Computational studies of structural motifs and cotranslational folding mechanisms in membrane and soluble proteins

Eleanor Law

New College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

Abstract

Membrane proteins are an important class of drug targets, making up at least 25% of proteins in the human genome. In this thesis I investigated two aspects of alpha-helical membrane protein structures.

Firstly, I investigated kinks in alpha-helices, many of which are thought to have functional roles. Kinks are changes of direction in helices, often defined in a binary fashion, but here I move towards defining them on a continuum. I found that kink angles are not generally a conserved property of homologues, pointing either to their not being functionally critical or to their function being related to conformational flexibility. I found correlation in kink angles and conformational change upon activation in GPCRs, reinforcing the belief that helix kinks are key, functional, flexible points in structures.

Secondly, I turned to the biogenesis of alpha-helical membrane proteins, and how this might be used to improve structure prediction. These proteins are inserted into the membrane during the process of translation by the ribosome, therefore the N-terminus may be able to adopt its tertiary fold before the C-terminus is translated. I found a weak signal in a non-redundant set of structures that membrane proteins exhibit asymmetry between the N- and C-termini. This might be expected if they are folding cotranslationally, as had been seen in soluble proteins. Motivated by this, I predicted the structures of membrane proteins using SAINT2, a cotranslational structure prediction program, and achieved promising results. I developed SAINT2-ScafFold, which folds proteins around a rigid N-terminus, but the accuracy of prediction of the remaining protein was no better than when the entire chain was sampled. A membrane potential was implemented in SAINT2, which slightly improves the accuracy of models generated.

Finally, the SAINT2-ScafFold method was applied to the completion of homology models that do not cover the entire target. An RMSD of less than 5 Å was achieved in more than half of the cases where a terminal transmembrane helix of membrane protein structures was predicted. This was an encouraging result for the prediction of membrane proteins from partial templates, and could easily be extended to soluble proteins.

Computational studies of structural motifs and cotranslational folding mechanisms in membrane and soluble proteins

Eleanor Law

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

This thesis is dedicated to my parents, who have made me the person I am today.

Acknowledgements

I would first like to thank my supervisor, Professor Charlotte Deane, who has been an essential source of expertise, wisdom and optimism throughout the course of my DPhil. I am extremely grateful that Charlotte has consistently made time for mundane questions and reading many drafts of this and other documents, even in her extremely busy schedule. Our meetings have always been enjoyable and a great encouragement, and have somehow digressed to a conversation about cycling more often than not. In addition, I would like to acknowledge the valuable guidance provided by my supervisors at UCB, Sebastian Kelm and Jiye Shi. I am grateful for the funding of both the EPSRC and UCB which has made this DPhil possible, and the staff of the DTC and Department of Statistics.

I would also like to thank the Oxford Protein Informatics Group (OPIG) for being such a welcoming group of people, and also for their contributions of ideas to my project. All have helped me to enjoy the last four years, but I will now mention a few individuals who have made an outstanding contribution to my work and happiness. Henry Wilman generously provided me with a solid foundation to start off my DPhil in the area of kink conservation. Saulo de Oliveira has been an incredibly reliable source of so much more than just answers to hundreds of questions about SAINT2. Clare West has brought a new lease of life to my enthusiasm for science that will make me very sad to leave. Saulo and Clare are brilliant people with whom to discuss scientific ideas and I have really loved the experience of sharing our new results with each other. Thanks to them both also for proofreading. The entire OPIG cryptic crossword team, and particularly Claire Marks, has helped me greatly by providing alternative mental stimulation at lunchtimes. I would also like to thank the Reliance Way crossword team and Ross Johnstone, who nurtured us through our first steps in cryptic crosswording. Two members of OPIG have been particularly close friends: Cristian Regep and Jinwoo Leem. Their company at unusual office hours and their generosity in bike-related items and unhealthy food has sustained me when the going got tough. I will dearly miss the daily conversations about the science of our various sports, and occasionally about the bigger questions of life.

I extend my thanks to all of the members of the Oxford University Squash Racquets Club, especially our coach Ben Rosec and the members of the women's

Blues squad over my time in Oxford. I am also grateful to anyone who has got up early to play with me before work to kick my brain into gear, notably James Mbewu who could always be called upon at short notice. In the recent months of thesis writing, the Cowley Road Condors have provided an excellent escape on two wheels and I would like to thank Cyndi Goh and Mimi Harrison for motivating me to test my limits and enjoy myself.

I would like to thank all of my family for their unwavering love and support, and all the amazing friends I have had the privilege of getting to know in my time in Oxford. In particular, I am grateful to Alan Lewis and Joe Hitchen, who have travelled this journey in parallel with me, Olga Sedelnikova and Nathan Jones. Thanks to my wonderful and generous sister Chrissie Law for proofreading.

I would finally like to thank the Postgrads community at St Aldates church, and the friends who have built me up in my faith over these years even through the challenges. I thank God for giving me my love of science, for sustaining and inspiring me, and for shaping me to be better able to bring the light of Jesus to the people I have met along the way.

Abstract

Membrane proteins are an important class of drug targets, making up at least 25% of proteins in the human genome. In this thesis I investigated two aspects of alpha-helical membrane protein structures.

Firstly, I investigated kinks in alpha-helices, many of which are thought to have functional roles. Kinks are changes of direction in helices, often defined in a binary fashion, but here I move towards defining them on a continuum. I found that kink angles are not generally a conserved property of homologues, pointing either to their not being functionally critical or to their function being related to conformational flexibility. I found correlation in kink angles and conformational change upon activation in GPCRs, reinforcing the belief that helix kinks are key, functional, flexible points in structures.

Secondly, I turned to the biogenesis of alpha-helical membrane proteins, and how this might be used to improve structure prediction. These proteins are inserted into the membrane during the process of translation by the ribosome, therefore the N-terminus may be able to adopt its tertiary fold before the C-terminus is translated. I found a weak signal in a non-redundant set of structures that membrane proteins exhibit asymmetry between the N- and C-termini. This might be expected if they are folding cotranslationally, as had been seen in soluble proteins. Motivated by this, I predicted the structures of membrane proteins using SAINT2, a cotranslational structure prediction program, and achieved promising results. I developed SAINT2-ScafFold, which folds proteins around a rigid N-terminus, but the accuracy of prediction of the remaining protein was no better than when the entire chain was sampled. A membrane potential was implemented in SAINT2, which slightly improves the accuracy of models generated.

Finally, the SAINT2-ScafFold method was applied to the completion of homology models that do not cover the entire target. An RMSD of less than 5 Å was achieved in more than half of the cases where a terminal transmembrane helix of membrane protein structures was predicted. This was an encouraging result for the prediction of membrane proteins from partial templates, and could easily be extended to soluble proteins.

Contents

List of Figures **xv**

List of Tables **xix**

List of Abbreviations **xxi**

1 Introduction **1**

1.1 Protein structure . . . 1
1.1.1 Primary structure . . . 1
1.1.2 Secondary structure . . . 2
1.1.3 Tertiary structure . . . 6
1.1.4 Quaternary structure . . . 6
1.1.5 Experimental determination of protein structure . . . 7
1.2 Protein folding . . . 8
1.2.1 Evidence for cotranslational folding . . . 10
1.3 Membrane Proteins . . . 12
1.3.1 Biological membranes . . . 12
1.3.2 Membrane protein structure . . . 13
1.3.2.1 Experimental determination . . . 14
1.3.2.2 Types of membrane protein . . . 15
1.3.2.3 Residue propensities for membrane layers . . . 16
1.3.3 Adding information to structurally characterised membrane proteins . . . 17
1.3.4 Alpha bundles . . . 17
1.3.4.1 Alpha-helical kinks and their identification . . . 18
1.4 Membrane protein structure prediction . . . 19
1.4.1 Transmembrane helix prediction . . . 20
1.4.2 Assessing the accuracy of three-dimensional models . . . 21
1.4.3 Template-based structure prediction . . . 23
1.4.3.1 Sequence alignment . . . 23
1.4.3.2 Coordinate generation . . . 24
1.4.4 De novo structure prediction . . . 25

1.4.4.1 Coevolutionary contact prediction 26
1.4.4.2 Fragment library generation 29
1.4.4.3 SAINT2 30
1.5 Thesis summary 33

2 Examining the conservation of kinks in alpha-helices 35
2.1 Background 36
2.1.1 Identification of kinks 36
2.1.2 Sequence, function and flexibility of kinks 38
2.2 Methods 41
2.2.1 Angle measurement by Kink Finder 41
2.2.2 Method of confidence interval estimation 42
2.2.3 Data sets 42
2.2.3.1 Identifying homologous helices 43
2.2.3.2 Identifying homologous aligned families of helices 45
2.2.3.3 Obtaining helix families for the seven transmembrane helices (TMHs) of G-protein coupled receptor structures (GPCRs) 45
2.2.4 Comparison of two helices using error estimation 46
2.2.4.1 Calculation of 'neighbouring sequence identity'. 46
2.2.5 Classification of families based on a 'most disrupted' site 47
2.3 Results 48
2.3.1 Confidence intervals of angles measured by Kink Finder . 48
2.3.2 Homologous helix pairs 50
2.3.2.1 Number of helix pairs found 50
2.3.2.2 Definition of pair classes 51
2.3.2.3 Presence of proline in kink pairs 51
2.3.3 Relationship between angle difference and sequence identity 53
2.3.4 Homologous helix families 53
2.3.4.1 Number of families found 53
2.3.4.2 Definition of family classes 56
2.3.4.3 Prevalence of proline in different family classes . 56
2.3.4.4 Relationship between angle variation and sequence conservation 59
2.3.5 G-protein coupled receptor case study 59
2.3.5.1 Angle variation relationship to sequence or flexibility 63
2.3.5.2 Correlation between kink angles 66
2.4 Discussion 66

3 Evidence for cotranslational folding in membrane proteins 71
3.1 Background . . . 71
3.1.1 Membrane protein insertion and folding *in vivo* . . . 72
3.1.2 Computational measures of cotranslational folding . . . 75
3.1.3 *De novo* membrane protein structure prediction . . . 77
3.1.4 Outline . . . 77
3.2 Methods . . . 78
3.2.1 Membrane protein sets . . . 78
3.2.1.1 Topology . . . 80
3.2.2 Statistical measures of cotranslational folding . . . 81
3.2.2.1 Mean central residue (MCR) . . . 81
3.2.2.2 NC_{cen} . . . 82
3.2.2.3 Sum of the log-transformed ratios (SLR) . . . 83
3.2.3 Stability of segments of native structures . . . 83
3.2.3.1 MPrelax protocol . . . 84
3.2.4 Comparison of sequence adjacent pairs according to orientation . . . 85
3.2.5 Prediction of membrane protein structures by SAINT2 . . . 87
3.3 Results and discussion . . . 88
3.3.1 Statistical measures of cotranslational folding . . . 88
3.3.2 Stability of segments of native structures . . . 91
3.3.3 Comparison of sequence adjacent pairs according to orientation . . . 94
3.3.4 Prediction of membrane protein structures by SAINT2 . . . 98
3.3.4.1 Relaxation and scoring by MPrelax . . . 103
3.4 Conclusions . . . 107

4 Adaptation of SAINT2 for membrane proteins 109
4.1 Methods . . . 111
4.1.1 SAINT2-ScafFold . . . 111
4.1.1.1 Implementation . . . 111
4.1.1.2 ScafFold scoring . . . 114
4.1.2 Membrane potential . . . 114
4.1.3 Datasets . . . 116
4.2 Results and discussion . . . 118
4.2.1 SAINT2-ScafFold performance . . . 118
4.2.1.1 Comparison of SAINT2-ScafFold and SAINT2-Wholly . . . 118
4.2.1.2 The effect of segment length on TM-score . . . 123

4.2.1.3 SAINT2-ScafFold without adjustment for the number of moves . . . 125
4.2.2 Implementation of the membrane potential . . . 125
4.2.2.1 Testing the membrane potential for decoy ranking 125
4.2.2.2 Testing the membrane potential during decoy generation . . . 131
4.2.2.3 Ranking of decoys generated with membrane potential . . . 136
4.2.2.4 Performance of decoy generation and ranking on test set . . . 138
4.2.3 Sampling efficiency . . . 140
4.3 Conclusions . . . 148

5 Completion of partial homology models **151**
5.1 Background . . . 151
5.2 Methods . . . 153
5.2.1 Prevalence of incomplete homology models for human membrane protein targets . . . 153
5.2.2 Native Structures completed by SAINT2-ScafFold . . . 153
5.2.2.1 Decoy ranking using linear models LM1 and LM2 156
5.2.3 Homology models completed by SAINT2-ScafFold . . . 156
5.3 Results and discussion . . . 158
5.3.1 Prevalence of incomplete homology models for human membrane protein targets . . . 158
5.3.2 Native structures completed by SAINT2-ScafFold . . . 162
5.3.3 Homology models completed by SAINT2-ScafFold . . . 170
5.4 Conclusions . . . 174

6 Conclusions and future work **177**
6.1 Kink evolution and flexibility . . . 177
6.2 Cotranslational folding in membrane proteins . . . 178
6.3 SAINT2-ScafFold as a model for cotranslational folding . . . 179
6.4 Adaptation of SAINT2 for membrane proteins . . . 180
6.5 Sampling efficiency . . . 181
6.6 Completion of partial homology models . . . 181

Bibliography **183**

Appendices

A Estimation of error in kink angle measurement **211**

B Tables of proline in kink pairs **215**

C Membrane protein datasets **219**
C.1 PDB codes in Set 1 only, total 72 219
C.2 PDB codes in both Set 1 and Set 2, total 21 221
C.3 PDB codes in Set 2 but not in Set1, total 34 222

List of Figures

1.1 The structure of an amino acid 2
1.2 The structures of the 20 amino acids 3
1.3 Backbone torsion angles 4
1.4 Ramachandran plot 4
1.5 Helices 5
1.6 Beta-sheets 5
1.7 A protein folding landscape 9
1.8 A phosphatidylcholine bilayer 13
1.9 A membrane protein coloured by membrane layer 16
1.10 Topology of alpha-helical membrane proteins 18
1.11 Growth of sequence data 19
1.12 Removal of predicted contacts which are incompatible with topology prediction 28
1.13 Forward and In vitro modes of SAINT2 32
1.14 Best TM-scores produced by In vitro and Forward modes 33

2.1 Examples of kinks showing associated features 36
2.2 Angle measurement by Kink Finder 42
2.3 Distribution of angles measured in helices from the membrane, soluble and GPCR data sets 44
2.4 Flowchart showing the classification of homologous helix pairs and families 47
2.5 Angle measurements and smoothing in an example family 48
2.6 The error for the maximum kink angles in the membrane and soluble helices 49
2.7 Examples of helix pairs which are not significantly different and significantly different 50
2.8 Difference in angle plotted against sequence identity 55
2.9 Distribution of sizes of families of at least five members. 56
2.10 Illustrations of a homologous helix family from each of the three main classes 57
2.11 Proline occurrence in helix families 58

2.12 Boxplot of standard deviation of angles in a family against mean sequence identity in the family 60
2.13 Scatterplot of standard deviation of angles in a family against mean sequence identity in the family 61
2.14 Distributions of angles measured at each site of the seven transmembrane helices in the GPCR family 62
2.15 GPCR kink angle variation 63
2.16 Bimodal angle distributions 64
2.17 Confidence intervals for angle measurements in bimodal angle populations 65
2.18 Correlation between kink angles in GPCR TMH 3 and 5 66

3.1 Structure of ribosome-Sec complex 73
3.2 Arrest peptides and transmembrane helix insertion 74
3.3 Chain length and transmembrane span distributions 79
3.4 Comparison between span identification methods 81
3.5 A conceptual visualisation of the statistical measures of cotranslational folding 82
3.6 Illustration of the extraction of segments 84
3.7 Length-dependence of the RosettaMP score 86
3.8 Distributions of statistical measures of cotranslational folding in Set2 90
3.9 Extraction and comparison of terminal segments 91
3.10 Histograms to show the distribution of $\Delta\bar{t}$ and ΔS 93
3.11 Average C-terminal TM-score to native plotted against average N-terminal TM-score to native 94
3.12 C-terminal length-normalised score plotted against N-terminal length-normalised score 95
3.13 C-terminal length-normalised score against N-terminal length-normalised score in four example PDB structures 96
3.14 Histogram to compare the interaction strength in the TMH pairs connected by an intracellular loop to those connected by an extracellular loop 98
3.15 Histogram to compare the loop length joining TMHs in the two groups of TMH pairs 99
3.16 Best TM-score produced by Forward and In vitro modes 100
3.17 Best TM-score produced by Forward and Reverse modes 101
3.18 Crystal structures and models of TatC 102
3.19 RosettaMP scores of the decoys before and after the MPrelax protocol 105

3.20 Distribution of distances from the centroid of relaxed decoys to the plane at the centre of the membrane 106

4.1 SAINT2-ScafFold 111
4.2 Move adjustment for SAINT2-ScafFold 113
4.3 Membrane potential pseudo-energy 115
4.4 Boxplots of the TM-scores of decoys generated by the Forward ScafFold and Wholly protocols 120
4.5 Boxplots of the TM-scores of decoys generated by the Reverse ScafFold and Wholly protocols 121
4.6 Comparison of models generated for two targets 122
4.7 Box and violin plots to show the effect of segment length on TM-score 124
4.8 Boxplots of the TM-scores of decoys generated by the Forward ScafFold with no move adjustment 126
4.9 Correlation between membrane potential and TM-score 129
4.10 Boxplots of the TM-scores of decoys generated by the Forward ScafFold protocol 132
4.11 Boxplots of the TM-scores of decoys generated by the Reverse ScafFold protocol 133
4.12 Comparisons of the TM-score of the best decoy generated using different weights 135
4.13 Acceptance ratio during Forward decoy generation for 1kqfC . . 141
4.14 Acceptance ratio during Forward decoy generation for 1kqfC, MP weight 0.5 144
4.15 Acceptance ratio during Forward decoy generation for 1kqfC, MP weight 1 145
4.16 Acceptance ratio during Forward decoy generation for 1kqfC, MP weight 10 146
4.17 Acceptance ratio difference map to compare with and without MP 147
4.18 Acceptance ratios at each residue averaged over all moves 148

5.1 Transmembrane helices in template against transmembrane helices in target 159
5.2 Template coverage of N-terminal transmembrane helices 160
5.3 Template coverage of C-terminal transmembrane helices 161
5.4 Relationship between transmembrane helices missing and sequence identity 163
5.5 Examples of decoys of different RMSDs 164
5.6 Rigid and flexible versions of SAINT2-ScafFold 164

5.7 SAINT2-ScafFold performance on completing the final helix from a native structure 165
5.8 Best RMSD for predicting one terminal helix by rigid or flexible . 166
5.9 Best of Top5 RMSD for predicting one terminal helix by rigid or flexible 167
5.10 Top ranked RMSD for predicting one terminal helix by rigid or flexible 168
5.11 Best of Top5 RMSD for predicting two terminal helices by rigid or flexible 169
5.12 Standard deviation of RMSD of decoys when predicted by rigid or flexible 170
5.13 Best of Top5 decoys generated by rigid mode, as ranked by LM1 compared to LM2 171
5.14 Comparison of Best of Top5 between Native and Homology ScafFold 172
5.15 Best of Top5 for rigid and flexible using the Homology set up . . 173
5.16 Best of Top5 ranked by LM1 and LM2 using the rigid Homology set up 174

A.1 Relating angle error to goodness of fit 212
A.2 Measured angle against goodness of fit for two kinks 213
A.3 Standard deviation of the difference between measured and true angle in ideal kinks 213
A.4 Angle measurement error for a range of values of quality of fit . 214

List of Tables

1.1 SAINT2 potentials and weights 32

2.1 Occurrence of proline in each class of aligned helix pairs 52
2.2 Correlation coefficients between angle difference and measures of sequence conservation 54
2.3 Number of helix families of each group size 56

3.1 The number of protein chains displaying potential cotranslational bias 89
3.2 The number of TMH pairs interacting, separated by whether the loop joining them is extracellular or intracellular 97
3.3 The number of correct decoys by each mode of SAINT2 104

4.1 ScafFold targets 117
4.2 Correlation coefficients between the components of the SAINT2 scoring function and TM-score 127
4.3 Coefficients for each component of the SAINT2 scoring function in linear models to predict TM-score 130
4.4 Correlation coefficients between the components of the SAINT2 scoring function and TM-score - decoys generated with membrane potential 137
4.5 Coefficients for each component of the SAINT2 scoring function in linear models to predict TM-score - decoys generated with membrane potential 137
4.6 ScafFold results using ranking by LM1 139

5.1 Dataset for testing the completion of the final one or two helices of a protein 155
5.2 Dataset for testing the completion of the final one or two helices of a protein 157

B.1 Number of aligned helix pairs and occurrence of proline in the soluble and membrane sets 216

B.2 Number of aligned helix pairs and occurrence of proline in the high quality soluble dataset 217

List of Abbreviations

AIC . . . . Akaike information criterion

AP . . . . . Arrest peptide

CAMEO . . Continuous automated model evaluation

CASP . . . Critical assessment of methods of protein structure prediction

CK . . . . . Conserved Kinked

CNS . . . . Crystallography and NMR System

CS . . . . . Conserved Straight

ΔG, $\Delta\Delta$G . Change in free energy, or change in free energy changes

EM . . . . . Electron microscopy

FILM . . . . Folding in lipid membranes

GDT-TS . . Global distance test – total score

GPCR . . . G-protein coupled receptor

kDa . . . . Kilodalton

LCP . . . . Lipidic cubic phase

LM1 . . . . Linear model 1

LM2 . . . . Linear model 2

MCR . . . . Mean central residue

MD . . . . . Molecular dynamics

MD . . . . . Molecular Dynamics

MP . . . . . Membrane potential

MP-T . . . . Membrane protein threader

mRNA . . . Messenger ribonucelic acid

MSA . . . . Multiple sequence alignment

NC . . . . . Not Conserved

NMR . . . . Nuclear magnetic resonance

NOEs . . . Nuclear Overhauser enhancements

OPM . . . . Orientations of proteins in membranes database

PDB . . . . Protein Data Bank

PDBTM . . Protein data bank of transmembrane proteins

PlmDCA . Pseudo-likelihood maximisation of Direct Coupling Analysis

PSI-BLAST Position-specific iterative basic local alignment search tool

PSICOV . . Protein sparse inverse covariance estimation program

RAPDF . . Residue-specific all-atom probability discriminatory function

RMSD . . . Root–mean square deviation

SAINT2 . . Sequential algorithm initiated at the nitrogen terminus 2

SLR . . . . Sum of the log-transformed ratios

TMH . . . . Transmembrane helix

TROSY . . Transverse relaxation optimized spectroscopy

WGS . . . . Whole genome shotgun

1 Introduction

1.1 Protein structure

Proteins are the molecules of life that carry out the majority of roles in biological cells. Their shapes dictate their responsibilities, ranging from transport proteins to control entry of molecules into a cell, to enzymes which catalyse reactions, to proteins which ensure the structural stability of cells and tissues. By understanding the structure of proteins, scientists have gained insights into their functions and mechanisms. It has also been possible to design molecules to modulate these functions. Therefore, the understanding of protein structure is of great importance to biochemistry and medicine.

1.1.1 Primary structure

A protein consists of a specific sequence of amino acids. The amino acids are chemically joined together by the reaction between the amino group and a carboxyl group to produce peptide bonds. This reaction is catalysed by the ribosome, which simultaneously reads mRNA to determine the specific order of amino acids to be synthesised. There are 20 different canonical amino acids encoded by mRNA, all of which are α-amino acids, i.e. they all have one carbon atom (C_α) between the amino and carboxyl groups (Figure 1.1). Figure 1.2 shows

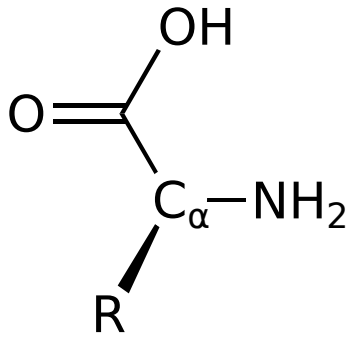


Figure 1.1: A generalised structure of the canonical amino acids, where R represents one of 20 different options (see Figure 1.2).

the chemical structure of the 20 amino acids. The only difference between amino acids is the the third group attached to the C_α, known as the side chain, which makes the C_α a chiral centre (with the exception of glycine where there is no side chain). In all biological organisms, only L-amino acids (as shown in Figure 1.1) are used by the ribosome to construct peptides. The different side chains give rise to different properties, as some are charged, some are uncharged but polar and capable of hydrogen bonding, and some are hydrophobic.

1.1.2 Secondary structure

In each amino acid residue unit, there are three bonds, shown in Figure 1.3. Torsion angles ϕ, ψ, and ω are used to measure the positions of these bonds. Every peptide bond is held in a planar shape by an extended π-bonding system, therefore the ω torsion angle displays very little variation between residues. The only rotatable bonds are those at the C_α atom, labelled by the ϕ and ψ angles, which can be displayed on a Ramachandran plot (Figure 1.4).

Steric clashes between groups forbid some areas of the plot, but favourable interactions make other regions more common. In the region marked "alpha-helix", this conformation of the peptide backbone, when repeated over several residues, allows hydrogen bonds to form between the N–H of the ith residue and the C=O of the $(i-4)$th residue (Figure 1.5). Slightly different angles give rise to tighter 3_{10} helices or looser π helices . In the region marked "beta sheet", this conformation leads to an extended chain, where two such chains, adjacent

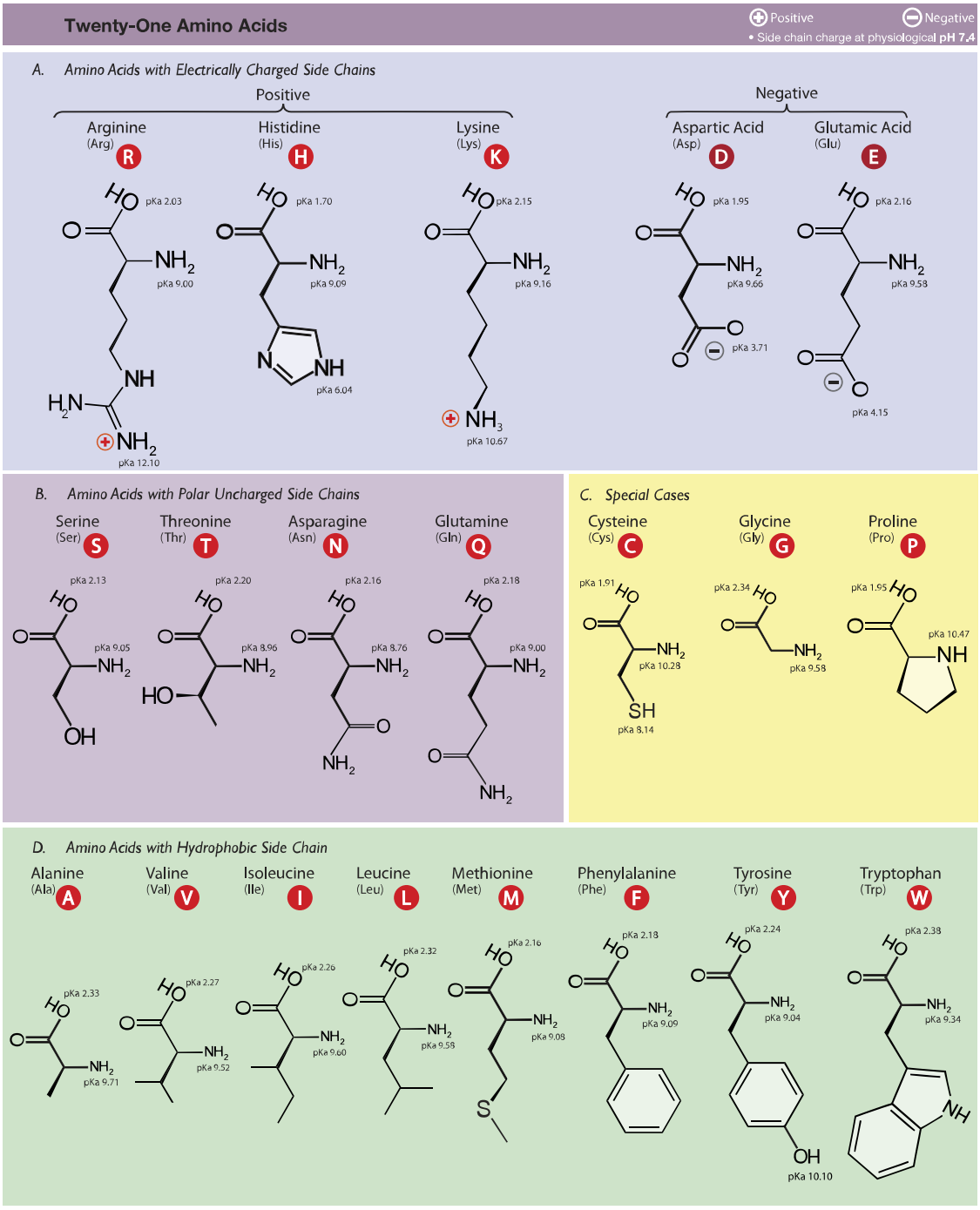


Figure 1.2: The chemical structures of the 20 amino acids encoded by mRNA to be covalently joined by the ribosome, grouped according to their characteristics. Adapted from an image by Dancojocari under Creative Commons (CC BY-SA 3.0 - https://creativecommons.org/licenses/by-sa/3.0/), via Wikimedia Commons

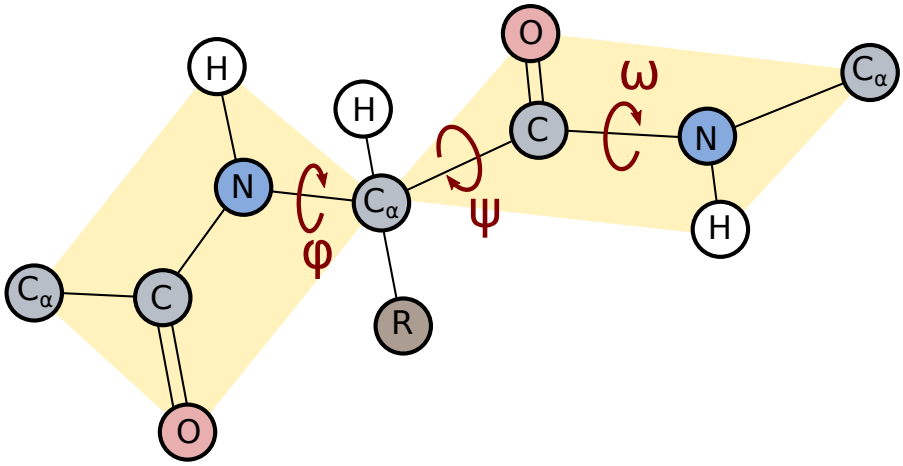


Figure 1.3: Backbone torsion angles. The approximately planar peptide bonds are shown in yellow.

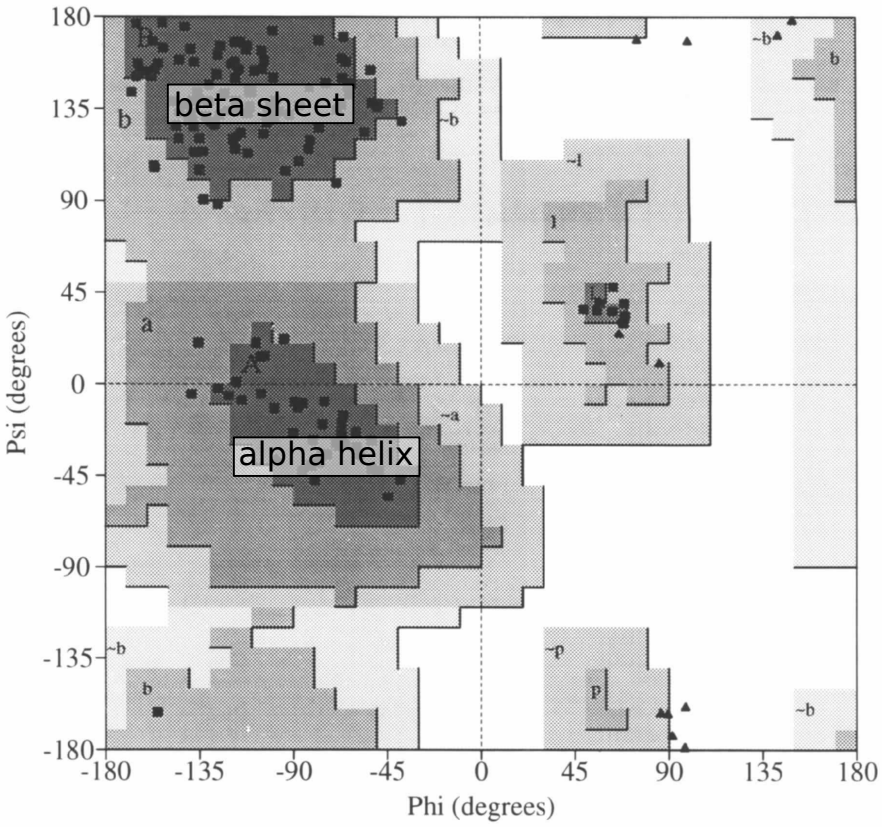


Figure 1.4: Ramachandran plot, adapted from Laskowski *et al.* (1993). Reproduced with permission of the International Union of Crystallography.

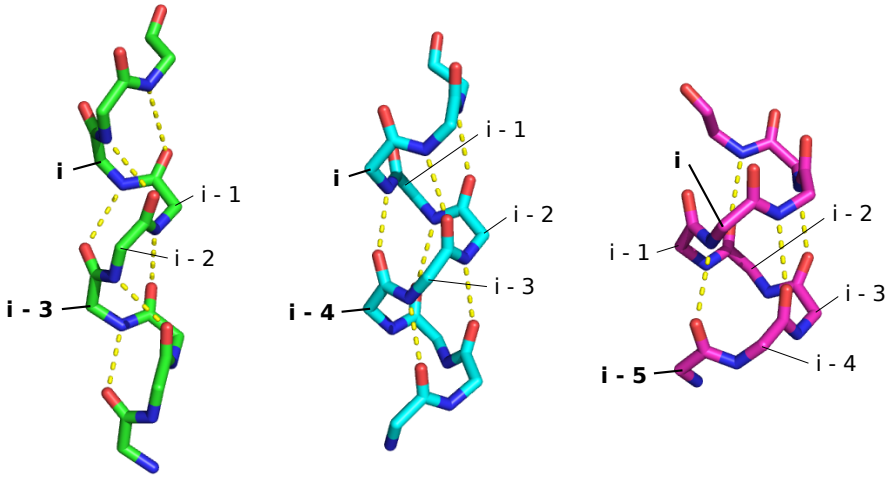


Figure 1.5: From left to right: 3_{10}-helix, α-helix, π-helix. Only backbone atoms are shown, each C_α labelled with the relative residue numbering. Hydrogen bonds are shown as dashed yellow lines.

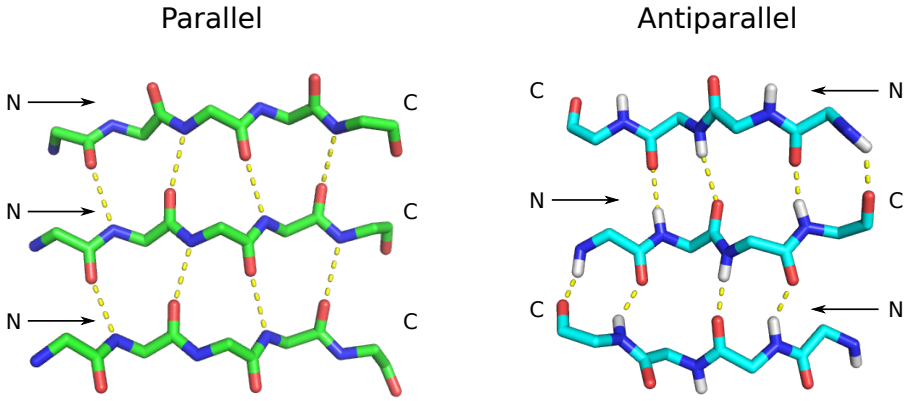


Figure 1.6: Parallel and antiparallel β-sheets. Hydrogen bonds are shown as dashed yellow lines, and the direction of each strand is shown, from N- to C-terminus.

to each other in either parallel or antiparallel orientation, can form regular hydrogen bonds between the chains (Figure 1.6). These regular patterns, called secondary structures, are the most common ways to satisfy the ability of the polar backbone groups to form hydrogen bonds. Residues in the structure which do not fall into either category are commonly referred to as loop or coil residues.

Computational methods can be used to annotate and analyse secondary structure in an unbiased way. Programs such as DSSP (Kabsch and Sander, 1983) use standard definitions for the hydrogen bonding patterns of α-helical, β-sheet or other features. DSSP is based on calculation of the theoretical energy

of the hydrogen bond interaction between potential hydrogen bonding partners, annotating any less than –0.5 kcal/mol. These methods are a useful starting point for analysis of helix geometry, by locating stretches of helical residues.

1.1.3 Tertiary structure

The secondary structures into which a peptide chain folds can be arranged against each other in very many ways. The three-dimensional arrangement of secondary structures and topology of connecting loops between them is referred to as the tertiary structure. A single polypeptide chain can often be divided into separate structural domains, which are lengths of the peptide that are able to fold independently, each having their own tertiary structure.

1.1.4 Quaternary structure

Many proteins consist of more than one peptide chain. Often, individual protein chains are able to fold independently to form domains, which then interact to form a complex. In other cases, complete folding of individual components is not possible without another part of the complex. The tertiary structures of the chains, which may be identical or different, are arranged in a specific orientation with respect to each other in order to carry out the function of the protein.

In this thesis, I focus on folding at the level of individual chains, and therefore tertiary structure. In the case of soluble proteins, it is possible to find many good examples of solved structures that consist of only one monomeric (single chain) protein. This makes it possible to focus on protein folding at the level of tertiary structure without complication from intermolecular effects. In membrane proteins (described in detail in Section 1.3), fewer are monomeric. In order to study tertiary folding in this thesis, I also use single peptide chains which may originally come from larger complexes. It is likely that most folding of a membrane protein can take place without the presence of a binding partner or another identical subunit, as the individual chains are synthesised separately and will not immediately be localised together in the membrane. However, it

should be considered that the structure adopted by a monomer of a complex in the membrane may be different from that in the final complex, and therefore it may be more difficult to predict without the additional inter-chain interactions that stabilise the native structure.

1.1.5 Experimental determination of protein structure

The Protein Data Bank (PDB) (Berman *et al.*, 2000) contains all published structures that have been experimentally determined. X-ray crystallography is the oldest and most common structure solution method, dating back to the solution of the structure of myoglobin (Kendrew *et al.*, 1958). The primary bottleneck of the technique is the challenge of obtaining large enough crystals. The main global measures of quality of an X-ray crystal structure are the resolution, dictating the scale of the smallest details which can be separated, and the R-factor, or R_{free}, which indicates the difference between the diffraction data observed and the diffraction data predicted by the model.

The second common structure solution method is NMR spectroscopy, and it has the advantage that crystals are not required. NMR is restricted to smaller proteins because tumbling rates are slower for larger proteins, and this leads to faster relaxation and increased linewidths. If a sample of protein can be prepared with ^{15}N and ^{13}C labelling, peaks can be assigned for backbone atoms. Spatial constraints can be obtained by observing nuclear Overhauser enhancements (NOEs) between nearby ^{1}H atoms, and these can be input into programs designed to sample conformations to satisfy these constraints, resulting in a model. More advanced techniques of energy input, for example TROSY, are used for larger proteins up to $\sim$ 50 kDa to improve sensitivity. Model quality metrics for NMR structures are not as well established as for crystallography, but the number of experimental restraint violations can be checked. Theoretical chemical shift data can be calculated from a given model but these are not routinely used for structure validation (Rosato *et al.*, 2013).

Increasingly, cryo-electron microscopy (cryo-EM) of single particles is becoming a popular option for larger proteins and complexes. Thousands of images of single particles are captured, oriented and averaged so that they can be combined together to produce an electron density map. Technological advances have improved the resolution to a range comparable to X-ray crystallography (< 4 Å), with a realistic number of particles (Li *et al.*, 2013).

