Protein Loop Structure Prediction

Yoonjoo Choi

St. Cross College

Oxford University

A thesis submitted for the degree of

Doctor of Philosophy in Statistics

Oct 2011

This dissertation is dedicated to three great mothers: my sister who gave birth to my dearly beloved nephew Tae-Yeon, my mother who brought me into existence and Mother Nature who governs all things.

Acknowledgements

First of all, I would like to give my sincere gratitude to my supervisor Charlotte. This dissertation would not exist without her. I also thank Professor Peter Jeavons at the computer science department who tossed my application form to Charlotte.

I am indebted to hundreds of people.

My parents and sister aided me both materially and spiritually. I should not forget my nephew, Tae-Yeon. I wrap up my gratitude to them since I will never finish if I list all the love they have given to me.

The department of statistics, St Cross college and Oxford University have funded me to undertake my DPhil study. Especially Ms Maureen Patricia Doherty did much so that I could get funding.

I have started my DPhil study with the two most wonderful jerks in the world: Waqar Ali, the grumbler and Sebastian Kelm, the vampire crow. I am grateful that the Kelms accepted my self-adoption to their family. I have been extremely happy to spend time with them. I would like to thank all former and current OPIG members.

My research career in protein structure prediction began at Korea Institute for Advanced Study. Professor Jooyoung Lee gave me a chance to learn protein bioinformatics. I am still surprised that he gave me such an opportunity although I had not had any experience before. Professor Keehyoung Joo was never tired of my silly tenacious questions and requests. It was one of the happiest periods in my life to be at the institute.

Dr Yoon-Jung Jang is a very very important person who directed a turning point in my future career. If I had not come across her noble spirit, I would not even have attempted to dive into a whole new world of protein

bioinformatics. I believe that she will realise her dream and save all the sick people in the world. If I have achieved anything in this field, I should acknowledge that it is all thanks to her virtue.

Each chapter is owed to many individuals.

I acknowledge that FREAD is not entirely my original algorithm, but my supervisor, Charlotte's. As always, all the basic ideas were already there. I was just the lucky one who could pick ripened fruit. Essentially all my work done is tiny modifications of great achievements honourable scientists have made. I had never thought that this work would become a publication. Dr Joon-Soo Ko and Professor Chaok Seok at Seoul National University encouraged me to publish this material and gave me insightful ideas.

I owe the antibody work to Dr Jiye Shi of UCB Celltech. He opened a new possibility and practical application of my prediction method. Although I had a basic idea of contact information, useful ideas came from Professor *Sir* Tom Blundell at Cambridge and his group members. His enormous databases and knowledge supported my basic idea and how to apply it to a real problem. I also thank the three Korean members of Tom's group who welcomed me warmly: Dr Semin Lee, Dr Sungsam Gong and Dr Jawon Song. The homology models of antibody structures I used were given by Professor Jeffrey Gray at Johns Hopkins University.

The loop stretch work is partially done by Sumeet Agarwal at the physics department. This work was initially inspired by Professor Hagai Meirovitch of Pittsburgh University who I met at ISMB/ECCB 2008, Stockholm.

Apart from the people mentioned above, to everybody I could not list because of the limited space, I am sending my remote hugs with lots of love.

Declaration

I herewith declare that I have produced this dissertation without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such.

This dissertation has not previously been presented or submitted in identical or similar form to any other institute or university.

Abstract

This dissertation concerns the study and prediction of loops in protein structures.

Proteins perform crucial functions in living organisms. Despite their importance, we are currently unable to predict their three dimensional structure accurately.

Loops are segments that connect regular secondary structures of proteins. They tend to be located on the surface of proteins and often interact with other biological agents. As loops are generally subject to more frequent mutations than the rest of the protein, their sequences and structural conformations can vary significantly even within the same protein family. Although homology modelling is the most accurate computational method for protein structure prediction, difficulties still arise in predicting protein loops. Protein loop structure prediction is therefore a bottleneck in solving the protein structure prediction problem.

Reflecting on the success of homology modelling, I implement an improved version of a database search method, FREAD. I show how sequence similarity as quantified by environment specific substitution scores can be used to significantly improve loop prediction.

FREAD performs appreciably better for an identifiable subset of loops (two thirds of shorter loops and half of the longer loops tested) than *ab initio* methods; FREAD's predictive ability is length independent. In general, it produces results within 2Å root mean square deviation (RMSD) from the native conformations, compared to an average of over 10Å for loop length 20 for any of the other tested *ab initio* methods.

I then examine FREAD's predictive ability on a specific type of loops called complementarity determining regions (CDRs) in antibodies. CDRs consist

of six hypervariable loops and form the majority of the antigen binding site. I examine CDR loop structure prediction as a general case of loop structure prediction problem. FREAD achieves accuracy similar to specific CDR predictors. However, it fails to accurately predict CDR-H3, which is known to be the most challenging CDR. Various FREAD versions including FREAD with contact information (ConFREAD) are examined. The FREAD variants improve predictions for CDR-H3 on homology models and docked structures.

Lastly, I focus on the local properties of protein loops and demonstrate that the protein loop structure prediction problem is a local protein folding problem. The end-to-end distance of loops (loop span) follows a distinctive frequency distribution, regardless of secondary structure elements connected or the number of residues in the loop. I show that the loop span distribution follows a Maxwell-Boltzmann distribution.

Based on my research, I propose future directions in protein loop structure prediction including estimating experimentally undetermined local structures using FREAD, multiple loop structure prediction using contact information and a novel *ab initio* method which makes use of loop stretch.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Proteins

Proteins play important roles in living organisms. They perform crucial functions in all biological processes. For example, protein enzymes act as catalysts in metabolism. Antibodies (another type of protein) are key agents in the immune system of vertebrates and many membrane proteins serve as receptors for cell signalling. In general, a protein's function is specified by its three dimensional structure.

1.2 Protein Structure

1.2.1 Amino Acids

Amino acids are the building blocks of proteins. The central atom of an amino acid is its α carbon. Four distinctive groups are linked to the α carbon: An amine group containing nitrogen and hydrogen atoms; a carboxyl group consisting of a carbon atom and an oxygen atom; an R group or a side chain that specifies the amino acid; and a hydrogen atom. These form a tetrahedral structure.

Due to an amino acid's tetrahedral nature and the four different substituents, two

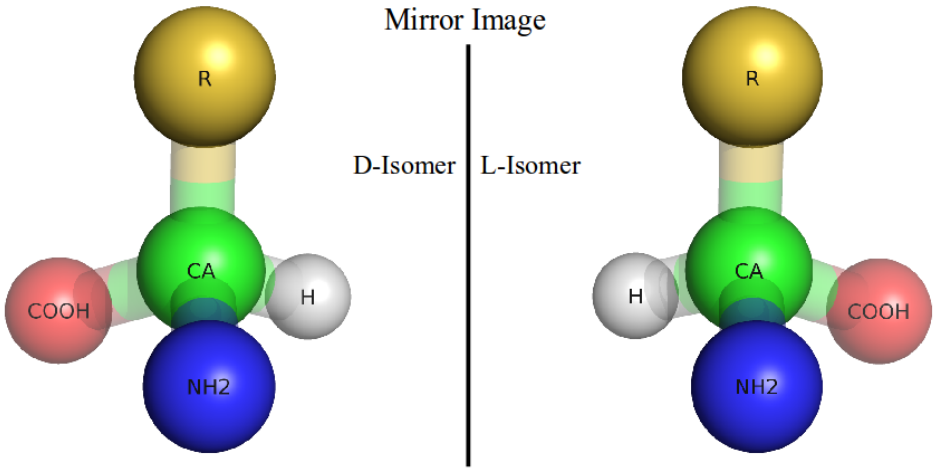


Figure 1.1: The chirality of amino acids
Natural proteins consist of only L amino acid types. The origin of the amino acid homochirality is unknown.

isomers are possible, L and D which are mirror images of one another (Figure 1.1). Natural proteins consist of only the L isomer.

The functions and structures of proteins are due to the specific chemical and physical properties of the side chains of amino acids. Each amino acid side chain has distinct features, such as size, shape, charge and hydrophobicity (Table 1.1 and Table A.1). The hydrophobicity of amino acids plays an important role in protein structure, called the hydrophobic effect. Water soluble globular proteins are likely to have hydrophobic amino acids on the inside while hydrophobic amino acids of membrane proteins are anchored into the lipid layer. The hydrophobic effect is thought to be one of the main driving forces of protein folding.

Alanine, leucine, isoleucine and valine have side chains consisting of only hydrogen

[1]There are many ways to calculate hydrophobicities. For example, free amino acids can be used to calculate hydrophobicities. Amino acids whose amine and carboxyl groups are blocked can also be used. Or propensities of each amino acid being in solvent exposed or buried areas of protein structures can be calculated. The hydrophobicity scale here is calculated using the last approach (Kyte and Doolittle, 1982). Amino acids are sorted in terms of the hydrophobicity.

[2]The unit is *Da* (Dalton).

Table 1.1: Standard amino acid codes and chemical properties

Amino Acid	Code	Hydrophobicity[1]	Polarity	Charge	Avr. Mass[2]
Arginine	Arg (R)	-4.5	polar	positive	174.20274
Lysine	Lys (K)	-3.9	polar	positive	146.18934
Asparagine	Asn (N)	-3.5	polar	neutral	132.11904
Aspartic acid	Asp (D)	-3.5	polar	negative	133.10384
Glutamic acid	Glu (E)	-3.5	polar	negative	147.13074
Glutamine	Gln (Q)	-3.5	polar	neutral	146.14594
Histidine	His (H)	-3.2	polar	neutral	155.15634
Proline	Pro (P)	-1.6	nonpolar	neutral	115.13194
Tyrosine	Tyr (Y)	-1.3	polar	neutral	181.19124
Tryptophan	Trp (W)	-0.9	nonpolar	neutral	204.22844
Serine	Ser (S)	-0.8	polar	neutral	105.09344
Threonine	Thr (T)	-0.7	polar	neutral	119.12034
Glycine	Gly (G)	-0.4	nonpolar	neutral	75.06714
Alanine	Ala (A)	1.8	nonpolar	neutral	89.09404
Methionine	Met (M)	1.9	nonpolar	neutral	149.20784
Cysteine	Cys (C)	2.5	nonpolar	neutral	121.15404
Phenylalanine	Phe (F)	2.8	nonpolar	neutral	165.19184
Leucine	Leu (L)	3.8	nonpolar	neutral	131.17464
Valine	Val (V)	4.2	nonpolar	neutral	117.14784
Isoleucine	Ile (I)	4.5	nonpolar	neutral	131.17464

and carbon atoms. They are nonpolar, hydrophobic and chemically unreactive. The side chains of serine and threonine contain hydroxyl groups (one oxygen atom covalently bound with a hydrogen atom). They are hydrophilic and polar. Due to the hydroxyl groups, they can be involved in hydrogen bonds. Note that the side chains of isoleucine and threonine are chiral and only one isomer is found.

Phenylalanine, tryptophan and tyrosine have aromatic rings in their side chains. Phenylalanine and tyrosine are very similar to each other. However, due to the hydroxyl group in the tyrosine side chain, its polarity is different from phenylalanine (tyrosine: polar and phenylalanine: nonpolar).

Aspartic acid and glutamic acid are polar and negatively charged due to their carboxyl groups. At neutral pH (7.0), their residues tend to be ionised and very polar. The side chains of asparagine and glutamine are very similar to aspartic acid and glutamic

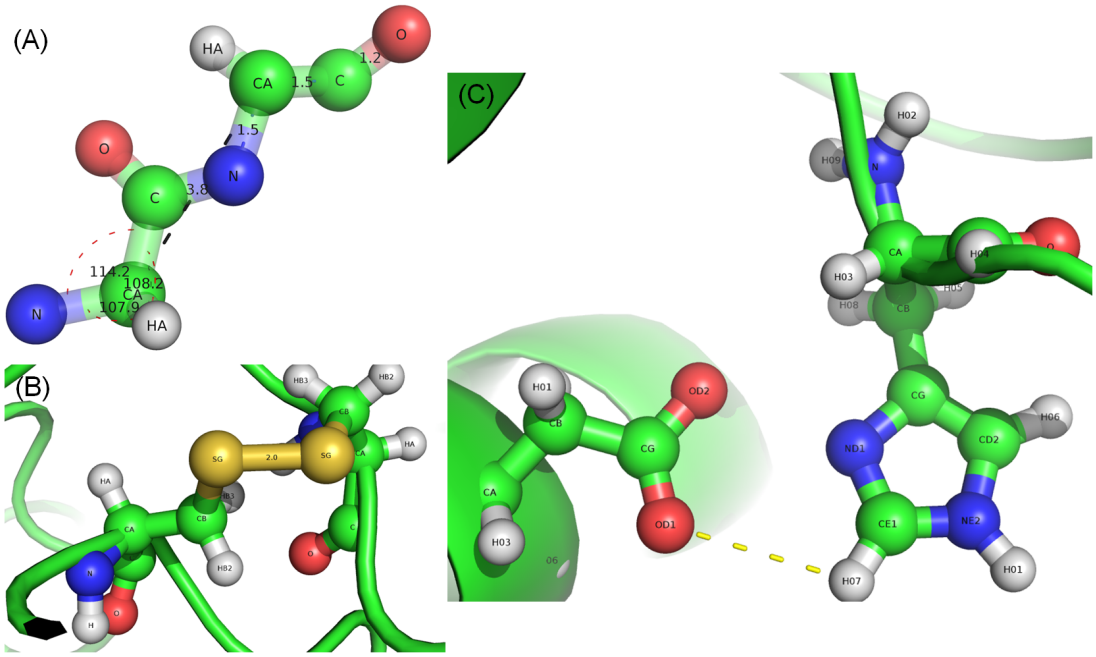


Figure 1.2: Bonding in Protein Structure
Chemical bonds play important roles in protein structures. (A) The peptide bond is a strong covalent bond between amino acids. The bond lengths and bond angles along the main chain of a polypeptide have limited ranges (the typical $C\alpha - C\alpha$ distance is approximately 3.8Å). (B) Covalent bonds between cystein side chains form disulfide bridges. The disulfide bond is most stable with a torsion angle ($C\beta - S - S - C\beta$) of $\pm 90°$. The typical distance between two sulfur atoms of a disulfide bond is approximately 2Å. (C) A salt bridge is a combination of two important noncovalent interactions: hydrogen bonding (yellow dotted line) and charge interactions. In this example, a negatively charged aspartic acid residue (left) is in an electrostatic interaction with a positively charged histidine residue (right).

acid except for the amide side chains instead of the carboxyl groups. They share similar properties and masses to the acidic amino acids, but do not have charges. Histidine, lysine and arginine have high pKa values (about 6, 10 and 12.5 respectively). Especially arginine and lysine are ionised under most physiological conditions and positively charged.

Cystein and methionine have a sulfur atom in their residues. Two cystein residues can form a disulfide bridge (Figure 1.2B). Glycine is the simplest amino acid which is not chiral. Due to the simplest side chain, glycine has less limitations in dihedral angles

(Section 1.2.2). The side chain of proline, on the other hand, is covalently bonded to the backbone nitrogen atom. This unique side chain formation restricts dihedral angle rotations.

The physicochemical properties of the amino acids are important drivers in protein folding. Generally the buried regions of soluble proteins are compactly packed and very well conserved. For example, buried water molecules and their interfaces are conserved among homologous protein members (Sreenivasan and Axelsen, 1992). If a buried amino acid of a protein whose side chain is small such as glycine, and if it is replaced by a large amino acid such as arginine during evolution, the protein may not maintain the same core structure.

Proteins are linear polymers composed of amino acids linked by peptides bonds. Peptide bonds are chemical linkages between the carboxyl group (COOH) of one amino acid and the amine group (NH_2) of another (Figure 1.2(A)). A peptide bond is formed when the two groups meet and a water molecule is expelled.

The backbone or main-chain of a protein therefore consists of four atoms: nitrogen, the α carbon, carbon, and oxygen. The first residue (or amino acid) of a protein has a free amine group (the N terminal of the protein). The final residue has a free carboxyl group (the C terminal of the protein).

A peptide chain has a planar conformation due to the partial double bond character between the carboxyl carbon and the backbone amine nitrogen atom (Pauling et al., 1951). All the bond lengths and angles in the backbone have limited ranges (Engh and Huber, 1991).

1.2.2 Dihedral Angles

The main degrees of freedom can be described by dihedral rotations along the bonds. There are two possible rotations per amino acid of the backbone. These dihedral angles are denoted ϕ ($C - N - C\alpha - C$) and ψ ($N - C\alpha - C - N$). If $\phi = \psi = 180$ degree,

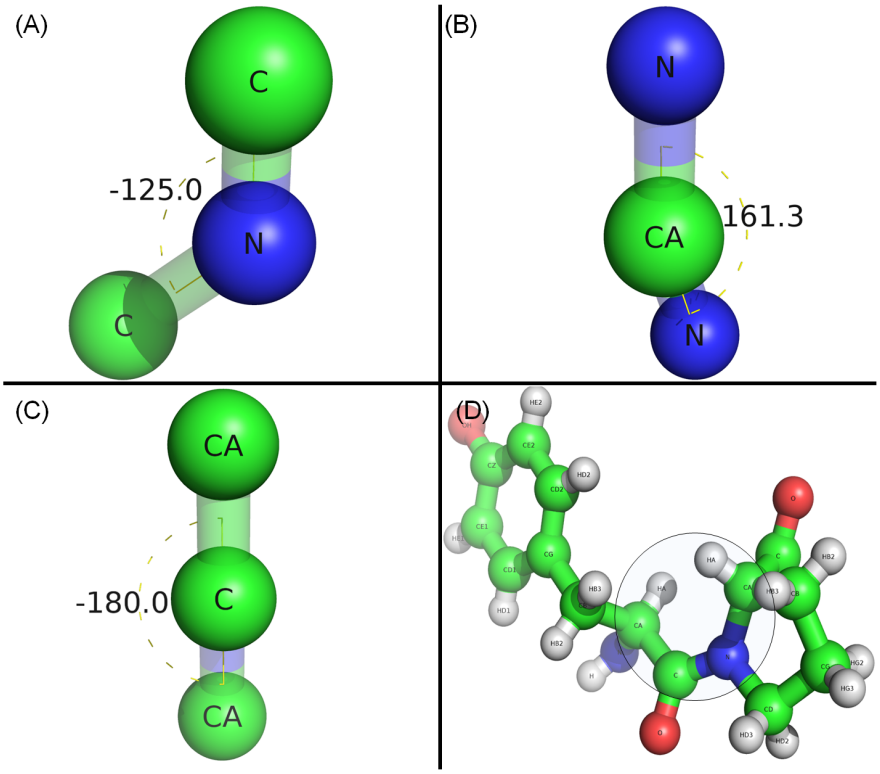


Figure 1.3: Dihedral angles
(A) ϕ angle (a rotation around the the $C-N-C\alpha-C$ bond) and (B) ψ angle (around the $N-C\alpha-C-N$ bond). (C) A peptide bond forms a plane between two adjacent residues and the ω angle between two residues connected by the peptide bond is around either 180° (*trans*) or 0° (*cis*). (D) The *cis* isomerisation is rare and mostly found in the X-Proline peptide bond. In this example, the ω angle between HIS-PRO is about 10°.

two adjacent residues are coplanar (Figure 1.3).

Ramakrishnan and Ramachandran (1965) described the Ramachandran plot (Figure 1.4), which displays combinations of ϕ and ψ dihedral angles in two dimensional space. Due to steric clashes among adjacent side chains, some Ramachandran areas are disallowed (the white blank areas in Figure 1.4). For example, The large disallowed region from roughly from 70° to 170° in ϕ dihedral angle is due to the steric clash between side-chain atoms and the backbone oxygen atom. Further distinctions can be calculated by examining the dihedral angles observed in known protein structures which let us identify favourable conformations. The Ramachandran plot also shows

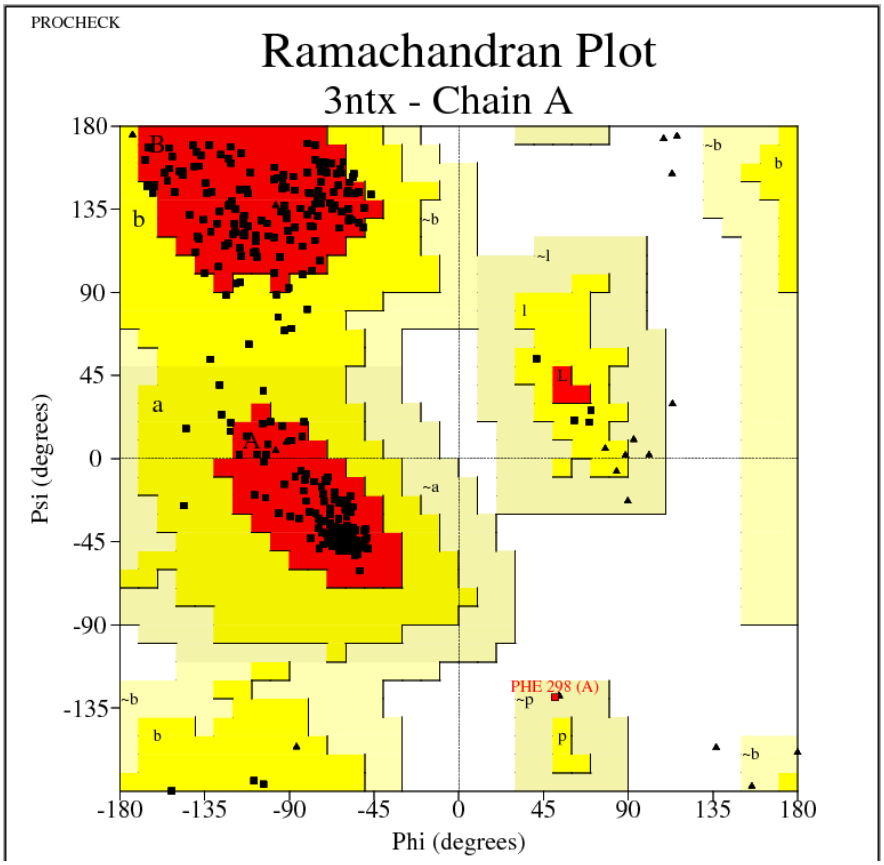


Figure 1.4: Ramachandran plot generated by PROCHECK (Laskowski et al., 1993)
The white areas are disallowed dihedral angle regions. Favourable angles are colored in red. "a" indicates the helical region, "b" shows β strands, and the "l" region is for left-handed helices. The black dots represent dihedral angles of the protein structure (PDB code 3NTX).

local patterns of a protein, i.e. secondary structures.

1.2.3 Levels of Protein Structure Organisation

There are four levels of protein structures. As a result of the peptide bonding of amino acids, a protein has an one-dimensional sequential order and forms a polypeptide chain (primary structure). Parts of the sequence fold into local structures (secondary structure). The local structures are packed together and form a unique three dimensional structure (tertiary structure). Several chains can be composed together and a func-

tioning protein is formed (quaternary Structure).

Primary Structure

Historically, the primary structure of a protein was firstly termed by Linderstrom-Lang (1952) as the unfolded peptide chain of a protein. In general, the primary structure is described by the sequence of residues along the protein. By convention, the primary structure reads from the amine group (N terminal) to the carboxyl group (C terminal).

The primary structure is critically important in protein folding. Anfinsen performed a series of experiments with bovine pancreatic RNase (Anfinsen and Haber, 1961; Anfinsen et al., 1961) and demonstrated that a protein structure is in thermodynamic equilibrium with its environment and the interactions between its amino acid side chains determine its three dimensional structure (Anfinsen, 1973). This hypothesis is the basic assumption of computer-based protein modelling.

Secondary Structure

Secondary structure in proteins is a repeating local pattern in three dimensional space. There are two main patterns; helix and β strand. Secondary structures are formally defined by hydrogen bonds (a non covalent electrostatic interaction).

Helices are rod-like helical structures where the side chains point outwards. They are further classified in terms of periodic hydrogen bonds: 3_{10} helices (hydrogen bond between the ith and $i+3$ residues), α-helices (the ith and $i+4$ residues) and π-helices (the ith and $i+5$ residues). Helices are mostly right-handed due to steric preferences. A β sheet is composed of two or more fully extended polypeptide chains called β strands. The adjacent β strands are bonded through hydrogen bonds. According to the directions of adjacent β strands, a β sheet can be parallel or anti-parallel. Secondary structures can be annotated using computational methods such as DSSP (Define Secondary Structure of Proteins) (Kabsch and Sander, 1983).

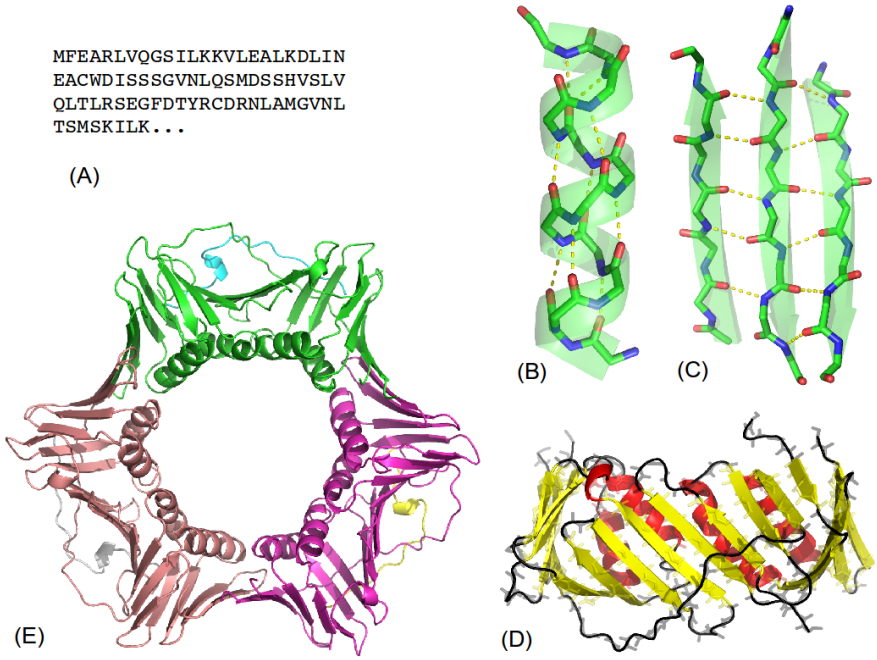


Figure 1.5: The levels of the protein structure organisation (PDB Code: 1AXC) (A) A sequence is the primary structure of the protein . The three dimensional structure of the protein consists of several regular secondary structures such as helices (B) and β sheets (C). (D) The secondary structures are linked by loops (coloured black) and form a tertiary structure. (E) Multiple folded proteins can be arranged together and constructs a multi-subunit complex, the quaternary structure.

In this dissertation, protein loops are defined as the segments of protein structure which connect the regular secondary structures. They are discussed further in Chapter 2.

Tertiary Structure

The tertiary structure of a protein is the overall shape of the polypeptide chain in three dimensional space. A tertiary structure usually consists of several elements of secondary structure. The folding to tertiary structure of water-soluble globular proteins is thought to be driven by the hydrophobic effect (Pace et al., 1996). Thus, the side chains of interior residues tend to be hydrophobic whereas those of exterior residues are largely hydrophilic in globular proteins. The tendencies of membrane proteins are

different from those of soluble proteins. For example, membrane-embedded regions typically have hydrophobic residues on the outside in order to interact with the lipids in the membrane (Stevens and Arkin, 1999).

Hydrogen bonds are also thought to be important to the tertiary structure folding. A folded structure forms many internal hydrogen bonds between atoms whereas, in an unfolded protein, the atoms in the hydrogen bonds are broken and form the bonds with surrounding water molecules (Myers and Pace, 1996). Hence, it is important to expel such water in the process of protein folding.

Quaternary Structure

The quaternary structure is a functioning protein that consists of multiple polypeptide chains. Each chain is called subunits and two or more chains are assembled together (dimer, trimer, tetramer, and so on). Each subunit can be identical or different. For example, in Figure 1.5(E), A human PCNA (Proliferating Cell Nuclear Antigen) consists of three identical chains (homo-tetramer).

Subunits are assembled in a similar way as in secondary or tertiary structures: noncovalent interactions such as hydrogen bonds or salt bridges, and covalent bonds like disulfide bridges. In general hydrophobic groups in each chain are joined together (Chothia and Janin, 1975).

1.3 Determination of Protein Structure

The Protein Data Bank (PDB) contains experimentally determined three dimensional structures of proteins (Sussman et al., 1998). About 99% of protein structures stored in the PDB are determined by X-ray crystallography (88%) and nuclear magnetic resonance spectroscopy (11%). In general, X-ray determined protein structures are considered to be more reliable to use (Bastolla et al., 2001).

1.3.1 X-ray Crystallography

X-ray crystallography determines a protein structure by the X-ray diffraction pattern of a crystallized protein. X-ray crystallography consists of several steps: preparing for crystals of a purified protein, measuring X-ray diffraction patterns of the crystal and interpreting the patterns.

The crystal is a three dimensional array of unit cells which contain proteins that adopt roughly identical orientations. Each diffracted X-ray is a sum of contributions of all scattered beams in the unit cells. The data collected from X-ray diffraction are a list of intensities. The intensities are recorded on a detector film and the center of the film is taken as the origin. The origin is not measurable as X-rays are obscured when they pass through the crystal.

The reflections on the film are described in an imaginary three dimensional space called reciprocal space. The axes of the reciprocal space is conventionally written as (h, k, l) and the center is written as $(h, k, l) = (0, 0, 0)$. h, k, l are always integer values that indicate how many lattices exist per unit cell and each individual reflection is assigned to be in the reciprocal space. Thus, individual reflections have 1) their own coordinates in the reciprocal space and 2) intensities (I_{hkl}).

As the reciprocal space is an inverse of Cartesian coordinate system, it has a unit of Å^{-1}. From this fact, one can estimate the Euclidean distance between reflections from the reciprocal space distance. If the furthermost reflection from the center is 1/2Å, it gives a resolution collected from the data (a resolution of 2Å). If the resolution of a protein is lower than 2.0Å, all the secondary structures and side chains of the protein are thought to be reliably positioned (Table 1.2).

Each reflection is essentially diffractions of electromagnetic waves. A reflection is a wave that many waves are superposed and so can be described as a Fourier sum. The Fourier sum for the reciprocal space is called the structure factor (F_{hkl}). The structure factor is proportional to the measurable quantity, intensity I_{hkl}, which is the amplitude

of the structure factor ($F_{hkl} \propto \sqrt{I_{hkl}}$).

The objects which diffracts X-rays are electron clouds of the protein in the crystal. The resulting image from X-ray crystallography is an electron cloud map. Using our knowledge about the protein sequence and the structures of amino acids, one can build a model from the electron cloud map. The model in the Cartesian space can be transformed into the reciprocal space and structure factors can be calculated. The R-factor (R) is the average fractional difference between the calculated structure factors (F_{calc}) and the originally observed structure factors (F_{obs}). Models are iteratively built and improved until the F_{calc} and F_{obs} agree.

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}. \tag{1.1}$$

$R = 0$ is ideal. A protein structure of $R \leq 0.2$ and resolution 2.0Å is in general considered to be a high-quality model. Morris et al. (1992) concluded that the most crucial measure of protein structure quality is the R-factor. If the R-factor of a protein structure is lower than 0.2, the structure is in general reliable (Blundell and Johnson, 1976; Morris et al., 1992).

Brünger (1992) proposed the free R-factor (R_{Free}) which takes randomly chosen intensities as a cross-validation subset. R_{Free} calculates how well the model agrees with the random subset which is not included in the model refinement. R_{Free} is supposed to be a better model quality indicator than R.

Table 1.2: Confidence in structural features of proteins determined by X-ray crystallography (Lesk, 1991)

	Resolution			
Structural feature	5 Å	3 Å	2.5 Å	2.0 Å
Chain tracing	-	Fair	Good	Good
Secondary structure	Helices fair	Fair	Good	Good
Side-chain conformations	-	-	Fair	Good
Orientation of peptide planes	-	-	Fair	Good

As mentioned earlier, a crystal is an array of unit cells and, in each unit cell, atom positions may not be placed in the same places. Physically, protein structures constantly move due to thermal vibration of backbone and side chain atoms. The B-factor or temperature factor of an atom in protein crystal structures reflects the fluctuation of its average position.

$$B_i = 8\pi^2\{u_i^2\} \simeq 78.96\{u_i^2\}, \tag{1.2}$$

where u_i is the mean square displacement of an atom i. For example, if the B-factor of atom i is 78.96Å^2, its total mean displacement is about 1Å. The B-factor is also used to estimate the quality of the structure and lower values are regarded as better. In WHATCHECK (Hooft et al., 1996), a programme used to assess protein structure

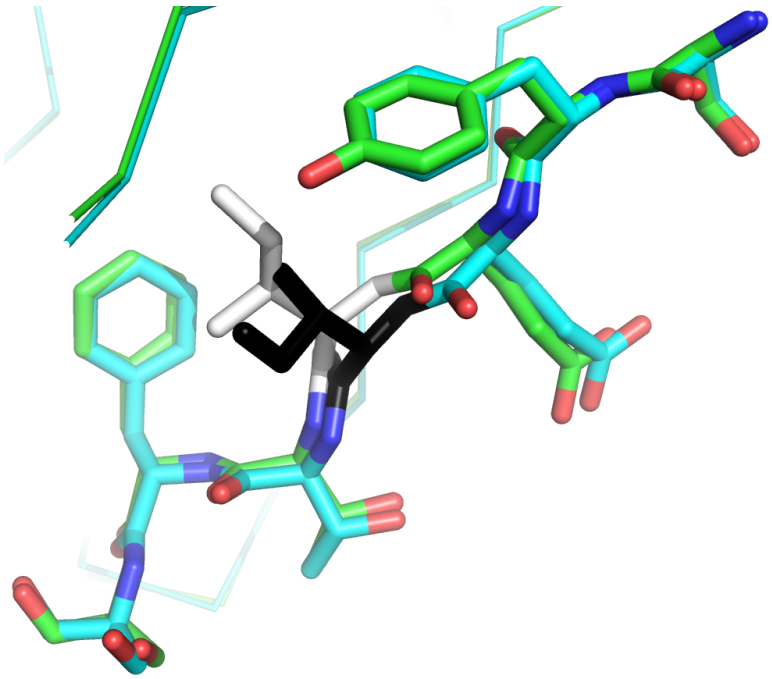


Figure 1.6: Structural differences between identical protein chains due to the crystal packing effect
A homodimer structure of plastocyanin has two identical chains (PDB Code: 3CVB). Chain A (green) and B (cyan) are superimposed. The crystal packing interfaces show different side chain structures, especially on Ile74 residues (white and black, respectively).

quality, the maximum acceptable average B-factor for buried atoms is approximately 24.

High B-factors may indicate thermal vibration or disorder, and cannot be identified in the electron maps. The common way is to remove such residues from the model. Terminal regions (especially N-terminus) are often omitted due to the effect. Structural distortions may occur due to other effects, such as low resolution, unexplained electron maps caused by unexpected molecules and crystal packing (Figure 1.6).

1.3.2 Nuclear Magnetic Resonance Spectroscopy

The basic idea of NMR spectroscopy is that some nuclei (for example, ^{1}H and ^{13}C) have net nuclear spins and behave like small magnets. Hence, when such nuclei are placed in a magnetic field, they are arranged in preferred orientations. The orientations can be changed by electromagnetic radiation. Different nuclei absorb different frequencies of energy. Therefore, specific nuclei can be detected by their characteristic energy absorption.

What NMR spectroscopy gives is a list of distance restraints. Using our knowledge about protein structures, models are built based on the distance restraints provided by NMR spectroscopy. Molecular Dynamics (MD) is often used to fit the restraints. During this process, several physically and chemically plausible models which satisfy the restraints are constructed. The final result of NMR structure determination is an ensemble of such models.

In the PDB, two types of NMR determined structures are found: multiple models and single averaged models (Figure 1.7). The averaged model is essentially an average of atom coordinates in the multiple NMR models. Simple averaging is highly likely to violate steric clashes and produce implausible bond lengths and angles. Thus, the averaged model is further refined to satisfy all the restraints. The root mean square deviations (RMSDs) for each atom are calculated and recorded as B-factors.

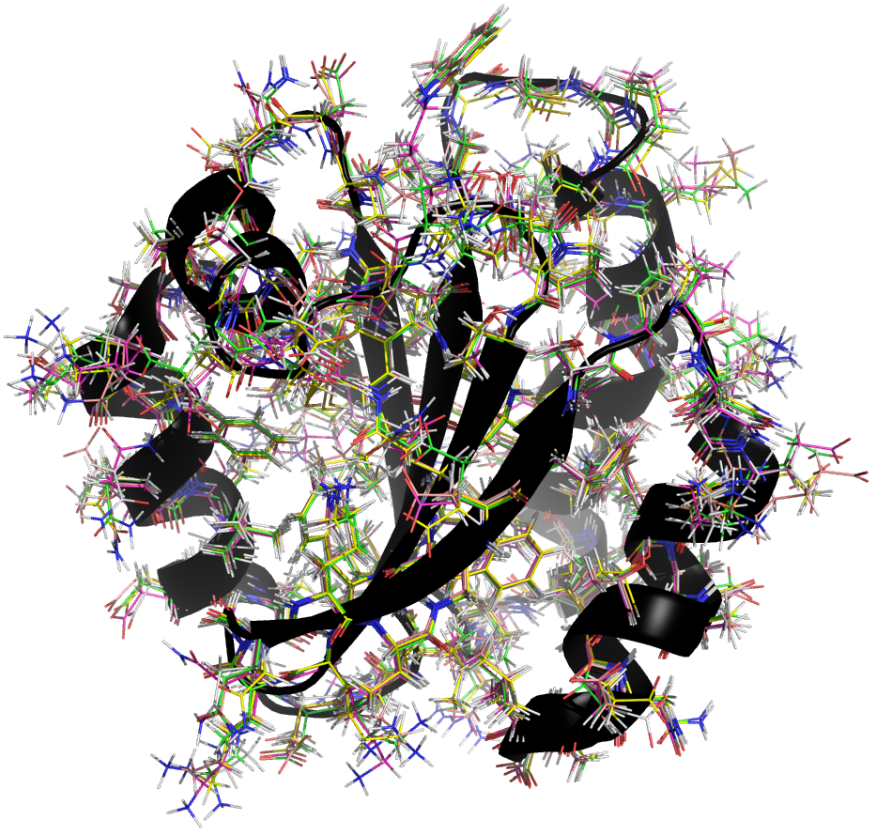


Figure 1.7: Protein structures determined by NMR spectroscopy
NMR spectroscopy generates multiple models derived from distance restraints. There are two PDB codes for Human thioredoxin structures determined by NMR in PDB (4TRX and 3TRX). 4TRX contains 33 models (First five models are displayed here in line representation) and 3TRX is an averaged model (black cartoon).

Just like B-factors, higher RMSDs of the averaged model may mean alternative conformations or static disorders. The primary reason for high RMSD values is the lack of sufficient distance restraints.

1.3.3 Computational Protein Structure Prediction

Experimental determination is the most accurate way to obtain protein structures. However, it is costly and time-consuming. There were 71158 protein structures deposited by Dec. 2010, including 20333 non redundant structures (sequence identity $< 30\%$) and 3427 human protein structures in that set. Compared to the number

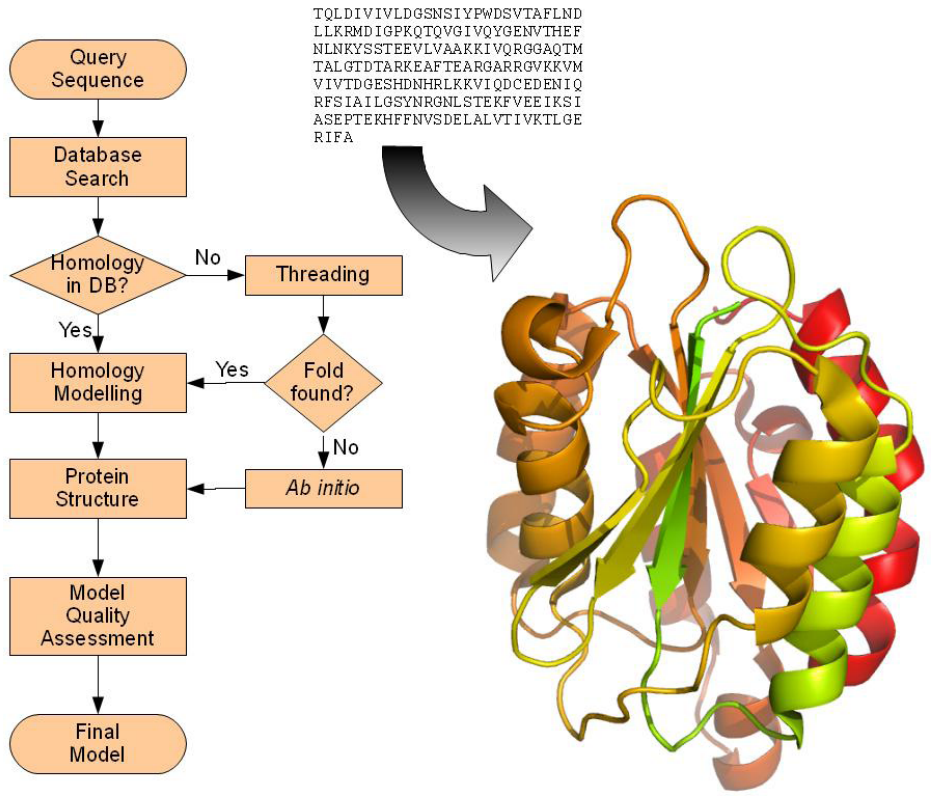


Figure 1.8: The general procedure of computational protein structure prediction
For a given target sequence, the first step is to identify homologous proteins which share high sequence similarities. If no homologous protein is found, threading is performed to find templates that have similar fold types. If no similar fold is found, *ab initio* methods are introduced. The sequence and structure displayed here is from PDB Code 1QCY.

of protein sequences known (about 12 million non-redundant sequences), only a small fraction of proteins have solved structures. Computational protein structure prediction is an alternative method to close this gap.

The general process of protein structure prediction consists of several steps (Figure 1.8). Current protein structure modelling techniques are examined at CASP (Critical Assessment of Techniques for Protein Structure Prediction) (Kryshtafovych et al., 2005; Moult et al., 2007, 2009), a biannual community-wide experiment on protein structure prediction. The main goal of CASP is to examine the ability and limitation of current

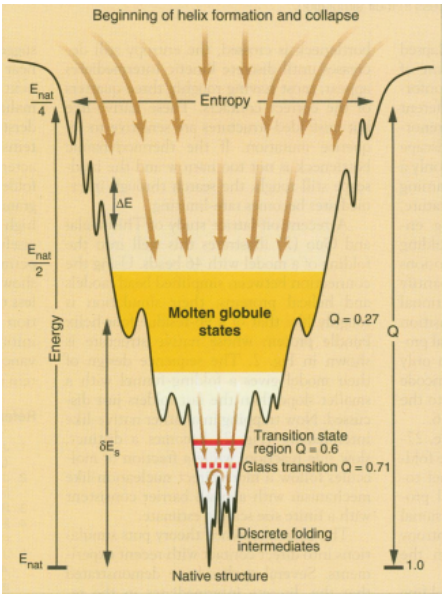


Figure 1.9: A schematic view of the energy funnel.
The folding funnel of a fast folding 60-residue helical protein. The native structure is in the lowest free energy. The figure was taken from Wolynes et al. (1995) (Figure 1).

computational modelling methods. Protein modellers perform blind predictions of given protein sequences while experimentalists determine the structures.

Ab initio Modelling

Modelling of protein structures is based on the principle that a protein adopts its native structure of the lowest free energy (Figure 1.9). Theoretically it is possible to compute the native structure when an exact force field is given and a sufficient amount of conformations are sampled.

ab initio methods make use of the principle in order to predict native structures. Based on our knowledges of physicochemical properties of amino acids, many structural conformations are sampled against a force field energy function [1]. Optimisation algorithms such as genetic algorithm, simulated annealing, conjugate gradient and so on are used to minimise sampled conformations. In general, *ab initio* methods are time-

[1]The general form of force field energy functions is described in Section 2.2.3

consuming and require huge computing powers. They are applied to small proteins which are known to fold fast (Freddolino et al., 2010; Moult et al., 2009).

There are several issues in the *ab initio* approach: sampling, force field energy and optimisation. As an energy landscape of a protein tends to be complicated, large-scale sampling is required. The recent advances in high performance MD simulations such as parallel computing and distributed computing show improvements in sampling speed (Klepsis et al., 2009). Exact force field energy functions are essential in *ab initio* modelling and simulation (Freddolino et al., 2009). For example, one of the most popular force field energy functions, CHARMM (Brooks et al., 1983; MacKerell et al., 1998), tends to overestimate helical structures (Best et al., 2008). It is possible that a protein folding pathway has many local minima near its global minimum (Figure 1.9). Apart from the automated optimisation techniques, human intuition may be useful to overcome this problem (Khatib et al., 2011).

Homology (Comparative) Modelling

The most successful and commonly used method is homology (comparative) modelling. Homology modelling is based on the fact that homologous proteins share very similar structures. Structures of homologous proteins are more conserved than sequences during evolution. Thus proteins of slightly dissimilar sequences can adopt similar structures. Homologous proteins are defined as a group of proteins which have a common ancestor. Although homology is not a synonym for high-sequence similarity, they are interchangeably used in general homology modelling. Generally >30% sequence identity is a standard cut-off value (Rost, 1999; Sanchez et al., 2000). Homology modelling takes such high sequence identity solved structure as templates and copy atom coordinates from the templates. If a template structure and the target sequence to be modelled have high sequence similarity, homology modelling may give reasonably accurate models, and is the most accurate computational modelling method (Baker and

Sali, 2001; Ginalski, 2006).

The first and most crucial step in homology modelling is to identify homologous proteins in a database. BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990), a fast sequence search algorithm, is the most popular sequence search method. Additionally PSI-BLAST (Position Specific Iterative BLAST) (Altschul et al., 1997) can be used to detect remotely related homologous protein sequences. PSI-BLAST creates sequence patterns called "profile" from the collected sequences found by BLAST and finds matched sequences again using the profile. When templates are identified, the target sequence and the template sequences are aligned. If the alignment fails, the final three dimensional protein structure will be wrong although highly homologous proteins are used as the templates.

