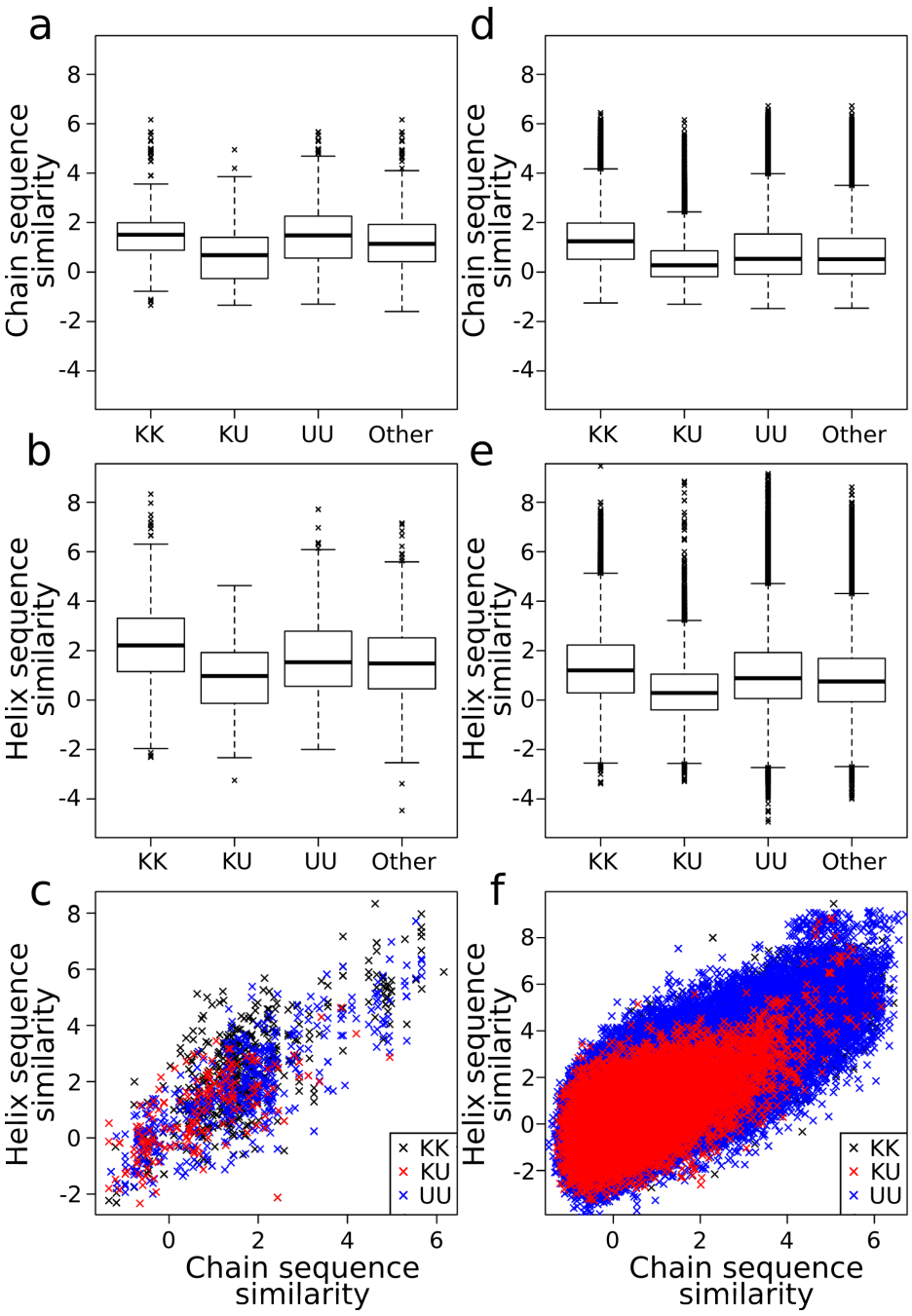


Figure 5.9: **Sequence similarity of helix pairs.** (a) and (d) Chain sequence similarity (b) and (e) Helix sequence similarity (c) and (f) Chain sequence similarity plotted against sequence similarity. (a)-(c) are the membrane helices, (d)-(f) the soluble protein helix pairs.

5.3.5 Proline and alignment gaps

Proline is the best sequence indicator of kinks - 85% of proline containing membrane helices are kinked, and 96% of proline containing soluble helices are kinked (see Chapter 4), if prolines in the first 4 residues of the helix are not counted. Another feature that has been suggested by a recent study to indicate the difference in the kinkedness of homologous helices is gaps in the alignment (Chen *et al.*, 2014). Table 5.1 shows the presence of proline and gaps in the aligned helix pairs, broken down by pair type.

There are many more examples of KU pairs that are P- than there are KU pairs that are -P, in both types of protein. Nearly half of the membrane KU pairs had a proline in the kinked helix, but no proline in the non-kinked helix (98 of 200). However, there are many helix pairs which have a proline in the helix with the larger angle (P- pairs) that are *both* kinked (65 KK, 98 KU), indicating the 'loss' of a proline only correlates with the loss of a kink in around 5 in 8 cases.

Gaps in the alignment of the helices are relatively infrequent (in only *c.* 12% and 14% of membrane and soluble helix pairs respectively). Their behaviour is different in membrane and soluble helices - whereas in membrane helices, a gap in an alignment of a kinked helix indicates a change in the kink in *c.* $^1/_3$ of pairs (48 KU, 103 KK), in soluble helices it indicates a change in the kink in *c.* $^3/_4$ of pairs (7500 KU, 2140 KK).

5.3.6 Helix families

This section describes a preliminary investigation of the helix families. Eleanor Law has continued the study, and we plan to publish this.

Figure 5.10 shows a network representation of the helix pairs within the membrane protein set. After the helix communities in the network were pruned to give families where ≥ 0.9 of the possible pairings were present, there were 28 families containing at least 5 helices, 13 with 8 or more helices, six with 10 or more helices, and the largest family had 15 helices. In the soluble helix set, after clustering, there were 856 families with at least 5 helices, 448 with 8 or more,

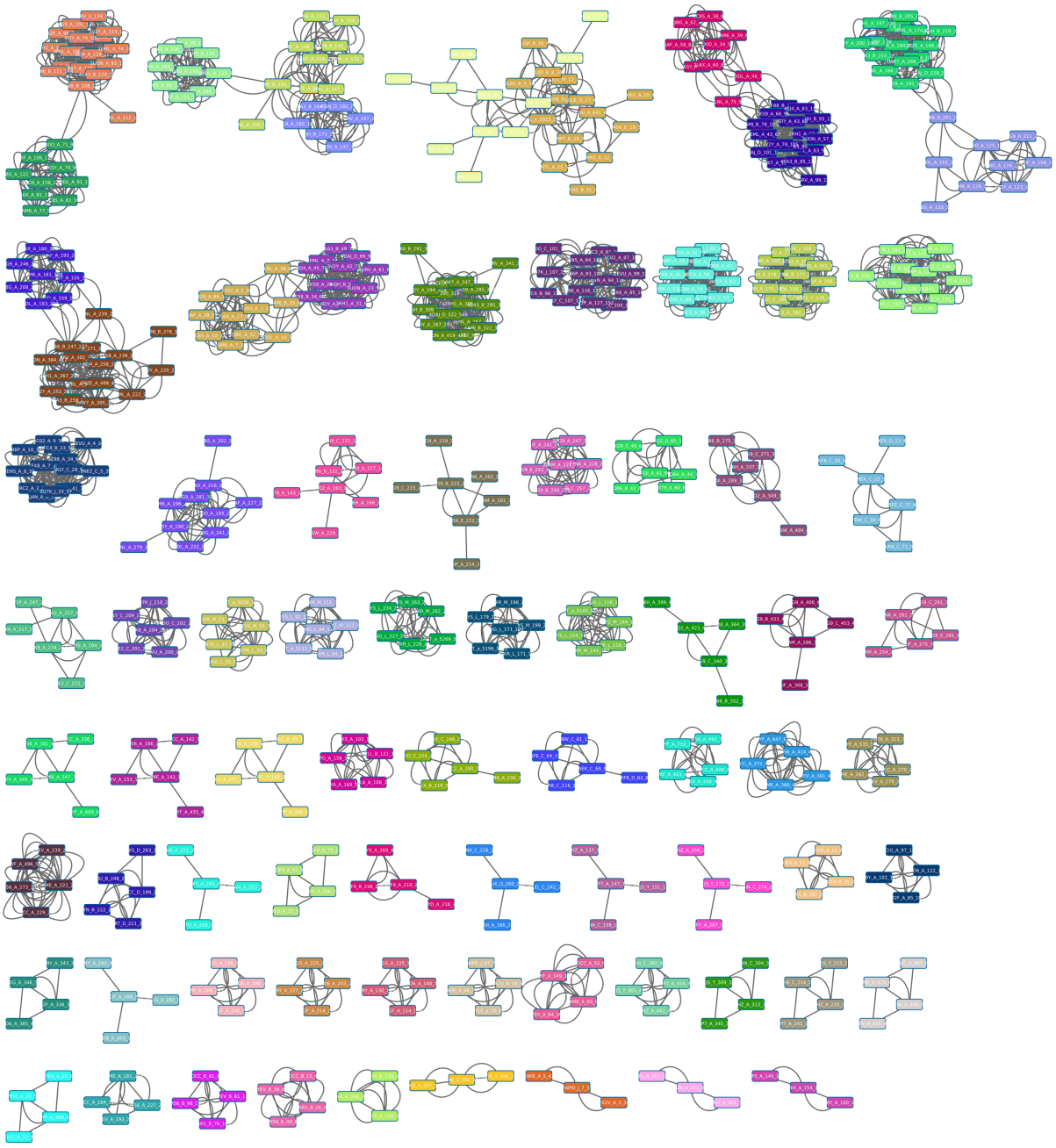


Figure 5.10: **Network of aligned helix pairs in the membrane data set.** Each node represents a helix, and each edge indicates that the two joined nodes (helices) are part of a helix pair. Helices in a community are coloured the same. Helix families were obtained from each community by iteratively removing the least connected helix in the community until every helix was connected to at least 90% of the other helices in the family.

306 with 10 or more, 189 with 15 or more, and the largest had 115 members.

Figures 5.11 and 5.12 show two example families, which correspond to the TM3 and TM6 helices in GPCRs respectively. TM3 (Figure 5.11) shows a non-conserved kink family, where the maximum kink angle varies from 9.9° to 43.6°. The sequence alignment shows some highly conserved residues - cysteine (position 3), serine (17), and aspartic acid, arginine, and tryptophan (27,28,29) (Figure 5.11d). The two helices with the largest angles come from the human A(2A) adenosine receptor and a reengineered version with an identical transmembrane domain (hence 100% sequence identity between the helices). The latter half of the 3EML helix sequence is similar to the others, however, it differs from the others in the first half. For example, 3EML has no aspartic acid (D) at position 10, and no charged residue at position 4 (where most of the other sequences have a charged or polar residue (K, R, D, E, or N). Interestingly, only one of the sequences contains a proline (2KS9, residue 10), and although that is close to the kink site (residues 6-8), the helix is not kinked.

TM6 (Figure 5.12) shows a helix family with a well conserved proline kink. The kink angles have a range of just 6.4° at the kink point (from 34° to 40.4°). The sequence alignment and logo show a highly conserved proline, at position 19 in the helix. There are also well conserved phenylalanine and tryptophan residues at positions 13 and 17. These aromatic residues are on the inside of the kink, which I observed frequently in membrane protein kinks in Chapter 4.

5.4 Discussion

5.4.1 Kink measurement error

This chapter described the first method for the characterisation of the error in α-helix kink angles. The error in the angle allows us to make comparisons between similar helices, and properly evaluate similarities or differences in their kinkedness.

The nature of kinks, as distorted parts of helices, makes it difficult to build a good statistical model to characterise error in the angles, so here we have used a heuristic approach. We have

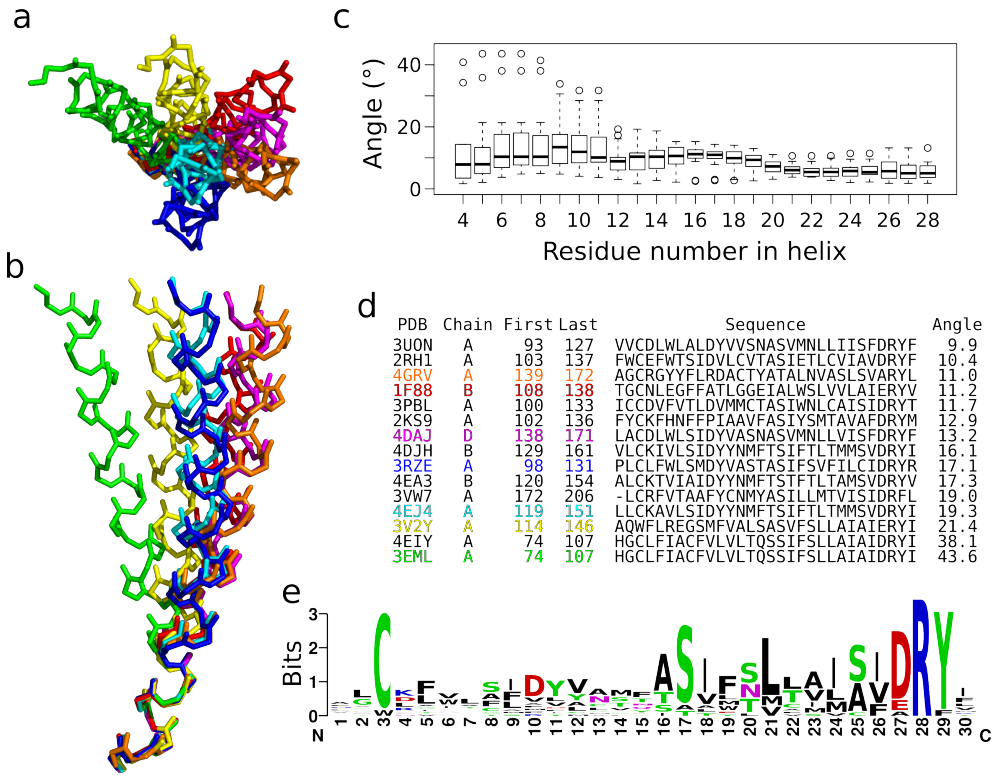


Figure 5.11: **A family of helices corresponding to the 3rd transmembrane helix in GPCRs.** (a) and (b) Top and side views of helices aligned by the six residues N-terminal to the kink. (c) Boxplots showing the range of kink angles at each position in the helix. Angles were smoothed over 3 residue windows. (d) Sequence alignment from Mammoth, ordered by maximum kink angle in helix. Helix names are coloured to match their representations in (a) and (b). Helices with black names are not shown in (a) and (b). (e) Sequence logo from sequence alignment.

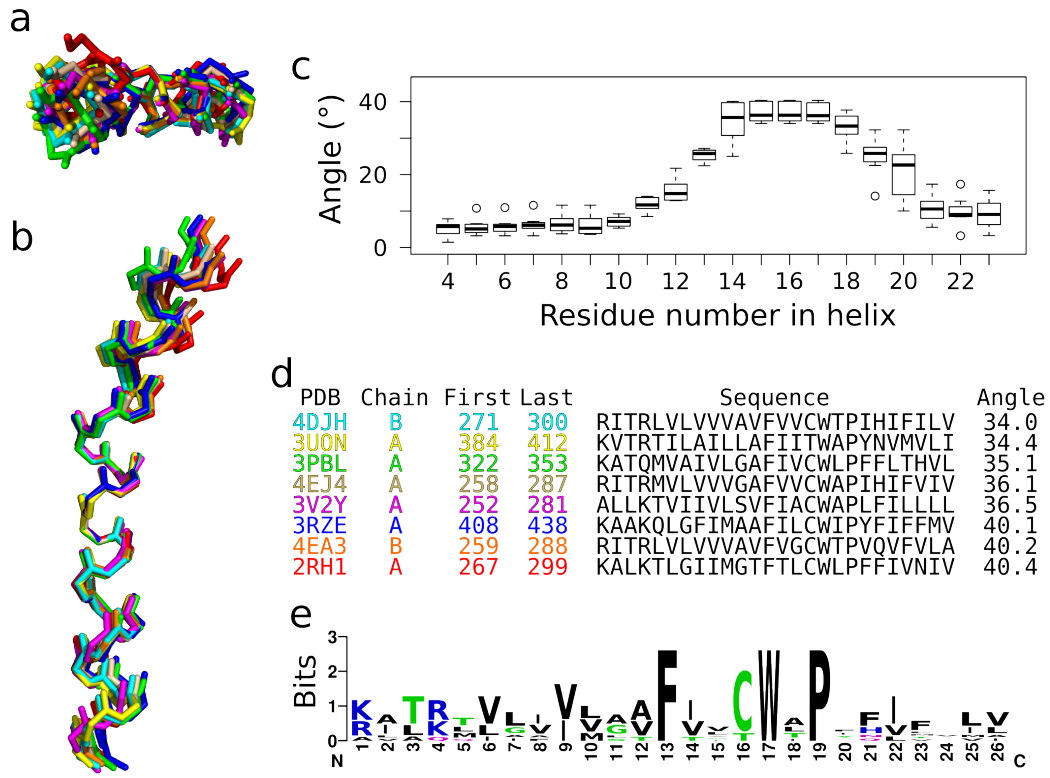


Figure 5.12: A family of helices corresponding to the 6^{th} transmembrane helix (TM6) in GPCRs. (a) and (b) Top and side views of helices aligned by the six residues N-terminal to the kink. (c) Boxplots showing the range of kink angles at each position in the helix. Angles are smoothed over a three residue window. (d) Sequence alignment from Mammoth, ordered by maximum kink angle in helix. Helix names are coloured to match their representations in (a) and (b). Helices with black names are not shown in (a) and (b). (e) Sequence logo from sequence alignment.

chosen one way to manipulate fits to measure the relationship between angle error (ϵ) and $r_n + r_c$. There are other potential methods, such as randomly moving the atoms in the helix by a small amount, or by substituting different sections of helices into 'ideal kinks', and observing the relationship between (ϵ) and $r_n + r_c$. It is very difficult to calculate the axis of regions of helix that are distorted, which means that although we can measure an angle, and calculate r, we cannot calculate the true angle, and so we cannot calculate ϵ. This method provides an estimate of the error, and importantly lets us properly evaluate the similarities and differences between kinks.

For the majority of helices, the maximum angle is estimated within five to eight degrees 95% of the time. The size of the error is correlated with the angle, with larger angles typically having worse axis fits, and thus larger errors. This degree of error is reasonable. It means that we can be reasonably confident that two kinks measured with a difference of 15° (or in the best case 10°), are significantly different. 10° is not a particularly large difference - indeed it is difficult to discern a difference visually between two kinks that differ in angle less than that.

The amount of error does not strongly correlate with the AHAH user responses. After normalising to take into account the effect of kink size on angle error, there is a weak effect - a higher proportion of participants annotate a helix as kinked when the error in that helix is lower. However, even when using the normalised errors, there is not a clear discrepancy between the errors in the kinked and unassigned helices.

5.4.2 Helix pairs

The sequence similarity of the helix pairs (Figure 5.9) suggests that the difference in kinks in homologous helices is not a helix-level difference. Although more differently kinked helices have lower helix sequence similarity, they also have lower chain sequence similarity. This indicates that the driving force for kink change is either in the whole chain, or in specific interactions that are not captured by the relatively coarse sequence similarity measure.

Proline is one such specific interaction. It is a reasonable, but by no means perfect, indicator

of a change in kinkedness of two homologous helices. It is a better indicator in soluble proteins - in membrane helix pairs that include at least one kinked helix, it is associated with the loss of a kink in 98 out of 254 cases, while in soluble helix pairs, kink loss occurs in 11665 of 13257 cases. This is similar to the results I saw in previous chapters, where proline acts as a moderately good indicator of a kink, and more so in soluble proteins than membrane proteins. The stronger influence that proline has in soluble proteins compared to membrane proteins is likely due to the membrane environment. The different hydrophobic and hydrophilic environments around different parts of the helices make it much more energetically costly to change the helix conformation. Consequently, it is more likely for proline to be incorporated without obvious deformation in a membrane helix than in a soluble helix. Gaps in the alignments also correlate with changes in kinks, but they are far less prevalent than proline residues.

5.4.3 Helix families

Our analysis identified 28 families of homologous membrane protein helices, and 856 families of homologous soluble protein helices. The two examples in Figures 5.11 and 5.12 show a conserved and a non-conserved kink. The TM6 helices contains a highly conserved FxxxWxP motif around the highly conserved kink. This correlates well with my findings in Chapter 4, where proline was an obvious indicator, and aromatic residues were frequently observed on the inside of the kink.

The TM3 helices show lower sequence similarity, and there is a non-conserved kink, which is clearly present in two helices. Several more of the helices show a degree of kinking close to the 20° region. The region around the kink (in the first half of the helix) has a higher sequence diversity than in the TM6 helices. The two kinked helices lack the charged residues that are common at positions 4 and 10 in the other helices.