Methods of structure determination can also give some indication of the flexibility of structures. In a population of images from cryo-EM, there may be many conformations of a structure, and if these can be clustered together, it is possible to see many different states of a protein and learn about its dynamics. In crystallography, multiple conformations with partial occupancy can be fitted into the electron density, and B-factors give an indication of where electron density is spread due to thermal fluctuations of an atom's position. Conformational changes of proteins can be studied by crystallising them with different ligands or conditions to stabilise alternative conformations. Homologues crystallised in different conformations have also been used to learn about possible alternative conformations of a protein (Narunsky *et al.*, 2015). I used a similar strategy to observe conformational variation in helices, which is the basis of Chapter 2. However, the techniques of structure determination do not usually allow the protein to be in near-native conditions. Therefore structural understanding is combined with biophysical techniques and molecular dynamics simulations in order to learn about native conformational flexibility and function of a given protein.

1.2 Protein folding

It is believed that all of the information needed to encode the three-dimensional structure of a protein is contained in its primary structure (amino acid sequence). This is due to their ability to fold *in vivo* or in cell-free systems with no assistance, and the ability of some proteins to refold from a denatured state *in vitro* (Anfinsen *et al.*, 1961). The link between sequence and structure had been

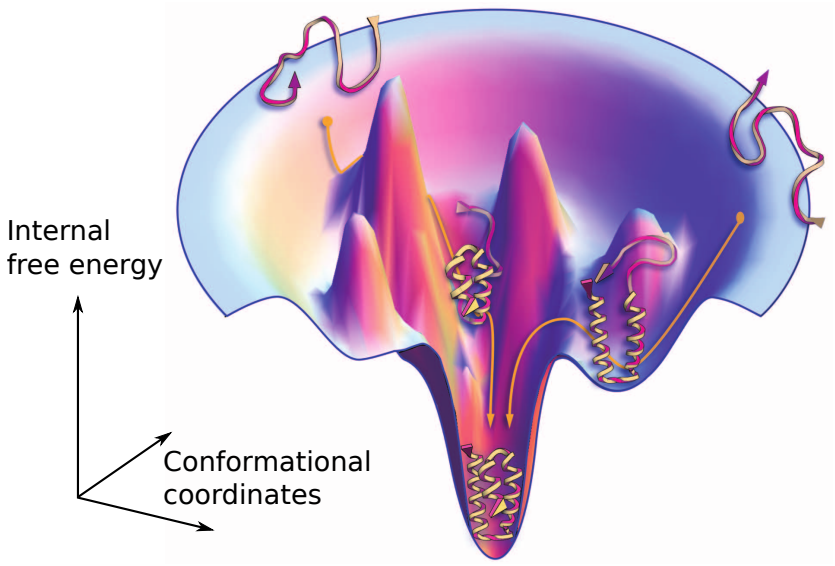


Figure 1.7: A protein folding landscape, adapted from Dill and MacCallum (2012). Reprinted with permission from AAAS.

associated with the belief that the native state of a protein is its thermodynamic global energy minimum, however an exhaustive search of all conformations is not realistically possible for a protein. The Levinthal paradox states that a single protein of 150 residues could never reach its native state in the observed folding times of a few seconds through a systematic search, assuming each rotatable bond can adopt only one of three conformations (Levinthal, 1969). Therefore it was proposed that most proteins follow a folding pathway which leads to a local energy minimum, which is the native state, in which the protein is kinetically trapped. In support of this, many proteins lose their activity over time, and others can adopt lower energy misfolded states, for example amyloids.

Though the native state of a protein does not need to be the global energy minimum, this does not imply that there is only one well defined folding pathway by which it is reached. A better description of a polypeptide's journey through conformational space is perhaps one of many possible routes over a high-dimensional folding landscape or funnel (Dill and Chan, 1997). Most of the process will consist of changes towards a lower energy state, but thermal energy allows "uphill" conformational changes also to take place (Figure 1.7). This is essential where the energy surface is rugged, as favourable interactions

formed in one intermediate must be broken in order to reach a more stable and native-like state.

The folding landscape *in vivo* may not be the same as *in vitro*, leading to alternative folding pathways. This enables more efficient folding in a cell, where folding occurs at a much faster rate than it does after chemical denaturation (Fedorov and Baldwin, 1999). In cells, proteins are also prevented from aggregating by chaperones that can bind to exposed hydrophobic surfaces on unfolded or partially folded proteins (Hartl and Hayer-Hartl, 2002).

1.2.1 Evidence for cotranslational folding

The other factor that could contribute to more efficient folding in cells is that protein chains do not suddenly appear in a cell in a fully extended conformation. More restricted folding pathways are available to the nascent (partially synthesised) chain as it grows during the process of translation, which may cause it to fold more efficiently. In *E. coli*, the rate of translation (around 50 ms per codon) is slower than the rate at which secondary structure can start to form (low ms scale) (Ellis *et al.*, 2010). The rate of translation in humans is even slower (Ingolia *et al.*, 2011). The difference between these rates is large enough to suggest that the nascent chain will be starting to fold before the whole protein is synthesised.

There is a large amount of experimental evidence supporting the relevance of cotranslational folding. A construct of three fluorescent protein half-domains was used to demonstrate the effect of cotranslational folding on the product produced (Sander *et al.*, 2014). The central domain could fold to produce a complete and functional fluorescent protein with either the N-terminal domain to emit yellow light, or the C-terminal domain to emit cyan light. When refolding after denaturation *in vitro*, the N- and C-terminal domains compete to produce an equal concentration of cyan and yellow fluorescent protein. However, when expressed in *E. coli*, the concentration of yellow fluorescent protein is approximately double that of cyan. The bias in favour of the central domain

folding with the N-terminal domain was increased with the addition of rare codons near the start of the third domain. This experiment shows that folding of entire domains and translation of those domains are occurring on a very similar timescale, resulting in the product observed being significantly affected by changes in translation speed.

As synonymous mutations have been shown to affect protein folding (e.g. Sander *et al.*, 2014; Buhr *et al.*, 2016), frequencies of the different redundant codons which encode the same amino acid have also been studied in detail. Measures have been developed to attempt to link this to the speed of translation in various different organisms (e.g. Reis *et al.*, 2004; Saunders and Deane, 2010b). The structure of mRNA has also been investigated, as the energy of its unfolding may also affect translation rates (Faure *et al.*, 2016). Differences in "speed" measures have been observed between secreted proteins and membrane proteins, suggesting that membrane proteins could have more time during translation in which to fold (Mahlab and Linial, 2014).

NMR has been used to probe cotranslational folding, investigating how structures of constructs which are missing their C-terminus resemble the final structure (Waudby *et al.*, 2013). When a population of ribosomes is stalled at the same point, the partially folded nascent chain shows the majority of the same cross-peaks that are formed in the final native structure. There are also examples of proteins which display enzymatic activity (Nicola *et al.*, 1999) and heme binding (Komar *et al.*, 1993) during the process of translation.

High-resolution structures of prokaryotic (Ban *et al.*, 2000) and eukaryotic (Ben-Shem *et al.*, 2010) ribosomes are available, which inform us better about the environment in which the protein folds. The tunnel of the ribosome from which the nascent chain emerges is 10–20 Å wide, so is able to accommodate an alpha-helix, which may allow a chain to form secondary structure before interacting with the rest of the extruded peptide.

Many simulations of cotranslational folding have also been carried out, using various coarse-grained approaches and simplifications of the system to reach

the timescales required (Trovato and O'Brien, 2016). Simulations have also been carried out on an atomic scale for a nascent membrane protein folding within the Sec translocon in which it is believed to fold (Ulmschneider *et al.*, 2015).

There is also evidence of cotranslational folding from bioinformatics, in the asymmetry of protein structures when the N-terminus is compared to the C-terminus. Protein structure prediction was more accurate for the 20 residues at the N-terminus than the 20 residues at the C-terminus for 79% of 493 predictions (Saunders and Deane, 2010a). It has also been observed that residues closer to the N-terminus are found near the centre of proteins of the α/β class and the distribution of contacts along a chain is asymmetric when comparing the N- and C-termini (Deane *et al.*, 2007).

1.3 Membrane Proteins

Much of this thesis focuses on one class of proteins in particular: membrane proteins. The differences between membrane proteins and soluble proteins arise from the way membrane proteins interact with their surroundings.

1.3.1 Biological membranes

Biological membranes surround every cell and every organelle within eukaryotic cells. They are a barrier to most proteins and polar molecules, and maintain the controlled environment within each compartment.

The biophysical properties of a membrane arise from its lipid composition. The lipids that make up a membrane have a charged or polar hydrophilic group, and a hydrophobic tail. The hydrophobic effect drives their assembly into a two-leaflet arrangement called a lipid bilayer, with the polar groups exposed to water and lipid tails buried within the membrane (Figure 1.8). It is the hydrophobic environment in the centre that prevents free movement of many substances across the membrane. Movement of ions and large polar molecules into the centre of the membrane requires highly unfavourable desolvation, therefore diffusion of either across the membrane is impossible or very slow.

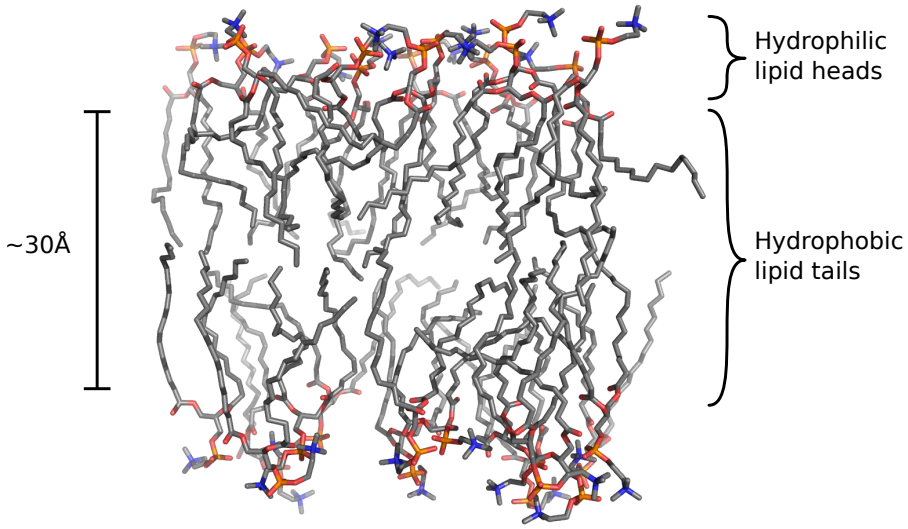


Figure 1.8: A phosphatidylcholine bilayer in stick representation and coloured by atom type in PyMOL (Schrödinger, 2015). The typical dimensions and position of different chemical groups have been calculated using X-ray and neutron diffraction (McCaughan and Krimm, 1982).

There is a large amount of variation between and within different membranes. Different membranes contain different lipids, for example phosphatidylserine, cardiolipin and cholesterol. Membranes display asymmetry in the form of differences between the two leaflets (van Meer *et al.*, 2008). Some membranes contain lipid rafts and areas of different lipid composition even within the same leaflet, and some membranes have a very high protein content.

However, the variation within and between different membranes is much less significant than the difference between a membrane and an aqueous environment. In order to learn about this greater difference, it is a reasonable first approximation to treat all membranes as similar. For some purposes it is also important to consider that membranes are asymmetric and that almost all proteins are only present in one orientation.

1.3.2 Membrane protein structure

For a protein, a membrane gives rise to a unique anisotropic environment in which to exist. Many proteins must span biological membranes to carry out their function. These functions include passive and active transport of substances

across the membrane, and transmission of signals across the membrane. Such integral membrane proteins are particularly important for medicine, accounting for over half of all drug targets (Overington *et al.*, 2006).

1.3.2.1 Experimental determination

Despite their importance, our knowledge of membrane protein structure lags behind that of soluble proteins. Membrane proteins are much more difficult to express in the quantities required for X-ray crystallography or NMR than soluble proteins. Conventional crystallography is very challenging, as membrane proteins are usually flexible. In addition, much of the protein surface exposes hydrophobic residues which would interact with lipids in the native membrane environment. To make membrane proteins soluble in water, detergents are required to mask the hydrophobic areas with charged groups. For a given protein, the best combination of different detergents is difficult to determine. Once solubilised, crystallisation is much more difficult because the surface area for forming crystal contacts is reduced by covering part of the surface with detergent. These solubilised membrane proteins are not usually suitable for study by NMR, as they are typically very large. Instead, magic angle spinning can be used with solid state samples to obtain structures (McDermott, 2009).

Recent developments are improving our ability to solve membrane protein structures, particularly by X-ray crystallography (Kang *et al.*, 2013). The transmembrane region of membrane proteins evolves at a slower rate than soluble proteins (Oberai *et al.*, 2009), so it has proved far easier to find point mutations which increase thermostability. Achieving greater thermostability increases the chance of obtaining crystals, and has been particularly successful in the case of G-protein coupled receptors (GPCRs) (e.g. Warne *et al.*, 2008; Tate, 2012). Another way to increase crystal stability is to allow the possibility of crystal contacts between the hydrophobic parts of membrane proteins, making crystals more compact. This requires a higher lipid content in the crystal, which can be achieved by growing them in mesophases such as the lipidic cubic phase (LCP)

(Caffrey *et al.*, 2012). LCP can be formed by increasing the ratio of lipids to water, therefore it is also an environment that better imitates the native membrane. Even with these advanced methods, it is often still impossible to obtain large and stable crystals. In these cases, femtosecond crystallography can be used on nanocrystals suspended in solution. This technique can even be carried out at room temperature as it eliminates the problem of radiation damage by capturing data before crystals are destroyed. Using X-ray free electron lasers, femtosecond crystallography has been used to study the mechanism of photosystem II (Kupitz *et al.*, 2014). Even if no crystals can be obtained, cryo-EM can be used effectively with membrane proteins and complexes, especially due to their larger size.

The membrane proteins of known 3D structure database (http://blanco.biomol.uci.edu/mpstruc/) is a manually curated database that currently contains 2312 structures of 720 unique membrane proteins. This can be contrasted with nearly 52,000 total structures in the PDB, non-redundant at 95% sequence identity (databases accessed 25th September 2017).

1.3.2.2 Types of membrane protein

Integral transmembrane membrane proteins fall into two groups: alpha-helical membrane proteins and beta-barrels. In the hydrophobic lipid tail environment of the membrane, an alpha-helix is the ideal way to stabilise the polar backbone, and expose only side chains. All backbone hydrogen bond donors and acceptors are satisfied (see Figure 1.5). Alpha-helical membrane proteins consist of one or more transmembrane helices that span the full width of the membrane. Occasionally, helices go only part of the distance across and the chain returns to the same side (Viklund *et al.*, 2006).

Beta-barrels are not as diverse and are only found in the outer membranes of bacteria, mitochondria and chloroplasts. In beta-barrels, the alternating orientation of side chains in a beta-sheet allows hydrophobic residues to point towards the lipid tails and polar residues to point inward. This creates a

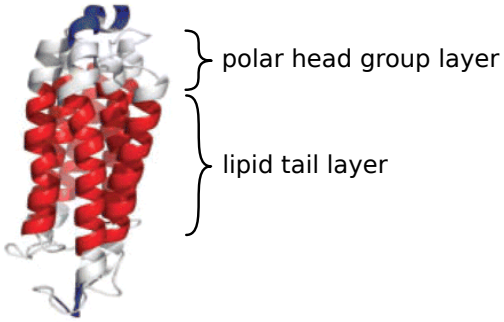


Figure 1.9: An alpha-helical membrane protein structure coloured by the membrane layer each residue is situated in, adapted from Kelm *et al.* (2009). Residues coloured dark blue are in the non-membrane, aqueous regions either side of the membrane.

hydrophilic channel which allows the passage of many types of molecule across the membrane.

1.3.2.3 Residue propensities for membrane layers

The membrane can be divided into a lipid tail layer at the centre, with a polar head group layer either side. Figure 1.9 shows an alpha-helical membrane protein, coloured according to which layer of the membrane each residue sits in. The energetic preference for amino acids to be found in different membrane layers can be calculated from the partitioning of peptides between aqueous and organic phases (Wimley *et al.*, 1996). Alternatively, propensities for amino acids to be in a given membrane layer can be calculated from solved structures (Ulmschneider *et al.*, 2005). Energetic changes of transition to the membrane phase have also been determined for each amino acid by experiments to measure the proportion of helices inserted (Hessa *et al.*, 2005). Insertion in this case was measured by glycosylation, which only takes place on the ER lumen side. These three methods are in good agreement and all show that hydrophobic residues are commonly found in the lipid tail layer. Aromatic residues are most commonly found at the interface between the lipid tail and polar head group layers. Based on these patterns, methods are available for embedding a crystal structure of a membrane protein into its likely orientation in a membrane. Some of these methods are based on minimising the calculated energy of the system by

rotating and translating the position of the protein relative to the membrane (Nugent and Jones, 2013; Lomize *et al.*, 2006; Kozma *et al.*, 2013). These are described in more detail in Chapter 3.

Others have used coarse-grained molecular dynamics simulations to determine the stable position of the protein in a bilayer and also any distortive effect that the protein has on the membrane (Sansom *et al.*, 2008; Stansfeld *et al.*, 2015). Based on the results of these insertions, iMembrane searches for homologues in these databases in order to transfer the annotation to a protein of unknown embedding by structural alignment (Kelm *et al.*, 2009).

1.3.3 Adding information to structurally characterised membrane proteins

The outcome of these embedding processes is an important input for other software which may perform relaxation of structures under a specific scoring system, docking of small molecules to membrane proteins, calculation of $\Delta\Delta G$ for specific mutations, and other predictions. In this thesis, I use the RosettaMP framework (Alford *et al.*, 2015; Leman *et al.*, 2017), which is able to perform some of these functions. Instead of using complete structures, the all-atom relaxation protocol of this framework is used in Chapter 3 to investigate the properties of partial segments of membrane proteins rather than entire structures.

1.3.4 Alpha bundles

In this thesis, I focus exclusively on alpha-helical integral membrane proteins, which are the majority of membrane proteins. These can be single-pass (see Type I and Type II in Figure 1.10), or multipass (polytopic in Figure 1.10). The number of spans and orientation in the membrane are referred to as the topology. Alpha-helical membrane proteins are known to be inserted into the membrane cotranslationally through a complex called the translocon. This begins the process of cotranslational folding, and is described in detail in Chapter 3.

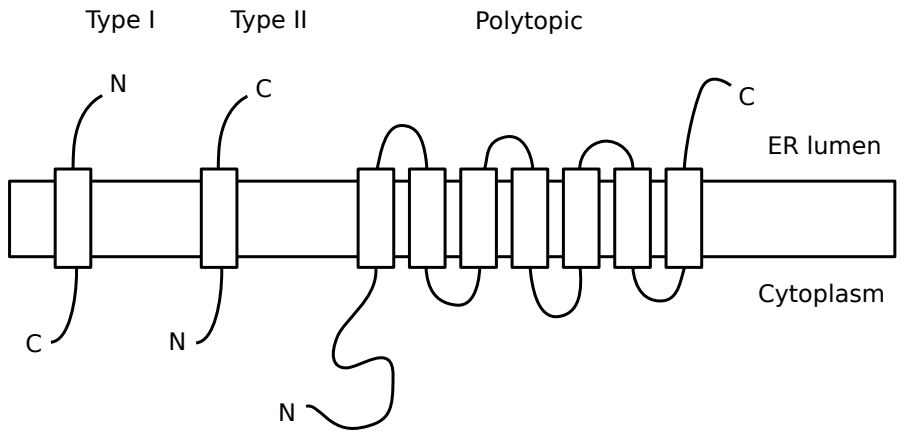


Figure 1.10: Topology of alpha-helical membrane proteins, with hydrophobic membrane spanning helices shown in blue. Single and multi-pass membrane proteins are oriented in a specific way, either with the N-terminus on the side of the cytoplasm or the side of the ER lumen (or extracellular/periplasmic side in prokaryotes).

1.3.4.1 Alpha-helical kinks and their identification

Alpha-helices in membrane proteins tend to be longer than helices in soluble proteins (Wilman *et al.*, 2014b), as they usually span the full width of the membrane. Over such lengths, helices are often not completely regular and they may display curving or abrupt changes of direction. The latter is usually associated with disruption of the backbone hydrogen bonding found in alpha-helices. These points in helices have been referred to as bends, kinks, and distortions, and several methods of identification of kinks are described in Chapter 2.

Kinks are thought to be important for structural flexibility, which allows conformational changes. A range of types of molecular dynamics simulations have shown kinks to be flexible both on their own in a membrane (Tieleman *et al.*, 2001; Hall *et al.*, 2009), and in ion channels (Choe and Grabe, 2009). Straighter and more kinked conformations have also been associated with different functional states in GPCRs (Bettinelli *et al.*, 2011; Deupi, 2012; Katritch *et al.*, 2013; Tehan *et al.*, 2014). This functional significance makes them an important motif to study, and in Chapter 2 I analyse their conservation between homologous protein structures to better understand their dynamics and evolution.

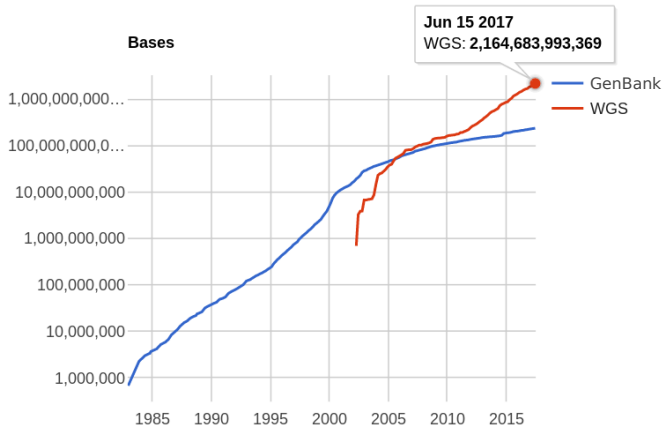


Figure 1.11: Growth of sequence data, represented by the number of bases in the GenBank and WGS (Whole Genome Shotgun) databases. Figure reproduced from https://www.ncbi.nlm.nih.gov/genbank/statistics/ (Benson *et al.*, 2013)

1.4 Membrane protein structure prediction

Given that protein structures are valuable for our understanding of their function and for drug design, efforts have been made to experimentally determine structures for many targets (e.g. Montelione, 2012). Despite structure solution methods improving and becoming more automated, the PDB is no longer growing exponentially (Berman *et al.*, 2012). The techniques described in Section 1.3.2.1 are particularly expensive and time consuming for membrane proteins and because of this they make up only 1-2% of the PDB (Koehler Leman *et al.*, 2015). Meanwhile, techniques for genome sequencing continue to accelerate (see Figure 1.11) and therefore a wealth of sequence data now dwarfs the structural data available. As discussed in Section 1.2, it is believed that all of the information needed to encode the structure of a protein is contained in its sequence. Entirely physics-based molecular dynamics simulations have generated correct structures for some fast-folding proteins (Lindorff-Larsen *et al.*, 2011), but simulations on these timescales of up to milliseconds can only be carried out on specialised hardware, and are difficult to scale up to longer proteins and longer folding times. Due to this, most ways to bridge this gap rely on knowledge-based approaches using the available structures. Therefore, protein structure prediction is particularly difficult for membrane

proteins as there are fewer structures to learn from. For the same reason, it is especially important because there are so many membrane proteins with no experimentally solved structure.

1.4.1 Transmembrane helix prediction

The membrane places an extra constraint on the conformations which need to be considered. The starting point for prediction of alpha-helical membrane protein structures is to accurately predict the locations of transmembrane helices (TMHs). The earliest methods of TMH prediction used a sliding window and hydrophobicity scales for amino acids described in Section 1.3.2.3 to predict regions which would favourably partition into the membrane.

To improve on the sliding window methods, other prediction programs use hidden Markov models (TMHMM, Krogh *et al.*, 2001; TMMOD, Kahsay *et al.*, 2005), neural networks (OCTOPUS, Viklund and Elofsson, 2008) and support vector machines (Memsat-SVM, Nugent and Jones, 2009). TOPCONS is a consensus approach which combines the results of several other predictors (Tsirigos *et al.*, 2015). One secondary structure prediction method, BCL::Jufo9D, combines secondary structure prediction with TMH prediction (Leman *et al.*, 2013). These methods are $\sim$ 95% accurate in predicting the correct number of transmembrane spans for transmembrane proteins.

After obtaining predictions of transmembrane spans, the next stage is prediction of their interactions within a helical bundle, to be followed by loop prediction for the residues joining them. 75% of membrane helix interactions can be clustered into just five helix packing configurations (Walters and DeGrado, 2006), suggesting that prediction of such interactions from sequence may be a good starting point. However, the complexity of interactions in membrane proteins is becoming clearer as more structures are solved (Li *et al.*, 2012; Zhang *et al.*, 2015). Currently, the best methods of structure prediction are similar to soluble protein structure prediction, but with some adaptations. These fall under two categories: template-based and template-free (or *de novo*) methods. Before

explaining these approaches, I describe the common methods of assessing the three-dimensional models generated.

1.4.2 Assessing the accuracy of three-dimensional models

In order to evaluate the accuracy of a method of predicting protein structures, soluble or membrane, software is typically trained on a set of known structures extracted from the PDB. For a more objective test, a large scale assessment and comparison of methods takes place every two years in the critical assessment of methods of protein structure prediction (CASP) (Moult *et al.*, 2011, 2014, 2016). CASP tests the ability of current methods to predict structures which have not yet been released, later comparing models to the true structures. This experiment ensures that the information available is representative of that for predicting any other protein whose structure is not yet solved.

A number of measures have been used to score the accuracy of a model against the experimental native structure. The simplest measure possible is a root mean squared deviation (RMSD) between all atoms in the native structure and their equivalent in the model. A superposition can be found to minimise the RMSD between a given set of pairs of equivalent atoms in two structures. RMSD can work well when the structures to be compared are already close, however it is dependent on the length of the structures to be compared (Betancourt and Skolnick, 2001).

TM-score is a measure of similarity based on RMSD, but normalised according to the length of the proteins, and then transformed to lie in the range (0,1] (a higher score corresponds to a better model):

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \tag{1.1}$$

where L_N is the number of residues in the native structure, L_T is the number of residues aligned to the template structure, d_i is the distance between the ith pair of aligned residues and d_0 is a length-dependent scaling factor to

normalise the match difference (Zhang and Skolnick, 2004). "Max" represents the function to return the highest possible value after optimising the spatial superposition. TM-align is a protocol to align two structures which may be of different lengths, where a resulting TM-score of > 0.5 can be interpreted to mean that the structures have the same fold (Zhang and Skolnick, 2005).

GDT-TS (Zemla, 2003) is a model quality score which combines together the similarity of two structures at different levels of accuracy. At each threshold of 1, 2, 4, and 8 Å, the superimposition which can bring the greatest number of C_α atom pairs within the threshold is found. The maximum number of residues within each threshold distance cut-off is summed and divided by the highest possible score (four times the protein length), to give a percentage. GDT-TS is dependent on the protein length therefore it does not provide an absolute measure of success for that target, but is useful for determining the best of a set of models for a given target.

While GDT-TS is a good all-round measure of model accuracy, recent CASP experiments have evaluated models based on a combination of scores, including amongst others a contact based score (Shi *et al.*, 2009), a consensus of ten scores (TenS, Kinch *et al.*, 2011), a score based on relative positioning of secondary structures (QCS, Cong *et al.*, 2011), and an alignment free local distance difference test (lDDT, Mariani *et al.*, 2013). The combination of scores helps to avoid the incentive to overtrain for a specific evaluation metric.

While this method of combining scores is important for large-scale prediction assessment, reports of protein structure prediction protocols frequently provide only one measure of model accuracy. The measure chosen is often TM-score (e.g. Michel *et al.*, 2014; Kosciolek and Jones, 2014; Jones *et al.*, 2015; Ovchinnikov *et al.*, 2017), and it is this score which I use for the majority of this thesis. In Chapter 5, I use RMSD, as this is commonly used for small parts of structures (Choi and Deane, 2010), and in my case the parts of the protein to be modelled are of similar size to each other.

1.4.3 Template-based structure prediction

The most accurate models of a target are obtained when there is a structure of a homologous protein in the PDB which can be identified and used as a template. BLAST was an early program for identifying homologues from a database, and is still a very fast way to find possible templates (Altschul *et al.*, 1990). Its speed is due to the use of short strings of amino acids to identify possible matches, therefore performing a local rather than global alignment. An "E-value" is calculated for each hit to indicate the expected number of matches in the database that are as significant as the hit.

PSI-BLAST was created as an extension of BLAST, where after searching the database once, the homologues detected below a specified E-value are combined to produce a sequence profile. Position-specific substitution scores are then used to search the database again (Altschul *et al.*, 1997). A specified number of iterations can be carried out, or the search ends when an iteration adds no new sequences to the profile. This method detects a greater number of homologues than BLAST as it is able to detect more remote homologous sequences. Currently, methods based on alignment of hidden Markov models (HMMs) are the leading methods in terms of sensitivity (e.g. Remmert *et al.*, 2011; Eddy, 2011). In this thesis I use PSI-BLAST to identify templates, as I do not require detection of remote homologues.

The two main considerations when choosing between the templates output by the above methods are the sequence similarity between the target and template, and the percentage of target residues which are covered by the template.

1.4.3.1 Sequence alignment

After choosing a suitable template, the sequences of the target and template may be accurately re-aligned in order to generate a high quality homology model. This is particularly a challenge in cases where the sequence identity is low.

In this thesis, I use MP-T, which is specifically designed for alignment of membrane proteins (Hill and Deane, 2013). MP-T extracts homologous

sequences for the template and target using PSI-BLAST, and uses up to 125 non-redundant sequences to build a multiple sequence alignment including the target and template. The template structure is embedded by iMembrane (Kelm *et al.*, 2009). Residues of the template structure are annotated according to their position in the membrane layer and contact to lipid tails or heads. Secondary structure is annotated by JOY (Mizuguchi *et al.*, 1998), and environment-specific substitution tables for each combination of annotations (Hill *et al.*, 2011) are used to optimally construct a multiple sequence alignment and thus align the template and target.

1.4.3.2 Coordinate generation

Medeller is a homology modelling method designed specifically for membrane proteins, and also uses the iMembrane annotation (Kelm *et al.*, 2010). The core region of a membrane protein that sits in the lipid tail membrane layer is better conserved between homologues. Therefore, coordinates from the template can be copied to the target, leading to an average RMSD of 1.97 Å in this region. Medeller then uses the knowledge-based loop modelling protocol FREAD (Choi and Deane, 2010) to complete regions that are not well conserved between the template and target. The programs iMembrane, MP-T and Medeller together make up the modelling pipeline Memoir, which is used in Chapter 5 to generate homology models.

Some methods also remodel the core section of a template structure to better fit the target sequence. These include a modelling approach that specifically aims to allow for changes in kink structure between template and target (Werner and Church, 2013). A library of possible kink structures from solved alpha-helical membrane proteins was sampled for each kink in the structure, and replacements were accepted if they did not cause steric clashes. Energy minimisation was run on each of many generated decoys, and they were scored so that the best model could be chosen. A different modelling pipeline first inspected multiple sequence alignments to determine whether helix structures were likely to have

changed from the template (Chen *et al.*, 2014). The core was remodelled if there were gaps in the alignment, prolines not aligned or missing in either template or target, or transmembrane helices predicted to be of different lengths. Sampling was carried out via a Monte Carlo method, allowing changes of angle at possible bend sites and constrained movement of helices relative to each other. This method reduced the RMSD in the transmembrane span regions when compared to Medeller, modelling some kink sites with < 1 Å RMSD.

Other available homology modelling programs, for example Modeller (Šali and Blundell, 1993), are not specific to membrane proteins but can be used to model them (Forrest *et al.*, 2006).

1.4.4 De novo structure prediction

If no template can be found for a target, *de novo* methods must be used to generate a prediction. Early *de novo* methods used a form of fragment-based structure prediction (e.g. Simons *et al.*, 1997), and this concept is still used by the most successful programs (e.g. Ovchinnikov *et al.*, 2016; Yang *et al.*, 2015; Xu and Zhang, 2012; Kinch *et al.*, 2016). These methods carry out a coarse-grained conformational search by combining fragments of many structures from the PDB. A fragment library is constructed so that there are a range of backbone conformations to choose from at each location along the protein chain. Each fragment is stored as torsion angles for a set of consecutive residues. Starting from an extended chain, random fragment substitutions are proposed, moving the backbone by replacing the torsion angles and recalculating coordinates of each atom. The new conformation is scored, and according to the sampling protocol, it may be rejected or accepted depending on the difference in score to the previous conformation and the temperature. The scoring functions usually involve some physics-based potentials, for example the Lennard-Jones potential, and a number of statistical potentials. The latter may include a residue-specific all-atom probability discriminatory function (RAPDF), which assigns an energy to every interaction between two atoms a measured distance apart (Samudrala

and Moult, 1998). This score is calculated based on the frequency of observations in PDB structures of those atoms in that distance bin compared to those atoms in all other distance bins. After thousands of moves, usually the lowest scoring conformation is saved, and the structure generated is called a decoy.

As there are so many possible combinations of fragments and parts of fragments, thousands of decoys are generated for a given target. Decoy structures are often converted to an all-atom representation with side chains and relaxed under a different scoring function from that used during decoy generation (e.g. Rohl *et al.*, 2004). Clustering methods can be used in order to choose popular but dissimilar possible models from the pool of decoys (e.g. Rohl *et al.*, 2004).

Some fragment-based approaches have been designed specifically for membrane proteins. One early method was RosettaMembrane (Yarov-Yarovoy *et al.*, 2006), which is a version of the standard soluble protein Rosetta *ab initio* protocol adapted to the membrane environment. An important development was the fast embedding process to find the optimal location for the model in the membrane. This embedding uses a Monte Carlo sampling approach that starts from the average direction of the predicted transmembrane helices as the membrane normal. The embedding makes it possible to use potentials that are specific to a particular layer of the membrane. A decoy is built up from a random pair of helices, adding one adjacent helix at a time, randomly chosen from one end or the other. The protocol achieved some success on small membrane proteins, and was later improved by using constraints from sequence or experiments (Barth *et al.*, 2009). BCL::MP-Fold is an approach based on sampling secondary structure arrangements, which has recently been adapted to use restraints inferred from electron paramagnetic resonance experiments (Fischer *et al.*, 2015).

1.4.4.1 Coevolutionary contact prediction

One area of research that has led to the improvement of *de novo* structure prediction methods is the prediction of contacts from correlated evolution of residues. Using these has led to methods which can achieve high accuracy on

long transmembrane protein domains, and also in soluble proteins. Notable success was achieved using an implementation of contact prediction as input for Rosetta for the most recent CASP experiments (Kinch *et al.*, 2016). Predicted contacts can be the basis of an additional potential in the scoring function, penalising contacts which are not satisfied (typically defined as < 8 Å). The precision of these predictions has greatly improved in the last ten years. The earliest methods calculated correlated evolution of residues between columns in a multiple sequence alignment (MSA) (Göbel *et al.*, 1994). Now, the leading methods use more sophisticated techniques, and a much greater volume of sequence data is available to construct MSAs.

Mean field direct coupling analysis (mfDCA) maximises entropy to extract coupled positions in the alignment (Marks *et al.*, 2011). The method aims to separate true interactions from spurious correlations that are due to indirect coupling or phylogenetic bias. This method was used in EVFold as input to CNS, a simulated annealing method for generating structures from NMR constraints, to generate models (Marks *et al.*, 2011). The same technique was used to fold transmembrane proteins, achieving a TM-score > 0.5 for 22 out of 25 benchmark proteins (Hopf *et al.*, 2012). To achieve these results, predicted contacts were filtered using the predicted transmembrane spans, as shown in Figure 1.12. Where two residues were predicted to be in contact, but the two residues were not positioned at a similar depth in the membrane, that contact was removed.

To achieve the same end, the protein sparse inverse covariance estimation program (PSICOV) inverts the sparse covariance matrix obtained from the MSA (Jones *et al.*, 2012). PSICOV contacts, like those from mfDCA, have been used to aid prediction of membrane proteins. They were used as input to the fragment-based structure prediction method, FILM3, after filtering out contacts by a method similar to that shown in Figure 1.12 (Nugent and Jones, 2012).

While good predictions were achieved by these methods, it was found that at least one sequence per residue is required to generate good contact predictions. Coverage of the full protein by the alignment is important, as

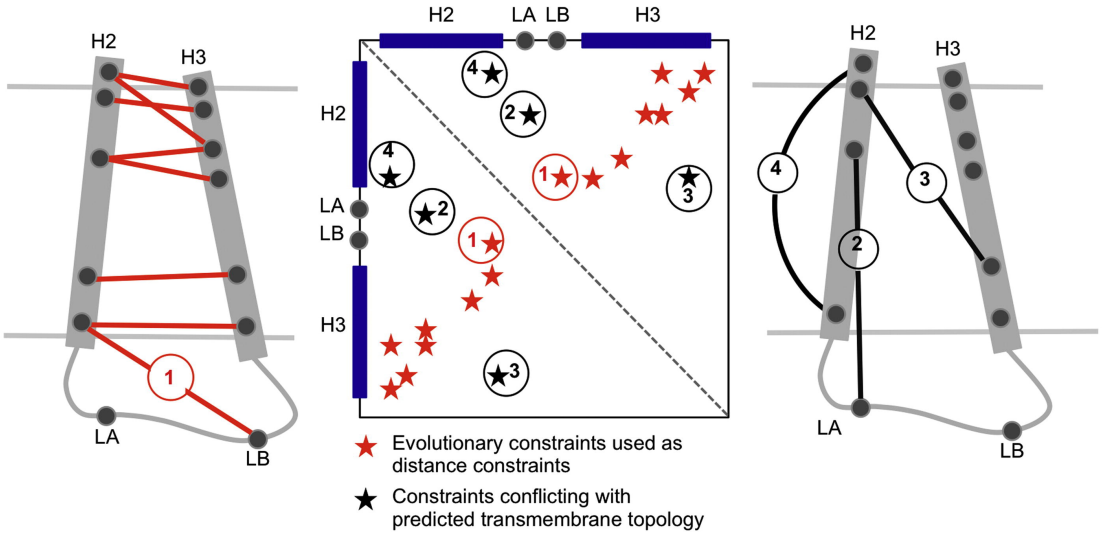


Figure 1.12: Removal of predicted contacts which are incompatible with topology (transmembrane span) prediction. Contacts retained are shown in red; those removed are shown in black. Figure reproduced from Hopf *et al.* (2012).

important coevolution information may be missed if the entire length of the target is not covered. The solved proteins used in benchmark tests are well characterised, and there are usually many sequences available. Typical prediction targets tend to be less studied, therefore it is common that few homologous sequences are available.

Requiring fewer sequences for the same accuracy, new machine learning methods predict contacts using a consensus of the above methods, and others such as CCMpred (Seemayer *et al.*, 2014). MetaPSICOV (Jones *et al.*, 2015), PConsC3 (Michel *et al.*, 2017) and RaptorX-Contact (Wang *et al.*, 2017) are among the programs using this strategy, combined with secondary structure prediction and predicted solvent accessibility. Neural networks or random forests are used to combine the features and predict contacts. When a number of contact prediction methods were used in the context of fragment-based structure prediction, MetaPSICOV achieved the most consistent results across different targets (de Oliveira *et al.*, 2017b). Almost all models were within 0.05 TM-score units of the best possible TM-score by other methods. MetaPSICOV was also among the best methods benchmarked on the proteins from CAMEO and

CASP11 assessments (Michel *et al.*, 2017). RaptorX-Contact was not included in the above tests, but has recently claimed the highest precision and recall compared to true contacts (Wang *et al.*, 2017). Metagenomic sequences have now been used to increase the number of effective sequences and expand the number of targets which can be attempted (Ovchinnikov *et al.*, 2017).

Some contact prediction methods have been trained specifically for membrane proteins, taking into account the additional constraints from Figure 1.12 as an input for the machine learning algorithms. Membrain is a machine learning method using PSICOV predicted contacts and transmembrane span prediction (Yang *et al.*, 2013). The number of models with TM-score > 0.5 produced by I-TASSER using this information was double the number achieved with no contacts. Another membrane specific method has been used to add predicted contact constraints to folding in BCL::Fold, reducing the RMSD from $\sim$ 6–8 Å to $\sim$ 4–7 Å (Teixeira *et al.*, 2017). They also explored the effect of introducing a number of correct contacts, which led to a further improvement for each target by around 2 Å on average.

Standard contact prediction methods have also been tested on membrane protein targets, and achieved good results even though they were trained on soluble proteins (Ovchinnikov *et al.*, 2016; Wang *et al.*, 2017).