During the above processes, substitution matrices play an important role. A substitution matrix describes values for all residue pairs. It is used to assign scores to each pair of residues in the alignment. An amino acid substitution matrix generally has 20 by 20 cells, each of which is occupied by a score. The simplest substitution matrix is the identity matrix which assumes that each amino acid does not change to any other amino acids. The most popularly used score matrices are based on the probabilities that amino acids mutate into other amino acids during evolution. PAM (Point Accepted Mutation, Dayhoff et al. (1978)) and BLOSUM (BLOck SUbstitution Matrix, Henikoff and Henikoff (1992)) are two most widely used substitution matrices. The two matrices are constructed in different ways. For example, BLOSUM is built using highly conserved regions without gaps whereas PAM uses global alignments including both conserved and mutable regions. The final matrices are logarithmic matrices of mutation probabilities called log-odds scores.

$$S_{ij} = a \cdot \log_b \left(\frac{Pr_{ij}}{q_i \cdot q_j} \right), \tag{1.3}$$

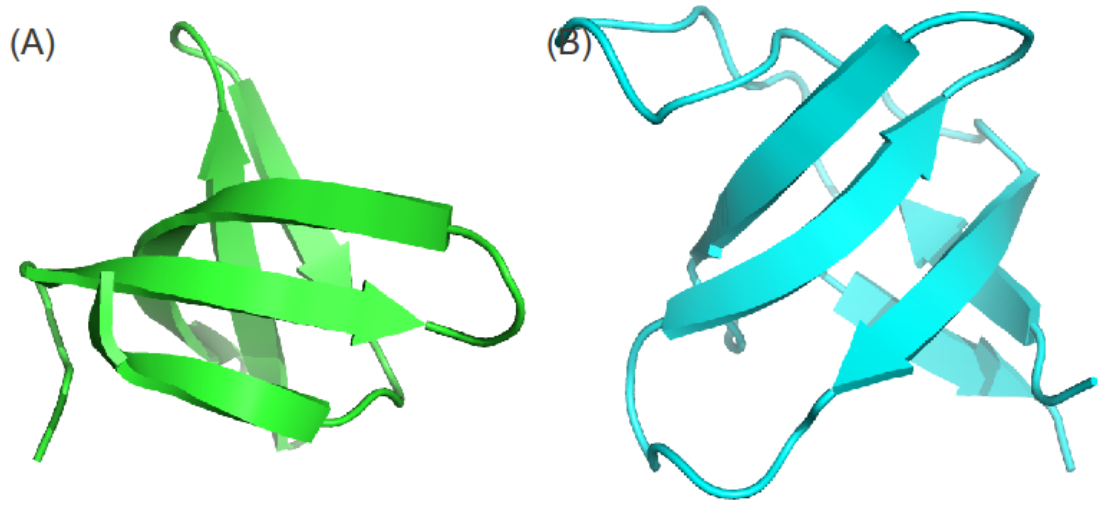


Figure 1.10: An example of proteins of non detectable homology, but similar structures
Two protein structures which share only about 14% sequence identity show similar structures. (A) dihydrofolate reductase (PDB Code: 1BIA) and (B) a kinase (1SHG).

where S_{ij} is the score, Pr_{ij} is the probability that the amino acid i is replaced by j, q_i and q_j are the background probabilities that the amino acids i and j in any protein sequences, a is the scaling factor and b is the log base.

The scaling factor a can vary. For example, the most widely used PAM and BLOSUM matrices, PAM250 and BLOSUM62, are scaled by $3/\log_2$ (third-bits) and $2/\log_2$ (half-bits) respectively. An alignment score is calculated by summation of each score from aligned pairs of amino acids.

Another crucial step in homology modelling is loop structure prediction. Loop regions are generally most variable in sequence and more subject to mutation. Thus they are usually unaligned and cannot be copied from templates. Loops often play functional roles such as ligand recognition and binding. Sometimes they are not determined by experimental methods as they are often mobile. Further details of protein loop structure prediction are discussed in Chapter 2.

Threading (Fold Recognition)

When no homologous protein structures are available in a modelling situation, there is an alternative template based approach, called threading (fold recognition). Proteins often adopt similar structures although the sequence similarity between the target and the template is not high. The idea is based on the general observation that protein structures have a limited number of folds. There are thought to be about 1200 different fold types (Andreeva et al., 2007; Greene et al., 2007) (Figure 1.10). These similar structures can be due to remote homology or the limited fold types caused by physical restraints.

Threading is a technique to identify such similar protein folds which do not have high sequence identities. When a target sequence is given, threading seeks to identify template structures in a database according to scoring functions. A common method is a use of substitution score matrices which take environment and structure information into account. FUGUE (Shi et al., 2010) has 64 scoring matrices which contain structural information such as solvent accessibility, hydrogen bonding and main chain conformations.

1.4 Outline of Dissertation

In Chapter 2, protein loops and their structure prediction are further discussed. Protein loop structure prediction using a database search method, FREAD, is introduced in Chapter 3. I demonstrate that one can accurately predict loop structure using a sequence similarity measure. Chapter 4 is an extension of Chapter 3. FREAD is tested on antibody complementarity determining regions. In this chapter, I demonstrate a method to predict interacting loops using contact information. In Chapter 5, I analyse the end-to-end distance of loops (span) and demonstrate that the span distribution follows a specific probability distribution (the Maxwell-Boltzmann distribution). I show

that the normalised span (loop stretch) is an informative measure indicating the difficulty level of protein loop structure prediction. From those results, I propose future directions of protein loop structure prediction in Chapter 6.

Chapter 2

Protein Loop Structure

2.1 Protein Loop

2.1.1 Definition

Apart from the helix and strand, there is another class of secondary structure which connects the two regular secondary structures. The discovery and classification of such structures, called "turn", date back to Venkatachalam (1968). Turns are short segments (2–6 residues in length) which have irregular patterns. They are further categorised in terms of the number of residues[1].

In an early study, irregular secondary structures longer than seven residues were classified as a novel class (Leszczynski and Rose, 1986). This type of structure was termed "loop" which came from its geometrical shape (typically, like the Greek letter Ω). Loops also connect two regular secondary structures, but in a less tight manner than turns.

The distinctive differences between turns and loops are the ranges of number of residues they have (turn: 2–6 residues, loop: >6 residues) and their tightness (Chou,

[1] δ- or 2-turn (Toniolo, 1980), γ- or 3-turn (Nemethy and Printz, 1972), β- or 4-turn (Venkatachalam, 1968), α- or 5-turn (Toniolo, 1980) and π- or 6-turn (Kim and Sussman, 1976). The numbers indicate residues of the turns.

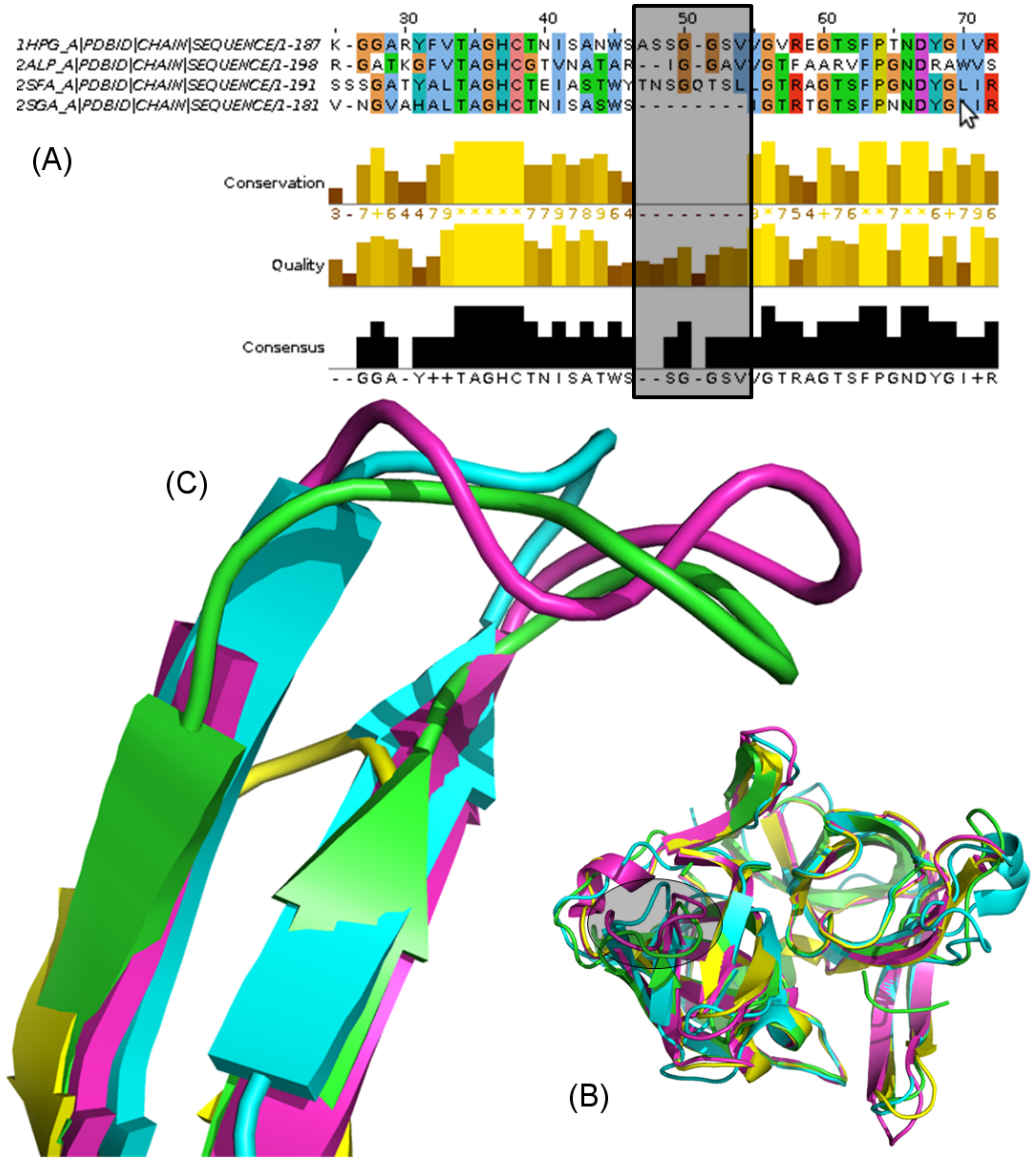


Figure 2.1: Loops from homologous proteins
Protein loops may have different sequences and structures within the same protein family. (A) The result of multiple sequence alignment of a protein family (Serine Proteinase) shows a gapped region. (B) The gapped region tends to be loops and (C) they appear to have different structures. Even in a homology modelling situation in which good templates are identified, it is difficult to predict the loop since definitive template structures are not given in the region.

2000; Leszczynski and Rose, 1986). The tightness is defined by structural features and becomes less meaningful in protein modelling when the structure is not given. Turns

and loops share a common feature: lack of repetitive hydrogen bond patterns. In this dissertation, turns are deemed short loops. Therefore, all local polypeptide chains, aside from helices and strands, are defined as loops. The regular secondary structures at either end of a loop are called "anchor regions".

A typical globular protein has on average a third of its residues in loops (Donate et al., 1996; Rose et al., 1983). They tend to be located on the surface of the protein and show far more variation between homologous protein structures than regular secondary structures do (Crawford et al., 1987; Rose, 1978).

The most successful computational modelling method, homology modelling, is based on the idea that homologous proteins share similar structures. However, due to larger differences in loop regions between homologous protein structures, difficulties arise in their prediction (Figure 2.1).

2.1.2 Classification

Despite the lack of pattern, loops are not random. Early studies of short turns and β hairpins[1] showed that these loops could be clustered into a few structural classes and their mirror images (Chou, 2000; Richardson, 1981; Sibanda and Thorton, 1985; Venkatachalam, 1968). A later study revealed that homologous proteins have similar loops of equivalent lengths (Greer, 1981).

Protein loop classifications have also been made across all loops (Burke et al., 2000; Donate et al., 1996; Espadaler et al., 2004; Oliva et al., 1997; Vanhee et al., 2011) or within specific protein classes such as antibody complementarity determining regions (Al-Lazikani et al., 1997; Chothia and Lesk, 1987). Most methods are based on local properties, such as anchor regions, the distance between the end points of the loop and the geometrical shape along the sequence (Kwasigroch et al., 1996; Ring et al., 1991; Vanhee et al., 2011; Wojcik et al., 1999) (Figure 2.2).

[1]Hairpins are a specific type of turn which changes the direction of the polypeptide chain. Note that β indicates an anti-parallel β sheet and a β hairpin does not have to be a β turn.

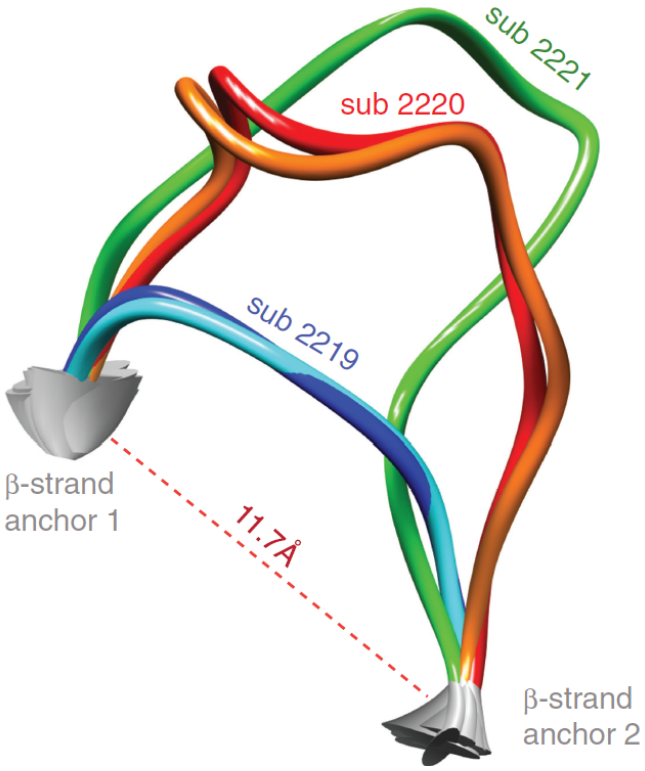


Figure 2.2: Example of protein loop classification
Loop BriX is a database of predefined protein loops. Loops are classified in terms of local properties. A superclass contains loops which have the same anchor type and similar loop end-to-end distances. The superclass is divided into several subclasses according to loop length and structural conformation. The figure was taken from Vanhee et al. (2011) (Figure 2A).

Loops can also be classified in terms of function. There is some evidence that a loop can have local functionality (Figure 2.3). Experiments have been carried out which support the idea that swapping a local loop sequence for a different functional loop sequence allows the new function to be taken on (Pardon et al., 1995; Toma et al., 1991; Wolfson et al., 1991). One common example of functional loop exchange is in the development of humanised antibodies (Queen et al., 1989; Riechmann et al., 1988).

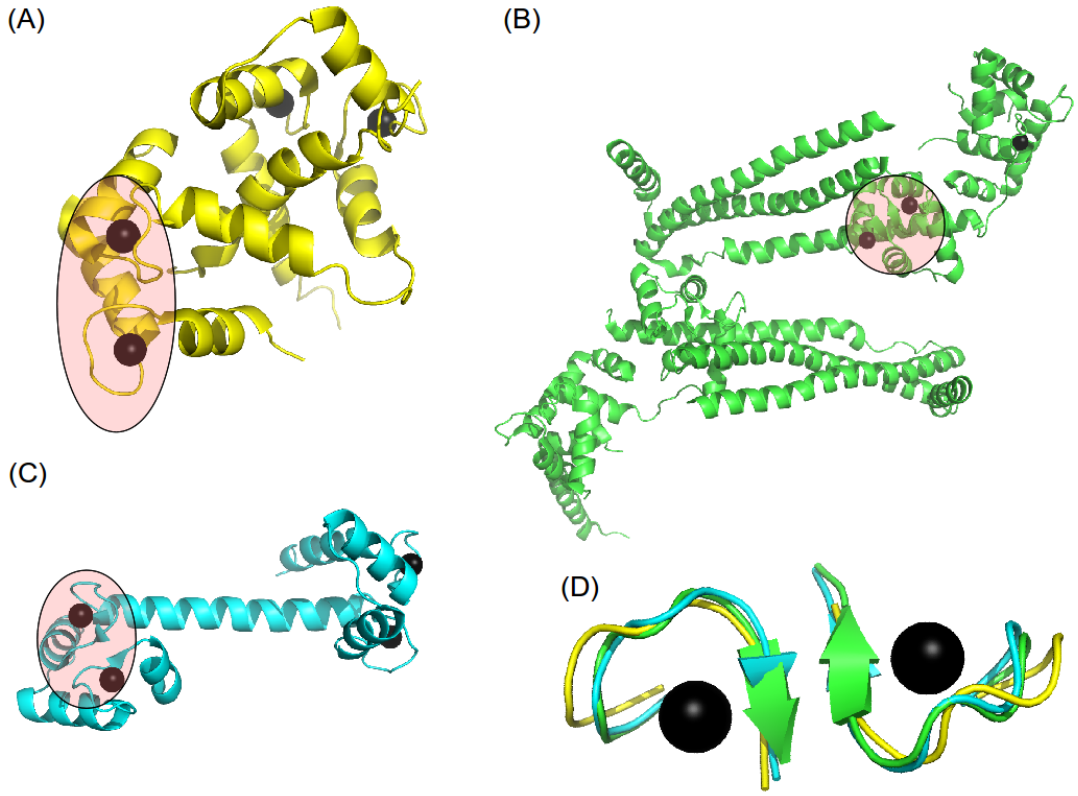


Figure 2.3: Calcium binding loops of EF hands
Although they have different overall structures ((A) 2BBM, (B) 1J1E and (C) 1OOJ), all the calcium binding functional loops look alike (D). Calcium ions are coloured black and the calcium binding regions are coloured in red (A–C).

2.2 Protein Loop Structure Prediction

Protein loop structure prediction generally consists of three stages: sampling, filtering and ranking (Figure 2.4). Candidate loops are sampled while avoiding redundancy and implausible structural conformations. A method then ranks the sampled candidate loops using scoring functions.

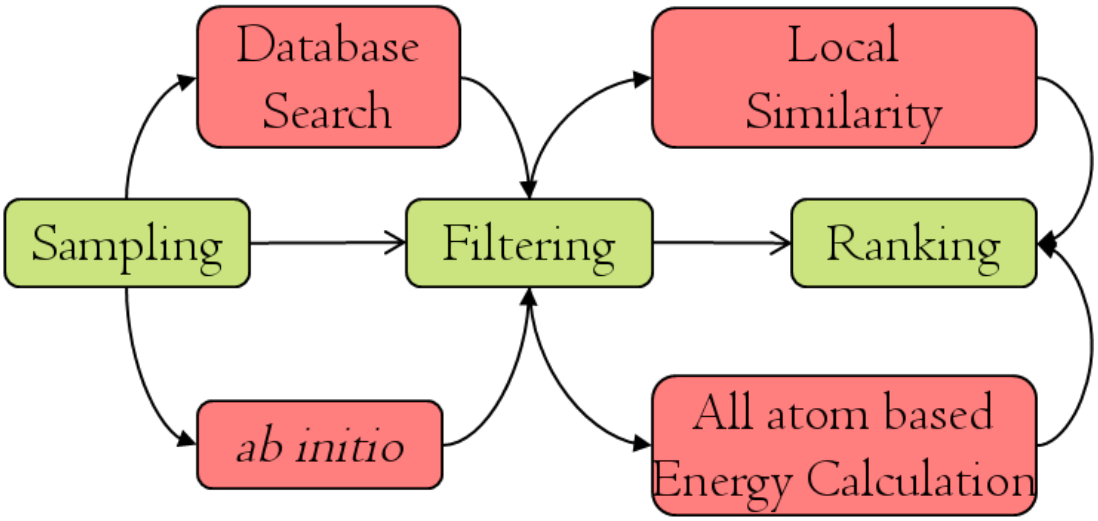


Figure 2.4: A general procedure of protein loop structure prediction
A method can take either a database search or *ab initio* approach. While or after sampling possible loop candidates, the redundancy of sampled structures and implausible conformations are avoided or eliminated in the filtering step. The final model is selected in the ranking step in accordance with scoring functions or local similarities.

2.2.1 Sampling

Database Search Loop Structure Prediction

Database search methods depend upon the assumption that similarities between local properties may suggest similar local structures. All database search methods work in a similar fashion using either a complete set or a classified set of loops and selecting predictions using features including sequence similarity, anchor geometry and some form of energy function. The quality of database search methods is highly dependent on the databases that they use.

Many database search methods use predefined loops in a similar fashion to those explained in section 2.1.2 (Fernandez-Fuentes et al., 2006a,b; Hildebrand et al., 2009; Michalsky et al., 2003; Peng and Yang, 2007; Wojcik et al., 1999). A database can contain entire protein structures without classifying loops in order to take into account ambiguous anchor regions in a modelling situation (Choi and Deane, 2010; Deane and

Blundell, 2001). Candidate loops can also be sampled from an artificially generated database (Deane and Blundell, 2000).

ab initio Loop Structure Prediction

Generally *ab initio* methods are referred to as those which do not use solved protein structure fragments *per se* for loop prediction. Candidate loops are generated and optimised against scoring functions (discussed in more detail in section 2.2.3). There have been numerous *ab initio* methods with different combinations of sampling methods, optimisation techniques and scoring functions[1].

The most common way to sample loop structures in *ab initio* methods is to use dihedral angle (ϕ, ψ) propensities. As backbone bond lengths and angles are nearly fixed, the main degrees of freedom are dihedral angle rotations. Each amino acid has distinct dihedral angle propensities and a loop is sampled based on the propensities of the sequence (Figure 2.5). Instead of using the standard dihedral angle set (ϕ, ψ), Sucha et al. (1995) and Spassov et al. (2008) proposed $(\psi_i,\ \phi_{i+1})$ dihedral angles to reflect neighbourhood residues.

The sampled fragment may well not connect the anchors. Hence a step to close the gap is required. The process is called loop closure. Loop closure is a mathematical technique to find dihedral angle rotations that steer a polypeptide chain to a desired position. Each dihedral angle is optimised in turn (Canutescu and Dunbrack Jr., 2003) or all dihedral angles alongside the loop are optimised at the same time (Hurst, 1994;

[1]High-temperature molecular dynamics (Bruccoleri and Karplus, 1990), simulated annealing optimisation (Higo et al., 1992), local interaction for a fast search (Finkelstein and Reva, 1992), CHARMM energy optimisation of fragments sampled using database search (van Vlijmen and Karplus, 1997), Generalised Born and AMBER potential energy (Rapp and Friesner, 1999), CHARMM22 force field with simulated annealing (Fiser et al., 2000), a colony energy term based on largely populated samples (Xiang et al., 2002), Monte Carlo simulated annealing (Rohl et al., 2004), OPLS-AA and the Generalised Born model (Jacobson et al., 2004), CHARMm with the Generalised Born model (Spassov et al., 2008) and fragment assembly accompanied with an analytic loop closure algorithm (Lee et al., 2010). In fact, some methods listed here do use database fragments as an initial model. The backbone atom coordinates of the initial model are re-positioned during optimisation. Therefore, more precisely, database search methods are referred to as those which use database fragments without further modifications to backbone atom coordinates.

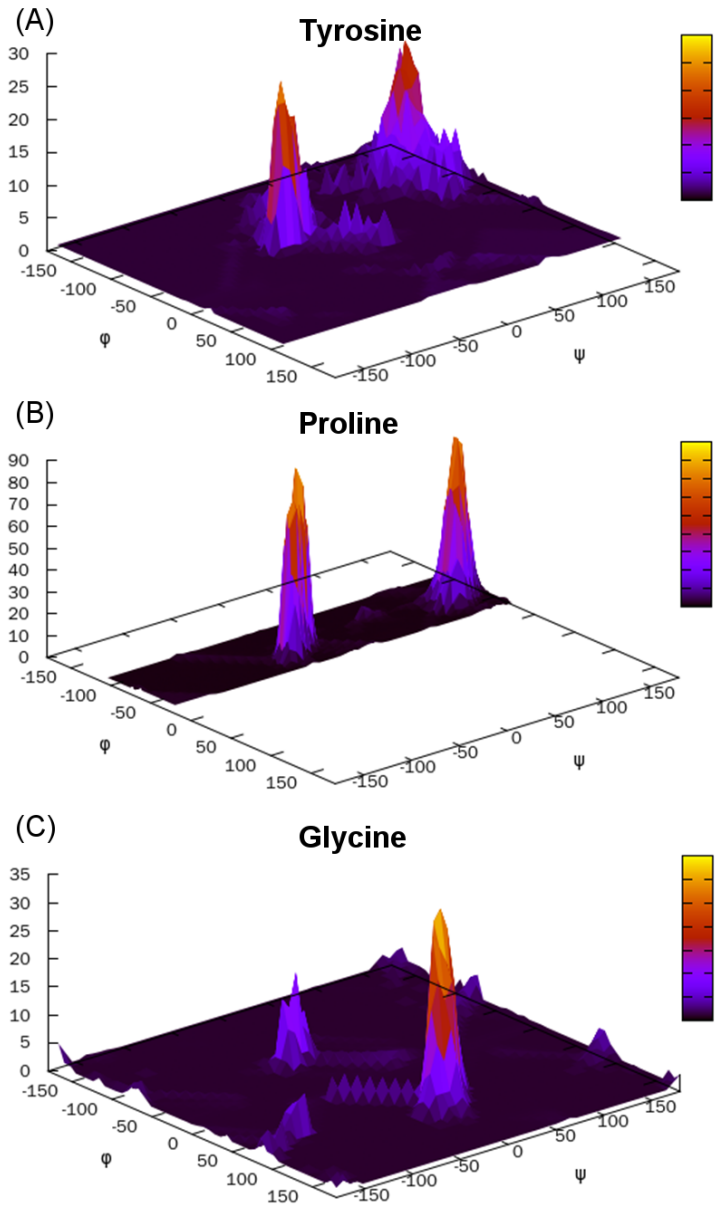


Figure 2.5: Heat maps of dihedral angle (ϕ, ψ) propensities of amino acids Each amino acid has different dihedral angle propensities. The regions in brighter colours are more probable. Generally amino acids show a propensity as in (A). (B) The sidechain of proline restricts dihedral rotations within limited ϕ angles at approximately -75°. (C) Glycine has only one hydrogen atom in the R group and more freedom of rotation. Positive ϕ angles are rare except for glycine.

Lee et al., 2010; Shenkin et al., 1987).

Table 2.1: The van der Waals radii of backbone atoms (Li and Nussinov, 1998)

Atom	Radius (Å)
Nitrogen	1.70
α Carbon	1.90
Carbon	1.75
Oxygen	1.49

2.2.2 Filtering

While loops are sampled, fragments which create steric clashes are removed in both database search and *ab initio* methods. The steric clashes can be checked using van der Waals interaction energy term. Typically the Lennard-Jones potential energy is used for the van der Waals interaction E_{vdw}.

$$E_{vdw} = \sum_{i<j} 4\epsilon \left(\frac{\sigma_{ij}}{r_{ij}^{12}} - \frac{\sigma_{ij}}{r_{ij}^{6}} \right), \tag{2.1}$$

where ϵ is the depth of the potential energy well, σ_{ij} is the collision diameter of the atom i and j, and r_{ij} is the distance between the two atoms.

In order to reduce computational time, pre-calculated van der Waals radii are frequently used. Each atom is considered to be a soft sphere with a defined radius (Table 2.1). Steric clashes are checked using these pre-calculated radii values allowing overlaps. High-resolution X-ray protein structures typically have no more than 30% overlaps between two atoms (Jacobson et al., 2004).

Due to the dihedral angle propensities, for a given loop sequence, *ab initio* methods may produce many highly similar structure conformations. For faster optimisation, the redundancy can be removed using a simple structure similarity cut-off value (DePristo et al., 2003) or a clustering algorithm (Jacobson et al., 2004). Database search methods use local properties as filters during sampling. Common filters are local sequence similarities with anchor matches.

2.2.3 Ranking

Even if near-native structures are obtained during the sampling and filtering steps, ranking is needed to select the best structure among the sampled models. The models are ranked using scoring functions. The basic assumption is that the native structure is in the global minimum of such scoring functions.

The scoring functions are in general categorised into three types: physics based energy function, statistical potential function and local similarity measures. A method can have one or more scoring functions. However, as *ab initio* methods do not use fragments from databases, local similarity measures are used only by database search methods.

Physics Based Energy Function

A general form of physics based energy function can be written as follows.

$$E_{\text{total}} = E_B + E_\theta + E_\phi + E_{vdw} + E_e. \tag{2.2}$$

The first three terms are contributions from internal energies. E_B is the bond length energy and E_θ is the bond angle energy. The third term E_ϕ is specifically for the backbone conformation (dihedral angles). The last two terms are non-bonded interaction energies. E_{vdw} is the van der Waals energy term described in eq. 2.1. The last term represents the electrostatic interactions of noncovalent bonds,

$$E_e = \frac{1}{4\pi\epsilon_0} \sum_{i<j} \frac{q_i q_j}{r_{ij}}, \tag{2.3}$$

where $\epsilon_0 = 8.854 \times 10^{-12} C^2 J^{-1} m^{-1}$ is the permitivity and q_i is the charge.

The parameters in the above terms are empirically determined. There are several packages for calculating this type of classical molecular dynamics force fields for macromolecules like proteins. The two most widely used energy functions are AM-

BER (Assisted Model Building with Energy Refinement) (Cornell et al., 1995) and CHARMM (Chemistry at HARvard Macromolecular Mechanics) (Brooks et al., 1983; MacKerell et al., 1998). OPLS-AA (Optimized Potentials for Liquid Simulations for All Atoms) (Jorgensen and Tirado-Rives, 1988) is a direct descendant of AMBER. All the force field energies are similar to one another except for parametrisations.

Soluble proteins *in vivo* are surrounded by many other molecules such as water. Precise energy can be calculated when the exact number of water molecules involved in the calculation (the explicit water model) are known. However, due to the lack of such knowledge and its computational expense, a mean force field energy of water molecules is popularly used (the implicit water model). In the implicit water model, a set of water molecules is regarded to make a mean force field. The Generalised Born (GB) model (Still et al., 1990) is an approximated implicit water model based on the assumption that soluble proteins in general are globular and can be regarded as spheres with dielectric charges. The GB model is written as

$$E_{GB} = \frac{1}{8\pi}\left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \sqrt{\alpha_i\alpha_j}\exp\left(-\left(\frac{r_{ij}}{2\sqrt{\alpha_i\alpha_j}}\right)^2\right)}}, \tag{2.4}$$

where ϵ_0 is the permittivity of vacuum space and ϵ is the dielectric constant of the solvent (in this case, water), q_i is the charge, r_{ij} is the distance between i and j atoms, α_i is the effective Born radius which describes how the atom is apart from the surface. If α_i and α_j are sufficiently small ($\alpha_i \simeq r_i$, i.e., they are on surface), the equation 2.4 is approximately the same as the electrostatic interaction. However, if they are large (buried), the energy contribution becomes smaller.

Statistical Potential Function

A statistical potential or knowledge-based potential is a score function which describes pairwise preferences of structural features in proteins. The statistical potential is based

on the assumption that the most probable protein structure of a given sequence can be sought using a large number of sequences with solved structures.

Note that there was a debate on the interpretation of the statistical potential. Sippl (1990) interpreted the statistical potential as a consequence of Boltzmann's principle (the potential of mean force).

$$\Delta F = -k_B T \ln \frac{Z}{Z_R}, \tag{2.5}$$

where ΔF is the free energy difference, k_B is the Boltzmann constant, T is the temperature, Z is the partition function and Z_R is the partition function of the reference states (native structures).

He assumed that native structures are in the lowest energy states under equilibrium and high probabilities of intra- and intermolecular atomic interactions in the native structures are related to actual physical quantities, such as energy, volume and temperature. The interpretation was broadly disputed (Ben-Naim, 1997; Moult, 1997) as the measurable properties from structure databases are not such physical properties and a similar equation can be obtained from a statistical theory without the assumption. Generally, the statistical interpretation is widely accepted.

The general form of statistical potentials is written using Bayesian inference (Simons et al., 1997).

$$Pr(X|A) = Pr(X) \cdot \frac{Pr(A|X)}{Pr(A)} \propto Pr(X) \cdot Pr(A|X), \tag{2.6}$$

where $Pr(X|A)$ is the conditional probability X given A and proportional to the likelihood $Pr(A|X)$ multiplied by the prior $Pr(X)$. The likelihood is assumed as a product of pairwise probabilities.

$$Pr(A|X) \simeq \prod_{ij} Pr(x_{ij}|A) \propto \prod_{ij} \frac{Pr(x_{ij}|A)}{Pr(x_{ij})}. \tag{2.7}$$

Generally the score $S(X)$ is the negative logarithm of the likelihood.

$$S(X) = -\sum_{ij} \ln\left(\frac{Pr(x_{ij}|A)}{Pr(x_{ij})}\right). \tag{2.8}$$

There are two reasons for the negative logarithm expression: 1) the formalism is the same form as the potential of mean force and 2) computationally easier to calculate.

The pairwise probability $Pr(x_{ij})$ can be from residue pairs (Miyazawa and Jernigan, 1996), all atom distances (Samudrala and Moult, 1998b; Zhou and Zhou, 2002), hydrophobic contribution (Narang, 2006) and dihedral angles (Rata et al., 2010) (See an example in Section 3.2.4).

Local Similarity

As *ab initio* methods do not directly use fragments from structure databases, ranking by local similarity measures is employed only by database search methods.

The most common and simplest ranking measure is anchor matches between the target and database anchor structures (Deane and Blundell, 2001). In general, database search methods rank predicted fragments using a combination of other local properties accompanied with anchor matches; sequence similarity measured by substitution score matrices (Choi and Deane, 2010; Fernandez-Fuentes et al., 2006a; Hildebrand et al., 2009; Michalsky et al., 2003), dihedral angle propensities of the predicted loop structures (Fernandez-Fuentes et al., 2006a) and artificial neural network (Peng and Yang, 2007), etc.

2.3 Measurement of Prediction Accuracy

"Accuracy" in protein structure prediction typically means how similar a predicted structure is to its native structure. One of the most popular and oldest measures is the root mean square deviation (RMSD).

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Tx_i - y_i)^2}, \tag{2.9}$$

where $(x_1, y_1), \cdots, (x_n, y_n)$ are coordinate sets of equivalent atom positions of the two structures and T is the transformation which minimises the RMSD value.

Although RMSD is arguably the most convenient and widely used measure, it has several problems and limitations. One of the problems is its dependence on the structure size. It is well known that RMSD increases proportionally to the radii of gyration (Hu et al., 1997). In CASP experiments, the GDT (Global Distance Test) score is used to evaluate how similar two structures are. In GDT-TS (GDT Total Score), a predicted structure is optimally superimposed onto its native structure and each atom distance is binned in several cutoff values (1, 2, 4 and 8Å) (Zelma et al., 2001). MAXSUB is an alternative measure for the structure similarity which gives a normalised value (0 to 1) (Siew et al., 2000).

However, both GDT and MAXSUB still show the size (or length) dependency (Zhang and Skolnick, 2004). In order to overcome the limitation, Zhang and Skolnick (2004) developed a size-independent similarity measure called TM-Score (Template Modelling Score).

$$\text{TM-Score} = max\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\left(\frac{d_i}{d_0}\right)}\right], \tag{2.10}$$

where n is the number of residues of two structures identical in length, d_i is the pairwise distance between corresponding residues and d_0 is a scale factor. The formalism of TM-Score and MAXSUB is nearly identical as they both came from the Levitt-Gerstein score (LG score) (Levitt and Gerstein, 1998). The main difference between the scores is d_0 (5Å for the LG score and 3.5Å for MAXSUB). TM-score takes $d_0 = 1.24\sqrt[3]{n-15} - 18$.

In protein loop structure prediction, RMSD is a conventional accuracy measure

despite its limitation. Although Zhang and Skolnick (2004) claimed that TM-Score is a length-independent similarity measure, it was tested only for proteins which are longer than typical loops (>30 residues in length) and which have regular secondary structures. Accuracy comparisons between loop structure prediction methods are meaningful when loops of the same length are predicted and compared.

The distribution of loop structure prediction accuracy is unknown as it depends on prediction methods. Typically, it is conventional to assume that the prediction accuracy is distributed normally. The average RMSD value is a conventional measure when a large scale benchmark is performed (Ko et al., 2011). Or individual prediction results can be compared (Karen et al., 2007).

In this dissertation, the measurement of accuracy is RMSD after superimposing anchor structures (global loop RMSD). All backbone atoms (N, Cα, C and O) are considered. For the large scale benchmarks performed in Chapter 3 and 4, the conventional average RMSD values are used to compared accuracies between prediction methods.

Chapter 3

FREAD: Protein Loop Structure Prediction Using a Database Search Algorithm

3.1 Introduction

Karen et al. (2007) benchmarked four commercial modelling protocols for loop structure prediction (two *ab initio*: MODELLER (Fiser et al., 2000) and PRIME – a commercial version of PLOP (Jacobson et al., 2004) and two database search: SYBYL (Tripos Inc., 2005) and ICM (Abagyan et al., 1994)). They reported that the *ab initio* methods performed better. Indeed, some published database search methods assert only that they are able to predict loops to the same level as *ab initio* (Michalsky et al., 2003; Peng and Yang, 2007). In general, *ab initio* loop prediction has been thought to be more effective (Fidelis et al., 1994; Lessel and Schomburg, 1999).

However, as the protein structure database has expanded dramatically but the number of entirely new folds has grown slowly, the coverage of loop structures has increased far faster. Fernandez-Fuentes and Fiser (2006) demonstrated that structure

fragments are almost saturated up to 12 residues in length. This observation encourages the re-evaluation of the predictive power of database search in loop modelling.

Here I describe an updated version of a database search algorithm, FREAD (Deane and Blundell, 2001). It is benchmarked against three *ab initio* methods, MODELLER, RAPPER (de Bakker et al., 2003; DePristo et al., 2003) and PLOP.

Two test sets were built for the benchmark; (1) the standard benchmark that uses native protein structures and (2) the CASP benchmark that uses models as the targets.

The standard benchmark examines the predictive ability of all the four protocols and evaluates the effect of the original FREAD's selection criteria on prediction. A set of selected 510 loops ranging from 4 to 20 residues in length (30 loops at each length) was used in the benchmark. Original FREAD is found to be the most accurate of the protocols tested in the standard benchmark. For example, at loop length 8, original FREAD predicts loops to 2.88Å RMSD on average (RAPPER: 2.88Å, MODELLER: 4.25Å and PLOP: 4.34Å).

The results of the original FREAD algorithm can be significantly improved by simply changing the substitution score cut-off. The new version of FREAD with a stricter cut-off decreases the number of predictions (by roughly a third for shorter loops and a half for longer loops). However, this is compensated for by the higher quality of prediction. The average global RMSD value (at length 8) of the predicted loops by FREAD with the stricter cut-off is 1.15Å (the number of predictions decreased from 30 to 18). The quality holds regardless of loop length (e.g., FREAD with the stricter cut-off still predicts to below 2Å for length 20 loops).

A more realistic test of the power of the protocols was performed using the best template based models (TBM) from CASP 7 and 8. During the eight rounds of CASP, homology modelling has been revealed as the most powerful method for computational structure prediction. Two hundred and twelve loops ranging from length 3 to 20 were defined on the models. All the modelling protocols including the new version of FREAD

were tested on this benchmark. In this benchmark, anchor structures are not correctly given and therefore the predictions are relatively worse than those in the standard benchmark where the correct anchor structures are given. To improve the coverage of FREAD, when FREAD fails to give a prediction, FREAD extends the target loop length. This process is possible as FREAD gives consistent prediction accuracy regardless of loop length. FREAD with the extended search was able to predict 127 of the 212 loop targets. It performed better than the original CASP models and the other protocols' predictions in 61 of the CASP targets.

3.2 Materials and Methods

3.2.1 Benchmark Test Sets

Standard benchmark test set

Chains of high resolution X-ray determined protein structures were selected using PISCES (Wang and Dunbrack, 2003) under the following criteria: Sequence identity percentage $\leq$90%, resolution $\leq$2Å, R-factor $\leq$0.2. The PDB chains were cleaned and annotated using JOY (Mizuguchi et al., 1998a). A loop structure was defined as a region between two secondary structures. These secondary structures must be at least three residues in length (Donate et al., 1996). Short (less than four residues) and long loops (longer than 20 residues) were discarded.

Pairwise sequence alignments of identical length loops were performed. If the sequence identity between two loops was greater than 40% (Fernandez-Fuentes and Fiser, 2006), the sequence with the lower B-factor was kept while the other was discarded. Two randomly selected sets were generated, each with 30 test loops of every length. Set one (Appendix Table B.1) was used to test all the methods and parametrise a new version of FREAD. Set two was used to check that overfitting of FREAD with the new parameter to set one had not occurred (a full list of Set Two is given in Appendix

Table B.2). In the standard benchmark, the native structures excluding their native loop coordinates were given to the modelling protocols.

CASP Benchmark Test Set

For each TBM target in CASP 7 and 8, the best models according to the LGA score (Kryshtafovych et al., 2007; Zelma, 2003) were selected. "Loop" regions were defined on these models. As the target template alignments that produced the CASP models were not given, the loop regions are assumed to be those which are predicted poorly. These are clearly not pure loops, but the segments do correspond to the standard operational problem of predicting the structurally variable regions of the model. As none of the modelling methods being benchmarked use databases apart from FREAD, these regions being perhaps "non-loop" should not affect their performance. If the distance between C_α atoms of a common residue in a model and its native structure is bigger than 5Å after global superposition, the region is defined as a "loop". To increase the numbers in the test set, three residue loops were also included. Short (less than three residues) and long loops (greater than 20 residues) were discarded (a full list is given in Appendix Table B.3).

Two hundred and twelve loops were defined in total ranging from 3 to 19 residues in length (no loops were found at length 20). Most loops were less than eight residues in length (151 out of 212).

3.2.2 Measurement of Accuracy

The measurement of accuracy used here is global RMSD of the main chain atoms (nitrogen, α carbon, carbon and oxygen), where the loop itself is not superimposed but the anchors are (two residues at each end of the loop).

To assess the ranking ability of the protocols, the top prediction (the model ranked as one by the modelling protocol) was also compared to the best prediction (the lowest

RMSD model in the final filtered set). The best prediction indicates the sampling power of a modelling protocol. The RMSD difference between the top loop model and the best loop model of a protocol gives an indication of its ranking ability.

3.2.3 *ab initio* Protocols

In this benchmark, FREAD was benchmarked against three widely used *ab initio* methods: MODELLER, RAPPER and PLOP.

MODELLER builds an initial model with uniformly spaced loop atoms on a straight line between N and C-anchors of the loop. The atom positions are randomized and relaxed until they reach protein-like structures. MODELLER optimises the initial models using pseudo-energy terms based on the CHARMM22 force field and statistical preferences of stereochemical features (i.e., bond lengths, bond angles and dihedral angles). The optimisation consists of three steps; conjugate gradient, molecular dynamics (MD) with simulated annealing and conjugate gradient again. In each step, loop atoms and other atoms including the neighbouring anchor atoms of the loop are separately optimised.

MODELLER offers several MD levels. The levels are divided in terms of its cooling speed. Empirically, the slowest MD level does not show significant improvement for shorter loops. In Fiser et al. (2000), a good compromise between accuracy and performance was achieved by building 50 loop decoys. I followed this procedure with "very slow" MD. The 50 decoys generated are ranked using a statistical potential, called DOPE (Shen and Sali, 2006).

RAPPER and PLOP build loops in a similar manner. RAPPER samples fragments from the N to C-anchor of the loop while PLOP generates fragments from the N and C-anchors simultaneously. They both build the loop by sampling from fine-grained dihedral angle libraries for each residue while avoiding steric clashes using the approximated van der Waals spheres of atoms. In a second step, they iteratively try to close

the loop (complete the fit to the anchors).

The RAPPER programme is not a full prediction protocol since it does not have a ranking method. In one of the series of articles (de Bakker et al., 2003), RAPPER decoys were scored using AMBER/GBSA force field. In Zhang et al. (2004), DFIRE shows similar discriminatory power to the MD function. In this benchmark, RAPPER was used as a sampler and a final model was selected using DFIRE. One thousand decoys per loop were generated for the RAPPER predictions following de Bakker et al. (2003) and DePristo et al. (2003).

The number of decoys generated by PLOP is not directly controllable as the generated decoys are clustered using the K-means algorithm to avoid redundancy. The number of clusters is set to be four times the number of loop residues. PLOP has an optional parameter, the hard sphere overlap factor (ofac). The ofac is the distance ratio between two atoms. In this benchmark, the ofac is set to be 0.7 (up to 30% of van der Waals radii overlap is allowed). PLOP also provides a crystal packing option (sym xtal). However, in Jacobson et al. (2004) and Karen et al. (2007), the crystal packing option did not show significant improvement in backbone conformations, but in sidechains. The crystal packing option was not introduced in this benchmark as the measurement of accuracy takes only into account backbone structures. The final model is chosen using the all-atom OPLS force field with the Surface Generalized Born model.

3.2.4 FREAD Algorithm

FREAD uses four main filters (Figure 3.1). First, a fast database search is performed using anchor Cα separations. Second, an environment specific substitution score filters out fragments that have low sequence similarities to the target loop. The environment specific substitution table was constructed by observing pairwise sequence alignments between loops in homologous proteins. The score table is a quantitative measure for the probability that an amino acid is substituted for another amino acid in a certain

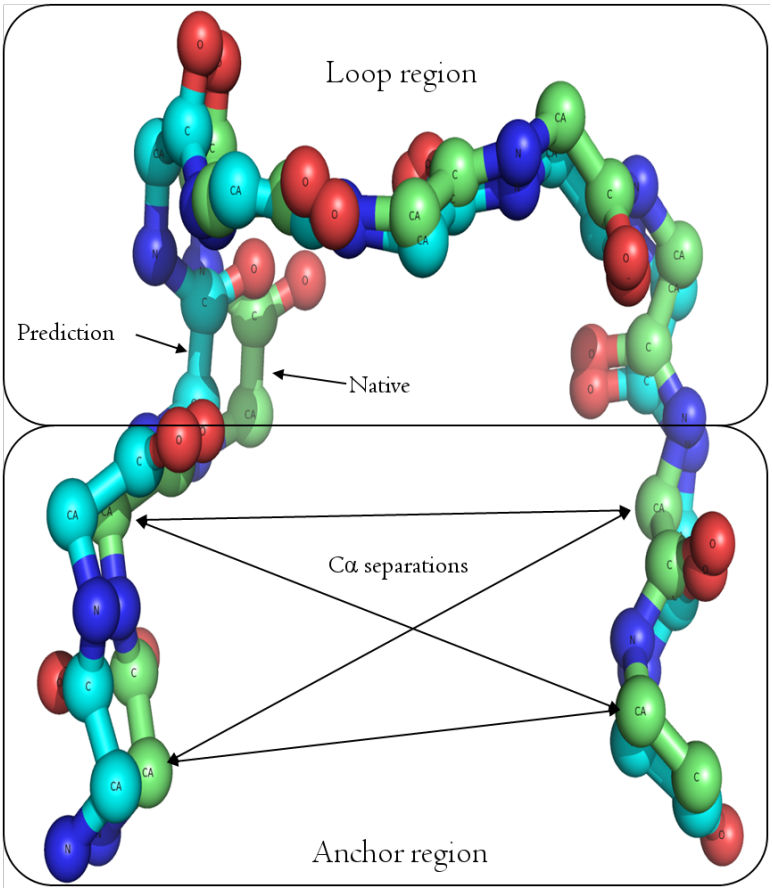


Figure 3.1: The FREAD algorithm
The FREAD algorithm consists of four steps. (1) Possible loop candidates in a database are sampled by checking the C_α separations of the anchor regions. (2) Loops whose sequences are dissimilar to the query loop sequence are filtered out (the sequence similarity measure is the environment specific substitution score (Section 3.2.5)). (3) The second filtering step using a statistical energy function removes implausible loop candidates. (4) The remaining loops are ranked in terms of the anchor RMSD.

environment (Lee and Blundell, 2009; Shi et al., 2010) (more details in section 3.2.5). Third, a statistical energy function (Samudrala and Moult, 1998a) filters out implausible loop decoys. Finally, the anchor RMSD, the RMSD between the target and the predicted structures for two residues either side of the loop, is calculated. All predicted loops are ranked in terms of the anchor RMSD.