These two examples suggest that there are differences between conserved and non-conserved kink families, but that a wider study is required to identify them. It is not clear if the differences in the kink conservation is a result of flexibility, where all helices can have all conformations,

or not.

Our approach to identifying families, by identifying homologous helices, and then detecting clusters within that network has yielded families corresponding to most of the seven transmembrane helices in GPCRs. This gives us confidence that it can identify other helix families, removing the need for manual curation that may bias the study. It also allows us to analyse the much larger set of soluble protein helices.

5.5 Conclusion

My error estimation module in Kink Finder allows the error in a kink angle to be estimated for the first time. Angles above 10° generally have an error in the region of ±6°, although this increases as the measured angle increases. I used this method to identify homologous aligned helix pairs within two data sets, one of membrane proteins, and one of soluble proteins.

I grouped these helix pairs into conserved kink (KK), unconserved kink (KU), unkinked (UU), and others (O). These four groups showed little difference in the sequence similarity, both at the chain and helix level, suggesting that changes in kinks are caused by specific residue effects, as opposed to helix or chain level sequence changes.

We then identified 28 membrane and 856 soluble homologous helix families by detecting communities within the helix pairs. I have shown examples of conserved and non-conserved families, which suggest that there may be sequence or residue signals that could be identified in a larger study. Further study is required to identify if and where kink non-conservation arises from flexibility, and where it arises from a change in the allowed conformations.

CHAPTER 6

Conclusions and future directions

In this thesis I have investigated methods to identify and define helix kinks, developed a method, Kink Finder, and used it to characterise kinks in soluble and membrane protein helix structures.

6.1 Kink Finder and B statistic

My Kink Finder method includes novel steps that consistently place kinks, and estimate angle error. The consistent positioning, based on the helix shape around the kink, revealed the hydrophobic/hydrophilic pattern around soluble helix kinks, and showed how kinks typically point into the solvent. The error estimation allowed the quantitative comparison of homolo-

gous helices for the first time, as described in Chapter 5. Like other methods, Kink Finder contains several heuristic parameters and thresholds. However, changing these parameters had no significant effects on the tests carried out here.

The B statistic method, that I developed in conjunction with Professor Kanti Mardia, provides a simple, statistically robust method devoid of arbitrary parameters. This, unlike other methods, produces a bimodal statistic, meaning that we could use the data itself to identify a threshold between kinked and unkinked helices. However, it is not able to identify the position of kinks. The method could be extended to include kink positioning, but this would necessitate the use of heuristic parameters.

6.2 Other kink finding methods

Previous studies of kinks have differed largely on the conclusions they drew about kinks, as well as the methods that they employ to detect them. In Chapter 3, I showed that the methods these studies use to identify kinks differ in both the kinked helices that they identify, and where they identify the kinks in these helices. Using my method to standardize the position of the kinks that these methods identify, leads to them giving far more consistent results, both in Chapter 3, and in the analysis of a much larger data set in Chapter 4. Previous methods did not consider the position of the kink relative to the helix, however they are generally biased towards identifying kink residues on the inside of the helix. This resulted in weak residue patterns, that became more obvious after Kink Finder's consistent kink positioning.

6.3 Crowdsourcing

Our crowdsourcing approach, AHAH, provided an alternative approach to kink finding. The results show that the classification of helices is often ambiguous, in terms of kinked, curved, or straight, particularly when differentiating between kinked and curved. The agreement with existing methods was moderate, with a particular discrepancy about the proportion of helices

that were identified as kinked. This indicates that many of the kinks identified by existing methods, including Kink Finder, are more subtle than the 'clear location where the direction of the helix changes' definition that we used in our study. Our study allowed us to generate a gold standard data set, which can be used for future kink identification methods training and testing.

All of the above results indicate that kink identification is difficult. Most methods rely on the assumption that the helix around the kink is relatively helical. As kinks are helix distortions, this is not necessarily true. The description 'trivial but difficult' (Richardson & Richardson, 1989) seems to apply equally well to the identification of kinks as it does to secondary structure assignment.

6.4 Kink characterisation

Throughout this thesis I have used a single summary statistic for kinks - the angle. This is a simplification, ignoring the diversity of distortions within kinks, such as tight and wide turns and missing hydrogen bonds.

However, the number of kinks (*c.* 400 in the membrane helix set) means that any subdivision of them is unlikely to provide robust conclusions. Even under my approach of considering all the kinks together, the only sequence patterns that can be observed are the use of proline and the hydrophobicity patterns for soluble kinks. Methods such as clustering the kink sequences revealed no clear patterns other than these. Grouping the amino acids by way of reduced alphabets (i.e. grouping the amino acids together), did not increase the sequence signals. Particularly with the membrane kinks, the number of charged amino acids was too small for meaningful conclusions to be drawn about them, regardless of the amino acid grouping.

It is tempting to think that breaking down kinks into more groups may reveal more patterns, but given a small enough set, random fluctuations are liable to over-interpretation. The way in which to sub-divide the kinks is also unclear.

6.5 Kink characteristics in soluble and membrane proteins

Using Kink Finder to analyse sets of helices from membrane and soluble proteins revealed that, allowing for differences in helix length, the kink properties of the membrane and soluble protein helices are far more similar than previously reported. Kinks are a feature of long (i.e. > 20 residues) protein helices, and the apparent difference in kink frequency is due to the enrichment of long helices in the set of membrane proteins. As with previous studies, proline is a strong kink indicator in both soluble and membrane helices, particularly when prolines in the first turn of the helix are ignored. Other sequence patterns are far less obvious, although patterns involving glycine, serine, and the aromatic amino acids were present in kinks identified by Kink Finder and kinks identified by MC-Helan. The soluble kinks show a strong hydrophobicity pattern, which is due to the location of kinks with respect to the solvent - they typically point out into the solvent.

This means that the residue patterns are connected to the overall shape of the protein i.e. kinks point out into the solvent, so that residues two before and after the kink residue are typically exposed to the solvent, and the kink residue is buried. Future studies should take this into account, possibly by normalising amino acid frequency based on the solvent accessibility. In membrane proteins this is likely to be even more complex, as residues on the surface of the protein can be exposed to hydrophilic or hydrophobic environments depending on their position relative to the membrane.

Residues patterns around kinks could be used to build a sequence based predictor of kinks. Prediction based on helix length and the presence of proline would likely perform comparatively well. Langelaan *et al.* (2010) found that the best predictor just used the four amino acids C-terminal to the kink position, i.e. where proline is most often found. It is not clear if any other sequence signal is strong enough to provide good prediction, although sequence based predictors (Kneissl *et al.*, 2011; Meruelo *et al.*, 2011; Seifert *et al.*, 2014) have been developed. Kneissl *et al.* (2011) obtained a balanced accuracy[1] of 0.82 for their best predictor using their data set.

[1]Balanced accuracy is the arithmetic mean of the sensitivity and specificity. Sensitivity is the proportion

However, predicting helices as kinked if they contained a proline in anywhere except the first four residues, and using the same data[1], yielded a balanced accuracy of 0.77[2]. Similarly, this prediction method applied to the membrane data set from Chapter 4 had a balanced accuracy of 0.80[3]. It is very likely that incorporating information about helix length into a proline-based predictor would result in performance better than that of published methods.

My work has shown that soluble kinks are similar to membrane kinks, suggesting that the soluble helices (particularly those with ≥ 20 residues), could be used to build and test a predictor on a much more diverse set of helices. This would provide a better test for a predictor, less prone to overfitting.

I found that many previously identified sequence patterns associated with kinks appear to be protein or family specific. Much work has been done on GPCRs, as well as some work on other kinked helix structures, and researchers have often identified features which they extrapolate to be true for all kinks. Amino acids, and amino acid motifs can be conserved for a variety of reasons, particularly when they occur in the interface between hydrophilic and hydrophobic layers of the membrane.

6.6 Kink conservation

Homologous helix pairs were identified in two helix sets, from soluble proteins, and membrane proteins. While there were many examples of helices with conserved kinkedness, many pairs of helices with non-conserved helices were also found. The conserved kink and non-conserved kink helix pairs differ similarly in their sequence and chain similarity. Although non-conserved kink pairs are, on average, from helices with lower sequence similarity than conserved kink pairs, they also come from less similar protein chains. This indicates that changes in kinks are not

of kinked helices that were predicted to be kinked, and specificity is the proportion of straight helices that are predicted to be straight.

[1]Kneissl *et al.* (2011) excluded the helices that they had classified as curved from their performance tests.

[2]Sensitivity 0.62 and specificity 0.91.

[3]Sensitivity 0.68, specificity 0.92. See Table 4.1 on page 131.

caused by changes to the helix as a whole, and suggests that they are either due to changes to the protein chain as a whole, or to single residue effects that are not captured by the relatively coarse measure of sequence similarity.

Work on GPCRs has indicated that at least some kinks are flexible (Bettinelli *et al.*, 2011; Katritch *et al.*, 2013; Sands *et al.*, 2006; Sansom & Weinstein, 2000; van der Kant & Vriend, 2014). However, further work is needed to identify if all kinks are flexible, or if only some of them are. Like many things in biology, I think it is likely that there is a spectrum of flexibility, i.e. there are very flexible kinks, very static kinks, and kinks everywhere in between. There are also likely to be a variety of modes of flexibility. Kink Finder may require further development to allow it to quantitatively identify kinks in flexible structures from NMR or molecular dynamics. The challenges faced in the comparison of kinks across families give a good indication of approaches for this further development of Kink Finder.

Data on the motion of proteins is rapidly increasing, both in terms of NMR structures, and molecular dynamics simulations. At the time of writing, the amount of available data is probably at, or approaching, the level required to undertake a bioinformatics investigation of kink flexibility. Once again, I am interested in studying many examples of kinks, to avoid the potential pitfall of over analysing the data. Of particular interest is the question of whether all kinks are flexible, or just some of them.

The approaches developed in Chapter 5 also allow the comparison of kinks within the same protein structure. There are many examples of protein structures where more than one protein is present in the asymmetric unit. This provides many structures of the same protein in different environments. This, along with the methods for kink comparison described in Chapter 5, would allow the investigation of how much the structure of kinks is affected by crystal packing. This question is yet unanswered, and the answer would be a important contribution to the field.

6.7 Kink prediction

Initially, my project envisioned the inclusion of kink prediction into a modelling pipeline (i.e. Memoir (Ebejer *et al.*, 2013)). There is a recent study that has included kink prediction in membrane protein modelling (Chen *et al.*, 2014), although the authors used a very simple approach to identifying where kinks may change between target and template - the loss or gain of proline at a position in the alignment, or the presence of a gap in the alignment. My results in Chapter 5 indicate that this will only be moderate accurate in terms of identifying when kinks will change.

The approach that I explored in Chapter 4, treating kinks as a single type and comparing across all families, was not fruitful. Proline is a clear and obvious feature of kinks, and although I identified a couple of weak sequence signals (aromatic motifs on the inside of kinks in particular), it is clear that many of the sequence signals claimed by other authors are artefacts of their methods. My work has shown that normalising for the number and the position (inside/outside) of kinks within helices explains some of the discrepancies between the results of previous studies. Indeed, most studies have ignored the strong correlation between helix length and kinkedness, as well as the fact that their kink finding methods generally tend to identify kink residues on the inside of the helix kink. The relationship is further complicated - helices tend to start on the inside of the protein, meaning the first residue is more often than not hydrophobic (Aurora & Rose, 1998). The amphipathic nature of helices in soluble proteins results in periodic fluctuations of amino acid propensities along the helix (Engel & DeGrado, 2004). Although this is not the case for helices within the hydrophobic core of the membrane, it would be surprising if similar effect was not present in parts of membrane helices within hydrophilic head group layer of the membrane. This results in a very complex situation - say, the fifth residue of a helix is likely to be buried, likely to be hydrophobic, and also likely to be a kink residue. Whereas, the sixth residue is likely to be exposed, hydrophilic (unless it is in the tail layer of the membrane), and not a kink residue.

One example of this difficulty is described here - do kinked and not-kinked helices have different solvent accessible surfaces (Kneissl *et al.*, 2011)? No, long and short helices have different solvent accessible surface areas, and long helices are more often kinked than short helices. This is trivially the case - most helices are amphipathic, with one buried and one solvent accessible face. They tend to start and end in contact with the rest of the protein - i.e. with a buried, hydrophobic residue - and so are $n + \frac{1}{2}$ turns long. Hence normally there are more buried residues ($\frac{n+1}{2}$ turns worth) in a helix than exposed ones ($\frac{n}{2}$ turns worth) - and the fewer turns there are, the more the ratio between them deviates from 1. Ergo, residues in shorter helices are, on average, more solvent exposed than those in long helices. Since shorter helices are less often kinked than long helices, then residues in not-kinked (typically short) helices are more solvent exposed than residues in kinked (typically long) helices.

Consequently, the propensity of an amino acid type to be in a helix depends on how exposed it is, where it is in the helix, and which layer of the membrane it is in. Much more data is required to properly allow for these relationships, so using this approach predict kinks within structures is not currently feasible. However, the finding that kinks occur in all types of protein, suggests that future prediction methods can (and should) be validated against soluble protein data, which will avoid some of the problems with previous research in this area. Alternatively, Chapter 5 provides a more realistic approach to prediction - using information from helices in similar proteins.

The work in Chapter 5 suggests that the best approach to kink prediction is to identify the characteristics of conserved and non-conserved kinks. These characteristics, along with the general characteristics of kinks identified in Chapter 4, could be used to build a sequence based predictor of kinks. Given an input sequence, this would utilise both general and family-specific kink information to provide a probability of the helix within the protein being kinked, and the likely location of the kinks. The kink predictions could then be used at various stages of the modelling process. Most simply, it could be used at the template selection stage, to select the template with the kinks that most closely match the predictions. It could also be used to score

the quality of models produced in the modelling process (as in Werner & Church (2013)). High confidence kink predictions could be used to build models of the protein core where no suitable templates are available. Conversely, kink characteristics could be used to develop methods to identify ways to engineer modifications to kinks in proteins. I think that identifying what characteristics are common to protein families with non-conserved kinks is most important to developing such a method. Obviously, adding or removing proline residues is most likely to effect a change in the kinkedness of a helix, but there are likely to be more subtle changes that could be made to effect such a change.

An extension of my work would be to explore how kink changes effect the three dimensional protein structure. This is important to structure prediction, as any method needs to know the likely changes given a change in the kink - are the helix termini fixed, or is the residues at the kink site that stay fixed in the structure, while the helix termini move? Similarly, the flexibility of the kink is important to structure prediction. In the most extreme case, where a highly flexible kink is predicted, a model building program should output more than one possible model, based on the likely helix and kink conformations.

6.8 Final words

I have investigated kinks, built three methods to identify them, and used one of these, Kink Finder, to characterise known structures of kinks, using novel positioning and error methods. I have shown that residue patterns around kinks are more complex than previously thought, being affected by the position of the kink relative to the rest of the protein. I have also demonstrated that soluble protein kinks are far more similar to those in membrane proteins than previously thought. Kinks feature in long (> 20 residues) helices regardless of the environment of the proteins.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2008). *Molecular Biology of the Cell*. Garland Science, New York, 5th edn. 1

Almén, M.S., Nordström, K.J.V., Fredriksson, R. & Schiöth, H.B. (2009). Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, **7**, 50. 21, 27

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402. 122

Åqvist, J. (1986). A simple way to calculate the axis of an α-helix. *Computers and Chemistry*, **10**, 97–99. 40, 42, 65, 66, 67, 80

Arunan, E., Desiraju, G.R., Klein, R.a., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D.C., Crabtree, R.H., Dannenberg, J.J., Hobza, P., Kjaergaard, H.G.,

Legon, A.C., Mennucci, B. & Nesbitt, D.J. (2011). Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure and Applied Chemistry*, **83**, 1637–1641. 10

Aurora, R. & Rose, G.D. (1998). Helix capping. *Protein Science*, **7**, 21–38. 35, 179

Baeza-Delgado, C., Marti-Renom, M.A. & Mingarro, I. (2013). Structure-based statistical analysis of transmembrane helices. *European Biophysics Journal*, **42**, 199–207. 35

Baker, L.A. & Baldus, M. (2014). Characterization of membrane protein function by solid-state NMR spectroscopy. *Current Opinion in Structural Biology*, **27C**, 48–55. 27

Bansal, M., Kumar, S. & Velavan, R. (2000). HELANAL: A program to characterize helix geometry in proteins. *Journal of Biomolecular Structure & Dynamics*, **17**, 811–819. vii, 4, 37, 41, 51, 85

Barlow, D.J. & Thornton, J.M. (1988). Helix geometry in proteins. *Journal of Molecular Biology*, **201**, 601–619. vi, 14, 36, 40, 45, 118

Barrett, P.J., Song, Y., Van Horn, W.D., Hustedt, E.J., Schafer, J.M., Hadzisel-imovic, A., Beel, A.J. & Sanders, C.R. (2012). The amyloid precursor protein has a flexible transmembrane domain and binds cholesterol. *Science*, **336**, 1168–1171. 3, 4, 118, 146

Bergelson, L. & Barsukov, L. (1977). Topological asymmetry of phospholipids in membranes. *Science*, **197**, 224–230. 19

Berman, H., Henrick, K., Nakamura, H. & Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, **35**, D301–D303. 24, 27, 28, 118, 119

Betancourt, M. & Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–9. 44

Bettinelli, I., Graziani, D., Marconi, C., Pedretti, A. & Vistoli, G. (2011). The approach of conformational chimeras to model the role of proline-containing helices on GPCR mobility: the fertile case of Cys-LTR1. *ChemMedChem*, **6**, 1217–1227. 4, 38, 44, 118, 178

Bill, R.M., Henderson, P.J.F., Iwata, S., Kunji, E.R.S., Michel, H., Neutze, R., Newstead, S., Poolman, B., Tate, C.G. & Vogel, H. (2011). Overcoming barriers to membrane protein structure determination. *Nature Biotechnology*, **29**, 335–340. 26

Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**, P10008. 156

Blundell, T.L., Barlow, D., Borkakoti, N. & Thornton, J.M. (1983). Solvent-induced distortions and the curvature of α-helices. *Nature*, **306**, 281–283. 29, 34, 141

Bowie, J.U. (2011). Membrane protein folding: how important are hydrogen bonds? *Current Opinion in Structural Biology*, **21**, 42–49. 23, 36, 46

Bowie, J.U. (2013). Membrane protein twists and turns. *Science*, **339**, 398–399. 118

Brändén, C.I. & Alwyn Jones, T. (1990). Between objectivity and subjectivity. *Nature*, **343**, 687–689. 25

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, **4**, 187–217. 40

Brünger, A. (1997). Free R value: Cross-validation in crystallography. *Methods in Enzymology*, **277**, 366–396. 25

Can, T. & Wang, Y. (2013). CTSS: A robust and efficient method for protein structure alignment based on local geometrical and biological features. *Proceedings of the 2003 IEEE Bioinformatics Conference.*, 169–79. 44

Cao, Z. & Bowie, J.U. (2012). Shifting hydrogen bonds may produce flexible transmembrane helices. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 8121–8126. 36, 45, 46, 118

Cartailler, J.P. & Luecke, H. (2004). Structural and functional characterization of π bulges and other short intrahelical deformations. *Structure*, **12**, 133–144. 29

Chakravarti, I., Laha, R. & Roy, J. (1967). *Handbook of Methods of Applied Statistics, Volume I*. John Wiley and Sons. 125

Chamberlain, A.K., Lee, Y., Kim, S. & Bowie, J.U. (2004). Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of Molecular Biology*, **339**, 471–479. 36

Chang, J.M., Di Tommaso, P., Taly, J.F. & Notredame, C. (2012). Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics*, **13 Suppl 4**, S1. 118

Chen, K.Y.M., Sun, J., Salvo, J.S., Baker, D. & Barth, P. (2014). High-resolution modeling of transmembrane helical protein structures from distant homologues. *PLoS Computational Biology*, **10**, e1003636. 29, 48, 146, 165, 179

Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.r.G.F., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K. & Stevens, R.C. (2007). High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science*, **318**, 1258–1265. 47

Conner, A.C., Hay, D.L., Simms, J., Howitt, S.G., Schindler, M., Smith, D.M., Wheatley, M. & Poyner, D.R. (2005). A key role for transmembrane Prolines in calcitonin receptor-like receptor agonist binding and signalling: Implications for family B G-protein-coupled receptors. *Molecular Pharmacology*, **67**, 20–31. 38

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. & Foldit Players (2010). Predicting protein structures with a multiplayer online game. *Nature*, **466**, 756–760. 86, 112

Cordes, F.S., Bright, J.N. & Sansom, M.S. (2002). Proline-induced distortions of transmembrane helices. *Journal of Molecular Biology*, **323**, 951–960. 45

Cuff, J.A. & Barton, G.J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–19. 33

Dahl, A.C.E., Chavent, M. & Sansom, M.S.P. (2012). Bendix: intuitive helix geometry analysis and abstraction. *Bioinformatics*, **28**, 2193–4. 44

Daleke, D.L. (2003). Regulation of transbilayer plasma membrane phospholipid asymmetry. *Journal of Lipid Research*, **44**, 233–242. 19