1.4.4.2 Fragment library generation

For many *de novo* structure prediction methods, the fragment library is an important part of the method as it dictates the possible conformations to choose from. The RMSD of fragments compared to the native structure at that position can be used to indicate their accuracy, with an RMSD < 1.5 Å considered a good fragment. The percentage of positions with a good fragment is referred to as coverage. Excess inaccurate fragments lead to wasted moves, therefore the overall proportion of fragments that are good (precision) is also important. NNMake has been used as the standard method of fragment library generation for the first stages of the *ab initio* protocol of Rosetta. Its libraries contain 200

fragments per position, all nine residues long (Gront *et al.*, 2011). Flib is a more recent method developed to consider different secondary structure types separately (de Oliveira *et al.*, 2015). Flib uses secondary structure prediction from PSIPRED (Jones, 1999) and torsion angles predicted by SpineX (Singh *et al.*, 2014) to score fragments and extract the best. Flib achieves better precision and almost equivalent coverage compared to NNMake, but with an average of only 26 fragments per position. For training and testing of *de novo* structure prediction methods, fragments from protein structures homologous to the target are removed, as these improve the quality of predictions and they would not be available for a target of unknown structure (de Oliveira *et al.*, 2015).

1.4.4.3 SAINT2

I have described secondary structure prediction, contact prediction and fragment libraries as contributors to the accuracy of *de novo* structure prediction. In addition to these components, a variety of sampling strategies are possible. The standard approach is a Monte Carlo method that begins from a fully extended chain, and which may use simulated annealing. Aiming to explore conformational space more effectively, EdaFoldAA (Simoncini and Zhang, 2013) learns what might be native-like conformations from other decoys and biases the search to those fragments. The sampling performance in this program has been analysed in terms of entropy to show that it explores a greater range of conformational space than Rosetta's low-resolution search (Kandathil *et al.*, 2016). UniCon3D (Bhattacharya *et al.*, 2016) is inspired by a foldon-like assembly building up from smaller units to construct a fold in a step-wise manner.

SAINT2 (de Oliveira *et al.*, 2015, 2017a), developed by our group, is a fragment-based protein structure predictor that aims to improve the efficiency of folding and accuracy of decoys by imitating biological folding. SAINT2 has Forward, Reverse and In vitro modes, illustrated in Figure 1.13. The Forward mode explores conformations in an analogous way to a cotranslationally folding protein, beginning with a nine-residue fragment at the N-terminus and

sampling conformations as the chain grows. There are two different kinds of step: extrusion and move. In an extrusion step, a random fragment is replaced at the growing C-terminal end of the peptide, but with one extra residue to elongate the chain. Extrusion steps are not scored and are always accepted. In a move step, a random location in the peptide is selected, a new fragment is proposed at that site, and the coordinates of atoms are recalculated for the new torsion angles at that position. The new conformation is scored, using a number of physics and knowledge-based potentials, with weights for each component shown in Table 1.1. The replacement is accepted or rejected depending on the score: if it is lower than the previous score, the move is accepted, but if it is higher, there is still a probability it will be accepted. In the Forward mode, 10,000 moves are carried out between extrusions, and these are distributed linearly so that there are more moves between extrusions when the peptide is longer. This ensures that a similar number of replacements occurs at each position throughout chain growth, as there are more positions to choose from later in the process. When all residues have been extruded, 1,000 further full length moves are performed to generate the final decoy.

The Reverse mode is identical to the Forward mode, except that it begins at the C-terminus and extensions are carried out towards the N-terminus.

The In Vitro mode is similar to all other fragment-based structure predictors. It begins from a complete and fully extended chain and performs move steps to build a decoy. All other aspects of the three modes are identical, all using fragments generated by the program Flib (de Oliveira *et al.*, 2015), described above.

For a test set of 41 diverse soluble proteins, the Forward mode produced a correct answer (TM-score > 0.5) in 17 cases, compared to only 13 cases for the In Vitro mode. Of the 13 cases where SAINT2 In Vitro produced a correct answer, 10 were predicted better by the Forward mode (Figure 1.14). For this study, 10,000 decoys were generated for each mode, using the same number of moves for each decoy and the same fragment library. With the same number of moves, the time to generate a single decoy by the Forward mode is only

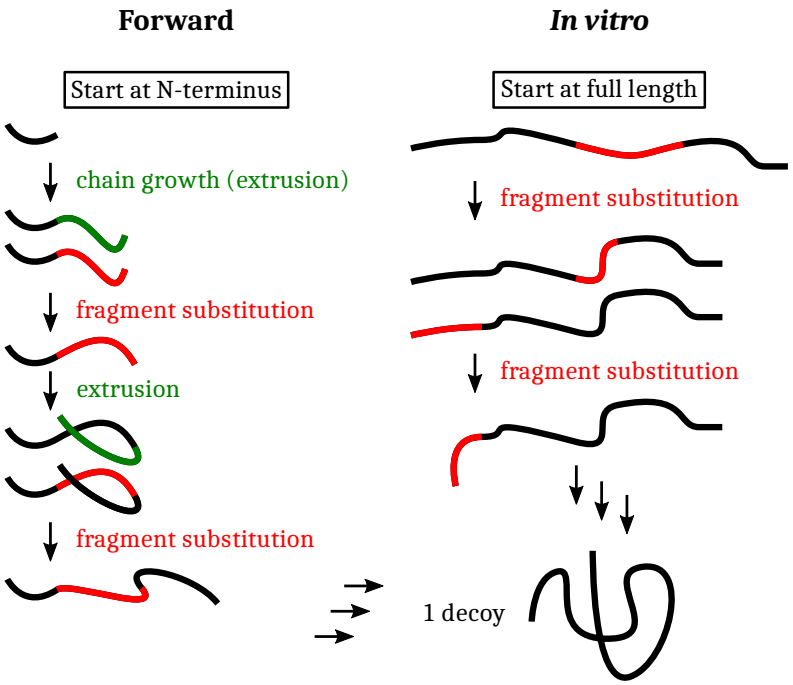


Figure 1.13: The Forward and In vitro modes of SAINT2. Forward uses 'move' (fragment substitution) and 'extrusion' (chain growth) steps, while In vitro performs only 'move' steps. Reverse mode is the same as Forward but starts at the C-terminus.

Potential	Short	Long
Lennard-Jones	0.505	0.304
RAPDF	0.156	0.303
Orientation	0.077	0.111
Solvation	0.262	0.282
Contact	1.000	1.000

Table 1.1: SAINT2 potentials and weights. Different weights are used to score short ($<$ 150 residues) and long ($\geq$ 150 residues) peptides. The raw potentials are transformed to have similar ranges, therefore weights are an indication of their relative importance.
The Lennard-Jones potential is the standard approximation to represent the physical attractive and repulsive forces between non-bonded atoms.
The RAPDF is as described in Section 1.4.4.
The Orientation score is a statistical potential similar to an RAPDF, but also binned according to the relative orientation of the side chains of the interacting residues.
The Solvation score is a statistical potential based on counts of neighbouring atoms, i.e. for an atom that is preferentially solvent exposed, lower counts would lead to lower score.
The Contact score is a penalty for every contact predicted by MetaPSICOV between pairs of residues $>$ 8 Å apart. The penalty is proportional to $(s - 8)$ where s is the distance between the residues in Å.

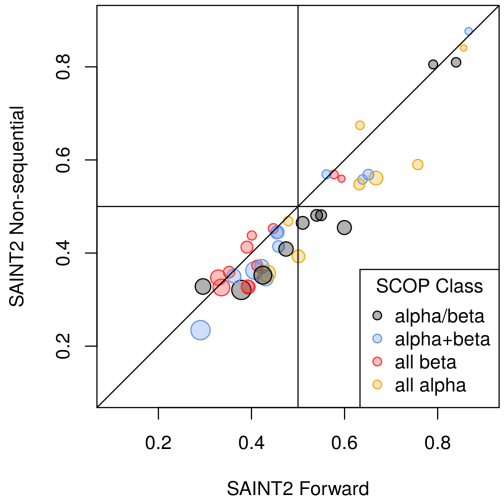


Figure 1.14: The TM-score of the best decoy produced by In vitro (Non-sequential) and Forward modes for a set of 41 soluble proteins. The majority of points are below the diagonal line, indicating that the Forward mode generates a better model in most cases. Figure reproduced from de Oliveira *et al.*, 2017a.

44–73% of the time taken by the In Vitro mode, as most calculations are made on a greatly reduced number of residues.

While this approach has been used to predict the structures of soluble proteins, it has not been tested for membrane protein targets. In this thesis I investigate its applicability to membrane proteins and go on to adapt the method for the purpose of completing partial structures.

1.5 Thesis summary

In this thesis, I use computational methods to study the features and folding of membrane proteins, making comparisons to soluble proteins where comparable data is available.

In Chapter 2, an analysis of helices in homologous proteins is carried out to investigate the conservation of kinks and kink angles. I identify that many kinks are not conserved between related structures and that they may be important for flexibility and function.

In Chapter 3, I find evidence of cotranslational folding in the structures of

membrane proteins, using statistical measures, and through comparing the performance structure prediction with SAINT2 by a sequential and *in vitro* method.

Chapter 4 describes the adaptation of SAINT2 to the membrane environment, and the development of a new version, SAINT2-ScafFold. SAINT2-ScafFold uses the principle that the N-terminus folds during the process of translation, therefore predicting by "folding" against it may improve the quality of prediction of the rest of the structure.

In Chapter 5, I use SAINT2-ScafFold for a new application, completion of templates that do not cover entire targets, in order to generate complete homology models.

Chapter 6 summarises the findings of my work, and suggests possible directions for future research in this area.

2

Examining the conservation of kinks in alpha-helices

This chapter is based on a paper published in PLoS One (Law *et al.*, 2016), of which I am the first author and for which I carried out the analysis of homologous pairs and families of alpha-helices. The paper also describes a method for estimating the error on kink angle measurements, which was developed by my co-author Henry Wilman. This method is described in detail in Appendix A.

Kinks are a structural feature of alpha-helices and are thought to have possible functional roles in at least four types of ion channel (Tieleman *et al.*, 2001; Alam and Jiang, 2009; England *et al.*, 1999; Ri *et al.*, 1999), four types of transporter (Ni *et al.*, 2011; D'Rozario and Sansom, 2008; Pebay-Peyroula *et al.*, 2003; Hilger *et al.*, 2009), gap junctions (Ri *et al.*, 1999), GPCRs (Bettinelli *et al.*, 2011; Deupi, 2012; Katritch *et al.*, 2013; Tehan *et al.*, 2014), and the amyloid precursor protein (Barrett *et al.*, 2012). Kinks have previously tended to be defined in a binary fashion. In the work described in this chapter, I have deliberately moved towards defining them on a continuum, which given the unimodal distribution of kink angles is a better description. From this perspective, I examine the

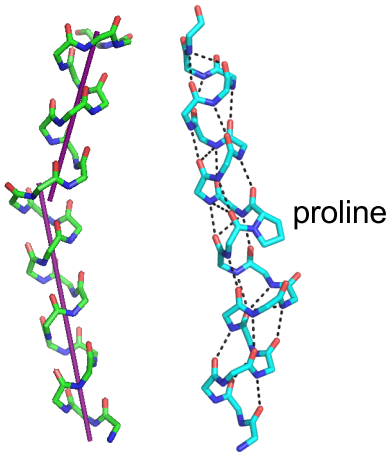


Figure 2.1: Examples of kinks showing associated features. Left: Change in the direction of the helix axis (shown in purple); Right: disruption of hydrogen bonds (shown as black dashed lines), especially at proline residues.

conservation of kinks in proteins, an important consideration in the homology modelling process.

2.1 Background

Disruptions of the ideal helix geometry in proteins, often called kinks or bends, frequently occur in transmembrane helices (TMHs) and the long helices in soluble proteins (Wilman *et al.*, 2014b). At a kink, the helix axis changes direction and the helical hydrogen bonding pattern is often broken (see Figure 2.1).

2.1.1 Identification of kinks

As they are motifs of interest, computational methods have been developed to identify them in order to assess their prevalence and understand what causes their formation. HELANAL is an early method which used a vector method to calculate the local helix axis and centre (origin) at each point along a helix, based on the locations of every set of four consecutive C_α atoms (Bansal *et al.*, 2000). If the angle between any two consecutive local helix axes is $> 20°$, the location is annotated as a kink. Helices without a kink are classified as either curved, straight or unclassified, based on the quality of fit by a circle or line to all the local helix origins. A later extension to the program, HELANAL-Plus,

considers a fit to the surface of a sphere instead of a circle for classification of curved helices (Kumar and Bansal, 2012).

MC-HELAN uses an alternative strategy to find helix segments and their axes, starting from helical seeds that must satisfy a number of geometric requirements (Langelaan *et al.*, 2010). An axis is fitted by a Monte-Carlo algorithm, and a further residue is added to the end of the helix segment if ϕ/ψ angles are helical and the backbone atoms are < 3 Å from the axis on average. Segments are located and extended iteratively from both the C-terminus and N-terminus. Any helix segments that are overlapping are defined as bends, while any transmembrane span that is made up of more than one helix segment with non-helix residues between is a disruption. This process leads to identification of more gradual changes of direction as bends, as the segments are long compared to the four-residue stretches used by HELANAL.

Prokink was developed to analyse the geometry around kinks in helices at the site of proline residues during molecular dynamics simulations (Visiers *et al.*, 2000). The bend angle is calculated together with the direction of bending and the tightness of winding of the helix. All three measures give an indication of structure changes over a trajectory.

Kink Finder fits cylinders to stretches of backbone atoms of six residues, and calculates angles between adjacent cylinders in order to locate kinks (Wilman *et al.*, 2014b). As for HELANAL, an angle of $> 20°$ is defined as a kink, but the distribution of angles found in helices is unimodal, indicating that kinks may not be well defined by a discrete classification.

In an attempt to overcome the contradictory computational methods, an orthogonal approach was taken by Kneissl *et al.* (2011), who manually assigned kinks. This was taken one step further by Wilman *et al.* (2014a) who used crowd sourcing for kink identification, combining the observations of 310 people. In most cases, there was not universal agreement on whether a helix was curved, kinked or straight, emphasising the difficulty of classification.

It is clear from the variety of existing methods of kink identification that there is not a simple and reliable way to annotate kinked helices. Even changes of direction of similar sizes can result from very different geometry and hydrogen bonding (Visiers *et al.*, 2000). Indeed, it may be more appropriate to avoid a binary classification, and in this chapter I pursue a comparative approach that is concerned with differences between helices as well as when they may be considered "kinked" on an absolute scale.

In addition, none of these methods, computational or manual, has any estimate of error on the kink angle measurements. Knowing the error on the kink calculation may help explain the discrepancies between definitions, and it also allows us to test whether kinks in different structures are statistically different. In this chapter, I use the Kink Finder method, and the heuristic error estimation method which has been implemented within it.

2.1.2 Sequence, function and flexibility of kinks

Even with the difficulties in kink definition, specific residues have been repeatedly associated with kinks. The most common of these is proline, which often occurs in the residues following a kink (e.g. Cordes *et al.*, 2002; Langelaan *et al.*, 2010; Wilman *et al.*, 2014b). This is thought to be due to the lack of an amide hydrogen atom on the nitrogen atom of proline. In a helix, this atom would usually form a hydrogen bond to the backbone carbonyl oxygen of the residue four earlier. Proline is found in many of the helices with the largest kink angles, however, up to two-thirds of kinked helices do not contain proline (e.g. Hall *et al.*, 2009; Langelaan *et al.*, 2010).

It has also been proposed that a proline could initially cause a kink to occur in a structure, after which it may mutate to another residue with the kink remaining (Yohannan *et al.*, 2004a). This hypothesis suggests that kinks will generally be conserved despite changes in sequence, and that both local sequence effects and more global interactions with neighbouring helices should be considered. Another indicator that kinks may be a conserved feature is

that kinks are often annotated as coil residues, and coil residues found in the central core of membrane helices (also including re-entrant helices) are often conserved (Kauko *et al.*, 2008).

The conservation of sequence and structure at the site of kinks indicates potential functional relevance, which is supported by experimental evidence and simulations in a number of cases. On this basis, it has been proposed that the gating of several ion channels is enabled by flexibility at conserved motifs: a PVP motif in the S6 helix of the Shaker Kv channel; a GxP motif in TM helix D5 of CLC channels (Tieleman *et al.*, 2001); a conserved glycine in NaK channels (Alam and Jiang, 2009); a GxxP motif in the Alm channel-forming peptide (Tieleman *et al.*, 2001); and a conserved proline in the M1 helix of the nicotinic acetylcholine receptor (England *et al.*, 1999; Dang *et al.*, 2000). In connexin32, the unusually large bend angle at a TP motif was reduced by mutation of the threonine to alternative amino acids, which caused closure of the channels at smaller voltages than for the native structure (Ri *et al.*, 1999; Sansom and Weinstein, 2000). In the human breast cancer resistance protein, an ABC transporter, mutation of a mid-helix proline to alanine affected substrate specificity (Ni *et al.*, 2011). It was therefore suggested that a molecular hinge induced by the proline could affect the structure of the drug-binding cavity or coupling of ATP hydrolysis and transport (Ni *et al.*, 2011; Rosenberg *et al.*, 2015). Structural and simulation studies in the glycerol-3-phosphate transporter (D'Rozario and Sansom, 2008), the ADP/ATP carrier (Pebay-Peyroula *et al.*, 2003), and the Na/proline symporter PutP (Hilger *et al.*, 2009) show that both proline and non-proline kinks appear to play a role in regulating access to binding cavities in transporters. In the amyloid precursor protein, a diglycine kink is said to contribute to the γ-secretase cleavage of C99 by allowing the helix to be flexibly curved for interaction with the protease (Barrett *et al.*, 2012; Cao *et al.*, 2017).

The crucial property of kinks for the function of the above proteins is their ability to create a flexible point within a structure. Molecular dynamics (MD) simulations on individual transmembrane helices (Tieleman *et al.*, 2001) have

shown a range of angles can be adopted by a single helix. MD on voltage-gated potassium channels (Choe and Grabe, 2009) showed a range of conformations for their helices, which could be significant in ion channel gating. Bettinelli *et al.* (2011) used a simplified model to explore conformational change in a large superfamily of membrane proteins: G-protein coupled receptors (GPCRs). They created conformational chimeras where the four proline-containing helices of GPCRs were independently able to adopt kinked or straight conformations, and the whole model was allowed to relax through MD simulations. They studied the stability of all 16 possible conformations and found that bound agonist favoured straight conformations, but bound antagonist favoured bent conformations. This suggests that kinks may have an important role in changes of conformation upon binding. This methodology was also applied to models of the human mAChR1 receptor, and used to improve virtual screening by using the different conformations produced (Pedretti *et al.*, 2015).

Flexibility of this type can also be seen by inspecting the differences between crystal structures of the same protein in different conformations. One case where this has been carried out is some of the subfamilies of GPCRs (Tehan *et al.*, 2014). The inactive and activated structures of the β_2-adrenergic receptor were compared and a 'swinging' and some unwinding of transmembrane helix (TMH) 6 was observed about the kink location (Katritch *et al.*, 2013; Tehan *et al.*, 2014), but it was shown to maintain the same bend angle. Alpha bulges, also known as π-turns, are a feature of TMH 2 and 5 in most GPCRs, and these are points where twisting and bending is seen when comparing inactive and activated structures of other receptors (van der Kant and Vriend, 2014). HELANAL has been used to compare the kinks in inactive and activated rhodopsin, opsin and the β_2-adrenergic receptor (Deupi, 2012). Differences in some of the kink angles were seen; however, as HELANAL has no method of estimating measurement error, the significance of these differences is difficult to evaluate.

These previous studies on kink flexibility mostly focus on one specific protein or model. In this study, I compared helices from homologues, rather than from

different structures of the same protein. This strategy made it possible to use a much larger set of data. The differences in angles seen may have been due to sequence differences between homologues, or they may show examples of kink flexibility. Using the error estimation method of Kink Finder, I was also able to test whether the angle differences are significant. I compared the angles of helices, without the need to classify them as "kinked" or "straight" according to an arbitrary threshold. I instead classified helices that display different angles as not conserved. In sets of pairs and families of homologous proteins, it was common to find homologous helices which were differently kinked. I then carried out an extended analysis of the seven transmembrane helices of GPCRs. The kinks in TMH 6 and 7 were well conserved, but the others showed greater variation. One receptor displayed a change of kink angle between agonist and antagonist bound structures, therefore my results also support the belief that kinks are functionally important.

2.2 Methods

2.2.1 Angle measurement by Kink Finder

Kink Finder (Wilman *et al.*, 2014b) was used to measure angles in helices. Kink Finder fits a cylinder to every 6-residue segment of a helix by minimising r, where

$$r = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(d_i - \bar{d}\right)^2} \tag{2.1}$$

m is the number of backbone atoms in the segment (24), d_i is the shortest distance from backbone atom i to the fitted helix axis, and $\bar{d}$ is the mean of all distances. The angle is measured between the axes of the cylinders fitted to adjacent segments, and this angle is assigned to the final residue of the first segment (Figure 2.2). Only helices 12 residues or longer can be analysed.

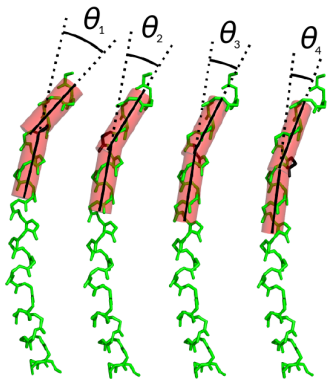


Figure 2.2: Angle measurement by Kink Finder. Cylinders, shown in red, are fitted to each six-residue segment of the helix. Angles θ_1, θ_2... are measured between the axes of adjacent cylinders, and allocated to the last residue of the first segment, shown in black. In this way an angle is assigned to every residue in the helix except the first five and the last six. Adapted from Wilman *et al.* (2014b).

2.2.2 Method of confidence interval estimation

The quality of fit, r, obtained by the cylinder fit method (Equation 2.1), can be used to estimate the error in the kink angle measurement. If the helix backbone is well fitted by a cylinder both before and after a kink, we have greater confidence in the calculation. Appendix A describes how this relationship was analysed by Henry Wilman to estimate the following statistical confidence interval of a measured kink angle, θ:

$$95\% \text{ confidence interval for true angle} = \theta \pm \varepsilon \tag{2.2}$$

I will refer to ε as the error, which is assumed to be symmetric, and which is calculated using the quality of fit on the N- (r_n) and C- (r_c) terminal sides of the kink:

$$\varepsilon = (6.349 \times \ln(r_n + r_c - 0.2937) + 13.15) \tag{2.3}$$

Thus ε is an estimate of the uncertainty in the kink angles measured by Kink Finder and provides a simple way to compare the angles in two helices.

2.2.3 Data sets

I gathered sets of soluble and membrane protein chains using the same methodology as in Wilman *et al.* (2014b). Soluble protein chains were obtained on

3rd March 2015 by filtering the protein data bank (PDB) (Berman *et al.*, 2003) using PISCES (Wang and Dunbrack, 2003) to select chains with $< 80\%$ sequence identity to each other, resolution < 5 Å, $40 <$ chain length < 1000 residues, and R-factor < 0.4. The set includes the first conformer of NMR structures but structures from electron microscopy experiments, and Cα atom-only structures were removed. Any protein chains in the PDBTM (Kozma *et al.*, 2013), Membrane Proteins of Known Structure Database (White and Wimley, 1999) or Orientations of Proteins in Membranes database (Lomize *et al.*, 2006) were removed to eliminate transmembrane structures. A second set of soluble proteins was obtained in the same way, but containing crystal structures only and with resolution < 2 Å and R-factor < 0.2. The results for this high quality data set were similar to those for the first set of soluble proteins, and are shown in Appendix B, Table B.2.

To create a set of membrane protein chains, a list of polytopic alpha-helical membrane protein PDB codes was taken from the Membrane Proteins of Known Structure Database on the 8th January 2015 and the PDBTM on 9th January 2015. After splitting into chains, a clean, non-redundant set was generated as described above for the soluble set. A redundant set of membrane protein chains was also kept where maximum sequence identity was 99%.

The G-protein coupled receptors database (GPCRDB) (Isberg *et al.*, 2014), accessed 4th November 2014, provided a set of 122 chains from crystal structures of G-protein coupled receptors of 28 different receptors. The database provides its own numbering scheme and structural alignment for these receptors.

For helices from each data set, the distribution of maximum angles measured by Kink Finder is shown in Figure 2.3. A list of PDB codes in each set is available in the online supplementary data package (Law *et al.*, 2016).

2.2.3.1 Identifying homologous helices

For the sets of soluble, membrane, and redundant membrane chains, helices were annotated using JOY (Mizuguchi *et al.*, 1998). Any helical segments

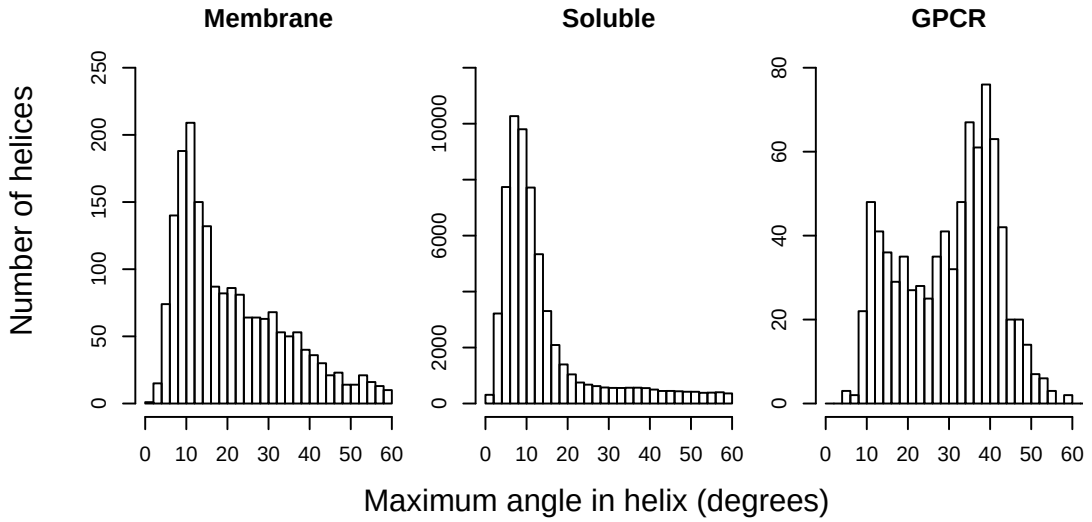


Figure 2.3: The distribution of maximum angles measured by Kink Finder in helices from the non-redundant membrane (from which MemPairs and MemFams were taken), non-redundant soluble (from which SolPairs and SolFams were taken) and GPCR data sets.

separated by only one or two residues were combined. The ends of helices were trimmed until they satisfied the criteria for a helical seed used by MCHELAN (Langelaan *et al.*, 2010). For membrane proteins, helices were only kept if at least one residue was annotated to be in the tail region of the membrane by iMembrane (Kelm *et al.*, 2009).

An all-against-all structural comparison of the protein chains in a set was carried out, using TM-align (Zhang and Skolnick, 2005). Pairs of helices were considered homologous if:

- the number of residues in the longer chain was no more than 50% greater than the number in the shorter chain
- the two structures shared the same fold, indicated by a TM-score of the alignment greater than 0.5 (Zhang and Skolnick, 2005)
- the sequence identity in the TM-align alignment was at least 10% (ignoring gaps)

- the two helices aligned in such globally similar structures had ends that were offset by no more than four residues in the TM-align alignment

This resulted in 629,524 homologous helix pairs in soluble chains (SolPairs), 4,104 pairs in the membrane chains (MemPairs), and 41,945 pairs in the redundant set of membrane chains (RMemPairs).

2.2.3.2 Identifying homologous aligned families of helices

Using the non-redundant membrane homologous helix pairs, a network of helices was constructed, where an edge connected each pair of homologous helices (defined in Section 2.2.3.1). Communities of related helices were extracted using the software of Traag *et al.* (2011), with resolution parameter $\lambda = 25$. All members of a community had to be connected to $> 90\%$ of the other members in order to be a helix family. If a community had any member with connectivity $< 90\%$, the member with the lowest connectivity was removed. Connectivity of other members was recalculated and the process repeated until all members satisfied the requirement of connectivity $> 90\%$. Families of fewer than five members were discarded, leaving 45 membrane helix families (MemFams). This process was repeated for soluble helices and 1258 soluble helix families (SolFams) were identified. For each helix family, a multiple structural alignment of the full protein chains was generated using MAMMOTH-Mult (Lupyan *et al.*, 2005).

2.2.3.3 Obtaining helix families for the seven transmembrane helices (TMHs) of G-protein coupled receptor structures (GPCRs)

The GPCRDB provides a structure-based alignment of each of the helices of GPCRs (Isberg *et al.*, 2014). The sequences in GPCRDB are the native sequences, but some GPCR structures contain mutations. In order to align these structures correctly, they were aligned to the receptor sequence provided without introducing any gaps. The GPCRDB alignments of helices were truncated to the first and last residues in each helix where more than half of the structures did not have a gap. TMH 7 was further shortened to start from residue 7x33,

as the consensus secondary structure annotated by JOY (Mizuguchi *et al.*, 1998) for the first four residues was not helical. This resulted in a helix family of at least 113 members for each of the GPCR TMHs.

2.2.4 Comparison of two helices using error estimation

Kink Finder (described in Section 2.2.1) was used to measure angles at all sites in all of the helices in the data sets. For each homologous helix pair, the site of the greatest angle in either helix was used as the most disrupted site for classification. This angle was compared to the largest angle in the other helix, within a window of ± 4 residues either side of the kink site. A window was used because it is both difficult to accurately define the position of a kink, and there may also be error in the alignment. If no angle was found in this window due to gaps, the helix pair was removed from the set. The pairs were then classified by the scheme shown in Figure 2.4 using the error estimate on each angle, ε (Equation 2.3). The "Not Conserved" class includes homologous helix pairs where both helices are kinked, but with significantly different angles.

When the difference between two angles is less than the sum of the errors on those angles, the difference is not significant. In these cases there may be a real difference between the angles but it is not found to be significant. This could lead to an underestimation of the number of kinks which are not conserved, classifying some as Conserved Straight, Conserved Kinked or Other.

2.2.4.1 Calculation of 'neighbouring sequence identity'.

I calculated the 'neighbouring sequence identity' of a homologous helix pair as a measure of sequence conservation among the residues in spatial proximity to the helices. For every helix in the data set, I found all surrounding residues in the chain that had at least one atom within 4 Å of an atom in the helix. For a pair of homologous helices, the 'neighbouring sequence identity' was the sequence identity over all the positions that were in the set of neighbouring residues for either helix and that were not gaps in the structural alignment.

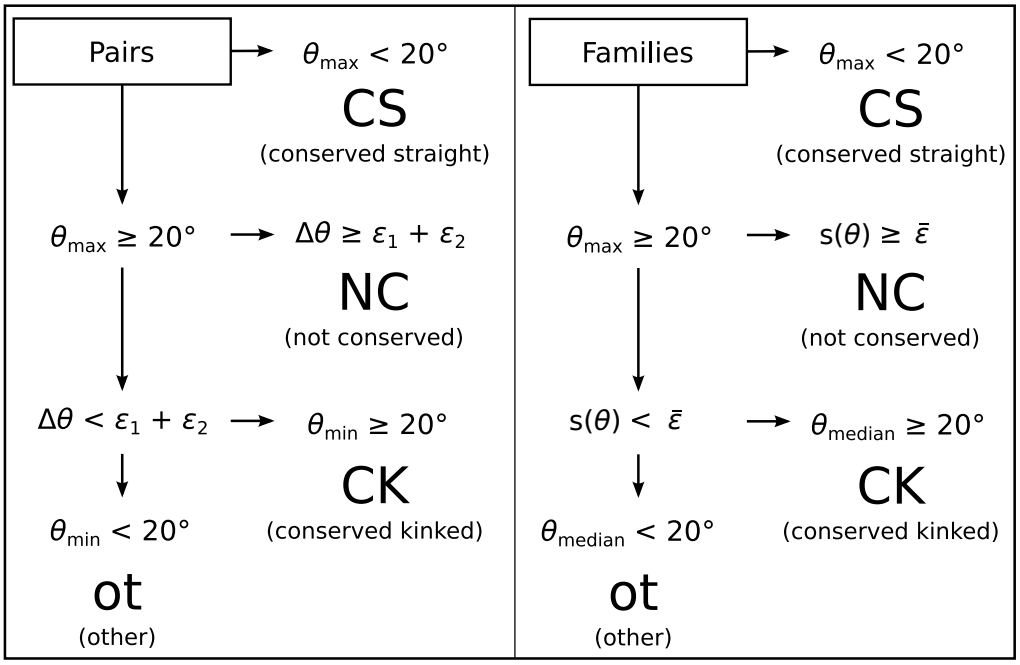


Figure 2.4: Flowchart showing the classification of homologous helix pairs and families. Pair classification uses the angles of the two helices at the most disrupted site (θ_{max}, θ_{min}), angle difference ($\Delta\theta$) and the error on each angle (ε). Family classification uses analogous statistics to obtain the same classes: the median angle (θ_{median}), standard deviation ($s(\theta)$), and mean error ($\bar{\varepsilon}$) of the angles in the family.

2.2.5 Classification of families based on a 'most disrupted' site

For the homologous helix families in the MemFam and SolFam sets, the MAMMOTH-Mult alignment was used to compare helices; for the GPCR set, the GPCRDB alignment was used. Angles were measured by Kink Finder at each residue, and assigned to that residue's position in the alignment. For each residue in a helix in the alignment, the maximum angle from a window of three residues was used as the smoothed angle for that residue (Figure 2.5). The smoothing allows for inaccurate alignment when comparing the angles around the sites of the largest angles. The angles of each of the helices in a family can then be compared at every site in the alignment.

Only one site of maximum disruption was used to classify a family. The site in the helix with the highest mean was chosen as the most disrupted site. Only sites in the alignment of the angle data which had at least five recorded angles after smoothing were considered. To determine the variation of angles at

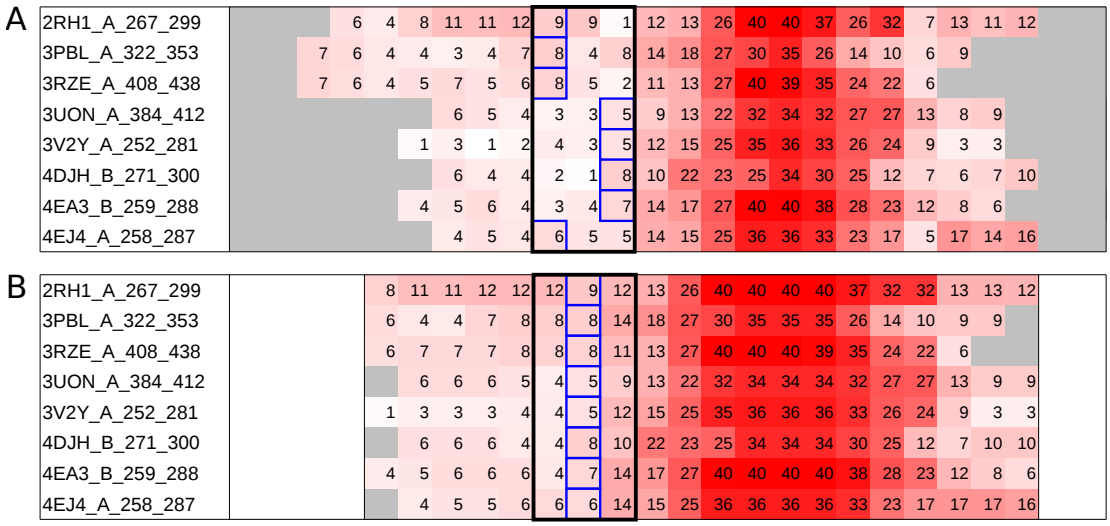


Figure 2.5: **A)** Angle measurements in degrees at each residue of every helix in an example family, in the positions those residues were located when aligned by MAMMOTH-Mult. **B)** Smoothed angles produced by taking the maximum angle (blue boxes) in a window of one residue either side of the position (black box).

the most disrupted site, the standard deviation was calculated. The standard deviation was compared to the mean error of angles at the most disrupted site in order to classify families as conserved or not. The flowchart in Figure 2.4 shows the method of classification for helix families.

All data for the homologous helix pairs and families is available in the online supplementary data package (Law *et al.*, 2016).

2.3 Results

2.3.1 Confidence intervals of angles measured by Kink Finder

In this chapter, I analyse whether helix kinks are conserved between homologues by comparing their angles. The first step is to calculate a confidence interval on the helix kink angles measured. I used the method within Kink Finder (Wilman *et al.*, 2014b) to calculate the confidence interval for every angle (see Appendix A). The typical range of error sizes is around 5–8°, and larger kinks are generally associated with larger errors (Figure 2.6), due to poorer fits. Figure 2.7A shows two helices that have a difference in angle of 15.6°. However, we cannot be

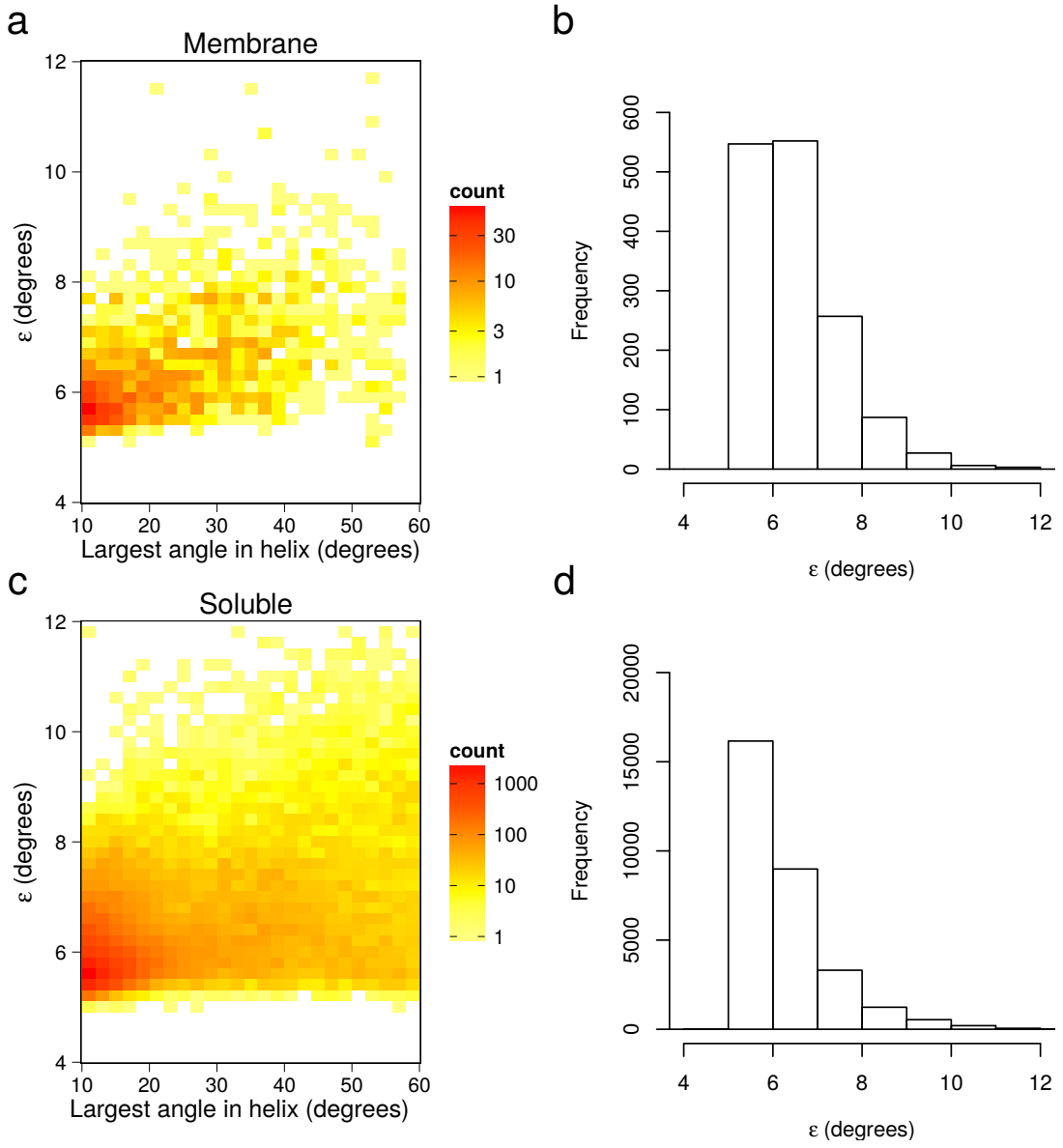


Figure 2.6: The error, ε, for the maximum kink angles in the membrane (a and b) and soluble (c and d) helices. Helices with maximum angle $\leq 10^{\circ}$ are not included. (a) and (c) Heat map showing the variation of ε with angle. Coloured using a log scale. (b) and (d) Histogram of ε.

sure in this case that the kink angles are different because the quality of fit is poor, and therefore the 95% confidence intervals ($\theta \pm$error, ε) are overlapping. In Figure 2.7B, two helices are shown that have angles that differ by only 11.6° but the confidence intervals do not overlap because the quality of fit is better. Thus we consider these helices to have significantly different kink angles.