FREAD Databases

The quality of FREAD prediction depends crucially on its database. The database is different from many traditional loop prediction databases (Donate et al., 1996; Espadaler et al., 2004; Hildebrand et al., 2009; Michalsky et al., 2003). In the FREAD database, all fragments of the protein are included, not just predefined loop regions. The database contains information of inter Cα distances up to 26 residues and their residue indices. When the anchor structures of a query loop are given, FREAD firstly seeks out matched fragments using the anchor Cα separations in the database (0.7Å cutoff).

The database for the standard benchmark was created using only X-ray determined protein structures selected by PISCES under the criteria: sequence identity $\leq$95%, resolution $\leq$2.7Å, R-factor $\leq$0.3. FREAD takes the coordinate information from the PDB files and builds a database containing C_α separations.

As this is a fairly comprehensive database, it is possible that the test loop structures are within it. FREAD may show the target structure itself as a prediction. In the standard benchmark, if FREAD predicts a fragment from the query, the structure is discarded. In this case, the second ranked prediction is regarded as the top prediction.

Additionally a database including only loop structures with their anchor regions was constructed (The loop definition used here is the same as that in the standard test set). If two such segments share 100% sequence identity, the sequence with the lower B-factor is stored in the database.

Different databases were used for the CASP model benchmark. For the CASP targets, only protein structures released before the CASPs started (For CASP7, before May 2006 and CASP8, before May 2008) were used.

Statistical Potential

The second and third filters of FREAD are the environment specific substitution score (ESSS) and a statistical potential function. As ESSS turns out to be the best indicator for accurate prediction, further details of ESSS are given in Section 3.2.5. The algorithm of the statistical potential FREAD uses was developed by Samudrala and Moult (1998a).

$$Pr(d_{ab}^{ij}|C) = \frac{Pr(d_{ab}^{ij}) \cdot Pr(C|d_{ab}^{ij})}{Pr(C)}, \tag{3.1}$$

where C is a "correct" structure and d_{ab}^{ij} is the distance between atoms i and j of amino acids a and b. The distance is divided into 19 bins ($d \leq$3Å, $d >$20Å and $n < d \leq n+1$ where $n = 3, 4, \cdots, 20$).

The score function is the probability of all atom pairwise distances that the structure is correct given a set of distances ($Pr(C|\{d_{ab}^{ij}\})$). Here, all the pairwise distances are approximated to be independent. Therefore the total probability is given as a multiple product of the probabilities of each pairwise distance. It is often practical to deal with sums than products by taking logarithms, especially when small numbers are involved in the product.

$$S(\{d_{ab}^{ij}\}) = -\sum_{ij} \ln \frac{Pr(d_{ab}^{ij}|C)}{Pr(d_{ab}^{ij})} \propto -\ln Pr(C|\{d_{ab}^{ij}\}). \tag{3.2}$$

In order to calculate the score, we need frequency distributions for all combinations of atom types ($Pr(d_{ab}|C)$ and $Pr(d_{ab})$). The distributions are directly calculated from X-ray determined high resolution PDB structures chosen under the criteria: sequence identity $\leq$ 30%, resolution $\leq$ 2.0Å and R-factor $\leq$ 0.2. This set contains 3865 PDB chains.

$Pr(d_{ab}|C)$ is given as follows.

$$Pr(d_{ab}|C) = \frac{N(d_{ab})}{\sum_d N(d_{ab})}, \tag{3.3}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a distance bin d. For example, if one observes that the distance between Cα of a glycine residue and Cβ of histidine in a structure is 6.2Å, the probability of this pair type is calculated using the frequency distributions. If there are 10 observations in the database between 6 and 7Å, and the total number of the pair in all the 19 bins is 100, then the probability is 0.1.

$Pr(d_{ab})$ is the probability of being the distance d in any atom types

$$Pr(d_{ab}) = \frac{\sum_{ab} N(d_{ab})}{\sum_d \sum_{ab} N(d_{ab})}. \tag{3.4}$$

Any intra-residue distances are discarded in the score summation and the frequency distributions.

Melding

Fragments chosen after the above filters are ranked in terms of anchor RMSD (the maximum cutoff value is 1Å). The loop fragments predicted by FREAD are melded onto the anchor structures. First the prediction and target anchor structures are superimposed (two residues for each anchor). The melding method used is identical to that used by SYBYL (Blundell et al., 1988; Tripos Inc., 2005). The melded coordinate M_i is

$$M_i = \frac{1}{n+1}(D \cdot P_i + ((n+1) - D) \cdot N_i), \tag{3.5}$$

where D is an integer given by the residue separation from the loop terminus to the anchor residue in question (in this case, 1 or 2), n is the number of anchor residues considered (n is set to be 2), P_i is the coordinate of anchor residues of the predicted

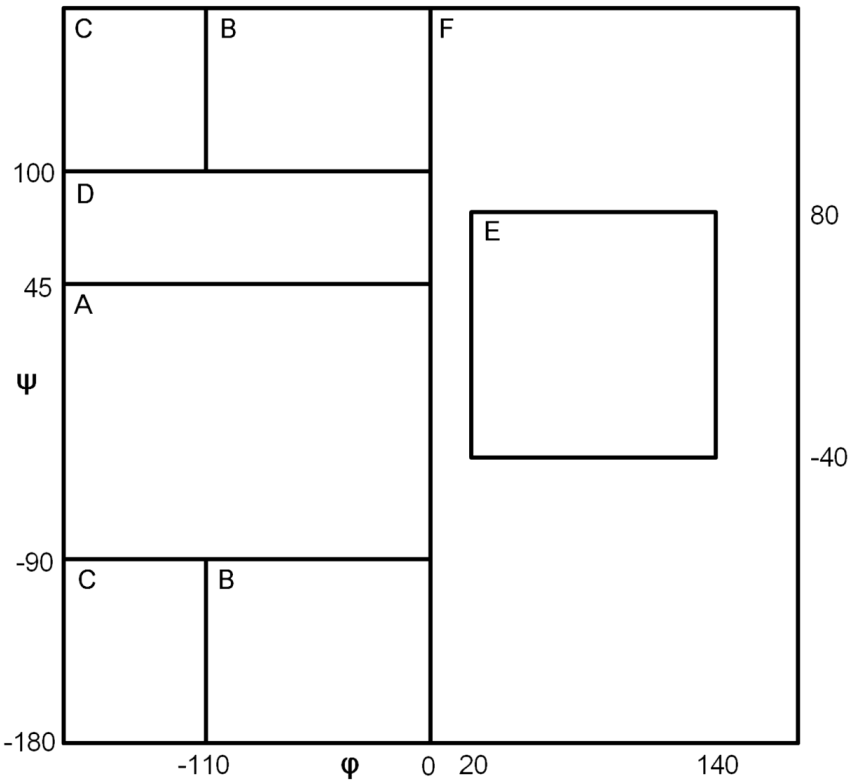


Figure 3.2: The dihedral angle areas defined for the dihedral angle specific substitution score
The six areas for the environmentally constrained substitution tables (described by the dihedral angle regions) are shown on the plot labelled A to F.

loop, N_i is the coordinate of the anchor residues of the target structure.

3.2.5 Environment Specific Substitution Score

The environment specific substitution score tables of FREAD were constructed by observing frequencies of amino acid substitutions in the loops of homologous protein pairs extracted from the HOMSTRAD database (Mizuguchi et al., 1998b). FREAD uses "environments" described by the six dihedral angle areas ($\{\phi, \psi\} \rightarrow \{(180, 180) \cup (180, 180)\}$: A) {(-80, 0)∩(90, 45)]; B) {(110, 0)∩{(180, 90)∪(100, 180)}}; C) {(-80, 110)∩{(180, -0)∪(100, 180)}}; D) {(-80, 0)∩(80, 100)}; E) {(20, 140)∩(-0, 80)}; F) the rest) (Figure 3.2).

To reduce multiple contributions from closely related members, sequences are clustered in a way similar to that used in the construction of BLOSUM matrices (Henrikoff and Henrikoff, 1992). Clustering is performed for the entire sequence using the overall percentage identities. For example, if the percentage threshold is set at 60% and the alignment contains three sequences A, B and C with the pairwise percentage identities between A and B being 70%, between A and C being 30% and between B and C being 30%, then A and B are clustered together. Then the contributions from the pairs (A, C) and (B, C) are averaged. For example, suppose there are three sequences A, B and C, and A and B are identical except for one residue (a and b respectively). If the corresponding residue in C is c, then $a \rightarrow c$ and $b \rightarrow c$ are counted as 1/2. Once this has been done, the raw substitution counts N_{ab}^{E} are calculated to be the frequency of amino acid a in environment E replaced by amino acid b. The probability is then given by

$$Pr(b|a,E) = N_{ab}^{E} / \sum_{c} N_{ac}^{E}, \tag{3.6}$$

where $P(b|a,E)$ is the probability that amino acid a in environment E is substituted by amino acid b. FREAD uses constrained tables, i.e. $N_{ab}^{E} = N_{ba}^{E}$. The substitution score (log-odds) for environment E is then created by

$$\begin{aligned} s(a, E \rightarrow b) &= \log\left(\frac{Pr(b|a,E)}{Pr(b)}\right) && (3.7) \\ &= \log\left(\frac{N_{ab}^{E} / \sum_{c} N_{ac}^{E}}{\sum_{a,E} N_{ab}^{E} / \sum_{a,b,E} N_{ab}^{E}}\right) && (3.8) \end{aligned}$$

where $s(a, E \rightarrow b)$ is the score for amino acid a in environment E replaced by amino acid b and q_b is the background probability of observing amino acid b. If the substitution matrices are used for comparing structure with sequence, the score matrices must

represent the odds ratio of the pair (a in structure, b in sequence) occurring in an alignment, as opposed to this match occurring by chance. The background probabilities, therefore, must be those for observing each amino acid residue in a sequence and are given by

$$Pr(b) = \sum_{a,E} N_{ab}^{E} / \sum_{a,b,E} N_{ab}^{E}. \tag{3.9}$$

Generally $s(a, E \rightarrow b) \neq s(b, E \rightarrow a)$. However, if each amino acid in a sequence is assumed to be in the same environment as that of the aligned residue in the structure, the background probability is then given by

$$Pr(b, E) = \sum_{a} N_{ab}^{E} / \sum_{a,b} N_{ab}^{E}. \tag{3.10}$$

This produces a symmetric matrix, i.e. $s(a, E \rightarrow b) = s(b, E \rightarrow a)$. Elements of the log-odds matrices are multiplied by a scaling factor of $3/\log 2$ and rounded to the nearest integer value (i.e., the log-odds scores are expressed in 1/3 bit units) (Deane and Blundell, 2001).

The total ESSS of a target loop is a linear sum of the score 3.8. For example, if the sequence of the loop is $ACDEF$ and that of a matched fragment is $STVWY$ (suppose all the residues of the fragment are in the dihedral angle area A), the total score is $1 - 7 - 6 - 5 + 4 = -13$ (See Table B.4).

Note that the formalism of the score functions in eq. 3.2 and 3.8 are essentially no different as they are based on Bayesian inference. The statistical potential measures statistical preferences of atom-pair distances and ESSS scores residue-pair mutations. The two scoring schemes assume that each pair event (atom-pairs or residue-pairs) is independent and the total probability is a product of individual pair probabilities. Hence the logarithmic formalism is a convenient choice to convert the multiplication into addition.

Table 3.1: The results of the four loop modelling protocols on the standard dataset

	MODELLER				RAPPER				PLOP				FREAD			
Length	Best	σ	Top	σ	Best	σ	Top	σ	Best	σ	Top	σ	Best	σ	Top	σ
4	1.13	0.61	1.73	1.42	**0.62**	0.41	***1.10***	0.77	1.62	1.37	1.79	0.98	0.72	0.88	1.29	1.14
5	1.51	0.57	2.30	1.29	**0.76**	0.35	***1.23***	0.57	2.56	1.31	2.76	1.45	0.78	0.56	2.19	2.02
6	1.61	0.57	2.38	1.67	0.89	0.79	1.92	1.67	2.71	1.73	3.25	1.88	**0.79**	0.54	***1.79***	1.37
7	1.93	0.80	3.44	2.35	1.18	0.52	2.60	1.26	3.31	1.93	3.73	1.82	**1.16**	0.97	***2.53***	2.34
8	2.19	0.80	4.25	2.63	1.40	0.72	***2.88***	1.62	4.01	2.89	4.34	2.03	*1.33*	0.99	2.88	2.37
9	2.60	0.70	4.31	2.31	**1.68**	0.66	***3.03***	1.18	4.15	1.94	5.58	1.93	1.69	1.23	3.08	2.60
10	2.60	0.65	5.69	4.03	**1.97**	0.75	***3.90***	1.84	4.55	3.71	6.41	3.64	1.97	1.48	4.25	3.58
11	2.96	0.96	5.34	2.64	**2.30**	1.22	4.63	2.58	4.76	2.35	6.52	2.80	2.61	1.51	***4.55***	3.63
12	3.54	0.84	7.18	3.73	2.74	1.04	5.10	2.37	6.09	3.92	6.86	3.12	**2.13**	1.67	***3.99***	3.88
13	3.55	0.92	6.96	3.68	3.18	1.17	5.72	3.14	5.82	3.25	7.86	3.63	**2.83**	2.29	***5.54***	4.25
14	3.89	1.24	7.24	2.52	**3.44**	1.61	***6.02***	2.39	6.08	3.29	8.37	4.07	3.48	1.99	6.07	4.36
15	3.99	0.94	7.93	3.26	**3.62**	1.28	6.41	3.18	5.77	2.99	9.60	3.71	4.02	3.26	***6.41***	5.05
16	4.76	1.42	8.65	3.63	**3.92**	1.71	***7.29***	2.68	7.40	3.41	9.86	3.34	5.61	5.07	7.50	6.15
17	4.52	1.10	9.61	3.96	4.75	2.31	***7.35***	3.14	7.71	4.29	9.00	3.40	**4.75**	2.77	7.84	5.27
18	4.86	1.55	7.64	3.40	4.65	1.40	7.56	3.03	7.74	4.28	10.54	4.45	**3.54**	2.85	***5.48***	5.64
19	5.45	1.71	10.52	4.71	**5.09**	1.36	9.10	4.25	8.72	4.76	11.51	4.96	5.14	3.27	***7.67***	5.27
20	5.47	1.50	10.49	5.12	5.97	2.52	10.64	3.73	8.43	4.85	11.14	4.19	**5.20**	4.59	***7.64***	6.4

The best global RMSDs (Best) and the top global RMSDs (Top) on the test set one, and their standard deviations (σ) are shown. The lowest best and top results are in bold.

3.3 Results

3.3.1 Standard Benchmark

When the top predictions made on all 510 test loops are considered, original FREAD produces the best global RMSD results at 10 of the lengths and RAPPER at the other 7 (Table 3.1). If best predictions (which indicate the sampling power) are considered, RAPPER and original FREAD also perform well. PLOP does not appear to produce accurate predictions, but in terms of ranking power, it is the most discriminatory of all the protocols with its top and best prediction in general very close.

One of the reasons that original FREAD and RAPPER may have higher accuracy on this native test set is that they never distort the given anchor structures, whereas PLOP and MODELLER may. This distortion is most marked for PLOP.

To ascertain which of original FREAD's filters affect its performance most, the relationship between these filters and the results produced were examined. Although anchor geometry has slight discriminatory power in the cases that anchors are nearly identical (Figure 3.3A), the most powerful filter was the environment specific substitution score (Figure 3.3C). The statistical potential score is shown to be a less significant discriminatory filter (Figure 3.3B). It should be noted that the two score functions are

based on the same formalism and built from sufficiently large amounts of data (see Section 3.2.5 and 3.2.4). The scores are calculated under the assumption that each probability of events is independent of each other, i.e., substitutions of corresponding residue positions are independent events (ESSS) and the statistical preference of an atom distance is not dependent upon other atoms distances. Although the assumption is not true, it is widely used and taken as a good approximation in general protein structure prediction. However, in protein loop structure prediction, the statistical potential is not a good approximation as the score was originally designed for larger structures than local loop segments. From here onwards, the statistical potential is not used as a filter. Therefore, if anchors are constrained (i.e., a prediction fits to a reasonable degree of accuracy onto the target), and if the score is above a cut-off of 25, it can be fairly certain that the loop is well predicted (within 2Å) regardless of loop length. The score cut-off is independent of loop length. This points to a strong dependence of original FREAD on highly similar loops.

It could be that the performance of original FREAD was over-estimated if the local loop structures FREAD uses for prediction come from the global templates people would use for TBM of the target sequence.

To test for such a bias, I examined the global (sequence and structure) similarity between the complete target structure and the complete structure from which any FREAD loop prediction was taken (Figure 3.4).

As can be seen in Figure 3.4, although there is global similarity in some cases, in a large percentage of cases (65%, sequence identity <30%), structurally similar loops can come from a globally distinct structure. This suggests that the local structure similarity can be quantified by local sequence similarity (as expressed by the environment dependent substitution tables) and the physical constraints of the anchors. Henceforth, "FREAD" indicates the new version of the FREAD algorithm with a cut-off of 25 on the environment specific substitution score.

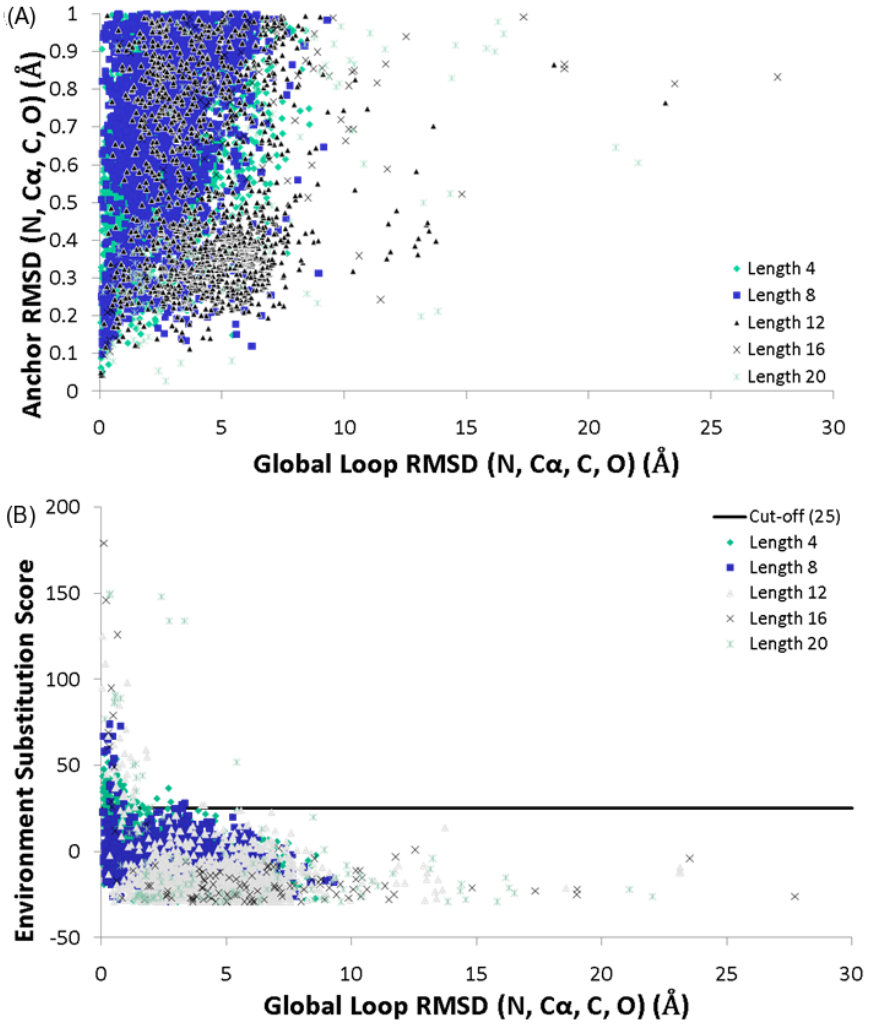


Figure 3.3: The effect of selection criteria on global loop RMSD
(A) Anchor RMSD (two residues either side of the loop) versus global loop RMSD. (B) The statistical potential score versus global loop RMSD. (C) Environment specific substitution score versus global loop RMSD for selected loop lengths (4, 8, 12, 16 and 20 in length). Each point in the plots is all the prediction made by FREAD.

The imposition of this cut-off decreases the number of predictions made by FREAD (Figure 3.5B) but significantly improves their accuracy (Figure 3.5A). The number of predictions drops by about a third for short loops and a half for longer loops. In total 286 loops out of the 510 are predicted (The change in average global loop RMSD versus coverage at different environment specific substitution score cut-offs is shown in Figure

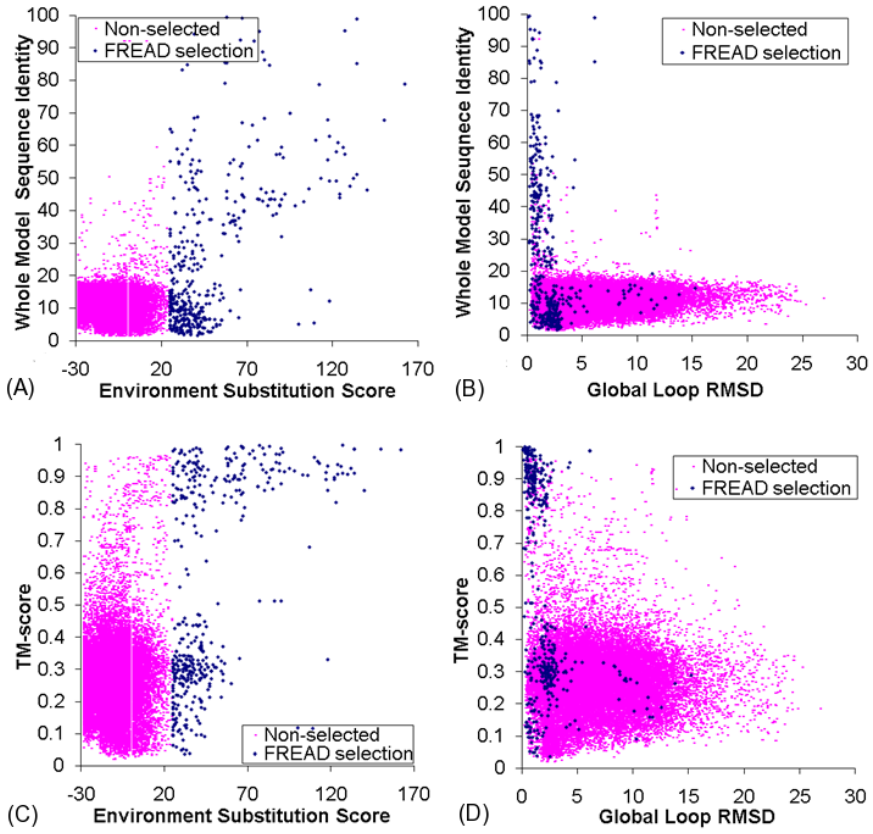


Figure 3.4: FREAD predictions often come from structures that are globally dissimilar to the target
The blue points are structures from which FREAD produced as top predictions are in the pink points are the other predictions. Compare global similarity between target and template in terms of sequence (A, B) and structure (C, D) with the local similarity quantified by environment substitution score (A, C) and global loop RMSD (B, D).

3.6). Instead of having accuracy decreasing with loop length, FREAD now has almost constant accuracy across loop lengths. For example, at loop length 8, the average RMSD becomes 1.15Å on the 18 loops now predicted (Original, for 30 loops, FREAD: 2.88Å, RAPPER: 2.88Å, MODELLER: 4.25Å and PLOP: 4.34Å).

The method change has also improved FREAD's ranking ability. The top and best predictions are now on average far closer to one another (0.18Å over all the test loops).

To identify whether FREAD's performance gain comes from the environment specific substitution score or the inclusion of nonloop fragments in the database, it was

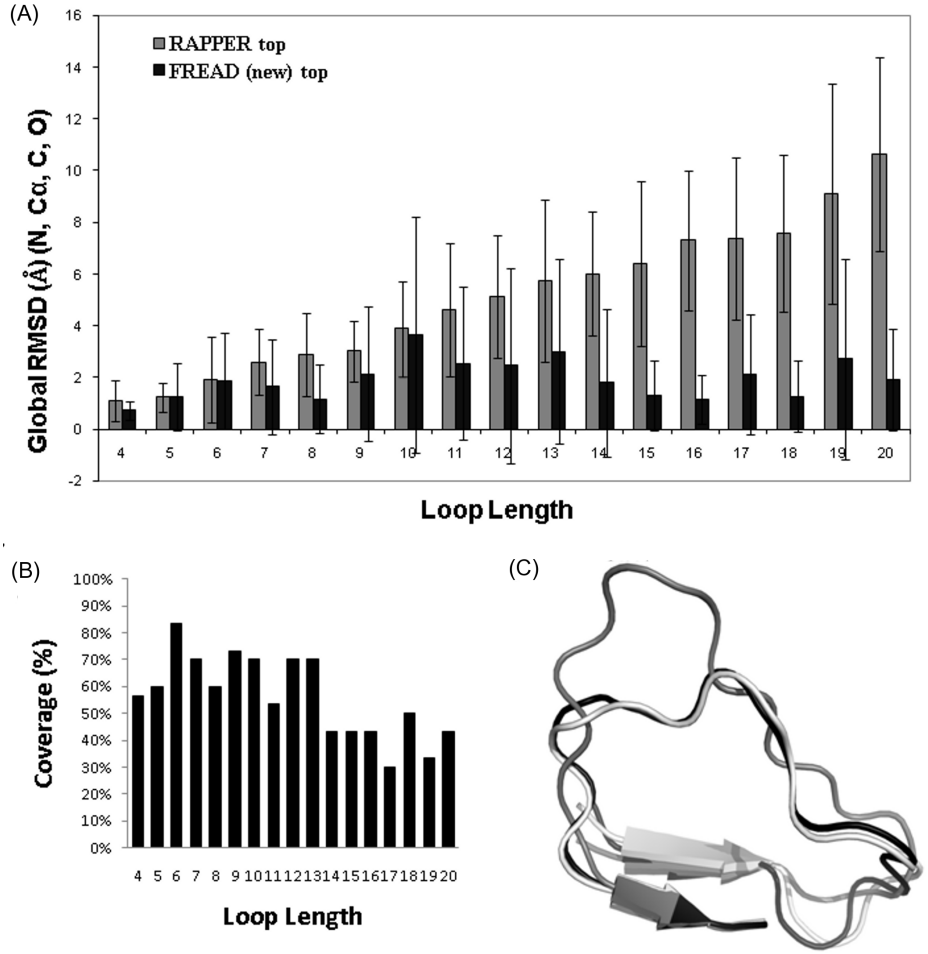


Figure 3.5: The predictive power of FREAD with the new environment specific substitution score cut-off
(A) Comparing average global RMSD for the subset loops predicted by FREAD with the new cut-off to the results of RAPPER, the best of the *ab initio* methods benchmarked here. FREAD is consistently accurate regardless of loop length. (B) The decrease in coverage made from original FREAD to FREAD with the new cut-off (the coverage of original FREAD is 100%). (C) An example prediction. The black structure is the native structure (2V9T, chain B 286–303). The gray loop is the top prediction by original FREAD (1WXR, chain A 816–833, the global loop RMSD against the native structure is 4.3Å). Its substitution score is –29 and anchor RMSD is 0.197Å. The white loop is the top prediction by FREAD with the new cut-off (1XKU, chain A 178–195, the global loop RMSD is 0.67Å). Its substitution score is 33 and anchor RMSD is 0.23Å. Despite the better anchor RMSD value of the grey loop structure, the total RMSD is worse than the top prediction of FREAD with the new cut-off.

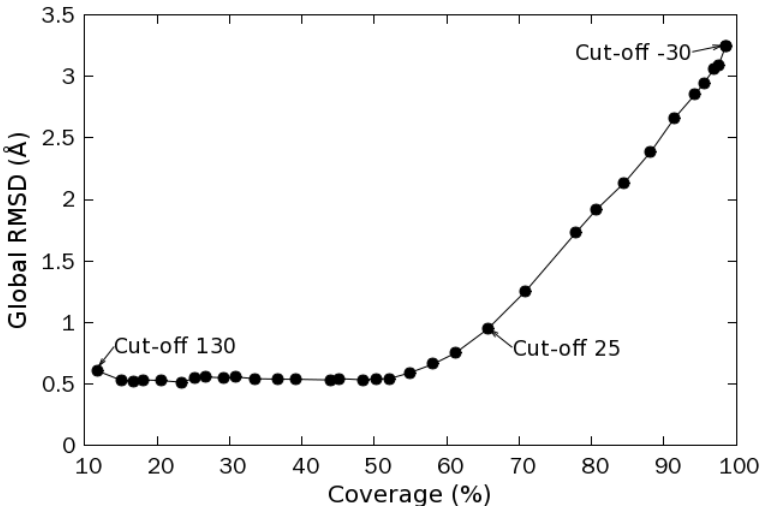


Figure 3.6: The change in average global loop RMSD of prediction across all lengths versus coverage for different environment specific substitution score cut-offs

tested with a database including only predefined loop fragments (FREAD-OL) for standard test set one. FREAD-OL, like FREAD, is length independent in terms of prediction accuracy and its accuracy is slightly better. However, the coverage of FREAD-OL is far lower. For example, FREAD-OL made only seven predictions compared to twenty two predictions for FREAD at length nine. Therefore, FREAD offers a balance between coverage (its coverage is lower than the original version, but higher than FREAD-OL) and accuracy (its accuracy is far higher than the original but marginally lower than FREAD-OL) (Figure 3.7).

To demonstrate that FREAD's predictive ability is not test set dependent, its performance on the test set two of entirely independent and sequence dissimilar loops was tested. On this set, it covers 52% of the loops and its accuracy remains very similar. For example, its average RMSD is 0.84Å on length eight loops (Figure 3.8).

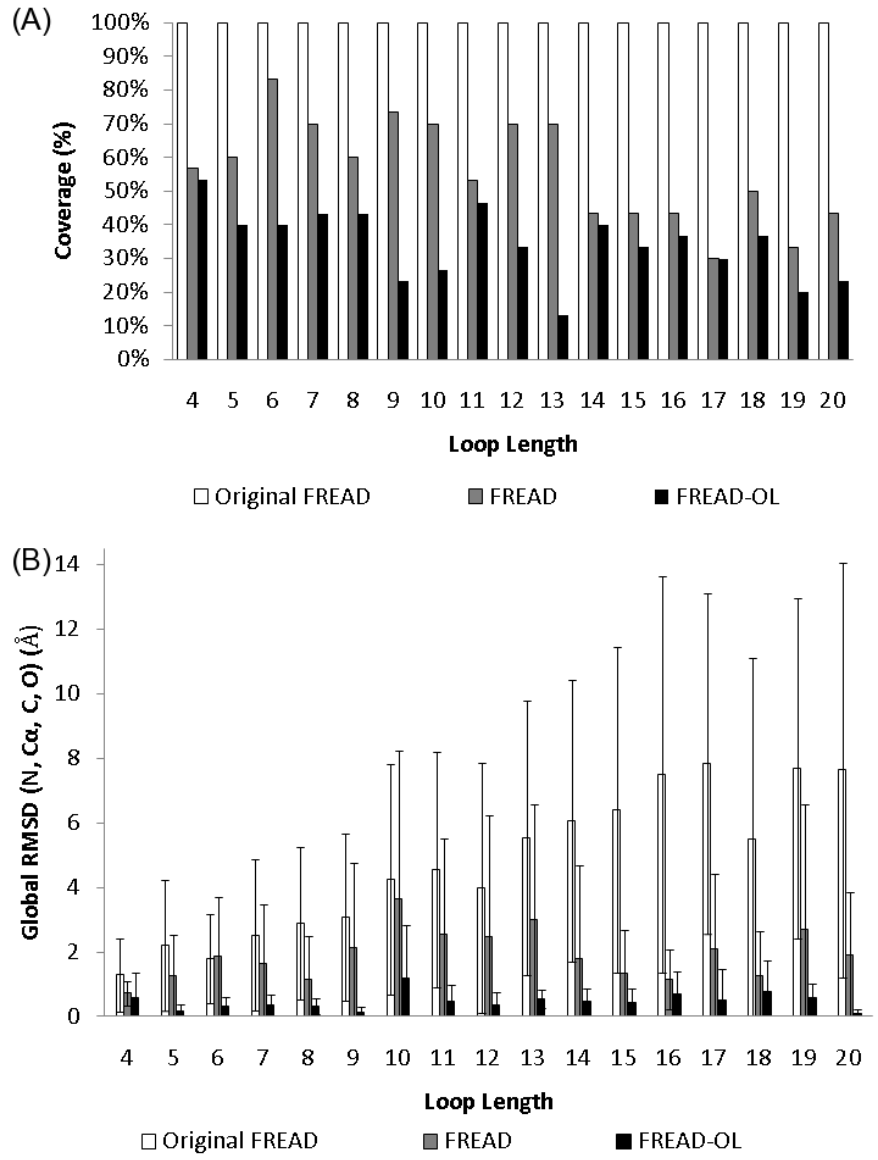


Figure 3.7: Comparison of database and environment specific substitution scores on prediction accuracy and coverage
(A) Coverage comparison between original FREAD, FREAD and FREAD-OL (FREAD using a database including only loop structures). (B) Accuracy comparison between the FREADs.

3.3.2 CASP Benchmark

To assess the predictive ability of FREAD in a more realistic setting, I used model structures computationally generated by CASP predictors. It means that anchor structures,

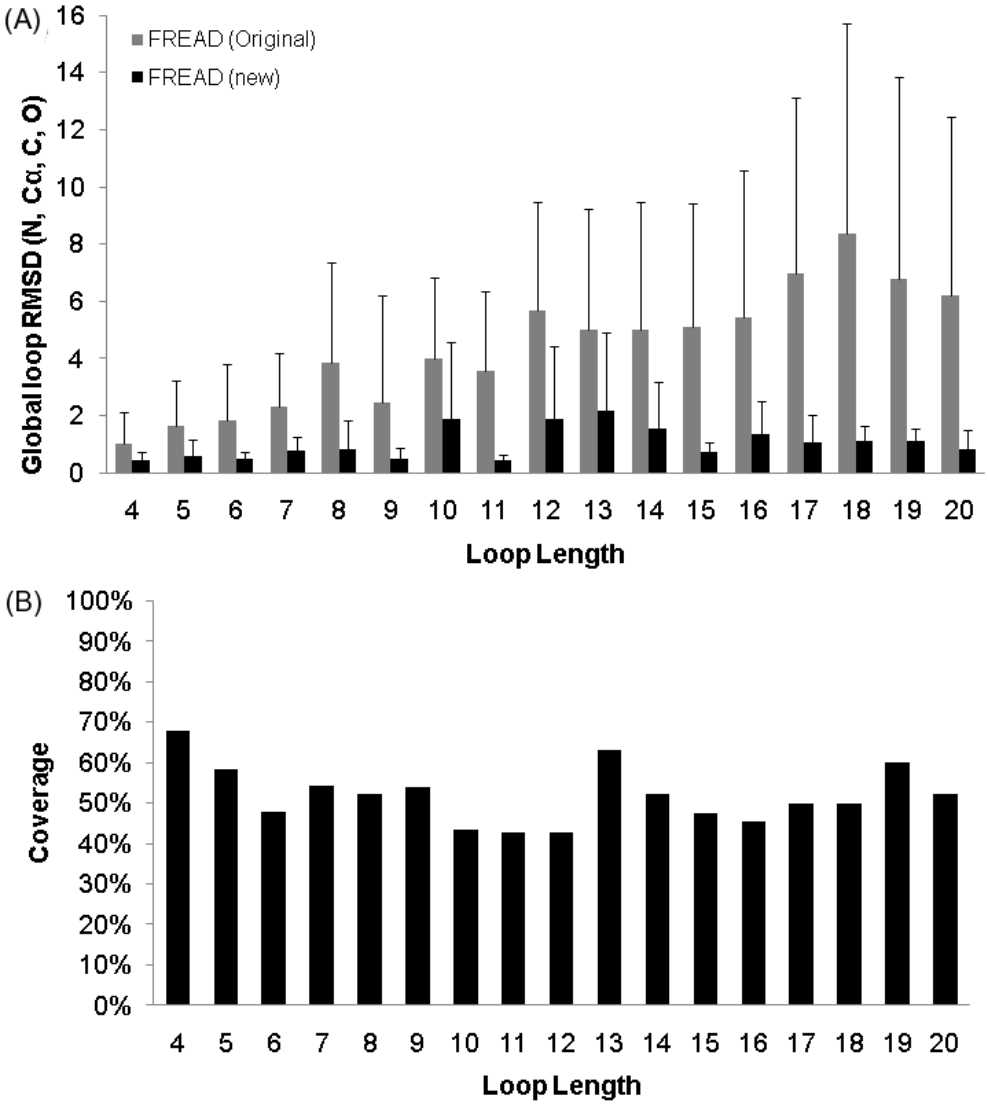


Figure 3.8: The predictive power of FREAD with the new environment specific substitution score cut-off for the standard benchmark test set two
(A) comparing average global RMSD for the subset loops predicted by original FREAD to FREAD. FREAD with the new cut-off still shows consistent accuracy regardless of loop length. (B) The decrease in coverage made from original FREAD to FREAD.

as well as the rest of the protein, are non-native.

Based on the standard benchmark, FREAD with a substitution score cut-off of 25 was run on loop test sets from CASP 7 and 8 (Appendix Table B.3). As described in section 3.2, the FREAD database used contains only structures available before the

CASPs in question.

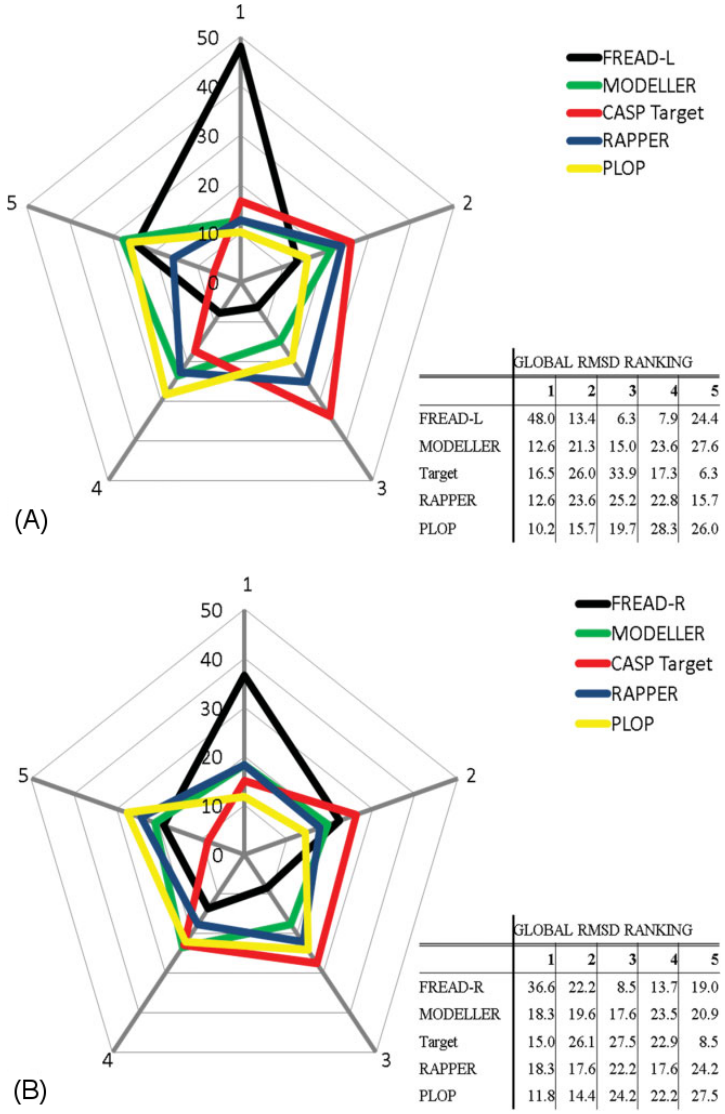
As the anchor structures of the CASP targets are not correctly given, variations of FREAD were tested. If FREAD cannot identify any matched fragments in the database, the loop region was extended by two, one at either end, up to length 26 until a prediction was achieved (FREAD-L). This was possible as FREAD retains consistent accuracy even as length increases. In a second version, FREAD-R, the anchor RMSD cut-off was relaxed (+0.1Å per cycle) and the Samudrala-Moult energy cut-off was relaxed (+5 per cycle), but the environment substitution score cut-off was fixed at 25. Relaxation of the cut-off values was carried out for three cycles. FREAD-R would lead in its extreme to a total dependence on sequence similarity.

FREAD-L (with length expansion) was able to predict 127 loop targets, while FREAD-R identified matched loops for 153 loop targets. This higher coverage for FREAD-R comes with a loss of accuracy (Figure 3.9).

Due to the small sample size, the average value of loop RMSD may not accurately represent the predictive powers of the protocols. Instead, I counted the ranks of loop prediction accuracies of the modeling protocols (Figure 3.9). Here the top predictions of each protocol along with the original CASP prediction are ranked one to five for accuracy. The frequency with which a protocol is ranked one can then be tallied. FREAD-L is ranked first on global RMSD 61 times (48%), while MODELLER and RAPPER are both ranked first 16 times (12.5%). The *ab initio* modelling protocols do not show significant improvements compared to the original CASP models, with the CASP models frequently ranking second or third.

Table 3.2: The correlation coefficients between global loop RMSD and GDT-TS score

GDT-TS	FREAD	MODELLER	Target	RAPPER	PLOP
Spearman's	-0.15	-0.16	-0.13	-0.15	-0.12
Pearson's	-0.18	-0.14	-0.09	-0.12	-0.10



	GLOBAL RMSD RANKING				
	1	2	3	4	5
FREAD-L	48.0	13.4	6.3	7.9	24.4
MODELLER	12.6	21.3	15.0	23.6	27.6
Target	16.5	26.0	33.9	17.3	6.3
RAPPER	12.6	23.6	25.2	22.8	15.7
PLOP	10.2	15.7	19.7	28.3	26.0

	GLOBAL RMSD RANKING				
	1	2	3	4	5
FREAD-R	36.6	22.2	8.5	13.7	19.0
MODELLER	18.3	19.6	17.6	23.5	20.9
Target	15.0	26.1	27.5	22.9	8.5
RAPPER	18.3	17.6	22.2	17.6	24.2
PLOP	11.8	14.4	24.2	22.2	27.5

Figure 3.9: Ranking of each method based on global RMSD of the top prediction using the CASP loop sets by FREAD-L and FREAD-R
The methods are ranked one to five according to the global RMSD of the top prediction. The five axes represent how often (in percent) each method was assigned a given ranking. For example, the black line in (A) shows that 48% of time FREAD-L makes the best (rank 1) prediction of the five methods and 24.4% of time it makes the worst (rank 5). (B) is the ranking results of FREAD-R.

As in this benchmark the structures are not correctly given, the quality of a loop prediction may be influenced by the quality of entire protein structure (Figure 3.10). The GDT-TS score (Zelma et al., 2001) of each model was calculated (Zhang and

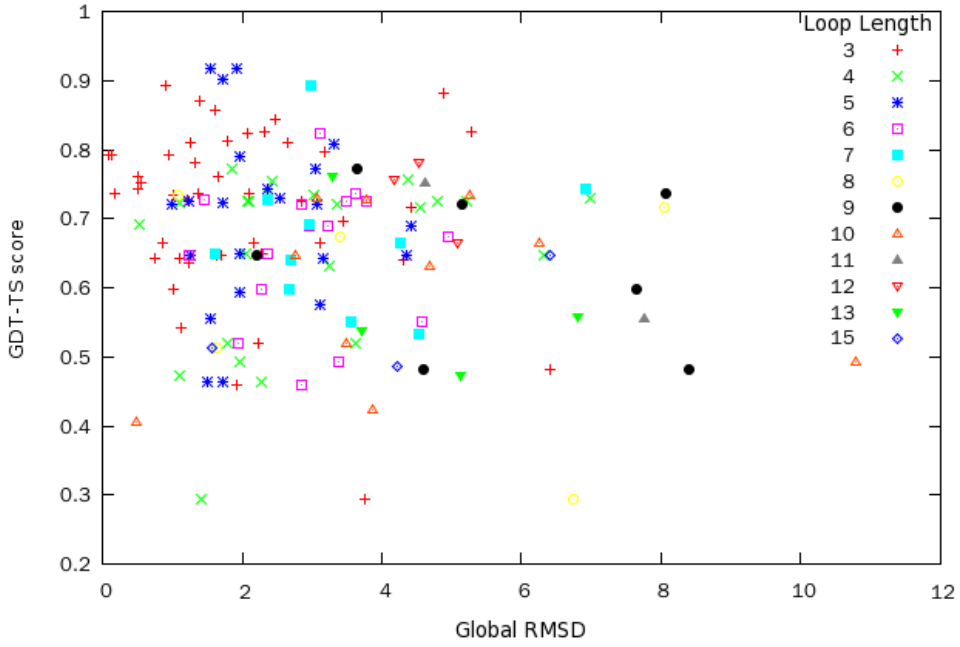


Figure 3.10: The correlation between the FREAD prediction quality and whole protein structure
Each point is the FREAD's top predictions on the CASP benchmark set. No correlation is found between the whole structure and the FREAD's predictions.

Skolnick, 2004) and compared to the corresponding global loop RMSD; no correlation between these scores was observed (Table 3.2).

Since the global structural quality does not affect prediction, it can be assumed that local anchor distortion might be an important factor. As Lessel and Schomburg (1999) pointed out, the critical weakness of using database search methods for loop prediction is that their prediction qualities can be influenced by the quality of anchor structures where the loop structure is inserted. This is still true in FREAD. The quality

Table 3.3: The correlation coefficients between global loop RMSDs and anchor RMSDs of target models and their native structures.

Anchor	FREAD	MODELLER	Target	RAPPER	PLOP
Spearman's	0.26 (0.49)	0.17 (0.08)	0.13 (0.11)	0.12 (0.25)	0.12 (0.09)
Pearson's	0.36 (0.57)	0.09 (-0.04)	0.13 (-0.01)	0.14 (0.13)	0.13 (0.00)

The numbers in the brackets are the correlation coefficients of loops longer than 7 residues in length.