Dasgupta, B. & Chakrabarti, P. (2008). pi-Turns: types, systematics and the context of their occurrence in protein structures. *BMC Structural B iology*, **8**, 39. 29

de Almeida, D.E. & Holoshitz, J. (2011). MHC molecules in health and disease: At the cusp of a paradigm shift. *Self/Nonself*, **2**, 43–48. 4, 38, 46

Deflorian, F. & Jacobson, K.A. (2011). Comparison of three GPCR structural templates for modeling of the P2Y12 nucleotide receptor. *Journal of Computer-Aided Molecular Design*, **25**, 329–338. 4

DeGrado, W.F., Gratkowski, H. & Lear, J.D. (2003). How do helix-helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. *Protein Science*, **12**, 647–665. 36, 46

Del Val, C., White, S.H. & Bondar, A.N. (2012). Ser/Thr motifs in transmembrane proteins: conservation patterns and effects on local protein structure and dynamics. *Journal of Membrane Biology*, **245**, 717–730. 45, 138

Deupi, X., Olivella, M., Govaerts, C., Ballesteros, J.A., Campillo, M. & Pardo, L. (2004). Ser and Thr residues modulate the conformation of pro-kinked transmembrane alpha-helices. *Biophysical Journal*, **86**, 105–115. 45, 46, 136, 137

Devillé, J., Rey, J. & Chabbert, M. (2008). Comprehensive analysis of the helix-X-helix motif in soluble proteins. *Proteins: Structure, Function, and Bioinformatics*, **72**, 115–135. 4, 38, 45, 46, 114, 118, 136, 137, 138, 139

Devillé, J., Rey, J. & Chabbert, M. (2009). An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. *Journal of Molecular Evolution*, **68**, 475–489. 44

Drews, J. (2000). Drug Discovery: A Historical Perspective. *Science*, **287**, 1960–1964. 21

Duclohier, H., Molle, G., Dugast, J.Y. & Spach, G. (1992). Prolines are not essential residues in the "barrel-stave" model for ion channels induced by alamethicin analogues. *Biophysical Journal*, **63**, 868–873. 38

Dunker, A.K., Romero, P., Obradovic, Z., Garner, E.C. & Brown, C.J. (2000). Intrinsic protein disorder in complete genomes. *Genome Informatics*, **11**, 161–71. 14

Ebejer, J.P., Hill, J.R., Kelm, S., Shi, J. & Deane, C.M. (2013). Memoir: template-based structure prediction for membrane proteins. *Nucleic Acids Research*, **41**, W379–83. 118, 146, 179

Eilers, M., Shekar, S.C., Shieh, T., Smith, S.O. & Fleming, P.J. (2000). Internal packing of helical membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5796–5801. 21, 46

Eilers, M., Patel, A.B., Liu, W. & Smith, S.O. (2002). Comparison of helix interactions in membrane and soluble α-bundle proteins. *Biophysical Journal*, **82**, 631–632. 21, 23

Engel, D.E. & DeGrado, W.F. (2004). Amino acid propensities are position-dependent throughout the length of alpha-helices. *Journal of Molecular Biology*, **337**, 1195–1205. 35, 179

Enkhbayar, P., Hikichi, K., Osaki, M., Kretsinger, R.H. & Matsushima, N. (2006). 310-helices in proteins are parahelices. *Proteins*, **64**, 691–9. 11, 14

Eveleigh, A., Jennett, C., Lynn, S. & Cox, A. (2013). "I want to be a Captain! I want to be a Captain": Gamification in the Old Weather Citizen Science Project. *Gamification '13: Proceedings of the First International Conference on Gameful Design, Research, and Applications.*, 79–82. 115

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M. & Berglund, L. (2010). Prediction of the human membrane proteome. *Proteomics*, **10**, 1141–1149. 27

Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J.P. & Bowie, J.U. (2004). Side-chain contributions to membrane protein structure and stability. *Journal of Molecular Biology*, **335**, 297–305. 23

Fodje, M. & Al-Karadaghi, S. (2002). Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering, Design, and Selection*, **15**, 353–358. 11, 14

Forrest, L.R., Tieleman, D.P. & Sansom, M.S. (1999). Defining the transmembrane helix of M2 protein from influenza A by molecular dynamics simulations in a lipid bilayer. *Biophysical Journal*, **76**, 1886–1896. 38

Fowler, P.W. & Sansom, M.S.P. (2013). The pore of voltage-gated potassium ion channels is strained when closed. *Nature Communications*, **4**, 1872. 4, 38, 44, 146

Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, **23**, 566–579. 32, 52

Geetha, V. & Munson, P. (1996). Simplified representation of proteins. *Journal of Biomolecular Structure & Dynamics*, **13**, 781–93. 43

Gimpelev, M., Forrest, L.R., Murray, D. & Honig, B. (2004). Helical packing patterns in membrane and soluble proteins. *Biophysical Journal*, **87**, 4075–4086. 21

Good, B.M. & Su, A.I. (2013). Crowdsourcing for bioinformatics. *Bioinformatics*, **29**, 1925–1933. 85

Hall, S.E., Roberts, K. & Vaidehi, N. (2009). Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *Journal of Molecular Graphics & Modelling*, **27**, 944–950. vii, 4, 37, 41, 44, 45, 46, 47, 51, 79, 110, 114, 136, 137, 138, 139

Hankamer, B., Barber, J. & Boekema, E.J. (1997). Structure and membrane organization of photosystem II in green plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, **48**, 641–671. 18

Hanson, A.J. & Thakur, S. (2012). Quaternion maps of global protein structure. *Journal of Molecular Graphics & Modelling*, **38**, 256–78. 43

Hanson, M.A., Roth, C.B., Jo, E., Griffith, M.T., Scott, F.L., Reinhart, G., Desale, H., Clemons, B., Cahalan, S.M., Schuerer, S.C., Sanna, M.G., Han, G.W., Kuhn, P., Rosen, H. & Stevens, R.C. (2012). Crystal structure of a lipid G protein-coupled receptor. *Science*, **335**, 851–5. 47

Hildebrand, P.W., Preissner, R. & Frömmel, C. (2004). Structural features of transmembrane helices. *FEBS letters*, **559**, 145–151. 21, 36

Hildebrand, P.W., Günther, S., Goede, A., Forrest, L., Frömmel, C. & Preissner, R. (2008). Hydrogen-bonding and packing features of membrane proteins: functional implications. *Biophysical Journal*, **94**, 1945–1953. 46

Hill, J.R. & Deane, C.M. (2013). MP-T: improving membrane protein alignment for structure prediction. *Bioinformatics*, **29**, 54–61. 52, 118

Hill, J.R., Kelm, S., Shi, J. & Deane, C.M. (2011). Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, **27**, i15–i23. 118, 155

Hischenhuber, B., Frommlet, F., Schreiner, W. & Knapp, B. (2012). MH(2)c: Characterization of major histocompatibility α-helices - an information criterion approach. *Computer Physics Communications*, **183**, 1481–1490. 4, 43

Hischenhuber, B., Havlicek, H., Todoric, J., Höllrigl-Binder, S., Schreiner, W. & Knapp, B. (2013). Differential geometric analysis of alterations in MH α-helices. *Journal of Computational Chemistry*, **34**, 1862–1879. 4

Holthuis, J.C.M. & Levine, T.P. (2005). Lipid traffic: floppy drives and a superhighway. *Nature Reviews. Molecular Cell Biology*, **6**, 209–220. 19

Hunte, C. & Michel, H. (2002). Crystallisation of membrane proteins mediated by antibody fragments. *Current Opinion in Structural Biology*, **12**, 503–508. 26

Hutchinson, E.G. & Thornton, J.M. (1996). PROMOTIF–a program to identify and analyze structural motifs in proteins. *Protein Science*, **5**, 212–220. 31

Jayasinghe, S., Hristova, K. & White, S.H. (2001). MPtopo: A database of membrane protein topology. *Protein Science*, **10**, 455–458. 59, 88

Jha, A.N., Vishveshwara, S. & Banavar, J.R. (2011). Amino acid interaction preferences in helical membrane proteins. *Protein Engineering, Design, and Selection*, **24**, 579–588. 35

Joh, N.H., Min, A., Faham, S., Whitelegge, J.P., Yang, D., Woods, V.L. & Bowie, J.U. (2008). Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. *Nature*, **453**, 1266–70. 23, 36

Johansson, A.C.V. & Lindahl, E. (2006). Amino-acid solvation structure in transmembrane helices from molecular dynamics simulations. *Biophysical Journal*, **91**, 4450–4463. 38

Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C. & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, **39**, D411–9. 29

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637. 29, 31, 52, 59, 119

Kaduk, C., Duclohier, H., Dathe, M., Wenschuh, H., Beyermann, M., Molle, G. & Bienert, M. (1997). Influence of proline position upon the ion channel activity of alamethicin. *Biophysical Journal*, **72**, 2151–2159. 38

Kahn, P.C. (1989). Defining the axis of a helix. *Computers & Chemistry*, **13**, 185–189. 41, 42, 64

Kandiraju, N., Dua, S. & Conrad, S. (2005). Dihedral angle based dimensionality reduction for protein structural comparison. In *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II*, 14–9, IEEE. 43

Karplus, M. (2014). CHARMM documentation for version c38b1. 40

Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–80. 122

Katritch, V., Reynolds, K.A., Cherezov, V., Hanson, M.A., Roth, C.B., Yeager, M. & Abagyan, R. (2009). Analysis of full and partial agonists binding to β2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *Journal of Molecular Recognition*, **22**, 307–318. 37

Katritch, V., Cherezov, V. & Stevens, R.C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annual Review of Pharmacology and Toxicology*, **53**, 531–556. 146, 178

Kauko, A., Illergård, K. & Elofsson, A. (2008). Coils in the membrane core are conserved and functionally important. *Journal of Molecular Biology*, **380**, 170–180. 24, 37, 46, 51

Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., Sarmenta, L., Blanchette, M. & Waldispühl, J. (2012). Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS ONE*, **7**, e31362. 86, 112

Keenan, T.W. & Morre, D.J. (1970). Phospholipid class and fatty acid composition of Golgi apparatus isolated from rat liver and comparison with other cell fractions. *Biochemistry*, **9**, 19–25. 19

Kelm, S., Shi, J. & Deane, C.M. (2009). iMembrane: homology-based membrane-insertion of proteins. *Bioinformatics*, **25**, 1086–1088. 52, 121, 148

Kelm, S., Shi, J. & Deane, C.M. (2010). MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics*, **26**, 2833–2840. 28, 47, 52, 146

Kessel, A. & Ben-Tal, N. (2011). *Introduction to Proteins: Structure, Function, and Motion*. CRC Press, Chapman & Hall. 14, 18, 19, 25

Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popovic, Z., Baker, D. & Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 18949–18953. 86, 112

King, S.M. & Johnson, W.C. (1999). Assigning secondary structure from protein coordinate data. *Proteins*, **35**, 313–20. 33

Klotz, I.M., Langerman, N.R. & Darnall, D.W. (1970). Quaternary structure of proteins. *Annual Review of Biochemistry*, **39**, 25–62. 4, 18

Kneissl, B., Mueller, S.C., Tautermann, C.S. & Hildebrandt, A. (2011). String kernels and high-quality data set for improved prediction of kinked helices in α-helical membrane proteins. *Journal of Chemical Information and Modeling*, **51**, 3017–3025. vii, 4, 37, 42, 44, 45, 47, 51, 55, 57, 59, 61, 79, 85, 86, 87, 88, 93, 99, 110, 112, 114, 115, 139, 144, 148, 176, 177, 180

Kneller, G.R. & Calligari, P. (2006). Efficient characterization of protein secondary structure in terms of screw motions. *Acta Crystallographica. Section D, Biological Crystallography*, **62**, 302–11. 43

Konagurthu, A.S., Lesk, A.M. & Allison, L. (2012). Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics*, **28**, 97–105. 32, 33

Kozma, D., Simon, I. & Tusnády, G.E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Research*, **41**, D524–D529. 119, 120, 148

Kuechler, A., Hauffa, B.P., Köninger, A., Kleinau, G., Albrecht, B., Horsthemke, B. & Gromoll, J. (2010). An unbalanced translocation unmasks a recessive mutation in the follicle-stimulating hormone receptor (FSHR) gene and causes FSH resistance. *European Journal of Human Genetics*, **18**, 656–661. 44

Kufareva, I., Rueda, M., Katritch, V., Stevens, R.C. & Abagyan, R. (2011). Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure*, **19**, 1108–1126. 4, 47, 51, 146

Kufareva, I., Katritch, V., Stevens, R.C. & Abagyan, R. (2014). Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure*, **22**, 1120–1139. 146

Kumar, P. & Bansal, M. (2012). HELANAL-Plus: a web server for analysis of helix geometry in protein structures. *Journal of Biomolecular Structure & Dynamics*, **30**, 773–783. vii, 4, 41, 54, 55, 57, 63, 64, 85, 86, 99, 112, 130

Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J.P. (1997). P-SEA: a new efficient assignment of secondary structure from Cα trace of proteins. *Bioinformatics*, **13**, 291–295. 33

Land, K., Slosar, A., Lintott, C., Andreescu, D., Bamford, S., Murray, P., Nichol, R., Raddick, M.J., Schawinski, K., Szalay, A., Thomas, D. & Vandenberg, J. (2008). Galaxy Zoo: the large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, **388**, 1686–1692. 85, 86

Langelaan, D.N., Wieczorek, M., Blouin, C. & Rainey, J.K. (2010). Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *Journal of Chemical Information and Modeling*, **50**, 2213–2220. vii, 4, 37, 42, 44, 45, 51, 54, 55, 57, 63, 85, 86, 99, 110, 112, 114, 118, 120, 122, 130, 176

Langelaan, D.N., Reddy, T., Banks, A.W., Dellaire, G., Dupré, D.J. & Rainey, J.K. (2013). Structural features of the apelin receptor N-terminal tail and first transmembrane segment implicated in ligand binding and receptor trafficking. *Biochimica et biophysica acta*, **1828**, 1471–83. 38

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291. 25

Lee, J. & Kim, S.H. (2009). Water polygons in high-resolution protein crystal structures. *Protein Science*, **18**, 1370–1376. 25

Leman, J.K., Mueller, R., Karakas, M., Woetzel, N. & Meiler, J. (2013). Simultaneous prediction of protein secondary structure and trans-membrane spans. *Proteins: Structure, Function, and Bioinformatics*, **81**, 1127–1140. 36

Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *Journal of Molecular Biology*, **114**, 181–239. 29

Linderstrø m Lang, K.U. (1952). *Lane Medical Lectures: Proteins and enzymes*. Stanford University Press. 4

Lingwood, D. & Simons, K. (2010). Lipid rafts as a membrane-organizing principle. *Science*, **327**, 46–50. 19

Lintott, C., Schawinski, K., Bamford, S., Land, K., Thomas, D., Edmondson, E., Masters, K., Robert, C., Raddick, M.J., Szalay, A., Andreescu, D., Murray, P. & Vandenberg, J. (2010). Galaxy Zoo 1 : Data Release of Morphological Classifications for nearly 900,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, **14**, 1–14. 112, 113

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P. & Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, **12**, 1–12. 112, 113

Liu, J., Tan, H. & Rost, B. (2002). Loopy proteins appear conserved in evolution. *Journal of Molecular Biology*, **322**, 53–64. 14

Lovell, S.C., Davis, I.W., Arendall, W.B.I., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S. & Richardson, D.C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Genetics*, **50**, 437–450. 42, 54, 120

Luengo-Oroz, M.A., Arranz, A. & Frean, J. (2012). Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, **14**, e167. 113

Lupyan, D., Leo-Macias, A. & Ortiz, A.R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263. 156

Mai, T.L. & Chen, C.M. (2014). Computational prediction of kink properties of helices in membrane proteins. *Journal of Computer-Aided Molecular Design*, **28**, 99–109. 47, 48, 118

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press and Elsevier Science, 1st edn. 61

Marsico, A., Henschel, A., Winter, C., Tuukkanen, A., Vassilev, B., Scheubert, K. & Schroeder, M. (2010a). Structural fragment clustering reveals novel structural and functional motifs in α-helical transmembrane proteins. *BMC Bioinformatics*, **11**, 204. 45, 136, 137, 143

Marsico, A., Scheubert, K., Tuukkanen, A., Henschel, A., Winter, C., Winnenburg, R. & Schroeder, M. (2010b). MeMotif: a database of linear motifs in α-helical transmembrane proteins. *Nucleic Acids Research*, **38**, D181–D189. 45, 136, 138

Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A.G. & Gibrat, J.F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Structural Biology*, **5**, 17. 10, 14, 32, 33, 52, 53

Maslennikov, I. & Choe, S. (2013). Advances in NMR structures of integral membrane proteins. *Current Opinion in Structural Biology*, **23**, 555–562. 27

Mazna, P., Grycova, L., Balik, A., Zemkova, H., Friedlova, E., Obsilova, V., Obsil, T. & Teisinger, J. (2008). The role of proline residues in the structure and function of human MT2 melatonin receptor. *Journal of pineal research*, **45**, 361–372. 38

Meruelo, A.D., Samish, I. & Bowie, J.U. (2011). TMKink: A method to predict transmembrane helix kinks. *Protein Science*, **20**, 1256–1264. vii, 4, 37, 41, 44, 45, 47, 51, 54, 58, 85, 99, 110, 114, 144, 176

Mezei, M. (2010). Simulaid: a simulation facilitator and analysis program. *Journal of Computational Chemistry*, **31**, 2658–2568. vii, 41

Middleton, D.A. (2007). Solid-state NMR spectroscopy as a tool for drug design: from membrane-embedded targets to amyloid fibrils. *Biochemical Society Transactions*, **35**, 985–990. 27

Mihajlovic, M. & Lazaridis, T. (2012). Charge distribution and imperfect amphipathicity affect pore formation by antimicrobial peptides. *Biochimica et Biophysica Acta*, **1818**, 1274–1283. 38

Mitra, K., Ubarretxena-Belandia, I., Taguchi, T., #, Mitra, K., Ubarretxena-Belandia, Iban Taguchi, T., Warren, G. & Engelman, D.M. (2004). Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4083–4088. 19

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623. 31, 52, 119, 123, 148

Mokrab, Y., Stevens, T.J. & Mizuguchi, K. (2010). A structural dissection of amino acid substitutions in helical transmembrane proteins. *Proteins*, **78**, 2895–907. 23, 24

Morein, S., Andersson, A.S., Rilfors, L. & Lindblom, G. (1996). Wild-type Escherichia coli cells regulate the membrane lipid composition in a "window" between gel and non-lamellar structures. *Journal of Biological Chemistry*, **271**, 6801–6809. 18

Muñoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. *Proteins*, **20**, 301–311. 34

Murray, D.T., Lu, Y., Cross, T.A. & Quine, J.R. (2011). Geometry of kinked protein helices from NMR data. *Journal of Magnetic Resonance*, **210**, 82–9. 43

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**, 536–540. 17

Newman, M.S. (1955). A notation for the study of certain stereochemical problems. *Journal of Chemical Education*, **32**, 344–347. 6

Nguyen, T., Wang, S., Anugu, V. & Rose, N. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*, **262**, 824–833. 86, 113

Ni, Z., Bikadi, Z., Shuster, D.L., Zhao, C., Rosenberg, M.F. & Mao, Q. (2011). Identification of proline residues in or near the transmembrane helices of the human breast cancer resistance protein (BCRP/ABCG2) that are important for transport activity and substrate specificity. *Biochemistry*, **50**, 8057–8066. 4, 118

Nugent, T. & Jones, D.T. (2011). Membrane protein structural bioinformatics. *Journal of Structural Biology*, **179**, 327–337. 4, 36, 118

Op den Kamp, J. (1979). Lipid asymmetry in membranes. *Annual Review of Biochemistry*, **48**, 47–71. 19

Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, **5**, 993–996. 21, 51

Pace, C.N. & Scholtz, J.M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal*, **75**, 422–427. 34

Page, R., Kim, S. & Cross, T. (2008). Transmembrane helix uniformity examined by spectral mapping of torsion angles. *Structure*, **16**, 787–797. 23, 36

Pal, L., Chakrabarti, P. & Basu, G. (2003). Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. *Journal of molecular biology*, **326**, 273–91. 35

Parvanta, C., Roth, Y. & Keller, H. (2013). Crowdsourcing 101: a few basics to make you the leader of the pack. *Health Promotion Practice*, **14**, 163–167. 85

Pauling, L., Corey, R.B. & Branson, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, **37**, 205–211. 11, 14, 29, 52

Preissner, R., Egner, U. & Saenger, W. (1991). Occurrence of bifurcated three-center hydrogen bonds in proteins. *FEBS Letters*, **288**, 192–196. 11

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). Numerical recipes in C (2nd ed.): the art of scientific computing. 67

Quine, J. (1999). Helix parameters and protein structure using quaternions. *Journal of Molecular Structure: THEOCHEM*, **460**, 53–66. 43

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 89, 125

Ranard, B.L., Ha, Y.P., Meisel, Z.F., Asch, D.A., Hill, S.S., Becker, L.B., Seymour, A.K. & Merchant, R.M. (2014). Crowdsourcing-Harnessing the Masses to Advance Health and Medicine, a Systematic Review. *Journal of General Internal Medicine*, **29**, 187–203. 85