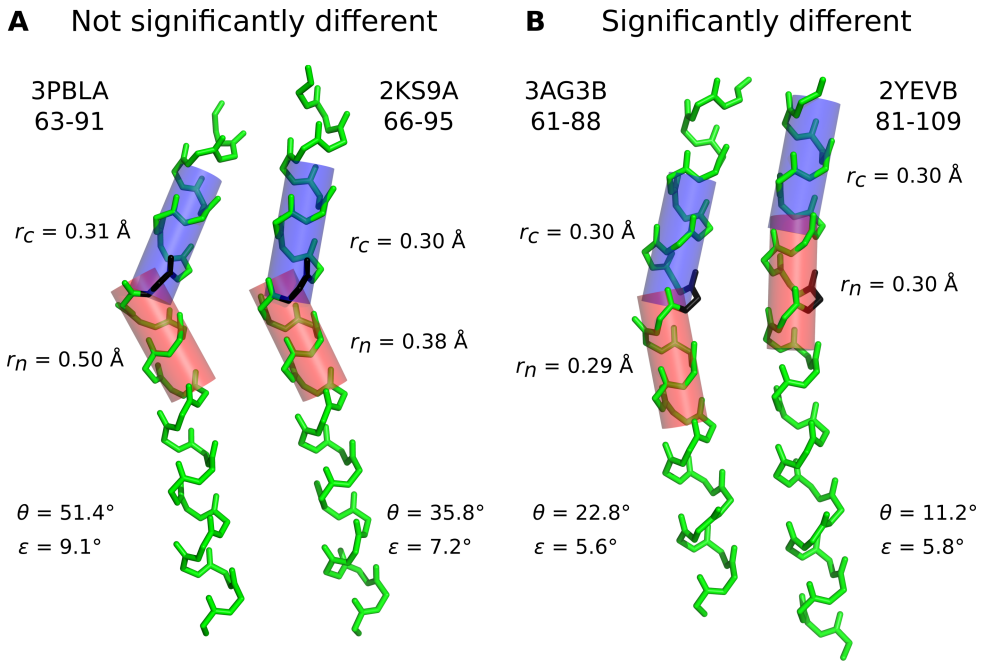


Figure 2.7: Two examples of helix pairs, which are (A) not significantly different and (B) significantly different. PDB code, chain identifier and residue numbers are given for each helix. The black residues are at the most disrupted site (see Section 2.2.4) in each helix pair. r_n and r_c give the quality of the cylinder fit (see Equation 2.1) to the backbone atoms on the N- (red) and C- (blue) terminal sides of the kink site. θ is the angle measured between the two cylinders. ε is the estimated error of the angle measurement, calculated from $r_n + r_c$ using Equation 2.3. If $\theta_{\max} - \theta_{\min} > \varepsilon_1 + \varepsilon_2$, the confidence intervals do not overlap therefore we consider the angles to be significantly different.

2.3.2 Homologous helix pairs

2.3.2.1 Number of helix pairs found

Non-redundant sets of 18,934 soluble and 392 membrane protein chains with at least one helix of 12 or more residues were collected as described in Section 2.2.3. From these, 4,104 aligned pairs of homologous membrane helices (MemPairs) and 629,524 aligned homologous soluble helix pairs (SolPairs) were extracted using the criteria in Section 2.2.3.1. Kink Finder was used to measure the angles and the error on these angles for all residues in every helix by the cylinder fit method found in Section 2.2.1.

2.3.2.2 Definition of pair classes

The homologous helix pairs were divided into four classes, based on the size and uncertainty of the kink angles in the two helices (Figure 2.4):

Conserved Straight $\theta_{max} < 20°$

Conserved Kinked: no significant angle variation $\theta_{min} > 20°$, $\theta_{max} - \theta_{min} < \varepsilon_1 + \varepsilon_2$

Not Conserved: significant angle variation $\theta_{max} > 20°$, $\theta_{max} - \theta_{min} > \varepsilon_1 + \varepsilon_2$

Other All other pairs

θ_{max} and θ_{min} are the larger and smaller of the measured angles in the helix pair; ε_1 and ε_2 are the errors of the two angles. The number of helix pairs in each of these classes is shown in Table 2.1. Conserved Straight (CS) is the most common class for both soluble and membrane proteins, with 77% of soluble helix pairs and 44% of membrane helix pairs belonging to this class. As expected, kinks are more common in the membrane protein set, probably because membrane helices are longer and longer helices are more frequently kinked (Wilman *et al.*, 2014b). Conserved Kinked (CK) pairs are more frequent than Not Conserved (NC) pairs in the membrane set ($29.0 \pm 1.5\%$ compared to $19.2 \pm 1.2\%$) but soluble proteins show the opposite trend ($6.38 \pm 0.05\%$ compared to $14.04 \pm 0.09\%$). Confidence intervals were estimated by extracting 1000 bootstrap samples from each data set, and taking the interval that included 95% of the bootstrap data.

2.3.2.3 Presence of proline in kink pairs

I tested whether the four classes are different in terms of proline occurrence, as proline is commonly associated with kinks. Proline was identified as present if it was at the position of the largest angle or in the four following residues. Table 2.1 shows the presence of proline in the aligned helix pairs, broken down by pair type (for a detailed breakdown see Appendix B, Table B.1).

	Membrane					Soluble				
	CK	CS	NC	other	total	CK	CS	NC	other	total
All	1189	1806	789	320	4104	40190	481388	88390	19556	629524
%	29.0	44.0	19.2	7.8	100.0	6.4	76.5	14.0	3.1	100.0
PP	14.4	0.9	1.9	1.8	19.0	1.3	0.0	0.4	0.1	1.8
P-	4.1	2.7	7.7	1.1	15.6	0.9	0.5	5.7	0.4	7.5
-P	4.4	0.2	1.3	0.8	6.7	0.8	0.1	0.3	0.0	1.2
- -	6.1	40.2	8.4	4.1	58.7	3.3	75.9	7.7	2.6	89.5

Table 2.1: The number of aligned helix pairs in each class, and occurrence of proline within that class. The helix pair classes conserved kinked (CK), conserved straight (CS), not conserved (NC), and other are defined in Figure 2.4. The frequency of proline in each class is given as a percentage of the total number of pairs. PP: proline in both helices; P-: proline in the helix with the larger kink angle; -P: proline in the helix with the smaller kink angle; - -: proline in neither helix. All percentages are rounded to one decimal place.

Proline is common at the most disrupted site in membrane helix pairs, occurring in both helices in $19.0 \pm 1.2\%$ of pairs, however it is much less common in soluble pairs and more often present in only one of the two helices ($7.5 + 1.2 = 8.69 \pm 0.06\%$) rather than in both ($1.82 \pm 0.03\%$). For both membrane and soluble proteins, when proline is present in both helices, most pairs are Conserved Kinked. When proline is not present in either helix at the most disrupted site, Conserved Straight is most common. In an experimental study of bacteriorhodopsin, two of three kinks still remained after mutation from proline to alanine (Yohannan *et al.*, 2004b). The authors also suggested that kinks could be predicted on the basis of proline at the location in homologous proteins, therefore suggesting that kinks are convserved where proline is present in a kinked helix but not present in its homologues. However, in my data, if a helix has a proline and is kinked and its partner helix does not have a proline, the pair may be Conserved Kinked or Not Conserved. Not Conserved (P-) is nearly as common as Conserved Kinked (P- and -P combined) for membrane helix pairs, and more common in soluble helix pairs, indicating that loss of proline is often accompanied by a significant reduction in kink angle.

2.3.3 Relationship between angle difference and sequence identity

While considering the angle difference between homologous helices, I investigated how it relates to the sequence identity between helices (Figure 2.8A), the sequence identity between neighbouring residues (Figure 2.8B, calculated as described in Section 2.2.4.1), and the sequence identity between the complete chains (Figure 2.8C), in all cases ignoring gaps. As expected, there is a trend for larger angle differences to be associated with lower sequence identity. Conversely, kinks are well conserved when sequence is conserved, whether on a local or global scale. For the non-redundant membrane set, the Spearman's rank correlation coefficients are very similar for Helix (-0.28), Neighbouring (-0.29) and Chain (-0.27) sequence identity. The partial correlation coefficients, when using other measures of sequence identity as controlling variables (see Table 2.2) are also similar to each other (~ -0.1), but show that each factor makes an independent contribution to kink angle changes.

This is also true for my non-redundant set of soluble proteins, but the correlation coefficients are weaker (~ -0.1). Conclusions cannot easily be drawn from this difference in correlation coefficients for soluble and membrane proteins, as the distribution of sequence identity is not the same for the two sets.

2.3.4 Homologous helix families

2.3.4.1 Number of families found

As homologous helix pairs showed large numbers of unconserved kinks, I built families of homologous helices in order to investigate kink conservation patterns across larger samples of related helices. Families contained at least five members, and Table 2.3 gives the number of families of various sizes, displayed graphically in Figure 2.9.

Mem/ Sol	Resolution cutoff (Å)	R factor cutoff	PISCES cull %SID	Proline kinks included	Number of helix pairs in dataset	Correlation coefficient with angle difference			Partial Correlation coefficient with angle difference					
						Helix	Neigh	Global	Helix (controlling for Global)	Global (controlling for Helix)	Helix (controlling for Neigh)	Neigh (controlling for Helix)	Neigh (controlling for Global)	Global (controlling for Neigh)
M	5	0.4	80	Yes	4105	-0.278	-0.290	-0.265	-0.134	-0.101	-0.110	-0.140	-0.132	-0.048
M	5	0.4	80	No	2412	-0.228	-0.225	-0.214	-0.107	-0.070	-0.103	-0.095	-0.084	-0.044
S	5	0.4	80	Yes	630333	-0.127	-0.115	-0.098	-0.087	-0.031	-0.079	-0.057	-0.063	-0.019
S	5	0.4	80	No	563959	-0.096	-0.101	-0.084	-0.058	-0.035	-0.050	-0.059	-0.058	-0.013
S	2	0.2	80	Yes	182860	-0.121	-0.107	-0.093	-0.083	-0.030	-0.077	-0.052	-0.057	-0.019
S	2	0.2	80	No	162111	-0.088	-0.094	-0.077	-0.054	-0.033	-0.046	-0.056	-0.054	-0.010

Table 2.2: Table of Spearman's rank correlation coefficients between angle difference ($|\theta_{\max} - \theta_{\min}|$) and each measure of sequence conservation. The three measures are helix sequence identity (Helix), neighbouring sequence identity (Neigh, see Section 2.2.4.1), and global sequence identity (Global). Partial correlation coefficients are also given for the three types of sequence identity, using each of the other measures of sequence identity as a controlling variable. Results are shown for each non-redundant data set, and also for each of these sets after any helix pair with proline at the kink site was removed.

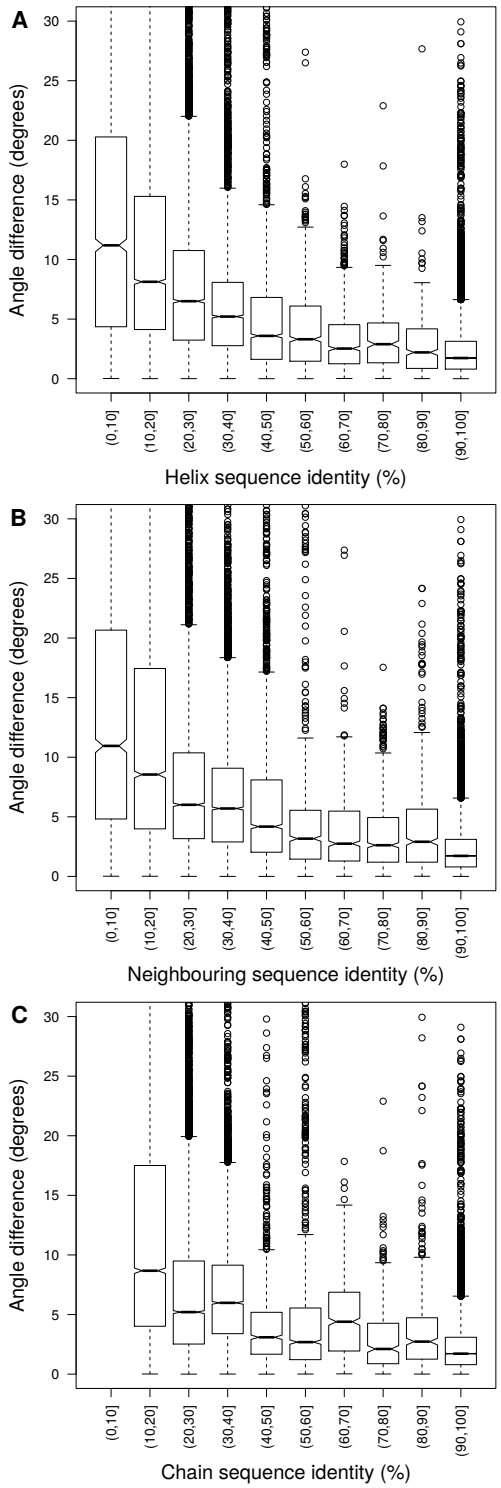


Figure 2.8: The difference in angle at the most disrupted site between helices in homologous helix pairs ($|\theta_{\max} - \theta_{\min}|$) plotted against the sequence identity between A) the homologous helix sequences, B) the residues in spatial proximity to the homologous helices (neighbouring residues) and C) the homologous chain sequences. Data from the redundant membrane protein set is shown so that the full range of sequence identity can be seen.

Group size	5	10	20	50	100
Membrane	45	18	0	0	0
Soluble	1258	473	170	28	4

Table 2.3: The number of helix families greater than or equal to each group size for the soluble and membrane protein sets.

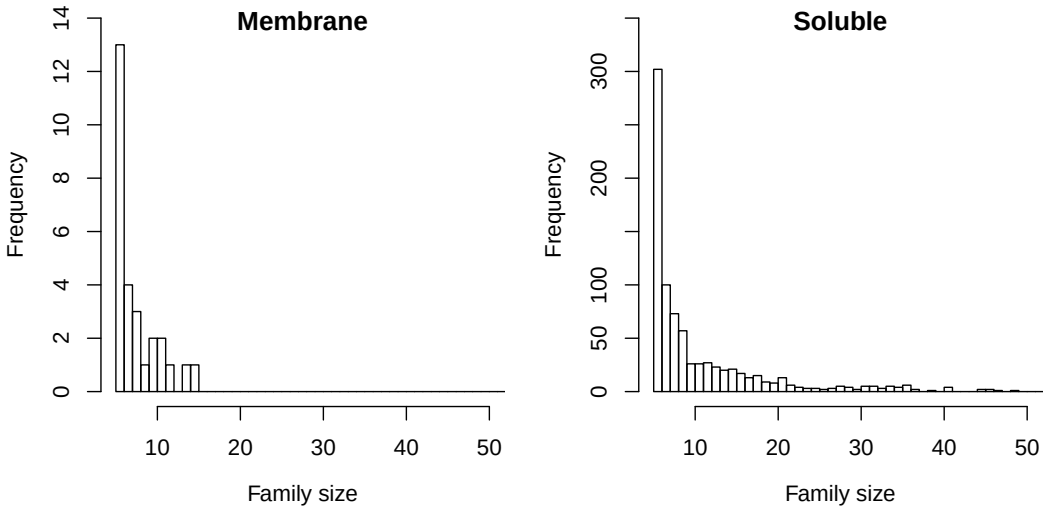


Figure 2.9: Distribution of sizes of families of at least five members.

2.3.4.2 Definition of family classes

Angles were measured for all members in a family and smoothed as described in Section 2.2.5. Each family was classified in an analogous way to the pair classification (Figure 2.4). Figure 2.10 shows an example of one family from each class. As with homologous helix pairs, Conserved Straight families are the most common class for both membrane and soluble families ($(19 \pm 6)/45$ and $(780 \pm 34)/1258$ respectively). In soluble proteins, Conserved Kinked families are rare (72 ± 16) compared to Not Conserved families (293 ± 30), but in membrane proteins they are equally common (12 ± 6 of each). Confidence intervals were estimated by the same bootstrap method described for the helix pairs.

2.3.4.3 Prevalence of proline in different family classes

The relationship between prolines and kinks in the homologous helix families gives similar patterns to those seen for pairs (Figure 2.11). If proline is found

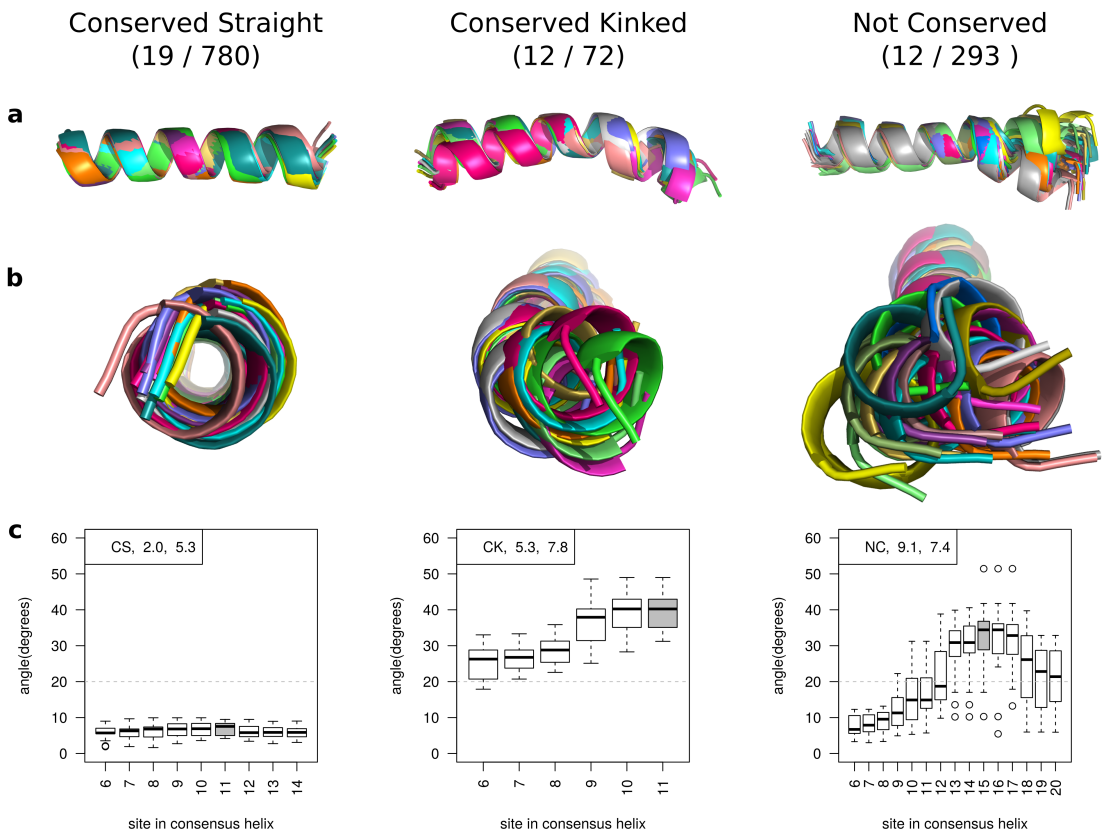


Figure 2.10: Illustrations of a homologous helix family from each of the three main classes. The number of families in each class is shown in brackets (membrane set / soluble set). 113 out of 1,258 soluble families and 2 out of 45 membrane families were classified as 'Other'. **(a)** and **(b)** Side and top view of helices in a family superimposed by aligning the residues prior to the kink site. **(c)** Boxplots to show the variation in angle after smoothing at each site in the helix. The grey box indicates the most disrupted site used to classify the helix (see Section 2.2.5). In the top left of each graph, the classification, $s(\theta)$ (standard deviation of angles), and $\bar{\varepsilon}$ (mean error) of the most disrupted site is given.

at the most disrupted site or the four following residues in every member of a family, it is a good indicator of kink conservation, as it was for helix pairs. Of the 10 membrane and soluble families where proline is fully conserved, 2 are classified as Not Conserved. For the 267 membrane and soluble helix families where proline is present in some but not all members, families are more frequently Not Conserved (175 ± 16) than Conserved Kinked (43 ± 12). Thus, once again proline conservation does not equal kink conservation in every case, and proline loss may or may not equal kink loss.

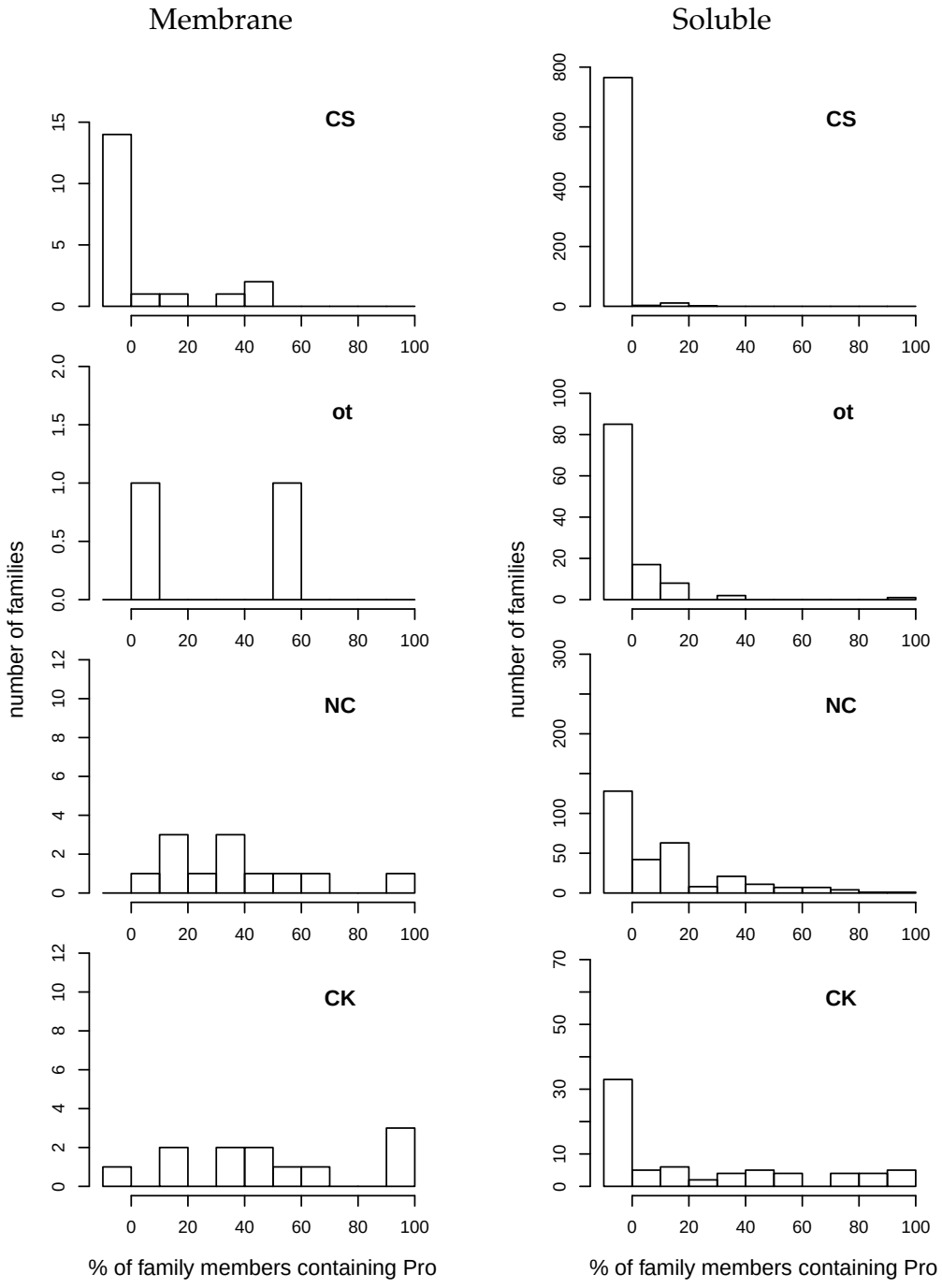


Figure 2.11: The percentage of family members containing proline at the most disrupted site in the helix family or in the four following residues, broken down by family class. The left-most bar represents families with no proline. CS: Conserved Straight, ot: Other, NC: Not Conserved, CK: Conserved Kinked.

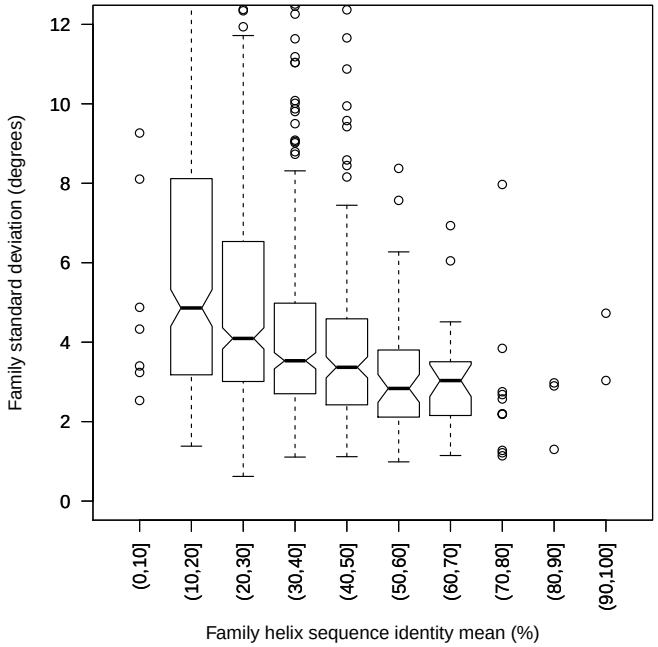
2.3.4.4 Relationship between angle variation and sequence conservation

In an analogous way to that used for pairs, I analysed the relationship between kink angle conservation and sequence conservation. I calculated the chain "sequence identity" for a family as the mean sequence identity between every pair of chains in the MAMMOTH-Mult alignment, ignoring gaps. Helix sequence identity was calculated in the same way using just the consensus helix positions. The relationship between family angle variation and sequence conservation (Figures 2.12, 2.13) is similar to that seen for pairs (Figure 2.8). There are very few data points at the higher end of the sequence identity range, even when combining all soluble and membrane data, as the family detection method led to most families containing at least some distant homologues. The partial correlation coefficient for helix sequence identity, given chain sequence identity as a controlling variable, is -0.16 and that for chain sequence identity, given helix sequence identity, is -0.11. This reinforces the suggestion found with the pairs that local and global sequence changes are both associated with kink angle changes.

2.3.5 G-protein coupled receptor case study

As a specific application of the methodology, I have carried out a detailed study of the transmembrane helices (TMHs) of GPCRs. The kinks in some GPCR helices are thought to be important for function (Katritch *et al.*, 2013), and many GPCR structures are available (Venkatakrishnan *et al.*, 2013). The structural alignment from the GPCRDB (Isberg *et al.*, 2014) was used to construct the smoothed angle profiles shown in Figure 2.14. The most disrupted site in each helix (see Section 2.2.5) is shown in grey, and the standard deviation and classification of each helix is given. Throughout this section, I refer to specific positions using the GPCRDB Class A numbering system, which is based on a structural alignment of all structures in the GPCRDB. The system agrees as far as possible with previous numbering based on a conserved residue in each helix, where the first number is the number of the helix; the second number is the

A



B

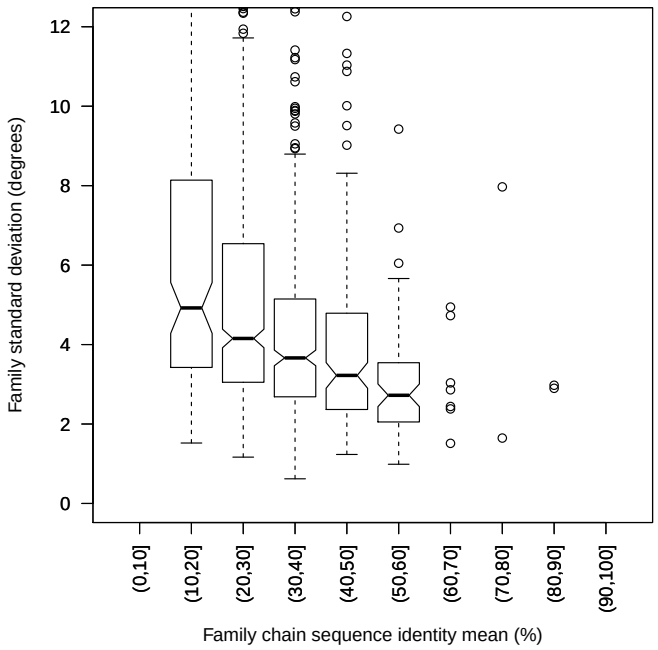


Figure 2.12: The standard deviation of angles at the most disrupted site in a family plotted against the family mean sequence identity between A) all pairs of homologous helix sequences and B) all pairs of homologous chain sequences. Data from the non-redundant membrane and soluble protein sets is combined, as the membrane set is small but appears to show a similar distribution to the soluble set (Figure 2.13).

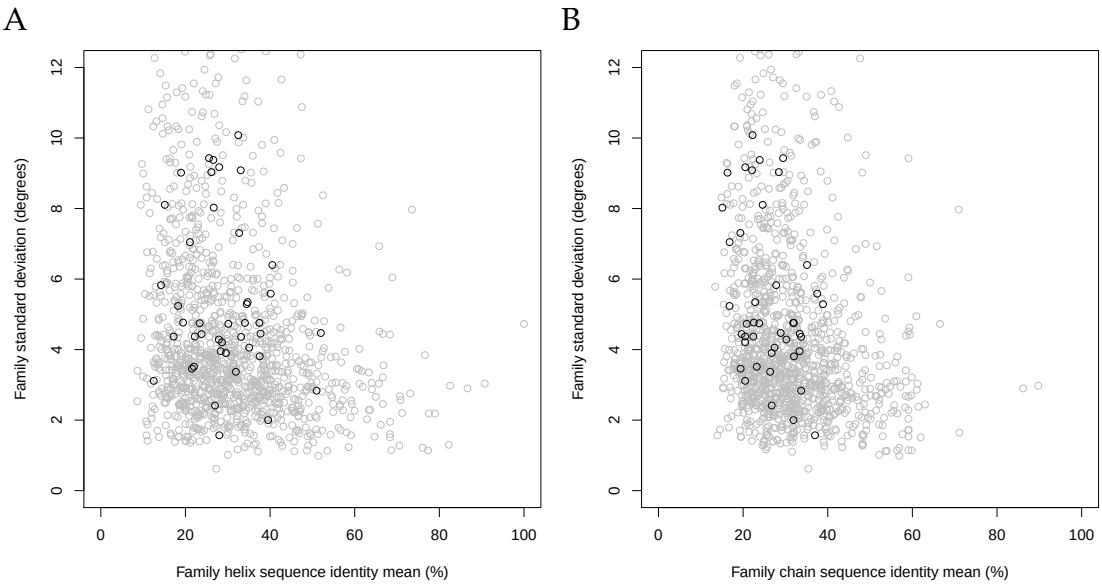


Figure 2.13: The standard deviation of angles at the most disrupted site in a family plotted against the family mean sequence identity between A) all pairs of homologous helix sequences and B) all pairs of homologous chain sequences. Data from the non-redundant membrane (black) and soluble (grey) protein sets.

position in the helix relative to the conserved residue, numbered 50. However, the structure-based alignment allows insertions or deletions at alpha-bulges, therefore providing a more accurate alignment. The GPCRDB numbering uses an 'x' separator, in contrast to the '.' separator used by sequence-based numbering.

TMH 1 is generally straight, but displays a wide variation in kink angles around residue 1x43. TMH 2 has a kink at residue 2x55 that is present in almost all members, but shows a wide range of angles and is therefore classified as Not Conserved. TMH 3 is straight for almost all members of the GPCR family, however there is a small group whose members show a kink angle of up to 50° at site 3x28. TMHs 4 and 5 have a kink in most members at the most disrupted site, but like TMH 2, these kinks take a wide range of angles. TMH 6 and 7 are the only helices that have a kink classified as Conserved. They also have the largest kink angles (Figure 2.15a). In TMH 6, only one member does not have a kink at position 6x47, and most other members are tightly grouped around 40°, with a standard deviation across the family of 5.8°. There also seems to be a Not Conserved kink present in a small number of members around residue

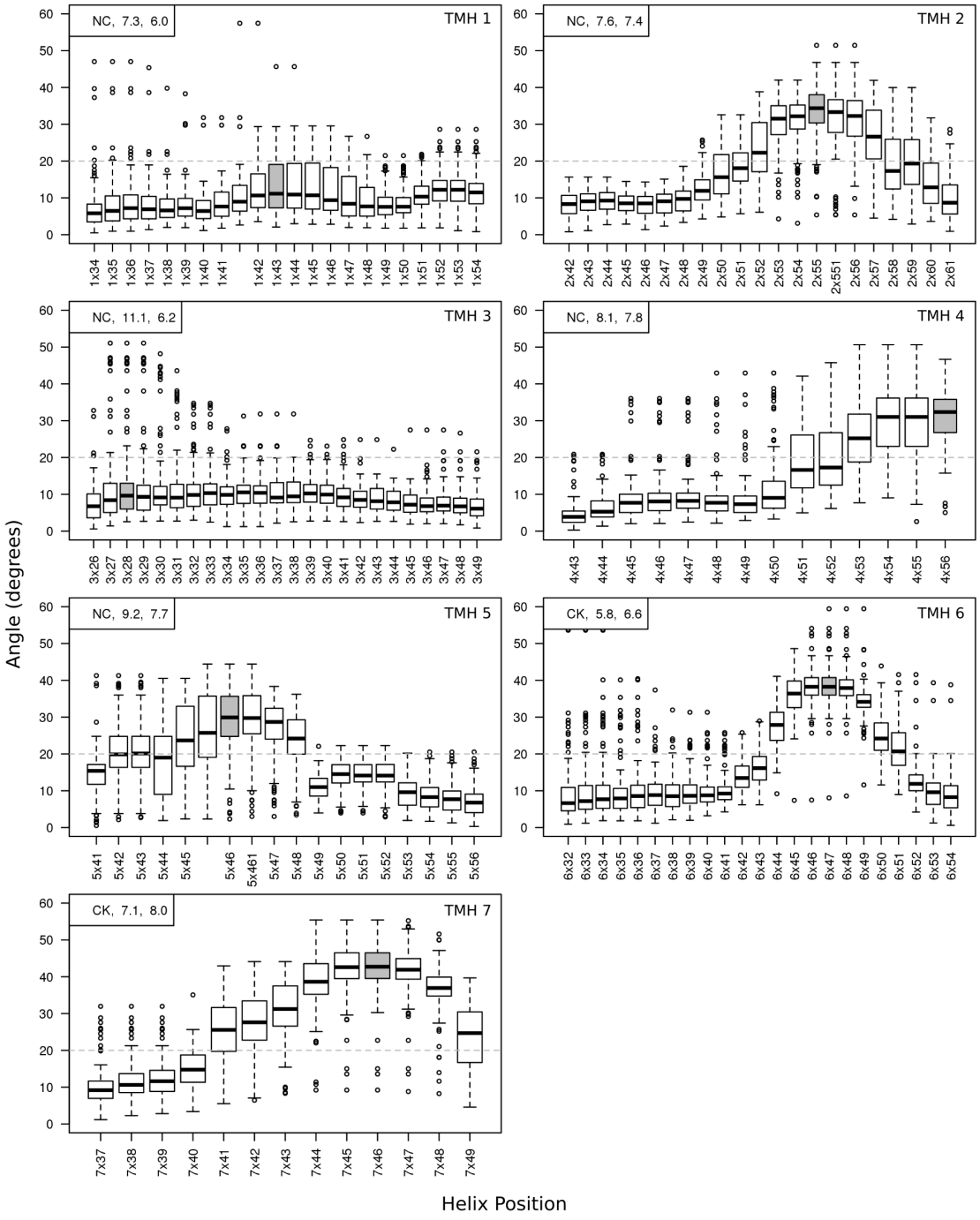


Figure 2.14: Distributions of angles measured at each site of the seven transmembrane helices in the GPCR family, after smoothing. The label at each site shown on the x-axis is the Class A numbering used in the GPCRDB (Isberg *et al.*, 2014). The broken grey line at 20° is the threshold for the definition of a kink. The most disrupted site in each helix (see Section 2.2.5) is shown in grey. In the top left of each graph, the classification, $s(\theta)$ (standard deviation of angles), and $\bar{\varepsilon}$ (mean error) of the most disrupted site is given.

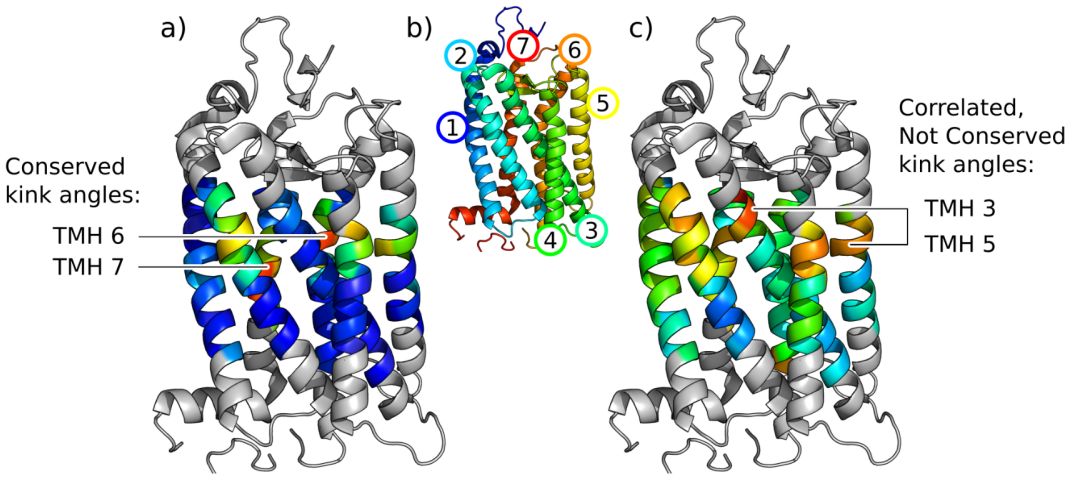


Figure 2.15: GPCR kink angle variation shown on the PDB structure 1F88, chain A. Colouring is by a) mean and c) standard deviation of angles at each site in the GPCR family, on a spectrum from the lowest values in blue to the highest in red. Grey residues have no angles measured as they are loop regions or within 6 residues of the end of the consensus helix (the minimum for a cylinder fit). b) is coloured by rainbow from N-terminus (blue) to C-terminus (red). The Conserved kink angles in TMH 6 and TMH 7 and the correlated kink angles at sites in TMH 3 and TMH 5 are labelled.

6x34, near the N-terminal end of the helix. The kink in TMH 7 at 7x46 shows a similar profile to TMH 6, however the standard deviation is slightly higher at 7.1°, though this is still less than the average error in this family of 8.0°.

Figure 2.15 shows the average size of angles measured at each site and their conservation across the family, presented on a structure of rhodopsin.