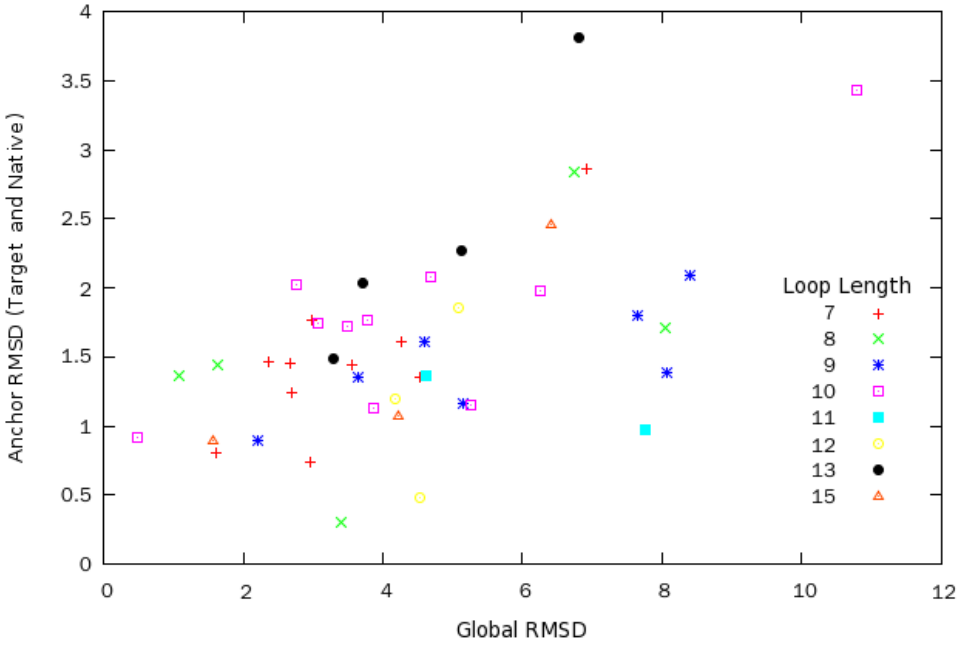


Figure 3.11: The correlation between the FREAD prediction quality and local anchor structure
There is a positive correlation between the anchor RMSD between native structure and its CASP model and accuracy.

of the prediction by any of the versions of FREAD is correlated with the discordance between anchor structures of a model and its native structure. The discordance was measured by anchor structure RMSD (two residues in each anchor) (Figure 3.11). The correlation is, however, only 0.49 and therefore does not explain all the inaccuracy in the prediction. It also appears stronger when the loop length is longer (Table 3.3). The *ab initio* methods do not show such correlation. However, despite this problem, FREAD-L still produces on average better predictions than the *ab initio* methods.

3.4 Conclusion

The era of database search loop prediction methods appeared to have passed as computer speed increased allowing *ab initio* methods to examine tens of thousands of potential conformations.

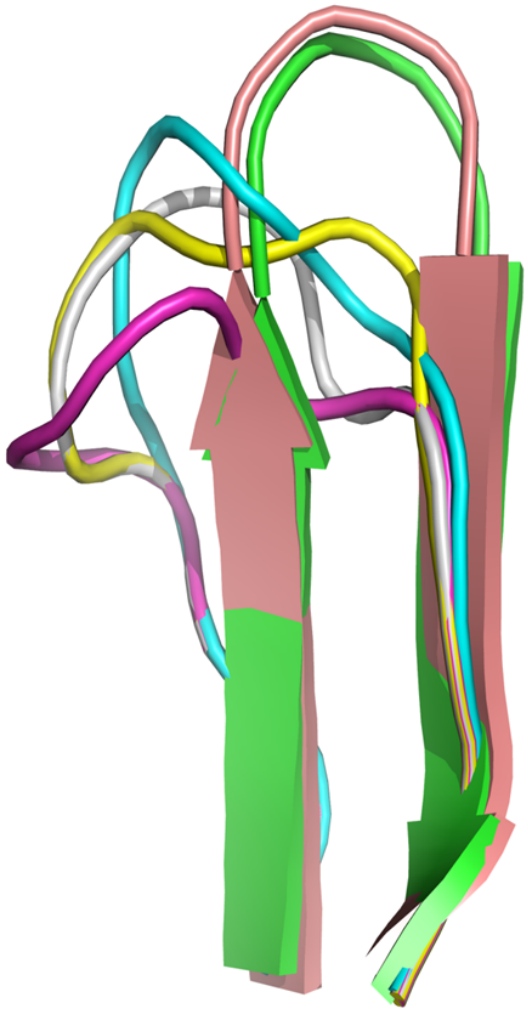


Figure 3.12: An example prediction on the CASP benchmark set using FREAD-L
The target loop is from 79 to 82 (T0427-D1, predicted by Jones-UCL). The native structure (3D3Y) is coloured pink and the original prediction by Jones-UCL is shown in white (the global loop RMSD against its native structure is 5.49Å). FREAD could not identify any matched fragments in the range. FREAD-L found a matched fragment in a longer range (77–84), which is shown in green (0.22Å). The *ab initio* methods, MODELLER (cyan, 4.78Å), RAPPER (yellow, 5.32Å) and PLOP (purple, 5.72Å), did not produce accurate results for the target loop.

In this chapter, FREAD, a database search loop prediction method, is re-evaluated and improved. It is shown that loop prediction using FREAD can be a powerful tool and, on average, outperforms three popular *ab initio* methods both on native structures

and CASP models. This re-examination of a database search method clearly shows that they still have significant power for a definable subset of loops both on real protein structures and on models.

The results as a whole demonstrate that local sequence similarity (as quantified by the environment dependent substitution tables) is the most powerful indicator of accuracy and can even remove the length dependence of prediction accuracy. Only six regions of the Ramachandran plot are used to describe structural type. It appears that this relatively simple split of structure alongside using only loop alignments is a better descriptor of loop structural similarity. The power of FREAD method also lies in the use of entire protein structures to select fragments from, rather than just loop segments as is used in many other database search methods.

One of the main problems of current database search methods for prediction on models is that they sample loops on fixed anchors. I attempted to overcome this weakness by extending loop lengths (FREAD-L) or relaxing anchor restrictions (FREAD-R). These processes can increase the number of predictions and still give better predictions than the *ab initio* protocols. These processes are possible as FREAD produces reasonable predictions regardless of loop length and its sampling speed is far faster than the *ab initio* methods. Multiple runs using FREAD are not unduly costly in terms of computational time. For example, a single FREAD prediction takes around 4 minutes on average (2.4 GHz Linux platform).

FREAD is, therefore, the best method among those benchmarked here when anchor structures are correctly given, and is on average better than the *ab initio* methods in real modeling situations when anchor structures can be inaccurate. FREAD will also continue to improve as the PDB continues expanding.

Chapter 4

Predicting Antibody Complementarity Determining Region Structures without Classification

4.1 Background

4.1.1 Antibody Structure

Antibodies are key elements in the immune system. They are also an important class of biological reagents which are widely used in therapeutic, diagnostic and research applications (Brekke and Sandlie, 2003; Morrison et al., 1984; Pavlou and Belsey, 2005; Reichert and Pavlou, 2004; Reichert et al., 2005).

The functional units of antibodies are immunoglobulin monomers. An immunoglobulin monomer consists of four polypeptide chains, two identical heavy chains and two identical light chains connected by disulfide bonds. The light chain is composed of

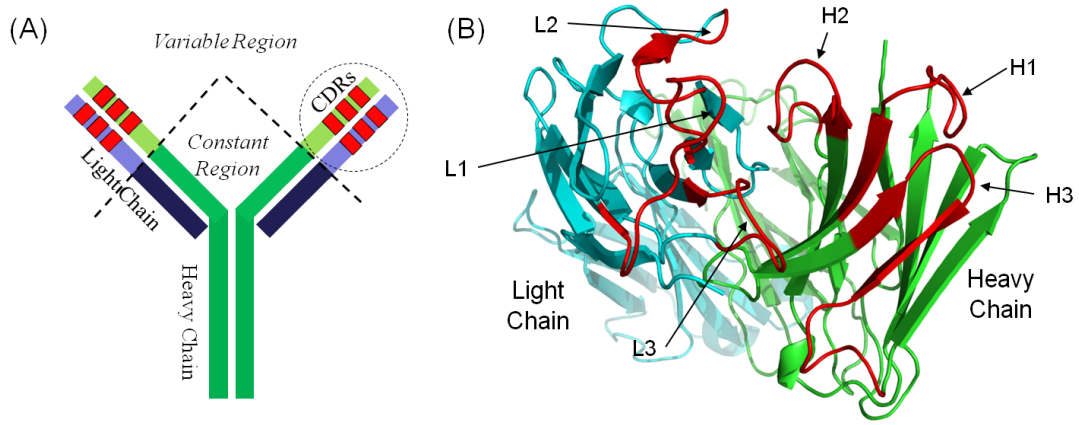


Figure 4.1: Antibody structure
(A) A schematic view of antibody structure. The circled region is magnified in (B) which is a cartoon view of the framework regions and six complementarity determining regions of a mouse antibody (1ACY).

domains which are around 220 amino acids long (there are about 440 amino acids in a heavy chain). They have a characteristic fold in which two β sheets create a sandwich shape, held together by interactions between conserved cysteines and charged amino acids.

The typical immunoglobulin structure schematically looks like a "Y" shape. It consists of three regions: a constant region and variable regions with hypervariable regions (Figure 4.1). The hypervariable regions form the binding site for the antigen called the Complementarity Determining Regions (CDRs). These vary in both structure and sequence. These CDRs are found in the two variable regions of the light (V_L) and heavy (V_H) chains and consist of three hypervariable loops in each of the light (L1, L2 and L3) and the heavy (H1, H2 and H3) chains. The variable region without the CDRs is called the framework region.

Light chains can be further classified into κ and λ classes in terms of different gene sources on which the light chains are encoded (κ: chromosome 2 and λ: chromosome 22).

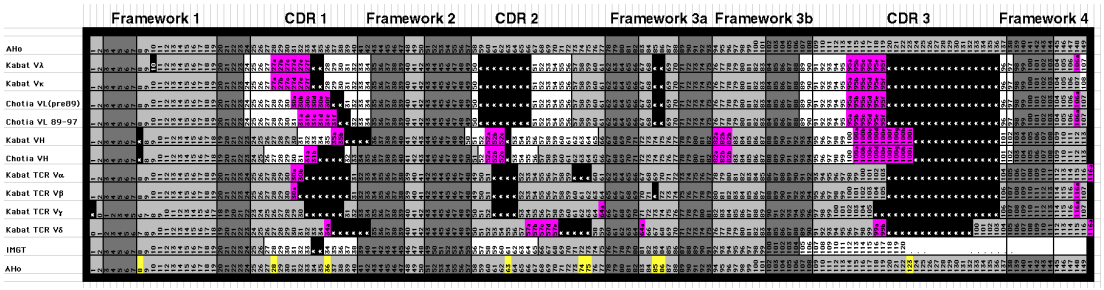


Figure 4.2: Antibody numbering schemes
The Kabat and Chothia numbering schemes annotate the two types of light chains and the heavy chain separately, whereas IMGT and AHo offer united numbering schemes. The yellow marks are reserved for insertions or deletions in the AHo numbering scheme. The other schemes annotate insertions or deletions unidirectionally (N to C). The white boxes indicate CDRs. This figure was taken from `http://www.bioc.uzh.ch/antibody/Numbering/NumFrame.html`

4.1.2 Numbering Scheme

In order to capture common sequence features within a protein superfamily, it is convenient to use the same nomenclature for the family members. Wu and Kabat (1970) analysed antibody sequences and found unusually high variations in certain regions which are later called the complementarity determining regions. They developed the Kabat numbering scheme based on highly conserved residues in antibody variable regions.

The Kabat numbering scheme is one of the most widely used. However, due to the small amount of antibody data available when the numbering scheme was established, errors in identifying CDR-L1 and H1 positions occurred. Chothia and Lesk (1987) corrected the positions to fit better with the three dimensional structure. Their revised numbering scheme is called the Chothia numbering scheme (Al-Lazikani et al., 1997; Chothia and Lesk, 1987; Chothia et al., 1989).

The Kabat and Chothia schemes have different numberings depending on the gene source (λ and κ light chains and the heavy chain). Lefranc (1997, 2011) proposed a unified numbering scheme, where the conserved residues have fixed positions (e.g. first cystein: 23, conserved tryptophan: 41, leucine: 89 and second cystein: 104). This

scheme (IMGT numbering) is used throughout this chapter unless stated otherwise. According to this numbering scheme, the CDRs are defined as follows: 27-38 (CDR-L1 and H1), 56-65 (CDR-L2 and H2) and 105-117 (CDR-L3 and H3).

The AHo numbering scheme (Honegger and Pluckthun, 2001) is an extended version of the IMGT scheme. They use bi-directional indels in the scheme (Figure 4.2).

4.2 Introduction

The Protein Data Bank (PDB) (Dutta et al., 2009) currently contains the structures of about 750 immunoglobulin structures. This enables good models to be created for the majority of antibody structures via homology modelling (framework region predicted to ∼1.2Å RMSD on average) (Sivasubramanian et al., 2009). Due to the high variability of the CDRs, these regions are predicted far less accurately (Al-Lazikani et al., 1997; Morea et al., 2000; Sivasubramanian et al., 2009). CDRs, however, are particularly important as they are the major contributors to the binding of the antigen (Davies and Metzger, 1983). Predicting the conformations of CDRs accurately is therefore vital for the understanding of antibody-antigen complexes and is of increasing in importance with the rise of therapeutic antibodies in healthcare (Walsh, 2006). Many diseases have benefited from therapeutic antibody drugs (Reichert and Pavlou, 2004) and homology modelling has already played an important role in the development of several such drugs (Schwede et al., 2009).

Despite the high sequence diversity of CDR loops, five of the six CDRs (L1, L2, L3, H1 and H2) are thought to have a set of limited structural conformations (canonical structures) (Chothia and Lesk, 1987). Reasonably accurate predictions can be made for these five non CDR-H3 loops using a set of sequence based canonical rules (Al-Lazikani et al., 1997; Bajorath and Sheriff, 1996; Morea et al., 2000; Ramsland et al., 1997). More recently the canonical structures have been updated (North et al., 2011) and it

has been shown that non CDR-H3 loops are largely predictable (estimated at 85%) using information derived from sequence, gene source and framework regions.

There have been several efforts to identify similar sequence rules for CDR-H3 (Martin and Thornton, 1996; Oliva et al., 1998; Shirai et al., 1999). However, no definitive canonical rules for all CDR-H3 loops have been found. Furthermore, there are many examples where different structural conformations and side chain arrangements of CDR-H3 loops occur because of their corresponding antigens (Manivel et al., 2000; Mundorff et al., 2000; Nguyen et al., 2003; Sethi et al., 2006; Wedemayer et al., 1997; Yin et al., 2003b, 2001). Some CDR-H3 loops in fact take on different structural conformations dependent on which antigen or which part of an antigen they bind (Schuermann et al., 2005; Yin et al., 2003a).

Much of the previous work on CDR structure prediction has been relying on the canonical rules. Using these rules, candidate CDR loops are selected from a database of CDR loop structures and grafted onto the rest of the antibody structure. If the query loop is not classifiable according to the rules then other strategies, such as *ab initio* methods, take over (Bruccoleri et al., 1988; Bruccoleri and Karplus, 1987; Mandal et al., 1996; Whitelegg and Rees, 2000).

Extensive benchmarks of CDR prediction have not been reported. Most methods have been tested on a small number of structures and focused on CDR-H3 loops. For example, ABGEN (Mandal et al., 1996) was tested on a set of 15 antibody structures and the CDR-H3 loops were predicted within the range of 0.98-4.06Å RMSD. WAM (Whitelegg and Rees, 2000) gave relatively accurate predictions of short CDR-loops (up to nine residues $\leq$ 1.7Å), but inaccurate results for longer loops based on 19 test structures. Marcatili et al. (2008) tested four examples and their results varied from 1.66–3.06Å for CDR-H3 loops of between 9 and 13 residues in length. A recent method RosettaAntibody was tested on a set of 54 antibody structures (Sivasubramanian et al., 2009). For non CDR-H3 loops, fragments were selected using BLAST (Altschul et al.,

1990) from a database of antibody structures and grafted onto the framework. For CDR-H3 loops, the *ab initio* Rosetta protocol (Rohl et al., 2004; Simons et al., 1997) was used, but with a database of fragments expanded to contain 230 antibody structures. Kinked CDR-H3 loops were specifically identified with sequence based rules (Shirai et al., 1999) and predicted using a special fragment library containing kinked conformations. The method gave reasonably accurate results for CDR-H3 loops (on average 2.91Å RMSD for all backbone atoms in model structures and 2.15Å in native structures).

In early studies, there were attempts to predict CDR structures using non-antibody structures as templates (de la Paz et al., 1986; Martin et al., 1989). As an extension of the idea, here I assume that the problems of the classification and prediction of antibody CDRs are in principle no different from those of loops whose anchor regions are anti-parallel beta sheets in soluble proteins. Here I demonstrate that CDRs can be predicted without prior classifications or knowledge of gene source.

The structural conformations of the CDR loops and the framework regions are known to be affected by their external environment (Braden and Poljak, 1995; Wilson and Stanfield, 1994) even without antigen binding (Pei et al., 1997; Rini et al., 1992). Although many CDRs may be predictable using sequence rules alone, these may not be sufficient to capture structural conformations in a modelling situation due to the high structural variability of CDR loops and its dependence upon surrounding environment. As FREAD predicts loops using only local similarities (local sequence and geometrical matches) and does not take into account such environmental effects, I developed an extended version of FREAD which uses contact profile to model the effects of surrounding environment.

FREAD was extensively tested on 97 non-redundant test structures. It produced accurate predictions for non CDR-H3 loops using a database containing only antibody structures (RMSD: 0.81Å (L1), 0.42Å (L2), 0.96Å (L3), 0.98Å (H1) and 0.88Å (H2) on

average), but relatively less accurate predictions for CDR-H3 loops (2.25Å on average). In order to overcome the relatively poor predictions of CDR-H3, I adapted FREAD to take greater account of sequence information (FREAD-S). This improved our results for CDR-H3 (RMSD: 1.38Å on average), but not for other CDRs, suggesting that CDR-H3 is the most sequence dependent of the CDRs. ConFREAD (FREAD-S with contact information) showed the best performance among the FREAD variants. However it caused a greater loss of prediction (for CDR-H3, 1.23Å on average with 70% coverage).

The FREAD variants were compared to RosettaAntibody on model structures. All the methods showed accurate results in non CDR-H3 predictions (about 1Å on average). CDR-H3 was once again less accurate (3.12 and 2.91Å for FREAD and RosettaAntibody respectively). FREAD-S and ConFREAD consistent with previous observations both achieved higher accuracy for CDR-H3 loops on models (RMSD: 2.07 and 2.91 Å for FREAD-S and RosettaAntibody, and 1.98 and 2.62Å for ConFREAD and RosettaAntibody respectively in the same subset).

Finally, in order to test the generic applicability, I modelled antigen-bound antibody structures using their corresponding antigen-free structures, computationally docking the antigen and then predicting the CDRs using FREAD. In this case, sequence-only rules may not discriminate which sampled fragments are best. In this test, for all CDR loops, ConFREAD gives the most accurate results among the FREAD variants (2.61, 3.5 and 1.35Å by FREAD, FREAD-S and ConFREAD respectively for native free against native bound structures in the same CDR-H3 subset).

4.3 Materials and Methods

4.3.1 Test Sets and CDR Definition

A total of 2009 antibody and antibody-related structures were collected using the union of the data in IMGT (Ver. 4.3.0) (Lefranc et al., 2009), immunoglobulin superfamily in

Table 4.1: A full list of the Native set

The Native Set with antigens																		
Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H
1NDM	A	B	1RIU	L	H	1AI1	L	H	1WEJ	L	H	1OB1	D	E	1MPA	L	H
3CFB	L	B	1C5C	L	H	1SM3	L	H	2W65	D	A	1KCR	L	H	1OSP	L	H
1N0X	M	H	1IKF	L	H	3DSF	L	H	1I9R	Y	K	2Z92	B	A	1A4K	A	H
1NC4	A	D	1T2Q	L	H	2IFF	L	H	1IFH	L	H	2R56	L	H	1E4X	L	H
1NCW	L	H	1FNS	L	H	1NSN	L	H	2AJX	L	H	1EAP	A	B	2MCP	L	H
1YUH	A	H	2NYY	C	D	1CFS	A	B	2CMR	L	H	1CE1	L	H	1NAK	L	I
35C8	L	H	2IPU	K	G	1FJ1	A	D	4FAB	L	H	2HKF	L	H	1MEX	L	H
1V7N	N	J	1F3D	J	K	1RZJ	L	H	1IND	L	H	1FIG	L	H	1TZG	M	H
1GGI	M	J	2HMI	C	D	2HRP	M	H	1YEE	L	H	1RJL	A	B	1ETZ	A	H
3BZ4	G	D	1OP5	K	M												
The Native Set without antigens																	
1E6O	L	H	1MCO	L	H	3FZU	L	C	1L7T	L	H	3CFE	A	B	1ZA6	G	D
2FJF	M	P	2IG2	L	H	1PG7	L	Z	2PR4	L	H	1F11	A	D	1RMF	L	H
1BM3	L	H	1AQK	L	H	1HKL	L	H	1UWG	X	Y	1MAM	L	H	1IAI	L	H
1NGZ	A	B	3EO9	L	H	2W9D	L	H	1NLD	L	H	2Q76	A	D	1NLB	L	H
1BZ7	A	B	1U6A	L	H	1PKQ	F	B	1Q9O	C	B	1A5F	L	H	3EYQ	C	D
1IGY	C	B	1MFE	L	H	1IGI	L	H	1F4Y	L	H	1MJU	L	H	1ADQ	L	H
1UM5	L	H	12E8	L	H	1DFB	L	H	1IT9	L	H	1P7K	A	H			

SCOP (Ver. 1.75) (Andreeva et al., 2007) and immunoglobulin homologous superfamily (2.60.40.10) in CATH (Ver. 3.3) (Greene et al., 2007). This union set was used to create a FREAD database excluding potential antibody structures (See section 4.3.4). Since the union set may contain non-immunoglobulins such as T-cell receptors, only the immunoglobulin structures which contain both heavy and light chains and CDRs annotated by IMGT were kept, leaving 588 immunoglobulin structures.

- Native Set: a non-redundant set created from the above structures. All the antibody structures in this set share less than 80% sequence identity in their variable regions. If any missing residues were found in any of the CDRs in the structure, the structure was discarded. This left 97 non-redundant antibody structures (56 antigen-bound structures and 41 antigen-free structures) which contain CDR-L1 through to H3 (The full list is given in Table 4.1).

- RA set: the CDR loops from the 54 structures used to test RosettaAntibody (Sivasubramanian et al., 2009). This set was used to compare FREAD to RosettaAntibody and test its prediction ability on modelled structures. Two sets were prepared; RA-Native set (54 native structures) and RA-Model set (54 homology models). (a full list is given in Table 4.2)

- Bound-Free set: Taken from Babor and Kortemme (2009). Twelve pairs of anti-

Table 4.2: A full list of the RA Set

Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H	Code	V_L	V_H
2DDQ	L	H	1DQQ	A	B	1Z3G	L	H	2AI0	L	H	1TET	L	H	1BQL	L	H
1CGS	L	H	1MLB	A	B	2C1P	L	H	2BDN	L	H	1FGN	L	H	1JPT	L	H
1A6T	A	B	1KEM	L	H	1QBL	L	H	1VFA	A	B	1IQD	A	B	1K4C	B	A
1JHL	L	H	2AEP	L	H	2FBJ	L	H	1IGT	A	B	2FD6	L	H	2ADF	L	H
2JEL	L	H	1YNT	A	B	1DBA	L	H	2B2X	L	H	1CLZ	L	H	2CJU	L	H
1FOR	L	H	1KB5	L	H	2AJU	L	H	1ZTX	L	H	1MCP	L	H	1NCA	L	H
2FJG	L	H	2H1P	L	H	1FPT	L	H	2ADG	A	B	1IGM	L	H	2FJH	L	H
2G5B	A	B	2H2P	D	C	1FBI	L	H	1BJ1	L	H	1WC7	L	H	1ZAN	L	H
2AJ3	A	B	2DQU	L	H	1F58	L	H	1HZH	L	H	1G9M	L	H	2B4C	L	H

body structures (a full list with RMSD differences is given in Table 4.3). The two sequences in each pair are 100% identical, but in one case, the structure contains an antigen and in the other it does not. The antigen-free structure is used as a model for the antigen-bound structure. The CDR conformations found in the bound structure were predicted using only the coordinates from the antigen-free structure. The antigen is positioned in the antigen-free structure by docking (See section 4.3.5).

In the original Babor and Kortemme (2009) set, there are 14 pairs. Here I exclude two of them: the 1AJ7–2RCS pair where the framework is very different between the two (C_α RMSD $>$ 3Å) and the 1FL5–1FL6 pair where residues are missing in the CDRs.

CDRs are defined as given in IMGT for both the Native set and the Bound-Free set. In the RA sets, the Chothia numbering scheme was used as it was used by RosettaAntibody in the original paper. In this case, the CDR loops of the light chain are composed of residues 24-34 (L1), 50-56 (L2) and 89-97 (L3), and the heavy chain CDR loops are residues 26-35 (H1), 50-56 (H2) and 95-102 (H3).

4.3.2 Measurement of Accuracy

The measure of accuracy used is global loop RMSD, calculated by superimposing entire structures except for the CDR loop regions and then calculating the RMSD of the loop main chain atoms (nitrogen, α carbon, carbon and oxygen). The predictions are

Table 4.3: The RMSD between CDRs of antigen-bound and antigen-free structures in the Bound-Free set after superimposing framework regions
The most different CDRs in each pair are in bold. The largest structure changes upon antigen binding occur in CDR-H loops. For the contact pattern changes upon binding, see Figure 4.10.

Free	Bound	L1	L2	L3	H1	H2	H3
1NGZ	1N7M	2.35	1.28	2.30	1.61	1.77	**2.65**
1D5I	1D6V	0.27	0.39	0.37	0.39	**1.99**	1.95
2A6J	2A6I	1.20	1.05	0.95	0.68	**2.17**	1.05
1Q9K	1Q9Q	1.70	0.77	0.46	0.81	0.42	**4.09**
1KCV	1KCS	0.34	0.26	0.93	0.89	0.83	**1.99**
1CR9	1CU4	0.56	0.43	0.64	1.55	1.09	**2.50**
1GGC	1GGI	1.43	1.48	1.28	2.03	1.18	**2.24**
1CGS	2CGR	5.25	2.53	3.92	**5.63**	4.30	2.79
1NBV	1CBV	1.25	0.73	1.02	**2.34**	1.65	1.76
1HIL	1IFH	1.59	1.34	0.57	**3.27**	2.27	3.12
1OAQ	1OAU	0.35	0.14	0.86	0.18	0.23	**1.59**
1MNU	1MPA	**2.43**	1.28	1.29	0.84	0.93	2.43

those ranked first by the modelling method. The method was tested using a leave one out cross validation. So in all cases, results from self-prediction were eliminated. Additionally, any self-predictions from both bound and free structures were removed in the Bound-Free set.

4.3.3 Contact Profile

A contact residue is defined in terms of the distance between a pair of C_β (C_α for glycine) atoms (including antigens if present). The atoms considered are either from the same chain of the residue of interest, from different chains or from the antigen. Within the same chain, ten residues at either side of the residues of interest are excluded from the contact calculation. The antigen can be a protein, peptide or hapten. If the antigen is a hapten, all heavy atoms in the hapten are considered in the contact calculation. This relatively simple definition is easy to calculate and does not depend on the quality of sidechain modelling. There are four possible contact types; 1) non-contact represented as "0" 2) external contact only (such as external chains or antigen)

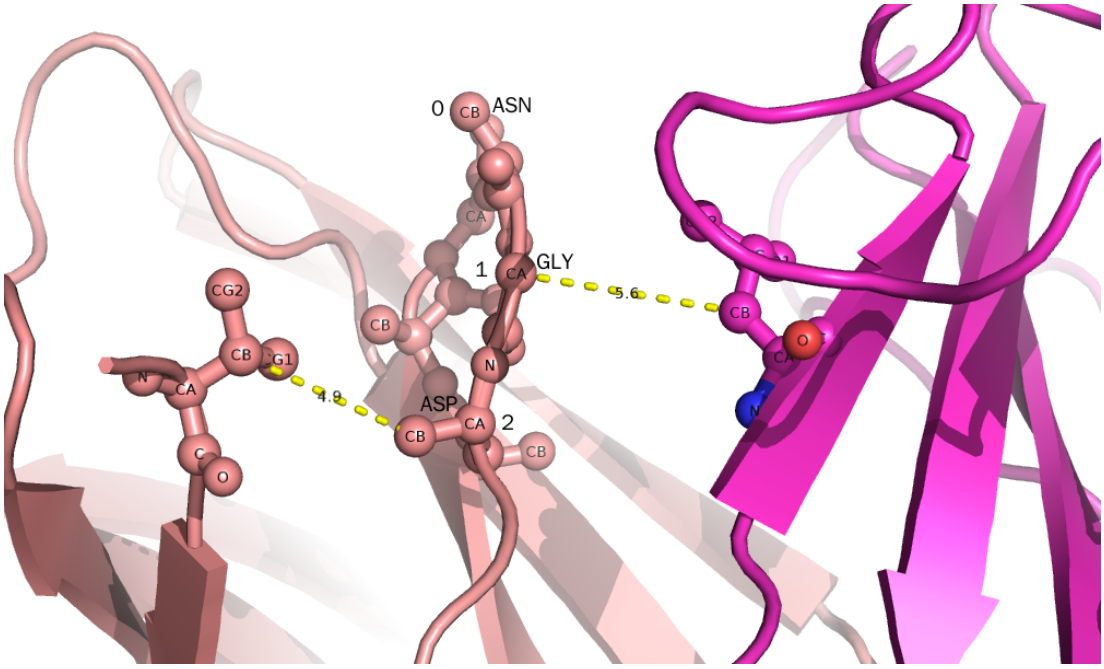


Figure 4.3: Contact profiles
The CDR-H3 loop in 1NSN has sequence TRGNGD. The asparagine residue H115 does not make any contact ("0"). The C_α atom in the glycine H116 is in an external contact ("1") with leucine L52 in the light chain. The aspartic acid residue 117 makes an internal contact ("2") with valine H2 in the heavy chain. The residue numbers here are assigned according to the IMGT numbering.

represented as "1", 3) internal chain contact only represented as "2" and 4) both external and internal contacts represented as "3" (See an example in Figure 4.3). The contact profiles of two loops can therefore be compared independently of sequence.

Contact profile information is available from three different sources.

a) The actual contact profile of the correct target fragment can be calculated from the full target structure; I term this the "target contact profile" (unavailable in a modelling situation)

b) The contact profile of a database fragment in its original structure termed the "database contact profile"

c) The contact profile calculated when a database fragment is grafted into the target

structure termed the "predicted contact profile"

Here I use only the database contact profile and the predicted contact profile both of which would be available in a modelling situation. Therefore the prediction made is entirely blind as to the target contact profile.

4.3.4 FREAD

FREAD Databases

Two separate databases were built.

- DB-I, which contains all the chains from the 2009 structures (identified as antibody and antibody-related).
- DB-E, which contains all known structures in the PDB with sequence identity $\leq$ 99%, resolution $\leq$ 3Å and R-factor $\leq$ 0.3. All the chains in DB-I are eliminated from DB-E leaving 28099 PDB chains (a structure database excluding antibody and antibody-related structures).

Selection Procedure and FREAD variants

In the standard FREAD protocol as described in Chapter 3, the maximum anchor RMSD cut-off value between the target and database anchors (two residues at each loop terminal) is set at 0.7Å and the ESSS cut-off is fixed at 25. All putative database fragments within these limits are then sorted and the one with the lowest anchor RMSD is selected as the first-ranked prediction. If FREAD is unable to identify a match of the same loop length within the database, it extends its search space by increasing the length of the loop region by two at a time (one at the N and one at the C terminal) until a prediction is made (stopped if length 26 reached). This extension is possible as FREAD has been shown to have relatively length independent accuracy.

Here I test a new variant of FREAD: FREAD-S. As described above, FREAD normally selects its first-ranked prediction according to anchor RMSD; in FREAD-S the list is ordered instead by the ESSS. This puts far more weight on the sequence component of the scoring system.

ConFREAD is an extended version of FREAD-S, which makes use of contact information. It acts as a filter on the database fragments in the prediction list given by FREAD. ConFREAD operates by running each FREAD prediction in turn (within the anchor RMSD and ESSS cut-off) and calculating both their database contact profile (the contact profile of the database fragment in the structure it was taken from) and their predicted contact profile (the profile of this database fragment when it is grafted onto the target structure). ConFREAD predicts only fragments which share 100% contact identity.

4.3.5 Idealised Docking

Twelve pairs of structures which share 100% sequence identity were chosen, but in each pair one member is bound to an antigen and the other is not (The Bound-Free set). The antigen-free structures are taken as initial models and their corresponding antigens (taken from their counterparts) are docked using ZDOCK (Chen et al., 2003). All the residues apart from the CDRs were blocked. Ten thousand antigen locations were generated and the best docked antigen, that closest to the native antigen position, was taken. After the docking step, ConFREAD was run on the antigen-free structures with the docked antigen. Self-predictions (fragments from both antigen-free and antigen bound structures) were discarded. The test protocol is outlined in Figure 4.4.

It should be noted that the aim of this study is not to test the docking method. The best docked antigen positions are taken as limitations of current protein docking programmes. The primary goal is to examine whether ConFREAD can produce accurate results even if the framework is approximate and the antigen is not exactly positioned.

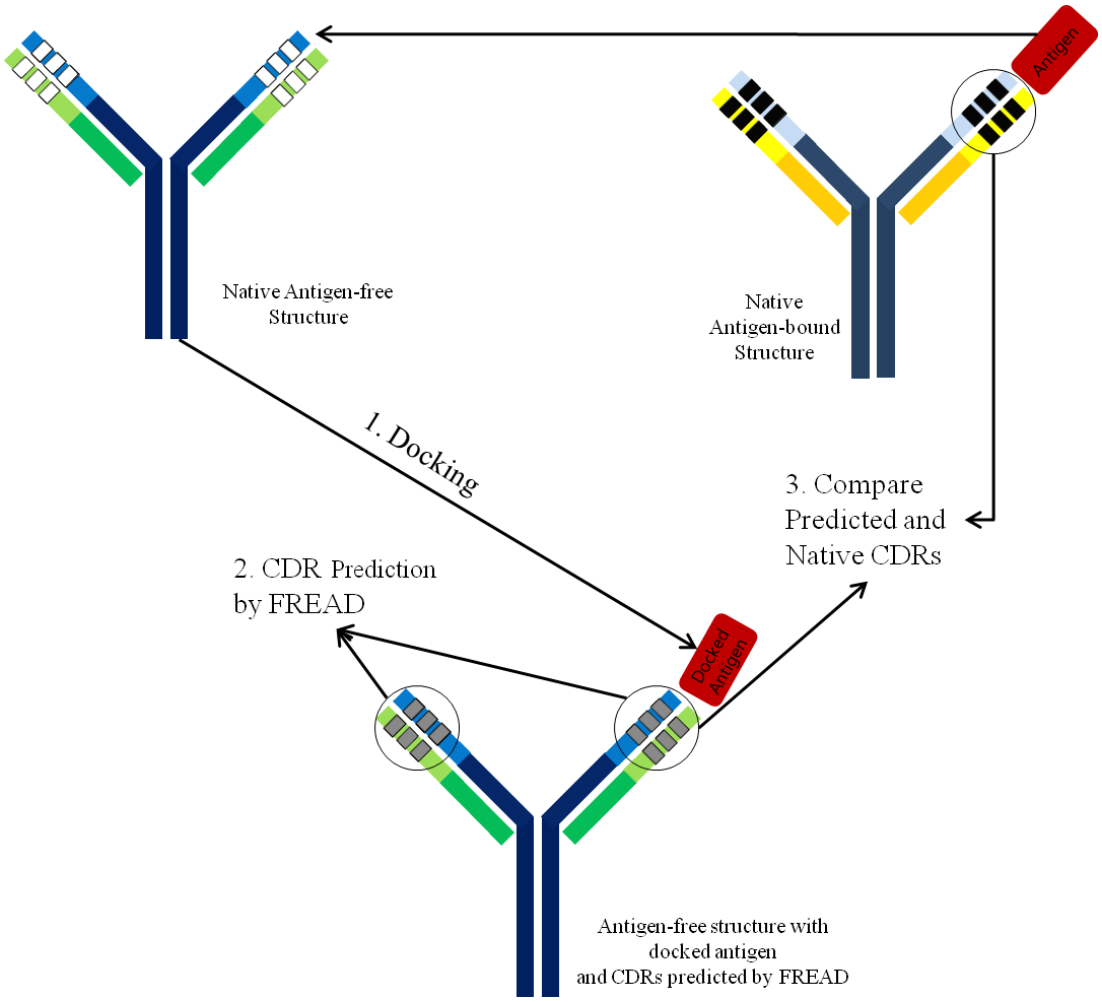


Figure 4.4: A schematic view of idealised docking and CDR prediction on the Bound-Free set
The antigen from the native antigen-bound structure is docked using ZDOCK to the antigen-free structure. Next the CDRs are predicted on the antigen-free framework in the presence of this approximately docked antigen. These predicted CDRs are then compared to those found in the native antigen bound structure.

4.4 Results

4.4.1 Using FREAD to Predict CDRs on Native Structures

Initially FREAD was used to predict the CDRs on a large non-redundant set of 97 structures (Native set, for more details see Materials and Methods) using DB-I (the

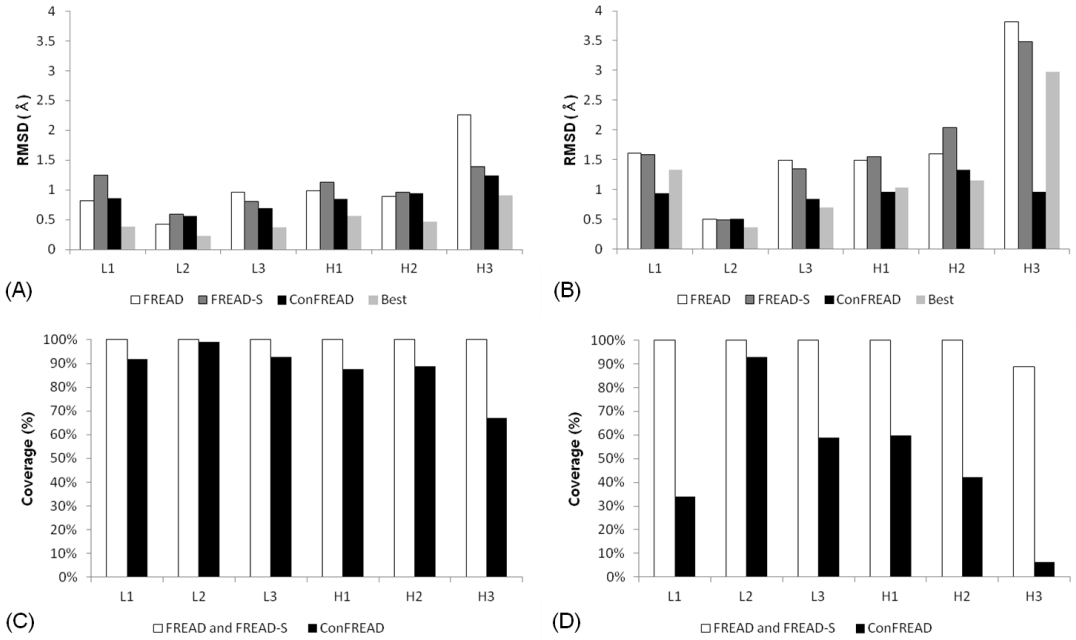


Figure 4.5: The results of CDR prediction on the Native Set
(A) Average global RMSD (DB-I), (B) Average global RMSD (DB-E), (C) Coverage (DB-I) and (D) Coverage (DB-E). Best refers to the average RMSD of the lowest RMSD structures FREAD produced. The detailed results including standard deviations are in Table C.1 and Table C.2.

database including antibody structures).

FREAD was able to predict all the cases in the Native set. As can be seen in Figure 4.5A, FREAD produced results of similar accuracy to those previously reported in Sivasubramanian et al. (2009) (RMSD: 0.81Å (L1), 0.42Å (L2), 0.96Å (L3), 0.98Å (H1), 0.88Å (H2) and 2.25Å (H3)). FREAD does not use any antibody specific knowledge such as conserved residues or structural classes to find matched fragments. The environment specific substitution score (ESSS, the main selection method in FREAD), which was developed for general loop structure prediction, appears able to accurately identify near native fragments for CDRs.

4.4.2 Investigating the applicability of the FREAD sequence score

It is noticeable that FREAD, like all methods, predicts CDR-H3 poorly, on average twice as badly as any of the other CDRs. In standard FREAD the first-ranked prediction is taken as the fragment with the lowest anchor RMSD among the predicted fragments which have a high sequence score (ESSS over 25). However, CDR loops are known to have high sequence-specificity and single mutations can cause changes in antigen affinity and structural conformation (Krause et al., 2011). A second version of FREAD was therefore tested, FREAD-S, which selects the fragment with the highest ESSS as the first-ranked prediction. Figure 4.5A shows that this method improved the accuracy of CDR-H3 prediction (RMSD: 2.25 $\rightarrow$ 1.38Å). However, such improvements were not seen for the other CDR loops.

4.4.3 Contact Profile and ConFREAD

As antibody CDRs change their structural conformations upon binding, it may be possible to use contacts between the antigen and the antibody and between different parts of the antibody to improve prediction. The contact profile describes the contacts of a fragment of protein structure with one value for each residue in the fragment. Each residue in CDR fragment is annotated between 0 and 3 dependent on the types of contacts that residue has (See section 4.3.3). For example, the different contact types can be seen in the CDR-H3 loops of chain B and H of 1XF2, an antibody-DNA complex (Figure 4.6). Chain H is actively involved in antigen binding while chain B is not. The two CDR-H3s of these chains share 100% sequence identity (VRGGYRPYYAMDY) whereas their contact profiles (2222100111212 and 2222012111212 for chain B and H respectively) share 76.9% contact identity. The structural distortion occurs where the contact information is different.

As shown in Figure 4.5A, using ConFREAD slightly improved prediction over FREAD-S and FREAD. However, it gives lower prediction coverage. For example,

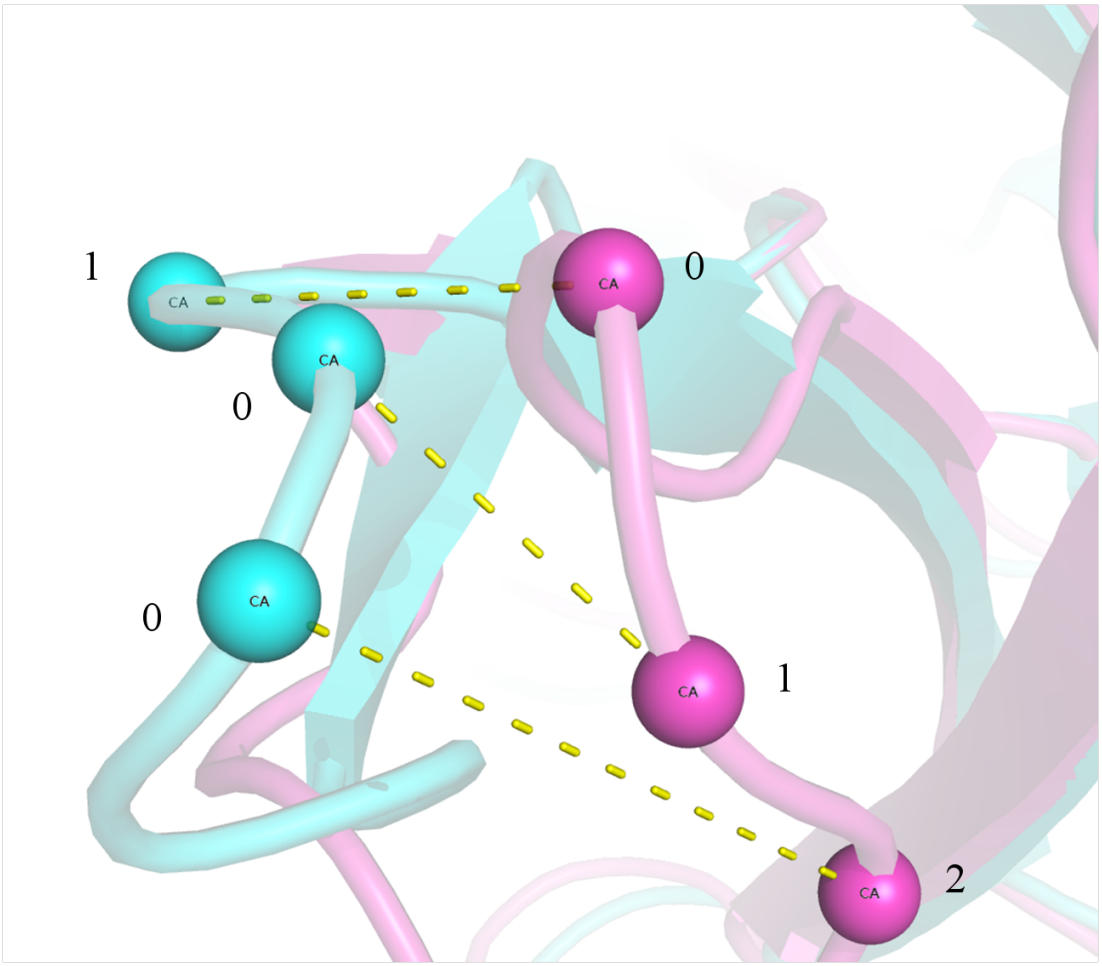


Figure 4.6: Structural difference and its relationship to contact profiles
Chain B (purple) and chain H (cyan) in 1XF2 share 100% sequence identity (VRGGYRPYYAMDY for CDR-H3), but have different structural conformations due to the antigen binding. Structural distortions occur on the residue pairs in different contacts (2222100111212 and 2222012111212 for chain B and H respectively).

in the case of CDR-H3 it drops from 100% (FREAD-S) to 70% (ConFREAD).

4.4.4 Predicting CDRs Using Non-Antibody Structures

If CDRs are no different from general loops, it may be possible to predict CDRs using non-antibody protein structures. The database excluding antibody structures (DB-E) was built and used to predict the CDRs of the Native Set.

FREAD and FREAD-S were able to predict most CDRs to a reasonable accuracy, but both perform worse than using a database containing antibody fragments (Figure 4.5B). In the case of CDR-H3, coverage is no longer 100% and accuracy has dropped (RMSD: 2.25→3.81Å for FREAD and 1.38→3.48Å for FREAD-S).

In the case of ConFREAD, with DB-E, accurate results are achieved but coverage drops substantially. ConFREAD's lack of coverage but highly accurate predictions for CDR-H3 using a database without antibody structures (6.1% coverage, 0.66Å average RMSD) may indicate that contacts are highly important for CDR-H3 shape and that such contacts are not seen in non-antibody structures.

It should be noted that all the fragments found for the CDR-H3 prediction using DB-E are from antibody-related structures (3CFB: 3EFD, 2W9D: 3H33, 1NLB: 3LS5, 1PKQ: 3GO1, 3EYQ: 3LS5, 2IPU: 3EYU). It is noticeable that non CDR-H3 predictions by ConFREAD are in the similar accuracy range to those using DB-I (RMSD: 0.86:0.93Å, 0.55:0.57Å, 0.68:0.83Å, 0.84:0.96Å, 0.94:1.33Å for CDR-L1, L2, L3, H1 and H2 respectively).

4.4.5 Predicting CDRs on Model Structures and Comparison to RosettaAntibody

The FREAD variants were also compared to RosettaAntibody. In order to demonstrate that FREAD is able to make predictions in a real-life modelling situation, CDR-H3 loops were predicted on the model set used in RosettaAntibody (the RA set). In this case, surroundings (such as antigens) were estimated by superimposing the model structure excluding the CDR loops onto the native structures.

FREAD and RosettaAntibody achieved similar accuracy on all CDR loops on the RA-Native set (Figure 4.7A). As seen previously on the Native set, FREAD-S did not improve the prediction for non CDR-H3 loops. However, FREAD-S did lead to improvement in CDR-H3 prediction (RMSD: 1.85→1.52Å).