Rasmussen, S.r.G.F., Choi, H.J., Fung, J.J., Pardon, E., Casarosa, P., Chae, P.S., Devree, B.T., Rosenbaum, D.M., Thian, F.S., Kobilka, T.S., Schnapp, A., Konetzki, I., Sunahara, R.K., Gellman, S.H., Pautsch, A., Steyaert, J., Weis, W.I. & Kobilka, B.K. (2011). Structure of a nanobody-stabilized active state of the $\beta 2$ adrenoceptor. *Nature*, **469**, 175–180. 47

Read, R.J., Adams, P.D., Arendall, W.B., Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J.S., Sheffler, W.H., Smith, J.L., Tickle, I.J., Vriend, G. & Zwart, P.H. (2011). A new generation of crystallographic validation tools for the Protein Data Bank. *Structure*, **19**, 1395–1412. 25

Rees, D.C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science*, **245**, 510–513. 21

Rey, J., Devillé, J. & Chabbert, M. (2010). Structural determinants stabilizing helical distortions related to Proline. *Journal of Structural Biology*, **171**, 266–276. 4, 38, 46, 47

Reyes, V.M. (2011). Representation of protein 3D structures in spherical (ρ, ÏŢ, θ) coordinates and two of its potential applications. *Interdisciplinary Sciences, Computational Life Sciences*, **3**, 161–74. 43

Ri, Y., Ballesteros, J.A., Abrams, C.K., Oh, S., Verselis, V.K., Weinstein, H. & Bargiello, T.A. (1999). The role of a conserved proline residue in mediating conformational changes associated with voltage gating of Cx32 gap junctions. *Biophysical Journal*, **76**, 2887–2898. 38

Richards, F.M. & Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**, 71–84. 32

Richardson, J.S. & Richardson, D.C. (1989). Principles and Patterns of Protein Conformation. In G.D. Fasman, ed., *Prediction of Protein Structure and the Principles of Protein Conformation*, chap. Principles, 1–95, Springer US, Boston, MA. 175

Rigoutsos, I., Riek, P., Graham, R.M. & Novotny, J. (2003). Structural details (kinks and non-α conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. *Nucleic Acids Research*, **31**, 4625–4631. 4, 37, 45, 51

Rochaix, J.D. (2014). Regulation and dynamics of the light-harvesting system. *Annual Review of Plant Biology*, **65**, 287–309. 18

Rosenbaum, D.M., Rasmussen, S.r.G.F. & Kobilka, B.K. (2009). The structure and function of G-protein-coupled receptors. *Nature*, **459**, 356–63. 47

Rosenbaum, D.M., Zhang, C., Lyons, J.A., Holl, R., Aragao, D., Arlow, D.H., Rasmussen, S.r.G.F., Choi, H.J., Devree, B.T., Sunahara, R.K., Chae, P.S., Gellman, S.H., Dror, R.O., Shaw, D.E., Weis, W.I., Caffrey, M., Gmeiner, P. & Kobilka, B.K. (2011). Structure and function of an irreversible agonist-β2 adrenoceptor complex. *Nature*, **469**, 236–240. 47

Rudolph, M.G., Stanfield, R.L. & Wilson, I.A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology*, **24**, 419–466. 38, 46

Sali, A. & Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815. 28

Sands, Z.A., Grottesi, A. & Sansom, M.S.P. (2006). The intrinsic flexibility of the Kv voltage sensor and its implications for channel gating. *Biophysical Journal*, **90**, 1598–1606. 178

Sansom, M.S.P. & Weinstein, H. (2000). Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. *Trends in Pharmacological Sciences*, **21**, 445–451. 4, 51, 118, 137, 178

Savitz, D., Sidel, V.W. & Solomon, A.K. (1964). Osmotic properties of human red blood cells. *The Journal of General Physiology*, **48**, 79–94. 26

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M. & Rost, B. (2011). Protein disorder–a breakthrough invention of evolution? *Current Opinion in Structural Biology*, **21**, 412–8. 14

Schrödinger LLC (2014). The PyMOL molecular graphics system, version 1.7.1.3. 13, 44

Schroeder, R.J., Ahmed, S.N., Zhu, Y., London, E. & Brown, D.A. (1998). Cholesterol and Sphingolipid Enhance the Triton X-100 Insolubility of Glycosylphosphatidylinositol-anchored Proteins by Promoting the Formation of Detergent-insoluble Ordered Membrane Domains. *Journal of Biological Chemistry*, **273**, 1150–1157. 19

Schwartz, T.W., Frimurer, T.M., Holst, B., Rosenkilde, M.M. & Elling, C.E. (2006). Molecular mechanism of 7TM receptor activation–a global toggle switch model. *Annual Review of Pharmacology and Toxicology*, **46**, 481–519. 4, 38, 118

Segrest, J.P., De Loof, H., Dohlman, J.G., Brouillette, C.G. & Anantharamaiah, G.M. (1990). Amphipathic helix motif: classes and properties. *Proteins*, **8**, 103–117. 34

Seifert, T., Lund, A., Kneissl, B., Mueller, S.C., Tautermann, C.S. & Hildebrandt, A. (2014). SKINK: a web server for string kernel based kink prediction in α-helices. *Bioinformatics*, DOI: 10.1093/bioinformatics/btu096. 4, 47, 176

Shi, J., Blundell, T.L. & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, **310**, 243–257. 123

Shi, L., Liapakis, G., Xu, R., Guarnieri, F., Ballesteros, J.A. & Javitch, J.A. (2002). β2 adrenergic receptor activation. Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *The Journal of biological chemistry*, **277**, 40989–40996. 38

Siegel, D.P., Cherezov, V., Greathouse, D.V., Koeppe, R.E., Killian, J.A. & Caffrey, M. (2006). Transmembrane peptides stabilize inverted cubic phases in a biphasic length-dependent manner: implications for protein-induced membrane fusion. *Biophysical Journal*, **90**, 200–211. 19

Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L. & Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic acids research*, **41**, D344–7. 144

Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M. & Orengo, C.A. (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research*, **41**, D490–8. 17

Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225. 28, 86

Sinclair, J.C. & Noble, M.E.M. (2004). Protein Lattice. 26

Skibba, R.A., Masters, K.L., Nichol, R.C., Zehavi, I., Hoyle, B., Edmondson, E.M., Bamford, S.P., Cardamone, C.N., Keel, W.C., Lintott, C. & Schawinski, K. (2012). Galaxy Zoo: the environmental dependence of bars and bulges in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, **423**, 1485–1502. 85

Sklenar, H., Etchebest, C. & Lavery, R. (1989). Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins*, **6**, 46–60. 33

Stevens, T.J. & Arkin, I.T. (1999). Are membrane proteins "inside-out" proteins? *Proteins*, **36**, 135–143. 118

Suchyna, T., Xu, L., Gao, F., Fourtner, C. & Nicholson, B. (1993). Identification of a Proline residue as a transduction element involved in voltage gating of gap junctions. *Nature*, **365**, 847–849. 4, 118

Sugeta, H. & Miyazawa, T. (1967). General method for calculating helical parameters of polymer chains from bond lengths, bond angles, and internal-rotation angles. *Biopolymers*, **5**, 673–679. 41, 44, 64

Tang, H., Wang, X.S., Hsieh, J.H. & Tropsha, A. (2012). Do crystal structures obviate the need for theoretical models of GPCRs for structure based virtual screening? *Proteins: Structure, Function, and Bioinformatics*, **80**, 1503–1521. 47

Tate, C., Stevens, R., Regan, L., Clarke, J., Miyano, M., Ago, H., Saino, H., Hori, T. & Ida, K. (2010). Internally bridging water molecule in transmembrane α-helical kink. *Current Opinion in Structural Biology*, **20**, 456–463. 46

Taylor, G.M. & Sanders, D.A. (1999). The role of the membrane-spanning domain sequence in glycoprotein-mediated membrane fusion. *Molecular Biology of the Cell*, **10**, 2803–2815. 38

Taylor, W.R. (2001). Defining linear segments in protein structure. *Journal of Molecular Biology*, **310**, 1135–1150. 33

The UniProt Consortium (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, **41**, D43–D47. 27, 122

Tieleman, D.P., Sansom, M.S. & Berendsen, H.J. (1999). Alamethicin helices in a bilayer and in solution: molecular dynamics simulations. *Biophysical Journal*, **76**, 40–49. 38

Tieleman, D.P., Shrivastava, I.H., Ulmschneider, M. & Sansom, M.S. (2001). Proline-induced hinges in transmembrane helices: possible roles in ion channel gating. *Proteins: Structure, Function, and Genetics*, **44**, 63–72. 4, 38, 118

Traag, V.A., Van Dooren, P. & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, **84**, 016114. 156

Tusnády, G.E., Dosztányi, Z. & Simon, I. (2004). Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, **20**, 2964–2972. 119, 120

Ulmschneider, M.B. & Sansom, M.S.P. (2001). Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1512**, 1–14. 35, 118

van der Kant, R. & Vriend, G. (2014). Alpha-bulges in G protein-coupled receptors. *International Journal of Molecular Sciences*, **15**, 7841–7864. 4, 178

van der Walt, S., Colbert, S.C. & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, **13**, 22–30. 61, 75

van Rossum, G. & de Boer, J. (1991). Interactively testing remote servers using the Python programming language. *CWI Quarterly*, **4**, 283–304. 89

Veenhoff, L.M., Heuberger, E.H. & Poolman, B. (2002). Quaternary structure and function of transport proteins. *Trends in Biochemical Sciences*, **27**, 242–249. 18

Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Computer Applications in the Biosciences : CABIOS*, **5**, 115–121. 123

Vingron, M. & Sibbald, P.R. (1993). Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 8777–8781. 123

Visiers, I., Braunheim, B.B. & Weinstein, H. (2000). Prokink: a protocol for numerical evaluation of helix distortions by proline. *Protein Engineering, Design, and Selection*, **13**, 603–606. vi, 37, 40, 51, 58, 85

von Ahn, L., Maurer, B., McMillen, C., Abraham, D. & Blum, M. (2008). re-CAPTCHA: human-based character recognition via Web security measures. *Science*, **321**, 1465–1468. 86

von Heijne, G. (2006). Membrane-protein topology. *Nature reviews. Molecular cell biology*, **7**, 909–18. 35

Walters, R.F.S. & DeGrado, W.F. (2006). Helix-packing motifs in membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13658–13663. 36

Wang, G. & Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591. 119

Wang, T.Y., Leventis, R. & Silvius, J.R. (2001). Partitioning of lipidated peptide sequences into liquid-ordered lipid domains in model and biological membranes. *Biochemistry*, **40**, 13031–13040. 19

Weber, M., Tome, L., Otzen, D. & Schneider, D. (2012). A Ser residue influences the structure and stability of a Pro-kinked transmembrane helix dimer. *Biochimica et Biophysica Acta*, **1818**, 2103–2107. 4, 44, 45, 118, 137

Werner, T. & Church, W.B. (2013). Kink characterization and modeling in transmembrane protein structures. *Journal of Chemical Information and Modeling*, **53**, 2926–2936. vii, 4, 37, 42, 45, 47, 51, 79, 114, 141, 143, 144, 181

White, S.H. & Wimley, W.C. (1999). Membrane protein folding and stability: physical principles. *Annual Review of Biophysics and Biomolecular Structure*, **28**, 319–365. 26, 28, 119, 120, 123, 148

William Humphrey, Andrew Dalke & Klaus Schulten (1996). VMD âĂŞ Visual Molecular Dynamics. *Journal of Molecular Graphics*, **14**, 33–38. 44

Wilman, H.R., Ebejer, J.P., Shi, J., Deane, C.M. & Knapp, B. (2014a). Crowdsourcing yields a new standard for kinks in protein helices. *Journal of Chemical Information and Modeling*, **54**, 2585–2593. 83

Wilman, H.R., Shi, J. & Deane, C.M. (2014b). Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins: Structure, Function, and Bioinformatics*, **82**, 1960–1970. 4, 86, 112, 114, 117

Wong, W.C., Maurer-Stroh, S., Schneider, G. & Eisenhaber, F. (2012). Transmembrane helix: simple or complex. *Nucleic Acids Research*, **40**, W370–W375. 36

Woolfson, D.N., Mortishire-Smith, R.J. & Williams, D.H. (1991). Conserved positioning of proline residues in membrane-spanning helices of ion-channel proteins. *Biochemical and Biophysical Research Communications*, **175**, 733–737. 38

Worth, C.L., Kleinau, G. & Krause, G. (2009). Comparative sequence and structural analyses of G-protein-coupled receptor crystal structures and implications for molecular models. *PloS ONE*, **4**, e7011. 47, 51

Yeagle, P.L. (2004). *The Structure of Biological Membranes*. CRC Press, 2nd edn. 19

Yohannan, S., Faham, S., Yang, D., Grosfeld, D., Chamberlain, A.K. & Bowie, J.U. (2004a). A Cα-H...O hydrogen bond in a membrane protein is not stabilizing. *Journal of the American Chemical Society*, **126**, 2284–2285. 23

Yohannan, S., Faham, S., Yang, D., Whitelegge, J.P. & Bowie, J.U. (2004b). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 959–963. 4, 45, 51, 118

Yohannan, S., Yang, D., Faham, S., Boulting, G., Whitelegge, J. & Bowie, J.U. (2004c). Proline substitutions are not easily accommodated in a membrane protein. *Journal of Molecular Biology*, **341**, 1–6. 45

Zacharias, J. & Knapp, E.W. (2014). Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC. *Journal of Chemical Information and Modeling*, **54**, 2166–2179. 31, 33

Zemla, A., Venclovas, C., Moult, J. & Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins: Structure, Function, and Genetics*, **37**, 22–29. 146

Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**, 2302–2309. 153

Zooniverse Team (2014). Zooniverse - Real Science Online. 85

Appendix A

This appendix contains a table of the gold standard kink data set created from AHAH, our crowdsourcing approach to kink identification.

Table A1: AHAH gold standard data set. Helix ID is in the format (e.g.) 3DHW_A_53_76, where 3DHW is the protein PDB code, A is the chain identifier, 53 is the PDB number of the first residue in the helix, and 76 is the PDB number of the final residue in the helix. The number of people who identified a helix as (e.g.) kinked, can be calculated by multiplying the proportion of kinked responses by the total number of responses. For example, 3DHW_A_53_76 was annotated as kinked by $0.935 \times 31 = 29$ participants, as curved by $0.065 \times 31 = 2$ participants, and as straight by $0.000 \times 31 = 0$ participants. A plain text version of this data set can be downloaded from http://www.stats.ox.ac.uk/research/proteins/resources#AHAH, along with the raw data collected from the AHAH web application.

		Proportion			Total Number
Helix ID	Classification	Kinked	Curved	Straight	of Responses
3DHW_A_53_76	KINK	0.935	0.065	0.000	31
3O0R_B_235_259	KINK	0.886	0.114	0.000	35
2WJN_M_143_165	KINK	0.865	0.081	0.054	37
3GIA_A_145_172	KINK	0.864	0.091	0.045	22
3DH4_A_281_313	KINK	0.844	0.125	0.031	32
2V50_A_428_453	KINK	0.828	0.138	0.034	29
1L0V_D_12_39	KINK	0.806	0.194	0.000	36
3B9Y_A_58_89	KINK	0.800	0.133	0.067	45
1JB0_A_438_467	KINK	0.800	0.200	0.000	35
1OKC_A_210_240	KINK	0.781	0.156	0.062	32

Continued on next page

Continued from previous page

2GFP_A_9_36	KINK	0.775	0.150	0.075	40
1V54_L_19_45	KINK	0.763	0.184	0.053	38
3K07_A_433_461	KINK	0.757	0.243	0.000	37
2GFP_A_322_345	KINK	0.750	0.139	0.111	36
1JB0_I_12_35	KINK	0.750	0.031	0.219	32
2V50_A_959_985	KINK	0.722	0.278	0.000	36
3K07_A_392_421	KINK	0.718	0.231	0.051	39
1V54_C_74_106	KINK	0.714	0.286	0.000	21
3PBL_A_64_92	KINK	0.711	0.263	0.026	38
2KDC_A_95_118	KINK	0.692	0.282	0.026	39
3K07_A_970_996	KINK	0.690	0.276	0.034	29
2BS2_C_122_150	KINK	0.689	0.289	0.022	45
1PW4_A_381_410	KINK	0.682	0.227	0.091	44
1U19_A_289_308	KINK	0.677	0.258	0.065	31
1BGY_C_77_103	KINK	0.667	0.233	0.100	30
2WSW_A_315_338	KINK	0.667	0.303	0.030	33
2VT4_A_324_344	KINK	0.667	0.333	0.000	30
1OKC_A_80_97	KINK	0.654	0.346	0.000	26
2HYD_A_122_158	KINK	0.645	0.323	0.032	31
2BL2_A_52_79	KINK	0.643	0.321	0.036	28
2GSM_A_93_123	KINK	0.639	0.306	0.056	36
2ZJS_Y_184_208	KINK	0.633	0.300	0.067	30
2ZW3_A_74_108	KINK	0.630	0.296	0.074	27
2R6G_G_160_182	KINK	0.628	0.372	0.000	43
3E86_A_89_110	KINK	0.625	0.175	0.200	40
2GFP_A_206_230	KINK	0.625	0.344	0.031	32
2JLN_A_162_189	KINK	0.625	0.300	0.075	40
1XQF_A_199_218	KINK	0.615	0.308	0.077	39
3GIA_A_270_305	KINK	0.614	0.341	0.045	44
1U19_A_152_171	KINK	0.606	0.182	0.212	33
2GSM_B_100_126	KINK	0.600	0.325	0.075	40
3K07_A_863_889	KINK	0.590	0.410	0.000	39
3ODU_A_275_302	KINK	0.583	0.361	0.056	36
3ODU_A_194_225	KINK	0.579	0.316	0.105	38
2J8C_M_146_167	KINK	0.568	0.432	0.000	37
2H8A_A_65_92	KINK	0.567	0.267	0.167	30
3O0R_B_270_296	KINK	0.562	0.094	0.344	32
2R6G_F_366_391	KINK	0.556	0.389	0.056	36
1XIO_A_72_91	KINK	0.556	0.200	0.244	45
1H68_A_72_90	KINK	0.553	0.184	0.263	38
1AR1_A_86_115	KINK	0.550	0.275	0.175	40
2B2F_A_115_136	KINK	0.550	0.300	0.150	40
1EHK_A_18_45	KINK	0.548	0.419	0.032	31
2HYD_A_13_41	KINK	0.545	0.364	0.091	33
1V54_B_16_46	KINK	0.543	0.257	0.200	35
1XIO_A_197_222	KINK	0.536	0.286	0.179	28
1YEW_B_58_83	KINK	0.533	0.467	0.000	30

Continued on next page

Continued from previous page

1BGY_C_346_376	KINK	0.531	0.438	0.031	32
3DHW_A_91_114	KINK	0.528	0.389	0.083	36
2A65_A_44_70	KINK	0.525	0.450	0.025	40
1ZCD_A_292_312	KINK	0.524	0.452	0.024	42
3K07_A_925_954	KINK	0.517	0.345	0.138	29
2F2B_A_5_33	KINK	0.514	0.371	0.114	35
3HFX_A_470_499	KINK	0.514	0.429	0.057	35
2ZY9_A_353_380	CURVED	0.045	0.909	0.045	22
1EZV_D_264_292	CURVED	0.111	0.889	0.000	27
1PW4_A_415_444	CURVED	0.088	0.853	0.059	34
3HFX_A_346_374	CURVED	0.088	0.853	0.059	34
2ONK_C_4_30	CURVED	0.158	0.816	0.026	38
2A65_A_338_371	CURVED	0.133	0.800	0.067	30
2WSW_A_345_375	CURVED	0.159	0.795	0.045	44
2W2E_A_42_70	CURVED	0.061	0.788	0.152	33
3HD6_A_244_270	CURVED	0.146	0.780	0.073	41
2RH1_A_104_135	CURVED	0.118	0.765	0.118	34
3ODU_A_107_138	CURVED	0.220	0.756	0.024	41
1AR1_A_220_249	CURVED	0.229	0.743	0.029	35
3B45_A_149_168	CURVED	0.143	0.743	0.114	35
2E74_A_80_106	CURVED	0.211	0.737	0.053	38
1XL4_A_110_135	CURVED	0.102	0.735	0.163	49
1PV7_A_347_374	CURVED	0.233	0.733	0.033	30
1BCC_E_29_62	CURVED	0.121	0.727	0.152	33
3M71_A_165_189	CURVED	0.265	0.706	0.029	34
1OTS_A_253_283	CURVED	0.273	0.697	0.030	33
1L7V_A_115_138	CURVED	0.214	0.690	0.095	42
3K3F_A_262_285	CURVED	0.265	0.676	0.059	34
3BEH_A_72_90	CURVED	0.297	0.676	0.027	37
3K07_A_1011_1038	CURVED	0.304	0.674	0.022	46
2GSM_A_27_55	CURVED	0.229	0.657	0.114	35
3B60_A_172_212	CURVED	0.257	0.657	0.086	35
2WJN_M_54_76	CURVED	0.156	0.656	0.188	32
1Q90_M_65_93	CURVED	0.125	0.650	0.225	40
3BZ1_C_270_291	CURVED	0.111	0.644	0.244	45
2J8C_M_263_286	CURVED	0.257	0.629	0.114	35
1AR1_A_372_392	CURVED	0.216	0.622	0.162	37
1Q90_B_80_106	CURVED	0.345	0.621	0.034	29
2ZT9_H_4_25	CURVED	0.265	0.618	0.118	34
2QTS_A_428_451	CURVED	0.353	0.618	0.029	34
2J8C_H_13_36	CURVED	0.154	0.615	0.231	39
1Q90_N_72_95	CURVED	0.174	0.609	0.217	23
3K3F_A_99_121	CURVED	0.091	0.606	0.303	33
3B4R_A_91_112	CURVED	0.368	0.605	0.026	38
1XQF_A_226_251	CURVED	0.233	0.605	0.163	43
2BL2_A_129_155	CURVED	0.400	0.600	0.000	35
3DDL_A_92_111	CURVED	0.406	0.594	0.000	32