2.3.5.1 Angle variation relationship to sequence or flexibility

In the GPCR set, multiple structures were available for some receptors. This made it possible to observe variation in angles for an individual receptor. I could also compare the distribution of angles for one receptor to the distributions of others. For example, in the Not Conserved TMH 1 at residue 1x43, the distribution of angles for rhodopsin was separated from most other GPCRs (Figure 2.16A). The kink at 1x43 is in a functionally significant region and present in just a few GPCRs (Langelaan *et al.*, 2013). Rhodopsin has a proline at residue 1x48, and in 30 of the 32 rhodopsin structures in the set, the angle at 1x43 is $> 16°$. There are 14 structures of 8 other GPCRs which also have angles

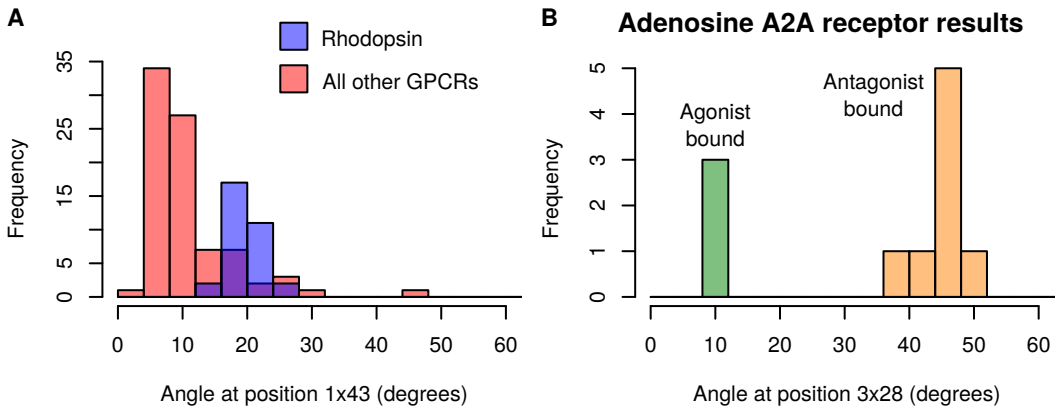


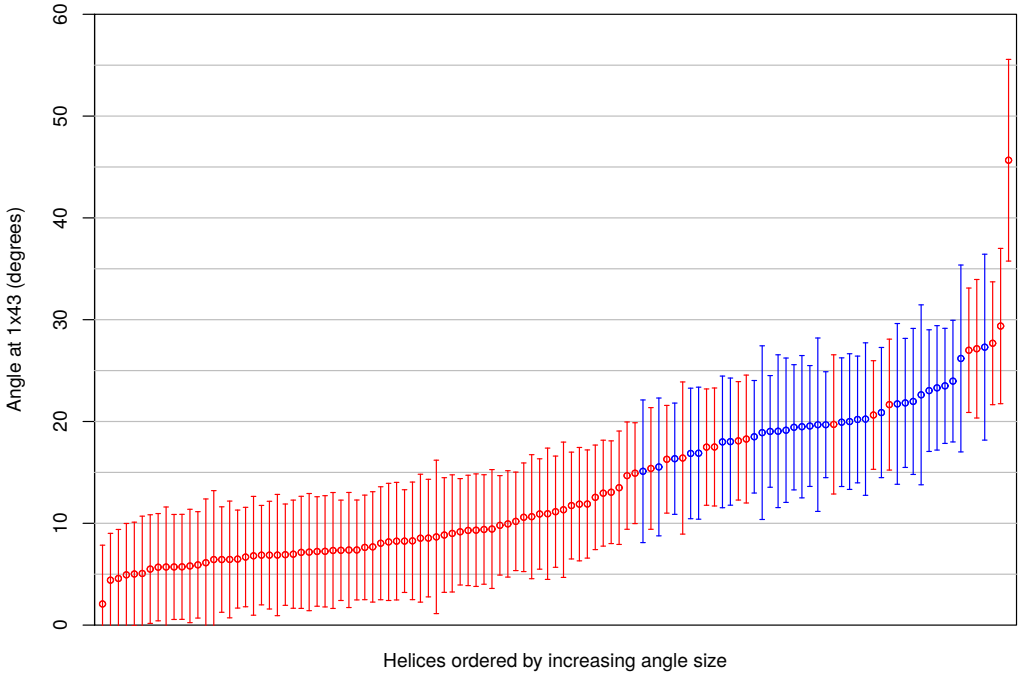
Figure 2.16: Bimodal angle distributions. **A)** Angle distribution at position 1x43 in all GPCR structures. Angles from rhodopsin structures are shown in blue (n=32); angles from all other structures shown in red (n=83). **B)** Angle distribution at position 3x28 in the human adenosine A_{2A} receptor. Agonist-bound receptors are shown in green (n=3); antagonist-bound receptors in orange (n=8). Figure 2.17 displays the errors for the angle data from both histograms.

$> 16°$. Three of these GPCRs have proline near the kink, but there is no obvious sequence causing the other five to be kinked. Proline is not present near this location in any of the non-rhodopsin structures with angles $< 16°$.

The full set of GPCR structures was also separated based on the type of ligand bound, but there was no difference between agonist, antagonist and inverse agonist structures overall. However, there were four receptors for which at least ten structures were available in the GPCRDB, so comparisons could be made between the different activation states of these individual receptors. One of these, the human adenosine A_{2A} receptor, displayed an angle change of over 30° at the most disrupted site in TMH 3. This site has the highest standard deviation of any angle in any of the seven GPCR helices (coloured red in Figure 2.15c).

In the case of the A_{2A} receptor, the angle change is from straight in the agonist-bound structures to kinked in the antagonist-bound structures (Figure 2.16B). This suggests that the change in angle is involved in the conformational change which occurs on activation of the receptor. The kink location in the helix is at the binding site for the natural ligand, therefore in this case its flexibility seems to be particularly important for the function of the receptor. This change of helix shape has previously been described qualitatively (Liu *et al.*, 2012).

A



B

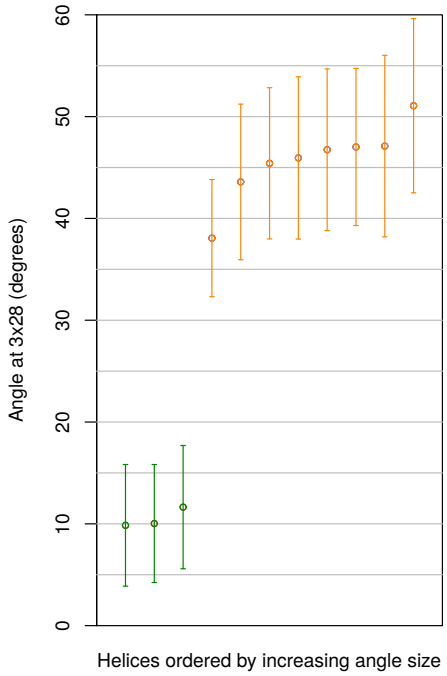


Figure 2.17: Angles from the histograms in Figure 2.16 ordered by magnitude and shown with their estimated 95% confidence intervals ($\pm\varepsilon$). **A)** Angles at position 1x43 in all GPCR structures. Angles from rhodopsin structures are shown in blue; angles from all other structures in red. **B)** Angles at position 3x28 in the human adenosine A_{2A} receptor. Agonist-bound receptors are shown in green (n=3); antagonist-bound receptors in orange (n=8).

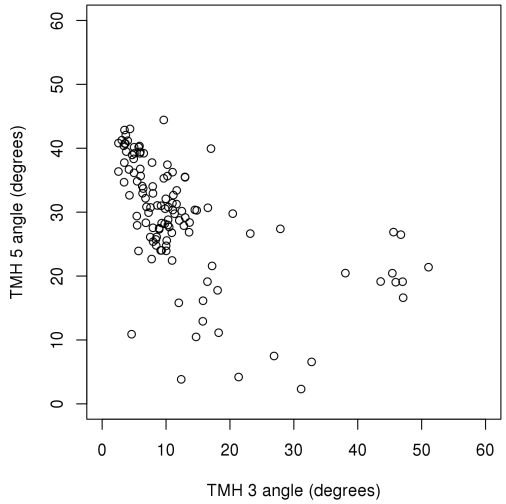


Figure 2.18: Correlation between the magnitude of angles at position 3x28 (TMH 3) and 5x46 (TMH 5) observed in all GPCR structures.

2.3.5.2 Correlation between kink angles

It is also possible to identify correlations between the kink angles in different helices. These could suggest concerted motion or interaction between the helices, where the kinking of one helix affects the structure of the other.

An example in GPCRs is TMH 3 and TMH 5, where the Spearman's rank correlation coefficient is 0.68 (Figure 2.18). The sites of these kinks are slightly separated in the structure with TMH 4 between them (Figure 2.15c). There is a weaker correlation between the angle at these sites and the TMH 4 kink: 0.31 for TMH 3/4 and -0.35 for TMH 4/5. These correlations suggest that change of conformation in one helix can influence the conformation of another.

2.4 Discussion

Using the error estimation method of Kink Finder, I have been able to compare helices and state whether their angles are different. Estimated on the basis of the quality of fit of cylinders to the helix either side of the kink, errors are usually between 5–8°, and tend to be slightly larger for larger kink angles.

Kink Finder uses a method of fitting cylinders to either side of the kink, which has the advantage of reliably finding changes of direction even in the

presence of non-canonical hydrogen bonding regions. This method requires a helix of 12 residues or more, but it is known that longer helices are more frequently kinked (Wilman *et al.*, 2014b). This results in a single metric, the kink angle, which allowed us to understand the error distribution and state the statistical significance of a difference in angles. There are other aspects of kink geometry such as the swivel angle which may not be conserved between the "Conserved Kinked" helices, therefore it is likely that kink conservation is overstated by these classifications.

In this work, for an overview of kink conservation, I chose to classify each helix pair or family using one most disrupted site. This avoids biasing the results, however it represents a simplification of the more detailed information shown in Figure 2.14. As I show for GPCRs, analysing all positions in a helix of interest reveals more about the system. In order to facilitate such analysis, these data are available for all of the helix families in the online supplementary data package (Law *et al.*, 2016).

Using the error estimation method, I have shown that kinks are not always conserved across structural homologues, i.e. they have significantly different angles. The different conformations seen in a pair or family can be explained by two possibilities, or a combination of both:

- The differences in sequence between two related proteins cause them to adopt different conformations.
- There is conformational flexibility at the kink, which could be important for function.

An investigation of the sequence dependence of the changes in angle provides evidence in support of the first option. The exact sequence drivers for kinks remain elusive. Even the loss of proline, the residue most closely associated with kinks, is associated with a significant reduction in kink angle in only 45% of cases. Changes in conformation would be important to understand when modelling a homologue with no proline where the template has a proline kink, or vice versa.

More generally, changes in angle are associated with changes in sequence. This relationship between kink conservation and sequence conservation is similar, whether considering only the residues in the homologous helices themselves, the residues in spatial proximity to the homologous helices, or the complete chains of the homologous proteins. Current kink modelling approaches assume that global effects are important for kink formation, especially for predicting the size of kinks (Werner and Church, 2013; Chen *et al.*, 2014). At the same time, sequence predictors make accurate kink predictions based on the primary sequence of the helix alone (Meruelo *et al.*, 2011; Langelaan *et al.*, 2010). My results suggest that both local and global factors are indeed independently important, but that one usually accompanies the other.

In my in-depth analysis of GPCRs, I also found evidence that kinks can significantly change conformation within a single protein. In the case of the human adenosine A_{2A} receptor, there were multiple structures in the data set, some with antagonists bound and others with agonists. Previous analysis of these structures reported a change in TMH 3 from kinked to straight between the two states (Liu *et al.*, 2012; Lebon *et al.*, 2011; Xu *et al.*, 2011). A rearrangement of water molecules in the binding site is associated with the conformational change (Liu *et al.*, 2012). I quantify the change from kinked to straight as an angle change of 30° and I verify the statistical significance of the difference between the inactivated and activated conformations. The conformational flexibility of this kink therefore appears to be functional.

I also found an example of strong correlation between the kink angles of two different helices across all GPCRs. This is consistent with the theory that it is the environment of a helix and not just its sequence which influences kink angle. I have shown this in the context of homologues, but it would also be interesting to see whether a similar effect can be observed in one protein through the course of a molecular dynamics trajectory.

Kinks are one of the possible deformations of a helix; others such as bulges and twisting may be equally functionally relevant. In the future, the examination

of all distortions within a single framework could improve our understanding of the link between structure and function.

Therefore, my investigation has shown significant and widespread variation of kink conformations which could point to their flexibility, and my results support the belief that kinks are important for functional changes of conformation.

In the next chapter, I will move from the analysis of local helix motifs to global features of alpha-helical membrane protein structures. The statistics calculated from whole proteins or segments of proteins will be used to explore whether these proteins may fold cotranslationally.

3

Evidence for cotranslational folding in membrane proteins

3.1 Background

Alpha-helical proteins are inserted into the membrane during the process of translation (Rapoport, 2007; Voorhees *et al.*, 2014; Ismail *et al.*, 2012), but it is not yet known to what extent the three-dimensional tertiary structure of a membrane protein takes shape during this time (Cymer and von Heijne, 2013). In this chapter, four different computational strands of evidence are presented to support the theory that alpha-helical membrane proteins fold cotranslationally. The fourth test, comparing the accuracy of the different modes of SAINT2 for membrane proteins, forms part of a paper of which I am a co-author (de Oliveira *et al.*, 2017a).

In this background section, I describe our current understanding of alpha-helical membrane protein folding, some computational measures for observing the potential for cotranslational folding in soluble proteins, *de novo* structure prediction methods in membrane proteins.

3.1.1 Membrane protein insertion and folding *in vivo*

Almost all membrane proteins are inserted into the membrane as the process of translation occurs via a membrane protein complex, the translocon (Rapoport, 2007). In prokaryotes, the translocon is called SecYEG and is located in the plasma membrane, while the eukaryotic homologue is Sec61 and is found in the rough endoplasmic reticulum. Bacteria have an alternative pathway which inserts some proteins after translation, independently of Sec. This pathway is mediated by YidC and is only used by proteins with one or two transmembrane helices (TMHs) (Dalbey *et al.*, 2014). YidC can alternatively form a complex with Sec and assist with cotranslational insertion of proteins. A crystal structure of YidC shows that it has a positively charged hydrophilic groove across part of the membrane on one side of its helix bundle (Kumazaki *et al.*, 2014). This groove is accessible from the cytoplasm, and it is proposed that it helps negatively charged loops to cross the membrane. In the case of the Sec complex, hydrophilic loops are able to pass through the centre of the complex, which is in the shape of a pore. One side of the pore can open to allow TMHs to pass laterally into the membrane, as shown in a crystal structure of the prokaryotic complex (Tsukazaki *et al.*, 2008). For the eukaryotic Sec61 complex, there is a high-resolution cryo-EM structure in complex with the ribosome during the process of translation (Voorhees *et al.*, 2014) (Figure 3.1). The ribosome tunnel is not tightly sealed to the translocon, which allows interaction between the nascent chain and the interface of the membrane as well as Sec.

The structural evidence described above demonstrates that the complexes are arranged in a way that makes cotranslational folding possible, but there are also experimental techniques which demonstrate that translation and insertion do occur simultaneously. The process of cotranslational insertion has been followed by arrest peptides (APs), which are sensitive to tension in the nascent chain (Ismail *et al.*, 2012; Cymer and von Heijne, 2013; Cymer *et al.*, 2014; Ismail *et al.*, 2015). Usually, an AP forms a specific conformation in the ribosome exit tunnel, which interacts with the ribosome to cause attenuation of translation to occur. A

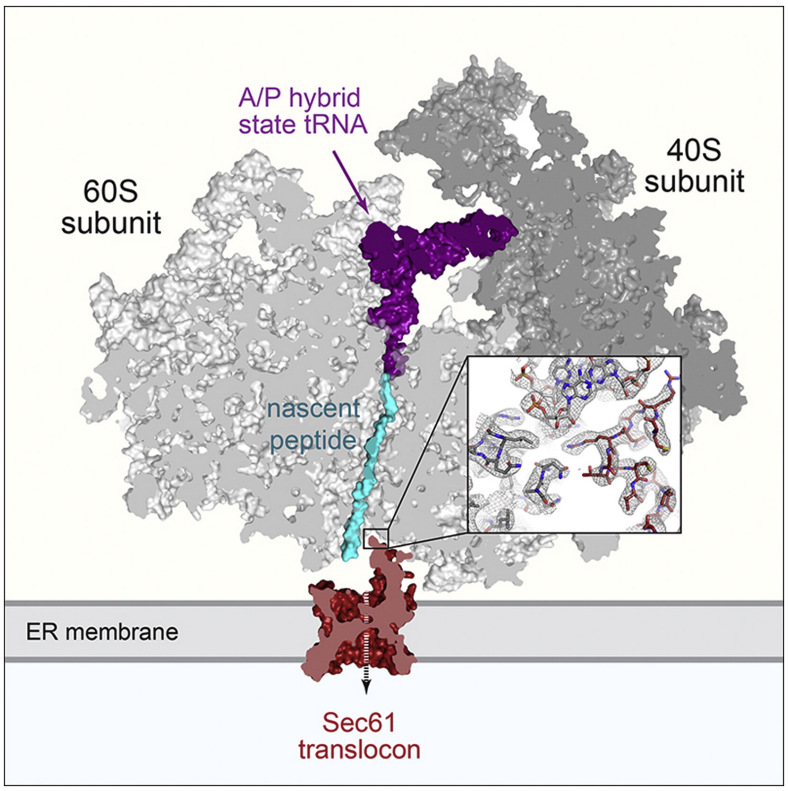


Figure 3.1: Structure of ribosome-Sec complex reproduced from Voorhees *et al.* (2014).

force on the nascent chain can prevent this, with a stronger pulling force leading to production of more full length peptide. Inserting a TMH into a construct shortly before an AP reduces attenuation, therefore tension is generated by insertion of the TMH into the membrane during translation. The effect on attenuation is dependent on the length of the linker between the TMH and the AP (Figure 3.2), and the hydrophobicity of the TMH. The force is greatest when the length of the linker between the TMH and the AP corresponds to the length which allows insertion of the TMH. The free energy change of the insertion of the TMH is negative, therefore this transition generates a force pulling on the nascent chain.

The traditional model for the folding of alpha-helical membrane proteins involves two stages (Popot and Engelman, 1990). The first is insertion of individual TMHs into the membrane, and the adoption of the correct secondary structure. This is then followed by rearrangement of the TMHs by lateral diffusion in the membrane, which results in the correct tertiary fold.

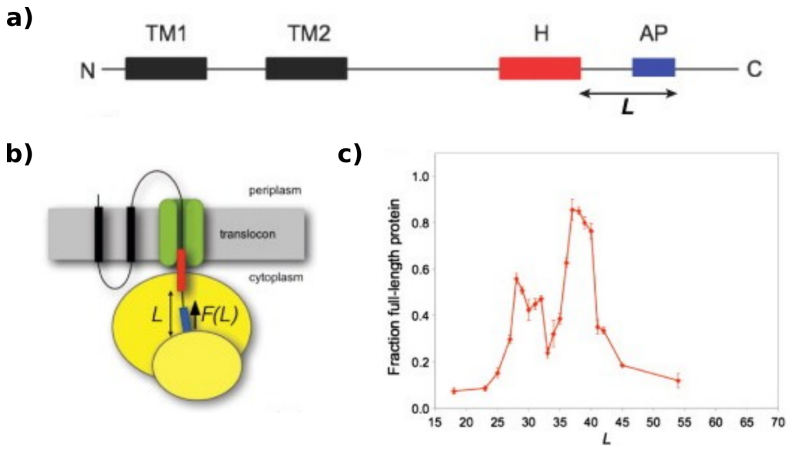


Figure 3.2: a) A construct featuring a transmembrane helix (H) and arrest peptide (AP), with a linker length L between them. b) Cartoon showing the position of H as the AP is translated, and the force $F(L)$ generated by H. c) Fraction of full length protein plotted against L. The greater the force, the greater the fraction of full length protein, as greater tension in the nascent chain reduces the probability of attenuation brought about by the AP. The peak at $L \approx 40$ is caused by the force generated from the favourable insertion of H. The earlier peak at $L \approx 30$ corresponds to an interaction between the N-terminal residues of H and Sec. Adapted from Cymer *et al.* (2015).

However, there are several lines of evidence which appear to contradict this model. Some TMHs are only mildly hydrophobic with charged or polar residues in the central layer of the membrane (De Marothy and Elofsson, 2015). In the complete native structure, such hydrophilic residues would usually be oriented towards the centre of the helix bundle, but in a single inserted TMH they will be exposed to hydrophobic lipid tails. TMHs containing such hydrophilic residues may not have a favourable ΔG of insertion into the membrane on their own, therefore they require interactions with other TMHs. For example, in the plant potassium channel KAT1, insertion of the S3 and S4 helices depends on interactions with helix S2 (Sato *et al.*, 2003). Mutations showed that negative charges in S2 interact with positive charges in S4 during simultaneous insertion of S3 and S4 together. Replacement of helix S2 with a homologous helix from the Shaker potassium channel prevented the correct folding and insertion (Zhang *et al.*, 2007).

TMHs can even interact before reaching the membrane. A crosslinking study in the related potassium channel Kv1.3 showed that a helix hairpin formed in the

folding vestibule of the ribosome (Tu *et al.*, 2014). Interactions during insertion are not just between TMHs, but also with the Sec complex. Crosslinking studies indicate that there is extensive interaction between inserted TMHs and with the translocon during the process of translation of an aquaporin (Sadlish *et al.*, 2005). Cryo-electron microscopy has provided structures of Sec with TMHs bound to the lateral gate location (Bischoff *et al.*, 2014). These interactions could provide ways to stabilise TMHs which would otherwise be unable to insert, and it has been proposed that TMHs may be inserted pair-wise during translation (Cymer *et al.*, 2015).

It is uncertain for what fraction of membrane proteins these folding mechanisms are important. Few proteins have been studied in detail, therefore the prevalence of inter-helix interactions and interactions between inserted TMHs and the Sec apparatus is not known. It is also uncertain how permanent the topology of a TMH is, once it has been inserted, as topology of individual TMHs can be reversed after insertion (Lu *et al.*, 2000). In one case it was observed that the orientation of an entire protein was changed when a single charged residue was added at the C-terminal end of the protein (Seppälä *et al.*, 2010). This evidence indicates that not only the tertiary structure, but also topology may be changed late in the folding process.

The experimental methods used on specific proteins are difficult to scale up and test on large sets. This makes it difficult to assess whether the cases investigated so far are exceptions, or typical of the majority of alpha-helical membrane proteins. To gain a wider perspective on their relevance, I have used computational methods to investigate the importance of cotranslational folding in membrane proteins.

3.1.2 Computational measures of cotranslational folding

Computational studies of cotranslational folding have previously been carried out for soluble proteins. Several methods analyse existing structures to determine whether they are more likely to be built up in the biologically relevant

direction of the N- to C-terminus than the reverse (e.g. Deane *et al.*, 2007; Saunders *et al.*, 2011). A number of "measures of cotranslation" can be calculated, which look for expected patterns such as an N-terminus which is more buried. In soluble proteins, these measures suggest that cotranslational folding may occur. For example, one measure, SLR (described in Section 3.2.2) shows that the distribution of contacts along a chain is not symmetrical for the α/β class (Deane *et al.*, 2007). In this chapter, three of the best characterised of these measures are evaluated for two sets of membrane proteins.

Another test for potential cotranslational folding is the use of fragment-based *de novo* structure prediction to build structures directionally. Fragment methods traditionally start with a fully extended complete protein (e.g. Rosetta, Raman *et al.*, 2009, and FRAGFOLD, Kosciolek and Jones, 2014). They then propose a substitution for a section of the protein chain with a conformation taken from a fragment library. The proposed move is accepted or rejected, with a probability determined by the difference between the scores of the current and proposed structures. After thousands of moves have been carried out, a decoy (potential structure) is generated. Typically thousands of decoys are generated by these methods and the highest scoring decoys, or representatives of the most popular clusters, are selected. Ellis *et al.* (2010) proposed a method adapted from Rosetta that built structures sequentially from the N- to C-terminus rather than starting from a fully extended chain. Its performance was similar to Rosetta, which is quite remarkable when considering the change from a global search to a more limited and greedy local search. The reverse direction, starting at the C-terminus, resulted in a lower mean accuracy of decoys in 64 out of 68 targets.

Our group has developed the program SAINT2 (de Oliveira *et al.*, 2015), which is an implementation of a cotranslational protein structure predictor, using the same principle as Ellis *et al.* (2010). SAINT2 has Forward, Reverse and In vitro modes, where In vitro is most similar to Rosetta in that it begins from a complete and fully extended chain. The Forward (cotranslational) mode starts at the N-terminus and adds to the chain while the Reverse mode starts at the

C-terminus (see Section 1.4.4.3 for a full description of the method). In soluble proteins, The Forward mode outperforms both the In vitro and Reverse modes.

The results from soluble structure prediction indicate that fragment-based decoy generation may be able to improve results by imitating the biological process of folding. They also indicate that cotranslational folding is relevant in a biological context for soluble proteins. In this chapter, I test SAINT2 on membrane proteins, and find a similar difference between the modes, suggesting that membrane proteins fold cotranslationally.

3.1.3 *De novo* membrane protein structure prediction

While cotranslational approaches have been used to predict the structures of soluble proteins, none of the current *de novo* membrane protein modelling methods explicitly use this strategy (see Section 1.4.4 for a full description). Leading methods start from an extended and complete protein chain, and conformational sampling is guided by contacts predicted from correlated evolution of residues (e.g. Ovchinnikov *et al.*, 2016; Nugent and Jones, 2012; Hopf *et al.*, 2012). These methods achieve some excellent predictions, but they do not attempt to imitate the biological folding pathway. By imitating biological folding, I aim to improve the accuracy of structure prediction. By investigating which changes to a structure prediction program achieve the best results, I also hope to find out which aspects of folding may be important *in vivo*.

3.1.4 Outline

This chapter lays out the evidence for cotranslational folding of membrane proteins using four approaches:

- Measures of cotranslational folding were calculated for two sets of membrane proteins and suggest a slight cotranslational bias.

- N- and C-terminal segments were extracted from membrane proteins and comparison of their scores after relaxation by RosettaMP showed that the N-terminal segments were more stable on average.
- Pairs of TMHs adjacent in sequence extracted from membrane proteins interacted more strongly if they were connected by an extracellular loop, compared to those connected by an intracellular loop.
- Protein structure prediction using SAINT2 produced more accurate decoys in the Forward mode than in the In vitro mode.

3.2 Methods

3.2.1 Membrane protein sets

Two non-redundant sets of alpha-helical membrane proteins with at least four transmembrane helices (TMHs) each were used in this chapter. Set1 is a set of sequence dissimilar chains, while Set2 was selected from the Orientations of Proteins in Membranes (OPM) database (Lomize *et al.*, 2006).

For Set1, the redundant set of 844 membrane protein chains from Chapter 2, culled to 99% sequence identity, was used as a starting point. The membrane layer of each residue was annotated using iMembrane (Kelm *et al.*, 2009), to establish which residues were in the lipid tail (T) layer of the membrane. Based on the distribution of lengths for stretches of consecutive T layer residues (Figure 3.3A), any stretch of 15 consecutive residues in the T layer of the membrane was defined as a span. Chains with fewer than four non-overlapping transmembrane spans were removed leaving 534 chains (the distribution of span counts is shown in Figure 3.3B). PISCES (Wang and Dunbrack, 2003) was used to cull this set at 20% sequence identity, resulting in a set of 93 chains.

To create Set2, I started with all polytopic alpha-helical transmembrane chains in the OPM database. I kept only continuous chains that had no gaps the crystal structure. A maximum of one chain was kept from each of the OPM families, taking the PDB with the best resolution if there was

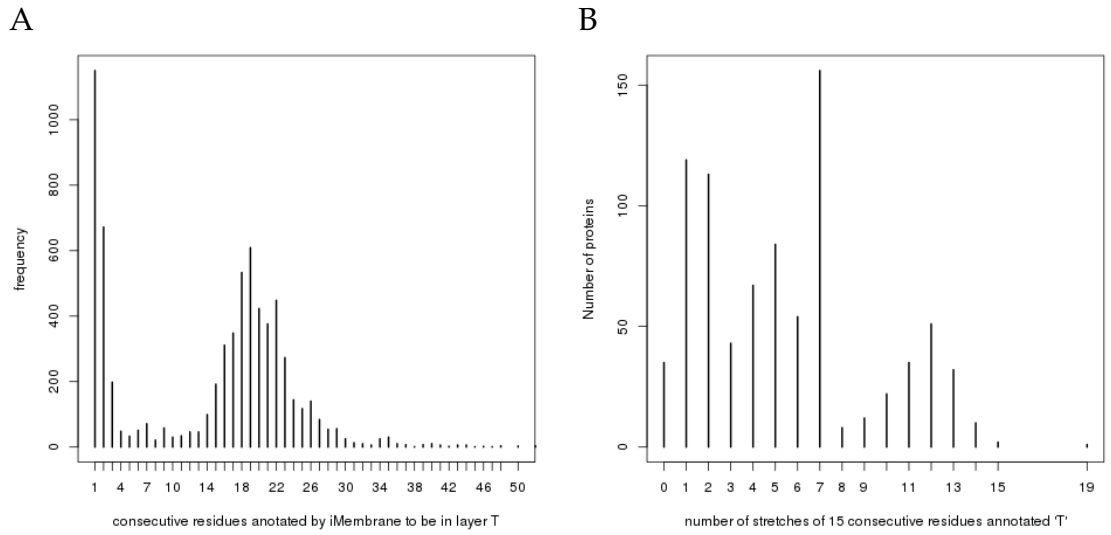


Figure 3.3: A) The length distribution for Set1 of stretches of consecutive residues annotated by iMembrane to be in the tail (T) layer of the membrane. B) The distribution of the number of spans in a protein for Set1, taking a span to be a stretch of 15 consecutive residues in the T layer of the membrane.

more than one X-ray structure. If there was no X-ray structure for a protein, the first NMR structure was used. This set was also culled using PISCES (Wang and Dunbrack, 2003), allowing a maximum sequence identity between chains of 20% to remove homologous chains within a single PDB file. The structures were inspected individually to remove any chains with fewer than four transmembrane spans, leaving 55 chains in Set2. Any chains which visibly included a soluble domain were removed, together with any chains that were not clearly a single transmembrane fold, for example if one TMH was separated from the main bundle in the structure. Of the remaining 39 proteins, I took the shortest 24 to create Set2A. These structures range in length from 132–385 residues.

Nine of the 39 proteins, evenly distributed across the range of lengths, were set aside to be a test set which would not be used for the training of scoring functions for SAINT2. Five of these test proteins were in Set2A ($\text{Set2A}_{\text{test}}$), so 19 short proteins made up the training set for scoring of SAINT2 decoys ($\text{Set2A}_{\text{train}}$).

For each of the above datasets, the PDB codes are given in Appendix A, together with the length, method of structure determination, resolution, R and R_{free}.

3.2.1.1 Topology

Three methods were used to annotate transmembrane spans of proteins at different stages: iMembrane (Kelm *et al.*, 2009), OPM spans (Lomize *et al.*, 2006), and the tool `mp_span_from_pdb` from the RosettaMP framework (Alford *et al.*, 2015), which uses the location of TMHs in the PDBTM (Kozma *et al.*, 2013) file. iMembrane inserts membrane proteins on the basis of sequence or structure homology to proteins which have been simulated in a membrane using coarse grained molecular dynamics simulations (Sansom *et al.*, 2008). I took strings of consecutive residues annotated to be in the tail layer of the membrane to be membrane spans, which resulted in a single span being assigned for any re-entrant loops. The OPM database predicts the membrane embedding of PDB entries using theoretical transfer energies of amino acids to the polarity profile of a membrane. OPM catalogues TMHs which span the membrane width reported for that protein. PDBTM predicts the insertion of membrane protein structures in a similar way to OPM. PDBTM provides PDB files with the coordinates transformed so that the protein is oriented optimally in a membrane with centre (0,0,0) and the z-axis as the membrane normal. Using the input of a structure already transformed into these membrane coordinates, RosettaMP's `mp_span_from_pdb` application uses DSSP (Kabsch and Sander, 1983) to locate helices in the membrane. In this application, residues between -15 Å to +15 Å are considered a span if the helix is at least three residues long, so they may not be complete spans. To include kinks and other distortions, helices separated by up to three loop residues are joined if they are oriented in the same direction.

The resulting annotations are compared for one membrane protein in Figure 3.4, demonstrating that there is some disagreement. In each section of this chapter, I tried to use the most appropriate annotation input for the task. OPM annotation was not available for all proteins in Set1 so iMembrane annotation was used, while the `mp_span_from_pdb` application was used to provide span information to RosettaMP.

```
iMem  NHHHHHHHHTTTTTTTTTTTTTTTTHHHHHHHHHHHHHHNNNNNNNNNNHNHHHHHHHHHTTTTTTTTTTTTTTTTTTHHHHHHHHHHHHHHHHTTT
OPM   LLSSSSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLSSSS
RosMP LLLLLLSSSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLSSSSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLSSSSS

iMem  TTTTTTTTTTTTTTHHHHHHHHHHHHHNNNNNNNNNNNNHHHHHHHHHTTTTTTTTTTTTTTTTTTHHHHHHHHHTTTTTTTTTTTTTTTTTTHH
OPM   SSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLSSSSSSSSSSSSSSSSSSLLLLLLLLLLSSSSSSSSSSSSSSSSSLL
RosMP SSSSSSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLLLLSSSSSSSSSSSSSSSSSSSSSSLLLLLLLLLLLLLLLSSSSSSSSSSSSSSLL

iMem  HHHHHHHTTTTTTTTTTTTTTTHH
OPM   LSSSSSSSSSSSSSSSSSSLLLL
RosMP LSSSSSSSSSSSSSSSSSSLLLL
```

Figure 3.4: Comparison between span identification methods. iMembrane (iMem), OPM, and RosettaMP (RosMP) annotations are shown for 4b4aA. iMembrane annotations are N = not in the membrane, H = lipid head layer, T = lipid tail layer; OPM and RosettaMP annotations are shown as S = transmembrane span, L = loop (any residue not in a span).

3.2.2 Statistical measures of cotranslational folding

For calculation of the following measures, I included only membrane spanning residues, as defined by iMembrane (Set1) or the OPM database (Set2), in order to assess the relative arrangement of TMHs only. Similar results were obtained when chains were truncated to the membrane spans and all loops between them.

I used three previously described statistical measures (Deane *et al.*, 2007; Saunders *et al.*, 2011) on the sets of membrane proteins to establish the presence of any potential cotranslational bias in the structures of the sets. The first two measures assess whether residues closer to the N-terminus are closer to the centre of a protein. This would be expected to be true under the hypothesis of cotranslational folding as these residues are folded first and therefore later residues could be folded around them. The third measure evaluates differences between the forward and reverse direction in terms of the number of contacts made by residues at each point along the chain. The measures are shown schematically in Figure 3.5.

3.2.2.1 Mean central residue (MCR)

The MCR (Saunders *et al.*, 2011) is a measure of where the η residues closest to the centroid of the protein occur along the sequence. Residues are weighted according to their distance to the centroid.

R_i gives the coordinates of the C_α atom of the ith residue along the protein chain of n residues. The protein's centroid, Z, is the mean position of all

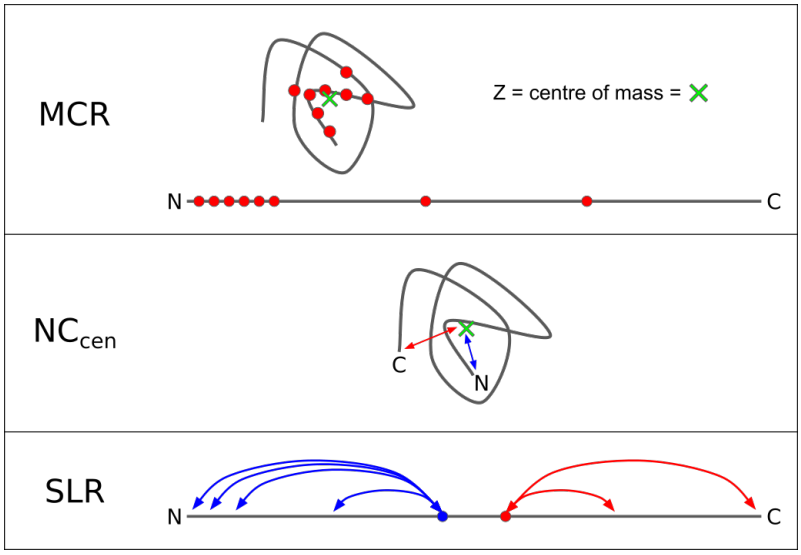


Figure 3.5: A conceptual visualisation of the statistical measures of cotranslational folding described in Section 3.2.2.

the heavy atoms in a chain of amino acids. The function $\delta(a,b)$ gives the distance between a and b. The function core(k) references the index of the kth closest residue to Z.

$$\text{MCR} = \frac{\sum_{i=1}^{\eta} \text{core}(i) W(i)}{n \sum_{i=1}^{\eta} W(i)} \tag{3.1}$$

where

$$W(i) = \frac{1}{\delta(R_{\text{core}(i)}, Z)} \tag{3.2}$$

A value of MCR < 0.5 indicates that the spatially central residues are nearer the N-terminus, which would be expected to occur more frequently in co-translationally folded proteins.

3.2.2.2 NC$_{cen}$

NC$_{cen}$ (Saunders *et al.*, 2011) indicates the difference in distance to the centroid between the two termini:

$$\text{NC}_{\text{cen}} = \log\left(\frac{\delta(R_1, Z)}{\delta(R_n, Z)}\right) \tag{3.3}$$

where n is the number of residues in the protein chain. A value of $NC_{cen} < 0$ indicates the N-terminus is closer to the centroid, which would be the expected bias in the case of cotranslational folding.

3.2.2.3 Sum of the log-transformed ratios (SLR)

SLR (Deane *et al.*, 2007) compares the number of contacts made to previous residues, starting from both ends of the chain. A_i^N denotes the number of previous contacts for a residue, defined as the number of residues from 1 to $i-6$ within 13 Å of residue i. A_i^C gives the equivalent number of contacts when numbering i from the C-terminal end. The measure calculates the ratio between A_i^N and A_i^C at each residue along the chain. Where either of these numbers is zero, the number of contacts is carried over and added to the total for the next residue, until the counts for N- and C-terminal ends under the grouping are both non-zero. i is the index of the ith group, which contains J_i residues, and I is the total number of groups. Logs are taken to map the range of possible values to the real line.

$$\mathrm{SLR} = \frac{1}{I}\sum_{i=1}^{l}\left(\log\left(\sum_{j=1}^{J_i} A_j^N\right) - \log\left(\sum_{j=1}^{J_i} A_j^C\right)\right) \tag{3.4}$$

Cotranslational folding would be expected to result in a greater number of positive SLR values.

3.2.3 Stability of segments of native structures

Set2, 55 non-redundant polytopic alpha-helical transmembrane chains, was used to look for the presence of semi-stable 'foldons' consisting of a subset of multiple transmembrane spans. All PDBTM files in the set were cleaned using Rosetta's `clean_pdb.py` script, which failed for two files, leaving 53 chains for analysis.

RosettaMP's `mp_span_from_pdb` application (Alford *et al.*, 2015) was used to locate each of the transmembrane spans in each protein chain from the PDBTM file (Kozma *et al.*, 2013). I extracted from each PDBTM file the p spans at the N-terminal end of the chain and their connecting loops, for integer values of

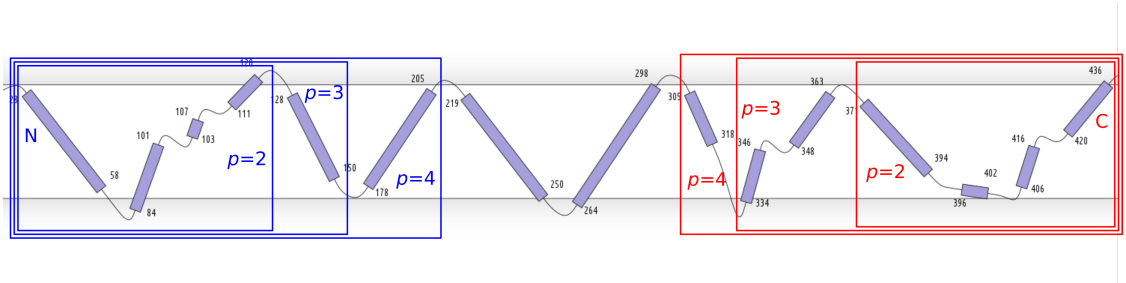


Figure 3.6: Illustration of the extraction of segments. Segments of two or more transmembrane spans (p being the number of spans) were taken from each end of the protein. The first three from each terminus are shown, and all other values up to one less than the total number of spans in the protein were also extracted. Topology map reproduced from Stansfeld *et al.* (2015).

p between 2 and $(q-1)$ inclusive, where q is the total number of spans in the chain (Figure 3.6). At the terminal end, and after the last included span, four further residues were included with the aim of adding hydrophilic residues to encourage the native membrane embedding. The equivalent segments at the C-terminal end were also extracted for a direct comparison.