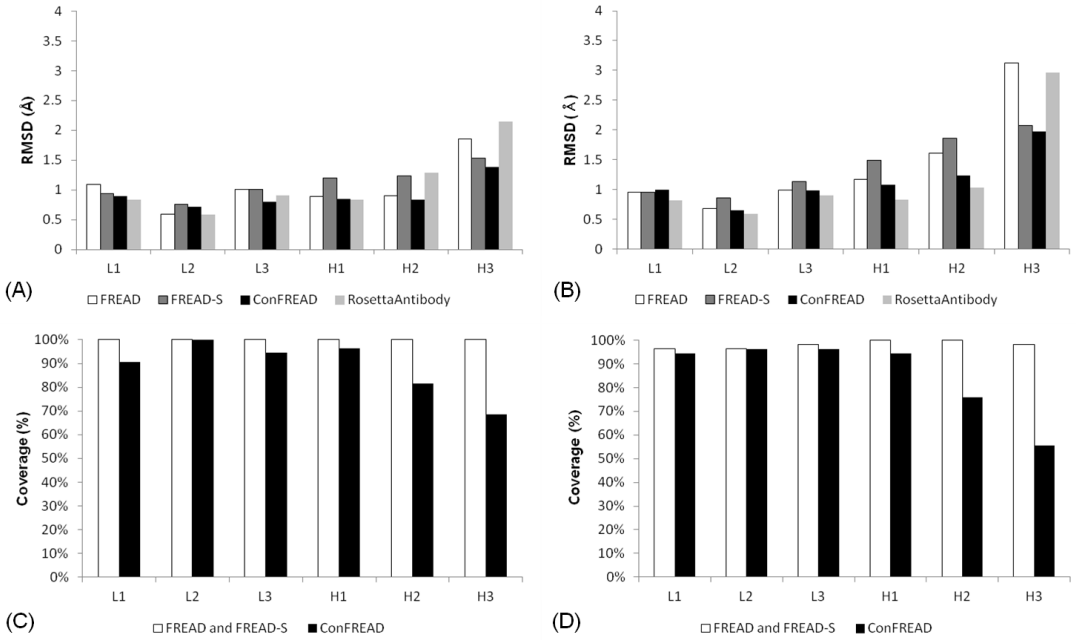


Figure 4.7: The results of the FREAD variants compared to RosettaAntibody on the RA sets
(A) Average global RMSD on the RA-Native set, (B) Average global RMSD on the RA-Model set, (C) Coverage (RA-Native set), (D) Coverage (RA-Model set). The detailed results including standard deviations are in Table C.3 and Table C.4.

Table 4.4: A full list of the results (global loop RMSD) of the CDR-H3 loops of RA-Native set. R.A. stands for RosettaAntibody.

Code	FREAD	FREAD-S	ConFREAD	R.A.	Code	FREAD	FREAD-S	ConFREAD	R.A.
2DDQ	4.06	3.82	3.82	1.7	1CLZ	1.89	0.60	0.61	1.0
1DQQ	0.64	0.37	1.01	0.3	2CJU	2.36	1.26	1.05	1.2
1Z3G	0.59	0.59	0.59	1.8	1FOR	2.08	2.94	2.03	1.3
2AI0	2.50	2.50	0.97	0.1	1KB5	3.49	2.29	3.13	1.7
1TET	0.28	0.28	-	0.8	2AJU	0.90	0.93	0.74	2.5
1BQL	0.32	0.32	0.32	1.5	1ZTX	2.02	1.82	3.03	2.8
1CGS	2.10	1.33	1.33	1.6	1MCP	1.46	0.44	0.44	1.1
1MLB	0.73	0.93	0.58	2.9	1NCA	6.03	0.67	0.67	1.6
2C1P	1.84	0.73	0.73	3.6	2FJG	0.50	2.02	0.50	2.6
2BDN	2.55	3.30	-	0.8	2H1P	2.03	2.29	-	3.1
1FGN	2.19	0.84	4.64	0.9	1FPT	1.72	3.55	-	3.6
1JPT	1.08	1.43	1.24	0.9	2ADG	1.95	0.32	0.32	3.7
1A6T	0.95	1.34	1.34	1.0	1IGM	1.90	5.51	-	3.1
1KEM	0.27	0.52	2.46	1.6	2FJH	4.09	0.93	-	1.7
1QBL	0.43	0.43	-	1.8	2G5B	0.36	0.36	3.63	1.9
1VFA	0.51	0.23	0.13	2.0	2H2P	2.71	0.54	0.54	2.5
1IQD	2.11	2.11	-	3.7	1FBI	4.35	0.51	-	5.7
1K4C	0.12	0.69	0.77	0.6	1BJ1	0.34	0.34	0.34	2.0
1JHL	2.13	2.13	1.87	0.9	1WC7	0.67	0.62	0.62	4.1
2AEP	2.29	2.61	-	1.0	1ZAN	6.10	6.26	4.62	4.4
2FBJ	3.61	1.60	1.60	1.1	2AJ3	0.45	0.66	-	3.3
1IGT	1.87	0.89	1.51	1.3	2DQU	0.48	0.48	0.48	2.7
2FD6	0.31	0.33	1.86	1.4	1F58	0.70	1.48	-	5.3
2ADF	2.34	2.43	-	2.0	1HZH	1.42	1.68	-	2.7
2JEL	0.95	2.42	-	2.0	1G9M	0.53	0.71	-	2.3
1YNT	0.12	0.12	0.12	3.2	2B4C	7.51	7.66	-	5.9
1DBA	1.11	0.55	0.60	0.7	**Mean**	1.85	1.53	1.38	2.15
2B2X	3.98	0.74	0.84	0.9	**STD**	1.62	1.54	1.24	1.33

Table 4.5: A full list of the results (global loop RMSD) of the CDR-H3 loops of RA-Model set. R.A. stands for RosettaAntibody.

Code	FREAD	FREAD-S	ConFREAD	R.A.	Code	FREAD	FREAD-S	ConFREAD	R.A.
2DDQ	4.20	3.68	-	1.4	1CLZ	1.01	1.26	4.32	2
1DQQ	0.71	0.71	1.12	0.3	2CJU	5.13	1.37	1.37	1.2
1Z3G	0.71	0.71	4.62	2.4	1FOR	3.81	2.22	2.05	2.1
2AI0	2.14	2.14	0.79	1.7	1KB5	2.99	2.86	4.50	3
1TET	6.58	3.33	-	1.1	2AJU	3.52	1.01	1.10	2.6
1BQL	4.37	4.74	-	1.6	1ZTX	2.60	2.97	-	1.9
1CGS	1.69	1.77	3.16	1.8	1MCP	3.19	1.11	-	2.9
1MLB	4.75	0.74	1.97	1.4	1NCA	11.56	0.82	-	3.4
2C1P	1.24	1.75	1.24	4.5	2FJG	2.82	2.49	-	2
2BDN	4.84	4.84	-	3	2H1P	1.96	2.28	-	1.7
1FGN	2.07	1.17	0.63	1.6	1FPT	3.62	3.14	3.72	2.2
1JPT	2.18	2.11	1.12	2.2	2ADG	4.82	1.52	4.61	4.2
1A6T	1.29	1.13	1.13	2	1IGM	1.92	5.53	-	4.5
1KEM	3.38	1.60	2.06	2.1	2FJH	3.85	1.07	4.37	3
1QBL	3.44	1.05	2.76	3.3	2G5B	0.84	0.84	3.65	3
1VFA	1.75	1.19	1.23	1	2H2P	4.17	4.17	-	1.6
1IQD	2.22	2.22	-	3.4	1FBI	4.33	4.63	-	5.8
1K4C	3.28	0.70	0.54	2.7	1BJ1	5.28	0.58	0.50	2.8
1JHL	1.36	2.07	1.60	1.1	1WC7	1.52	1.48	1.87	2.7
2AEP	2.89	2.89	-	1.2	1ZAN	4.90	6.37	-	4.3
2FBJ	3.74	1.68	-	2.7	2AJ3	4.60	4.57	-	3.8
1IGT	3.12	1.84	-	1.7	2DQU	5.02	0.35	0.35	3.1
2FD6	3.57	1.08	-	1.9	1F58	2.58	2.00	-	13.6
2ADF	1.89	2.33	-	3	1HZH	0.74	1.19	0.74	6.4
2JEL	1.90	2.80	-	1.4	1G9M	0.69	0.78	0.68	4.6
1YNT	4.00	0.51	0.51	3.6	2B4C	-	-	-	5.6
1DBA	0.77	1.17	1.02	4	**Mean**	3.12	2.07	1.98	2.96
2B2X	3.87	1.41	-	7.8	**STD**	1.88	1.40	1.44	2.08

On the RA-Model set, for CDR-H3, FREAD-S gave 98% coverage and an RMSD of 2.07Å compared to 2.91Å for RosettaAntibody in the same subset (Figure 4.7B). This level of accuracy on modelled structures with an approximate framework is encouraging. ConFREAD once again improves prediction accuracy with a drop in coverage. The improvement in accuracy is however only marginal (Table 4.4 and Table 4.5 for detailed results).

4.4.6 Predicting CDRs of Antigen-Bound Structures Using Antigen-Free Structures as Templates

In the Bound-Free Set, the antigen-free structure was used as a template to predict its CDR loops in a bound form. In each case the two structures share 100% sequence identity but only one structure has a bound-antigen. In this set, predicting CDRs using only sequence based approaches may not be successful. ZDOCK was run on the free structure where all the residues apart from the CDR residues were blocked to bind. Ten thousand antigen positions were generated and the best placed antigen was chosen.

Here I am examining only the current limit of antibody-antigen docking sampling. The purpose of this study is not benchmarking a docking method, but investigating structural transitions of local CDR regions upon environment changes and how one can capture the structural variability. Note that only ConFREAD makes use of the antigen position.

The FREAD variants were run on this free structure with the docked antigen. It was assumed that the bound CDR forms are unknown. Hence, the self-predictions from both free and bound structures were eliminated (Otherwise FREAD would give the free structure as the top prediction for any CDR of the bound structure).

For all CDRs, ConFREAD gave the best prediction on average (Figure 4.8). Again, the most significant improvement can be seen in CDR-H3 prediction (Table 4.6). For non CDR-H3 loops, FREAD showed similar accuracies to those for native CDR loops of the free structures. However, for CDR-H3, although the coverage decreased, ConFREAD successfully discriminated bad fragments and gave better results than using the native free CDRs.

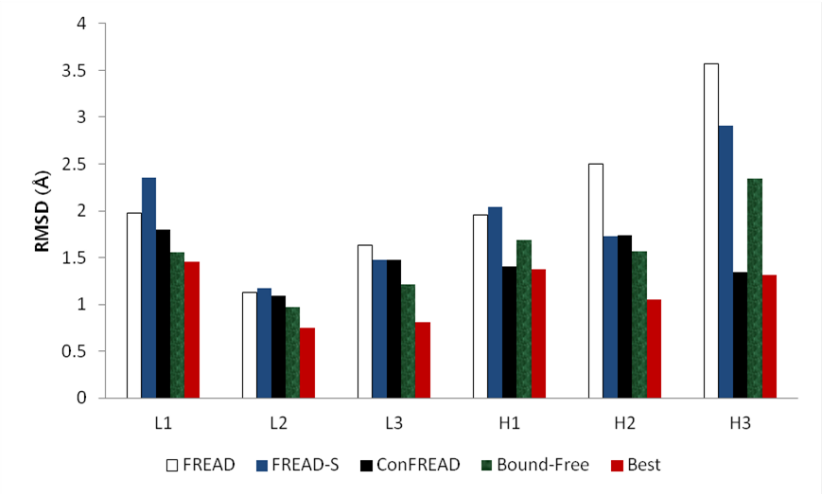


Figure 4.8: The results of the FREAD variants on the Bound-Free set
Best refers to the average RMSD of the lowest RMSD structures FREAD produced. Detailed results including standard deviations are in Table C.5.

Table 4.6: The prediction results of the Bound-Free set for CDR-H3. The second last column is for the best fragments found using FREAD and the last column is differences between the centroids of the native antigen and the docked antigen. Free-Bound refers to the difference between the antigen free CDR-H3 and its bound counterpart.

Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best	Antigen Position
1NGZ	1N7M	2.65	1.20	5.63	1.50	0.71	21.66
1D5I	1D6V	1.95	5.22	1.63	0.73	0.73	28.06
2A6J	2A6I	1.05	4.28	2.18	-	1.39	0.74
1Q9K	1Q9Q	4.09	0.60	4.37	1.21	0.60	25.54
1KCV	1KCS	1.99	1.66	0.63	0.63	0.63	0.89
1CR9	1CU4	2.50	2.50	2.95	-	1.35	0.57
1GGC	1GGI	2.24	2.88	0.60	-	0.60	1.74
1CGS	2CGR	2.79	5.59	3.17	-	2.19	24.93
1NBV	1CBV	1.76	3.28	2.64	-	2.25	24.82
1HIL	1IFH	3.12	3.08	3.08	3.08	2.50	0.64
1OAQ	1OAU	1.59	3.93	5.65	0.92	0.36	0.71
1MNU	1MPA	2.43	8.66	2.39	-	2.39	1.54
Mean		2.35	3.57	2.91	1.35	1.31	
STD		0.79	2.20	1.66	0.91	0.82	

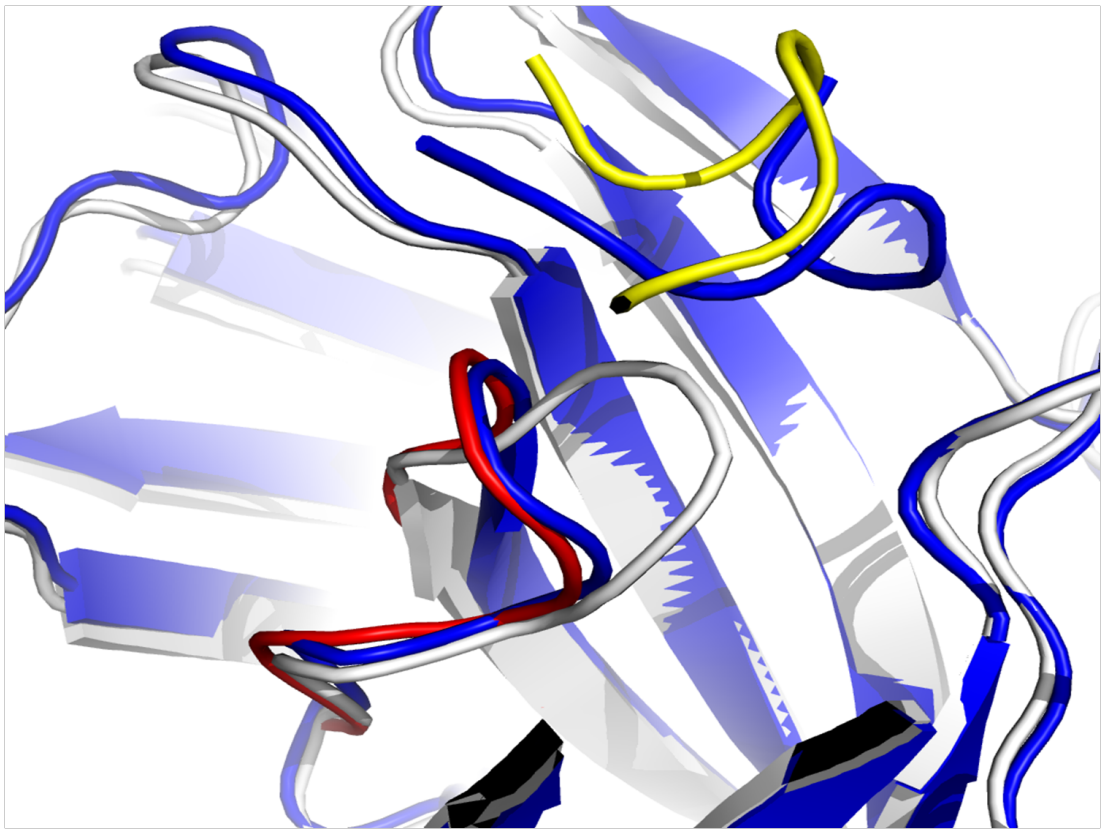


Figure 4.9: Predicting CDR-H3 using a docked antigen
The native antigen and bound CDR-H3 of the bound structure (1KCS) are blue and its free structure (1KCV) is white (1.99Å). Their contact profiles are 223311012 and 222211012 respectively. The yellow structure is the docked antigen and the prediction made by ConFREAD is red (0.63Å).

Although the antigens were not correctly placed, ConFREAD was still able to give relatively accurate predictions (Figure 4.9 and Table 4.6). This is probably due to the roughness of the contact profile method, since the contact profile is given only by atomic distances and does not depend on exact antigen positions.

4.5 Discussion

CDR structure prediction has been reasonably successful for five of the six CDR loops in antibodies, through the use of the canonical rules. The sixth CDR, CDR-H3 has in general proved more challenging and is predicted less accurately. I approach CDR loop structure prediction as a special case of the general loop structure prediction problem using FREAD, a database search method, without relying on any classifications of

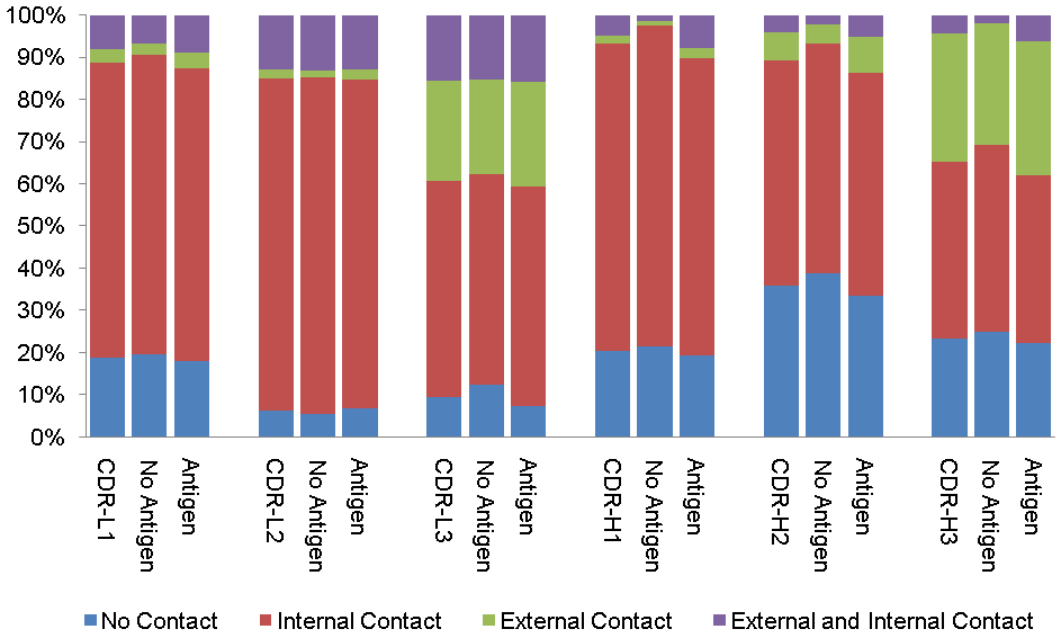


Figure 4.10: Contact profile composition of CDRs in the Native set
Each bar represents contact type compositions in single CDRs on average. The first bars are sums of antigen bound structures and unbound structures. The second bars are contact type frequencies of antibody structures where antigens are not present. The third bars are from antigen bound structures.

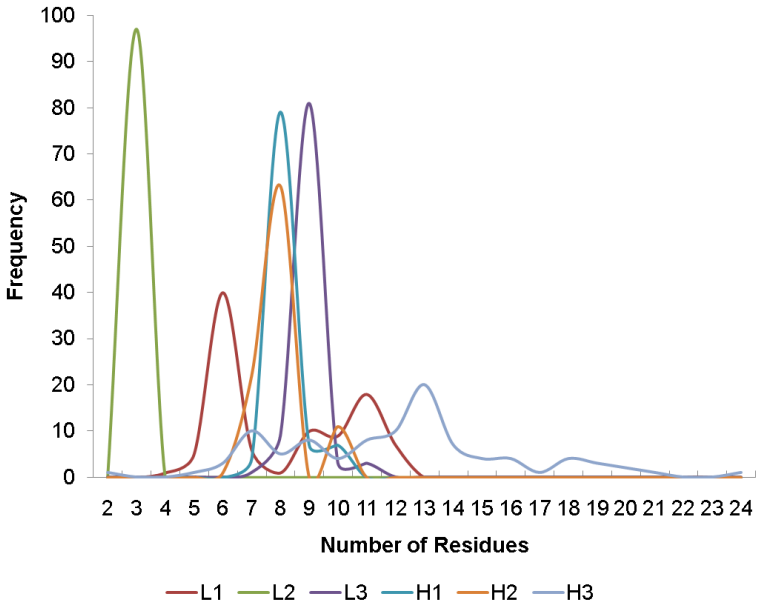


Figure 4.11: CDR length distribution

antibodies. On a non-redundant set of 97 antibody structures, I show that CDRs can be predicted successfully using FREAD. However, CDR-H3 loops are predicted least accurately. In order to overcome this, I modified FREAD to focus more on sequence similarity in prediction (FREAD-S). FREAD-S improved prediction for CDR-H3, but not for the other CDR loops.

CDR loops are supposed to be actively involved in antigen interactions and sometimes interact with other parts of the antibody structure. CDR-H3 can also take on multiple structural conformations dependent on its antigen. Figure 4.10 shows how contact types are composed in single CDRs. In fact, CDR-H3 and CDR-L3 are both involved more in external contacts than the others. About 40% of those CDR types have external contacts whereas this is only about 10% in the other CDRs.

However, CDR-L3's contacts do not change much upon antigen binding. All the CDRs in the light chain do not show significant differences in contact patterns between antigen bound structures and unbound structures. This suggests that most external contacts in CDR-Ls are with heavy chains and they have relatively smaller contributions

to antigen-binding. On the other hand, CDR-Hs show notable differences in contact compositions.

In this chapter, I have examined multiple FREAD variants on multiple datasets and situations under the assumption that CDR prediction is no different from general loop structure prediction. In any test set, non CDR-H3 loops are accurately predicted using any FREAD variant and the contact filter does not show significant improvements in prediction accuracy. This may be due to the contact compositions of non CDR-H3 loops. CDR-L1, L2, H1 and H2 generally have internal contacts or no contacts.

The conservation of contact patterns in terms of antigen binding and the fact that non CDR-H3 loops have few external contacts mean that most of their structural conformations can be captured using sequence rules. However, in the case of CDR-H3, the changes in contact patterns upon antigen binding and the wide length variation would make classification and prediction difficult (See Figure 4.11).

Most CDRs can be predicted using local similarities such as local sequence and geometrical matches. However, predicting CDRs with the contact information (ConF-READ) is a more generic method as shown in the prediction using antibody excluded database and the tests on model structures and bound-free pairs. The backbone structure of CDRs can be captured as long as the contact information is correctly given, even though external environments are not exactly known (the idealised docking).

Chapter 5

How Long Is a Piece of Loop?

Acknowledgement

This work was done in collaboration with Sumeet Agarwal (Department of Physics, Oxford University). He analysed the loop span distribution and generated random numbers from the Maxwell-Boltzmann distribution.

Nomenclature

In this chapter, proteins are divided into two main classes: membrane and soluble proteins. Loops from membrane protein structures are called "membrane loops" and those from soluble proteins structures are referred to as "soluble loops". Loops connecting anti-parallel β sheets are termed "anti-parallel β loop".

"span" (l) refers to the physical spatial distance between two end C_α atoms of a loop. Maximum loop span (l_{max}) is the longest distance that a set of residues (n) span.

The maximum span l_{max} is a function of the number of residues n and calculated as follows.

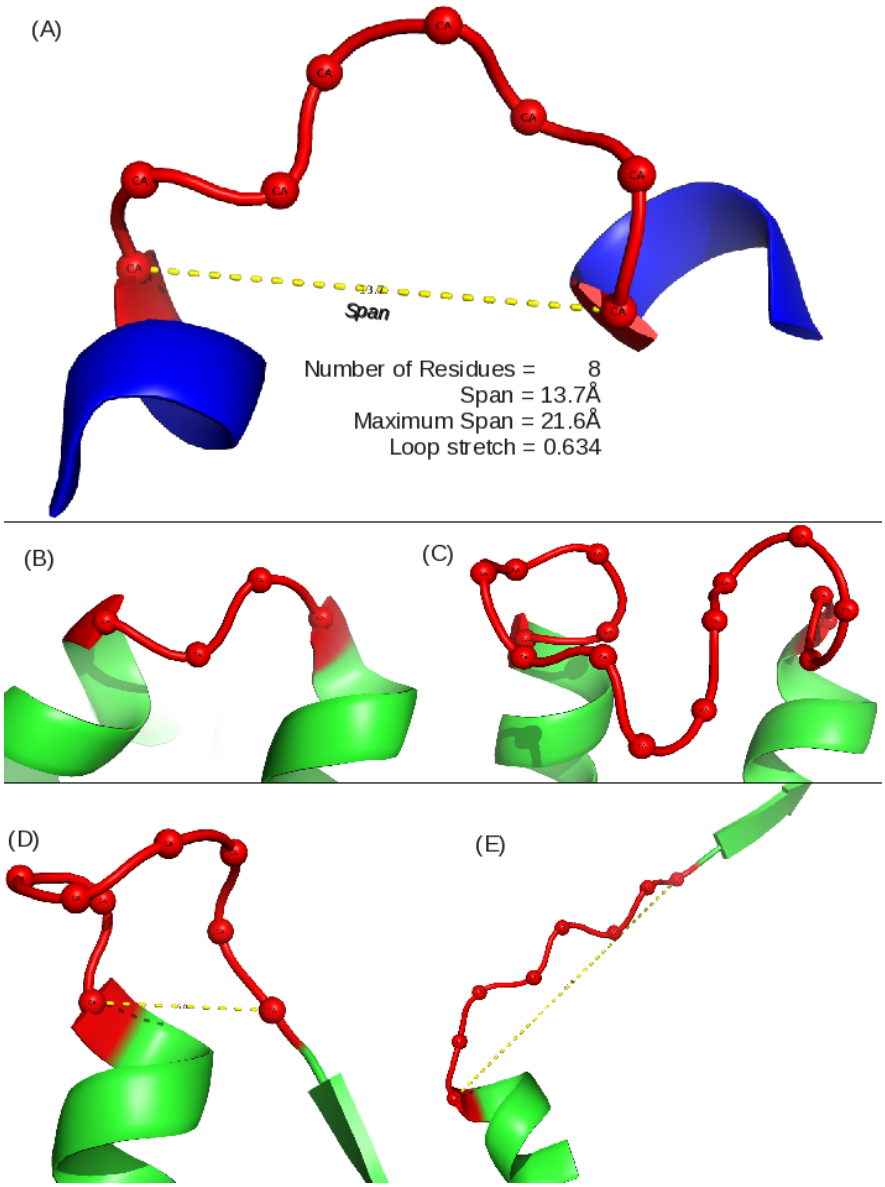


Figure 5.1: Loop description
(A) Span (l) is the end-to-end C_α distance of a loop. The maximum span (l_{max}) is defined by its number of residues. In this example (2J9O Chain A, 198–205), the loop has 8 residues and thus its maximum span is 21.6Å. Loop stretch (λ) is the normalised span (13.7/21.6 $\simeq$ 0.634). A loop can be shorter (Figure B: 3G0M Chain A, 16–19) or longer (Figure C: 2PYW Chain A, 321–336) in terms of its number of residues. Figure D and E show loops of different loop stretches. The loop in 1JI1 (Chain A, 250–257) is more contracted than Figure A (0.31) and that in 3CJE (Chain A, 148–155) is nearly fully stretched (0.99).

$$l_{max}(n) = \begin{cases} \gamma \cdot (n/2 - 1) + \delta & \text{if } n \text{ is even,} \\ \gamma \cdot (n-1)/2 & \text{if } n \text{ is odd,} \end{cases}$$

where $\gamma = 6.046$Å and $\delta = 3.46$Å (Flory, 1998; Tastan et al., 2009). "Loop stretch" (λ) is the normalised loop span of the observed span between two C_α atoms at each end of a loop in a protein structure over the loop's maximum span. If the span between two terminal C_α atoms in the loop is l, the loop stretch λ of the loop is defined as a normalised span,

$$\lambda \equiv \frac{l}{l_{max}}. \tag{5.1}$$

The values of γ and δ are theoretical approximations and therefore λ of some loops may occasionally be larger than 1. Similar notations are found in Ring et al. (1991) and Tastan et al. (2009).

A loop is stretched or contracted in terms of loop stretch and short or long in respect of the number of residues (Figure 5.1).

5.1 Introduction

Protein structures are hypothesised to be in thermodynamic equilibrium with their environment (Anfinsen, 1973). In other words, the primary determinant of a protein structure is its atomic interactions, i.e., sequence. An analogous conjecture has arisen at the local scale: the modelling of protein loops is often considered a mini protein folding problem (Fiser et al., 2000; Nagi and Regan, 1997).

As seen in Chapter 2, loop classifications and database search prediction methods are based on this conjecture. They classify or predict loops using local properties, e.g., secondary structures from which the loop starts and finishes (anchor region), distance

between the end points of the loop or anchors (span) and loop sequence similarity. In spite of partial successes in the classifications and prediction methods, it is unclear whether or not the assumptions underlying these approaches are valid.

Here, I focus on a specific local property of protein loop structure: span, which is the distance between the two terminal C_α atoms of the loop. I analyse the loop span distribution under various conditions such as anchor types, the number of residues and different protein classes (soluble and membrane proteins). The nature of the distribution is broadly similar across different protein classes or anchor types, except for loops linking anti-parallel β sheets (anti-parallel β loops). In particular, the most frequently occuring span appears to stay the same irrespective of the number of residues. These observations suggests that loop span is an independent local property and is distributed irrespectively of other local properties and global structures. I demonstrate that the observed span distribution can be largely explained by a simple model of random fluctuations with a given length scale, based on the Maxwell-Boltzmann distribution.

I also show that the normalised loop span (loop stretch λ) is an informative indicator for protein loop structure prediction. It is believed that the accuracy of loop structure prediction depends on the number of residues, i.e., the larger the number of residues, The more difficult it is to predict loops accurately (Choi and Deane, 2010; Karen et al., 2007). I show that the accuracy of loop structure prediction is also dependent on loop stretch. Fully stretched loops ($\lambda \simeq 1$) can be predicted accurately using any type of method. However contracted loops ($\lambda \ll 1$) are harder to predict. In fact, shorter loops tend to be more stretched whereas longer loops are likely to be more contracted. Thus it is suggested that loop stretch should be addressed in practical modelling situations and loop structure prediction should be concerned with predicting highly contracted loops.

5.2 Material and Methods

5.2.1 Loop Definition

Secondary structures are annotated using JOY (Mizuguchi et al., 1998a). A loop structure is defined as any region between two regular secondary structures that are at least three residues in length (Donate et al., 1996). Short (less than 4 residues in length) loops were discarded. Redundancy was checked using sequence identity. If a pair of loops share over 40% sequence identity (Fernandez-Fuentes and Fiser, 2006), the loop which has a higher average B-factor was discarded.

5.2.2 Loops from Membrane Protein Structures

Membrane proteins (3789 chains) were taken from PDBTM (Tusnady et al., 2004). A membrane layer is defined from -20 to 20Å (Scott et al., 2008) and loops whose two end C_α atoms are outside the layer were discarded. A total of 1027 non-redundant membrane loops were defined. Henceforth, any loops from the membrane protein structures are called membrane loops.

5.2.3 Loops from Soluble Protein Structures

All protein chains determined by X-ray crystallography which share less than 99% sequence identity (<3.0 resolution and <0.3 R-factor) were collected using PISCES (Wang and Dunbrack, 2003). All the 3789 membrane chains were removed. In order to get rid of potential membrane protein structures in the list, PSI-BLAST (Altschul et al., 1997) was run on the 3789 membrane chains iteratively 5 times. Any chains found (e-value <0.001) were removed from the soluble protein structure list. A total of 25191 non-redundant soluble loops were defined from 27717 soluble protein chains.

5.2.4 Comparison of Loop Structure Prediction Methods

Test Set

Two test sets were prepared for comparison.

- The first set includes non-redundant loops of 8 residues. They were binned every 0.1 loop stretch. In each bin, 40 test loops were randomly selected. A total of 320 test loops from 0.2 to 1 in loop stretch were used (A full list is given in Table D.1). MODELLER (Fiser et al., 2000) as an *ab initio* method and FREAD (Choi and Deane, 2010) as a database search method were benchmarked on this test set.

- The second set consists of loops of 6∼10 residues. Each number of residues includes two classes: contracted loops ($\lambda < 0.4$) and stretched loops ($\lambda > 0.95$). A total of 346 test loops were identified (58, 72, 110, 58 and 48 loops respectively). The number of the contracted and stretched loops are the same in each number of residues, e.g., 55 contracted test loops for loops of 8 residues (A full list is given in Table D.2).

The measurement of accuracy is loop RMSD of all backbone atoms (N, C_α, C and O) after superimposing anchor structures.

MODELLER Setting

The default loop refinement script was used. One hundred loop models were sampled under the molecular dynamics level of *slow*. The DOPE potential energy (Shen and Sali, 2006) was used for model quality assessment.

FREAD Setting

A database was constructed using the 27717 soluble protein chains defined above. All the parameters were set as default (the environment substitution score cut-off value $\geq$ 25). Any results from self-prediction were eliminated.

5.3 Results

5.3.1 Loop Span Distribution

The number of residues in a loop is distributed in a similar fashion regardless of anchor types except for anti-parallel β loops (Figure 5.2A). However, despite the range of the number of residues, the loop span distributions show a nearly constant mode ($\simeq$ 15Å) across loops with varying numbers of residues (Figure 5.2B). Figure 5.2C shows how loop spans are distributed in different anchor types. Again, apart from anti-parallel β loops which are physically constrained by hydrogen bonds, the loop span distributions do not change upon anchor structures. These results suggest that the loop span is distributed independently of local anchor structures.

The loop span distribution also does not alter when considering different protein classes. In Figure 5.2D, loop spans of membrane loops and soluble loops are distributed in a similar manner.

5.3.2 Maxwell-Boltzmann Distribution for Loop Span Distribution

From the observations, it appears that loop span is distributed independently of local anchor structures or global protein class. Here I assume that a protein loop is an independent unit of the protein structure and the span is determined regardless of any other effects including sequence or the rest of the structure.

A model for the loop span distribution is established under the hypothesis that the two end points of loops fluctuate in three dimensional space, following the Maxwell-

Boltzmann distribution. Two constraints are imposed in this model: the minimum span l_{min}, since the end points cannot approach each other too closely and the maximum span l_{max}. Within these constraints, the span oscillates according to a normal distribution with a given length-scale l_{mode} in three dimensional space. Thus, under this model, the

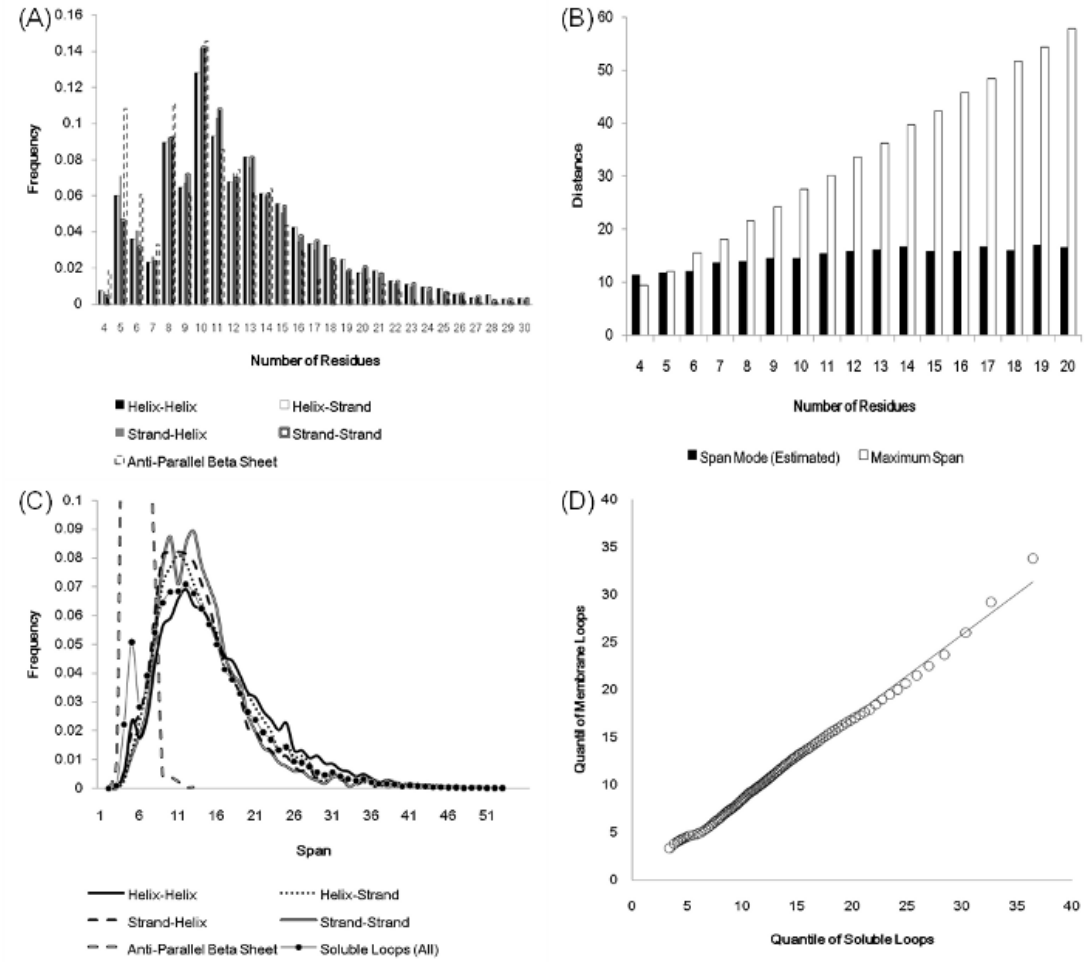


Figure 5.2: Loop span distribution
(A) The frequency distribution of loops containing different numbers of residues is similar across all the anchor types except for anti-parallel β loops. (B) The modes of the loop span distribution stay nearly the same ($\simeq$ 15Å). The modes were estimated by fitting the Beta distribution. (C) The loop span distribution in terms of the anchor secondary structure do not show differences except for anti-parallel β loops. The upper part of the anti-parallel β loop span distribution is omitted in the figure. (D) The Q–Q plot shows that the loop span distributions of soluble and membrane proteins are from the same probability distribution.

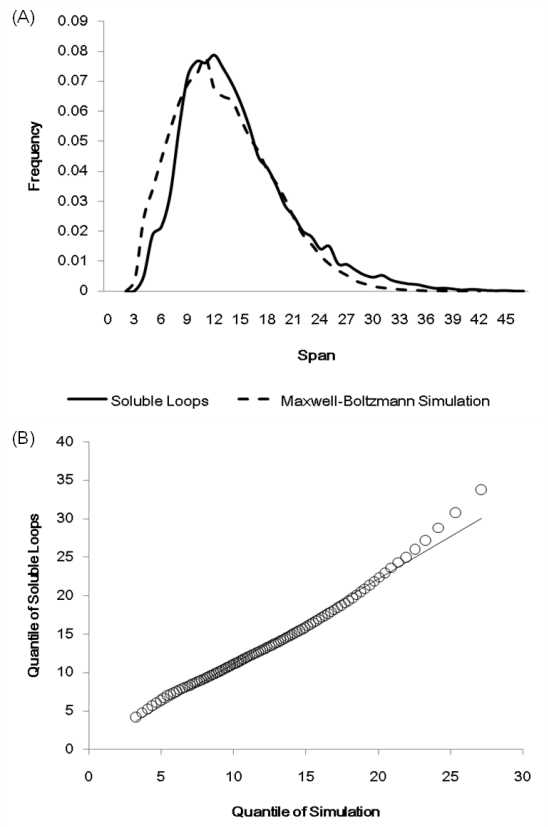


Figure 5.3: Loop span distribution and the Maxwell-Boltzmann distribution (A) The loop span distribution (solid line) from soluble loops shows a good fit with the Maxwell-Boltzmann distribution (dotted line). In the soluble loops, all anti-parallel β loops and longer loops ($>$ 20 residues) were eliminated. (B) The Q–Q plot suggests that they follow nearly the same distribution except for the slight mismatch in the extended span region.

loop span l of n residues is distributed as

$$l = \sqrt{l_x^2 + l_y^2 + l_z^2}, \qquad \text{where } l_x, l_y, l_z \sim \mathcal{N}\left(0, \frac{l_{mode}^2}{2}\right) \tag{5.2}$$

subject to the constraints that $l \geq l_{min}$ and $l \leq l_{max}(n)$, as stated above. The variance of $l_{mode}^2/2$ corresponds to a modal span distance of l_{mode}. Thus there are two parameters to be determined in our model: l_{min} and l_{mode}. l_{min} is set to 3.8Å, which is the typical distance between two neighbouring C_α atoms in a protein chain, while l_{mode} is set to

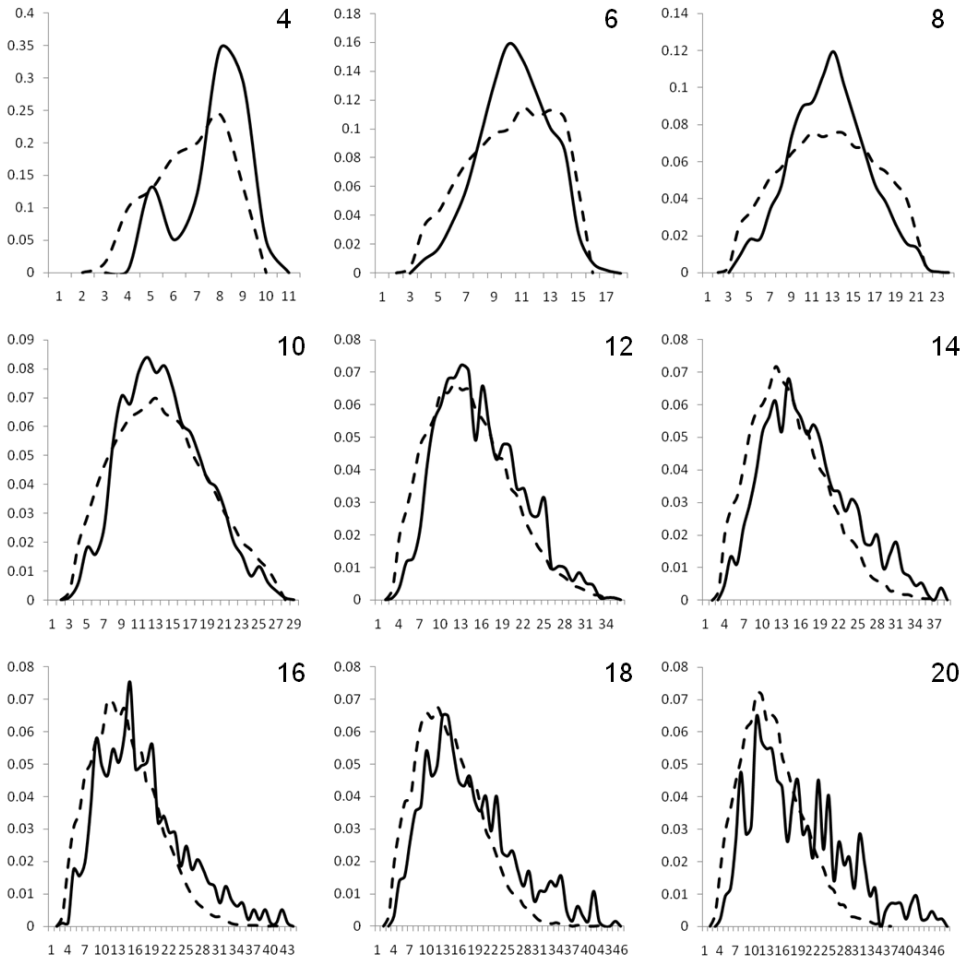


Figure 5.4: Loop span distributions of different number of residues and their corresponding Maxwell-Boltzmann distributions
The dotted lines are from the simulation of the Maxwell-Boltzmann distribution and the solid lines are from actual data. Only even number of residues are shown. The X-axis is loop span and Y-axis is the relative frequency.

an estimate of the empirical mode, obtained by fitting a Gaussian smoothing kernel to the observed loop span distribution.

As there are not many longer loops in the data set, loops longer than 20 residues were discarded. In addition, all anti-parallel β loops were eliminated due to their physical constraints. These eliminations left 23,499 soluble loops. Having set the two parameters l_{min} and l_{mode}, loop spans were generated 10 times per model in accordance

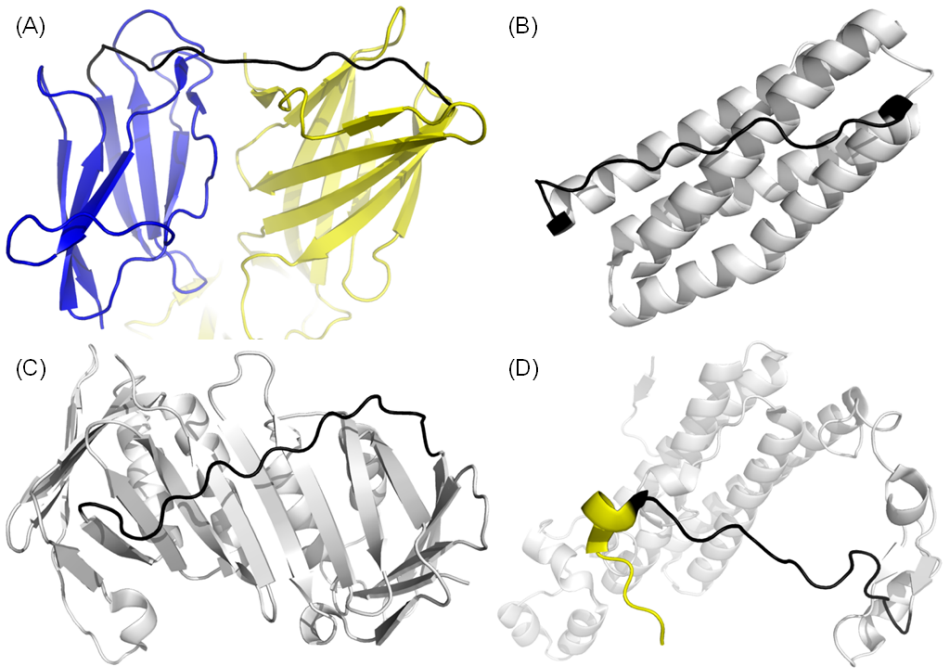


Figure 5.5: Long and extended loops
There are mainly four classes of exceptionally long and stretched loops (black): (A) Domain linker (1OK8, Chain A 291-305), (B) local secondary structure packing (1YV1, Chain A 88–103), (C) global fold (3F1W, Chain A 118–134) and (D) terminal regions (1O9I, Chain A 243–258)

with the Maxwell-Boltzmann distribution, preserving the observed distribution of the number of residues.

The simulation outcome is depicted in Figure 5.3A. The two distributions show the same shape and the quantile comparison in Figure 5.3B indicates that they are statistically similar. In order to investigate the mismatches in the tailed region, the distribution is split in terms of the number of residues (Figure 5.4).