Continued on next page

Continued from previous page

1KQF_C_13_36	CURVED	0.282	0.590	0.128	39
3PBL_A_363_385	CURVED	0.290	0.581	0.129	31
1XFH_A_80_106	CURVED	0.108	0.568	0.324	37
1RWT_A_56_87	CURVED	0.297	0.568	0.135	37
1LNQ_A_73_96	CURVED	0.400	0.567	0.033	30
1E12_A_60_80	CURVED	0.282	0.564	0.154	39
1JB0_M_6_28	CURVED	0.317	0.561	0.122	41
1BCC_C_78_104	CURVED	0.441	0.559	0.000	34
3BZ1_Z_2_28	CURVED	0.425	0.550	0.025	40
2BS2_C_203_236	CURVED	0.162	0.541	0.297	37
2KDC_A_33_47	CURVED	0.268	0.537	0.195	41
1BCC_C_347_376	CURVED	0.452	0.524	0.024	42
1JB0_L_44_65	CURVED	0.238	0.524	0.238	21
3BZ1_B_449_475	CURVED	0.379	0.517	0.103	29
2WIT_A_395_425	CURVED	0.414	0.517	0.069	29
3KP9_A_101_123	CURVED	0.200	0.514	0.286	35
2E74_F_4_28	CURVED	0.314	0.514	0.171	35
1ZOY_D_39_62	CURVED	0.405	0.514	0.081	37
1FFT_A_380_400	CURVED	0.349	0.512	0.140	43
3GIA_A_323_337	STRAIGHT	0.000	0.000	1.000	29
1BGY_K_17_34	STRAIGHT	0.000	0.020	0.980	49
2B2F_A_215_240	STRAIGHT	0.000	0.022	0.978	46
1RHZ_A_76_89	STRAIGHT	0.028	0.000	0.972	36
2ZXE_A_992_1012	STRAIGHT	0.000	0.029	0.971	35
2J8C_L_34_55	STRAIGHT	0.030	0.000	0.970	33
1KPL_A_216_232	STRAIGHT	0.000	0.030	0.970	33
3KCU_A_115_134	STRAIGHT	0.000	0.033	0.967	30
3B45_A_96_112	STRAIGHT	0.000	0.033	0.967	30
1RC2_A_162_177	STRAIGHT	0.033	0.000	0.967	30
3A7K_A_147_168	STRAIGHT	0.036	0.000	0.964	28
2Z73_A_150_168	STRAIGHT	0.000	0.036	0.964	28
3GD8_A_232_249	STRAIGHT	0.000	0.045	0.955	44
1RHZ_C_31_49	STRAIGHT	0.028	0.028	0.944	36
3EHZ_A_200_212	STRAIGHT	0.029	0.029	0.943	35
3HQK_A_123_142	STRAIGHT	0.029	0.029	0.943	35
1PV7_A_382_398	STRAIGHT	0.029	0.029	0.943	35
1KQF_C_112_131	STRAIGHT	0.000	0.059	0.941	34
3D31_C_194_210	STRAIGHT	0.070	0.000	0.930	43
2F2B_A_224_243	STRAIGHT	0.024	0.048	0.929	42
3BZ1_B_239_257	STRAIGHT	0.053	0.026	0.921	38
3M71_A_15_30	STRAIGHT	0.081	0.000	0.919	37
2B2F_A_86_104	STRAIGHT	0.059	0.029	0.912	34
3C02_A_224_240	STRAIGHT	0.059	0.029	0.912	34
1Z98_A_200_212	STRAIGHT	0.022	0.067	0.911	45
1WPG_A_896_911	STRAIGHT	0.044	0.044	0.911	45
3MP7_A_242_259	STRAIGHT	0.047	0.047	0.907	43
3HD6_A_345_360	STRAIGHT	0.062	0.031	0.906	32

Continued on next page

Continued from previous page

3KCU_A_32_56	STRAIGHT	0.062	0.031	0.906	32
1RHZ_A_170_187	STRAIGHT	0.094	0.000	0.906	32
3KCU_A_162_183	STRAIGHT	0.051	0.051	0.897	39
2R9R_B_162_182	STRAIGHT	0.034	0.069	0.897	29
2VPZ_C_21_40	STRAIGHT	0.103	0.000	0.897	29
3DIN_C_304_322	STRAIGHT	0.053	0.053	0.895	38
3KP9_A_74_88	STRAIGHT	0.043	0.064	0.894	47
2W2E_A_201_215	STRAIGHT	0.071	0.036	0.893	28
2NQ2_A_99_112	STRAIGHT	0.027	0.081	0.892	37
1WPG_A_832_852	STRAIGHT	0.083	0.028	0.889	36
1V54_C_157_182	STRAIGHT	0.037	0.074	0.889	27
3BZ1_H_30_49	STRAIGHT	0.000	0.111	0.889	27
2WSW_A_230_248	STRAIGHT	0.000	0.111	0.889	36
3BZ1_B_203_217	STRAIGHT	0.023	0.091	0.886	44
1RC2_A_205_226	STRAIGHT	0.029	0.086	0.886	35
1JB0_B_651_671	STRAIGHT	0.088	0.029	0.882	34
2WIT_A_490_509	STRAIGHT	0.040	0.080	0.880	25
2A65_A_196_214	STRAIGHT	0.061	0.061	0.879	33
1NEK_C_69_95	STRAIGHT	0.091	0.030	0.879	33
2ONK_C_232_251	STRAIGHT	0.024	0.098	0.878	41
3HD6_A_8_28	STRAIGHT	0.024	0.098	0.878	41
1Q90_D_128_147	STRAIGHT	0.050	0.075	0.875	40
1VGO_A_111_131	STRAIGHT	0.053	0.079	0.868	38
2GSM_A_137_155	STRAIGHT	0.067	0.067	0.867	30
1J4N_A_143_159	STRAIGHT	0.054	0.081	0.865	37
1VGO_A_15_35	STRAIGHT	0.000	0.138	0.862	29
1PV7_A_288_307	STRAIGHT	0.028	0.111	0.861	36
3HFX_A_54_70	STRAIGHT	0.083	0.056	0.861	36
1Q16_C_183_197	STRAIGHT	0.095	0.048	0.857	42
2ZXE_A_954_969	STRAIGHT	0.071	0.071	0.857	28
1UAZ_A_90_107	STRAIGHT	0.098	0.049	0.854	41
1UAZ_A_17_36	STRAIGHT	0.025	0.125	0.850	40
3BZ1_B_136_156	STRAIGHT	0.075	0.075	0.849	53
2B2F_A_247_261	STRAIGHT	0.000	0.152	0.848	33
1Q90_L_5_26	STRAIGHT	0.043	0.109	0.848	46
3MP7_A_410_426	STRAIGHT	0.115	0.038	0.846	26
1JB0_X_11_30	STRAIGHT	0.026	0.132	0.842	38
1YMG_A_39_58	STRAIGHT	0.028	0.139	0.833	36
3DHW_A_188_205	STRAIGHT	0.000	0.167	0.833	36
1Q16_C_125_145	STRAIGHT	0.059	0.118	0.824	34
1L0V_D_99_115	STRAIGHT	0.036	0.143	0.821	28
2E74_E_3_28	STRAIGHT	0.077	0.103	0.821	39
2J8C_M_114_138	STRAIGHT	0.079	0.105	0.816	38
3HQK_A_385_404	STRAIGHT	0.053	0.132	0.816	38
1Q90_A_253_278	STRAIGHT	0.097	0.097	0.806	31
3KLY_A_161_181	STRAIGHT	0.056	0.139	0.806	36
3O0R_B_381_414	STRAIGHT	0.075	0.125	0.800	40

Continued on next page

Continued from previous page

2R6G_F_278_307	STRAIGHT	0.050	0.150	0.800	40
2R6G_F_40_57	STRAIGHT	0.103	0.103	0.795	39
1XQF_A_349_379	STRAIGHT	0.069	0.138	0.793	29
1YMG_A_84_106	STRAIGHT	0.043	0.174	0.783	46
1KPL_A_422_438	STRAIGHT	0.094	0.125	0.781	32
1JB0_A_672_688	STRAIGHT	0.167	0.056	0.778	36
3B45_A_228_241	STRAIGHT	0.083	0.139	0.778	36
3PBL_A_147_165	STRAIGHT	0.065	0.161	0.774	31
3ORG_A_238_255	STRAIGHT	0.100	0.133	0.767	30
3M71_A_226_244	STRAIGHT	0.118	0.118	0.765	34
1OKC_A_177_199	STRAIGHT	0.105	0.132	0.763	38
1EYS_H_13_33	STRAIGHT	0.103	0.154	0.744	39
3EHZ_A_254_280	STRAIGHT	0.086	0.171	0.743	35
3K07_A_895_917	STRAIGHT	0.171	0.086	0.743	35
3JYC_A_159_183	STRAIGHT	0.038	0.231	0.731	26
1JB0_B_39_68	STRAIGHT	0.108	0.162	0.730	37
1RC2_A_81_104	STRAIGHT	0.111	0.167	0.722	36
2GFP_A_160_175	STRAIGHT	0.100	0.180	0.720	50
2VPZ_C_227_245	STRAIGHT	0.125	0.156	0.719	32
1UAZ_A_112_132	STRAIGHT	0.143	0.143	0.714	28
2ZXE_A_909_932	STRAIGHT	0.097	0.194	0.710	31
3C02_A_10_35	STRAIGHT	0.097	0.194	0.710	31
1RHZ_A_402_423	STRAIGHT	0.062	0.229	0.708	48
3DIN_C_393_412	STRAIGHT	0.088	0.206	0.706	34
1WPG_A_965_988	STRAIGHT	0.265	0.029	0.706	34
1JB0_L_75_99	STRAIGHT	0.108	0.189	0.703	37
1C3W_A_38_61	STRAIGHT	0.212	0.091	0.697	33
3MP7_A_434_458	STRAIGHT	0.125	0.188	0.688	32
3MP7_A_148_170	STRAIGHT	0.200	0.120	0.680	25
2VL0_A_261_284	STRAIGHT	0.147	0.176	0.676	34
2VPZ_C_148_166	STRAIGHT	0.147	0.176	0.676	34
1H68_A_97_115	STRAIGHT	0.135	0.189	0.676	37
1ZOY_C_120_139	STRAIGHT	0.056	0.278	0.667	36
3BZ1_C_155_179	STRAIGHT	0.200	0.143	0.657	35
2V50_A_899_917	STRAIGHT	0.276	0.069	0.655	29
1XQF_A_41_61	STRAIGHT	0.034	0.310	0.655	29
2NQ2_A_236_256	STRAIGHT	0.152	0.212	0.636	33
1EYS_L_95_118	STRAIGHT	0.098	0.268	0.634	41
3B8C_A_242_262	STRAIGHT	0.200	0.178	0.622	45
1ZCD_A_150_174	STRAIGHT	0.111	0.278	0.611	36
2ZJS_Y_217_236	STRAIGHT	0.171	0.220	0.610	41
1JB0_A_534_558	STRAIGHT	0.280	0.120	0.600	25
1JB0_B_521_546	STRAIGHT	0.103	0.310	0.586	29
2FBW_C_36_61	STRAIGHT	0.111	0.306	0.583	36
2VV5_A_64_86	STRAIGHT	0.152	0.273	0.576	33
1JB0_F_67_88	STRAIGHT	0.071	0.357	0.571	28
2FYN_A_362_380	STRAIGHT	0.162	0.270	0.568	37

Continued on next page

Continued from previous page

1RC2_A_37_53	STRAIGHT	0.094	0.344	0.562	32
1WPG_A_752_780	STRAIGHT	0.094	0.344	0.562	32
1VGO_A_87_104	STRAIGHT	0.324	0.118	0.559	34
1VGO_A_43_64	STRAIGHT	0.316	0.132	0.553	38
2GFP_A_237_255	STRAIGHT	0.275	0.175	0.550	40
2Z73_A_33_60	STRAIGHT	0.108	0.351	0.541	37
1EYS_L_180_206	STRAIGHT	0.103	0.379	0.517	29
1V54_G_14_38	UNASSIGNED	0.500	0.235	0.265	34
2FYN_A_91_117	UNASSIGNED	0.500	0.412	0.088	34
1ZOY_C_86_111	UNASSIGNED	0.500	0.324	0.176	34
2VT4_A_157_177	UNASSIGNED	0.488	0.140	0.372	43
3HQK_A_348_373	UNASSIGNED	0.487	0.385	0.128	39
2WJN_M_199_224	UNASSIGNED	0.485	0.333	0.182	33
3BZ1_C_109_133	UNASSIGNED	0.452	0.405	0.143	42
1PW4_A_350_372	UNASSIGNED	0.433	0.267	0.300	30
2BHW_A_56_87	UNASSIGNED	0.424	0.485	0.091	33
1XQF_A_126_147	UNASSIGNED	0.408	0.490	0.102	49
3K07_A_362_383	UNASSIGNED	0.405	0.216	0.378	37
2R9R_B_256_275	UNASSIGNED	0.405	0.500	0.095	42
2GFP_A_100_126	UNASSIGNED	0.400	0.433	0.167	30
1V54_A_408_434	UNASSIGNED	0.400	0.343	0.257	35
2GSM_A_451_477	UNASSIGNED	0.395	0.372	0.233	43
1FFT_A_55_79	UNASSIGNED	0.394	0.485	0.121	33
3BEH_A_187_226	UNASSIGNED	0.389	0.250	0.361	36
1VGO_A_207_229	UNASSIGNED	0.375	0.219	0.406	32
1EHK_A_293_326	UNASSIGNED	0.371	0.429	0.200	35
1BGY_E_34_60	UNASSIGNED	0.343	0.371	0.286	35
1FFT_A_188_213	UNASSIGNED	0.341	0.366	0.293	41
1JB0_B_579_607	UNASSIGNED	0.323	0.452	0.226	31
1V54_A_372_396	UNASSIGNED	0.308	0.256	0.436	39
2ZJS_Y_105_132	UNASSIGNED	0.306	0.444	0.250	36
1BGY_C_35_54	UNASSIGNED	0.306	0.250	0.444	36
1RHZ_A_256_278	UNASSIGNED	0.300	0.425	0.275	40
1UAZ_A_172_197	UNASSIGNED	0.293	0.390	0.317	41
2B2F_A_37_56	UNASSIGNED	0.292	0.417	0.292	24
1V54_C_16_36	UNASSIGNED	0.286	0.262	0.452	42
2VL0_A_205_222	UNASSIGNED	0.281	0.375	0.344	32
3GIA_A_85_112	UNASSIGNED	0.273	0.394	0.333	33
1C3W_A_203_224	UNASSIGNED	0.257	0.286	0.457	35
2ZJS_Y_271_288	UNASSIGNED	0.242	0.485	0.273	33
2WIT_A_104_119	UNASSIGNED	0.238	0.357	0.405	42
1H68_A_156_179	UNASSIGNED	0.237	0.368	0.395	38
1EZV_C_111_134	UNASSIGNED	0.231	0.385	0.385	26
3DDL_A_13_36	UNASSIGNED	0.227	0.341	0.432	44
1BCC_C_321_341	UNASSIGNED	0.219	0.312	0.469	32
3HD6_A_123_141	UNASSIGNED	0.214	0.500	0.286	42
1NKZ_A_13_36	UNASSIGNED	0.211	0.289	0.500	38

Continued on next page

Continued from previous page

2FYN_A_126_149	UNASSIGNED	0.189	0.351	0.459	37
3HD6_A_387_415	UNASSIGNED	0.188	0.479	0.333	48
2FYN_C_12_35	UNASSIGNED	0.184	0.474	0.342	38
1C3W_A_132_152	UNASSIGNED	0.154	0.359	0.487	39
1ZCD_A_358_380	UNASSIGNED	0.152	0.364	0.485	33
1KPL_A_78_97	UNASSIGNED	0.143	0.371	0.486	35
3B9Y_A_321_338	UNASSIGNED	0.119	0.452	0.429	42
3B45_A_172_192	UNASSIGNED	0.083	0.417	0.500	36

Appendix B

Supplementary figures and tables for Chapter 4. This includes figures made from the kinks identified by the modified MC-Helan method, and a table of the PDB chains used in the membrane data.

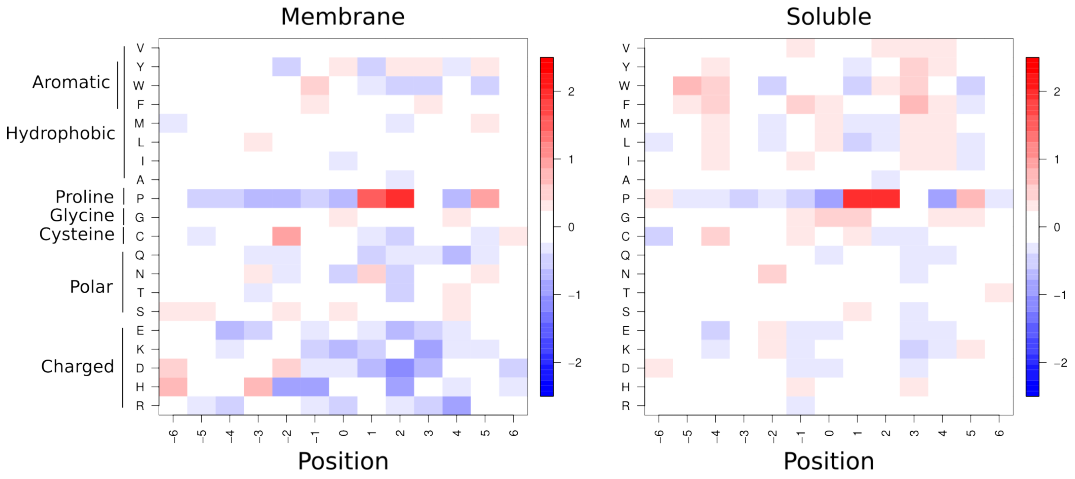


Figure B1: **Amino acid propensities for membrane (left) and soluble (right) kinks, identified by MC-Helan.** Each value is the propensity for a given residue to be at a given position relative to a kink. These are calculated using sequence profiles. Position 0 is the kink residue, position +1 is one residue towards the C-terminus from the kink residue, as shown in Figure 4.4. Red indicates positive propensities, while blue indicates negative propensities. Residues with positive propensities are found more frequently at that position in kinked helices than in helices in general.