3.2.3.1 MPrelax protocol

The MPrelax protocol (Alford *et al.*, 2015) was used to minimise the energy of each native structure segment, using as input the transmembrane spans located by the `mp_span_from_pdb` application. The MPrelax protocol is based on Rosetta's FastRelax (Tyka *et al.*, 2011), which carries out cycles of backbone perturbation, sidechain repacking and gradient-based local minimisation. The repulsive component of the score is increased from one cycle to the next. MPrelax first positions the centre of the membrane at the centre of mass of the transmembrane spans. Then it performs eight cycles of FastRelax, using an all-atom membrane score function (Barth *et al.*, 2007). The best (lowest) score from all of the cycles is kept, and the conformation given as output. This protocol was run 10 times for each segment.

The TM-score between each relaxed model and the segment from the native structure was calculated, enforcing the correct alignment between the two

structures. I used the following measure to compare the relaxed segments at the N- and C-terminus that have the same number of spans:

$$\Delta \bar{t} = \bar{t}_N - \bar{t}_C \tag{3.5}$$

where $\bar{t}_N$ is the mean TM-score for the 10 N-terminal models and $\bar{t}_C$ is the mean for the C-terminal models.

The RosettaMP scores indicate the relative energy of each conformation of a given system, and cannot be directly interpreted by comparing to scores from another structure. However, the scores obtained appeared to scale approximately with the length of the chain (Figure 3.7A), due to many of the terms in the scoring function being calculated per residue. Therefore it seems reasonable to compare the N- and C-terminal segment scores, with the following adjustment for length:

$$\Delta S = \frac{\bar{s}_N}{n_N} - \frac{\bar{s}_C}{n_C} \tag{3.6}$$

where $\bar{s}_N$ is the mean RosettaMP score for the 10 N-terminal models and $\bar{s}_C$ is the mean for the C-terminus; n_N and n_C are the numbers of residues in the N- and C-terminal segments respectively. This simple length normalisation is not perfect (Figure 3.7B), but it is an improvement. The overall distributions of the lengths of N- and C-terminal segments were similar, therefore differences in length should not lead to an overall bias.

3.2.4 Comparison of sequence adjacent pairs according to orientation

In order to compare TMH pairs that were joined by an intracellular loop to those joined by an extracellular loop, I used only chains from Set2 for which topology was annotated in the OPM. The spans defined by the OPM database were used to define TMHs. I extracted a set of 441 pairs of TMHs that were adjacent in sequence, and kept only the 278 pairs in which both TMHs were at least 18 residues long. The OPM PDB file contains transformed coordinates which embed the protein into a membrane with centre $z = 0$. Whether the

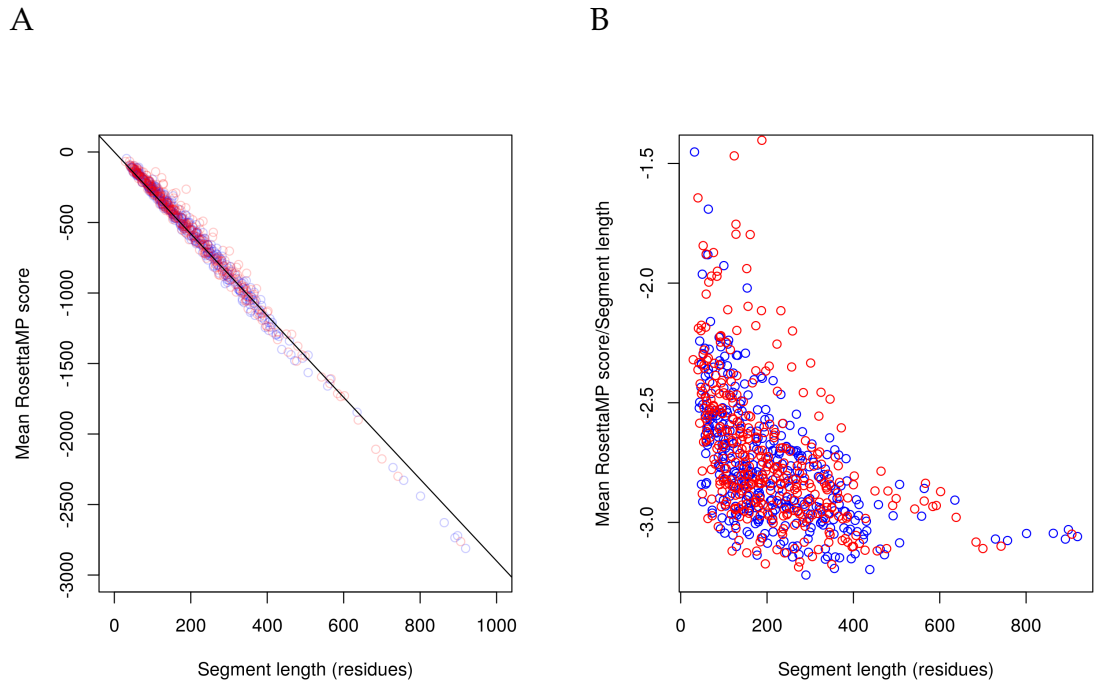


Figure 3.7: Length-dependence of the RosettaMP score. A) Mean RosettaMP score ($\bar{s}$) plotted against segment length (n) for all N- (blue) and C- (red) terminal segments. The black regression line has a fixed y-intercept of zero. B) Mean RosettaMP score divided by segment length $\left(\frac{\bar{s}}{n}\right)$ plotted against segment length (n).

extracellular side of the membrane was positive or negative was determined using the topology of the N-terminus annotated in the OPM and the z-coordinate of the first atom in the PDB file. A pair of TMHs was defined to be joined by an extracellular loop if the following were true: the last residue in the first TMH and the first residue of the second TMH had a C_α z-coordinate of magnitude > 8 on the extracellular side; the first residue of the first TMH and the last residue of the second TMH had a C_α z-coordinate of magnitude > 8 on the intracellular side. To be defined as a pair of TMHs joined by an intracellular loop, the opposite was required. This process resulted in 141 extracellular loop pairs, 117 intracellular loop pairs and 20 other pairs which fitted neither definition.

A measure of 'interaction strength' between two helices has previously been used to curate datasets of interacting pairs of helices (Zhang *et al.*, 2015). The reciprocal of the distance was calculated between every residue in the first helix and every residue in the second. If two residues were > 25 Å apart, this value

was set to zero. A window of n residues was chosen in each helix to maximise the average of inverse distances between them:

$$M = \frac{1}{n^2} \sum_{i=a}^{a+n-1} \sum_{j=b}^{b+n-1} x_{ij} \tag{3.7}$$

where M is the mean inverse distance (interaction strength), n is the window size ($n = 12$ was used here and by Zhang *et al.*, 2015), a and b are the starting residues of the window in each helix, and x_{ij} is the inverse of the distance in angstroms between the C_α atoms of residues i and j, or zero if they are > 25 Å apart. The maximum value of M was taken, after it was calculated for all values of a and b from 1 to $L - n + 1$, where L is the length of the particular helix.

3.2.5 Prediction of membrane protein structures by SAINT2

SAINT2 (de Oliveira *et al.*, 2015), is a fragment-based protein structure predictor developed in our group, described in detail in Section 1.4.4.3. SAINT2 has Forward, Reverse and In vitro modes, illustrated in Figure 1.13. The Forward and Reverse modes have two different kinds of step: extrusion and move. Extrusion inserts a random fragment at the growing end of the peptide (C-terminus in the Forward mode; N-terminus in Reverse) with one extra residue to elongate the chain. In a move step, a random location in the peptide is selected, a new fragment is proposed at that site, and the replacement is accepted or rejected depending on the score. When all residues have been extruded, further move steps are performed to generate the final decoy. The In vitro mode is similar to all other fragment-based structure predictors. It begins from a complete and fully extended chain and performs move steps to build a model. The SAINT2 score is similar to most other fragment-based structure predictors, containing all the standard elements including contact predictions (Jones *et al.*, 2015), RAPDF, Lennard-Jones, solvation and orientation. The fragments used by SAINT2 are from the program FLIB (de Oliveira *et al.*, 2015).

SAINT2 was run in all three modes on the 24 proteins in Set2A. 10,000 decoys were generated in each mode. Each decoy was scored by running TM-align to obtain a TM-score against the native structure. A TM-score > 0.5 is considered to be a correct answer, as this indicates that the structures share the same fold (Xu and Zhang, 2010).

3.3 Results and discussion

I present four different approaches to investigate whether cotranslational folding is likely to be important for the formation of tertiary structure in membrane proteins. Two sets of membrane proteins were used: one set of 93 sequence dissimilar membrane protein chains (Set1), and another of 55 chains from the Orientations of Proteins in Membranes (OPM) database (Lomize *et al.*, 2006) (Set2).

3.3.1 Statistical measures of cotranslational folding

The statistical measures of cotranslational folding described in Section 3.2.2 have previously been used to indicate that the structures of protein domains show a significant bias towards features expected for cotranslational folding (Deane *et al.*, 2007; Saunders *et al.*, 2011). Here, the same measures are used on membrane proteins specifically.

Each measure was calculated for every chain, for the subset of residues which were in membrane. For Set1, these were the residues annotated to be in the tail layer of the membrane by iMembrane (Kelm *et al.*, 2009). For Set2, these were all residues which were part of a transmembrane span annotated by the OPM (Lomize *et al.*, 2006). Table 3.1 shows the number in each sample which had a value greater than or less than the expected mean (given no cotranslational bias) for each measure. The values of the measures for each chain can be categorised into those showing potential cotranslational folding features and those that do not. In the case of MCR, a value > 0.5 indicates the bias expected in cotranslationally folding proteins. $NC_{cen} < 0$ and $SLR > 0$ are also indicators of pro-cotranslational bias. In Set1, for MCR and SLR, the

		Set 1		Set 2	
		number	%	number	%
MCR	< 0.5	47	51	32	58
	≥ 0.5	46	49	23	42
NC_{cen}	< 0	57	61	34	62
	≥ 0	36	39	21	38
SLR	> 0	45	48	33	60
	≤ 0	48	52	22	40

Table 3.1: The number of protein chains for which the statistical measure was less than or greater than the expected value under no cotranslational bias. The group which is expected to be larger in the case of cotranslational bias is shaded in grey.

proteins were evenly split between showing cotranslational bias and not. For NC_{cen} in Set1, and all measures in Set2, more proteins were found with values in the range characteristic of cotranslational folding. The percentage of structures showing cotranslational bias was similar to that previously observed in soluble proteins, where 56% had an MCR < 0.5 and 58% had an NC cen < 0 (Saunders *et al.*, 2011). In soluble proteins, much of the bias was seen in proteins of the α/β SCOP class (Deane *et al.*, 2007; Saunders *et al.*, 2011), with the All-α class displaying little or no bias. Therefore, it is interesting that I observed a small bias in transmembrane alpha-helical bundles, though the difference was not statistically significant due to the small size of the sets.

Set2 was more carefully curated, so might be a better representation of membrane proteins. The distribution of each measure in Set2 is shown in Figure 3.8, indicating the cotranslational bias in each case. Figure 3.8D shows the relationship between two of the measures, MCR and SLR. While there is a tendency for a protein to show cotranslational bias in one measure if it also does in the other, the correlation is weak. This indicates that the two measures are evaluating somewhat different features of structures, and because of this they give complementary evidence that a cotranslational bias exists.

One measure which would not necessarily be expected to show as clear a pattern in membrane proteins as in soluble proteins is the NC_{cen}, as the termini

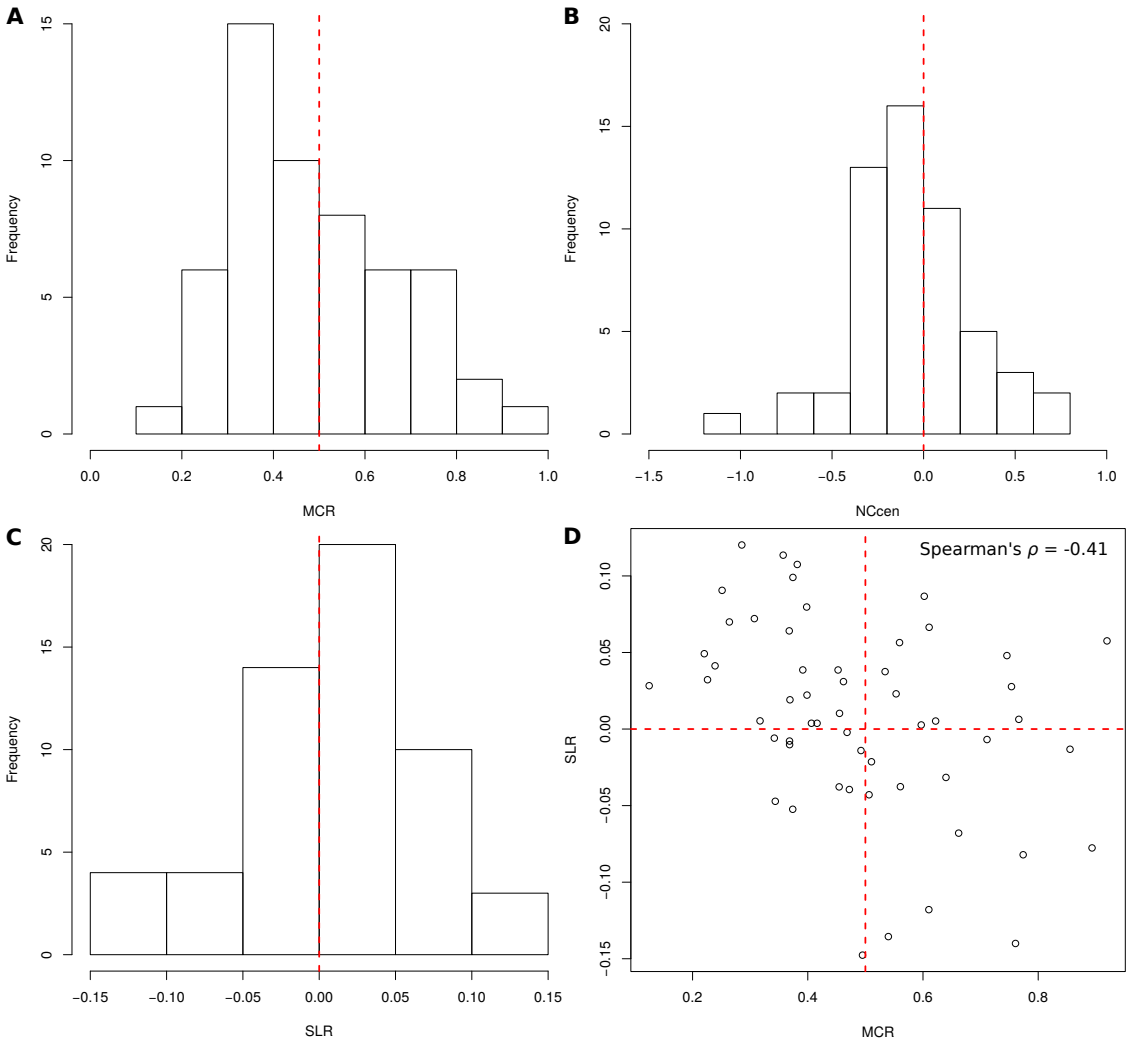


Figure 3.8: Distributions of statistical measures of cotranslational folding in Set2. The dashed red lines indicate the expected mean value of each statistic if no cotranslational bias exists. A) Mean central residue (MCR). B) Relative distance of the N- and C-termini to the centroid of the protein (NC_{cen}). C) Sum of the log ratios, and indication of bias in the distribution of contacts along the chain (SLR). D) SLR plotted against MCR, with the Spearman's rank correlation coefficient shown.

of membrane proteins will be outside the membrane. This might increase the distance to both termini, therefore making the ratio between them closer to one. Despite this, the end of an N-terminal helix is still closer to the centre of the protein than the end of a C-terminal helix.

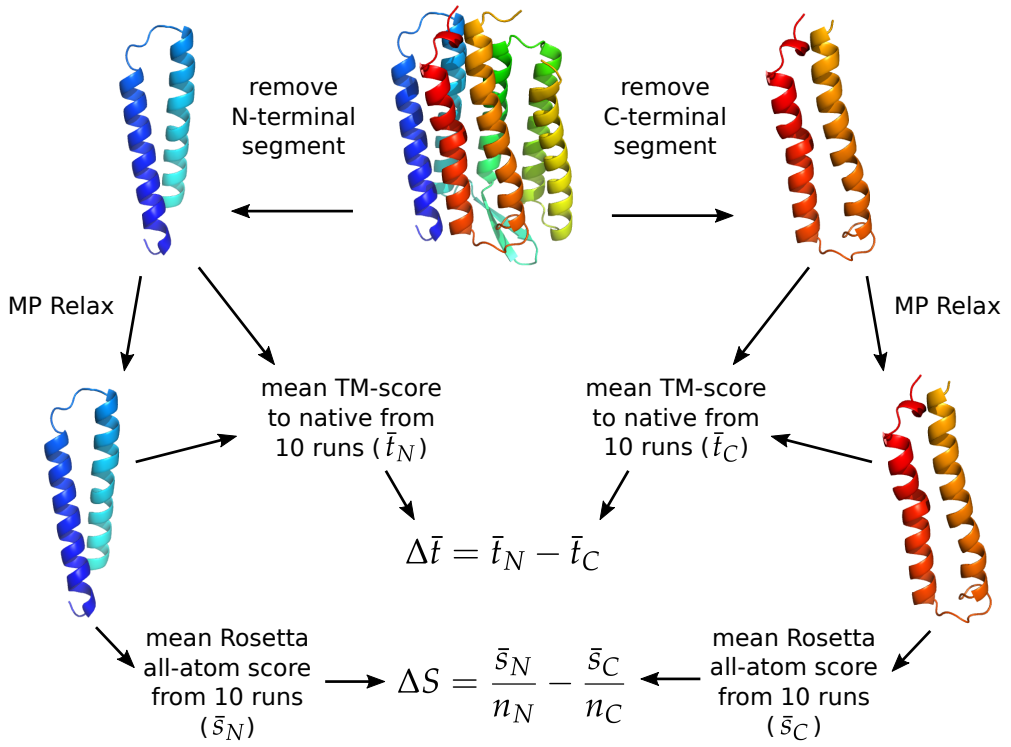


Figure 3.9: Extraction and comparison of membrane protein terminal segments. n_N and n_C are the number of residues in the N- and C-terminal segments respectively.

3.3.2 Stability of segments of native structures

I obtained segments, partial protein structures containing a subset of the transmembrane spans, from the N- and C-terminal ends of each protein chain in Set2 as described in Section 3.2.3 and Figure 3.6. The structure segments were relaxed by the method described in Section 3.2.3.1. Segments of an equivalent size from each end of a structure were compared, on the basis of TM-score to the native segment and RosettaMP score, giving $\Delta\bar{t}$ and ΔS respectively (Figure 3.9). A positive value of $\Delta\bar{t}$ implies that the relaxed N-terminal segment is closer to the native structure than the relaxed C-terminal segment (Equation 3.5). Figure 3.10A shows that no overall bias is seen for $\Delta\bar{t}$, and Figure 3.10C shows an inconsistent pattern at different sizes of segment.

To calculate ΔS, a length normalisation is used to make comparisons between termini (see Equation 3.6). Figure 3.10B shows a marginal bias towards N-terminal segments having more negative normalised scores, i.e. being more

stable. This is consistent with cotranslational folding where N-terminal segments form intermediate foldons, while C-terminal segments are not required to be stable in the membrane environment without the N-terminus also present. Using the null hypothesis $\Delta S = 0$, a t-test for the overall distribution of all lengths of segments gives the p-value 1.7×10^{-5}. However, the data for all values of p cannot be be considered independent as segments of different lengths are taken from the same structure, therefore this test could be misleading.

Figures 3.11 and 3.12 show how the scores for the segments from each terminus compare at different values of p (number of spans). The TM-scores shown in Figure 3.11 appear to be symmetrically distributed about the line $y = x$, and most span lengths have a high average TM-score for both terminal segments. A greater proportion of short segments have lower TM-scores, which may partly cause the greater spread of values for $\Delta\bar{t}$ in Figure 3.10C at low values of p. The cotranslational bias for N-terminal segments to be more stable can be seen in Figure 3.12, where the majority of points are above the line $y = x$. ΔS is given by the difference between the normalised RosettaMP scores for the two termini $\left(\frac{\bar{s}_N}{n_N}, \frac{\bar{s}_C}{n_C}\right)$, and so these are compared in Figure 3.12. The majority of points are above the line $y = x$, which means that the N-terminal segment has a lower score and is therefore more stable.

Figure 3.13 shows, for four example PDB structures, how the relationship between $\frac{\bar{s}_C}{n_C}$ and $\frac{\bar{s}_N}{n_N}$ changes as the number of spans, p, in the segment increases. In the case of 3qe7A (uracil transporter UraA) and 3rkoL (one chain from the transmembrane domain of respiratory complex I), the cotranslational bias is present at any length but is more pronounced to start with. It may be that the difference in stability is greater for the first two spans, or the change may be caused by the method of normalisation for different lengths. It is likely to be the former case in 3rkoL, where the C-terminus is separated from the rest of the chain, interacting only with other chains in the complex and therefore not stabilised by interactions with other helices in the same chain. 4ky0C (glutamate transporter GltTk) shows the reverse bias, and 4m48A (dopamine transporter

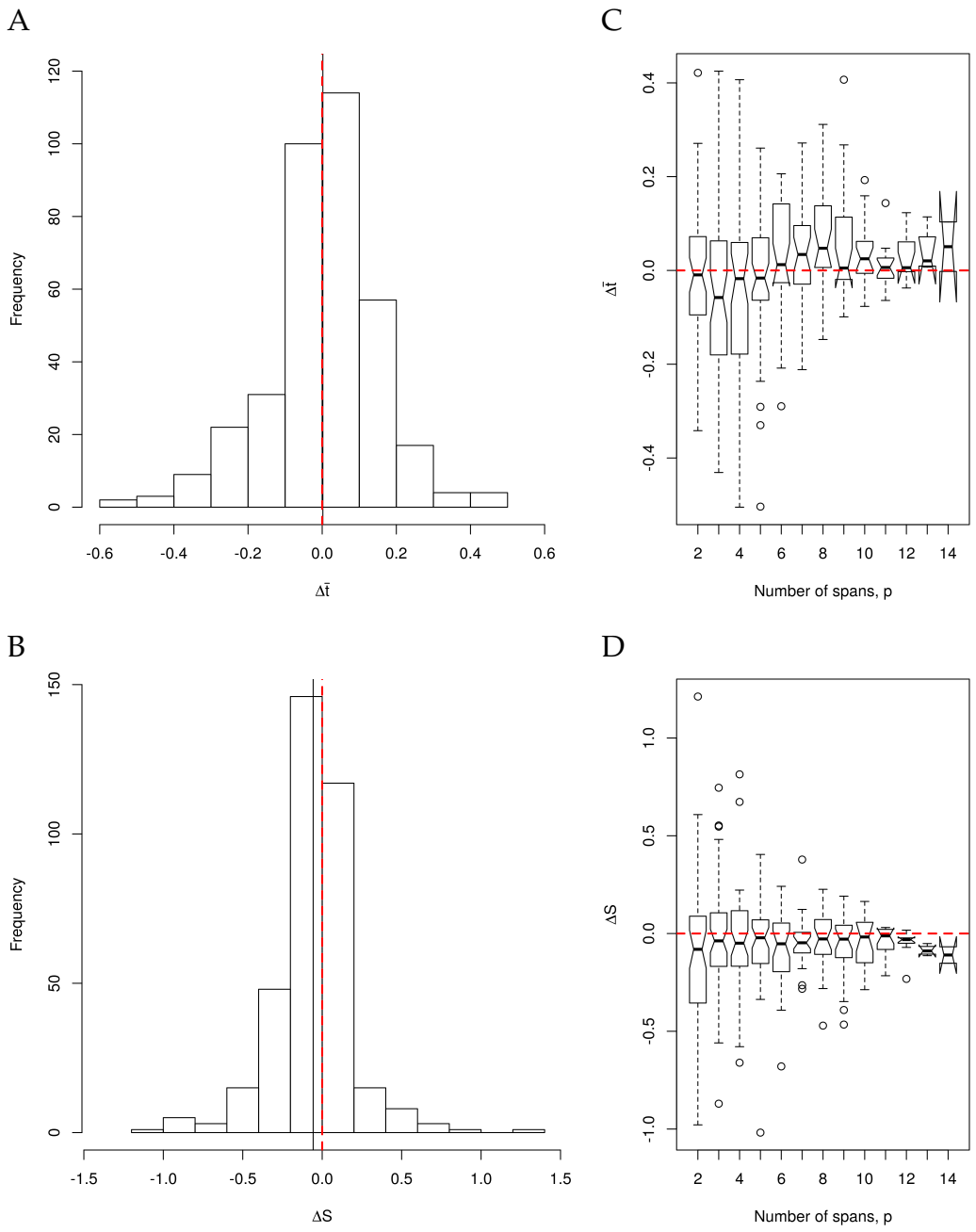


Figure 3.10: (A, B) Histograms to show the distribution of A) $\Delta\bar{t}$ and B) ΔS (both defined in Section 3.2.3.1) for all proteins in Set2 over all values of p (the number of spans in a segment). $\Delta\bar{t}$ is the difference between the N-terminus and C-terminus in the average TM-score between the relaxed model and native segment, positive values indicating N-terminal segment is closer to the native structure. ΔS is the difference between the termini in the average RosettaMP high-resolution score divided by segment length, negative values indicating that the N-terminal segment has a lower energy. The mean value for each statistic is shown by a vertical black line. (C, D) Boxplots of C) $\Delta\bar{t}$ and D) ΔS against the number of spans, p. Sample sizes decrease towards the higher values of p, as fewer proteins in the set have enough spans. The expected value for both statistics under the assumption of no bias is zero, indicated by the dashed red line.

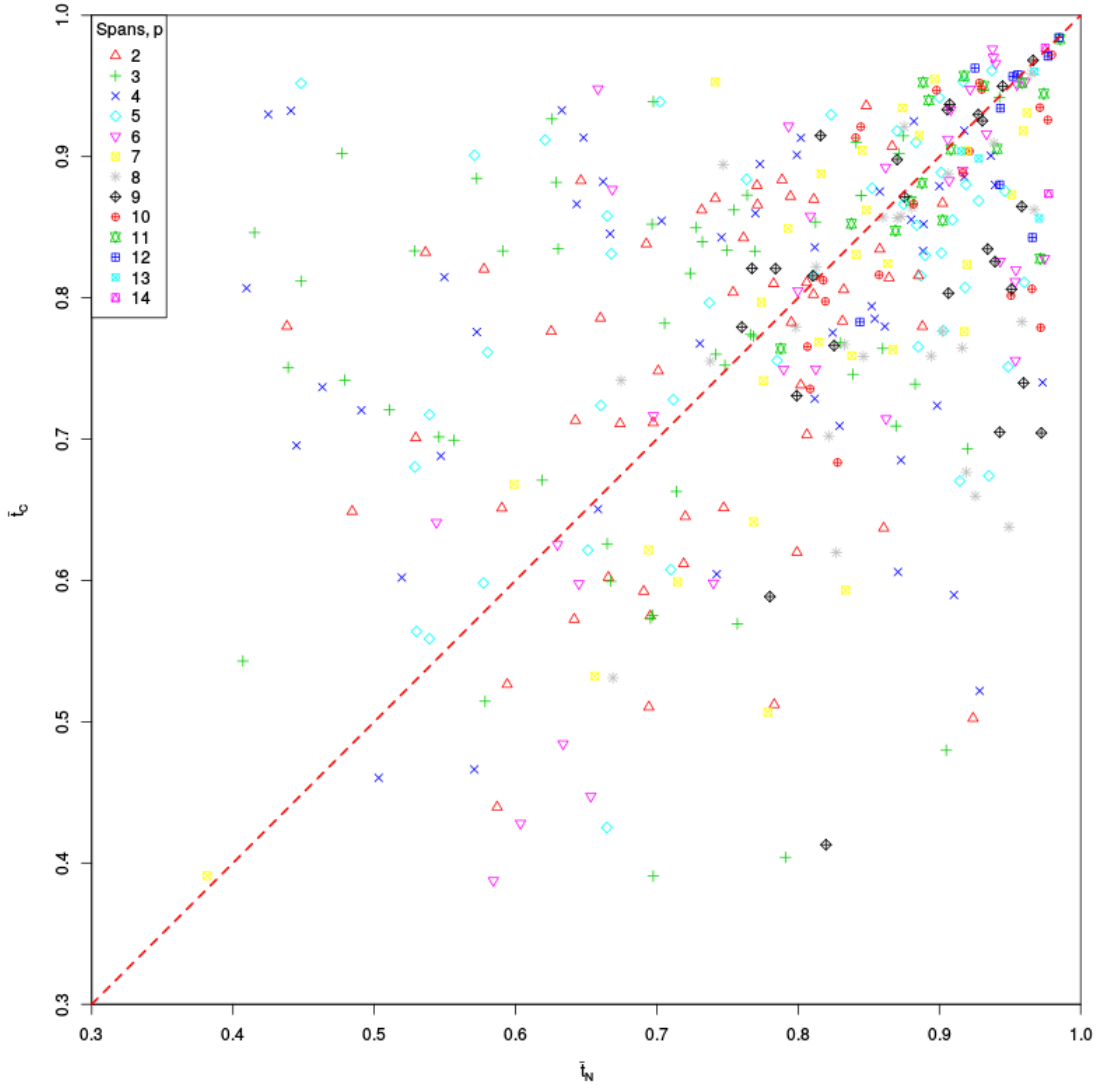


Figure 3.11: Average C-terminal TM-score to native ($\bar{t}_C$) plotted against average N-terminal TM-score to native ($\bar{t}_N$), with points coloured by the number of spans, p, in the segment. For points below the dashed red line, the N-terminal segment remains closer to the native structure than the equivalent C-terminal segment.

of Drosophila) is closer to neutral, but the more stable terminus changes from C to N as more spans are added to the structure.

3.3.3 Comparison of sequence adjacent pairs according to orientation

In order to investigate the importance of insertion of transmembrane helices (TMHs) in pairs, I analysed pairs of TMHs that were adjacent in the sequence

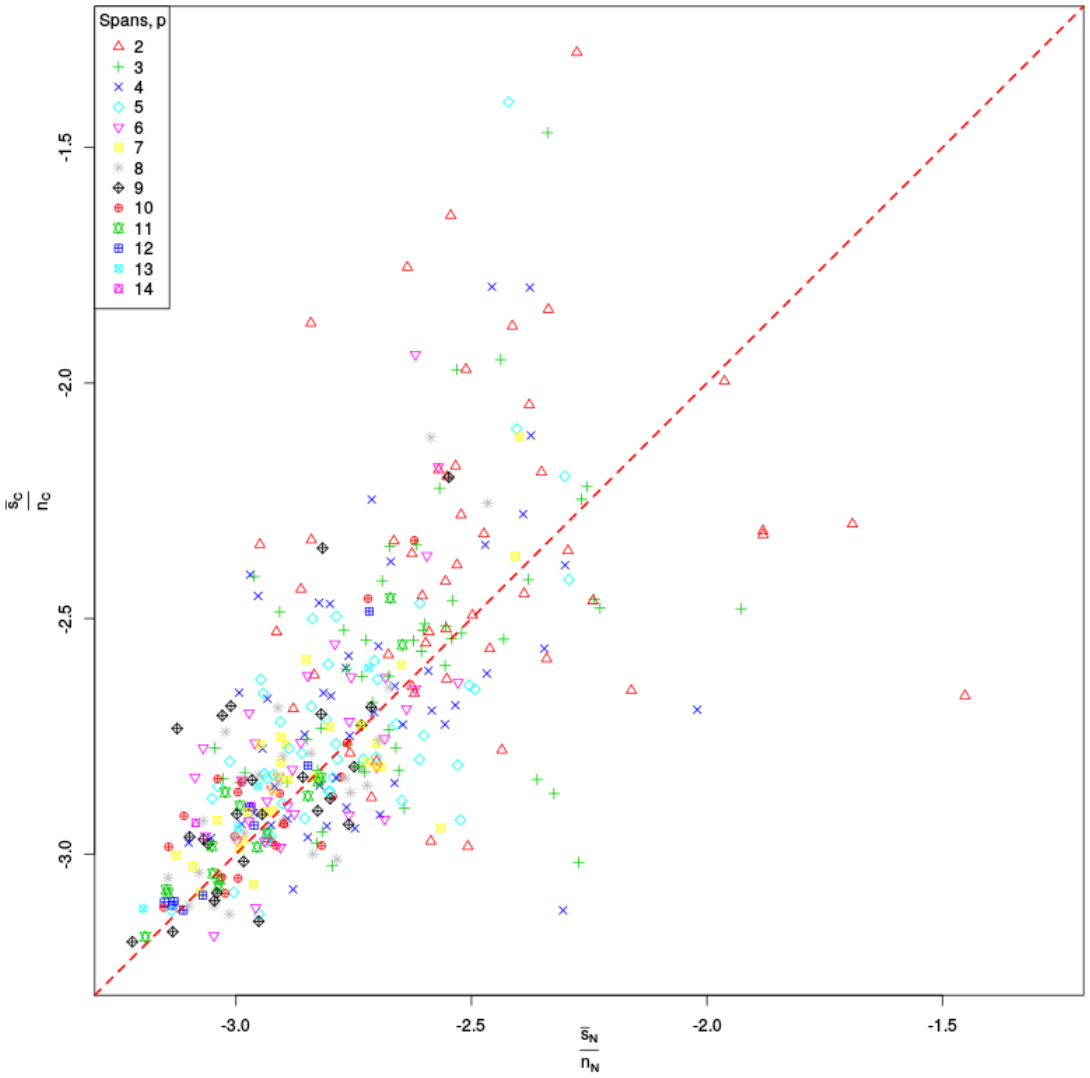


Figure 3.12: C-terminal length-normalised score $\left(\frac{\bar{s}_C}{n_C}\right)$ plotted against N-terminal length-normalised score $\left(\frac{\bar{s}_N}{n_N}\right)$, with points coloured by the number of spans, p, in the segment. Points above the dashed red line have a lower (better) N-terminal score than C-terminal score.

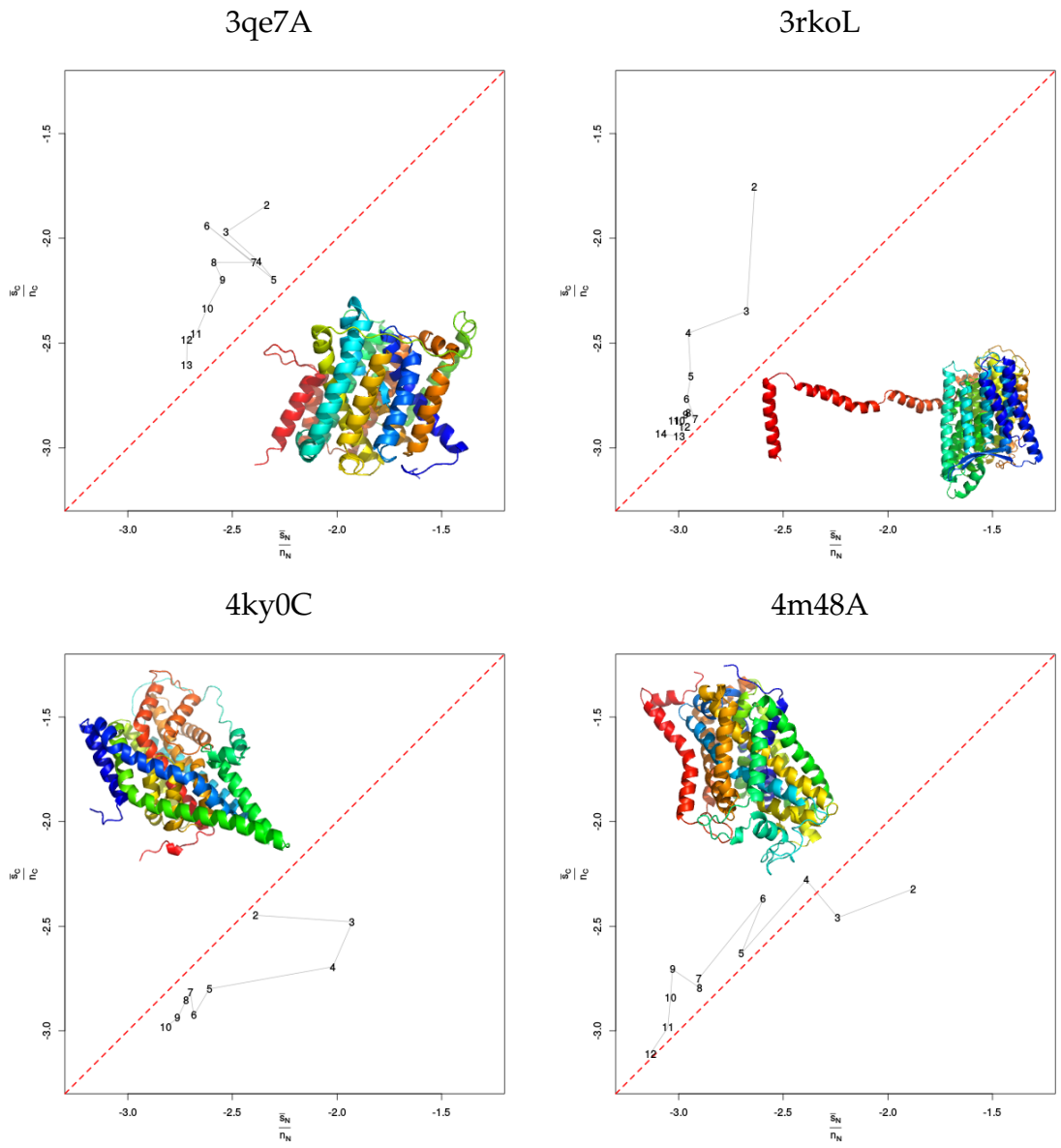


Figure 3.13: C-terminal length-normalised score $\left(\frac{\bar{s}_C}{n_C}\right)$ plotted against N-terminal length-normalised score $\left(\frac{\bar{s}_N}{n_N}\right)$, for every value of p for four example PDB structures. Points are labelled with the number of spans, p, in the segment. Each structure is shown in cartoon representation, coloured by rainbow from N-terminus (blue) to C-terminus (red).

	extracellular		intracellular	
	number	%	number	%
$M \leq 0.065$	36	26	40	34
$M > 0.065$	105	74	77	66

Table 3.2: The number and percentage of TMH pairs from Set2 which are interacting, defined as interaction strength (M) > 0.065, in the set of TMH pairs joined by an extracellular loop and the set of TMH pairs joined by an intracellular loop.

of a protein chain. From Set2, I extracted a set of 141 TMH pairs joined by an extracellular loop, and a set of 117 joined by an intracellular loop (see Section 3.2.4). Only the pairs joined by an extracellular loop would have an opportunity to interact prior to insertion from the intracellular side, in order to be inserted together.

Figure 3.14 compares the distribution of interaction strengths (see Section 3.2.4) of the two groups of TMH pairs. A Mann-Whitney test indicated that the interaction strength was greater for pairs joined by an extracellular loop (p-value = 9×10^{-5}). Using the generous cut-off for interaction strength used by Zhang *et al.* (2015) to detect interacting pairs, the majority of both groups is defined to be interacting (Table 3.2). The fraction of non-interacting pairs is higher for the group of pairs with an intracellular loop, and among the interacting pairs, the extracellular loop group has a greater number of very strong interactions (> 0.1, Figure 3.14).