These longer loops in the tail region are further investigated by manual inspection. Such loops are categorised into four types.

- Domain linkers (Figure 5.5A)
- Local effects such as secondary structure packing (Figure 5.5B)

- Specific fold types (e.g., β sandwiches) (Figure 5.5C)
- Terminal regions (Figure 5.5D)

5.3.3 Protein Loop Structure Prediction and Loop Stretch

Loop length is known to be related to stability (Nagi and Regan, 1997) and the accuracy of most loop modelling techniques. However, if a loop contains many residues but is highly stretched, it will be predicted relatively accurately, as it can take on only a small number of different conformations. Here I calculated the normalised loop span, loop stretch and its effect on protein loop structure prediction.

The loop stretch distributions of protein classes and anchor types are no different (Figure 5.6). Figure 5.6D displays how loop stretch frequencies are distributed in different number of residues, demonstrating that the number of residues is negatively correlated with loop stretch, i.e., the longer a loop is, the more contracted. This result suggests that the actual problem of protein loop structure prediction should be in loop stretch. In other words, the relatively accurate prediction of short loops may be due to their high loop stretches.

In order to check the relationship between accuracy and loop stretch, I chose two loop modelling programmes, which have different sampling methods, for the first test set. This test set contains loops of only 8 residues in length and there are 40 non-redundant loops in every 0.1 loop stretch bin. MODELLER (Fiser et al., 2000) is a popular protein structure prediction programme which has a built-in *ab initio* loop modelling module. FREAD (Choi and Deane, 2010) is a database search method which samples candidate loops depending on local properties and rank predictions based on sequence similarity. The measurement of accuracy is RMSD of all backbone atoms.

The average accuracy of MODELLER shows a negative linear correlation against loop stretch (Figure 5.7A). In the case of fully stretched loops ($\lambda > 0.95$), MODELLER can produce consistently accurate predictions, but its predictions worsen as the target

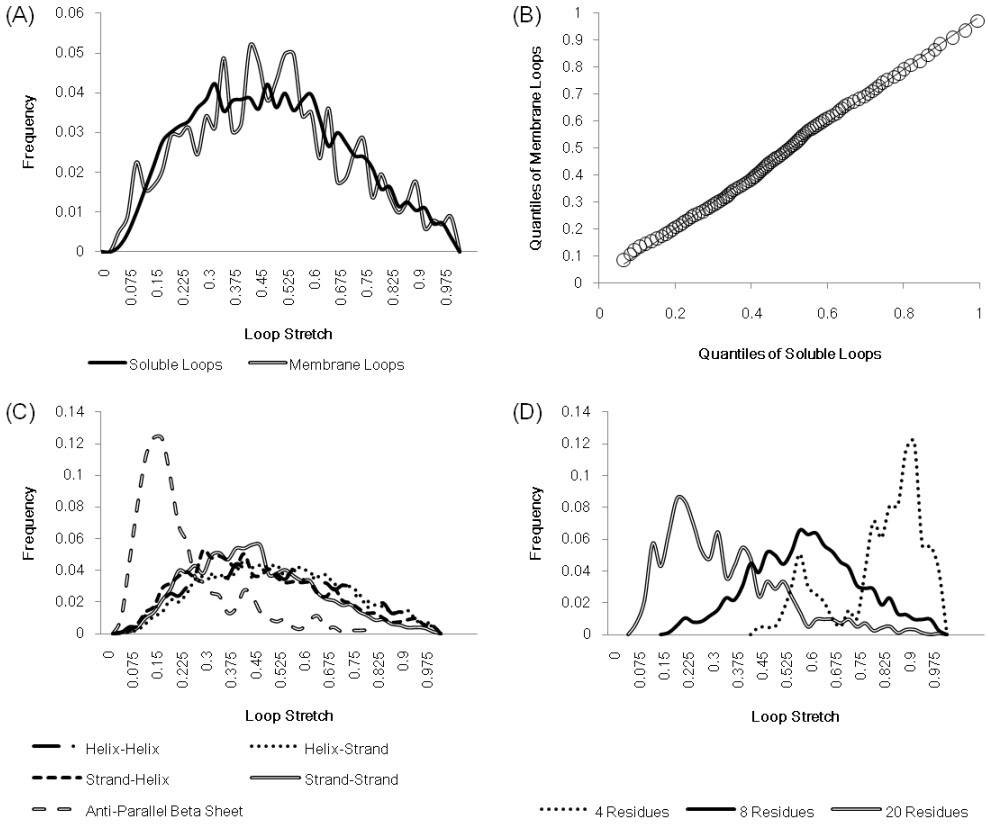


Figure 5.6: Loop stretch distribution
The loop stretch distributions of protein classes and anchor types are no different. ((A) and (B)) The loop stretch distributions of membrane and soluble loops show the same frequency distribution. (C) The loop stretch distribution does not change upon anchor types except for anti-parallel β loops. (D) Shorter loops tend to be more stretched whereas longer loops are likely to be more contracted.

loops are less stretched. FREAD shows stable predictions across all loop stretch bins except for highly contracted loops ($\lambda < 0.4$) (Figure 5.7B).

In order to assess the effect of loop stretch in loop structure prediction, MODELLER was re-examined on the second set. In this test set, loops were divided into two main classes: contracted loops ($\lambda < 0.4$) and stretched loops ($\lambda > 0.95$). This test set contains loops of 6–10 residues in length. MODELLER produced consistently accurate results on this test set for fully stretched loops ($\lambda > 0.95$), but failed to accurately predict contracted loops (Figure 5.7C).

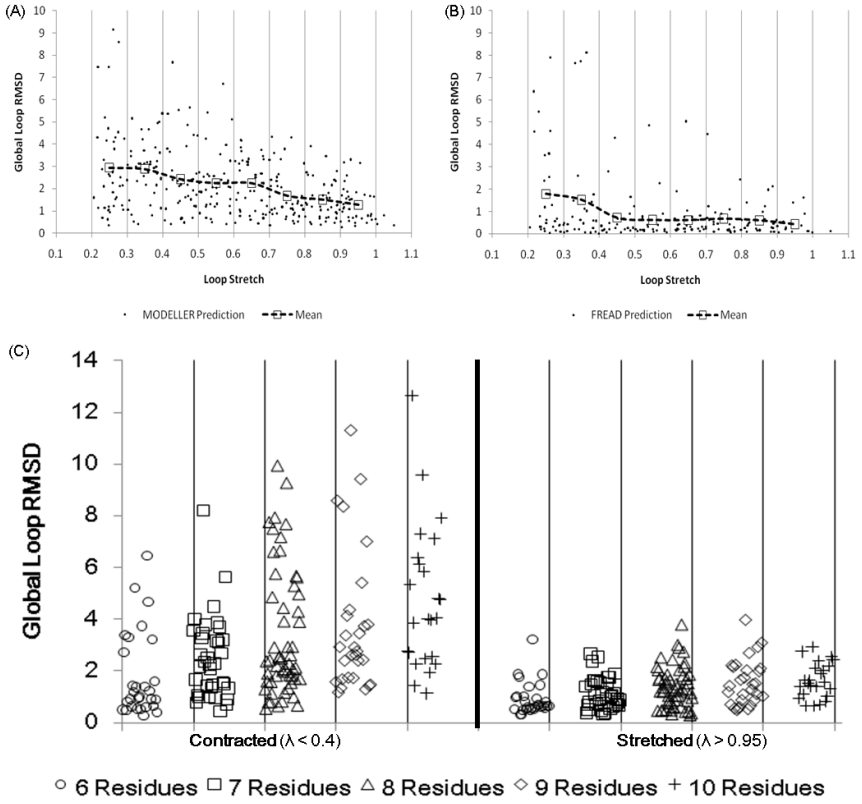


Figure 5.7: Effect of loop stretch on protein loop structure prediction
The accuracy of protein loop structure prediction methods does not only depend on the number of residues, but also on loop stretch. MODELLER (A) and FREAD (B) both show accurate results when the target loop is stretched on the first set (including loops of 8 residues in length only). MODELLER shows worse prediction as loop stretch decreases whereas FREAD gives consistent accuracy on loop stretch. However both methods fail to predict very contracted loops ($\lambda < 0.4$). (C) On the second set, MODELLER gives accurate predictions regardless of the number of residues for stretched target loops, whereas its prediction quality is not reliable for contracted loops.

5.4 Conclusion

In this chapter, I focus on a specific local property (span) and the modes of loop span distribution appear to be independent of the number of residues. Loop span shows a distinct frequency distribution which does not depend on anchor types and protein classes. From the observations, I hypothesised that loop span is given independently of the other effects and tried fitting the loop span distribution with the

Maxwell-Boltzmann distribution. The loop span distribution appears to correspond to a truncated Maxwell-Boltzmann distribution.

It has been believed that the accuracy of protein loop structure prediction is related to the number of residues. Many report that longer loops are harder to predict. I show that the primary determinant of prediction accuracy in fact appears to be loop stretch. As fully stretched loops have limited degrees of freedom, it is not difficult to predict such loops. However, contracted loops are shown to be more difficult to predict.

Therefore, one can formulate the sampling difficulty in protein loop structure prediction. As traditionally known, the sampling difficulty (S) is known to be proportional to the number of residues (n).

$$S \propto n. \tag{5.3}$$

Loop stretch λ is also an informative measure of accuracy. As seen in Figure 5.7C, fully stretched loops can be predicted accurately regardless of the number of residues and the number of residues becomes meaningful when a target loop is contracted. Thus, the sampling difficulty in eq. 5.3 can be rewritten as follows.

$$S \propto n\,(1-\lambda)\,. \tag{5.4}$$

It should be noted that the true nature of the independence of loop span is not understood. However, one can speculate that this result is a piece of evidence that supports the ideas that supersecondary structures[1] are folding units (Fernandez-Fuentes et al., 2010) and protein loop structure prediction is a mini protein folding problem.

[1] A supersecondary structure is defined as two consecutive regular secondary structures and a loop that links them.

Chapter 6

Conclusion and Future Directions

As the number of solved structures expands, homology modelling will become more available and more powerful. However, as CASPs have shown, homology modelling still has problems in predicting regions where no template is available. These regions tend to be loops.

In this dissertation, I have described a revised version of FREAD, a database search method for loop structure prediction. By examining FREAD, I find that there is a certain value of local sequence similarity which ensures structural similarity. FREAD was tested on large test sets and generally predicted loops to within 2Å on average regardless of the number of residues. Its predictive ability also encompasses a more specific class of protein loops, antibody CDR loops.

An important question arises from these predictions: Is protein loop structure prediction a global or local problem?

FREAD uses only local information such as anchor matches and sequence similarity and does not care about the entire protein structure. From the benchmark on CASP model structures, I found no correlation between prediction accuracy and global structure match.

Then, I focused on a local property: loop span, and found that the loop span is

Table 6.1: Analogy between protein structure prediction and protein loop structure prediction

Protein Structure	Protein Loop Structure
Thermodynamic equilibrium with the external environments	Physical equilibrium with the local environment, such as anchor structures
Template selection based on global sequence similarity (Typically 30% sequence identity)	Prediction based on local sequence similarity

an independent property. The loop span frequency was distributed indifferently to protein classes or anchor structure elements (except for anti-parallel β loops). Unlike the previous observations and benchmarks of protein loop structure prediction reports, I demonstrate that prediction accuracy of protein loop structure depends not only on the number of residues, but also on loop stretch.

Therefore, one can draw an analogy between the hypotheses of global protein folding and local loop folding (Table 6.1). Folded proteins (loops) are in thermal equilibrium with their environment. Changes of environment (loop stretch) lead to structural changes. The primary determinant of a protein (loop) structure under the same environment (loop stretch and anchor types) is its (local) sequence. Homologous proteins with similar (local) sequences show similar (local) structures.

My results suggest the following three future applications.

Estimating Experimentally Undetermined Local Structures Using FREAD One of the advantages of using a database search method is that one is informed of where the predicted fragments come from. As FREAD gives accurate predictions, FREAD can be applied to estimate experimentally undetermined local structures (See Figure 6.1 for an example).

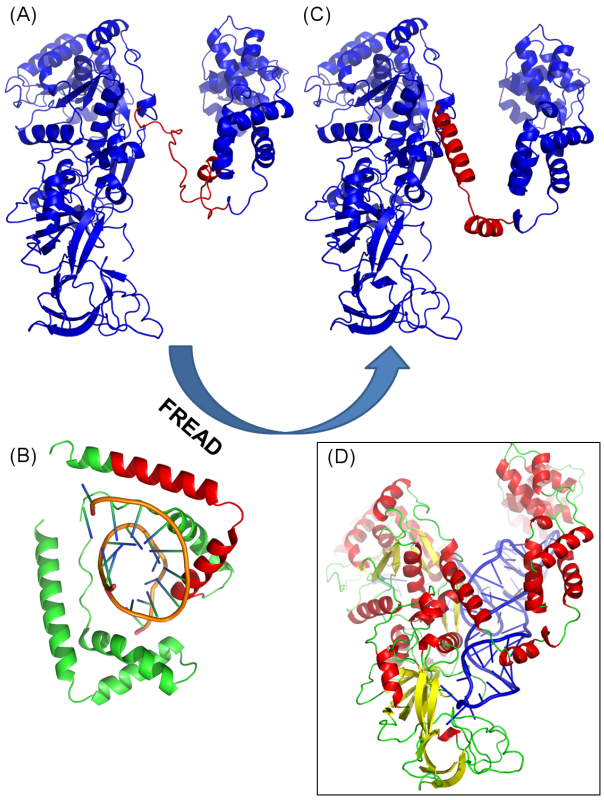


Figure 6.1: An example of function estimation using FREAD
(A) A target structure with a missing link (in red). (B) The link was predicted using FREAD and a possible fragment was found in 1QRV (36–65, Chain A). 1QRV is a DNA-complex and one can estimate that the missing link may have a similar function to 1QRV. (C) The predicted fragment was grafted on the target structure. (D) The target structure was later found to be a tRNA binding protein. Figure D was generated by Edward Snell at the State University of New York, Buffalo.

Predicting Interacting Loops Using Contact Profiles Little work has been done on predicting interacting loops simultaneously so far. Recently, Danielson and Lill (2010) reported an *ab initio* prediction method to predict partial segments (6, 9 and 12 residues) of multi-interacting loops in a protein structure. The prediction quality of multi-interacting loops is significantly lower than that of single loop structures. As FREAD accompanied with the contact profile method (ConFREAD) gave accurate results on more plastic loop structures (CDR), it could perhaps be extended to predicting interacting loops.

A Novel *ab initio* Method Using Loop Stretch In my study of loop stretch, I showed that current *ab initio* methods are able to predict only fully stretched loops accurately. The exact reason for the inaccuracy in less stretched loop prediction should be further investigated. However, one clear cause is the larger conformational space available to less stretched loops. Therefore more intensive sampling is required in order to predict contracted loops.

Most *ab initio* methods sample candidate loops based on dihedral propensities (Section 2.2.1). Therefore the sampling power of an *ab initio* method heavily depends on the quality of dihedral angle libraries. Dihedral angle propensities are different in terms of loop stretch (Figure 6.2) and more precise sampling may be achievable using dihedral angle libraries specific to loop stretch.

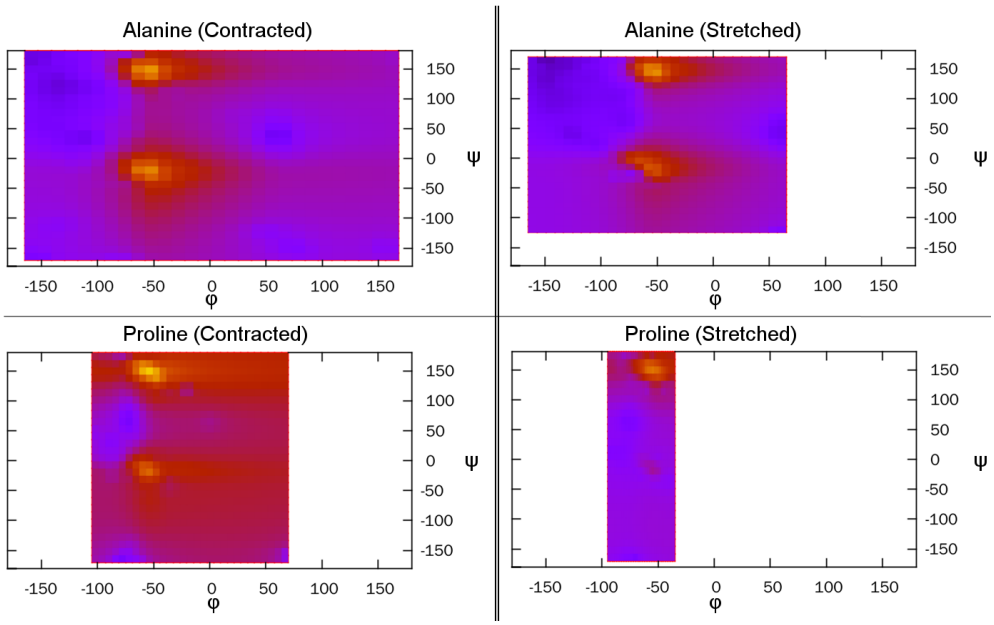
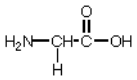
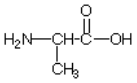
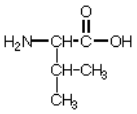
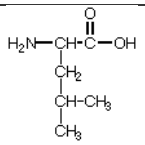
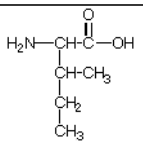
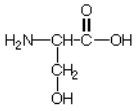
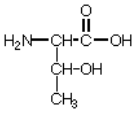
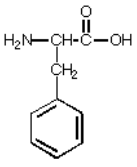
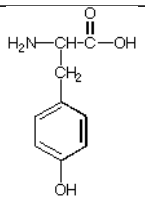
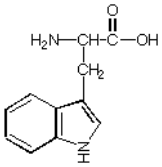
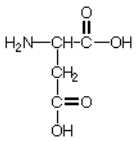
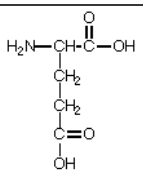
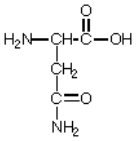
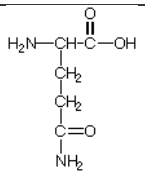
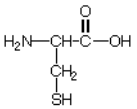


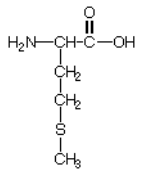
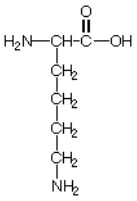
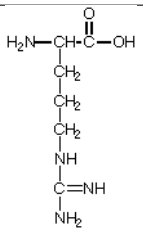
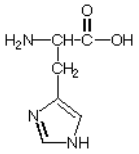
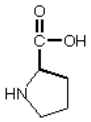
Figure 6.2: Different dihedral angle propensities in terms of loop stretch
The same amino acid has different dihedral angle propensities in terms of loop stretch. The dihedral angle areas of stretched loops ($\lambda > 0.8$) are limited whereas those of contracted loops (< 0.2) are widely distributed. Negative ψ angles are rare in stretched loops.

Appendix A

Introduction

Table A.1: Amino acid structures. The figures were taken from `http://www.biomed.curtin.edu.au/biochem/tutorials/AAs/AA.html`

				
Glycine	Alanine	Valine	Leucine	Isoleucine
				
Serine	Threonine	Phenylalanine	Tyrosine	Tryptophan
				
Asparate	Glutamate	Asparagine	Glutamine	Cystein

				
Methionine	Lysine	Arginine	Histidine	Proline

Appendix B

FREAD Benchmark Test Set

Table B.1: The full list of the set one of the loops in the standard loop benchmark (PDB code, chain, loop length and starting residue)

1FNL	A	4	95	2APJ	B	5	51	2I5V	O	6	91	2NQT	A	7	48
1OXX	K	4	134	2GSD	A	5	174	2FZW	A	6	336	1WA3	E	7	145
1M2J	A	4	11	2BDR	A	5	86	1T6G	D	6	156	2IFQ	B	7	69
3CLS	D	4	135	2QED	A	5	51	1UWC	A	6	205	2R8O	B	7	601
2ELC	A	4	168	3CMS	A	5	33	2BJK	A	6	406	2GKE	A	7	118
2FVY	A	4	41	2I3H	B	5	125	2D51	A	6	25	1V5D	A	7	134
1MG5	A	4	39	2HQJ	A	5	25	1WKR	A	6	179	2G8S	B	7	155
2D81	A	4	272	1T61	A	5	184	2PS2	C	6	245	1KWF	A	7	169
2Z0J	A	4	117	2FCR	A	5	7	2O2P	B	6	793	1YS1	X	7	111
1WAP	N	4	48	1RA0	A	5	338	1IUA	A	6	17	3CMS	A	7	184
1HT6	A	4	258	1YS1	X	5	283	1J31	D	6	223	1BRF	A	7	22
2AXW	B	4	86	1PX5	B	5	112	2VB1	A	6	37	1Z2N	X	7	126
1KJV	A	4	250	2NTP	B	5	175	2EJN	B	6	72	2RB9	D	7	154
2OZL	A	4	185	2OSS	A	5	76	1MV8	D	6	378	2CVD	B	7	102
1W3O	A	4	63	2APJ	D	5	217	3CLM	A	6	65	2DM6	A	7	293
2FXF	A	4	120	2BBK	J	5	267	2P6C	B	6	11	2QJJ	C	7	237
1KQF	A	4	915	1X46	A	5	21	2R2A	B	6	179	3CMS	A	7	143
2UXY	A	4	181	1RA0	A	5	391	2OS0	A	6	7	2FMA	A	7	140
2I47	A	4	330	2I7D	A	5	41	1R45	C	6	118	1WTA	A	7	171
2HO3	C	4	261	2H14	A	5	310	2OTU	E	6	97	1F5V	A	7	221
1WST	A	4	285	2NTP	A	5	118	1WZA	A	6	250	1WKQ	A	7	20
2GF9	A	4	143	1B8P	A	5	123	2VBA	C	6	297	1OXX	K	7	264
3B4U	B	4	80	2OZL	D	5	95	2C43	A	6	302	2CYG	A	7	147
1J0H	A	4	118	1OYG	A	5	319	1W32	A	6	115	2QJW	A	7	127
1DUN	A	4	45	2AEE	A	5	64	1W9H	A	6	354	2E11	B	7	11
2H6L	A	4	34	2ON5	H	5	35	2YVT	A	6	40	1MML	A	7	138
1RU4	A	4	292	1ZKP	A	5	34	1SVI	A	6	115	2QG6	A	7	99
1FBA	A	4	219	1KHB	A	5	185	1K92	A	6	269	1R6X	A	7	195
3CMS	A	4	21	2PBP	A	5	93	2V27	A	6	255	1PN2	C	7	21
1NDB	B	4	469	2B49	A	5	793	2CN3	B	6	87	2H6E	A	7	24

1HN0	A	8	580	2BBK	J	9	288	1DY2	A	10	91	1UMG	A	11	332
2Z3H	A	8	21	2UW1	B	9	50	2H61	B	10	40	2C1V	A	11	214
1SW0	A	8	170	1OX0	A	9	89	2UZI	H	10	99	1EI5	A	11	418
1ZZ1	C	8	309	2B82	B	9	113	2FT6	A	10	9	2A6V	B	11	188
2YQU	A	8	430	1QVZ	B	9	49	1RXQ	D	10	85	2E5F	A	11	126
3BIQ	A	8	134	1K7J	A	9	174	2DR1	B	10	39	1F9Y	A	11	132
1NWP	A	8	84	1Y0P	A	9	156	1NDB	A	10	216	2AQ8	A	11	149
3BWU	D	8	24	1FNL	A	9	73	1V0Z	D	10	169	1NB9	A	11	90
2PW4	A	8	147	2BV4	A	9	95	1BQC	A	10	27	1JU2	B	11	373
1RKQ	B	8	146	2H8G	A	9	139	1R7A	A	10	470	2RIK	A	11	187
2BEM	B	8	149	1IV8	A	9	43	1M6S	A	10	119	1H03	Q	11	67
1RH9	A	8	360	1UN3	A	9	14	2C1S	B	10	54	2AYD	A	11	301
1JUH	A	8	20	2FAR	A	9	594	2Q8R	H	10	9	2CKI	A	11	160
1QFM	A	8	639	2PRD	A	9	75	2JDA	A	10	16	2GB4	A	11	31
2DSN	B	8	314	2JBA	A	9	40	1ZK7	A	10	307	2Z0J	F	11	7
1DJT	B	8	38	1KJV	B	9	12	1XV2	A	10	141	1R6X	A	11	257
2OJ6	D	8	432	2GDQ	B	9	337	1OOH	B	10	56	2QP8	B	11	419
2GAG	A	8	572	1PX0	D	9	80	2FDV	B	10	112	2RGM	A	11	322
1HDH	B	8	13	2VEF	A	9	112	1B6G	A	10	13	2GAK	B	11	179
2H3B	B	8	300	1MKA	B	9	114	2RDQ	A	10	230	1QWL	A	11	407
1ENF	A	8	178	1EZG	A	9	18	2Q4G	Y	10	148	1PA2	A	11	220
1PK9	C	8	206	1CQX	B	9	210	2CST	B	10	123	2FR5	C	11	27
2B1X	A	8	310	2Z98	A	9	93	1P1J	A	10	518	2SIC	E	11	33
1TAF	A	8	32	2DSN	A	9	339	3CLM	A	10	297	2OT4	B	11	250
1GWM	A	8	54	1KQF	A	9	528	1NVR	A	10	246	1JIG	C	11	80
1UI0	A	8	162	1WOQ	B	9	134	1YOC	B	10	109	1JHD	A	11	370
1VL2	C	8	114	2CB5	A	9	90	1DBF	C	10	78	2PEZ	A	11	92
1TJY	A	8	91	1H1N	A	9	130	1Q7E	A	10	259	2IWK	A	11	395
2VO9	C	8	125	2OB5	A	9	130	2I51	B	10	44	2CAR	B	11	141
2IW1	A	8	277	2JHF	A	9	92	1U5D	B	10	118	1UG6	A	11	174
1P0I	A	12	224	1DW0	B	13	31	3CLS	D	14	190	1Y4W	A	15	109
1QW9	B	12	213	2RB7	B	13	324	1QJW	B	14	302	2FAF	A	15	535
1ALV	B	12	180	1AJ8	A	13	302	1JHD	A	14	295	2GZ1	B	15	94
2BW8	B	12	111	1SFF	B	13	240	2DCF	A	14	203	2Z9W	A	15	236
2PET	A	12	124	1EC7	C	13	160	1Y2K	A	14	351	1QWO	A	15	178
3BC8	A	12	421	1VRM	A	13	276	1OOY	A	14	320	2J6G	A	15	207
2YWI	B	12	172	1C9O	A	13	33	2CD7	A	14	55	2NUW	B	15	251
2DVM	D	12	71	2DDT	B	13	91	1VMF	C	14	68	2CDU	A	15	313
2G0W	A	12	77	3B8Z	A	13	321	1OCK	B	14	63	2EAB	B	15	596
1VBU	B	12	647	1RX0	A	13	208	1JP4	A	14	153	2AB0	B	15	113
1QTW	A	12	146	1J31	C	13	40	2FHF	A	14	882	1MG4	A	15	133
1Q7F	B	12	771	1HT6	A	13	108	2NT0	A	14	343	1JNR	C	15	213
1KQF	A	12	678	1JV1	A	13	109	1X25	B	14	32	2NAC	B	15	323
2RDE	A	12	190	2RFM	A	13	88	2D0W	A	14	98	1ZL0	B	15	198
1JIX	A	12	221	2GOU	A	13	235	2BJK	B	14	388	1R6X	A	15	161
3B8X	A	12	184	1VGM	B	13	140	2BC0	A	14	432	2O0R	A	15	199
1UAY	A	12	117	2P0A	A	13	232	1T5O	B	14	153	1V73	A	15	93
1YNP	A	12	15	1SAU	A	13	35	1CNV	A	14	188	2EO5	A	15	47
1YPB	I	12	53	2RIJ	A	13	149	2NW8	B	14	34	2QXX	B	15	43
1UDH	A	12	124	1Q8F	C	13	66	1HXH	C	14	87	1P5X	A	15	157
1L8N	A	12	444	1B8A	A	13	269	2GB4	A	14	182	1LQA	B	15	49
1T2W	C	12	121	2INU	C	13	347	1YQZ	A	14	415	1GCI	A	15	74

1TO4	B	12	120	1F60	A	13	345	1ONW	A	14	134	2YZH	A	15	118
1V1H	C	12	345	2J7P	A	13	86	2DEB	A	14	394	1MPG	A	15	143
1LOP	A	12	86	1MUC	A	13	317	1N2Z	A	14	161	1CCW	D	15	172
3BZW	C	12	202	2RC3	C	13	98	1V7Z	F	14	221	1E4C	P	15	59
2Q99	A	12	124	1LAM	A	13	9	1LLF	A	14	437	1BLX	B	15	30
1OZ2	A	12	231	1ZED	A	13	158	3CLS	C	14	14	1LLF	A	15	388
1YPH	C	12	69	1C2A	A	13	26	2QUL	A	14	213	2CCH	A	15	233
1U60	D	12	89	1V0Z	C	13	386	1O91	C	14	622	2DEP	A	15	298
2RB9	D	16	188	2C0H	A	17	278	1VLB	A	18	682	1HH8	A	19	168
1K7C	A	16	7	2QMJ	A	17	200	2QEW	A	18	196	1M2X	C	19	220
3CJY	A	16	129	1SH7	B	17	180	1WDP	A	18	224	1EK6	A	19	137
1Q7E	A	16	126	1AJ2	A	17	140	1U6R	B	18	62	1R03	A	19	77
2UUY	A	16	184	1ODZ	A	17	79	2G6F	X	18	14	1C7S	A	19	479
2DVT	B	16	233	3B8I	E	17	118	2Q3U	B	18	121	1XLQ	B	19	30
2ZDR	A	16	329	2OH5	A	17	91	2OW6	A	18	870	1NC5	A	19	193
2NTP	B	16	141	2GJU	B	17	138	2V9T	B	18	286	1VLG	F	19	66
1BRT	A	16	57	1G9G	A	17	158	1VA4	C	18	119	1PFV	A	19	517
2PGZ	C	16	158	1WX1	A	17	116	2F2B	A	18	34	2ORD	A	19	254
1PJC	A	16	153	1JFX	A	17	164	1JO8	A	18	6	1PA2	A	19	112
1UCS	A	16	41	1Y9Z	B	17	157	1UG6	A	18	37	1BRT	A	19	123
1P5X	A	16	70	1YS7	A	17	87	2PZM	A	18	121	1K7H	A	19	96
2RFG	B	16	258	1OWL	A	17	159	2J1G	C	18	214	1W7C	A	19	617
1RDQ	E	16	273	1T64	A	17	264	1XG0	A	18	16	2Q3U	A	19	120
1DXE	B	16	112	1Y12	B	17	10	2A28	C	18	6	1OI2	B	19	143
2AAX	A	16	746	1YDY	B	17	131	2PTR	A	18	287	1ZCJ	A	19	279
2V8H	D	16	205	2PYW	B	17	65	3BI1	A	18	674	1HDH	A	19	136
2H3B	A	16	231	1DBX	A	17	97	2BJQ	A	18	15	2AQ2	B	19	87
1A12	C	16	92	2Q8K	A	17	88	2NT0	A	18	186	2GMN	A	19	200
2RHW	A	16	70	1F61	A	17	99	1EK6	B	18	190	1A58	A	19	45
2R37	A	16	169	1KAP	P	17	202	2RKV	A	18	188	2APJ	A	19	76
2EAB	A	16	615	2C6U	A	17	163	2IVF	A	18	801	2Z30	A	19	370
2GOU	A	16	23	2BJF	A	17	258	1UMD	C	18	267	2I57	B	19	59
2J3X	A	16	110	1NFF	A	17	192	1XD3	C	18	77	2Z1M	C	19	131
2OLM	A	16	94	2FHF	A	17	900	1OQ1	D	18	84	1IEJ	A	19	172
2Q62	D	16	71	2FAR	A	17	814	2Z6O	A	18	103	1OAC	B	19	607
1CXP	C	16	310	1DS1	A	17	124	1CPQ	A	18	55	1WMD	A	19	293
2PHJ	A	16	36	2VBA	B	17	218	2B1X	C	18	395	2BRY	B	19	342
1NBA	B	16	95	1UPS	A	17	37	2Z8G	A	18	435	1ODM	A	19	119
2PMQ	A	20	345	2OT4	A	20	162	2B82	B	20	185	1VBI	A	20	138
2QMC	B	20	399	1K7H	B	20	96	1EEX	A	20	516	1GQI	B	20	159
1V0Z	D	20	224	1WLG	A	20	159	2IXT	B	20	167	1JI1	B	20	12
1T64	A	20	137	2D81	A	20	301	1XCR	B	20	38	1F2T	A	20	49
2DJF	B	20	298	1W8O	A	20	151	2QQJ	A	20	309	1JNR	C	20	126
1IOO	B	20	38	1XVY	A	20	261	2ECU	A	20	81	1QGJ	A	20	108
3BDI	A	20	62	2G8S	A	20	168	1TVN	A	20	259	2GTD	D	20	162
1Y6V	A	20	112	1HJS	A	20	304								

Table B.2: The full list of the set two of the loops in the standard loop benchmark (PDB code, chain, loop length and starting residue)

1R6J	A	4	222	1V2D	A	5	110	1W6G	A	6	357	2C1I	A	7	326
1MXR	A	4	46	1ODM	A	5	274	2HO3	A	6	93	1H65	B	7	193
1W2W	J	4	268	1R6J	A	5	231	2P5K	A	6	32	6CEL	A	7	133
1OFL	A	4	241	1PN0	D	5	492	1VGG	C	6	49	2FOZ	A	7	25
2DXC	L	4	155	1BRF	A	5	6	1PMM	D	6	425	2JDI	D	7	39
3CMS	A	4	206	1VLP	A	5	270	1A9X	A	6	797	1RK6	A	7	249
1WVF	A	4	121	1Y93	A	5	143	2Z1A	A	6	274	1HFE	M	7	66
2MCG	1	4	131	1V0L	A	5	44	2VBF	A	6	474	2HEU	C	7	351
3BOJ	A	4	81	2R8O	B	5	451	1DQG	A	6	29	1O7E	B	7	102
2E0T	A	4	171	1IAB	A	5	179	8ACN	A	6	154	1YYD	A	7	50
1ESI	A	4	94	2ET1	A	5	90	1KV7	A	6	142	2B4Y	B	7	175
1XRU	A	4	5	1NXM	B	5	87	1OYG	A	6	240	2AD6	C	7	368
3BB0	A	4	469	2JER	H	5	116	1LAM	A	6	329	2CBP	A	7	34
1FJ2	B	4	49	2HQJ	A	5	25	2HXT	A	6	300	1WVF	A	7	271
2R85	A	4	235	1QVE	A	5	73	2BJF	A	6	294	2I8G	B	7	59
1VLP	D	4	170	1SMR	E	5	33	2FYF	B	6	353	1YRR	A	7	247
2DDX	A	4	258	2ET1	A	5	68	1FXD	A	6	25	1SR4	A	7	91
2BHU	A	4	340	1NDB	A	5	446	1E19	A	6	104	1NXM	B	7	15
2D81	A	4	180	1ODM	A	5	46	1NKG	A	6	15	3BZW	E	7	186
2CB0	A	4	111	2A2Q	T	5	41	1RVK	A	6	207	2Z9W	B	7	56
2OFC	B	4	95	1IDS	A	5	119	2E4T	A	6	241	1LNI	B	7	83
1L6W	J	4	166	2GAI	A	5	350	1RJD	B	6	190	2E7Z	A	7	404
2JER	H	4	58	1MDL	A	5	238	1PG4	A	6	304	1EU1	A	7	552
2V0C	A	4	160	3BI1	A	5	177	1BQC	A	6	154	1WVE	A	7	46
2E6F	B	4	183	1IDS	B	5	54	1N1T	A	6	251	2I3H	A	7	133
2BLN	A	4	74	2V3G	A	5	71	2O5G	A	6	39	1W9H	A	7	75
2HO3	A	4	112	1E4Y	B	5	8	1Y07	A	6	106	2GC7	I	7	52
1NNH	A	4	73	1W66	A	5	191	2FQP	D	6	80	2O07	B	7	274
1PPO	A	4	135	2MCG	1	5	13	1YKD	A	6	360	2RB9	B	7	154
1KCZ	A	4	266	1RA0	A	5	391	1O04	E	6	422	3BM1	A	7	146
1CI9	A	8	338	2O6X	A	9	59	2QZU	A	10	375	1UG6	A	11	174
1M22	A	8	348	1UF5	A	9	44	1CG5	A	10	44	1K3I	A	11	486
1M22	A	8	141	1BYR	A	9	88	2CHO	A	10	265	1V5V	A	11	73
2HQS	A	8	159	2MCG	1	9	110	2QFE	A	10	701	1Q6H	B	11	130
8ACN	A	8	209	1I9Z	A	9	581	2CZQ	B	10	15	1P1X	A	11	168
1G0O	D	8	168	2JAE	B	9	150	2NOO	A	10	134	1R4P	A	11	56
1AOH	B	8	57	3SIL	A	9	57	1LUC	B	10	215	1B8P	A	11	90
1KV9	A	8	536	1ARB	A	9	90	2QEW	A	10	128	1V0E	D	11	675
2HK0	A	8	11	1KJV	B	9	12	1U6R	B	10	189	1ITU	A	11	41
1PKH	A	8	169	1CVR	A	9	403	2OR2	A	10	45	2P51	A	11	226
2F9F	A	8	94	1LZJ	A	9	159	1VEF	A	10	77	2ELC	C	11	79
1YKI	C	8	202	3BKX	B	9	106	1MI3	D	10	284	1BX4	A	11	250
1UOH	A	8	35	2R2D	A	9	144	1LYV	A	10	352	1OJJ	A	11	97
2IBJ	A	8	36	1ESW	A	9	306	1JZ8	B	10	641	2BC0	A	11	191
1GPE	B	8	129	1RG8	A	9	123	1NOF	A	10	255	2QZU	A	11	149
1G6S	A	8	264	1Q0R	A	9	209	1W6G	A	10	285	2QNI	A	11	123
2CNQ	A	8	233	1LOE	C	9	9	1UFY	A	10	77	1NHS	A	11	180
1WBA	A	8	83	1Y6V	A	9	66	1HY7	A	10	168	1HDH	B	11	207
2OZ3	A	8	329	1RM6	B	9	53	1YZV	A	10	84	1QF8	A	11	16

1F61	B	8	39	1MUG	A	9	44	1YG9	A	10	52	1GKP	E	11	284
1DQZ	B	8	754	2UW1	B	9	50	1XLQ	A	10	72	1GQV	A	11	114
2BV5	A	8	460	1PJ5	A	9	735	1V7W	A	10	487	1XG4	A	11	211
1OTK	B	8	25	1K20	B	9	153	1QH8	D	10	465	1RU4	A	11	354
2H98	A	8	146	1IAG	A	9	85	1M3Z	D	10	127	7A3H	A	11	67
2C2P	A	8	85	1RL0	A	9	152	2AD6	C	10	434	1SXV	A	11	46
2P2S	B	8	156	1UQR	L	9	113	1OA4	A	10	109	1KQ3	A	11	114
2D4V	D	8	230	2QPX	A	9	37	1JTP	B	10	100	3BFM	A	11	111
2UU8	A	8	97	1QKS	A	9	248	1OYG	A	10	364	1YS7	B	11	188
2B6D	A	8	492	1W23	B	9	232	1TA3	A	10	223	2IMQ	X	11	263
1J0H	A	8	462	1GWE	A	9	134	1U6E	B	10	80	1VHE	A	11	124
1W5R	A	12	96	2HLJ	A	13	43	1JK7	A	14	18	2GSO	B	15	405
1AE9	A	12	297	1F60	A	13	232	1RM6	D	14	141	3BFM	A	15	130
2I47	B	12	264	2GF9	A	13	45	1B25	B	14	559	2FMP	A	15	301
1JDL	A	12	93	2V0H	A	13	185	1YFQ	A	14	240	1HN0	A	15	682
2DXC	J	12	46	1ET9	A	13	142	1JNR	C	14	56	1G8K	E	15	109
1W96	B	12	493	2Q0S	A	13	185	1K55	C	14	139	1THG	A	15	401
1XCR	B	12	219	1ZZE	A	13	232	2BKL	B	14	452	1EDG	A	15	35
2IHT	B	12	350	2I51	A	13	141	1T9H	A	14	221	2GCO	A	15	28
2O04	A	12	17	2ISW	A	13	302	1JHD	A	14	295	2B6D	A	15	551
1EQC	A	12	20	1SR4	C	13	71	2IVF	A	14	782	1SVB	A	15	146
1YQ2	E	12	335	1ITW	A	13	643	2B4P	B	14	286	1HP1	A	15	172
1OEW	A	12	171	2O9A	C	13	87	2A9S	A	14	107	1MG4	A	15	133
1E5K	A	12	13	1YS1	X	13	291	1UWW	A	14	59	2PE4	A	15	372
2CVD	D	12	136	1YO3	B	13	53	2JE8	B	14	171	2NW2	A	15	143
1V7W	A	12	84	1GNX	A	13	280	1MV8	C	14	320	1W85	D	15	28
1GYO	B	12	11	1ICP	A	13	241	1XU9	A	14	119	2HRG	A	15	152
2G8J	A	12	53	1YC9	A	13	127	2C0H	A	14	40	1HYO	B	15	643
1N1T	A	12	592	2ERX	A	13	29	1Q92	A	14	195	1Y4W	A	15	109
3BIX	B	12	570	1XX1	A	13	74	2GAG	A	14	744	2UVJ	A	15	289
1QPC	A	12	455	1RZL	A	13	74	2JAM	B	14	235	2INU	B	15	187
2YWI	B	12	172	2D1S	A	13	225	2OQZ	A	14	212	2NO4	B	15	196
1RCQ	A	12	157	1QKS	A	13	70	1LAM	A	14	128	2Z9W	B	15	236
2BO4	B	12	211	1D4O	A	13	117	1GQI	A	14	445	1QWO	A	15	178
1DJ0	A	12	103	2EAB	B	13	683	2HXT	A	14	19	1NHS	A	15	419
1LUC	A	12	158	2D51	A	13	91	1GTE	D	14	593	1ZOV	B	15	33
1L3S	A	12	322	2UUY	A	13	93	1HFE	L	14	277	1BHE	A	15	121
1USL	D	12	35	1VGM	B	13	140	1OFW	B	14	245	2O07	B	15	253
2FNU	B	12	179	1J1N	A	13	199	1GQI	B	14	361	1KB0	A	15	421
1EU1	A	12	159	2GSD	A	13	307	1ITW	A	14	158	2J0A	A	15	216
1DS1	A	12	282	2P2W	A	13	134	1QNR	A	14	202	1YQW	R	15	180
2NTP	B	16	141	2OH5	A	17	91	1WVE	A	18	202	2GMN	A	19	200
1VZY	A	16	90	1PN0	B	17	174	2F2B	A	18	34	1UF5	A	19	126
1WDP	A	16	399	2D4V	D	17	350	1UMH	A	18	123	2BSJ	B	19	54
2Z30	A	16	346	2OLR	A	17	173	2OZL	A	18	257	1DMR	A	19	739
1IZC	A	16	12	3B96	A	17	39	2TPS	A	18	151	1W7C	A	19	617
1JQ5	A	16	306	2HXW	B	17	168	2DBQ	A	18	134	1OFL	A	19	180
2IQ7	D	16	110	1W9P	A	17	96	1FCQ	A	18	185	1C7S	A	19	479
1Y0E	A	16	144	1JDW	A	17	339	1WVF	A	18	88	2J9C	C	19	37
2RBD	B	16	72	2FFU	A	17	89	1NPY	A	18	182	1JNR	D	19	104
1KCZ	B	16	70	1CS1	B	17	174	1YQ2	E	18	100	1HDH	B	19	246
2VJQ	B	16	153	1KIC	A	17	71	1EYB	A	18	98	2CJG	A	19	316

2V27	A	16	104	1GA8	A	17	245	1GAI	A	18	293	1HBN	E	19	60
1OR0	D	16	242	1Y37	B	17	58	2RKV	A	18	188	2H88	O	19	55
1VPS	D	16	177	1G9G	A	17	158	1RP0	A	18	71	1DPE	A	19	68
1H1N	A	16	276	2JIG	A	17	144	1UG6	A	18	37	1HDH	A	19	136
1LVW	B	16	7	3BB7	A	17	333	2OSX	A	18	281	1QL0	A	19	150
2G8J	A	16	296	1QMG	B	17	573	1ON3	E	18	64	1OCY	A	19	501
1H1N	B	16	201	2R2D	E	17	21	2H4P	A	18	169	2I0K	A	19	118
1WDV	B	16	92	2QMJ	A	17	200	2F5V	A	18	76	1RM6	C	19	31
1VEM	A	16	289	1N28	B	17	22	1GZ7	B	18	484	2CB2	F	19	52
2VJQ	A	16	389	2CN3	B	17	147	1PL3	A	18	147	1IOO	B	19	38
2ISA	C	16	94	2H8Z	A	17	267	2JD1	B	18	190	1LKO	A	19	64
2AKA	B	16	53	2HVM	A	17	104	1T3Q	B	18	5	2J1N	C	19	105
1OFW	B	16	17	1GVF	B	17	174	2FAF	A	18	215	1Y0P	A	19	475
2CHO	B	16	545	2HOX	B	17	204	1RA0	A	18	63	1Y5I	A	19	74
1T3T	A	16	617	1HL2	A	17	262	1E25	A	18	213	2A0M	A	19	146
1ITW	A	16	704	1LZL	A	17	223	1QSA	A	18	360	1KWG	A	19	482
1GTE	D	16	318	1M15	A	17	90	2EVE	A	18	77	1XLQ	A	19	30
1IM5	A	16	95	1CCW	B	17	402	1MSK	A	18	912	1DLJ	A	19	241
1DDJ	C	16	728	2CZQ	A	17	161	1QKS	A	18	99	1WMW	D	19	219
1W8O	A	20	151	2D81	A	20	301	1H4R	A	20	151	1MJ5	A	20	62
2Q0Z	X	20	237	2AFW	A	20	321	1Y7W	A	20	252	2QP2	A	20	270
1M0Z	A	20	179	1VR5	B	20	359	1OYC	A	20	289	2ISA	B	20	270
3BIX	A	20	509	1F2T	A	20	49	1G2O	C	20	142	1Q6Z	A	20	274
1UX6	A	20	890	2CN3	A	20	395	3BDI	A	20	62	1W2L	A	20	21
2O7I	A	20	13	1Y7B	B	20	268	1DHK	A	20	295	1QHF	A	20	115
2EZ9	A	20	544	1ELV	A	20	496	1P4K	C	20	345	1NE7	D	20	138
1Y0P	A	20	48	3BI1	A	20	539								

Table B.3: The full list of the loops in the CASP benchmark. The residue numbers are from the corresponding native structures (all from chain A)

Target	PDB	Loop Range				
T0291	2GSF	601-608	771-786			
T0292	2JAV	77-80				
T0297	2HSJ	79-81	137-144	179-182		
T0301	2H9F	28-36	51-67	153-155		
T0303	2HSZ	200-202				
T0313	2H58	483-485	636-641	660-663	682-689	
T0318	2HB6	10-21	81-88	99-101	118-135	140-143
T0322	2HBO	40-42	64-66			
T0325	2I5I	199-205	215-231			
T0326	2H2W	79-81				
T0328	2HAG	232-235				
T0330	2HCF	162-164				
T0334	2OAL	42-46	98-102			
T0339	2HDY	339-343	349-352	366-376	386-394	
T0341	2HO4	21-23				
T0342	2I5T	54-63				
T0362	2HX5	55-57	60-62			

T0364	2HLJ	45-54	117-119				
T0365	2IIU	72-74	109-113	148-155			
T0370	2IAB	104-108	130-135				
T0371	2HX1	208-210					
T0373	2HR3	27-33	85-90	139-143			
T0376	2HMC	144-148	231-233	255-258	278-283		
T0380	2HQ7	55-57	106-108				
T0389-D1	2VSW	72-76					
T0394-D1	3DCY	72-79	105-107	110-115	212-228		
T0397-D2	3D4R	136-143					
T0399-D1	3D4E	145-155	168-172				
T0401-D1	3D5P	48-50	70-72	104-108			
T0402-D1	3DB0	50-56	69-71				
T0406-D1	3DI5	63-68					
T0407-D1	3E38	79-88	132-140	159-173	220-222	225-228	
T0407-D2	3E38	283-285					
T0411-D1	3D1P	67-69	77-88				
T0412-D1	3D3O	64-66	83-91	129-134			
T0413-D1	3D0K	59-61	216-230	270-275			
T0414-D1	3D0J	30-39	67-70	111-113	118-129		
T0415-D1	3D6W	114-116					
T0417-D1	3D3S	25-31	129-131	141-143			
T0419-D2	3CZP	292-295	377-379	386-395	427-432	436-439	
T0420-D1	3D5N	125-129	146-149	155-160			
T0421-D1	3CZQ	161-168	175-189				
T0422-D1	3D8B	474-476	509-514	531-536			
T0425-D1	3CZX	56-58	126-128	131-139			
T0427-D1	3D3Y	27-30	79-82	91-97			
T0429-D2	3DB3	231-236	262-268				
T0430-D1	2VUW	519-527					
T0430-D2	2VUW	756-770					
T0431-D1	3DAX	38-49					
T0431-D2	3DAX	114-123	155-164	193-196	224-226		
T0432-D1	3DAI	1065-1069					
T0436-D1	3D6K	45-47	59-64	104-106	113-127	204-210	235-243
T0440-D1	3DCP	53-57	63-65	191-197			
T0441-D2	3D8U	223-228					
T0443-D1	3DEE	74-86					
T0447-D1	3DO6	155-157	275-277	386-392	537-539		
T0448-D1	3DC7	67-79	142-144				
T0449-D1	3DCD	58-62	65-68	97-99	197-201	258-271	
T0456-D1	3DFA	46-50					
T0456-D2	3DFA	138-142	156-168				
T0457-D1	3DEV	46-48	65-76				
T0457-D2	3DEV	226-240					
T0460-D1	2K4N	8-20	40-49				
T0461-D1	3DH1	92-94					
T0463-D1	3DHN	44-47	78-83	121-125	147-151	181-189	
T0464-D1	2K5R	40-42					
T0477-D1	3DKP	269-273	276-284	303-306			
T0478-D1	3D19	47-50	63-75				
T0479-D1	3DKZ	62-71	115-118	136-141			

T0483-D1	3DLS	1035-1049				
T0486-D1	2VX2	182-184				
T0487-D1	3DLB	481-486	507-512	584-588		
T0487-D3	3DLB	32-38	72-74			
T0487-D5	3DLB	334-340	386-390			
T0489-D1	3DL1	96-98	133-144	159-177	191-200	
T0490-D1	3DME	59-62	217-222	250-253	321-324	
T0491-D1	3DM4	81-83				
T0493-D1	3DMN	615-622	652-655			
T0494-D1	2VX3	215-217	429-431	434-443		
T0495-D1	3DNX	74-78	85-88	100-104		
T0497-D1	3DMB	43-45				
T0501-D1	3DMA	69-85	108-113	129-134		
T0501-D2	3DMA	246-257	310-314			
T0503-D1	3DN7	9-12	81-85			
T0504-D1	3DLM	34-37				
T0507-D1	3DO8	78-80				
T0509-D1	3DR5	54-57				
T0510-D1	3DOA	71-76	105-114			
T0511-D1	3E03	41-49	151-161	170-172	209-213	
T0512-D1	3DSM	49-65	107-109	140-144	208-214	267-276	302-304
T0513-D1	3DUP	195-201	258-260			

Table B.4: The environments specific substitution matrices in the six dihedral angle regions (Figure 3.2). There are 21 amino acid types including the cystein-cystein couple (J). The corresponding dihedral angle regions are in italic bold on the top left of each substitution score matrix.