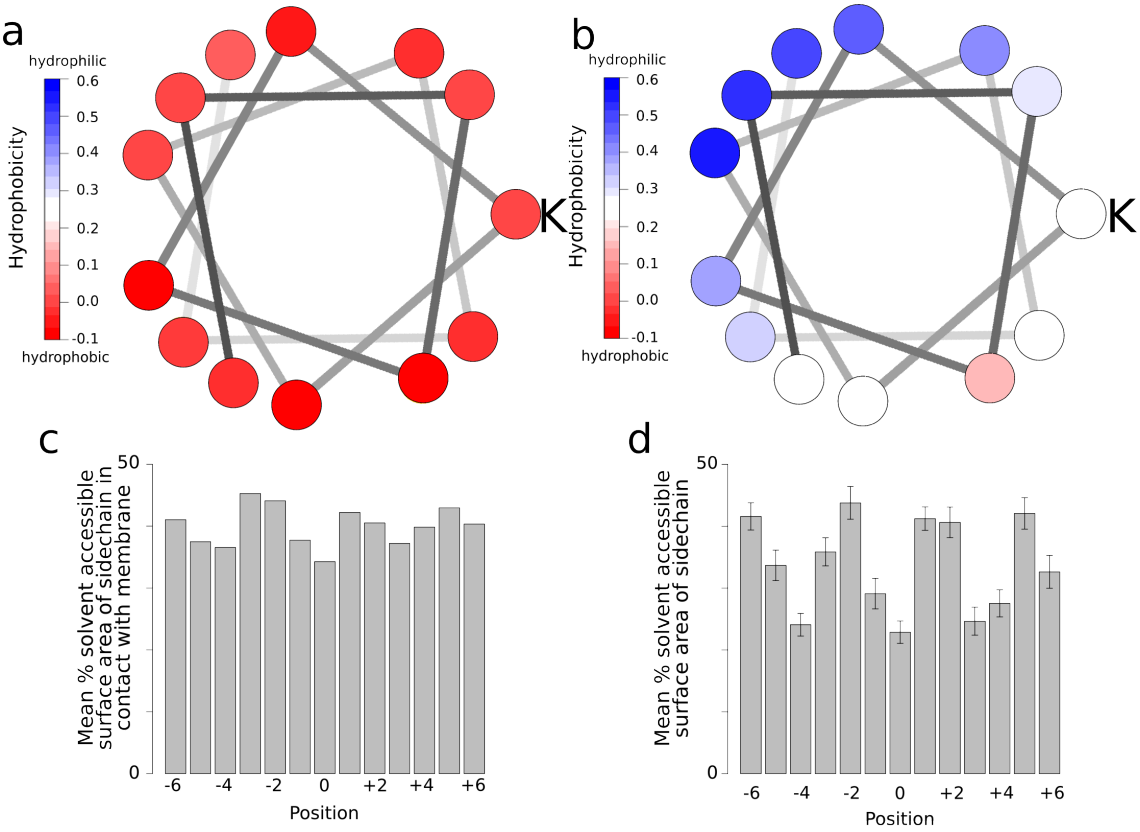


Figure B2: **Hydrophobicities, solvent accessible surface areas, and membrane contacts for kinks identified by MC-Helan.** (a) Standard helical wheel diagram showing the average hydrophobicity of residues around membrane kinks. K indicates the kink residue (position 0 in Figure 4.4). (b) Standard helical wheel diagram for soluble kinks. K indicates the kink residue. (c) Average percentage of residue in contact with the membrane in kinks. (d) Average solvent accessible percentage of residues in soluble kinks. Bars show 2 s.d. from 50 length-matched samples.

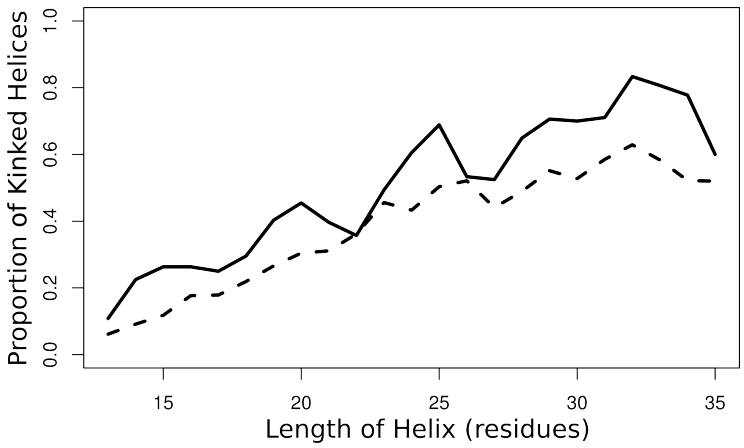


Figure B3: **Proportion of kinked helices, as determined by MC-Helan. The proportion of helices of a given length that are kinked (with an angle over 12°) is similar for membrane (solid line) and soluble (dashed line) proteins. In the same way that the maximum kink angle in each helix (as defined by Kink Finder) increases as the helix gets longer, the probability of a helix being kinked (as defined by MC-Helan) increases as the helix gets longer.**

Table B2: **Table of amino acid motif frequency in kinks and randomly selected parts of straight helices.** Motifs that occur in more than 10% of kinks, or are more than twice as frequent in kinks than straight helices are highlighted in bold. First section: Proline containing motifs. Second section: motifs highlighted by other authors. Third section: aromatic and polar motifs. Fourth section: small residue motifs. Ar = aromatic residue (F, Y or W). This is a version of Table 4.3, but using kinks identified by MC-Helan.

	Membrane			Soluble		
Motif	% Kinks	% Straight Helices	Ratio	% Kinks	% Straight Helices	Ratio
[AVILMFYW]xxxP	**27.2**	0.6	**49.3**	9.5	0.0	**453.4**
xxxxP	**41.9**	1.0	**43.6**	**27.7**	0.1	**344.3**
[ST]P	3.7	0.0	**81.5**	2.4	0.2	**12.8**
[DR]P	0.8	0.2	**3.4**	2.6	0.2	**17.4**
xP	**44.3**	7.0	**6.4**	**29.4**	2.0	**14.9**
xxxP	**42.5**	1.8	**23.9**	**28.1**	0.2	**172.1**
[AVILMFYW]xP	**32.1**	3.1	**10.4**	**19.1**	0.4	**46.5**
GxP	2.9	0.0	**128.0**	0.9	0.0	**36.3**
xxP	**43.8**	3.5	**12.4**	**28.7**	0.7	**38.8**
P[ST]	4.1	0.1	**60.1**	2.5	0.6	**4.5**
Px	**43.3**	11.0	**3.9**	**30.0**	5.5	**5.5**
P	**45.0**	11.3	**4.0**	**30.6**	5.5	**5.6**
[VALT]LWx[AG]YP	0.3	0.0	NA	0.0	0.0	NA
GHPxVY[FI]	0.3	0.0	NA	0.0	0.0	NA
[ST][ST]	**12.6**	9.8	1.3	9.4	7.3	1.3
[ST]xx[ST]	9.9	10.8	0.9	9.1	7.0	1.3
[ST]xxx[ST]	9.3	7.8	1.2	8.7	6.8	1.3
GxxGxxxG	1.1	1.3	0.8	0.1	0.0	**2.1**
GxxxGxxG	1.1	0.6	1.8	0.1	0.0	**5.2**
ArxxxSxxxAr	0.7	0.4	1.9	0.3	0.1	**3.2**
LSAxF	0.3	0.3	0.9	0.1	0.0	**6.0**
WLF[ST]	0.3	0.0	NA	0.0	0.0	NA
ArxxxAr	**17.1**	12.7	1.3	9.0	5.9	1.5
ArxxxArxxAr	2.9	0.2	**13.7**	1.1	0.4	**2.4**
ArxxxArxxxAr	3.6	0.4	**8.0**	1.0	0.4	**2.7**
ArxxArxxxAr	2.1	1.1	2.0	1.0	0.4	**2.6**
[RHDEK]xxx[RHDEK]	8.5	9.5	0.9	**50.6**	58.5	0.9
[ASG]xxx[ASG]	**39.7**	47.0	0.8	**26.3**	33.3	0.8
[STNQ]xxx[STNQ]	**17.6**	18.0	1.0	**25.1**	24.7	1.0
G_0 or G_{+1}	**16.0**	19.8	0.8	**10.9**	6.2	1.7

Table B3: **P values for Kolmogorov-Smirnov tests on the distribution of maximum kink angles in helices of given lengths.** P values ≤ 0.05 indicate that the distributions of angles for soluble and membrane helices (of a given length) are different.

Helix Length (residues)	p value
13	3.9×10^{-07}
14	0.0004
15	0.0002
16	0.003
17	0.0006
18	0.317
19	0.303
20	0.174
21	0.866
22	0.018
23	0.212
24	0.757
25	0.917
26	0.329
27	0.323
28	0.084
29	0.033
30	0.246
31	0.315
32	0.113
33	0.594
34	0.335
35	0.758
36	0.667
37	0.429
38	0.016
39	0.900
40	0.800

Table B4: **Table of membrane proteins in the data set.** Code = PDB Code; C = Chain identifier; St. = Number of first residue in helix; End = Number of last residue in helix; Res. = Resolution (Angstroms); R = R-Factor. NAs in the last two columns indicate the protein structure was solved by a method other that X-ray diffraction.

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
1A91	A	2	38	NA	NA	1A91	A	47	78	NA	NA

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
1AIG	L	33	56	2.600	0.21	1AIG	L	84	111	2.600	0.21
1AIG	L	116	139	2.600	0.21	1AIG	L	171	198	2.600	0.21
1AIG	L	227	249	2.600	0.21	1AR1	B	26	58	2.700	0.21
1AR1	B	76	104	2.700	0.21	1BCC	D	199	227	3.160	0.27
1BCC	E	30	58	3.160	0.27	1BE3	C	34	51	3.000	0.26
1BE3	C	137	149	3.000	0.26	1BE3	C	172	200	3.000	0.26
1BE3	C	225	242	3.000	0.26	1BE3	C	287	307	3.000	0.26
1BE3	C	319	339	3.000	0.26	1BE3	C	345	376	3.000	0.26
1BE3	C	76	102	3.000	0.26	1BE3	C	110	130	3.000	0.26
1BE3	G	36	70	3.000	0.26	1BE3	J	18	46	3.000	0.26
1C17	M	101	125	NA	NA	1C17	M	137	166	NA	NA
1C17	M	202	228	NA	NA	1C17	M	235	263	NA	NA
1C8S	A	10	31	2.000	0.17	1C8S	A	38	61	2.000	0.17
1C8S	A	82	100	2.000	0.17	1C8S	A	202	221	2.000	0.17
1C8S	A	177	191	2.000	0.17	1C8S	A	106	127	2.000	0.17
1C8S	A	133	151	2.000	0.17	1DXR	H	12	35	2.000	0.19
1DXR	L	33	54	2.000	0.19	1DXR	L	84	111	2.000	0.19
1DXR	L	116	139	2.000	0.19	1DXR	L	171	198	2.000	0.19
1DXR	L	228	250	2.000	0.19	1DXR	M	53	76	2.000	0.19
1DXR	M	111	137	2.000	0.19	1DXR	M	143	166	2.000	0.19
1DXR	M	262	284	2.000	0.19	1DXR	M	198	225	2.000	0.19
1DXR	M	241	253	2.000	0.19	1E7P	C	76	97	3.100	0.28
1E7P	C	202	236	3.100	0.28	1E7P	C	22	46	3.100	0.28
1E7P	C	121	149	3.100	0.28	1E7P	C	171	193	3.100	0.28
1EHK	B	6	38	2.400	0.22	1EK9	A	23	36	2.100	0.21
1EK9	A	78	93	2.100	0.21	1EK9	A	229	242	2.100	0.21
1EYS	H	11	33	2.200	0.23	1EYS	L	33	55	2.200	0.23
1EYS	L	92	117	2.200	0.23	1EYS	L	124	147	2.200	0.23
1EYS	L	234	254	2.200	0.23	1EYS	L	179	205	2.200	0.23
1EYS	M	55	76	2.200	0.23	1EYS	M	112	137	2.200	0.23
1EYS	M	144	167	2.200	0.23	1EYS	M	263	283	2.200	0.23
1EYS	M	199	224	2.200	0.23	1EZV	C	32	51	2.300	0.22
1EZV	C	75	102	2.300	0.22	1EZV	C	173	202	2.300	0.22
1EZV	C	226	244	2.300	0.22	1EZV	C	288	308	2.300	0.22
1EZV	C	320	340	2.300	0.22	1EZV	C	348	379	2.300	0.22
1EZV	C	111	134	2.300	0.22	1EZV	C	138	150	2.300	0.22
1EZV	E	51	79	2.300	0.22	1EZV	G	61	81	2.300	0.22
1F88	B	34	63	2.800	0.19	1F88	B	71	98	2.800	0.19
1F88	B	108	137	2.800	0.19	1F88	B	153	166	2.800	0.19
1F88	B	201	223	2.800	0.19	1F88	B	247	276	2.800	0.19
1F88	B	291	308	2.800	0.19	1FX8	A	7	33	2.200	0.20

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
1FX8	A	41	59	2.200	0.20	1FX8	A	145	167	2.200	0.20
1FX8	A	178	193	2.200	0.20	1FX8	A	204	216	2.200	0.20
1FX8	A	234	253	2.200	0.20	1FX8	A	85	105	2.200	0.20
1H2S	B	24	49	1.930	0.23	1H2S	B	53	81	1.930	0.23
1IJD	A	15	35	3.000	0.24	1IJD	B	8	33	3.000	0.24
1J4N	A	6	34	2.200	0.27	1J4N	A	51	69	2.200	0.27
1J4N	A	94	115	2.200	0.27	1J4N	A	141	158	2.200	0.27
1J4N	A	170	184	2.200	0.27	1J4N	A	217	229	2.200	0.27
1JB0	J	13	32	2.500	0.20	1JB0	L	43	64	2.500	0.20
1JB0	L	74	99	2.500	0.20	1JB0	L	114	138	2.500	0.20
1KF6	C	106	128	2.700	0.23	1KF6	C	20	49	2.700	0.23
1KF6	C	67	89	2.700	0.23	1KF6	D	12	41	2.700	0.23
1KF6	D	99	115	2.700	0.23	1KF6	D	61	88	2.700	0.23
1KPL	B	33	65	3.000	0.26	1KPL	B	78	99	3.000	0.26
1KPL	B	128	141	3.000	0.26	1KPL	B	152	165	3.000	0.26
1KPL	B	215	232	3.000	0.26	1KPL	B	330	348	3.000	0.26
1KPL	B	444	456	3.000	0.26	1KPL	B	171	187	3.000	0.26
1KPL	B	253	283	3.000	0.26	1KPL	B	288	305	3.000	0.26
1KPL	B	357	377	3.000	0.26	1KPL	B	421	438	3.000	0.26
1KQF	B	248	277	1.600	0.18	1KQF	C	12	36	1.600	0.18
1KQF	C	50	75	1.600	0.18	1KQF	C	111	133	1.600	0.18
1KQF	C	145	174	1.600	0.18	1KZU	B	5	35	2.500	0.23
1L7V	A	3	32	3.200	0.26	1L7V	A	56	81	3.200	0.26
1L7V	A	93	106	3.200	0.26	1L7V	A	228	250	3.200	0.26
1L7V	A	276	295	3.200	0.26	1L7V	A	305	323	3.200	0.26
1L7V	A	114	138	3.200	0.26	1L7V	A	142	167	3.200	0.26
1L7V	A	190	213	3.200	0.26	1LGH	A	19	38	2.400	0.21
1LGH	B	10	41	2.400	0.21	1M56	A	26	55	2.300	0.24
1M56	A	92	115	2.300	0.24	1M56	A	136	154	2.300	0.24
1M56	A	186	213	2.300	0.24	1M56	A	227	257	2.300	0.24
1M56	A	273	305	2.300	0.24	1M56	A	313	326	2.300	0.24
1M56	A	342	370	2.300	0.24	1M56	A	450	476	2.300	0.24
1M56	A	491	520	2.300	0.24	1M56	A	381	400	2.300	0.24
1M56	A	414	444	2.300	0.24	1M56	B	49	82	2.300	0.24
1M56	B	98	127	2.300	0.24	1M56	C	17	36	2.300	0.24
1M56	C	42	67	2.300	0.24	1M56	C	73	106	2.300	0.24
1M56	C	197	228	2.300	0.24	1M56	C	238	259	2.300	0.24
1M56	C	134	155	2.300	0.24	1M56	C	161	187	2.300	0.24
1M56	D	19	49	2.300	0.24	1MM4	A	7	19	NA	NA
1N7L	A	25	50	NA	NA	1NEK	C	103	128	2.600	0.25
1NEK	C	22	51	2.600	0.25	1NEK	C	69	96	2.600	0.25

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
1NKZ	A	15	35	2.000	0.17	1OCC	A	12	41	2.800	0.20
1OCC	A	52	87	2.800	0.20	1OCC	A	95	115	2.800	0.20
1OCC	A	142	169	2.800	0.20	1OCC	A	184	212	2.800	0.20
1OCC	A	229	261	2.800	0.20	1OCC	A	270	283	2.800	0.20
1OCC	A	299	327	2.800	0.20	1OCC	A	407	433	2.800	0.20
1OCC	A	448	477	2.800	0.20	1OCC	A	336	359	2.800	0.20
1OCC	A	372	401	2.800	0.20	1OCC	B	15	45	2.800	0.20
1OCC	B	61	87	2.800	0.20	1OCC	C	73	105	2.800	0.20
1OCC	C	193	223	2.800	0.20	1OCC	C	234	254	2.800	0.20
1OCC	C	16	34	2.800	0.20	1OCC	C	40	65	2.800	0.20
1OCC	C	129	152	2.800	0.20	1OCC	C	156	183	2.800	0.20
1OCC	D	77	102	2.800	0.20	1OCC	G	13	37	2.800	0.20
1OCC	I	12	52	2.800	0.20	1OCC	K	13	34	2.800	0.20
1OCC	L	18	44	2.800	0.20	1OCC	M	12	41	2.800	0.20
1OCR	J	26	54	2.350	0.20	1ORQ	C	29	50	3.200	0.25
1ORQ	C	60	78	3.200	0.25	1ORQ	C	147	170	3.200	0.25
1ORQ	C	207	236	3.200	0.25	1PV6	A	7	35	3.500	0.29
1PV6	A	43	70	3.500	0.29	1PV6	A	168	185	3.500	0.29
1PV6	A	221	247	3.500	0.29	1PV6	A	254	284	3.500	0.29
1PV6	A	290	306	3.500	0.29	1PV6	A	312	338	3.500	0.29
1PV6	A	142	157	3.500	0.29	1PV6	A	345	374	3.500	0.29
1PV6	A	378	398	3.500	0.29	1PW4	A	20	52	3.300	0.30
1PW4	A	122	146	3.300	0.30	1PW4	A	154	178	3.300	0.30
1PW4	A	191	206	3.300	0.30	1PW4	A	253	276	3.300	0.30
1PW4	A	290	315	3.300	0.30	1PW4	A	321	338	3.300	0.30
1PW4	A	351	372	3.300	0.30	1PW4	A	382	408	3.300	0.30
1PW4	A	415	443	3.300	0.30	1PW4	A	65	78	3.300	0.30
1PW4	A	95	109	3.300	0.30	1Q90	A	249	280	3.100	0.22
1QLE	C	17	34	3.000	0.23	1QLE	C	54	74	3.000	0.23
1QLE	C	80	113	3.000	0.23	1QLE	C	206	232	3.000	0.23
1QLE	C	247	266	3.000	0.23	1QLE	C	141	163	3.000	0.23
1QLE	C	169	194	3.000	0.23	1QLE	D	17	46	3.000	0.23
1RC2	A	2	24	2.500	0.23	1RC2	A	36	53	2.500	0.23
1RC2	A	81	103	2.500	0.23	1RC2	A	131	151	2.500	0.23
1RC2	A	162	176	2.500	0.23	1RC2	A	200	224	2.500	0.23
1RH5	C	30	47	3.200	0.24	1RHZ	A	23	40	3.500	0.25
1RHZ	A	76	88	3.500	0.25	1RHZ	A	104	127	3.500	0.25
1RHZ	A	137	163	3.500	0.25	1RHZ	A	170	183	3.500	0.25
1RHZ	A	210	228	3.500	0.25	1RHZ	A	256	275	3.500	0.25
1RHZ	A	313	332	3.500	0.25	1RHZ	A	367	394	3.500	0.25
1RHZ	A	401	422	3.500	0.25	1RHZ	B	3	21	3.500	0.25

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
1RHZ	B	31	64	3.500	0.25	1S5L	x	2015	2043	3.500	0.30
1SQQ	K	8	35	3.000	0.23	1UYN	X	788	810	2.600	0.23
1VF5	D	23	42	3.000	0.26	1VRY	A	2	14	NA	NA
1VRY	A	34	51	NA	NA	1WP1	A	66	82	2.560	0.26
1WP1	A	131	147	2.560	0.26	1WP1	A	278	294	2.560	0.26
1WP1	A	339	355	2.560	0.26	1WRG	A	12	41	NA	NA
1XIO	A	3	26	2.000	0.23	1XIO	A	34	56	2.000	0.23
1XIO	A	71	91	2.000	0.23	1XIO	A	159	185	2.000	0.23
1XIO	A	195	223	2.000	0.23	1XIO	A	99	121	2.000	0.23
1XIO	A	125	153	2.000	0.23	1XL6	A	47	69	2.850	0.27
1XL6	A	108	135	2.850	0.27	1XME	A	11	45	2.300	0.22
1XME	A	65	97	2.300	0.22	1XME	A	105	125	2.300	0.22
1XME	A	143	172	2.300	0.22	1XME	A	181	212	2.300	0.22
1XME	A	221	255	2.300	0.22	1XME	A	262	274	2.300	0.22
1XME	A	292	326	2.300	0.22	1XME	A	347	368	2.300	0.22
1XME	A	380	408	2.300	0.22	1XME	A	415	444	2.300	0.22
1XME	A	463	492	2.300	0.22	1XME	A	527	551	2.300	0.22
1XRD	A	10	41	NA	NA	1YCE	a	3	44	2.400	0.20
1YCE	a	50	79	2.400	0.20	1YEW	A	185	207	2.800	0.27
1YEW	A	232	256	2.800	0.27	1YEW	B	14	43	2.800	0.27
1YEW	B	64	79	2.800	0.27	1YEW	B	90	104	2.800	0.27
1YEW	B	141	164	2.800	0.27	1YGM	A	33	54	NA	NA
1YQ3	D	5	29	2.200	0.17	1YQ3	D	33	58	2.200	0.17
1YQ3	D	62	86	2.200	0.17	1YST	H	12	34	3.000	0.23
1Z98	A	34	63	2.100	0.18	1Z98	A	74	92	2.100	0.18
1Z98	A	161	181	2.100	0.18	1Z98	A	199	213	2.100	0.18
1Z98	A	116	139	2.100	0.18	1Z98	A	236	262	2.100	0.18
1ZCD	A	13	29	3.450	0.30	1ZCD	A	59	83	3.450	0.30
1ZCD	A	156	174	3.450	0.30	1ZCD	A	181	199	3.450	0.30
1ZCD	A	223	236	3.450	0.30	1ZCD	A	247	271	3.450	0.30
1ZCD	A	358	370	3.450	0.30	1ZOY	D	98	121	2.400	0.23
1ZOY	D	38	62	2.400	0.23	1ZOY	D	66	91	2.400	0.23
1ZRT	D	221	249	3.500	0.30	1ZZA	A	13	34	NA	NA
1ZZA	A	64	81	NA	NA	2AXT	a	5036	5053	3.000	0.23
2AXT	a	5143	5165	3.000	0.23	2AXT	a	5269	5294	3.000	0.23
2AXT	a	5111	5136	3.000	0.23	2AXT	a	5196	5221	3.000	0.23
2AXT	z	5003	5028	3.000	0.23	2AXT	z	5037	5061	3.000	0.23
2B2F	A	3	27	1.720	0.18	2B2F	A	85	104	1.720	0.18
2B2F	A	187	208	1.720	0.18	2B2F	A	214	241	1.720	0.18
2B2F	A	246	260	1.720	0.18	2B2F	A	269	292	1.720	0.18
2B2F	A	336	366	1.720	0.18	2B2F	A	35	53	1.720	0.18