These observations support the hypothesis that pairs of TMHs may interact before insertion, and even be inserted together, as this could be linked to stronger interactions in the extracellular loop group. In contrast, the intracellular loop group cannot obviously be inserted from the intracellular side in a concerted fashion, therefore these interactions may form after rearrangement of TMHs within the membrane. Differences in interaction strength could have been caused by shorter loops in one set making those adjacent TMHs closer and more likely to interact, however there is no trend for extracellular loops to be shorter (Figure 3.15). It was also previously seen that the average loop length between

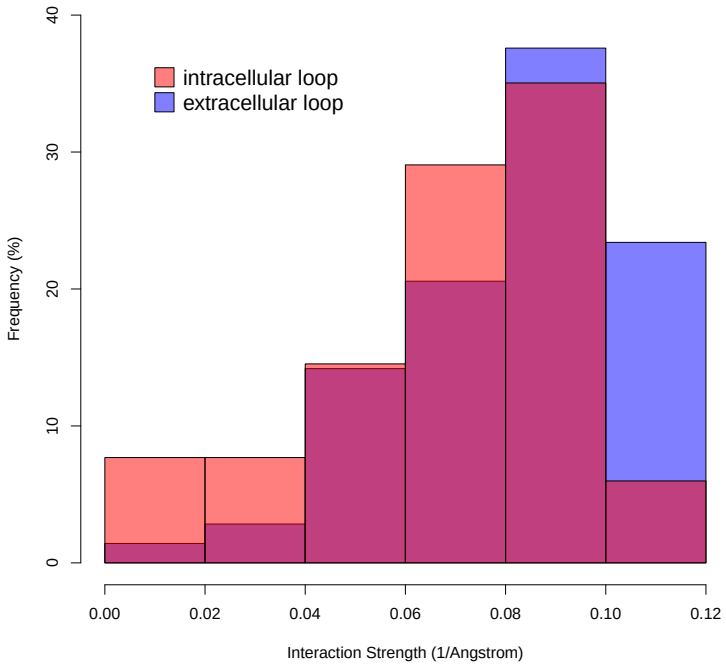


Figure 3.14: Histogram to compare the interaction strength (see Section 3.2.4) in the TMH pairs connected by an intracellular loop to those connected by an extracellular loop. The frequency is calculated as a percentage of the number in the sample (extracellular = 141, intracellular = 117).

interacting sequence-adjacent TMHs is no different to that for non-interacting sequence-adjacent TMHs (Gimpelev *et al.*, 2004).

3.3.4 Prediction of membrane protein structures by SAINT2

SAINT2 is a fragment based structure prediction program which has obtained good results for soluble proteins, demonstrating that the Forward method slightly outperforms the Reverse method and greatly outperforms In vitro. The difference between the modes suggests that it is successfully simulating the folding pathway, otherwise the greedy nature of the conformational search would lead to entrapment in a local minimum and less accurate results. Here I have tested it on membrane proteins, postulating that a similar result could be an indication that membrane proteins are also folding cotranslationally. I used the three versions of SAINT2 to test whether a sequential approach is a more

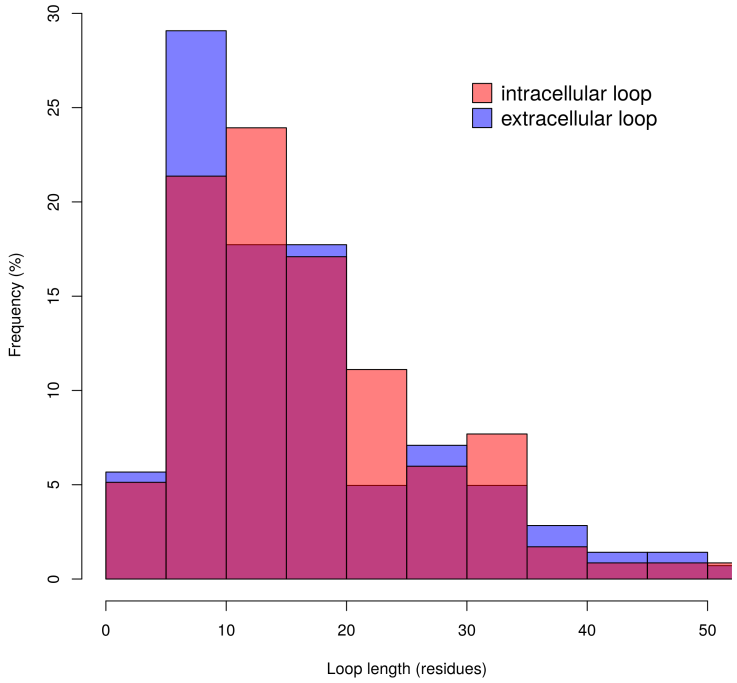


Figure 3.15: Histogram to compare the loop length joining TMHs in the two groups of TMH pairs. The frequency is calculated as a percentage of the number in the sample (extracellular = 141, intracellular = 117).

efficient way to sample the conformational space and generate accurate decoys.

I generated 10,000 decoys for each of the 24 targets in transmembrane protein Set2A using SAINT2 Forward, SAINT2 Reverse and SAINT2 In vitro. For each version, identical fragment libraries, residue-residue contacts and number of moves to generate a decoy were used. Figure 3.16 compares the best TM-score of all decoys (TM-score Best) produced by each of SAINT2 Forward and SAINT2 In vitro. For 19 out of 24 proteins, SAINT2 Forward produced a more accurate decoy. In two cases, the improvement in TM-score Best for Forward over In vitro was > 0.15. There were five proteins where SAINT2 Forward produced a correct answer (TM-score > 0.5) but SAINT2 In vitro did not. Two of these cases were the two longest transmembrane proteins (292 and 324 residues) for which a correct decoy was produced by any version. For two proteins, SAINT2 In vitro produced a correct answer while SAINT2 Forward did not. In soluble proteins, SAINT2 Forward always produces a correct decoy for targets where SAINT2

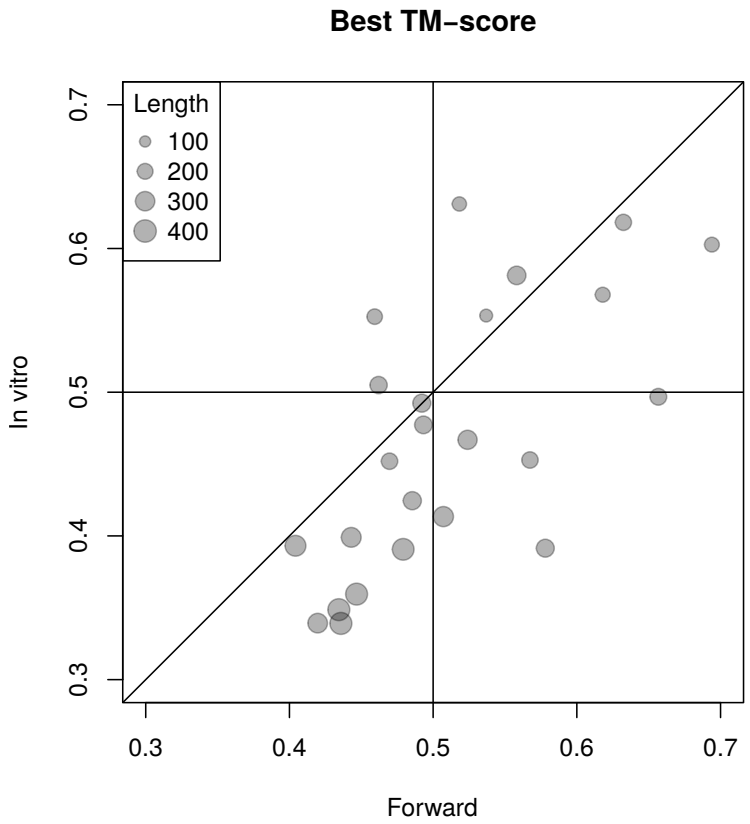


Figure 3.16: Comparison between the best TM-score produced out of a set of 10,000 decoys by SAINT2 Forward (x-axis) and SAINT2 In vitro (y-axis). SAINT2 Forward is performing better where points are below the diagonal line.

In vitro is successful, so in this respect SAINT2 Forward is more consistent for soluble proteins than membrane proteins. For the set of membrane proteins, the sequential method does not always lead to improvement, but overall the improvement in TM-scores for SAINT2 Forward over SAINT2 In vitro is similar to the improvement seen in soluble proteins.

The Rosetta *ab initio* membrane protocol uses an incremental but bi-directional method to build decoys (Yarov-Yarovoy *et al.*, 2006). This aims to improve the efficiency of sampling by allowing each TMH to adopt a position that spans the membrane as it is added. The trajectory to a conformation in which all TMHs span the membrane could be more difficult to achieve starting from an extended chain, as in the Rosetta protocol for soluble proteins (Rohl *et al.*, 2004). If this is the only advantage of the sequential method, we would expect no overall

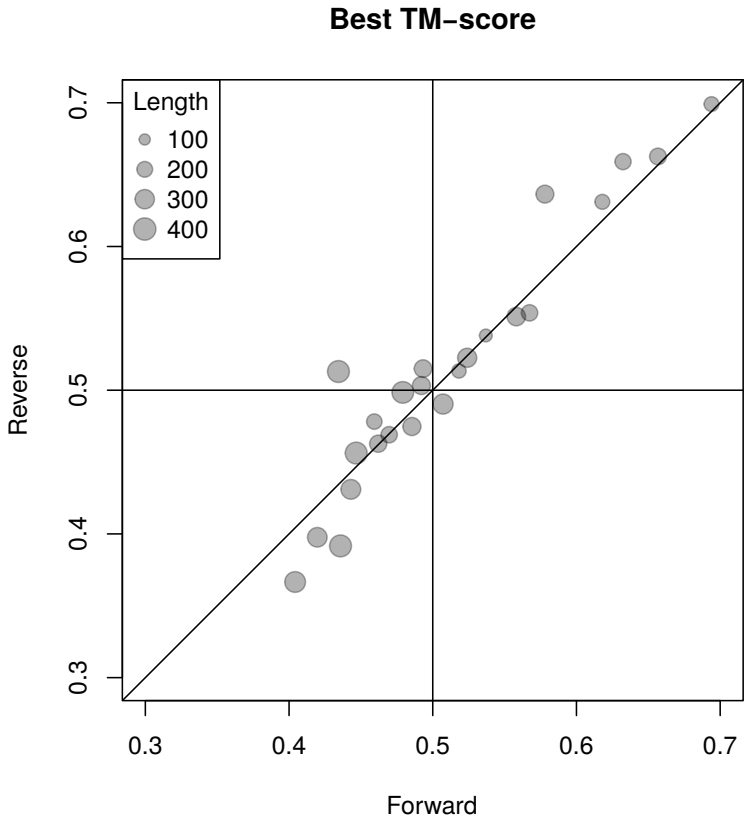


Figure 3.17: Comparison between the best TM-score produced out of a set of 10,000 decoys by SAINT2 Forward (x-axis) and SAINT2 Reverse (y-axis). SAINT2 Forward is performing better where points are below the diagonal line.

difference between SAINT2 Forward and SAINT2 Reverse, which performs the same protocol as SAINT2 Forward but in the non-biological C- to N-terminal direction. Figure 3.17 shows that SAINT2 Forward and SAINT2 Reverse perform very similarly on this set of membrane proteins. It is therefore unclear to what extent the biological significance of cotranslational folding is important, in addition to the benefit of an incremental method for sampling efficiency.

The largest differences between the SAINT2 Forward and SAINT2 In vitro were shown by PDB code 3klyA, a pentameric formate channel, and 4b4aA, which is TatC, part of the twin-arginine transport pathway which transports folded proteins across the membranes of bacteria and chloroplasts (Rollauer *et al.*, 2012). Correct models were produced for TatC by both SAINT2 Forward and SAINT2 Reverse, but not SAINT2 In vitro (Figure 3.18).

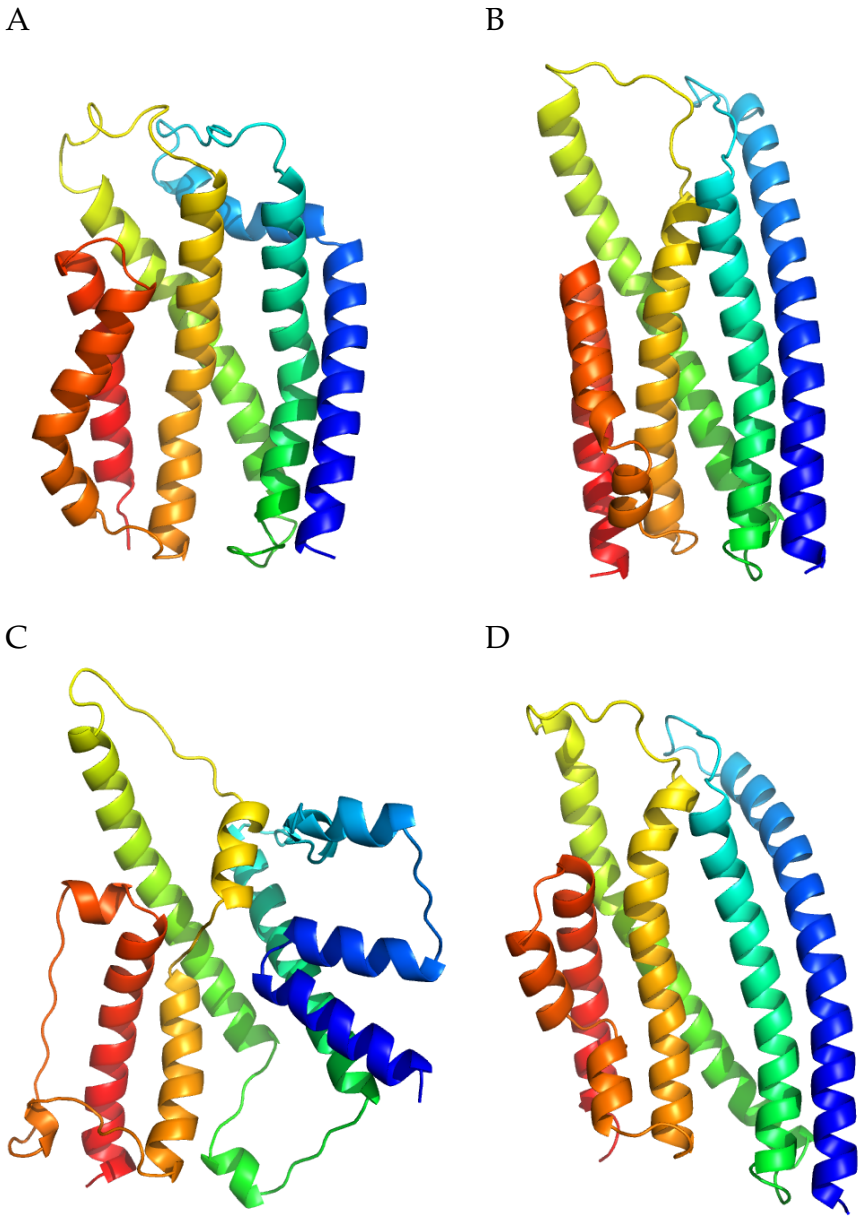


Figure 3.18: Structures of TatC (PDB code 4b4aA, length 225 residues) coloured by rainbow from N-terminus (blue) to C-terminus (red). All structures are viewed from the same angle, as aligned to the crystal structure. A) Crystal structure. B) Best model by SAINT2 Forward (TM-score 0.66 to the crystal structure). C) Best model by SAINT2 In vitro (TM-score 0.49). D) Best model by SAINT2 Reverse (TM-score 0.63).

3.3.4.1 Relaxation and scoring by MPrelax

While some correct answers were produced by SAINT2, there tended to be very few correct decoys per target (Table 3.3). This may be due to the lack of any membrane-specific component in the scoring system. Only secondary structure prediction and contacts can help to guide the TMHs to their correct positions. To begin training SAINT2 to perform better on membrane proteins, the set was divided into a training ($Set2A_{train}$) and test set ($Set2A_{test}$), with only the training set targets used to measure success (Section 3.2.1). To attempt to distinguish between good and bad decoys, I scored 1000 Forward decoys for the targets in $Set2A_{train}$ with the membrane protein low-resolution scoring function in the RosettaMP framework. The results of this score are shown in Figures 3.19A and B for two targets. These are typical targets in the set, like the majority of targets showing hardly any noticeable correlation between the score and the accuracy of the decoy. However, the scores are very high, and therefore not typical of conformations generated in programs which use the RosettaMP score. These results suggest that relaxation could improve the scores and perhaps their correlation.

The MPrelax protocol used for the segments in Section 3.3.2 was applied again to a subset of 1000 decoys per target in $Set2A_{train}$. Spans from the native structure were used as input, and just one relaxed model was produced for each decoy. This time only one run per decoy was used due to the computational demand of processing such a large number of structures. The time taken to perform one run of MPrelax on one core ranged from around 3 minutes for a length of 132 residues to roughly 40 minutes for 385 residues. Figures 3.19C and D show the relaxed scores obtained, plotted against the TM-score between the native and the relaxed decoy. An improved correlation was observed for most targets in the training set, when compared to the unrelaxed scores. However, this appears to be mostly due to the TM-scores of some relaxed decoys being worse than the original decoys as the original decoy partially unfolds during

PDB code and chain identifier	Length (residues)	In vitro	Forward	Reverse
1orsC	132	44	147	101
4huqS	164	86	6	4
3rlbA	176	38	3702	1753
2xowA	179	5	99	318
4a2nB	192	5	0	0
4o6yA	210	38	436	434
1kqfC	216	0	158	91
3b4rB	216	0	0	0
4b4aA	225	0	467	916
1e12A	239	1	0	0
2vpzC	250	0	0	1
3klyA	257	0	74	76
2dyrC	259	0	0	1
2w2eA	263	0	0	0
4od5A	274	10	28	28
1okcA	292	0	3	2
4n7wA	307	0	0	0
3m73A	313	0	0	0
2qi9A	324	0	1	0
4ezcA	345	0	0	0
1zcdA	376	0	0	0
1u7gA	383	0	0	0
4bwzA	384	0	0	1
3cx5C	385	0	0	0

Table 3.3: The number of correct decoys (TM-score > 0.5) for each mode of SAINT2 for each target in Set2A, out of 10,000 generated for each mode.

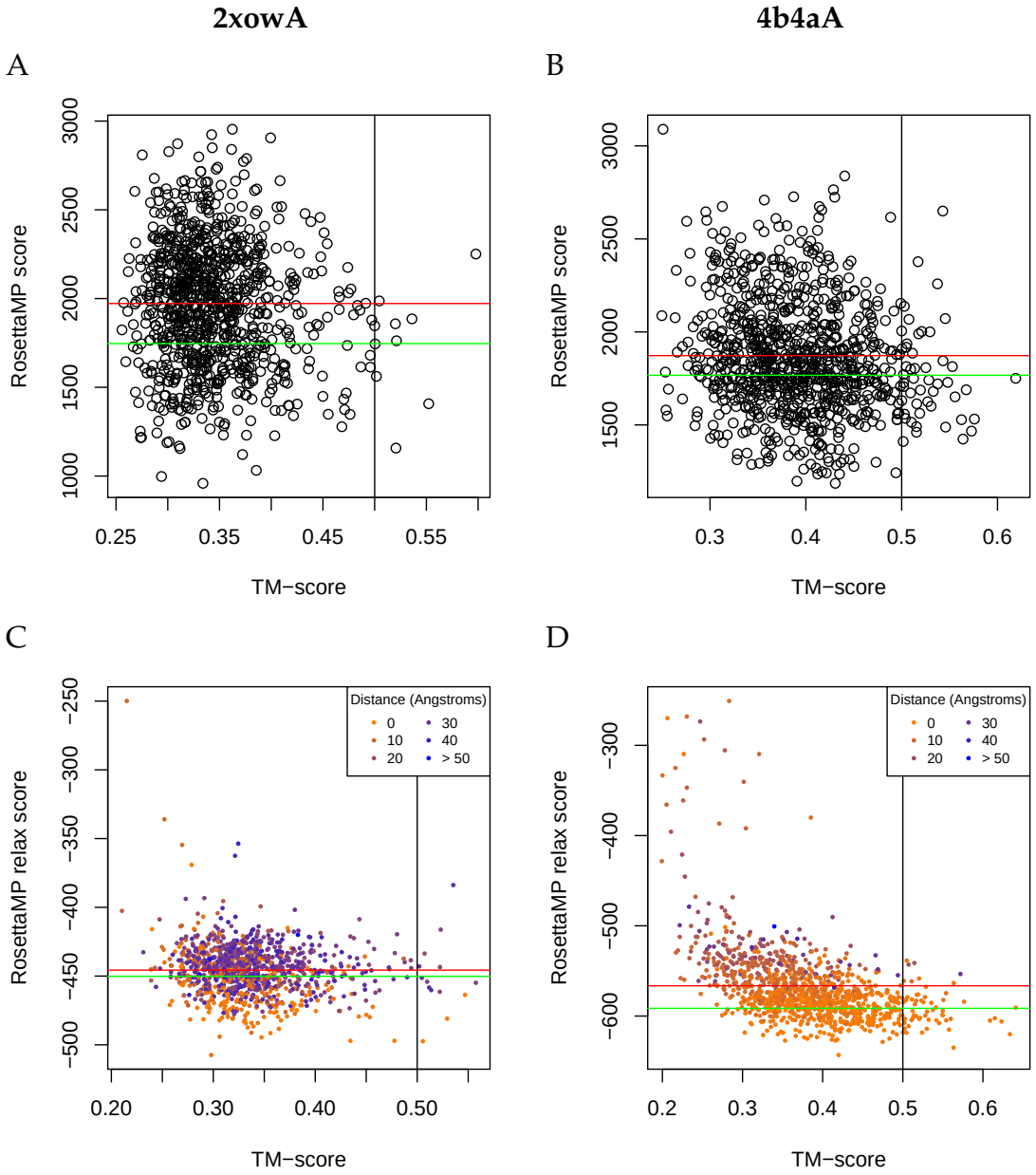


Figure 3.19: A) and B) RosettaMP score of the decoy plotted against TM-score between the decoy and the native structure for two targets in the training set. C) and D) High-resolution scores at the end of the MPrelax protocol, plotted against TM-score between the relaxed decoy and the native structure. Decoys are coloured by the distance between the protein centroid and the plane at the centre of the membrane, from orange (centroid at the centre of the membrane) to blue (centroid > 50 Å from the membrane). The green line indicates the mean Rosetta score for decoys with TM-score ≥ 0.5; the red line is the mean for decoys with TM-score < 0.5. 2xowA is a structure of GlpG, an intramembrane rhomboid protease; 4b4aA is a structure of TatC, part of the twin-arginine transport pathway.

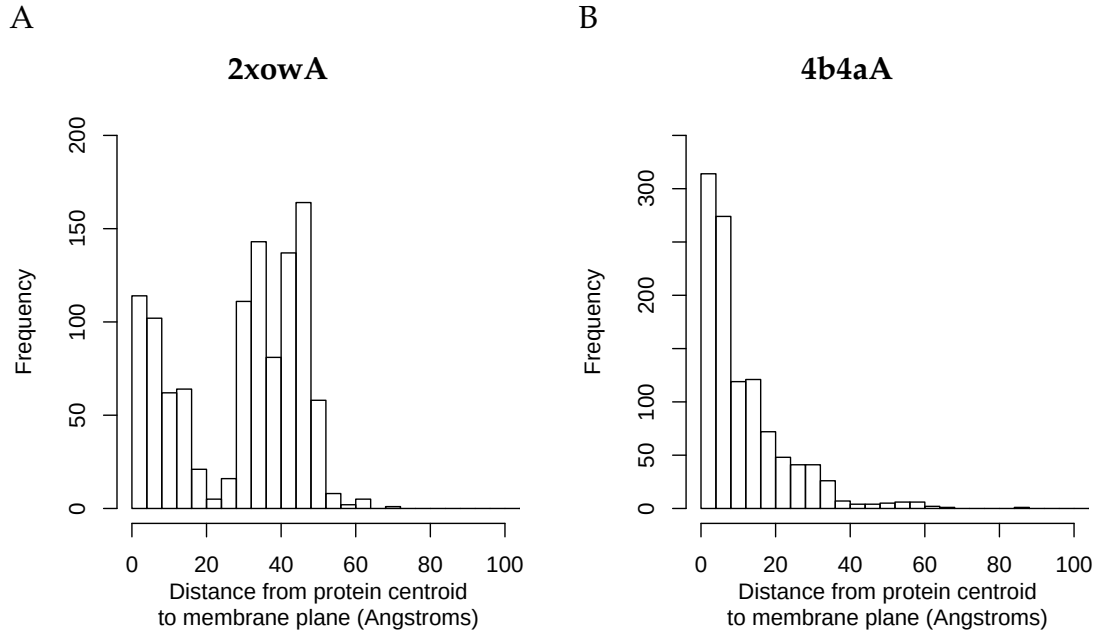


Figure 3.20: The distribution of distances from the centroid of the relaxed decoy to the plane at the centre of the membrane for the two targets in Figure 3.19.

relaxation. These structures are therefore easily separated from the rest of the population which remain folded.

The relative success of embedding each decoy in the membrane also affected the scores. I used the distance between the protein centroid and the plane at the centre of the membrane as a proxy for success in embedding each decoy. Figure 3.20 shows the distribution of distances for the two given targets, and this distance is also displayed in Figures 3.19C and D, ranging from orange to blue. 4b4aA was usually reasonably well embedded, but less than half of the decoys for 2xowA were successfully embedded (distance $<$ 20 Å). It was also very common for the majority of decoys not to be embedded for other targets. For 4b4aA, the unsuccessfully embedded targets had worse RosettaMP scores and TM-scores. By contrast, in 2xowA, while the poor embedding leads to higher RosettaMP scores, this did not correspond to worse TM-scores. For other targets, it was even the case that better RosettaMP scores were achieved by worse embedded decoys, therefore there was no consistent pattern.

3.4 Conclusions

In this chapter, I performed a series of tests to see if the tertiary structure of membrane proteins could be adopted during the process of translation.

Three different statistical measures calculated from membrane protein structures were suggestive of a bias consistent with cotranslational folding. Unfortunately, the number of structures in the non-redundant data set Set2 was too small to be confident that the differences observed were statistically significant.

Segments consisting of several transmembrane spans were extracted from this set of membrane proteins so that the N- and C-terminal parts of the structures could be compared. I found that the N-terminal segments displayed marginally more stable and foldon-like behaviour than their equivalent C-terminal counterparts, when relaxed using RosettaMP (Alford *et al.*, 2015).

A suggested mechanism of alpha-helical membrane protein folding is that pairs of transmembrane helices (TMHs) could be inserted into the membrane together (Cymer *et al.*, 2015). I separated pairs of consecutive TMHs into those that could have been inserted as a pair from the cytoplasmic side of the membrane, and those that could not. The group which could be inserted pairwise interacted more strongly than those that could not.

I also used protein structure prediction by SAINT2 to investigate possible differences in folding by a non-directional method and a sequential method. SAINT2 Forward was more effective than SAINT2 In vitro, therefore the directional algorithm which better resembles cotranslational folding was the more successful approach. Disappointingly, the Reverse method performs similarly, and therefore it is unclear what aspect of the Forward method is causing it to perform better, and how large a role biological cotranslational folding may have. At least one correct answer was produced for nearly half of the decoys, but the current challenge is differentiating between the very small number of correct decoys ($< 5\%$) and the incorrect decoys. The MPrelax protocol did not appear to select correct decoys with any more success than the raw

low-resolution membrane scoring function used directly on the SAINT2 output or the SAINT2 score itself. The following chapters describe my work on the implementation of a membrane embedding throughout decoy generation, in order to increase the number of good decoys. As part of this development, I began with the implementation of a protocol to start from an embedded part of the native structure to investigate how this may affect folding.

4

Adaptation of SAINT2 for membrane proteins

This chapter describes the process of adapting SAINT2 for the prediction of alpha-helical membrane proteins. Membrane proteins pose a different challenge from the soluble proteins on which SAINT2 has previously been trained and tested. Membrane proteins exist in a different biophysical environment, meaning alternative scores are appropriate. They also tend to be longer than soluble proteins (Koehler Leman *et al.*, 2015), therefore they are usually outside of the length range (roughly $<$ 150 residues) of proteins predicted well by *de novo* fragment-based modelling. The results in Section 3.3.4 show that correct answers can be obtained for some shorter membrane proteins, without any adaptation of the original SAINT2 program.

Other *de novo* predictors have been adapted specifically for prediction of membrane protein structures (e.g. Hopf *et al.*, 2012; Ovchinnikov *et al.*, 2015; Teixeira *et al.*, 2017), with most making use of predicted contacts inferred from correlated evolution of residues. These approaches have led to high accuracy (TM-score $>$ 0.7) predicted structures for a large number of proteins for which a great amount of sequence data is available (Ovchinnikov *et al.*, 2017). However,

none of the existing programs attempt to imitate the biological folding process of membrane proteins.

In this chapter, I aim to build a *de novo* predictor specific to membrane proteins that uses information specific to their folding. In soluble proteins, the biology-mimicking sequential protocol of SAINT2 produces much more accurate decoys than the alternative In vitro method, and I hope to make similar improvements for the prediction of membrane proteins. I also aim to learn more about the biological folding process, by observing which computational strategies are most successful.

In order to reduce the complexity of the task, instead of predicting a complete membrane protein structure, I have developed a method of decoy generation that builds on a part of the native structure, SAINT2-ScafFold. I use this as a simplified model for cotranslational folding, because at the point of extrusion of the second half of a protein, the first half is already inserted into the membrane. This N-terminal section may already be able to adopt the same conformation as it forms when it is part of the fully extruded and folded native structure. SAINT2-ScafFold was found to perform similarly to the original fully sampling SAINT2 in prediction of the remaining protein, though there were cases where it leads to an improvement.

SAINT2-ScafFold was then used to test a new potential for scoring residues according to their depth in the membrane, which improved the accuracy of decoys. From the large number of decoys generated, acceptance ratios could be calculated for move steps at every stage of chain growth, and in every location in the peptide. The maps of acceptance ratios provide insight into the sampling process, and the effect of the new membrane potential on the acceptance of proposed moves.

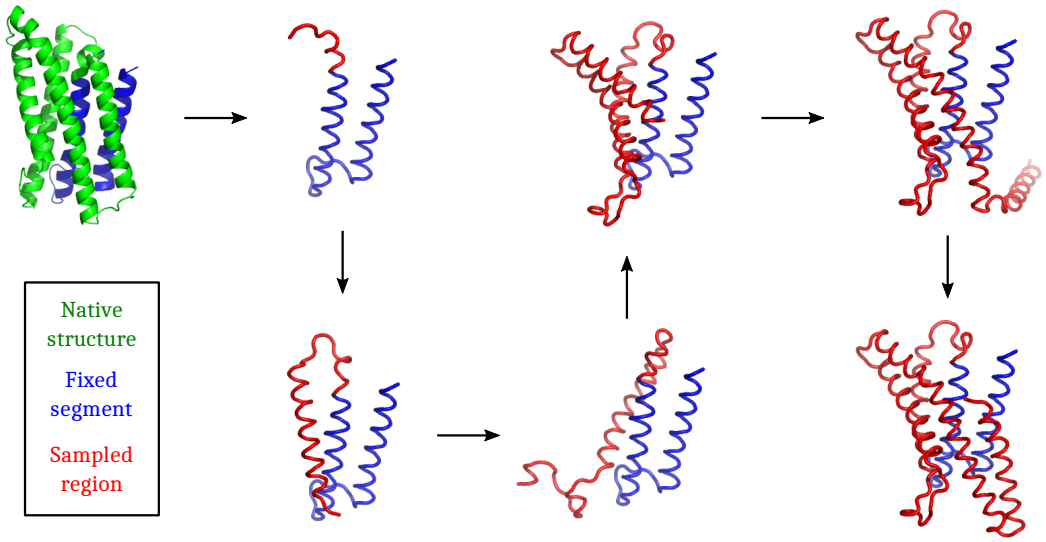


Figure 4.1: SAINT2-ScafFold builds from a segment of the native structure.

4.1 Methods

4.1.1 SAINT2-ScafFold

4.1.1.1 Implementation

SAINT2-ScafFold is a new version of SAINT2 (described in Section 1.4.4.3) that was implemented to build on a native structure segment that is not conformationally sampled during the process of decoy generation. The user provides an argument for the number of residues to be read from the native PDB structure, which creates the peptide object from which to begin the decoy generation (Figure 4.1). Exactly the same fragment library built for use by the original full length sampling version of SAINT2 (SAINT2-Wholly) can be used for SAINT2-ScafFold. The first step is an extrusion of one residue, using a fragment from the library that ends one residue past the end of the segment. As the minimum length of a fragment in the fragment libraries is six residues, the new fragment includes residues within the segment. In order to ensure no change to the segment, only the torsion angles for the non-segment residues are replaced. This is the case for any move steps that overlap the segment, in addition to the extrusion steps. In this way, for residues not in the segment,

all of the torsion angles available to SAINT2-Wholly in the original fragment library can also be used by the ScafFold protocol.

The algorithm continues using the same protocol as SAINT2-Wholly (Section 1.4.4.3), but an adjustment is made in relation to the number of move steps between each extrusion. As the peptide grows, therefore increasing the number of available residues to sample, the probability of each move changing the torsion angles of a given residue decreases. In SAINT2-Wholly, a different number of moves is made after each extrusion; this ensures an equivalent amount of sampling is performed on a given residue between successive extrusions at the start and end of the simulation. The 10,000 growth moves are allocated between extrusions so that the number of moves at a given length increases linearly with length.

In SAINT2-ScafFold, the number of moves is proportional to the length of peptide which is currently being actively sampled. Regardless of the overall length of the protein, SAINT2-Wholly uses 10,000 moves. In order to make a comparison between ScafFold and Wholly, a similar number of moves must be proposed in the second half of the protein at each length in both versions. Figure 4.2 shows how the appropriate total number of growth moves can be calculated. The gradient of the line is the same for both methods, as the number of moves proposed at a given length is proportional to the peptide length at that point. As the triangles are similar, the ratio of the number of moves is equal to the square of the ratio of the number of residues to be sampled:

$$\text{adjusted growth moves} = \frac{M_g \left(L - S\right)^2}{L^2} \tag{4.1}$$

where M_g is the number of growth moves in SAINT2-Wholly (usually 10,000), L is the length of the full peptide, and S is the length of the segment. $(L - S)$ is therefore the length of peptide to be sampled in SAINT2-ScafFold.

On top of the 10,000 moves during extrusion, SAINT2-Wholly performs 1,000 moves once the entire structure has been extruded (full length moves). For full

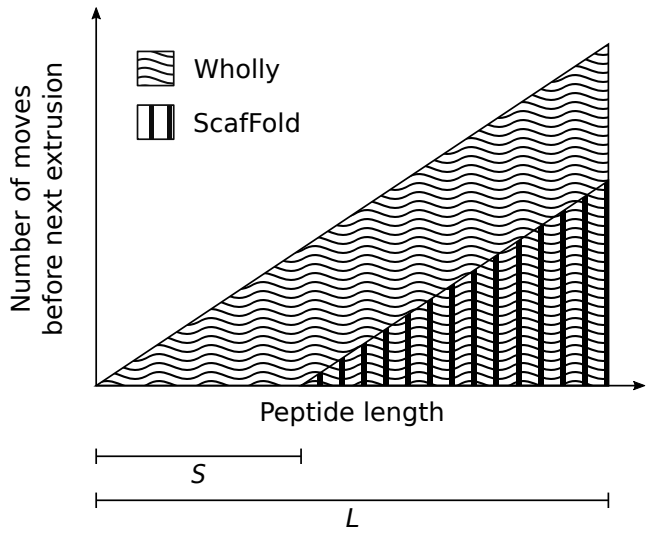


Figure 4.2: Move adjustment for SAINT2-ScafFold, compared to SAINT2-Wholly. The area shown for each method represents the total number of moves.

length moves, the adjustment is based on the ratio of the lengths of residues available to sample when the peptide is fully extruded:

$$\text{adjusted full length moves} = \frac{M_f\,(L - S)}{L} \tag{4.2}$$

where M_f is the number of full length moves in SAINT2-Wholly (usually 1,000).

The final change to the program in the implementation of SAINT2-ScafFold was to change the frame of reference for the recalculation of atom coordinates. In the original SAINT2, after torsion angles are replaced, atom coordinates are recalculated one by one, beginning at the C-terminal end of the replaced fragment. This facilitates implementation of features such as a ribosome wall-like potential, where the ribosome is a plane from which the C-terminus emerges. In my case, I wished to use the membrane, or existing segment, as the frame of reference, so it did not need be repositioned after every move. This simplified the later use of a membrane potential based on the membrane embedding position of the segment. Therefore, the direction of calculation of atom positions was changed to proceed from the N- to C-terminus.

The Reverse mode functions were changed in an analogous way, so that the C-terminal segment remains static as moves and extrusions are carried out in the N-terminal section of the peptide.

4.1.1.2 ScafFold scoring

In order to compare the SAINT2-ScafFold and Wholly versions, I used the accuracy of prediction of the non-segment section of the protein. TM-align was used to calculate the TM-score between the native structure and each decoy, after extracting from both only the residues not in the ScafFold segment. In order to compare the performance of prediction by SAINT2-ScafFold with different length segments, the set of non-segment residues which were common to all set-ups were used for scoring. In other words, this comparison was made by scoring only the residues whose conformations were sampled in the runs with the longest segment.

4.1.2 Membrane potential

A membrane potential was implemented based on the knowledge-based potential developed by Nugent and Jones (2013). This method embeds a structure in a membrane by scoring a range of positions and choosing the lowest energy position to output as the correct embedding. The potential takes the form of a table of pseudo-energies for each amino acid at different depths in the membrane, divided into 1.5 Å bins over a full membrane thickness of 48 Å (Figure 4.3). The energies for all residues in a structure are summed according to their depths to give a total energy for a position of the protein relative to the membrane.

The SAINT2-ScafFold protocol begins from a scaffold segment of the native structure, which was embedded in the membrane by the PDBTM in Section 3.2.1.1. Therefore this membrane position was used as a frame of reference and other possible orientations were not sampled during decoy generation. In order to avoid the requirement of using a specific orientation of the protein in the membrane (i.e. N-terminus intracellular or extracellular) in addition to defined or predicted transmembrane spans, I used a simplified symmetric version of this potential (Figure 4.3). For each distance from the centre of the membrane, I took the average of the energy at that distance either side of the membrane centre.

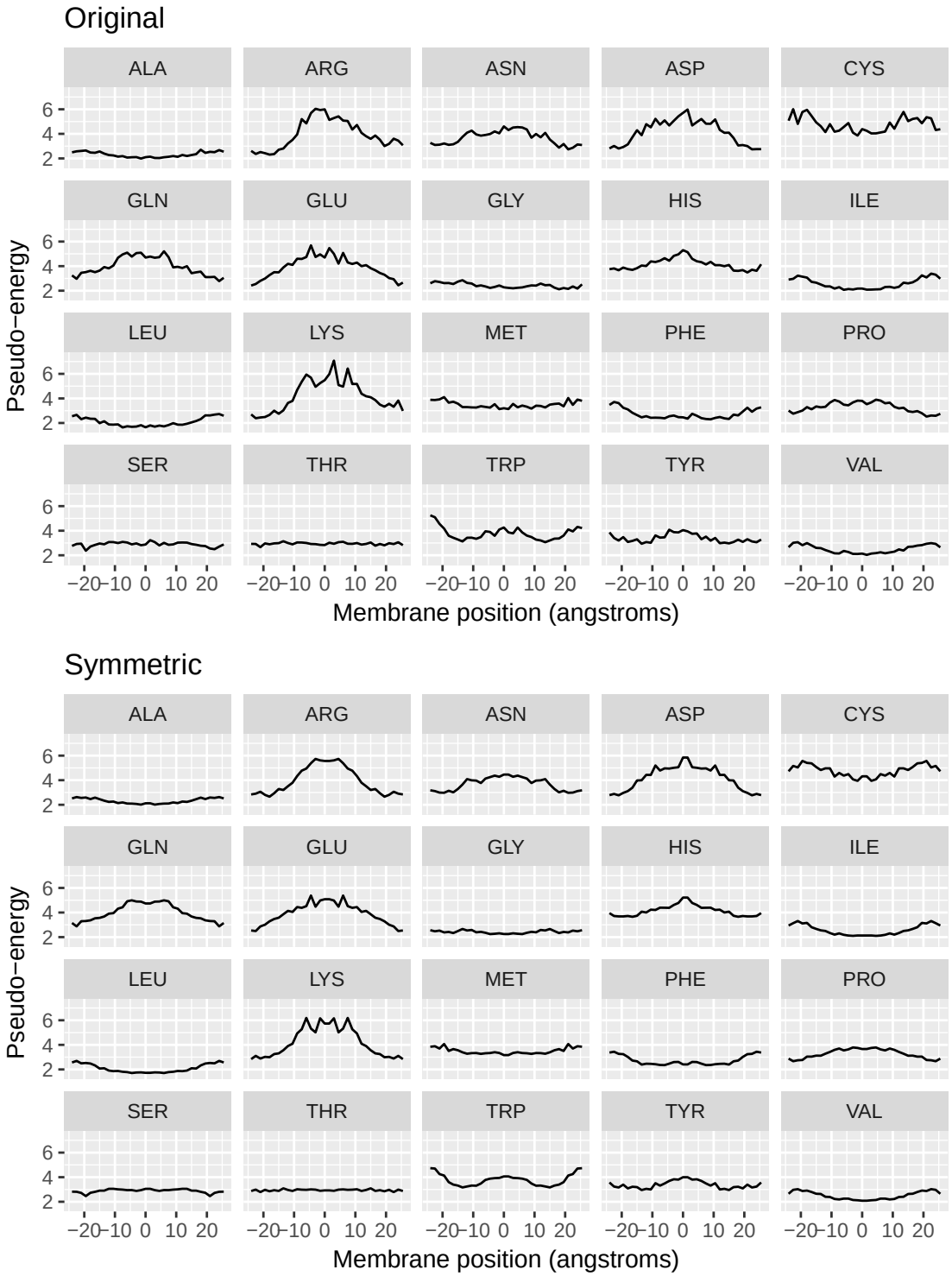


Figure 4.3: Membrane potential pseudo-energy for each amino acid. "Original" shows the pseudo-energies published by Nugent and Jones (2013); "Symmetric" shows the pseudo-energies after averaging the values at ± each distance from the centre.