A	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	4	-6	-2	-1	-3	2	-3	-2	-1	-2	-1	-2	-1	-1	-2	1	0	0	-4	-3	1
C	-6	21	-9	-8	-10	-8	-9	-8	-10	-6	-10	-4	-11	-6	-9	-5	-7	-9	-48	-5	0
D	-2	-9	7	2	-7	-1	-2	-7	0	-7	-7	1	-1	1	-1	0	-1	-6	-6	-4	-6
E	-1	-8	2	6	-5	-2	-1	-5	1	-5	-5	0	-1	2	0	-1	-1	-4	-5	-3	-6
F	-3	-10	-7	-5	9	-5	-2	0	-5	2	1	-5	-7	-4	-5	-4	-3	0	3	4	-2
G	2	-8	-1	-2	-5	9	-3	-5	-2	-5	-4	-1	-2	-1	-2	1	-1	-3	-4	-4	0
H	-3	-9	-2	-1	-2	-3	11	-4	-1	-4	-3	1	-3	1	0	-2	-2	-3	-2	2	-4
I	-2	-8	-7	-5	0	-5	-4	7	-4	3	3	-5	-4	-4	-4	-4	-1	4	-2	-1	0
K	-1	-10	0	1	-5	-2	-1	-4	5	-4	-3	0	-1	2	3	0	-1	-3	-6	-3	-6
L	-2	-6	-7	-5	2	-5	-4	3	-4	6	3	-4	-5	-3	-3	-4	-3	1	-1	-1	0

M	-1	-10	-7	-5	1	-4	-3	3	-3	3	8	-4	-5	-1	-3	-3	-1	1	-1	0	-1
N	-2	-4	1	0	-5	-1	1	-5	0	-4	-4	7	-2	1	0	1	0	-3	-5	-2	-4
P	-1	-11	-1	-1	-7	-2	-3	-4	-1	-5	-5	-2	11	-1	-2	0	-2	-3	-6	-4	-5
Q	-1	-6	1	2	-4	-1	1	-4	2	-3	-1	1	-1	6	1	0	-1	-3	-4	-2	-5
R	-2	-9	-1	0	-5	-2	0	-4	3	-3	-3	0	-2	1	7	-1	-1	-4	-3	-2	-4
S	1	-5	0	-1	-4	1	-2	-4	0	-4	-3	1	0	0	-1	5	2	-2	-5	-3	-1
T	0	-7	-1	-1	-3	-1	-2	-1	-1	-3	-1	0	-2	-1	-1	2	6	0	-5	-3	0
V	0	-9	-6	-4	0	-3	-3	4	-3	1	1	-3	-3	-3	-4	-2	0	6	-3	-2	2
W	-4	-48	-6	-5	3	-4	-2	-2	-6	-1	-1	-5	-6	-4	-3	-5	-5	-3	15	3	-6
Y	-3	-5	-4	-3	4	-4	2	-1	-3	-1	0	-2	-4	-2	-2	-3	-3	-2	3	10	-2
J	1	0	-6	-6	-2	0	-4	0	-6	0	-1	-4	-5	-5	-4	-1	0	2	-6	-2	15

B	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	7	0	-3	-1	-4	0	-2	-2	-1	-3	-3	-3	-1	-1	-2	0	-1	0	-3	-4	2
C	0	19	-6	-4	-8	-16	-45	-10	-10	-10	-18	-10	-8	-11	-12	-8	-6	-15	-2	-9	-3
D	-3	-6	8	2	-7	-3	-1	-7	-1	-7	-5	3	-3	-1	-2	1	-2	-6	-5	-4	-3
E	-1	-4	2	8	-5	-4	-1	-4	2	-5	-2	0	-2	3	0	-1	-1	-4	-5	-5	-5
F	-4	-8	-7	-5	10	-6	0	1	-5	1	1	-4	-5	-3	-6	-5	-4	-1	4	6	-3
G	0	-16	-3	-4	-6	11	-3	-8	-4	-7	-6	-4	-4	-3	-6	0	-3	-7	-5	-4	-2
H	-2	-45	-1	-1	0	-3	13	-6	-1	-4	-1	1	-5	1	1	-2	-2	-5	-3	1	1
I	-2	-10	-7	-4	1	-8	-6	7	-3	3	1	-5	-4	-5	-4	-6	-2	4	-4	-2	-1
K	-1	-10	-1	2	-5	-4	-1	-3	7	-4	-1	0	-1	2	4	-2	0	-4	-5	-3	-3
L	-3	-10	-7	-5	1	-7	-4	3	-4	7	3	-5	-5	-3	-4	-5	-4	1	-2	-2	-3
M	-3	-18	-5	-2	1	-6	-1	1	-1	3	10	-3	-5	3	-2	-4	-2	1	-1	-1	-3
N	-3	-10	3	0	-4	-4	1	-5	0	-5	-3	9	-4	0	-1	2	1	-4	-6	-3	-4
P	-1	-8	-3	-2	-5	-4	-5	-4	-1	-5	-5	-4	7	-2	-3	-2	-3	-4	-9	-5	-4
Q	-1	-11	-1	3	-3	-3	1	-5	2	-3	3	0	-2	9	2	-1	-1	-4	-3	-4	-1
R	-2	-12	-2	0	-6	-6	1	-4	4	-4	-2	-1	-3	2	10	-2	-2	-4	-4	-3	-6

S	0	-8	1	-1	-5	0	-2	-6	-2	-5	-4	2	-2	-1	-2	7	3	-4	-5	-4	0
T	-1	-6	-2	-1	-4	-3	-2	-2	0	-4	-2	1	-3	-1	-2	3	7	-1	-7	-3	-1
V	0	-15	-6	-4	-1	-7	-5	4	-4	1	1	-4	-4	-4	-4	-4	-1	6	-4	-3	1
W	-3	-2	-5	-5	4	-5	-3	-4	-5	-2	-1	-6	-9	-3	-4	-5	-7	-4	15	4	-7
Y	-4	-9	-4	-5	6	-4	1	-2	-3	-2	-1	-3	-5	-4	-3	-4	-3	-3	4	11	-4
J	2	-3	-3	-5	-3	-2	1	-1	-3	-3	-3	-4	-4	-1	-6	0	-1	1	-7	-4	16

C	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	6	-12	-4	-3	-4	2	-3	-2	-2	-2	0	-3	-25	-2	-3	0	-1	0	-6	-4	3
C	-12	17	-13	-5	-4	-4	-5	-11	-8	-15	-2	-9	-19	-3	-6	-9	-9	-9	-7	-5	-5
D	-4	-13	12	3	-5	-3	-1	-8	-1	-7	-6	4	-20	-2	-2	1	-1	-8	-5	-3	-10
E	-3	-5	3	9	-6	-3	-1	-5	2	-5	-2	0	-22	4	1	0	0	-3	-5	-3	-6
F	-4	-4	-5	-6	8	-6	-1	0	-5	0	-1	-4	-24	-4	-6	-5	-4	-2	2	3	-2
G	2	-4	-3	-3	-6	10	-5	-7	-3	-5	-4	-2	-23	-4	-5	-1	-2	-4	-5	-6	0
H	-3	-5	-1	-1	-1	-5	11	-5	1	-5	-4	0	7	2	1	0	0	-5	-4	0	-2
I	-2	-11	-8	-5	0	-7	-5	5	-5	2	1	-5	-26	-5	-5	-6	-3	2	-3	-3	-1
K	-2	-8	-1	2	-5	-3	1	-5	8	-5	-2	1	4	3	5	0	1	-3	-7	-3	-4
L	-2	-15	-7	-5	0	-5	-5	2	-5	6	3	-5	-25	-4	-3	-6	-3	1	-3	-3	0
M	0	-2	-6	-2	-1	-4	-4	1	-2	3	9	-1	-20	1	-4	-3	-2	0	-5	-2	1
N	-3	-9	4	0	-4	-2	0	-5	1	-5	-1	10	-21	1	1	2	1	-5	-4	-3	-4
P	-25	-19	-20	-22	-24	-23	7	-26	4	-25	-20	-21	11	-21	6	4	7	-28	-20	-24	-17
Q	-2	-3	-2	4	-4	-4	2	-5	3	-4	1	1	-21	9	3	0	0	-4	-6	-3	-7
R	-3	-6	-2	1	-6	-5	1	-5	5	-3	-4	1	6	3	9	-1	0	-4	-7	-2	-2
S	0	-9	1	0	-5	-1	0	-6	0	-6	-3	2	4	0	-1	7	2	-5	-7	-3	-2
T	-1	-9	-1	0	-4	-2	0	-3	1	-3	-2	1	7	0	0	2	6	-2	-7	-3	-1
V	0	-9	-8	-3	-2	-4	-5	2	-3	1	0	-5	-28	-4	-4	-5	-2	4	-6	-3	0
W	-6	-7	-5	-5	2	-5	-4	-3	-7	-3	-5	-4	-20	-6	-7	-7	-7	-6	14	1	-4
Y	-4	-5	-3	-3	3	-6	0	-3	-3	-3	-2	-3	-24	-3	-2	-3	-3	-3	1	8	-3

J	3	-5	-10	-6	-2	0	-2	-1	-4	0	1	-4	-17	-7	-2	-2	-1	0	-4	-3	13

D	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	8	-3	-5	-2	-5	0	-4	-5	-4	-2	0	-2	-4	-6	-2	3	2	0	-7	-2	2
C	-3	19	-43	-3	-3	-36	-8	-5	-3	-10	-34	-42	-35	0	-1	-1	-34	-11	1	-39	-1
D	-5	-43	6	0	-12	-10	-4	-8	-5	-9	-4	0	-2	-7	-4	-3	-5	-10	-15	-5	-6
E	-2	-3	0	10	-8	-7	-11	-8	1	-5	2	-1	-4	3	-4	-2	-3	-5	-8	-3	1
F	-5	-3	-12	-8	9	-2	-3	1	-46	1	0	-7	-4	-2	-3	-6	-4	-2	-1	3	-3
G	0	-36	-10	-7	-2	13	-8	-41	-9	-8	-9	-7	2	1	-8	-1	3	-43	0	-1	-37
H	-4	-8	-4	-11	-3	-8	9	-9	-3	-7	-43	0	-6	-1	1	-2	-2	-11	-4	2	-40
I	-5	-5	-8	-8	1	-41	-9	10	-6	5	3	-9	-6	-11	-6	-8	-3	7	-4	-10	1
K	-4	-3	-5	1	-46	-9	-3	-6	10	-9	-7	-1	5	2	3	0	2	-5	-10	-1	-4
L	-2	-10	-9	-5	1	-8	-7	5	-9	8	3	-5	-7	-4	-5	-4	-6	4	-7	-1	0
M	0	-34	-4	2	0	-9	-43	3	-7	3	12	-10	-2	0	-7	-4	-4	-5	0	-8	6
N	-2	-42	0	-1	-7	-7	0	-9	-1	-5	-10	6	-3	-3	-1	2	-2	-6	-14	-3	2
P	-4	-35	-2	-4	-4	2	-6	-6	5	-7	-2	-3	13	0	-5	-6	-2	-8	-41	-6	-36
Q	-6	0	-7	3	-2	1	-1	-11	2	-4	0	-3	0	8	4	2	-1	-2	5	-5	-4
R	-2	-1	-4	-4	-3	-8	1	-6	3	-5	-7	-1	-5	4	9	-2	-1	-1	2	-5	-39
S	3	-1	-3	-2	-6	-1	-2	-8	0	-4	-4	2	-6	2	-2	8	1	-10	-3	-6	2
T	2	-34	-5	-3	-4	3	-2	-3	2	-6	-4	-2	-2	-1	-1	1	12	1	-4	1	-36
V	0	-11	-10	-5	-2	-43	-11	7	-5	4	-5	-6	-8	-2	-1	-10	1	9	-3	-3	3
W	-7	1	-15	-8	-1	0	-4	-4	-10	-7	0	-14	-41	5	2	-3	-4	-3	12	-8	-4
Y	-2	-39	-5	-3	3	-1	2	-10	-1	-1	-8	-3	-6	-5	-5	-6	1	-3	-8	9	-40
J	2	-1	-6	1	-3	-37	-40	1	-4	0	6	2	-36	-4	-39	2	-36	3	-4	-40	13

E	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	11	5	-1	-3	1	-2	-3	-1	-1	-2	-8	-3	-18	1	-3	3	5	-2	-1	-3	6
C	5	21	-5	-36	-33	-6	-37	-25	-39	-2	-30	-8	-11	-36	-38	-37	-28	-27	-28	-35	11

D	-1	-5	8	2	-4	-3	-3	-36	0	-6	-1	2	-21	-1	-5	0	2	-37	-1	-1	-3
E	-3	-36	2	11	-5	-2	-4	-31	0	-5	-36	0	-17	5	0	-2	6	-32	-9	-3	-1
F	1	-33	-4	-5	15	-5	2	-28	-2	6	1	0	-14	0	1	-3	0	-30	11	8	-2
G	-2	-6	-3	-2	-5	1	-3	-9	-3	-6	-5	-3	-32	-4	-3	-2	-3	-5	-4	-7	-5
H	-3	-37	-3	-4	2	-3	10	-32	2	2	0	3	-17	3	3	-3	2	-33	-35	5	-2
I	-1	-25	-36	-31	-28	-9	-32	18	4	13	-25	1	-6	3	-33	2	9	12	10	8	-25
K	-1	-39	0	0	-2	-3	2	4	7	1	3	2	-19	4	4	-1	-36	5	-2	-3	1
L	-2	-2	-6	-5	6	-6	2	13	1	14	9	-2	-16	0	1	-4	1	10	-34	6	-1
M	-8	-30	-1	-36	1	-5	0	-25	3	9	17	0	-11	4	1	3	6	-27	-28	1	9
N	-3	-8	2	0	0	-3	3	1	2	-2	0	5	-25	-1	0	1	2	-6	-8	1	-1
P	-18	-11	-21	-17	-14	-32	-17	-6	-19	-16	-11	-25	8	17	-18	-18	-9	-7	-9	-15	-10
Q	1	-36	-1	5	0	-4	3	3	4	0	4	-1	17	12	2	1	-34	5	0	-6	7
R	-3	-38	-5	0	1	-3	3	-33	4	1	1	0	-18	2	11	-4	-35	5	0	1	2
S	3	-37	0	-2	-3	-2	-3	2	-1	-4	3	1	-18	1	-4	10	5	-34	2	-10	2
T	5	-28	2	6	0	-3	2	9	-36	1	6	2	-9	-34	-35	5	14	-25	-26	-32	7
V	-2	-27	-37	-32	-30	-5	-33	12	5	10	-27	-6	-7	5	5	-34	-25	20	14	5	-26
W	-1	-28	-1	-9	11	-4	-35	10	-2	-34	-28	-8	-9	0	0	2	-26	14	20	7	-28
Y	-3	-35	-1	-3	8	-7	5	8	-3	6	1	1	-15	-6	1	-10	-32	5	7	14	2
J	6	11	-3	-1	-2	-5	-2	-25	1	-1	9	-1	-10	7	2	2	7	-26	-28	2	16

F	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	12	14	2	-29	-23	-2	1	-24	-6	-25	13	-11	7	6	8	-35	10	0	-19	4	14
C	14	-13	-24	-19	-13	-8	-16	-15	-23	-16	19	-23	16	-19	-18	-25	-19	-13	-9	-17	21
D	2	-24	13	3	6	-2	-27	-1	-34	-27	-26	8	5	3	-29	1	-5	3	-21	-28	-29
E	-29	-19	3	17	-19	-4	12	-21	10	-22	16	4	10	-25	-24	6	8	-19	-16	-23	-24
F	-23	-13	6	-19	-13	-2	-16	19	-23	-16	-15	-23	16	-19	-18	9	-20	-13	-10	17	-18
G	-2	-8	-2	-4	-2	0	-10	-10	-5	-2	-43	-3	-13	-5	-3	-4	-5	-5	-4	-5	-47
H	1	-16	-27	12	-16	-10	24	-17	-26	-19	-17	-26	13	-22	-21	9	-22	18	-12	-20	-21

I	-24	-15	-1	-21	19	-10	-17	21	-24	19	-16	-8	15	-21	-20	12	-21	20	-11	-19	-20
K	-6	-23	-34	10	-23	-5	-26	-24	16	-26	10	-33	6	13	11	-4	5	-22	-19	12	-28
L	-25	-16	-27	-22	-16	-2	-19	19	-26	21	-17	8	13	-22	-21	-28	-22	-16	-12	-20	-21
M	13	19	-26	16	-15	-43	-17	-16	10	-17	-16	-25	15	-21	-20	-27	-21	20	-11	-19	19
N	-11	-23	8	4	-23	-3	-26	-8	-33	8	-25	15	6	-5	-28	-1	8	8	-19	-27	-28
P	7	16	5	10	16	-13	13	15	6	13	15	6	45	10	11	4	10	16	20	12	11
Q	6	-19	3	-25	-19	-5	-22	-21	13	-22	-21	-5	10	17	14	-2	-6	-19	-16	14	-24
R	8	-18	-29	-24	-18	-3	-21	-20	11	-21	-20	-28	11	14	14	-13	9	11	-14	12	-23
S	-35	-25	1	6	9	-4	9	12	-4	-28	-27	-1	4	-2	-13	14	4	-25	-21	5	-30
T	10	-19	-5	8	-20	-5	-22	-21	5	-22	-21	8	10	-6	9	4	19	-2	-16	-23	-24
V	0	-13	3	-19	-13	-5	18	20	-22	-16	20	8	16	-19	11	-25	-2	20	-9	-17	-18
W	-19	-9	-21	-16	-10	-4	-12	-11	-19	-12	-11	-19	20	-16	-14	-21	-16	-9	31	-13	-14
Y	4	-17	-28	-23	17	-5	-20	-19	12	-20	-19	-27	12	14	12	5	-23	-17	-13	19	-22
J	14	21	-29	-24	-18	-47	-21	-20	-28	-21	19	-28	11	-24	-23	-30	-24	-18	-14	-22	21

Appendix C

Antibody Complementarity Determining Region Loop Structure Prediction

Table C.1: The detailed results of CDR prediction on the Native set using DB-I.

CDR		Mean	STD	Median	Coverage
L1	FREAD	0.81	0.75	0.60	100.0%
	FREAD-S	1.25	1.05	0.97	100.0%
	ConFREAD	0.86	0.80	0.69	91.8%
	Best	0.39	0.39	0.69	100.0%
L2	FREAD	0.42	0.34	0.35	100.0%
	FREAD-S	0.59	0.38	0.56	100.0%
	ConFREAD	0.56	0.36	0.47	99.0%
	Best	0.22	0.22	0.47	100.0%
L3	FREAD	0.96	0.88	0.75	100.0%
	FREAD-S	0.80	0.62	0.64	100.0%
	ConFREAD	0.69	0.52	0.53	92.8%
	Best	0.37	0.33	0.53	100.0%
H1	FREAD	0.99	0.88	0.77	100.0%
	FREAD-S	1.13	0.96	0.80	100.0%
	ConFREAD	0.84	0.67	0.65	87.6%
	Best	0.56	0.74	0.65	100.0%
H2	FREAD	0.88	0.88	0.61	100.0%
	FREAD-S	0.96	0.87	0.67	100.0%
	ConFREAD	0.94	0.91	0.62	88.7%
	Best	0.46	0.56	0.62	100.0%

H3	FREAD	2.26	2.09	1.99	100.0%
	FREAD-S	1.38	1.56	0.71	100.0%
	ConFREAD	1.24	1.14	0.69	67.0%
	Best	0.91	0.97	0.69	100.0%

Table C.2: The detailed results of CDR prediction on the Native set using DB-E.

CDR		Mean	STD	Median	Coverage
L1	FREAD	1.61	1.99	0.87	100.0%
	FREAD-S	1.59	1.98	0.86	100.0%
	ConFREAD	0.94	0.77	0.78	34.0%
	Best	1.32	1.99	0.78	100.0%
L2	FREAD	0.50	0.45	0.38	100.0%
	FREAD-S	0.48	0.33	0.39	100.0%
	ConFREAD	0.51	0.31	0.42	92.8%
	Best	0.36	0.33	0.42	100.0%
L3	FREAD	1.49	1.43	1.11	100.0%
	FREAD-S	1.34	1.38	0.95	100.0%
	ConFREAD	0.84	0.41	0.76	58.8%
	Best	0.69	0.51	0.76	100.0%
H1	FREAD	1.49	1.35	1.01	100.0%
	FREAD-S	1.55	1.49	1.02	100.0%
	ConFREAD	0.96	0.58	0.77	59.8%
	Best	1.03	1.26	0.77	100.0%
H2	FREAD	1.60	1.11	1.34	100.0%
	FREAD-S	2.04	1.47	1.62	100.0%
	ConFREAD	1.33	0.68	1.29	42.3%
	Best	1.15	0.86	1.29	100.0%
H3	FREAD	3.81	1.99	3.46	88.7%
	FREAD-S	3.48	2.21	3.13	88.7%
	ConFREAD	0.97	0.66	0.61	6.2%
	Best	2.97	1.90	0.61	88.7%

Table C.3: The detailed results of CDR prediction on the RA-Native set.

CDR		Mean	STD	Median	Coverage
L1	FREAD	1.09	0.73	0.90	100.0%
	FREAD-S	0.94	0.46	0.86	100.0%
	ConFREAD	0.90	0.51	0.73	90.7%
	RosettaAntibody	0.83	0.36	0.78	100.0%
L2	FREAD	0.59	0.31	0.51	100.0%
	FREAD-S	0.75	0.42	0.68	100.0%
	ConFREAD	0.72	0.38	0.68	100.0%
	RosettaAntibody	0.58	0.25	0.54	100.0%
L3	FREAD	1.00	0.58	0.93	100.0%
	FREAD-S	1.01	0.80	0.80	100.0%
	ConFREAD	0.80	0.67	0.61	94.4%
	RosettaAntibody	0.91	0.53	0.83	100.0%

H1	FREAD	0.88	0.41	0.82	100.0%
	FREAD-S	1.20	0.72	1.13	100.0%
	ConFREAD	0.84	0.59	0.69	96.3%
	RosettaAntibody	0.83	0.41	0.84	100.0%
H2	FREAD	0.90	0.62	0.73	100.0%
	FREAD-S	1.23	0.93	1.01	100.0%
	ConFREAD	0.83	0.72	0.69	81.5%
	RosettaAntibody	1.28	1.49	0.93	100.0%
H3	FREAD	1.85	1.62	1.78	100.0%
	FREAD-S	1.53	1.54	0.91	100.0%
	ConFREAD	1.38	1.24	0.84	68.5%
	RosettaAntibody	2.15	1.33	1.80	100.0%

Table C.4: The detailed results of CDR prediction on the RA-Model set.

CDR		Mean	STD	Median	Coverage
L1	FREAD	0.75	0.51	0.70	96.3%
	FREAD-S	0.95	0.57	0.86	96.3%
	ConFREAD	0.99	0.58	0.83	94.4%
	RosettaAntibody	0.82	0.35	0.76	100.0%
L2	FREAD	0.68	0.42	0.61	96.3%
	FREAD-S	0.86	0.43	0.85	96.3%
	ConFREAD	0.65	0.24	0.61	96.3%
	RosettaAntibody	0.59	0.25	0.54	100.0%
L3	FREAD	0.99	0.90	0.79	98.1%
	FREAD-S	1.13	0.94	0.83	98.1%
	ConFREAD	0.98	0.60	0.84	96.3%
	RosettaAntibody	0.90	0.53	0.83	100.0%
H1	FREAD	1.17	1.01	0.82	100.0%
	FREAD-S	1.49	0.96	1.16	100.0%
	ConFREAD	1.08	0.60	0.86	94.4%
	RosettaAntibody	0.82	0.41	0.82	100.0%
H2	FREAD	1.61	1.70	0.95	100.0%
	FREAD-S	1.86	1.75	1.41	100.0%
	ConFREAD	1.23	0.73	1.23	75.9%
	RosettaAntibody	1.03	0.55	0.91	100.0%
H3	FREAD	3.12	1.88	3.12	98.1%
	FREAD-S	2.07	1.40	1.68	98.1%
	ConFREAD	1.98	1.44	1.31	55.6%
	RosettaAntibody	2.96	2.08	2.65	100.0%

Table C.5: The prediction results of the Bound-Free set for non CDR-H3.

CDR-L1						
Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	2.35	2.51	2.30	2.30	2.13
1D5I	1D6V	0.27	0.31	2.79	0.98	0.27
2A6J	2A6I	1.20	1.80	1.34	1.00	0.71
1Q9K	1Q9Q	1.70	1.40	1.76	2.43	1.28

1KCV	1KCS	0.34	0.52	1.10	0.95	0.34
1CR9	1CU4	0.56	1.65	2.10	1.86	0.39
1GGC	1GGI	1.43	1.83	2.15	2.03	1.41
1CGS	2CGR	5.25	5.94	5.94	-	5.64
1NBV	1CBV	1.25	3.01	3.48	2.16	1.67
1HIL	1IFH	1.59	1.27	1.69	1.69	1.23
1OAQ	1OAU	0.35	0.95	0.72	-	0.31
1MNU	1MPA	2.43	2.55	2.95	2.62	2.04
Mean		1.56	1.98	2.36	1.80	1.45
STD		1.38	1.49	1.38	0.63	1.48

CDR-L2

Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	1.28	1.39	1.27	0.96	0.77
1D5I	1D6V	0.39	0.44	0.40	0.40	0.32
2A6J	2A6I	1.05	1.37	1.14	1.14	0.76
1Q9K	1Q9Q	0.77	0.79	1.14	1.14	0.49
1KCV	1KCS	0.26	0.29	1.03	0.29	0.25
1CR9	1CU4	0.43	0.45	0.46	0.44	0.32
1GGC	1GGI	1.48	1.55	1.44	1.55	1.39
1CGS	2CGR	2.53	3.11	3.20	3.20	2.52
1NBV	1CBV	0.73	1.23	0.73	0.73	0.25
1HIL	1IFH	1.34	1.32	1.34	1.34	1.09
1OAQ	1OAU	0.14	0.25	0.36	0.36	0.10
1MNU	1MPA	1.28	1.40	1.56	1.56	0.74
Mean		0.97	1.13	1.17	1.09	0.75
STD		0.67	0.79	0.76	0.81	0.68

CDR-L3

Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	2.30	2.38	2.63	2.01	1.50
1D5I	1D6V	0.37	1.07	0.77	2.21	0.26
2A6J	2A6I	0.95	1.08	1.28	1.36	0.54
1Q9K	1Q9Q	0.46	0.41	0.41	0.43	0.29
1KCV	1KCS	0.93	0.90	0.55	0.71	0.44
1CR9	1CU4	0.64	1.24	1.23	1.14	0.38
1GGC	1GGI	1.28	1.89	1.12	1.10	0.55
1CGS	2CGR	3.92	3.82	4.17	4.29	3.03
1NBV	1CBV	1.02	1.94	2.05	1.08	0.80
1HIL	1IFH	0.57	1.95	0.62	0.85	0.52
1OAQ	1OAU	0.86	1.30	1.01	0.57	0.24
1MNU	1MPA	1.29	1.57	1.84	1.94	1.11
Mean		1.22	1.63	1.47	1.47	0.81
STD		0.99	0.88	1.08	1.05	0.79

CDR-H1

Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	1.61	2.47	1.87	1.42	0.71
1D5I	1D6V	0.39	0.41	0.92	0.37	0.36
2A6J	2A6I	0.68	1.44	1.77	0.87	0.57
1Q9K	1Q9Q	0.81	0.71	1.56	0.49	0.42
1KCV	1KCS	0.89	1.34	0.83	-	0.68
1CR9	1CU4	1.55	2.32	2.57	-	2.32
1GGC	1GGI	2.03	2.97	1.97	1.97	1.41

1CGS	2CGR	5.63	4.55	4.55	-	4.55
1NBV	1CBV	2.34	2.27	2.38	1.85	1.85
1HIL	1IFH	3.27	3.27	3.27	3.27	2.85
1OAQ	1OAU	0.18	0.84	0.89	1.07	0.15
1MNU	1MPA	0.84	0.82	1.92	1.33	0.68
Mean		1.69	1.95	2.04	1.40	1.38
STD		1.53	1.25	1.07	0.89	1.31

CDR-H2						
Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	1.77	1.82	2.46	2.46	1.30
1D5I	1D6V	1.99	2.24	2.76	2.76	1.37
2A6J	2A6I	2.17	7.97	1.22	0.39	0.39
1Q9K	1Q9Q	0.42	0.41	0.74	0.74	0.33
1KCV	1KCS	0.83	1.02	1.18	1.18	0.68
1CR9	1CU4	1.09	0.84	1.81	0.90	0.69
1GGC	1GGI	1.18	1.40	1.40	-	0.77
1CGS	2CGR	4.30	5.08	4.25	5.18	3.48
1NBV	1CBV	1.65	1.12	1.51	2.21	1.12
1HIL	1IFH	2.27	2.62	1.87	-	1.67
1OAQ	1OAU	0.23	1.64	0.28	0.30	0.15
1MNU	1MPA	0.93	3.88	1.24	1.24	0.65
Mean		1.57	2.50	1.73	1.73	1.05
STD		1.09	2.18	1.04	1.48	0.89

CDR-H2						
Free	Bound	Free-Bound	FREAD	FREAD-S	ConFREAD	Best
1NGZ	1N7M	2.65	1.20	5.63	1.50	0.71
1D5I	1D6V	1.95	5.22	1.63	0.73	0.73
2A6J	2A6I	1.05	4.28	2.18	-	1.39
1Q9K	1Q9Q	4.09	0.60	4.37	1.21	0.60
1KCV	1KCS	1.99	1.66	0.63	0.63	0.63
1CR9	1CU4	2.50	2.50	2.95	-	1.35
1GGC	1GGI	2.24	2.88	0.60	-	0.60
1CGS	2CGR	2.79	5.59	3.17	-	2.19
1NBV	1CBV	1.76	3.28	2.64	-	2.25
1HIL	1IFH	3.12	3.08	3.08	3.08	2.50
1OAQ	1OAU	1.59	3.93	5.65	0.92	0.36
1MNU	1MPA	2.43	8.66	2.39	-	2.39
Mean		2.35	3.57	2.91	1.35	1.31
STD		0.79	2.20	1.66	0.91	0.82

Appendix D

Loop Stretch Test Set

Table D.1: The first test set for loop stretch. All test loops have 8 residues in length. There are 40 test loops in every 0.1 loop stretch (λ) bin (N: N-anchor, C: C-anchor, α: Helix, β: Strand and λ: Loop stretch).

Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1R8S	A	138	145	β	α	0.2059	1W78	A	139	146	β	α	0.3009
2QF4	A	133	140	β	β	0.2162	1QFM	A	631	638	β	α	0.3019
1RI6	A	203	210	β	β	0.2177	1M2D	A	84	91	α	α	0.3063
2HSJ	A	190	197	α	α	0.2180	1PJX	A	222	229	β	β	0.3100
2FSR	A	88	95	β	β	0.2298	3GRD	A	40	47	β	β	0.3109
2ZUX	A	453	460	β	β	0.2301	3BON	A	150	157	β	β	0.3115
1YOC	A	50	57	α	α	0.2334	1JI1	A	247	254	α	β	0.3131
3CQL	A	157	164	α	α	0.2336	1S9U	A	22	29	α	α	0.3182
1FN9	A	259	266	β	β	0.2364	3BBB	A	106	113	α	β	0.3244
2I5V	O	126	133	β	β	0.2370	1C5E	A	32	39	β	β	0.3273
2HC1	A	99	106	β	β	0.2420	1OR0	B	115	122	β	β	0.3313
2GMN	A	19	26	β	β	0.2455	2W5N	A	302	309	β	β	0.3321
2WHJ	A	158	165	β	α	0.2459	1JND	A	321	328	β	β	0.3375
2G2C	A	117	124	β	β	0.2472	2OLM	A	53	60	α	β	0.3406
2EHZ	A	188	195	β	β	0.2475	3DAQ	A	247	254	α	α	0.3429
3ENU	A	73	80	β	β	0.2476	1OK0	A	65	72	β	β	0.3464
1VRM	A	39	46	α	α	0.2480	2I5V	O	183	190	β	β	0.3471
2Q0I	A	160	167	β	β	0.2484	2O4U	X	246	253	β	β	0.3479
2QED	A	116	123	β	β	0.2486	1UWF	A	43	50	β	β	0.3523
2VSM	A	350	357	β	β	0.2512	2PVB	A	15	22	α	α	0.3563
1KNM	A	59	66	β	β	0.2513	2Z0J	A	138	145	β	α	0.3588
2E7Z	A	3	10	β	β	0.2586	1QW9	A	344	351	β	β	0.3602
1LZL	A	199	206	α	α	0.2594	2J2J	A	139	146	β	β	0.3643
1UPS	A	328	335	β	β	0.2599	1JI1	A	465	472	β	α	0.3646
3H0N	A	152	159	β	β	0.2600	2PRV	A	61	68	β	α	0.3647
3FJZ	A	140	147	β	β	0.2618	1RU4	A	272	279	β	β	0.3704
2HO3	A	228	235	β	β	0.2627	3BIO	A	107	114	β	α	0.3752

2QZU	A	371	378	β	β	0.2631	1WDP	A	240	247	α	α	0.3754
1Q35	A	216	223	β	β	0.2670	2CKK	A	70	77	β	β	0.3786
2FNO	A	10	17	β	α	0.2673	3E9K	A	140	147	α	β	0.3816
1E9G	A	58	65	β	β	0.2704	1ZJA	A	117	124	α	α	0.3829
2ZKM	X	548	555	α	β	0.2742	1GKM	A	135	142	β	α	0.3837
1G0S	A	129	136	β	β	0.2758	2H98	A	105	112	β	α	0.3844
2CL2	A	120	127	β	β	0.2787	2OBL	A	221	228	β	α	0.3851
1OEW	A	183	190	β	β	0.2810	2CXN	A	409	416	β	α	0.3895
1J0P	A	69	76	α	α	0.2853	2P02	A	132	139	β	α	0.3916
1VHE	A	291	298	α	β	0.2876	2ZKM	X	351	358	α	β	0.3942
2OSX	A	61	68	α	α	0.2885	3C1Q	A	63	70	α	α	0.3947
1PJX	A	255	262	β	β	0.2914	2CN3	A	314	321	β	β	0.3950
1Y43	B	9	16	β	β	0.3000	1QL0	A	170	177	β	β	0.3986
1ITX	A	144	151	β	α	0.4059	2VZP	A	59	66	β	β	0.5020
1ODM	A	170	177	α	β	0.4060	1L7A	A	291	298	β	α	0.5021
2VLQ	B	89	96	α	β	0.4093	2IW1	A	273	280	β	α	0.5027
1VHE	A	11	18	α	α	0.4110	1KWG	A	281	288	α	α	0.5036
2Z3H	A	16	23	α	β	0.4128	1NKG	A	243	250	α	α	0.5080
1WUI	S	232	239	α	α	0.4143	2EIY	B	206	213	β	α	0.5152
2BJK	A	317	324	α	β	0.4202	2CWS	A	209	216	β	β	0.5214
3SIL	A	241	248	β	β	0.4235	1P3D	A	69	76	β	α	0.5214
2J2J	A	4	11	β	β	0.4254	2FI1	A	6	13	β	α	0.5214
3B8D	A	149	156	β	α	0.4280	2I0K	A	195	202	α	β	0.5219
2GKE	A	161	168	β	α	0.4281	2IW1	A	83	90	β	β	0.5222
2C71	A	166	173	β	α	0.4410	3FVS	A	56	63	α	α	0.5255
1GUD	A	63	70	β	α	0.4434	1MDL	A	86	93	α	α	0.5265
1JND	A	11	18	α	α	0.4444	2VK2	A	61	68	β	α	0.5306
1GKM	A	105	112	α	α	0.4458	3BMV	A	474	481	β	β	0.5331
1V7W	A	352	359	α	α	0.4483	2R9F	A	75	82	α	α	0.5357
1UCS	A	23	30	β	α	0.4505	2I7D	A	5	12	β	α	0.5372
3BFV	A	181	188	β	α	0.4513	3E9K	A	422	429	β	α	0.5386
2CNQ	A	247	254	β	β	0.4527	3CUZ	A	245	252	α	β	0.5399
1CHD	A	126	133	β	α	0.4533	3D32	A	73	80	α	β	0.5440
1DS1	A	160	167	β	β	0.4571	2QV8	A	119	126	β	β	0.5459
1RU4	A	186	193	β	β	0.4606	2OXN	A	4	11	β	α	0.5514
1Q6Z	A	383	390	α	β	0.4606	1UFY	A	47	54	β	α	0.5524
1FJ2	A	201	208	β	α	0.4633	3H6J	A	350	357	β	β	0.5530
2AQ5	A	238	245	β	β	0.4666	2AQ5	A	157	164	β	β	0.5560
2IW1	A	199	206	β	α	0.4691	1CSH	A	230	237	α	α	0.5616
1XRU	A	230	237	β	β	0.4696	1HT6	A	359	366	α	β	0.5653
3DAQ	A	105	112	β	α	0.4714	1H97	A	94	101	α	α	0.5693
1UAS	A	128	135	β	α	0.4759	2VFR	A	406	413	α	α	0.5726
1TJY	A	64	71	β	α	0.4793	2BWR	A	328	335	α	β	0.5782
3FDY	A	392	399	β	β	0.4814	1VLA	A	29	36	α	α	0.5792
2VHK	A	114	121	β	β	0.4840	2V8T	A	235	242	β	α	0.5832
2OXG	A	66	73	β	β	0.4853	1B8O	A	190	197	β	α	0.5834
1SU8	A	191	198	α	α	0.4858	1Y4W	A	482	489	β	β	0.5850
2D81	A	29	36	α	β	0.4873	1WKR	A	144	151	β	β	0.5886
3CCD	A	6	13	β	α	0.4880	2UYT	A	127	134	α	α	0.5896
1M55	A	151	158	α	β	0.4937	1T3Y	A	98	105	α	β	0.5900
1ODM	A	159	166	α	α	0.4941	1GWM	A	103	110	β	β	0.5902
1GNL	A	70	77	α	α	0.4948	2E6F	A	84	91	α	β	0.5902

1U4G	A	102	109	β	β	0.4984	3G7R	A	99	106	α	α	0.5949
1YAC	A	133	140	β	α	0.6038	1Y93	A	11	18	β	α	0.7003
1GA6	A	300	307	α	α	0.6039	2WAA	A	276	283	β	α	0.7009
3E2D	A	225	232	β	α	0.6077	1E58	A	202	209	α	β	0.7014
1HDH	A	8	15	β	α	0.6121	16PK	A	332	339	β	α	0.7035
3F1L	A	153	160	α	α	0.6149	1IOM	A	79	86	α	α	0.7037
1RQB	A	197	204	β	α	0.6160	2BHU	A	540	547	α	α	0.7047
1Q0R	A	268	275	β	α	0.6161	2VFR	A	66	73	β	β	0.7063
3G28	A	4	11	α	α	0.6171	3GMV	X	82	89	α	α	0.7066
1PX0	A	202	209	α	α	0.6174	1SU8	A	324	331	β	α	0.7083
1NKG	A	95	102	β	β	0.6191	1Q0R	A	122	129	β	α	0.7130
2HZL	A	203	210	β	β	0.6199	2ZUX	A	354	361	β	β	0.7152
1EEX	A	130	137	α	β	0.6210	1DYP	A	101	108	β	β	0.7203
2GGC	A	156	163	α	α	0.6214	1H4A	X	155	162	α	β	0.7213
2D81	A	57	64	α	α	0.6244	1TBF	A	206	213	α	α	0.7214
1GA6	A	13	20	α	β	0.6321	1QFM	A	571	578	β	α	0.7222
2GAI	A	434	441	α	α	0.6373	1U4G	A	89	96	α	β	0.7258
1DK8	A	99	106	α	α	0.6391	1F5V	A	87	94	α	α	0.7321
1GQI	A	689	696	α	α	0.6403	1H4A	X	66	73	α	β	0.7329
2WHJ	A	84	91	β	α	0.6410	3B8D	A	255	262	α	β	0.7359
2JE8	A	378	385	β	α	0.6439	1PN2	A	222	229	β	β	0.7424
1VLP	A	174	181	β	α	0.6458	3H09	A	440	447	β	β	0.7459
1UAS	A	244	251	β	α	0.6470	3EIX	A	200	207	α	α	0.7475
1H16	A	326	333	α	β	0.6565	2ZQ0	A	470	477	β	α	0.7479
3FDY	A	431	438	β	β	0.6602	2V3Z	A	419	426	β	α	0.7604
3H7C	X	109	116	β	α	0.6609	1GCI	A	16	23	α	β	0.7619
2V03	A	56	63	α	β	0.6609	2BHU	A	395	402	β	α	0.7650
1ATZ	A	111	118	β	α	0.6686	2H1V	A	294	301	β	α	0.7651
2OIZ	A	346	353	β	β	0.6754	3BMV	A	365	372	α	α	0.7663
1O7Q	A	224	231	α	α	0.6779	1GNL	A	394	401	β	α	0.7663
3D03	A	46	53	β	α	0.6780	1OXX	K	326	333	β	β	0.7676
2VBK	A	138	145	β	β	0.6782	1PJX	A	299	306	β	α	0.7700
2RFQ	A	173	180	β	α	0.6784	2CYJ	A	52	59	α	β	0.7716
2ZUX	A	317	324	α	β	0.6815	7A3H	A	181	188	α	β	0.7786
3DQP	A	64	71	β	α	0.6817	1ODZ	A	98	105	β	α	0.7798
2PQC	A	150	157	β	β	0.6827	1NKG	A	467	474	β	β	0.7801
1LFW	A	262	269	β	α	0.6833	2JG0	A	168	175	α	α	0.7835
2E6F	A	16	23	β	α	0.6841	2GZQ	A	62	69	β	β	0.7903
1PQ7	A	160	167	α	β	0.6892	1K2X	A	128	135	α	α	0.7914
3BVX	A	188	195	β	α	0.6935	1OK0	A	42	49	β	β	0.7951
3DG9	A	179	186	β	α	0.6999	1K0M	A	188	195	α	α	0.7980
3HID	A	98	105	β	α	0.8002	1RQB	A	374	381	α	α	0.9002
2HO3	A	208	215	β	β	0.8018	3B8I	A	255	262	α	α	0.9036
3GN6	A	84	91	α	β	0.8045	1E4C	P	102	109	α	α	0.9037
3BPT	A	134	141	α	α	0.8058	2VLQ	B	78	85	α	α	0.9055
1LLF	A	322	329	α	β	0.8090	3BZW	A	81	88	α	β	0.9077
1VLP	A	36	43	β	α	0.8090	3B5M	A	137	144	β	α	0.9079
3GVE	A	53	60	β	α	0.8145	1U5P	A	137	144	α	α	0.9080
2EIY	B	62	69	α	α	0.8149	1LVM	A	9	16	α	α	0.9094
1GNL	A	507	514	β	α	0.8166	2P4H	X	282	289	α	α	0.9098
1OQ1	A	50	57	β	β	0.8184	2INU	A	70	77	β	α	0.9113
1R8S	E	68	75	α	α	0.8192	2VUW	A	173	180	α	α	0.9156

2AYH	A	191	198	α	β	0.8214	2LIS	A	102	109	α	α	0.9172
3DEL	B	96	103	β	α	0.8220	1VKM	A	234	241	α	α	0.9188
1MUN	A	128	135	α	α	0.8407	3CXU	A	245	252	α	β	0.9225
1UAS	A	179	186	β	α	0.8416	1CQX	A	39	46	α	α	0.9228
2E6F	A	66	73	β	α	0.8433	2EZ9	A	325	332	α	α	0.9237
1HP1	A	54	61	β	α	0.8442	2HY7	A	187	194	β	α	0.9240
1J1N	A	146	153	α	α	0.8444	1NNF	A	66	73	α	α	0.9249
2V3V	A	150	157	α	α	0.8507	3H87	C	38	45	α	α	0.9259
1UWF	A	82	89	β	β	0.8540	3FHL	A	278	285	β	α	0.9288
2QMQ	A	138	145	β	α	0.8587	2IYA	A	129	136	β	α	0.9308
2OIZ	D	64	71	β	β	0.8590	3CUZ	A	430	437	α	α	0.9348
1WUI	L	357	364	β	α	0.8604	2GMN	A	115	122	α	β	0.9408
1NZJ	A	235	242	α	α	0.8614	1H16	A	312	319	α	α	0.9409
2H98	A	56	63	β	β	0.8646	2C2U	A	87	94	α	α	0.9443
1KA1	A	68	75	β	α	0.8655	2WAA	A	127	134	β	β	0.9444
2ABK	A	128	135	α	α	0.8656	1VQZ	A	235	242	α	β	0.9552
1B5E	A	138	145	α	β	0.8689	1QLW	A	123	130	α	α	0.9593
1DPE	A	5	12	β	α	0.8734	1YKI	A	198	205	α	α	0.9680
2IYA	A	205	212	β	α	0.8750	1L6P	A	50	57	β	β	0.9745
1NKG	A	453	460	α	β	0.8825	1W2W	B	103	110	α	α	0.9772
1HP1	A	372	379	α	β	0.8836	2QMC	B	157	164	β	β	0.9835
2Z4U	A	111	118	β	α	0.8868	2Z26	A	247	254	β	α	0.9849
3E03	A	145	152	β	α	0.8873	3CJE	A	135	142	α	β	0.9858
3BOE	A	140	147	β	β	0.8878	1VYR	A	337	344	α	α	0.9873
1O9R	A	73	80	α	α	0.8893	2IVF	A	166	173	α	α	0.9942
2R8O	A	634	641	β	α	0.8895	1GWE	A	218	225	β	α	0.9989
3BWH	A	230	237	α	α	0.8899	2JE8	A	542	549	β	α	1.0033
2PQC	A	222	229	β	β	0.8910	2PKF	A	38	45	α	β	1.0228
3BOE	A	114	121	α	α	0.8973	1GK9	A	122	129	α	α	1.0498

Table D.2: The second test set for loop stretch. This test set consists of loops of 6–10 residues. In each number of residues, the numbers of contracted ($\lambda < 0.4$) and stretched ($\lambda > 0.95$) loops are the same (N: N-anchor, C: C-anchor, α: Helix, β: Strand and λ: Loop stretch).