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
2B2F	A	114	136	1.720	0.18	2B6P	A	10	31	2.400	0.24
2B6P	A	39	59	2.400	0.24	2B6P	A	83	107	2.400	0.24
2B6P	A	127	148	2.400	0.24	2B6P	A	160	174	2.400	0.24
2BG9	A	406	432	4.000	1.00	2BG9	A	219	237	4.000	1.00
2BG9	A	247	266	4.000	1.00	2BG9	B	223	239	4.000	1.00
2BG9	B	250	274	4.000	1.00	2BG9	B	288	301	4.000	1.00
2BG9	B	433	465	4.000	1.00	2BG9	C	291	314	4.000	1.00
2BG9	C	453	477	4.000	1.00	2BG9	C	233	250	4.000	1.00
2BG9	C	257	280	4.000	1.00	2BG9	E	285	305	4.000	1.00
2BG9	E	454	471	4.000	1.00	2BG9	E	222	236	4.000	1.00
2BG9	E	253	277	4.000	1.00	2BHW	A	55	86	2.500	0.22
2BHW	A	124	143	2.500	0.22	2BHW	A	170	200	2.500	0.22
2BL2	A	12	45	2.100	0.19	2BL2	A	52	78	2.100	0.19
2BL2	A	86	122	2.100	0.19	2BL2	A	128	155	2.100	0.19
2C3E	A	8	37	2.800	0.25	2C3E	A	210	238	2.800	0.25
2C3E	A	275	290	2.800	0.25	2C3E	A	74	98	2.800	0.25
2C3E	A	108	126	2.800	0.25	2C3E	A	169	200	2.800	0.25
2E74	A	35	55	3.000	0.23	2E74	A	79	105	3.000	0.23
2E74	A	114	135	3.000	0.23	2E74	A	177	209	3.000	0.23
2E74	B	39	57	3.000	0.23	2E74	B	94	116	3.000	0.23
2E74	B	127	145	3.000	0.23	2EVU	A	4	33	2.300	0.19
2EVU	A	52	72	2.300	0.19	2EVU	A	99	122	2.300	0.19
2EVU	A	143	164	2.300	0.19	2EVU	A	176	188	2.300	0.19
2EVU	A	200	212	2.300	0.19	2EVU	A	227	243	2.300	0.19
2EZ0	A	33	63	3.540	0.26	2EZ0	A	75	97	3.540	0.26
2EZ0	A	152	165	3.540	0.26	2EZ0	A	215	231	3.540	0.26
2EZ0	A	253	282	3.540	0.26	2EZ0	A	290	305	3.540	0.26
2EZ0	A	330	348	3.540	0.26	2EZ0	A	359	378	3.540	0.26
2EZ0	A	387	399	3.540	0.26	2EZ0	A	421	434	3.540	0.26
2EZ0	A	171	189	3.540	0.26	2FBW	C	34	62	2.100	0.19
2FBW	C	81	109	2.100	0.19	2FBW	C	116	138	2.100	0.19
2FYN	A	46	66	3.200	0.22	2FYN	A	91	118	3.200	0.22
2FYN	A	126	148	3.200	0.22	2FYN	A	189	217	3.200	0.22
2FYN	A	328	345	3.200	0.22	2FYN	A	363	381	3.200	0.22
2FYN	A	390	412	3.200	0.22	2FYN	A	251	269	3.200	0.22
2FYN	B	222	251	3.200	0.22	2FYN	C	12	36	3.200	0.22
2GFP	A	10	33	3.500	0.28	2GFP	A	43	56	3.500	0.28
2GFP	A	208	226	3.500	0.28	2GFP	A	242	257	3.500	0.28
2GFP	A	270	282	3.500	0.28	2GFP	A	294	312	3.500	0.28
2GFP	A	321	343	3.500	0.28	2GFP	A	357	377	3.500	0.28
2GFP	A	100	126	3.500	0.28	2GFP	A	134	150	3.500	0.28

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
2GFP	A	160	175	3.500	0.28	2HYD	A	278	317	3.000	0.26
2HYD	A	12	41	3.000	0.26	2HYD	A	52	106	3.000	0.26
2HYD	A	111	159	3.000	0.26	2HYD	A	164	209	3.000	0.26
2HYD	A	218	272	3.000	0.26	2IC8	A	201	217	2.100	0.24
2IC8	A	229	241	2.100	0.24	2IC8	A	251	268	2.100	0.24
2IC8	A	95	112	2.100	0.24	2IC8	A	152	167	2.100	0.24
2IC8	A	171	192	2.100	0.24	2IUB	F	278	307	2.900	0.26
2IUB	F	328	342	2.900	0.26	2JAF	A	28	50	1.700	0.24
2JAF	A	58	81	1.700	0.24	2JAF	A	106	125	1.700	0.24
2JAF	A	131	153	1.700	0.24	2JAF	A	158	188	1.700	0.24
2JAF	A	193	216	1.700	0.24	2JAF	A	227	252	1.700	0.24
2K37	A	5	26	NA	NA	2K73	A	42	61	NA	NA
2K73	A	69	96	NA	NA	2K73	A	142	162	NA	NA
2K73	A	12	35	NA	NA	2K9P	A	39	72	NA	NA
2K9P	A	80	103	NA	NA	2K9Y	A	537	553	NA	NA
2KNC	B	701	739	NA	NA	2KS1	B	148	169	NA	NA
2KS9	A	29	58	NA	NA	2KS9	A	66	95	NA	NA
2KS9	A	102	135	NA	NA	2KS9	A	144	162	NA	NA
2KS9	A	193	220	NA	NA	2KS9	A	239	273	NA	NA
2KS9	A	285	307	NA	NA	2KSD	A	28	45	NA	NA
2KSE	A	15	33	NA	NA	2KSE	A	160	180	NA	NA
2KSY	A	5	27	NA	NA	2KSY	A	34	54	NA	NA
2KSY	A	123	148	NA	NA	2KSY	A	155	179	NA	NA
2KSY	A	70	91	NA	NA	2KSY	A	95	114	NA	NA
2L0J	A	25	44	NA	NA	2L2T	A	51	76	NA	NA
2L35	A	9	29	NA	NA	2L35	A	39	55	NA	NA
2L6X	A	27	51	NA	NA	2L6X	A	60	85	NA	NA
2L6X	A	91	113	NA	NA	2L6X	A	122	139	NA	NA
2L6X	A	180	204	NA	NA	2L9U	A	641	671	NA	NA
2LCK	A	16	40	NA	NA	2LCK	A	214	237	NA	NA
2LCK	A	267	295	NA	NA	2LCK	A	83	107	NA	NA
2LCK	A	116	131	NA	NA	2LCK	A	172	196	NA	NA
2LKG	A	15	41	NA	NA	2LKG	A	52	69	NA	NA
2LKG	A	79	97	NA	NA	2LLY	A	8	30	NA	NA
2LLY	A	37	58	NA	NA	2LLY	A	71	92	NA	NA
2LLY	A	110	129	NA	NA	2LNL	A	38	66	NA	NA
2LNL	A	75	98	NA	NA	2LNL	A	112	136	NA	NA
2LNL	A	150	173	NA	NA	2LNL	A	207	225	NA	NA
2LNL	A	239	267	NA	NA	2LNL	A	279	307	NA	NA
2LOM	A	30	46	NA	NA	2LOM	A	65	81	NA	NA
2LOR	A	65	95	NA	NA	2LOR	A	19	52	NA	NA

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
2NQ2	A	7	24	2.400	0.22	2NQ2	A	53	86	2.400	0.22
2NQ2	A	99	112	2.400	0.22	2NQ2	A	235	256	2.400	0.22
2NQ2	A	311	327	2.400	0.22	2NQ2	A	117	139	2.400	0.22
2NQ2	A	147	170	2.400	0.22	2NQ2	A	194	214	2.400	0.22
2NQ2	A	282	303	2.400	0.22	2NR9	A	116	131	2.200	0.24
2NR9	A	166	191	2.200	0.24	2NR9	A	10	26	2.200	0.24
2NR9	A	63	82	2.200	0.24	2NR9	A	86	108	2.200	0.24
2OAR	A	69	99	3.500	0.32	2OAR	A	13	44	3.500	0.32
2ONK	C	2	29	3.100	0.26	2ONK	C	35	77	3.100	0.26
2ONK	C	169	198	3.100	0.26	2ONK	C	130	152	3.100	0.26
2ONK	C	226	249	3.100	0.26	2Q6H	A	166	183	1.850	0.20
2Q6H	A	191	214	1.850	0.20	2Q6H	A	241	255	1.850	0.20
2Q6H	A	276	290	1.850	0.20	2Q6H	A	337	368	1.850	0.20
2Q6H	A	449	477	1.850	0.20	2Q6H	A	483	509	1.850	0.20
2Q6H	A	11	36	1.850	0.20	2Q6H	A	44	71	1.850	0.20
2Q6H	A	88	123	1.850	0.20	2Q6H	A	375	390	1.850	0.20
2Q6H	A	399	423	1.850	0.20	2Q7M	C	3	35	4.250	0.24
2Q7M	C	117	138	4.250	0.24	2Q7M	C	50	73	4.250	0.24
2Q7M	C	81	100	4.250	0.24	2QDZ	A	10	26	3.150	0.29
2QFI	A	129	141	3.800	0.32	2QKS	A	103	132	2.200	0.23
2QKS	A	42	65	2.200	0.23	2QTS	E	57	69	1.900	0.21
2QTS	E	427	460	1.900	0.21	2R6G	F	69	90	2.800	0.24
2R6G	F	277	306	2.800	0.24	2R6G	F	484	503	2.800	0.24
2R6G	F	18	35	2.800	0.24	2R6G	F	40	55	2.800	0.24
2R6G	F	365	388	2.800	0.24	2R6G	F	426	447	2.800	0.24
2R9R	H	221	243	2.400	0.21	2R9R	H	254	274	2.400	0.21
2R9R	H	381	416	2.400	0.21	2R9R	H	160	183	2.400	0.21
2R9R	H	279	296	2.400	0.21	2R9R	H	321	346	2.400	0.21
2RH1	A	103	136	2.400	0.20	2RH1	A	147	169	2.400	0.20
2RH1	A	197	229	2.400	0.20	2RH1	A	267	298	2.400	0.20
2RH1	A	31	59	2.400	0.20	2RH1	A	67	95	2.400	0.20
2RH1	A	305	326	2.400	0.20	2UUI	A	5	32	2.000	0.20
2UUI	A	105	141	2.000	0.20	2UUI	A	44	72	2.000	0.20
2UUI	A	76	98	2.000	0.20	2VV5	A	29	58	3.450	0.29
2VV5	A	63	84	3.450	0.29	2VV5	A	111	125	3.450	0.29
2W1P	A	41	69	1.400	0.16	2W1P	A	82	103	1.400	0.16
2W1P	A	127	151	1.400	0.16	2W1P	A	168	188	1.400	0.16
2W1P	A	200	215	1.400	0.16	2W1P	A	247	265	1.400	0.16
2WIE	A	5	41	2.130	0.20	2WIE	A	46	76	2.130	0.20
2WLL	B	121	149	3.650	0.25	2WLL	B	60	83	3.650	0.25
2WPD	J	7	36	3.430	0.29	2WPD	J	47	73	3.430	0.29

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
2WSC	4	42	55	3.300	0.36	2WSW	A	128	161	2.290	0.21
2WSW	A	312	337	2.290	0.21	2WSW	A	344	375	2.290	0.21
2WSW	A	404	434	2.290	0.21	2WSW	A	16	30	2.290	0.21
2WSW	A	103	117	2.290	0.21	2WSW	A	187	222	2.290	0.21
2WSW	A	229	247	2.290	0.21	2WSW	A	446	465	2.290	0.21
2WSW	A	469	499	2.290	0.21	2WSW	A	33	48	2.290	0.21
2WSW	A	52	71	2.290	0.21	2WSW	A	252	275	2.290	0.21
2WSW	A	279	292	2.290	0.21	2X2V	A	3	33	2.500	0.19
2X2V	A	38	68	2.500	0.19	2X79	A	60	86	3.800	0.28
2X79	A	105	133	3.800	0.28	2X79	A	210	231	3.800	0.28
2X79	A	244	273	3.800	0.28	2X79	A	307	328	3.800	0.28
2X79	A	336	356	3.800	0.28	2X79	A	428	448	3.800	0.28
2X79	A	143	157	3.800	0.28	2X79	A	163	189	3.800	0.28
2XND	J	18	34	3.500	0.26	2XND	J	51	70	3.500	0.26
2XQ2	A	11	28	2.730	0.25	2XQ2	A	280	313	2.730	0.25
2XQ2	A	453	472	2.730	0.25	2XQ2	A	479	501	2.730	0.25
2XQ2	A	522	544	2.730	0.25	2XQ2	A	549	572	2.730	0.25
2XQ2	A	85	108	2.730	0.25	2XQ2	A	163	177	2.730	0.25
2XQ2	A	187	209	2.730	0.25	2XQ2	A	349	378	2.730	0.25
2XQ2	A	392	413	2.730	0.25	2XQ2	A	423	447	2.730	0.25
2XQ2	A	124	156	2.730	0.25	2YEV	A	64	99	2.360	0.17
2YEV	A	107	125	2.360	0.17	2YEV	A	153	179	2.360	0.17
2YEV	A	193	224	2.360	0.17	2YEV	A	239	272	2.360	0.17
2YEV	A	279	292	2.360	0.17	2YEV	A	308	336	2.360	0.17
2YEV	A	416	442	2.360	0.17	2YEV	A	459	489	2.360	0.17
2YEV	A	558	573	2.360	0.17	2YEV	A	580	597	2.360	0.17
2YEV	A	617	643	2.360	0.17	2YEV	A	724	756	2.360	0.17
2YEV	A	765	787	2.360	0.17	2YEV	A	25	52	2.360	0.17
2YEV	A	345	368	2.360	0.17	2YEV	A	381	409	2.360	0.17
2YEV	A	657	681	2.360	0.17	2YEV	A	686	715	2.360	0.17
2YEV	B	30	62	2.360	0.17	2YEV	B	81	109	2.360	0.17
2YIU	B	248	277	2.700	0.24	2YIU	C	19	39	2.700	0.24
2YN6	A	261	285	3.310	0.23	2YN6	A	202	220	3.310	0.23
2YN6	A	228	252	3.310	0.23	2YVX	A	278	303	3.500	0.29
2YVX	A	323	343	3.500	0.29	2YVX	A	352	374	3.500	0.29
2YVX	A	389	414	3.500	0.29	2YVX	A	429	443	3.500	0.29
2ZJS	Y	151	177	3.200	0.25	2ZJS	Y	183	205	3.200	0.25
2ZJS	Y	215	236	3.200	0.25	2ZJS	Y	272	288	3.200	0.25
2ZJS	Y	309	327	3.200	0.25	2ZJS	Y	356	388	3.200	0.25
2ZJS	Y	401	417	3.200	0.25	2ZJS	Y	14	31	3.200	0.25
2ZJS	Y	105	132	3.200	0.25	2ZJS	Y	78	92	3.200	0.25

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
2ZT9	C	252	283	3.000	0.23	2ZW3	A	23	42	3.500	0.34
2ZW3	A	73	105	3.500	0.34	2ZW3	A	132	156	3.500	0.34
2ZW3	A	185	215	3.500	0.34	2ZXE	G	22	38	2.400	0.25
3AM6	A	77	92	3.200	0.29	3AM6	A	101	122	3.200	0.29
3AM6	A	129	150	3.200	0.29	3AM6	A	161	187	3.200	0.29
3AM6	A	196	226	3.200	0.29	3AM6	A	7	30	3.200	0.29
3AM6	A	39	61	3.200	0.29	3AON	A	31	69	2.000	0.20
3AON	A	127	162	2.000	0.20	3AQP	A	4	23	3.300	0.30
3AQP	A	326	353	3.300	0.30	3AQP	A	364	390	3.300	0.30
3AQP	A	395	428	3.300	0.30	3AQP	A	449	468	3.300	0.30
3AQP	A	621	650	3.300	0.30	3AQP	A	265	292	3.300	0.30
3AQP	A	296	319	3.300	0.30	3AQP	A	565	586	3.300	0.30
3AQP	A	594	615	3.300	0.30	3AQP	A	658	685	3.300	0.30
3AQP	A	692	725	3.300	0.30	3ARC	j	10	31	1.900	0.17
3AYF	A	20	46	2.500	0.24	3AYF	A	248	274	2.500	0.24
3AYF	A	300	329	2.500	0.24	3AYF	A	349	379	2.500	0.24
3AYF	A	386	413	2.500	0.24	3AYF	A	435	462	2.500	0.24
3AYF	A	469	482	2.500	0.24	3AYF	A	496	529	2.500	0.24
3AYF	A	535	548	2.500	0.24	3AYF	A	572	601	2.500	0.24
3AYF	A	609	639	2.500	0.24	3AYF	A	647	673	2.500	0.24
3AYF	A	683	716	2.500	0.24	3AYF	A	735	763	2.500	0.24
3B4R	A	19	31	3.300	0.25	3B4R	A	40	63	3.300	0.25
3B4R	A	96	112	3.300	0.25	3B4R	A	125	138	3.300	0.25
3B4R	A	163	185	3.300	0.25	3B4R	A	192	217	3.300	0.25
3B8C	A	242	256	3.600	0.35	3B8C	A	630	650	3.600	0.35
3B8E	A	123	137	3.500	0.28	3B8E	A	276	289	3.500	0.28
3B8E	A	310	322	3.500	0.28	3B8E	A	755	787	3.500	0.28
3B8E	A	851	866	3.500	0.28	3B8E	A	913	928	3.500	0.28
3B8E	A	945	963	3.500	0.28	3B8E	A	987	1003	3.500	0.28
3B9Y	A	34	59	1.850	0.16	3B9Y	A	66	89	1.850	0.16
3B9Y	A	102	120	1.850	0.16	3B9Y	A	130	148	1.850	0.16
3B9Y	A	198	220	1.850	0.16	3B9Y	A	227	253	1.850	0.16
3B9Y	A	260	274	1.850	0.16	3B9Y	A	283	310	1.850	0.16
3B9Y	A	325	338	1.850	0.16	3B9Y	A	343	369	1.850	0.16
3B9Y	A	170	182	1.850	0.16	3C02	A	9	35	2.050	0.18
3C02	A	43	61	2.050	0.18	3C02	A	134	156	2.050	0.18
3C02	A	166	183	2.050	0.18	3C02	A	87	107	2.050	0.18
3C02	A	221	248	2.050	0.18	3C1G	A	97	119	2.300	0.19
3C1G	A	125	147	2.300	0.19	3C1G	A	168	180	2.300	0.19
3C1G	A	199	219	2.300	0.19	3C1G	A	225	252	2.300	0.19
3C1G	A	258	272	2.300	0.19	3C1G	A	348	380	2.300	0.19

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
3C1G	A	7	32	2.300	0.19	3C1G	A	282	300	2.300	0.19
3C1G	A	313	330	2.300	0.19	3C1G	A	40	58	2.300	0.19
3CHX	A	239	259	3.900	0.34	3CHX	B	24	49	3.900	0.34
3CHX	B	146	166	3.900	0.34	3CHX	C	69	85	3.900	0.34
3CHX	C	100	122	3.900	0.34	3CHX	C	144	163	3.900	0.34
3CWB	E	32	58	3.510	0.28	3CWB	G	33	68	3.510	0.28
3CX5	D	263	296	1.900	0.24	3CX5	I	18	43	1.900	0.24
3D31	C	242	255	3.000	0.25	3D31	C	14	41	3.000	0.25
3D31	C	140	165	3.000	0.25	3D31	C	193	216	3.000	0.25
3D31	C	56	86	3.000	0.25	3D9S	A	8	32	2.000	0.16
3D9S	A	84	108	2.000	0.16	3D9S	A	128	150	2.000	0.16
3D9S	A	161	174	2.000	0.16	3D9S	A	204	221	2.000	0.16
3D9S	A	42	56	2.000	0.16	3DDL	A	91	110	1.900	0.25
3DDL	A	119	141	1.900	0.25	3DDL	A	151	171	1.900	0.25
3DDL	A	183	208	1.900	0.25	3DDL	A	222	257	1.900	0.25
3DDL	A	10	38	1.900	0.25	3DDL	A	46	72	1.900	0.25
3DIN	C	82	95	4.500	0.28	3DIN	C	111	139	4.500	0.28
3DIN	C	159	178	4.500	0.28	3DIN	C	186	198	4.500	0.28
3DIN	C	214	228	4.500	0.28	3DIN	C	274	286	4.500	0.28
3DIN	C	304	321	4.500	0.28	3DIN	C	350	380	4.500	0.28
3DIN	C	392	412	4.500	0.28	3DIN	D	32	57	4.500	0.28
3DIN	E	15	30	4.500	0.28	3DIN	E	55	70	4.500	0.28
3EAM	A	201	213	2.900	0.20	3EAM	A	221	244	2.900	0.20
3EAM	A	254	281	2.900	0.20	3EAM	A	286	314	2.900	0.20
3EGW	C	3	30	1.900	0.17	3EGW	C	50	68	1.900	0.17
3EGW	C	182	197	1.900	0.17	3EGW	C	154	167	1.900	0.17
3EGW	C	124	147	1.900	0.17	3EGW	C	83	111	1.900	0.17
3EML	A	74	106	2.600	0.20	3EML	A	118	138	2.600	0.20
3EML	A	174	204	2.600	0.20	3EML	A	222	258	2.600	0.20
3EML	A	7	33	2.600	0.20	3EML	A	43	66	2.600	0.20
3EML	A	267	290	2.600	0.20	3EMO	A	1019	1035	3.000	0.22
3GI9	C	10	25	2.480	0.25	3GI9	C	122	138	2.480	0.25
3GI9	C	144	166	2.480	0.25	3GI9	C	40	62	2.480	0.25
3GI9	C	85	112	2.480	0.25	3GI9	C	184	211	2.480	0.25
3GI9	C	221	244	2.480	0.25	3GI9	C	271	304	2.480	0.25
3GI9	C	322	336	2.480	0.25	3GI9	C	340	365	2.480	0.25
3GI9	C	373	395	2.480	0.25	3GI9	C	399	422	2.480	0.25
3HD6	A	10	29	2.100	0.17	3HD6	A	59	77	2.100	0.17
3HD6	A	85	109	2.100	0.17	3HD6	A	122	140	2.100	0.17
3HD6	A	148	169	2.100	0.17	3HD6	A	215	231	2.100	0.17
3HD6	A	243	270	2.100	0.17	3HD6	A	280	292	2.100	0.17

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
3HD6	A	303	330	2.100	0.17	3HD6	A	345	359	2.100	0.17
3HD6	A	385	416	2.100	0.17	3HD7	F	226	272	3.400	0.24
3HZQ	A	15	45	3.820	0.29	3HZQ	A	65	87	3.820	0.29
3KBC	A	13	29	3.510	0.27	3KBC	A	37	71	3.510	0.27
3KBC	A	83	106	3.510	0.27	3KBC	A	151	169	3.510	0.27
3KBC	A	175	218	3.510	0.27	3KBC	A	225	251	3.510	0.27
3KBC	A	261	275	3.510	0.27	3KBC	A	314	329	3.510	0.27
3KBC	A	377	414	3.510	0.27	3KLY	C	28	54	2.100	0.18
3KLY	C	63	83	2.100	0.18	3KLY	C	107	134	2.100	0.18
3KLY	C	160	181	2.100	0.18	3KLY	C	187	202	2.100	0.18
3KLY	C	246	277	2.100	0.18	3KLY	C	209	224	2.100	0.18
3KP9	A	21	43	3.600	0.26	3KP9	A	72	88	3.600	0.26
3KP9	A	103	123	3.600	0.26	3KP9	A	131	147	3.600	0.26
3KP9	A	157	179	3.600	0.26	3KVN	A	314	337	2.500	0.21
3KZI	F	19	40	3.600	0.30	3M6E	A	10	24	2.650	0.20
3M6E	A	99	120	2.650	0.20	3M6E	A	261	286	2.650	0.20
3M6E	A	295	308	2.650	0.20	3M6E	A	209	228	2.650	0.20
3M6E	A	46	64	2.650	0.20	3MC9	B	335	351	2.200	0.20
3MP7	A	30	45	2.900	0.28	3MP7	A	116	130	2.900	0.28
3MP7	A	147	168	2.900	0.28	3MP7	A	176	193	2.900	0.28
3MP7	A	241	259	2.900	0.28	3MP7	A	287	307	2.900	0.28
3MP7	A	345	368	2.900	0.28	3MP7	A	410	427	2.900	0.28
3MP7	A	434	455	2.900	0.28	3MP7	B	29	58	2.900	0.28
3NE2	C	5	34	3.000	0.25	3NE2	C	53	73	3.000	0.25
3NE2	C	100	123	3.000	0.25	3NE2	C	146	164	3.000	0.25
3NE2	C	176	189	3.000	0.25	3NE2	C	201	213	3.000	0.25
3NE2	C	222	243	3.000	0.25	3O7Q	A	25	48	3.140	0.22
3O7Q	A	151	171	3.140	0.22	3O7Q	A	196	228	3.140	0.22
3O7Q	A	259	287	3.140	0.22	3O7Q	A	117	144	3.140	0.22
3O7Q	A	62	86	3.140	0.22	3O7Q	A	90	112	3.140	0.22
3O7Q	A	294	319	3.140	0.22	3O7Q	A	324	343	3.140	0.22
3O7Q	A	349	373	3.140	0.22	3O7Q	A	380	402	3.140	0.22
3O7Q	A	415	429	3.140	0.22	3OB6	B	42	66	3.000	0.24
3OB6	B	275	306	3.000	0.24	3OB6	B	324	339	3.000	0.24
3OB6	B	83	111	3.000	0.24	3OB6	B	122	141	3.000	0.24
3OB6	B	225	246	3.000	0.24	3OB6	B	352	374	3.000	0.24
3OB6	B	146	159	3.000	0.24	3OB6	B	383	403	3.000	0.24
3OB6	B	407	426	3.000	0.24	3ORG	A	278	292	3.500	0.26
3ORG	A	321	350	3.500	0.26	3ORG	A	359	371	3.500	0.26
3ORG	A	430	451	3.500	0.26	3ORG	A	460	473	3.500	0.26
3ORG	A	90	129	3.500	0.26	3ORG	A	135	158	3.500	0.26

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
3ORG	A	210	225	3.500	0.26	3ORG	A	237	255	3.500	0.26
3P5N	A	11	27	3.600	0.27	3P5N	A	60	77	3.600	0.27
3P5N	A	84	100	3.600	0.27	3P5N	A	110	130	3.600	0.27
3P5N	A	154	182	3.600	0.27	3PBL	A	35	56	2.890	0.24
3PBL	A	63	91	2.890	0.24	3PBL	A	148	166	2.890	0.24
3PBL	A	186	215	2.890	0.24	3PBL	A	322	352	2.890	0.24
3PBL	A	100	132	2.890	0.24	3PBL	A	362	385	2.890	0.24
3PIK	A	63	79	2.300	0.20	3PIK	A	122	138	2.300	0.20
3PIK	A	269	285	2.300	0.20	3PIK	A	330	346	2.300	0.20
3PL9	A	90	116	2.800	0.28	3PL9	A	145	163	2.800	0.28
3PL9	A	190	215	2.800	0.28	3PUX	G	11	39	2.300	0.23
3PUX	G	81	110	2.300	0.23	3PUX	G	260	281	2.300	0.23
3PUX	G	212	227	2.300	0.23	3Q7K	J	33	56	2.800	0.22
3Q7K	J	65	83	2.800	0.22	3Q7K	J	107	135	2.800	0.22
3Q7K	J	161	182	2.800	0.22	3Q7K	J	188	203	2.800	0.22
3Q7K	J	247	265	2.800	0.22	3Q7K	J	210	225	2.800	0.22
3QBG	A	62	86	1.800	0.23	3QBG	A	122	141	1.800	0.23
3QBG	A	146	168	1.800	0.23	3QBG	A	241	270	1.800	0.23
3QBG	A	32	55	1.800	0.23	3QBG	A	174	200	1.800	0.23
3QBG	A	208	231	1.800	0.23	3QF4	A	270	311	2.900	0.22
3QF4	A	13	41	2.900	0.22	3QF4	A	51	98	2.900	0.22
3QF4	A	210	263	2.900	0.22	3QF4	A	109	151	2.900	0.22
3QF4	A	158	197	2.900	0.22	3QF4	B	308	333	2.900	0.22
3QF4	B	35	66	2.900	0.22	3QF4	B	75	120	2.900	0.22
3QF4	B	236	285	2.900	0.22	3QF4	B	133	174	2.900	0.22
3QF4	B	177	221	2.900	0.22	3RGB	C	51	72	2.800	0.27
3RGB	C	89	113	2.800	0.27	3RGB	C	127	162	2.800	0.27
3RGB	C	171	197	2.800	0.27	3RGB	C	257	269	2.800	0.27
3RIF	A	214	230	3.340	0.25	3RIF	A	242	264	3.340	0.25
3RIF	A	275	301	3.340	0.25	3RIF	A	308	338	3.340	0.25
3RKO	A	17	42	3.000	0.23	3RKO	A	64	89	3.000	0.23
3RKO	A	97	119	3.000	0.23	3RLB	B	7	25	2.000	0.21
3RLB	B	78	91	2.000	0.21	3RLB	B	107	131	2.000	0.21
3RLB	B	143	172	2.000	0.21	3RLB	B	53	68	2.000	0.21
3RW0	A	1116	1151	2.950	0.27	3RW0	A	1198	1216	2.950	0.27
3RW0	A	1001	1032	2.950	0.27	3RW0	A	1042	1065	2.950	0.27
3RZE	A	33	54	3.100	0.22	3RZE	A	63	89	3.100	0.22
3RZE	A	98	130	3.100	0.22	3RZE	A	143	162	3.100	0.22
3RZE	A	188	216	3.100	0.22	3RZE	A	450	467	3.100	0.22
3RZE	A	408	437	3.100	0.22	3RZE	A	1126	1140	3.100	0.22
3SFE	C	37	65	2.810	0.27	3SFE	C	84	113	2.810	0.27

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
3SFE	C	120	141	2.810	0.27	3SPG	A	156	185	2.610	0.20
3SPG	A	79	108	2.610	0.20	3SYA	A	169	196	2.980	0.24
3SYA	A	91	119	2.980	0.24	3TDO	C	60	78	2.200	0.18
3TDO	C	83	95	2.200	0.18	3TDO	C	101	128	2.200	0.18
3TDO	C	180	195	2.200	0.18	3TDO	C	202	214	2.200	0.18
3TDO	C	224	253	2.200	0.18	3TDO	C	2	20	2.200	0.18
3TDO	C	24	55	2.200	0.18	3TDO	C	136	149	2.200	0.18
3TDO	C	153	175	2.200	0.18	3TUF	A	105	126	2.260	0.17
3TUF	A	139	164	2.260	0.17	3TUI	A	3	40	2.900	0.24
3TUI	A	94	112	2.900	0.24	3TUI	A	186	211	2.900	0.24
3TUI	A	51	63	2.900	0.24	3TUI	A	68	82	2.900	0.24
3TUI	A	144	163	2.900	0.24	3UDC	A	17	49	3.350	0.25
3UDC	A	60	87	3.350	0.25	3UDC	A	92	125	3.350	0.25
3UG9	A	158	174	2.300	0.21	3UG9	A	246	271	2.300	0.21
3UG9	A	281	313	2.300	0.21	3UG9	A	86	107	2.300	0.21
3UG9	A	120	138	2.300	0.21	3UG9	A	186	206	2.300	0.21
3UG9	A	211	241	2.300	0.21	3UON	A	93	126	3.000	0.23
3UON	A	137	166	3.000	0.23	3UON	A	184	213	3.000	0.23
3UON	A	384	411	3.000	0.23	3UON	A	23	49	3.000	0.23
3UON	A	57	85	3.000	0.23	3UON	A	419	442	3.000	0.23
3UX4	A	3	21	3.260	0.24	3UX4	A	26	51	3.260	0.24
3UX4	A	142	159	3.260	0.24	3UX4	A	170	187	3.260	0.24
3UX4	A	76	95	3.260	0.24	3UX4	A	101	122	3.260	0.24
3V2Y	A	42	72	2.800	0.23	3V2Y	A	79	104	2.800	0.23
3V2Y	A	114	145	2.800	0.23	3V2Y	A	158	174	2.800	0.23
3V2Y	A	200	231	2.800	0.23	3V2Y	A	252	280	2.800	0.23
3V2Y	A	294	311	2.800	0.23	3V6I	B	73	102	2.250	0.23
3VOU	A	74	98	3.200	0.29	3VOU	A	22	44	3.200	0.29
3VR8	C	71	94	2.810	0.23	3VR8	C	116	142	2.810	0.23
3VR8	C	149	179	2.810	0.23	3VR8	D	116	141	2.810	0.23
3VR8	D	56	76	2.810	0.23	3VR8	D	82	106	2.810	0.23
3VW7	A	172	205	2.200	0.22	3VW7	A	216	235	2.200	0.22
3VW7	A	266	296	2.200	0.22	3VW7	A	305	338	2.200	0.22
3VW7	A	347	373	2.200	0.22	3VW7	A	92	131	2.200	0.22
3VW7	A	136	163	2.200	0.22	3ZUX	A	251	276	2.200	0.20
3ZUX	A	284	308	2.200	0.20	3ZUX	A	15	27	2.200	0.20
3ZUX	A	36	54	2.200	0.20	3ZUX	A	67	91	2.200	0.20
3ZUX	A	125	137	2.200	0.20	3ZUX	A	155	183	2.200	0.20
3ZUX	A	188	214	2.200	0.20	3ZUX	A	217	245	2.200	0.20
4A2N	B	4	28	3.400	0.24	4A2N	B	41	61	3.400	0.24
4A2N	B	75	99	3.400	0.24	4A2N	B	127	142	3.400	0.24

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
4A2N	B	150	172	3.400	0.24	4AMJ	B	205	236	2.300	0.20
4AMJ	B	278	313	2.300	0.20	4AMJ	B	33	67	2.300	0.20
4AMJ	B	78	104	2.300	0.20	4AMJ	B	111	142	2.300	0.20
4AMJ	B	322	343	2.300	0.20	4AMJ	B	156	178	2.300	0.20
4APS	A	15	46	3.300	0.27	4APS	A	110	136	3.300	0.27
4APS	A	216	241	3.300	0.27	4APS	A	248	266	3.300	0.27
4APS	A	285	310	3.300	0.27	4APS	A	329	342	3.300	0.27
4APS	A	355	376	3.300	0.27	4APS	A	388	401	3.300	0.27
4APS	A	425	439	3.300	0.27	4APS	A	452	470	3.300	0.27
4APS	A	53	75	3.300	0.27	4APS	A	84	103	3.300	0.27
4APS	A	145	171	3.300	0.27	4APS	A	175	198	3.300	0.27
4AYX	A	428	469	2.900	0.23	4AYX	A	167	201	2.900	0.23
4AYX	A	208	257	2.900	0.23	4AYX	A	268	309	2.900	0.23
4AYX	A	369	422	2.900	0.23	4AYX	A	312	357	2.900	0.23
4B4A	A	60	88	3.500	0.25	4B4A	A	101	128	3.500	0.25
4B4A	A	146	175	3.500	0.25	4B4A	A	6	27	3.500	0.25
4B4A	A	31	44	3.500	0.25	4B4A	A	181	202	3.500	0.25
4B4A	A	207	225	3.500	0.25	4DAJ	D	138	170	3.400	0.25
4DAJ	D	182	210	3.400	0.25	4DAJ	D	229	254	3.400	0.25
4DAJ	D	494	513	3.400	0.25	4DAJ	D	66	94	3.400	0.25
4DAJ	D	101	129	3.400	0.25	4DAJ	D	522	544	3.400	0.25
4DJH	B	219	256	2.900	0.23	4DJH	B	271	299	2.900	0.23
4DJH	B	57	86	2.900	0.23	4DJH	B	93	121	2.900	0.23
4DJH	B	129	160	2.900	0.23	4DJH	B	309	332	2.900	0.23
4DJH	B	173	196	2.900	0.23	4DJI	A	41	65	3.190	0.31
4DJI	A	103	117	3.190	0.31	4DJI	A	127	146	3.190	0.31
4DJI	A	154	180	3.190	0.31	4DJI	A	235	259	3.190	0.31
4DJI	A	364	393	3.190	0.31	4DJI	A	410	429	3.190	0.31
4DJI	A	289	321	3.190	0.31	4DL0	G	9	58	2.910	0.21
4DVE	A	3	23	2.090	0.19	4DVE	A	55	70	2.090	0.19
4DVE	A	149	179	2.090	0.19	4DVE	A	90	109	2.090	0.19
4DVE	A	118	143	2.090	0.19	4DXW	B	12	27	3.050	0.24
4DXW	B	38	64	3.050	0.24	4DXW	B	196	224	3.050	0.24
4DXW	B	117	154	3.050	0.24	4EA3	B	49	76	3.010	0.25
4EA3	B	85	113	3.010	0.25	4EA3	B	120	153	3.010	0.25
4EA3	B	164	188	3.010	0.25	4EA3	B	214	241	3.010	0.25
4EA3	B	259	287	3.010	0.25	4EA3	B	295	320	3.010	0.25
4EIY	A	174	208	1.800	0.18	4EIY	A	220	257	1.800	0.18
4EIY	A	4	33	1.800	0.18	4EIY	A	43	67	1.800	0.18
4EIY	A	74	106	1.800	0.18	4EIY	A	111	138	1.800	0.18
4EIY	A	267	290	1.800	0.18	4EIY	A	293	305	1.800	0.18

Continued on next page

Continued from previous page

Code	C	St.	End	Res.	R	Code	C	St.	End	Res.	R
4EJ4	A	45	75	3.400	0.26	4EJ4	A	83	110	3.400	0.26
4EJ4	A	207	238	3.400	0.26	4EJ4	A	258	286	3.400	0.26
4EJ4	A	119	150	3.400	0.26	4EJ4	A	162	186	3.400	0.26
4EJ4	A	295	318	3.400	0.26	4EZC	B	49	63	2.360	0.20
4EZC	B	142	161	2.360	0.20	4EZC	B	173	186	2.360	0.20
4EZC	B	214	227	2.360	0.20	4EZC	B	301	326	2.360	0.20
4EZC	B	85	104	2.360	0.20	4EZC	B	249	268	2.360	0.20
4F4L	C	7	27	3.490	0.27	4F4L	C	74	92	3.490	0.27
4FC4	B	80	92	2.400	0.20	4FC4	B	98	125	2.400	0.20
4FC4	B	177	190	2.400	0.20	4FC4	B	219	248	2.400	0.20
4FC4	B	56	73	2.400	0.20	4FC4	B	150	172	2.400	0.20
4FC4	B	3	20	2.400	0.20	4FC4	B	23	49	2.400	0.20
4G1U	A	7	26	3.010	0.27	4G1U	A	55	79	3.010	0.27
4G1U	A	232	253	3.010	0.27	4G1U	A	279	298	3.010	0.27
4G1U	A	309	325	3.010	0.27	4G1U	A	92	104	3.010	0.27
4G1U	A	115	139	3.010	0.27	4G1U	A	145	168	3.010	0.27
4G1U	A	191	218	3.010	0.27	4GBY	A	125	152	2.810	0.23
4GBY	A	160	186	2.810	0.23	4GBY	A	191	217	2.810	0.23
4GBY	A	277	306	2.810	0.23	4GBY	A	370	384	2.810	0.23
4GBY	A	313	334	2.810	0.23	4GBY	A	341	363	2.810	0.23
4GBY	A	443	461	2.810	0.23	4GBY	A	8	29	2.810	0.23
4GBY	A	50	80	2.810	0.23	4GBY	A	84	103	2.810	0.23
4GBY	A	405	423	2.810	0.23	4GRV	A	61	87	2.800	0.23
4GRV	A	99	126	2.800	0.23	4GRV	A	139	171	2.800	0.23
4GRV	A	187	207	2.800	0.23	4GRV	A	233	265	2.800	0.23
4GRV	A	302	331	2.800	0.23	4GRV	A	341	373	2.800	0.23
4H33	A	15	37	3.100	0.28	4H33	A	68	92	3.100	0.28
4HKR	A	237	270	3.350	0.28	4HKR	A	276	323	3.350	0.28
4HKR	A	145	178	3.350	0.28	4HKR	A	192	215	3.350	0.28