6 Residues													
Contracted ($\lambda < 0.4$)							Stretched ($\lambda > 0.95$)						
Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1B7M	A	100	105	α	α	0.2725	2ETX	A	69	74	α	α	0.9503
1JMX	A	100	105	α	α	0.2787	1NQ7	A	31	36	α	α	0.9504
2ZEZ	A	30	35	α	β	0.2865	1M0S	A	174	179	β	α	0.9507
1FGJ	A	465	470	α	α	0.2892	1DJ0	A	191	196	β	α	0.9508
1ORD	A	231	236	α	β	0.2892	1YHV	A	190	195	α	α	0.9509
1U8E	A	85	90	β	β	0.2909	1T15	A	197	202	α	α	0.9511
1QLH	A	282	287	α	β	0.2961	1EG7	A	91	96	β	α	0.9517
3GVD	A	54	59	α	α	0.2973	1C0F	A	217	222	α	α	0.9522
2W20	A	272	277	β	α	0.3132	2B94	A	172	177	α	α	0.9533
3GHA	A	51	56	α	β	0.3240	3DA4	A	150	155	α	α	0.9562
1PZN	A	214	219	α	α	0.3270	2I9E	A	29	34	α	β	0.9595

1J0A	A	145	150	α	β	0.3358	4CMS	A	55	60	α	α	0.9648
2RDX	A	267	272	β	α	0.3366	1SVT	A	134	139	α	α	0.9660
1XPM	A	107	112	β	α	0.3407	3EG7	A	136	141	α	β	0.9674
1PPJ	D	52	57	α	α	0.3434	1W88	A	330	335	α	α	0.9676
1UA4	A	236	241	α	α	0.3437	2QIO	A	191	196	β	α	0.9705
4CEL	A	368	373	β	α	0.3449	2VHX	A	360	365	α	α	0.9758
3C70	A	9	14	β	α	0.3590	1F20	A	205	210	α	β	0.9823
2A0U	A	208	213	β	α	0.3592	1VLB	A	307	312	β	α	0.9832
1O04	A	187	192	β	α	0.3607	2VBF	A	489	494	α	α	0.9838
2JLG	A	628	633	α	α	0.3626	1XDY	A	199	204	β	α	0.9839
1NH1	A	100	105	β	α	0.3679	2IP6	A	66	71	α	α	0.9863
1A8Q	A	26	31	β	α	0.3723	2EB1	A	15	20	α	α	0.9912
1DMU	A	286	291	α	β	0.3727	1DJE	A	263	268	α	α	0.9927
2C8J	A	86	91	β	α	0.3759	2FQ6	A	236	241	α	α	0.9942
3D3U	A	345	350	α	β	0.3863	1N0U	A	556	561	β	β	0.9946
1X6L	A	463	468	α	β	0.3882	1XYN	A	84	89	β	β	0.9957
2H57	A	55	60	β	α	0.3957	1S5J	A	428	433	α	α	0.9964
5CPA	A	67	72	β	α	0.3975	1Z28	A	205	210	α	α	0.9970

7 Residues													
Contracted ($\lambda < 0.4$)							Stretched ($\lambda > 0.95$)						
Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1UUR	A	403	409	α	α	0.2873	1GK2	A	308	314	α	β	0.9506
1FRF	S	228	234	α	α	0.3047	3DRW	A	397	403	α	α	0.9511
1KCX	A	356	362	α	α	0.3061	3EJ5	X	191	197	α	α	0.9530
2BA1	A	131	137	β	β	0.3078	2EAD	A	123	129	α	β	0.9556
1YIO	A	54	60	β	α	0.3081	1RK6	A	243	249	β	α	0.9560
1JGI	A	340	346	α	β	0.3137	2OOL	A	110	116	β	α	0.9582
3HJC	A	309	315	β	α	0.3147	2I9D	A	100	106	β	α	0.9596
1JHL	A	36	42	α	β	0.3158	1MRG	A	236	242	α	α	0.9606
2FPR	A	147	153	β	α	0.3176	3DRA	B	44	50	α	α	0.9607
3C8U	A	29	35	β	α	0.3366	1G1Y	A	121	127	α	α	0.9668
1VKB	A	84	90	α	β	0.3382	1QP9	A	39	45	α	α	0.9672
1J71	A	189	195	β	β	0.3459	1UC8	A	252	258	β	α	0.9687
2AM1	A	410	416	β	α	0.3467	1XV2	A	217	223	α	α	0.9708
3ESM	A	25	31	β	β	0.3475	3DGK	A	230	236	α	α	0.9709
1RKI	A	26	32	α	β	0.3501	1NM1	A	193	199	α	α	0.9739
1A0I	A	53	59	β	α	0.3513	2A5D	B	34	40	α	α	0.9811
1RCE	A	37	43	α	α	0.3538	3EEZ	A	316	322	α	α	0.9814
1IVY	A	370	376	β	α	0.3540	2CUY	A	77	83	α	β	0.9821
4CMS	A	187	193	β	β	0.3551	3DDS	A	471	477	β	α	0.9822
1YOO	A	252	258	β	α	0.3598	1EEX	A	17	23	α	α	0.9835
1T0I	A	94	100	β	α	0.3714	1YQ2	A	761	767	β	α	0.9836
3FEQ	A	388	394	α	β	0.3721	3CNJ	A	374	380	β	β	0.9861
3CC6	A	68	74	α	β	0.3734	1T1U	A	94	100	β	α	0.9874
1KY3	A	152	158	β	α	0.3742	2YY7	A	293	299	α	α	0.9891
1V6T	A	35	41	β	α	0.3762	1BGX	T	395	401	α	α	0.9899
2ZKI	A	78	84	β	α	0.3766	3C9A	A	219	225	β	β	0.9901
1IAY	A	149	155	α	β	0.3778	3CMM	A	849	855	α	α	0.9922
2YX9	A	434	440	β	β	0.3780	2FSQ	A	170	176	α	β	0.9948
2Z9S	A	19	25	β	β	0.3819	2JG1	A	112	118	β	α	0.9961
2A39	A	347	353	β	α	0.3839	1H8P	A	45	51	α	β	1.0123
1KBB	A	479	485	β	β	0.3847	2RFP	A	78	84	α	α	1.0161

3FJU	B	48	54	α	β	0.3860	2G84	A	19	25	α	α	1.0250
1TCA	A	243	249	α	α	0.3882	3D6R	A	12	18	α	β	1.0297
1G59	A	180	186	β	α	0.3893	3CNJ	A	148	154	α	α	1.0308
2FP3	A	141	147	α	α	0.3949	1AEC	A	43	49	α	α	1.0361
1U6Z	A	95	101	α	β	0.3959	3E0K	A	117	123	α	α	1.0388

8 Residues													
Contracted ($\lambda < 0.4$)							Stretched ($\lambda > 0.95$)						
Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1P2Z	A	766	773	α	α	0.1876	3DBH	B	197	204	α	α	0.9506
1F6B	A	166	173	β	α	0.1945	1QLM	A	127	134	α	β	0.9515
1QWO	A	275	282	α	α	0.1957	1VJG	A	151	158	α	α	0.9516
2A5D	A	143	150	β	α	0.2024	3ELJ	A	257	264	α	α	0.9525
2V5X	A	16	23	α	α	0.2052	2AN0	A	319	326	α	α	0.9543
1ORD	A	166	173	α	α	0.2060	3GED	A	81	88	β	α	0.9550
1JXH	A	12	19	β	α	0.2086	1S48	A	110	117	α	α	0.9565
1SVI	A	174	181	β	α	0.2089	1QTN	A	33	40	α	α	0.9567
2H5E	A	255	262	β	α	0.2155	1SQI	A	278	285	α	α	0.9571
1M1Z	A	414	421	α	α	0.2175	3CF6	E	661	668	α	α	0.9573
2HSJ	A	187	194	α	α	0.2180	1PJT	A	209	216	α	β	0.9577
1PNF	A	18	25	β	β	0.2192	1CIP	A	145	152	α	β	0.9586
1Q4U	A	47	54	α	α	0.2220	3H5Q	A	422	429	β	β	0.9618
2EIS	A	13	20	α	α	0.2243	1RLZ	A	244	251	α	α	0.9641
2FN4	A	148	155	β	α	0.2246	3DM7	A	133	140	β	α	0.9644
3EUC	A	184	191	α	β	0.2249	1X2G	A	245	252	α	β	0.9648
2B3Z	A	42	49	β	α	0.2275	2GP4	A	566	573	α	α	0.9657
2FX5	A	236	243	α	α	0.2286	2BII	A	28	35	β	α	0.9662
1CNS	A	160	167	α	α	0.2303	1W6Q	A	75	82	β	β	0.9713
3BWS	A	249	256	β	β	0.2312	1NO7	A	416	423	α	β	0.9724
5EAS	A	107	114	α	α	0.2315	2RCN	A	238	245	α	α	0.9732
1RTZ	A	81	88	α	β	0.2329	1I4N	A	22	29	α	α	0.9748
2WBL	C	154	161	β	α	0.2339	3CUX	A	397	404	α	α	0.9786
3E1E	A	42	49	α	α	0.2341	1VYS	X	339	346	α	α	0.9805
2DSL	A	30	37	α	α	0.2346	1T5O	A	278	285	α	α	0.9813
2BME	A	149	156	β	α	0.2357	2R14	A	337	344	α	α	0.9814
2QF7	A	778	785	α	α	0.2368	2HSJ	A	127	134	β	α	0.9831
1KG7	A	191	198	α	α	0.2371	3CJE	A	138	145	α	β	0.9858
1CC1	S	245	252	α	α	0.2373	1MTY	D	500	507	α	α	0.9863
2OV9	A	117	124	α	α	0.2377	2FK6	A	152	159	β	α	0.9879
1G12	A	103	110	α	α	0.2394	3EUH	A	11	18	α	α	0.9905
2OAF	A	19	26	α	α	0.2413	1FC9	A	170	177	β	β	0.9910
2NT3	A	53	60	β	α	0.2417	2GOU	A	338	345	α	α	0.9913
1WVU	A	146	153	α	α	0.2423	2IP4	A	312	319	α	β	0.9917
1KHV	A	133	140	α	α	0.2426	1JJ2	O	52	59	β	α	0.9919
3ES3	A	37	44	α	α	0.2443	2VWJ	A	284	291	α	α	0.9920
1RXD	A	96	103	β	α	0.2470	1TQ8	A	113	120	β	α	0.9924
1Z6O	A	62	69	α	α	0.2478	3HB6	A	215	222	β	α	0.9925
2FUJ	A	15	22	α	α	0.2491	1JYO	E	79	86	α	α	0.9950
1Z2I	A	241	248	α	α	0.2497	1F89	A	146	153	α	β	0.9952
2HLJ	A	17	24	α	α	0.2532	2BDW	A	224	231	α	α	0.9958
1MR1	C	22	29	α	β	0.2547	1QO7	A	36	43	α	α	0.9963
1AFV	B	118	125	α	α	0.2562	2VDR	B	251	258	β	α	0.9975
1Z68	A	662	669	β	α	0.2601	2GZA	A	102	109	β	α	0.9995

2QG7	A	113	120	α	β	0.2608	2JE8	A	529	536	β	α	1.0033
1P59	A	188	195	α	α	0.2628	1Z8X	A	181	188	α	α	1.0047
2A5Y	B	437	444	α	β	0.2641	3BU2	A	16	23	β	β	1.0106
1MVL	A	63	70	α	α	0.2645	2BKA	A	48	55	β	α	1.0129
2F1N	A	139	146	β	α	0.2658	1NG2	A	43	50	β	α	1.0153
2FN3	A	397	404	β	α	0.2668	2IUF	A	231	238	β	α	1.0214
2DCM	A	624	631	β	α	0.2671	1ZHX	A	356	363	α	β	1.0322
2FNO	A	15	22	β	α	0.2673	2GUP	A	91	98	α	α	1.0352
2RDM	A	57	64	β	α	0.2689	2BPA	1	352	359	α	α	1.0368
1LNS	A	51	58	α	α	0.2691	1GK9	A	122	129	α	α	1.0498
1EWK	A	397	404	α	α	0.2695	1QGC	1	19	26	α	α	1.0697

9 Residues													
Contracted ($\lambda < 0.4$)							Stretched ($\lambda > 0.95$)						
Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1VC4	A	86	94	β	α	0.1662	1LTK	A	277	285	α	β	0.9534
3BZ1	C	189	197	α	α	0.1840	1TLV	A	57	65	α	α	0.9550
1ILE	A	388	396	β	β	0.1845	2PKJ	A	313	321	α	α	0.9585
1JDP	A	15	23	β	α	0.1921	2E18	A	242	250	α	α	0.9590
1XWS	A	65	73	α	β	0.1964	2W5Y	A	8	16	α	α	0.9640
1VB3	A	98	106	β	α	0.1970	1YVF	A	77	85	α	α	0.9662
1VI9	A	10	18	β	α	0.1993	2AG5	A	81	89	β	α	0.9700
1SU8	A	150	158	α	α	0.2080	2V26	A	272	280	α	α	0.9702
2VA1	A	15	23	α	α	0.2098	3DNU	A	327	335	β	α	0.9721
2IUX	A	236	244	α	α	0.2110	1OFL	A	346	354	α	β	0.9723
2DE6	A	67	75	β	α	0.2142	1TY7	B	181	189	α	β	0.9796
3DI5	A	32	40	α	α	0.2152	1V7M	X	103	111	α	β	0.9804
2IP2	A	83	91	α	α	0.2195	1ZMT	A	79	87	β	α	0.9814
1N1B	A	113	121	α	α	0.2196	2ERK	A	259	267	α	α	0.9821
1PIG	A	123	131	α	α	0.2214	1MJO	A	21	29	α	α	0.9885
1I1J	A	13	21	β	β	0.2258	1R89	A	255	263	α	α	0.9900
1XF1	A	705	713	β	β	0.2263	2F7F	A	193	201	α	α	0.9906
1RZ2	A	106	114	β	α	0.2271	3E03	A	97	105	β	α	0.9916
1SQJ	A	341	349	α	β	0.2291	2I1O	A	173	181	α	α	0.9944
2V8V	A	325	333	α	β	0.2293	1RQP	A	180	188	α	β	0.9978
1BJF	A	49	57	α	α	0.2313	2UUU	A	90	98	α	α	1.0049
2OK5	A	410	418	α	β	0.2324	3GBE	A	370	378	α	α	1.0051
1R5T	A	24	32	α	β	0.2328	2EW8	A	87	95	β	α	1.0063
2ORW	A	111	119	β	α	0.2367	1QQP	1	19	27	α	α	1.0190
3C2G	A	484	492	α	α	0.2380	1OSY	A	14	22	α	β	1.0217
1BRW	A	111	119	β	α	0.2387	2V6G	A	324	332	α	α	1.0248
1T3I	A	144	152	β	α	0.2401	3D7L	A	65	73	β	α	1.0278
2ZYL	A	52	60	β	α	0.2404	1DPC	A	13	21	α	α	1.0467
1Z8X	A	9	17	β	α	0.2407	1YBV	A	101	109	β	α	1.0697

10 Residues													
Contracted ($\lambda < 0.4$)							Stretched ($\lambda > 0.95$)						
Code	Chain	Start	End	N	C	λ	Code	Chain	Start	End	N	C	λ
1O7D	A	37	46	α	α	0.1271	1UKC	A	217	226	β	α	0.9510
2E8Y	A	473	482	α	α	0.1299	1SH0	A	91	100	α	α	0.9528
1FF9	A	383	392	β	α	0.1381	1UYR	A	142	151	β	α	0.9532
1H9A	A	290	299	β	α	0.1455	2Z8F	A	347	356	α	α	0.9552
2QQM	A	358	367	β	β	0.1462	1YB1	A	91	100	β	α	0.9567

2Q40	A	244	253	β	α	0.1542	1O2D	A	311	320	α	α	0.9602
1I6I	A	106	115	α	α	0.1578	3CKJ	A	299	308	β	α	0.9611
1I0A	A	425	434	α	α	0.1584	1GZ6	A	96	105	β	α	0.9622
2Z3I	A	44	53	β	α	0.1609	1SQH	A	158	167	α	β	0.9676
1G66	A	11	20	β	α	0.1627	1KKE	A	41	50	β	β	0.9696
3F2V	A	115	124	α	α	0.1640	2BLN	A	182	191	α	α	0.9723
1XRS	B	110	119	α	α	0.1684	1NY5	A	134	143	α	α	0.9747
3G0T	A	126	135	α	β	0.1688	2FBY	A	7	16	α	β	0.9747
2FFJ	A	181	190	β	α	0.1708	2BIB	A	304	313	α	β	0.9791
2D1S	A	437	446	α	α	0.1764	2DY1	A	368	377	α	β	0.9812
1OMW	A	219	228	α	β	0.1765	2DJI	A	182	191	α	α	0.9821
1TZ7	A	363	372	α	α	0.1766	1IDJ	A	317	326	β	α	0.9855
1DP4	A	10	19	β	α	0.1772	1XV2	A	139	148	α	β	0.9889
1UWC	A	176	185	α	β	0.1792	2DJF	B	138	147	α	α	0.9941
2DE0	X	138	147	β	α	0.1794	1JA9	A	90	99	β	α	1.0076
1RJ9	A	111	120	β	α	0.1803	1YG8	A	108	117	β	α	1.0102
1I4N	A	81	90	β	α	0.1815	2FGT	A	181	190	β	α	1.0126
1K7C	A	188	197	α	α	0.1823	1UP8	A	28	37	α	α	1.0148
1Y0P	A	71	80	α	α	0.1826	2YVL	A	213	222	β	α	1.0371

Bibliography

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM - A New Method for Protein Modeling and Design: Application to Docking and Structure Prediction from the Distorted Native Conformation. *J. Compt. Chem*, 15:488–506. 38

Al-Lazikani, B., Lesk, A. M., and Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, 273(4):927–948. 25, 67, 68

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410. 19, 69

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402. 19, 94

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S., Hubbard, T., Chothia, C., and Murzin, A. (2007). Data growth and its impact on the scop database: new developments. *Nucleic Acids Research*, 36:D419–D425. 21, 72

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181:223–230. 8, 92

Anfinsen, C. B. and Haber, E. (1961). Studies on the Reduction and Re-formation of Protein Disulfide Bonds. *J. Biol. Chem*, 236:1361–1363. 8

Anfinsen, C. B., Haber, E., Sela, M., and White Jr., F. H. (1961). The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *PNAS*, 47:1309–1314. 8

Babor, M. and Kortemme, T. (2009). Multi-constraint computational design suggests that native sequences of germline antibody h3 loops are nearly optimal for conformational flexibility. *PROTEINS*, 75(4):846–858. 72, 73

Bajorath, J. and Sheriff, S. (1996). Comparison of an antibody model with an x-ray structure: the variable fragment of br96. *PROTEINS*, 24(2):152–157. 68

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96. 18

Bastolla, U., Farwer, J., Knapp, E. W., and Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the protein data bank. *EMBO*, 44:79–96. 10

Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys*, 107:3698–3706. 34

Best, R. B., Buchete, N. V., and Hummer, G. (2008). Are current molecular dynamics force fields too helical? *Biophysical Journal*, 95:L07–L09. 18

Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M. (1988). 18th Sir Hans Krebs Lecture, Knowledge-based protein modelling and design. *Eur. J. Biochem*, 172:513–520. 47

Blundell, T. L. and Johnson, L. N. (1976). *PROTEIN CRYSTALLOGRAPHY*. ACADEMIC PRESS, London. 12

Braden, B. C. and Poljak, R. J. (1995). Structural features of the reactions between antibodies and protein antigens. *FASEB*, 9(1):9–16. 70

Brekke, O. H. and Sandlie, I. (2003). Therapeutic antibodies for human diseases at the dawn of the twenty-first century. *Nat Rev Drug Discov.*, 2(1):52–62. 65

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.*, 4:187–217. 18, 33

Bruccoleri, R. and Karplus, M. (1990). Conformational Sampling Using High-Temperature Molecular Dynamics. *Biopolymers*, 29:1847–1862. 29

Bruccoleri, R. E., Haber, E., and Novotny, J. (1988). Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature*, 335(6190):564–568. 69

Bruccoleri, R. E. and Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 26(1):137–168. 69

Brünger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475. 12

Burke, D. F., Deane, C. M., and Blundell, T. L. (2000). Browsing the sloop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, 16(6):513–519. 25

Canutescu, A. A. and Dunbrack Jr., R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.*, 12:963–972. 29

Chen, R., Li, L., and Weng, Z. (2003). Zdock: an initial-stage protein-docking algorithm. *PROTEINS*, 52(1):80–87. 77

Choi, Y. and Deane, C. M. (2010). Fread revisited: Accurate loop structure prediction using a database search algorithm. *PROTEINS*, 78(6):1431–1440. 28, 35, 93, 95, 101

Chothia, C. and Janin, J. (1975). Principles of protein–protein recognition. *Nature*, 256:705–708. 10

Chothia, C. and Lesk, A. M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.*, 196:901–917. 25, 67, 68

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., and Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, 342:877–883. 67

Chou, K. C. (2000). Prediction of tight turns and their types in proteins. *Anal Biochem*, 286(1):1–16. 23, 25

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K. J., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197. 33

Crawford, I. P., Niermann, T., and Kirschner, K. (1987). Prediction of Secondary Structure by Evolutionary Comparison: Application to the α Subunit of Tryptophan Synthase. *PROTEINS*, pages 118–129. 25

Danielson, M. L. and Lill, M. A. (2010). New computational method for prediction of interacting protein loop regions. *PROTEINS*, 78(7):1748–1759. 107

Davies, D. R. and Metzger, H. (1983). Structural basis of antibody function. *Annu Rev Immunol.*, 1:87–117. 68

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. *National Biomedical Research Foundation*, 5:345–352. 19

de Bakker, P. I. W., DePristo, M. A., Burke, D. F., and Blundell, T. L. (2003). Ab Initio Construction of Polypeptide Fragments: Accuracy of Loop Decoy Discrimination by an All-Atom Statistical Potential and the AMBER Force Field With the Generalized Born Solvation Model. *PROTEINS*, 41:21–40. 39, 43

de la Paz, P., Sutton, B. J., Darsley, M. J., and Rees, A. R. (1986). Modelling of the combining sites of three anti-lysozyme monoclonal antibodies and of the complex between one of the antibodies and its epitope. *EMBO*, 5(2):415–425. 70

Deane, C. M. and Blundell, T. L. (2000). A Novel Exhaustive Search Algorithm for Predicting the Conformation of Polypeptide Segments in Proteins. *PROTEINS*, 40:135–144. 29

Deane, C. M. and Blundell, T. L. (2001). CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.*, 10:599–612. 28, 35, 39, 50

DePristo, M. A., de Bakker, P. I. W., Lovell, S. C., and Blundell, T. L. (2003). Ab Initio Construction of Polypeptide Fragments: Efficient Generation of Accurate, Representative Ensembles. *PROTEINS*, 51:41–55. 31, 39, 43

Donate, L. E., Rufino, S. D., Canard, L., and Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci.*, 5:2600–2616. 25, 40, 45, 94

Dutta, S., Burkhardt, K., Young, J., Swaminathan, G. J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H. M. (2009). Data deposition and annotation at the worldwide protein data bank. *Mol Biotechnol.*, 42(1):1–13. 68

Engh, R. A. and Huber, R. (1991). Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Cryst*, A47:392–400. 5

Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J. E., and Oliva, B. (2004). Archdb: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Research*, 32:D185–D188. 25, 45

Fernandez-Fuentes, N., Dybas, J. M., and Fiser, A. (2010). Structural characteristics of novel protein folds. *PLoS Computational Biology*, 6(4):e1000750. 104

Fernandez-Fuentes, N. and Fiser, A. (2006). Saturating representation of loop conformational fragments in structure databanks. *BMC Structural Biology*, 6:15. 38, 40, 94

Fernandez-Fuentes, N., Olivia, B., and Fiser, A. (2006a). A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Research*, 34:2085–2097. 28, 35

Fernandez-Fuentes, N., Zhai, J., and Fiser, A. (2006b). Archpred: a template based loop structure prediction server. *Nucleic Acids Research*, 34:W173–W176. 28

Fidelis, K., Stern, P., Bacon, D., and Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.*, 7:953–960. 38

Finkelstein, A. V. and Reva, B. A. (1992). Search for the stable state of a short chain in a molecular field. *Protein Eng.*, 5:617–624. 29

Fiser, A., Do, R. K. G., and Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci.*, 9:1753–1773. 29, 38, 42, 92, 95, 101

Flory, P. J. (1998). *Statistical Mechanics of Chain Molecules*. Hanser. 92

Freddolino, P. L., Harrison, C. B., Liu, Y., and Schulten, K. (2010). Challenges in protein-folding simulations. *Nature Physics*, 6:751–758. 18

Freddolino, P. L., Park, S., Roux, B., and Schulten, K. (2009). Force field bias in protein folding simulations. *Biophysical Journal*, 96:3772–3780. 18

Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16(2):172–177. 19

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thorton, J. M., and Orengo, C. A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35:D291–D297. 21, 72

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.*, 153(4):1027–1042. 25

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *PNAS*, 89(22):10915–10919. 19

Henrikoff, S. and Henrikoff, J. (1992). Amino-acid substitution matrices from protein blocks. *PNAS*, 89:10915–10919. 49

Higo, J., Collura, V., and Garnier, J. (1992). Development of an Extended Simulated Annealing Method: Application to the Modeling of Complementary Determining Regions of Immunoglobulins. *Biopolymers*, 32:33–43. 29

Hildebrand, P., Goede, A., Bauer, R., Gruening, B., Ismer, J., Michalsky, E., and Preissner, R. (2009). Superlooper–a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Research*, 37:W571–W574. 28, 35, 45

Honegger, A. and Pluckthun, A. (2001). Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J. Mol. Biol.*, 309(3):657–670. 68

Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996). Errors in protein structures. *Nature*, 381(6580):272. 13

Hu, W., Godzik, A., and Skolnick, J. (1997). Sequence–structure specificity–how does an inverse folding approach work? *Protein Eng.*, 10(4):317–331. 36

Hurst, T. (1994). Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.*, 34:190–196. 29

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., and Friesner, R. A. (2004). A Hierarchical Approach to All-Atom Protein Loop Prediction. *PROTEINS*, 55:351–367. 29, 31, 38, 43

Jorgensen, W. and Tirado-Rives, J. (1988). The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666. 33

Kabsch, W. and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22:2577–2637. 8

Karen, A. R., Weigelt, C. A., Nayeem, A., and Krystek Jr., S. R. (2007). Loopholes and missing links in protein modeling. *Protein Sci.*, 16:1–14. 37, 38, 43, 93

Khatib, F., DiMaio, F., Group, F. C., Group, F. V. C., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popovic, Z., Jaskolski, M., and Baker, D. (2011). Crystal structure of a monomeric retroviral

protease solved by protein folding game players. *Nature Structural and Molecular Biology*, 18:1175–1177. 18

Kim, S. H. and Sussman, J. L. (1976). Turn is a conformational pattern in rna loops and bendsπ. *Nature*, 260:645–646. 23

Klepsis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19(2):120–127. 18

Ko, J., Lee, D., Park, H., Coutsias, E. A., Lee, J., and Seok, C. (2011). The FALC-Loop web server for protein loop modeling. *Nucleic Acids Research*, 39:W210–W214. 37

Krause, J. C., Ekiert, D. C., Tumpey, T. M., Smith, P. B., Wilson, I. A., and Crowe, J. E. J. (2011). An insertion mutation that distorts antibody binding site architecture enhances function of a human antibody. *MBio*, 2(1):mBio.00345–10. 80

Kryshtafovych, A., Prlic, A., Dmytriv, Z., Daniluk, P., Milostan, M., Eyrich, V., Hubbard, T., and Fidelis, K. (2007). New tools and expanded data analysis capabilities at the protein structure prediction center. *PROTEINS*, 69(Suppl 8):19–26. 41

Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. (2005). Progress Over the First Decade of CASP Experiments. *PROTEINS*, Suppl 7:225–236. 16

Kwasigroch, K., Chomilier, J., and Mornon, J. (1996). A global taxonomy of loops in globular proteins. *J. Mol. Biol.*, 259(4):855–872. 25

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132. 2

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291. ix, 7

Lee, J., Lee, D., Park, H., Coutsias, E. A., and Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *PROTEINS*, 78(16):3428–3436. 29, 30

Lee, S. and Blundell, T. L. (2009). Ulla: a program for calculating environment-specific amino acid substitution tables. *Bioinformatics*, 25:1976–1977. 44

Lefranc, M. P. (1997). Unique database numbering system for immunogenetic analysis. *Immunol. Today*, 18(11):509. 67

Lefranc, M. P. (2011). Antibody nomenclature: from imgt-ontology to inn definition. *MAbs*, 3(1):1–2. 67

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Reginier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). Imgt, the international immunogenetics information system. *Nucleic Acids Research*, 37:D1006–D1012. 71

Lesk, A. M. (1991). *Protein Architecture: A Practical Approach*. Oxford University Press, New York. xii, 12

Lessel, U. and Schomburg, D. (1999). Importance of Anchor Group Positioning in Protein Loop Prediction. *PROTEINS*, 37:56–64. 38, 61

Leszczynski, J. F. and Rose, G. D. (1986). Loops in Globular Proteins: A Novel Category of Secondary Structure. *Science*, 234:849–855. 23, 24

Levitt, M. and Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *PNAS*, 95:5913–5920. 36

Li, A. and Nussinov, R. (1998). A set of van der waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *PROTEINS*, 32(1):111–127. xii, 31

Linderstrom-Lang, K. U. (1952). *Proteins and Enzymes*. Lane Medical Lectures. Stanford University Press. 8

MacKerell, A. D. J., Bashford, D., Bellott, M., Dunbrack, R. L. J., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E. I., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., D. Yin, and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616. 18, 33

Mandal, C., Kingery, B. D., Anchin, J. M., Subramaniam, S., and Linthicum, D. S. (1996). Abgen: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol.*, 14(3):323–328. 69

Manivel, V., Sahoo, N. C., Salunke, D. M., and Rao, K. V. (2000). Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity*, 13(5):611–620. 69

Marcatili, P., Rosi, A., and Tramontano, A. (2008). Pigs: automatic prediction of antibody structures. *Bioinformatics*, 24(17):1953–1954. 69

Martin, A. C., Cheetham, J. C., and Rees, A. R. (1989). Modeling antibody hypervariable loops: a combined algorithm. *PNAS*, 86(23):9268–9272. 70

Martin, A. C. and Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.*, 263(5):800–815. 69

Michalsky, E., Goede, A., and Preissner, R. (2003). Loops In Proteins (LIP) – a comprehensive loop database for homology modeling. *Protein Eng.*, 16:979–985. 28, 35, 38, 45

Miyazawa, S. and Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644. 35

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998a). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, 14:617–623. 40, 94

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998b). Homstrad: a database of protein structure alignments for homologous families. *Protein Sci.*, 7:2469–2471. 48

Morea, V., Lesk, A. M., and Tramontano, A. (2000). Antibody modeling: implications for engineering and design. *Methods*, 20(3):267–279. 68

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., and Thorton, J. M. (1992). Stereochemical Quality of Protein Structure Coordinates. *PROTEINS*, 12:345–364. 12

Morrison, S. L., Johnson, M. J., Herzenberg, L. A., and Oi, V. T. (1984). Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *PNAS*, 81(21):6851–6855. 65

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Current Opinion in Structural Biology*, 7:194–199. 34

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., and Tramontano, A. (2007). Critical assessment of methods of protein structure prediction - Round VII. *PROTEINS*, 69:3–9. 16

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., and Tramontano, A. (2009). Critical

assessment of methods of protein structure prediction - round viii. *PROTEINS*, Suppl 9:1–4. 16, 18

Mundorff, E. C., Hanson, M. A., Varvak, A., Ulrich, H., Schultz, P. G., and Stevens, R. C. (2000). Conformational effects in biological catalysis: an antibody-catalyzed oxy-cope rearrangement. *Biochem.*, 39(4):627–632. 69

Myers, J. K. and Pace, C. N. (1996). Hydrogen bonding stabilizes globular proteins. *Biophysical Journal*, 71(4):2033–2039. 10

Nagi, A. D. and Regan, L. (1997). An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold Des*, 2(1):67–75. 92, 101

Narang, P. (2006). Protein structure evaluation using an all-atom energy based empirical scoring function. *J Biomol Struct Dyn*, 23:385–406. 35

Nemethy, G. and Printz, M. P. (1972). The γ turn, a possible folded conformation of the polypeptide chain. comparison with the β turn. *Macromolecules*, 5(6):755–758. 23

Nguyen, H. P., Seto, N. O., MacKenzie, C. R., Brade, L., Kosma, P., Brade, H., and Evans, S. V. (2003). Germline antibody recognition of distinct carbohydrate epitopes. *Nat Struct Biol.*, 10(12):1019–1025. 69

North, B., Lehmann, A., and Dunbrack, R. L. J. (2011). A new clustering of antibody cdr loop conformations. *J. Mol. Biol.*, 406(2):228–256. 68

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. E. (1997). An Automated Classfication of the Structure of Protein Loops. *J. Mol. Biol.*, 266:814–830. 25

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. E. (1998). Automated classification of antibody complementarity determining region 3 of the

heavy chain (h3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.*, 279(5):1193–1210. 69

Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB*, 10:75–83. 9

Pardon, E., Haezebrouck, P., De Baetselier, A., Hooke, S., Fancourt, K., Desmet, J., Dobson, C., Van Dael, H., and Joniau, M. (1995). A ca(2+)-binding chimera of human lysozyme and bovine alpha-lactalbumin that can form a molten globule. *J. Biol. Chem*, 270(18):10514–10524. 26

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *PNAS*, 37:205–211. 5

Pavlou, A. K. and Belsey, M. J. (2005). The therapeutic antibodies market to 2008. *Eur J Pharm Biopharm.*, 59(3):389–396. 65

Pei, X. Y., Holliger, P., Murzin, A., and Williams, R. L. (1997). The 2.0-a resolution crystal structure of a trimeric antibody fragment with noncognate vh-vl domain pairs shows a rearrangement of vh cdr3. *PNAS*, 94(18):9637–9642. 70

Peng, H. and Yang, A. (2007). Modeling protein loops with knowledge-based prediction of sequence-structure alignment. *Bioinformatics*, 23:2836–2842. 28, 35, 38

Queen, C., Schneider, W., Selick, H., Payne, P., Landolfi, N., Duncan, J., Avdalovic, N., Levitt, M., Junghans, R., and Waldmann, T. (1989). A humanized antibody that binds to the interleukin 2 receptor. *PNAS*, 86(24):10029–10033. 26

Ramakrishnan, C. and Ramachandran, G. N. (1965). Stereochemical Criteria for Polypeptide and Protein Chain Conformations: II. Allowed Conformations for a Pair of Peptide Units. *Biophysical Journal*, 5:909–933. 6

Ramsland, P. A., Guddat, L. W., Edmundson, A. B., and Raison, R. L. (1997). Diverse binding site structures revealed in homology models of polyreactive immunoglobulins. *J Comput Aided Mol Des.*, 11(5):453–461. 68

Rapp, C. S. and Friesner, R. A. ((1999)). Prediction of Loop Geometries Using a Generalized Born Model of Solvation Effects. *PROTEINS*, 35:173–183. 29

Rata, I. A., Li, Y., and Jakobsson, E. (2010). Backbone statistical potential from local sequence-structure interactions in protein loops. *J Phys Chem B*, 114:1859–1869. 35

Reichert, J. and Pavlou, A. K. (2004). Monoclonal antibodies market. *Nat Rev Drug Discov.*, 3(5):383–384. 65, 68

Reichert, J. M., Rosensweig, C. J., Faden, L. B., and Dewitz, M. C. (2005). Monoclonal antibody successes in the clinic. *Nat Biotechnol.*, 23(9):1073–1078. 65

Richardson, J. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, 34:167–339. 25

Riechmann, L., Clark, M., Waldmann, H., and Winter, G. (1988). Reshaping human antibodies for therapy. *Nature*, 332(6162):323–327. 26

Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E. (1991). Taxonomy and Conformational Analysis of Loops in Proteins. *J. Mol. Biol.*, 224:685–699. 25, 92

Rini, J. M., Schulze-Gahmen, U., and Wilson, I. A. (1992). Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science*, 255(5047):959–962. 70

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93. 29, 70

Rose, G. D. (1978). Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature*, 272:586–590. 25

Rose, G. D., Young, W. B., and Gierasch, L. M. (1983). Interior turns in globular proteins. *Nature*, 304:654–657. 25

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, 12:85–94. 18

Samudrala, R. and Moult, J. (1998a). A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure. *J. Mol. Biol.*, 279:287–302. 44, 46

Samudrala, R. and Moult, J. (1998b). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916. 35

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000). Protein structure modeling for structural genomics. *Nature Structural Biology*, 7:Suppl. 986–990. 18

Schuermann, J. P., Prewitt, S. P., Davies, C., Deutcher, S. L., and Tanner, J. J. (2005). Evidence for structural plasticity of heavy chain complementarity-determining region 3 in antibody-ssdna recognition. *J. Mol. Biol.*, 347(5):965–978. 69

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H. M., Jones, D., Brenner, S., Burley, S. K., Das, R., Dokholyan, N. V., Dunbrack, R. L. J., Fidelis, K., Fiser, A., Godzik, A., Huang, Y., Humblet, C., Jacobson, M. P., Joachimiak, A., Krystek, S. R. J., Kortemme, T., Kryshtafovych, A., Montelione, G. T., Moult, J., Murray, D., Sanchez, R., Sosnick, T. R., Standley, D. M., Stouch, T., Vajda, S., Vasquez, M., Westbrook, J. D., and Wilson, I. A. (2009). Outcome of a workshop on applications of protein models in biomedical research. *Structure*, 17(2):151–159. 68

Scott, K. A., Bond, P. J., Ivetac, A., Chetwynd, A. P., Khalid, S., and Sansom, M. S. P. (2008). Coarse-grained md simulations of membrane protein-bilayer self-assembly. *Structure*, 16(4):621–630. 94

Sethi, D. K., Agarwal, A., Manivel, V., Rao, K. V., and Salunke, D. M. (2006). Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity*, 24(4):429–438. 69

Shen, M. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15:2507–2524. 42, 95

Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H., and Levinthal, C. (1987). Predicting Antibody Hypervariable Loop Conformation. I. Ensembles of Random Conformations for Ringlike Structures. *Biopolymers*, 26:2053–2085. 30

Shi, J., Blundell, T. L., and Mizuguchi, K. (2010). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure- dependent gap penalties. *J. Mol. Biol.*, 310:243–257. 21, 44

Shirai, H., Kidera, A., and Nakamura, H. (1999). H3-rules: identification of cdr-h3 structures in antibodies. *FEBS Lett.*, 455:188–197. 69, 70

Sibanda, B. L. and Thorton, J. M. (1985). β-Hairpin families in globular proteins. *Nature*, 316:170–174. 25

Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785. 36

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268(1):209–225. 34, 70

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213(4):859–883. 34

Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J. (2009). Toward high-resolution homology modeling of antibody fv regions and application to antibody-antigen docking. *PROTEINS*, 74(2):497–514. 68, 69, 72, 79

Spassov, V. Z., Flook, P. K., and Yan, L. (2008). LOOPER: a molecular mechanics-based algorithm for protein loop prediction. *Protein Eng.*, 21:91–100. 29

Sreenivasan, U. and Axelsen, P. H. (1992). Buried water in homologous serine proteases. *Biochem.*, 31:12785–12791. 5

Stevens, T. J. and Arkin, I. T. (1999). Are membrane proteins "inside-out" proteins? *PROTEINS*, 36(1):135–143. 10

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.*, 112:6127–6129. 33

Sucha, S., Dubose, R. F., March, C. J., and Subashini, S. (1995). Modeling protein loops using a phi(i+1), psi(i) dimer database. *Protein Sci.*, 4:1412–1420. 29

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., and Abola, E. E. (1998). Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Cryst.*, D54:1078–1084. 10

Tastan, O., Klein-Seetharaman, J., and Meirovitch, H. (2009). The effect of loops on the structural organization of alpha-helical membrane proteins. *Biophysical Journal*, 96(6):2299–2312. 92

Toma, S., Campagnoli, S., Margarit, I., Gianna, R., Grandi, G., Bolognesi, M., De Filippis, V., and Fontana, A. (1991). Grafting of a calcium-binding loop of thermolysin to bacillus subtilis neutral protease. *Biochem.*, 30(1):97–106. 26

Toniolo, C. (1980). Intramolecularly hydrogen-bonded peptide conformations. *CRC Crit Rev Biochem*, 9(1):1–44. 23

Tripos Inc. (2005). SYBYL. 38, 47

Tusnady, G. E., Doztanyi, Z., and Simon, I. (2004). Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics*, 20(17):2964–2972. 94

van Vlijmen, H. W. T. and Karplus, M. (1997). PDB-based Protein Loop Prediction: Parameters for selection and Methods for Optimization. *J. Mol. Biol.*, 267:975–1001. 29

Vanhee, P., Verschueren, E., Baeten, L., Stricher, F., Serrano, L., Rousseau, F., and Schymkowitz, J. (2011). Brix: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Research*, 39:D435–D442. 25, 26

Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. v. conformation of a system of three linked peptide units. *Biopolymers*, 6(10):1425–1436. 23, 25

Walsh, G. (2006). Biopharmaceutical benchmarks 2006. *Nat Biotechnol.*, 24(7):769–776. 68

Wang, G. and Dunbrack, R. L. J. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, 12:1589–1591. 40, 94

Wedemayer, G. J., Patten, P. A., Wang, L. H., Schultz, P. G., and Stevens, R. C. (1997). Structural insights into the evolution of an antibody combining site. *Science*, 276(5319):1665–1669. 69

Whitelegg, N. R. and Rees, A. R. (2000). Wam: an improved algorithm for modelling antibodies on the web. *Protein Eng.*, 13(12):819–824. 69

Wilson, I. A. and Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Current Opinion in Structural Biology*, 4(6):857–867. 70

Wojcik, J., Mornon, J.-P., and Chomilier, J. (1999). New Efficient Statistical Sequence-dependent Structure Prediction of Short to Medium-sized Protein Loops Based on an Exhaustive Loop Classification. *J. Mol. Biol.*, 289:1469–1490. 25, 28

Wolfson, A., Kanaoka, M., Lau, F., and Ringe, D. (1991). Insertion of an elastase-binding loop into interleukin-1 beta. *Protein Eng.*, 4:313–317. 26

Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. (1995). Navigating the folding routes. *Science*, 267(5204):1619–1620. 17

Wu, T. and Kabat, E. (1970). An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.*, 132(2):211–250. 67

Xiang, Z., Soto, C. S., and Honig, B. (2002). Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *PNAS*, 99(11):7432–7437. 29

Yin, J., 4th. Beuscher, A. E., Andryski, S. E., Stevens, R. C., and Schultz, P. G. (2003a). Structural plasticity and the evolution of antibody affinity and specificity. *J. Mol. Biol.*, 330(4):651–656. 69

Yin, J., Andryski, S. E., 4th. Beuscher, A. E., Stevens, R. C., and Schultz, P. G. (2003b). Structural evidence for substrate strain in antibody catalysis. *PNAS*, 100(3):856–861. 69

Yin, J., Mundorff, E. C., Yang, P. L., Wendt, K. U., Hanway, D., Stevens, R. C., and

Schultz, P. G. (2001). A comparative analysis of the immunological evolution of antibody 28b4. *Biochem.*, 40(36):10764–10773. 69

Zelma, A. (2003). Lga: a method for finding 3d similarities in protein structures. *Nucleic Acids Research*, 31:391–399. 41

Zelma, A., Venclovas, C., Moult, J., and Fidelis, K. (2001). Processing and evaluation of predictions in casp4. *PROTEINS*, 45(Suppl 5):13–21. 36, 60

Zhang, C., Liu, S., and Zhou, Y. (2004). Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci.*, 13:391–399. 43

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *PROTEINS*, 57:702–710. 36, 37, 60

Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, 11:2714–2726. 35