The potential was normalised for the length of the peptide by dividing by the number of residues in the chain, and multiplied by a factor of 1,000 so that the standard deviation of values across decoys was comparable to other scores.

To establish the best weight to use to combine the membrane potential with the other potentials, I used subsets of the decoy populations generated by SAINT2-ScafFold (adjusted moves) in both Forward and Reverse directions. In order to learn from all decoy sets together, despite very different ranges of potentials, each potential for each set of decoys was independently normalised. For each decoy, a z-score for each potential was calculated by subtracting the mean value of the potential for the decoy set and dividing by the standard deviation for that set. To adjust the TM-score ranges to be more similar, I used equal numbers of good decoys (TM-score > 0.5) and bad decoys from each decoy set. A sample of the size of the smaller group (good or bad) was taken from the larger group. R was used to calculate a matrix of correlation coefficients and to fit a linear regression model for various combinations of the SAINT2 score components, with TM-score as the output variable. The Akaike information criterion (AIC) was also calculated in R for each model.

The coefficients reported in Section 4.2.2.1 were also calculated without equalising the numbers of good and bad decoys, and with an additional requirement for bad decoys to be between 0.4 and 0.5, to remove the very worst decoys and focus on those closer to the native structure. For each of these methods of subsetting the data, and for the two combined, a similar overall ranking of the importance of each potential was obtained.

4.1.3 Datasets

Separate training and test sets were used, similar to the sets of shorter membrane proteins described in Section 3.2.1 ($\text{Set2A}_{\text{train}}$ and $\text{Set2A}_{\text{test}}$). Table 4.1 lists all targets used, their lengths, the number of residues in each segment, and the length to be predicted (non-segment).

	PDB code	length	TMHs	Forward		Reverse	
				segment length	length to predict	segment length	length to predict
Training set	1orsC	132	4	56	76	51	81
	2xowA	179	6	75	104	43	136
	4a2nB	192	5	61	131	65	127
	4o6yA	210	6	56	154	56	154
	1kqfC	216	4	77	139	106	110
	3b4rB	216	7	56	160	48	168
	4b4aA	225	6	84	141	45	180
	3klyA	257	7	63	194	69	188
	2w2eA	263	8	93	170	48	215
	4od5A	274	9	48	226	43	231
	1okcA	292	6	94	198	84	208
	4n7wA	307	10	48	259	57	250
	2qi9A	324	10	78	246	49	275
	4ezcA	345	12	53	292	74	271
	1zcdA	376	12	73	303	45	331
	4bwzA	384	13	40	344	65	319
	3cx5C	385	8	102	283	63	322
Test set	3rlbA	176	6	43	133	74	102
	1e12A	239	7	58	181	63	176
	2vpzC	250	8	64	186	41	209
	2dyrC	259	7	56	203	62	197
	3m73A	313	10	64	249	58	255
	1u7gA	383	11	62	321	74	309

Table 4.1: ScafFold targets, lengths, and segment lengths for each direction of prediction. For each target, the segment length includes up to the end of the first two TMHs, either starting at the N-terminus (Forward) or C-terminus (Reverse).

4.2 Results and discussion

4.2.1 SAINT2-ScafFold performance

ScafFold is a new version of SAINT2, created in order to simplify the prediction of membrane proteins and investigate folding from a possible intermediate in the folding pathway. SAINT2-ScafFold begins from a segment of the native structure, and completes the structure through a series of move and extrusion steps (Section 4.1.1.1).

4.2.1.1 Comparison of SAINT2-ScafFold and SAINT2-Wholly

ScafFold was initially tested on the training set of membrane proteins described in Section 4.1.3. For each target, the segment chosen for each protein was the length of the peptide up to the end of the second transmembrane span, identified by the RosettaMP `mp_span_from_pdb` (Alford *et al.*, 2015) annotation (see Section 3.2.1.1). These segments were used to test Forward ScafFold, while the final two transmembrane spans were used as a segment for Reverse ScafFold. For both directions, the adjustments to the number of growth moves and full length moves described in Section 4.1.1.1 were used. 10,000 decoys were generated in each direction using identical fragment libraries and predicted contacts to those used for full length structure prediction (SAINT2-Wholly) in Section 3.3.4. TM-scores were calculated between the non-segment part of each decoy and the same part of the native structure, and the SAINT2-Wholly decoys generated in Section 3.3.4 were rescored in the same way (see Section 4.1.1.2).

Figure 4.4 compares the distributions of TM-scores seen for each method. SAINT2-ScafFold had little effect on the distribution of TM-scores for each protein when compared to SAINT2-Wholly. I would have expected that a correct starting segment should aid the folding of subsequent helices into the correct positions, but a consistent improvement was not seen. For a few targets, particularly the shorter ones such as 2xowA and 4o6yA, the median TM-score was higher using ScafFold, however it was only for 4o6yA that there was a

noticeable improvement in the accuracy of the best decoy. There are targets where the inverse is true and ScafFold generated a worse best decoy, for example 4b4aA. Figure 4.5 shows similar results for the Reverse direction.

In Figure 4.4, the TM-score distributions are also shown of 5,000 decoys generated by a third method: SAINT2 for the second half of the protein only, using the adjusted number of moves that was used for ScafFold. This method provides context to the differences seen between SAINT2-Wholly and SAINT2-ScafFold. For 4o6yA and 1kqfC, it is worse by around 0.1 TM-score units, and for 1orsC it actually performs far better, but in most cases there are no large differences. It seems that the lack of any N-terminus altogether can be detrimental, but in these cases, it is not always true that provision of the correct N-terminus (i.e. SAINT2-ScafFold) is better than allowing the full chain to be sampled. For example, this is not the case in 1kqfC and 4od5A.

Figure 4.6 shows the best decoys produced by SAINT2-Wholly and SAINT2-ScafFold for the targets 4b4aA and 4o6yA. For 4b4aA, the segment includes a short interfacial helix which bends towards the C-terminal helices. In the native structure, the helices pack closely against each other; however, the best ScafFold decoy is very loosely packed. It may be that there were no appropriate fragments for the loop between transmembrane helices three and four, which must avoid clashes with the static segment. Using SAINT2-Wholly, there is more flexibility, and in order to get a more correct C-terminus, the N-terminal conformation is quite different from the native structure. The models for this target suggest that a lack of flexibility in the segment may pose a particular challenge when subsequent loops interact extensively with it, and the fragment library may not be diverse or accurate enough to allow this. The lack of flexibility may also generally affect the ability to pack helices closely, as the segment side of an interaction between two helices cannot compensate for an odd shape in the other helix. Another possible issue that may cause SAINT2-ScafFold to perform poorly on some targets is a difficult "take-off" point at the end of a segment. The take-off point is the fixed position and direction of the final residue of the

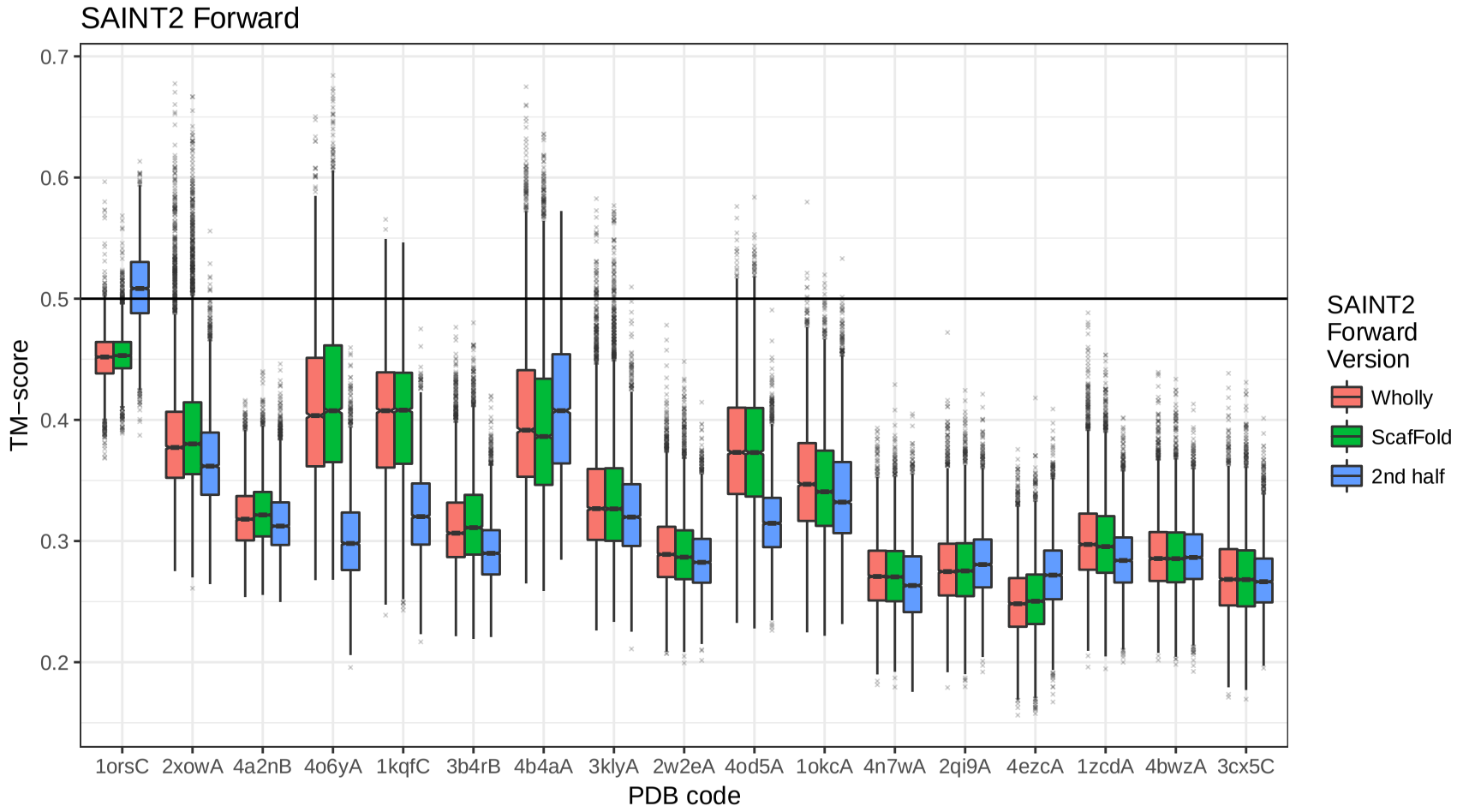


Figure 4.4: Boxplots to compare the TM-scores of the non-segment region of decoys in the Forward mode. The three methods shown are SAINT2-Wholly (full sampling), ScafFold, and non-segment region only (2nd half). For each method, the results for 5,000 decoys are shown. This enables a direct comparison to the 2nd half test, for which only 5,000 decoys were generated. Targets are ordered by increasing length.

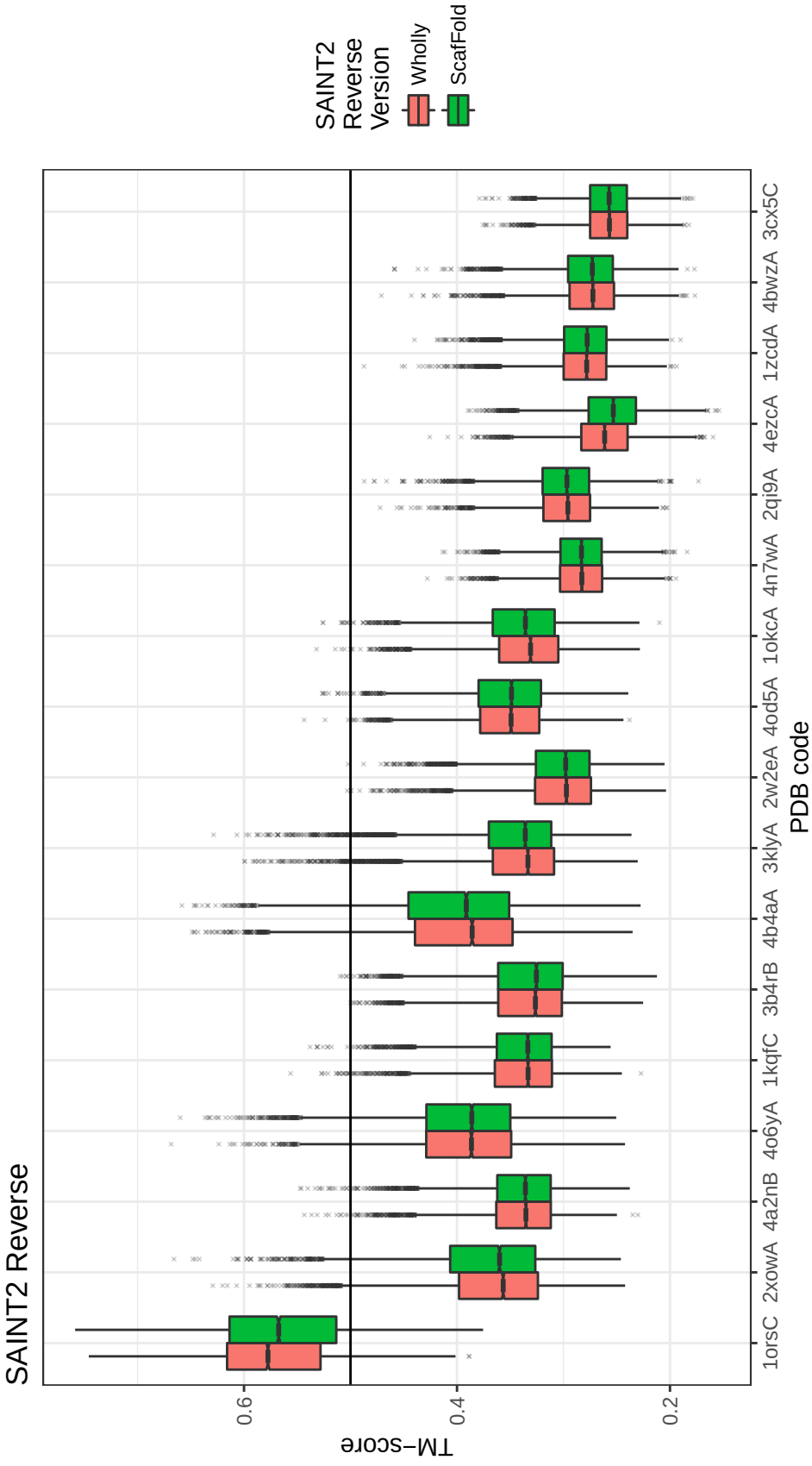


Figure 4.5: Boxplots to compare the TM-scores of the non-segment region of decoys in the Reverse mode. The two methods shown are SAINT2-Wholly (full sampling) and ScafFold. For each method, the results for 10,000 decoys are shown. Targets are ordered by increasing length.

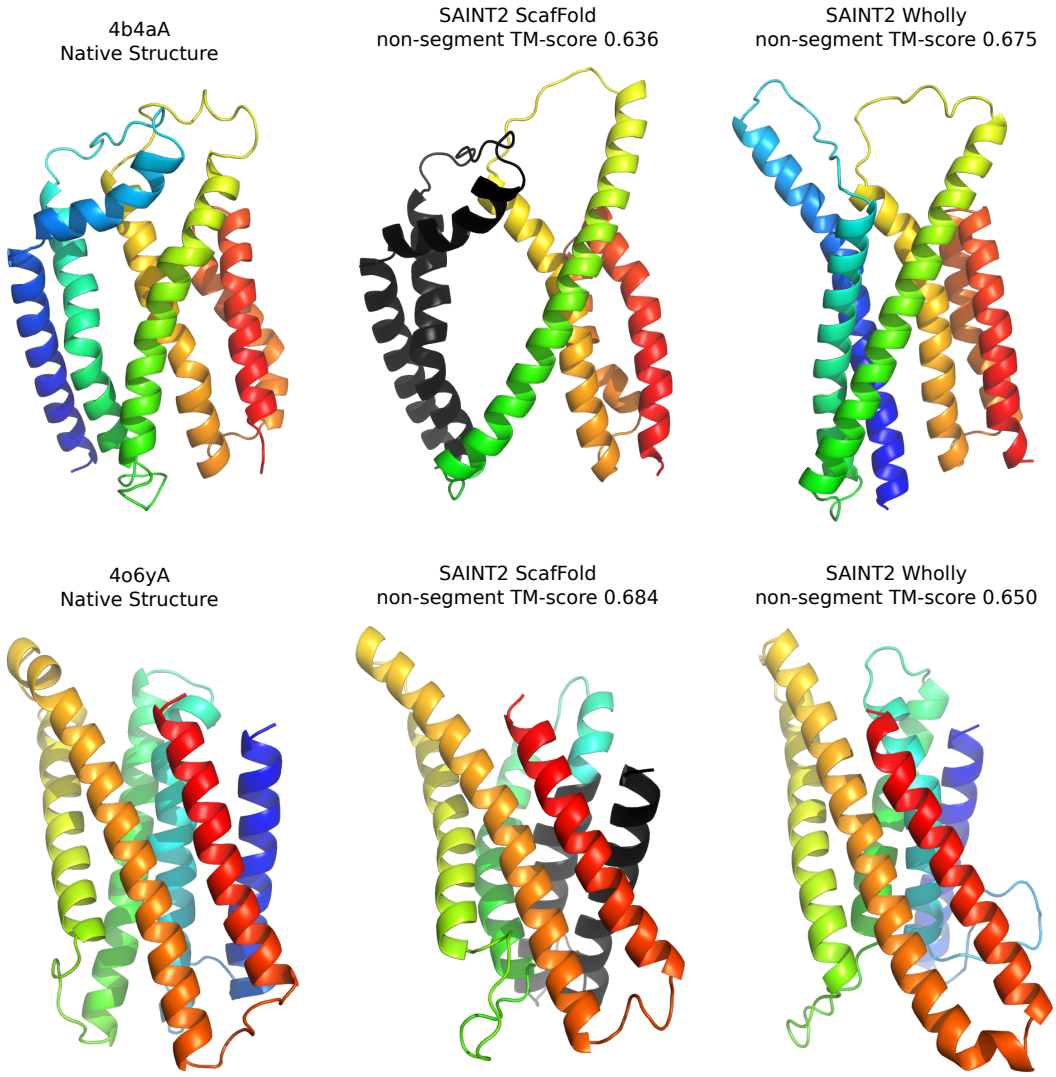


Figure 4.6: Comparison of models generated in the Forward direction by SAINT2-ScafFold and SAINT2-Wholly for two targets. The segment region at the N-terminus is shown in black.

segment, and in some cases it may require an unusual conformation of the backbone in the loop so that the following secondary structure is positioned correctly. This is especially challenging due to the limitations of using a fragment library that may not include good fragments for the following loop. In addition to these problems, it is also possible that SAINT2 scoring does not work as well when "unrelaxed" parts of native structures are used during modelling.

On the other hand, 4o6yA demonstrated the expected improvement for the ScafFold protocol, where the presence of the segment appears to aid the

assembly of the remaining structure. Transmembrane helices five and six in the ScafFold model pack closely against the segment. In the best SAINT2-Wholly model, the first transmembrane helix is behind helix two and does not interact with the C-terminus.

4.2.1.2 The effect of segment length on TM-score

In order to check that SAINT2-ScafFold was functioning in the intended way, 4b4aA and 4o6ya, for which the greatest difference was seen between SAINT2-ScafFold and SAINT2-Wholly, were investigated more thoroughly. Both proteins were predicted using SAINT2-ScafFold from increasing sizes of segments, generating 4,600 decoys in each case. The shortest segment, 10 residues, should have a negligible effect, and is important to show that the ScafFold implementation has not caused any functions of the original SAINT2 to fail in their intended purpose. The other lengths of segment were from the N-terminus to the end of the first transmembrane span, and to two residues into the second transmembrane span. For every length of segment, TM-scores were calculated in the same way, over the residues not included in the longest segment.

Figure 4.7 shows that the difference in TM-scores between different segment lengths was very marginal. Segment lengths of zero (SAINT2-Wholly) and 10 gave almost the same results, therefore the implementation of SAINT2-ScafFold appeared not to confer a disadvantage. Increasing the segment size to the end of the first span also changed very little. This extension of the segment does not carry any information about the positioning of the second transmembrane helix and therefore it is not surprising that results are similar to having no segment. When the loop to the start of the second span is included in the segment, there are a couple of very good outlying best decoys for 4b4aA, but no significant changes to the distribution for either target. Therefore almost all of the difference seen in the last section between SAINT2-ScafFold and SAINT2-Wholly is introduced with the addition of the second span to the segment. Due

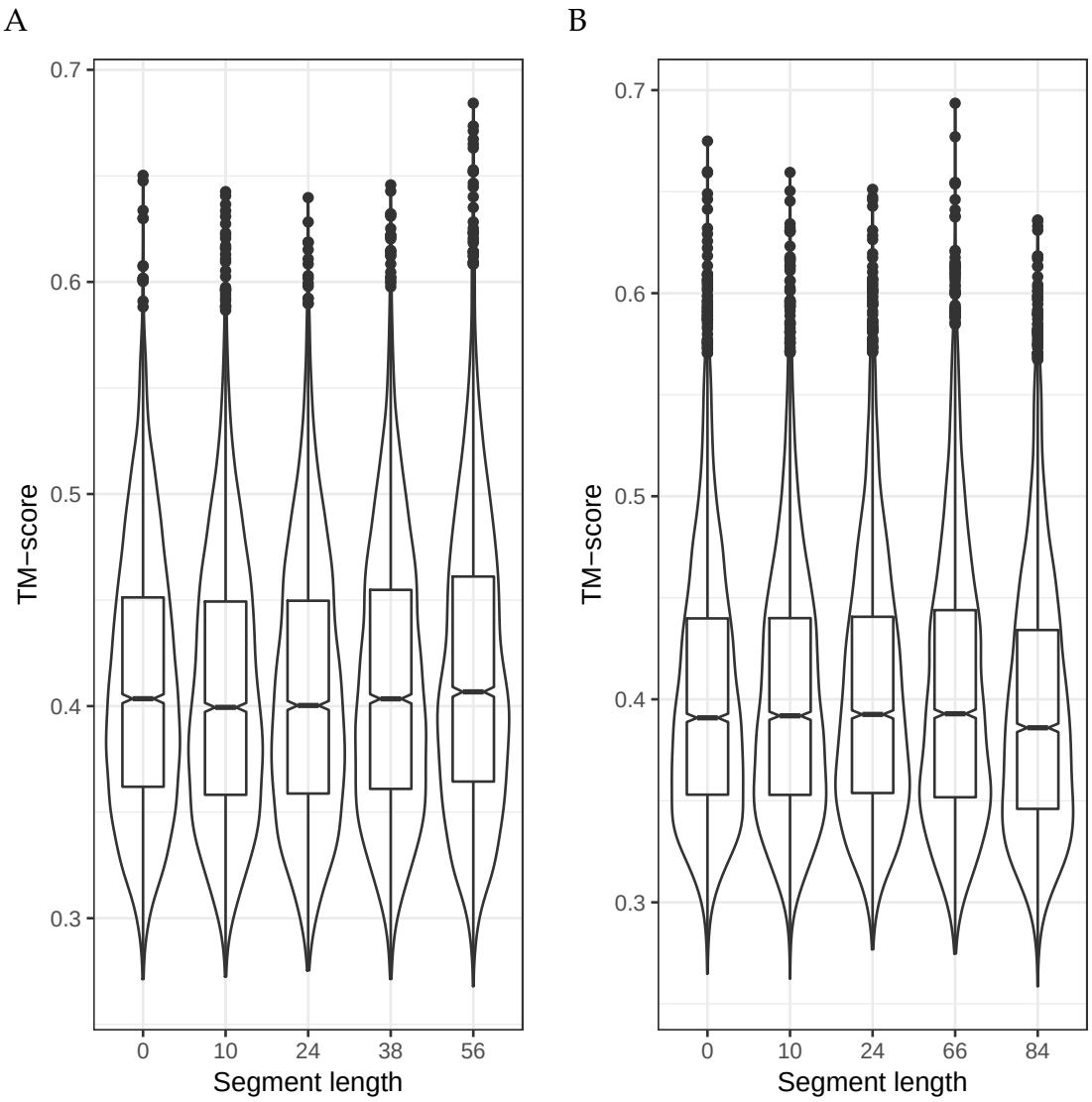


Figure 4.7: Box and violin plots for two targets predicted in Forward mode showing the effect of segment length on TM-score. For every decoy, the TM-score was calculated over the residues that were not part of the longest segment length on the x-axis for that protein, i.e. the residues whose conformations were sampled under every set up tested for that protein. The segment lengths are 0 (SAINT2-Wholly), 10, end of first transmembrane span (24 in both cases), two residues into the second transmembrane span, end of second transmembrane span. For each length, the results for 4,600 decoys are shown. A) 4o6yA, scored from residue 57 onwards. B) 4b4aA, scored from residue 85 onwards.

to this, it seems likely that the take-off point for the remainder of the protein is a critical factor in the success or failure of SAINT2-ScafFold.

4.2.1.3 SAINT2-ScafFold without adjustment for the number of moves

To investigate one further difference between SAINT2-Wholly and SAINT2-ScafFold, 5,000 decoys were generated by SAINT2-ScafFold without the adjusted moves protocol. Figure 4.8 compares ScafFold with 10,000 moves (full moves) to the results shown previously for SAINT2-ScafFold and SAINT2-Wholly. For 4b4aA and 1okcA, SAINT2-ScafFold (full moves) recovers the performance seen for SAINT2-Wholly. For the longer targets, there is little difference, particularly because the sampled region includes the majority of the protein so the move adjustment does not greatly reduce the number of moves.

4.2.2 Implementation of the membrane potential

In this section, I test whether use of a membrane potential improves prediction. I use the SAINT2-ScafFold protocol as that allows the membrane position of the protein to be estimated. Residues not in the segment can then be scored according to their propensity to be found at a specific depth in the membrane. Such a knowledge-based membrane potential incorporates the physical differences between side chains and the energy of transferring them from an aqueous to hydrophobic phase. A membrane potential should improve the adoption of the correct fold, as it should encourage transmembrane helices to adopt positions crossing the "membrane", and also approximately parallel to the previous helices. As described in Section 4.1.2, I implemented a potential based on that described by Nugent and Jones (2013), using their knowledge-based energies for each amino acid at different depths in the membrane.

4.2.2.1 Testing the membrane potential for decoy ranking

I first tested whether the membrane potential could be used to find good decoys within a population of decoys generated under the standard SAINT2 potential.

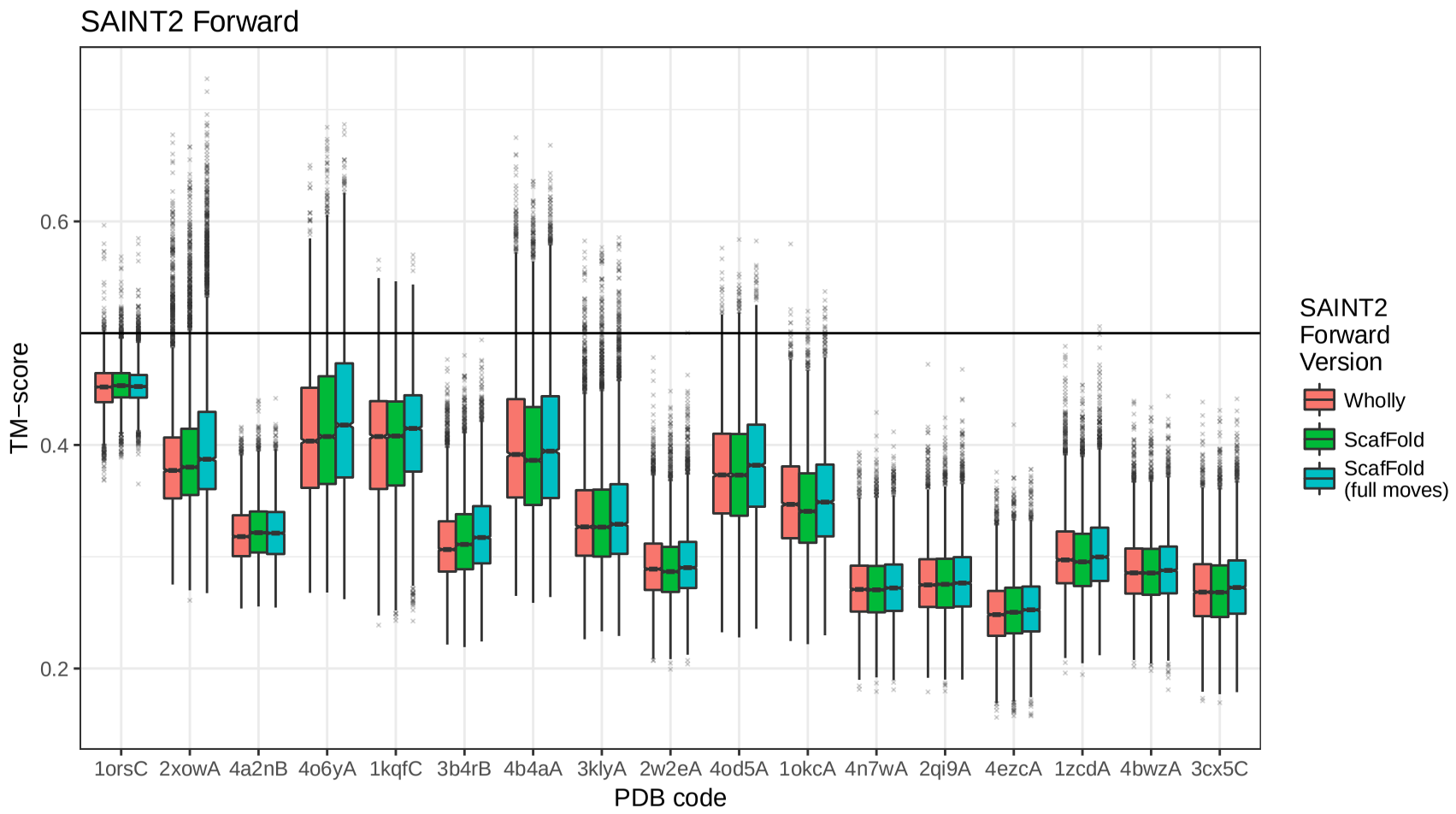


Figure 4.8: Boxplots to compare the TM-scores of the non-segment region of decoys. The three methods shown are SAINT2-Wholly (full sampling), ScafFold, ScafFold (full moves) with no adjustment i.e. 10,000 moves. For each method, the results for 5,000 decoys are shown. Targets are ordered by increasing length.

	TM-score	Solvation	Orientation	RAPDF	Lennard-Jones	Contact	Membrane
TM-score	1.000	-0.052	-0.056	-0.107	-0.063	-0.428	-0.183
Solvation	-0.038	1.000	0.329	0.737	0.003	0.263	-0.020
Orientation	-0.056	0.312	1.000	0.318	0.019	0.188	0.036
RAPDF	-0.098	0.728	0.304	1.000	0.054	0.331	-0.019
Lennard-Jones	-0.094	0.004	0.018	0.065	1.000	0.077	0.015
Contact	-0.456	0.248	0.185	0.316	0.095	1.000	0.145
Membrane	-0.183	-0.024	0.040	-0.019	0.021	0.157	1.000

Table 4.2: Correlation coefficients between the components of the SAINT2 scoring function and TM-score. The Pearson correlation coefficient is given in the upper triangle, and Spearman's rank correlation coefficient is given in the lower triangle. Coefficients are shaded according to a scale from -1 (red) to 1 (blue) through white (0).

The sets of decoys generated by SAINT2-ScafFold in both Forward and Reverse directions described in Section 4.2.1 were re-scored to evaluate the membrane potential for the completed decoys. In this section, I calculated the TM-score of the entire decoy, including the fixed segment, against the full native structure. This ensured that the relative positioning of the additional helices compared to the first two embedded helices was accounted for in the score. The distributions of TM-scores were quite different for the different populations of decoys, due to the range of difficulty of the targets. Therefore, in order to combine the results for many different proteins, I used a number of ways to subset and normalise the data, described in Section 4.1.2. Initially, the correlation was evaluated between the components of the score and the TM-score against the native structure, giving the coefficients in Table 4.2. A negative correlation indicates that a potential could be useful within decoy generation by SAINT2, as a higher TM-score against the native structure should have a lower, more favourable SAINT2 score.

The membrane potential had a correlation coefficient of ~ -0.18 against TM-score, stronger than all other potentials except the contact potential. Figure 4.9 shows the correlation between the membrane potential and TM-score for three typical individual sets of decoys, and before normalisation, the combined set of all decoys from different proteins (A). The plot for all decoys demonstrates the need to normalise each decoy set individually, as some sets universally had higher membrane potential scores. For some sets of decoys, the correlation was strong, and in other sets there was almost no correlation. The solvation potential has by far the weakest correlation with TM-score of all the score components. This score would not be expected to show good correlation as it was trained on soluble proteins. Soluble proteins will show different trends for surface-exposed residues, as the membrane environment allows hydrophobic residues to be exposed while an aqueous environment does not.

In order to establish the contribution of each potential in predicting the TM-score of a decoy, linear models were built using different combinations of the normalised potentials, shown in Table 4.3. Akaike information criterion (AIC) values for each model are also given. It is clear that such a model is not the best reflection of the relationship between the potentials and the TM-score. However, it is a simple way to assess how decoys could be ranked by summing the potentials according to different weights and to establish the most important potentials for this purpose.

The coefficients for the solvation potential were positive, which indicated that worse decoys were being scored more favourably by this potential. There is no justification for using the solvation potential with a negative weight, therefore this implied that the solvation potential was not useful for ranking decoys. It is also unlikely to be helpful in the process of decoy generation, therefore it was not used in the following sections when decoys were generated to test the membrane potential.

When looking at the linear models that do not include a solvation potential or contact potential, the orientation and RAPDF potentials had negative coefficients,

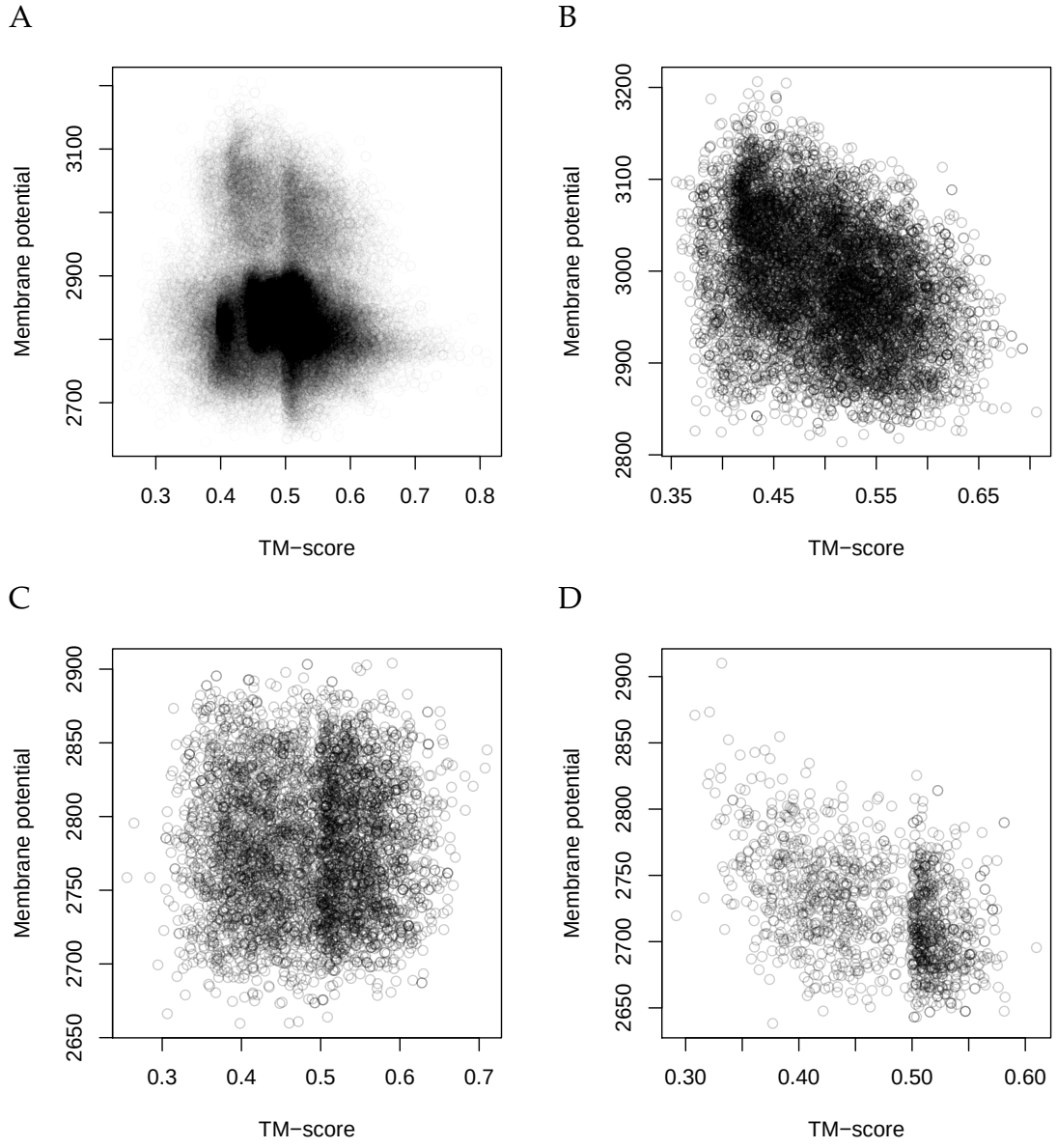


Figure 4.9: Correlation between membrane potential and TM-score in A) All decoy sets B) 1kqfC (Forward mode) C) 4b4aA (Reverse mode) D) 3b4rB (Forward mode). For each set of decoys, an equal number of good (TM-score > 0.5) and bad (TM-score < 0.5) decoys is shown.

Intercept	0.4926	0.4918	0.4888	0.4885	0.4926	0.4918	0.4888	0.4884
Solvation					0.0041	0.0038	0.0050	0.0048
Orientation	-0.0017	-0.0011	0.0011	0.0013	-0.0021	-0.0015	0.0006	0.0008
RAPDF	-0.0065	-0.0070	0.0024	0.0017	-0.0094	-0.0097	-0.0011	-0.0016
Lennard-Jones	-0.0039	-0.0037	-0.0021	-0.0021	-0.0037	-0.0035	-0.0019	-0.0019
Contact			-0.0297	-0.0284			-0.0298	-0.0284
Membrane		-0.0124		-0.0082		-0.0124		-0.0082
AIC	-149581	-151600	-160611	-161654	-149676	-151685	-160781	-161812

Table 4.3: Coefficients for each component of the SAINT2 scoring function in linear models to predict TM-score. Each column represents a separate linear regression model, and where a score component is not included in a given model, the cell is left blank. Coefficients are shaded according to a scale from -0.03 (red) to 0.03 (blue) through white (0). AIC values are shaded according to a scale from the least negative (white) to the most negative (green).

as would be expected. The positive coefficients when the contact potential was included could be due to the correlation between these two potentials and the contact potential, and overtraining on this set of decoys. As the orientation, RAPDF and Lennard-Jones potentials are important parts of the model when the contact potential is not included, these potentials were used during decoy generation in the following sections to test the membrane potential. As these potentials form the core of the SAINT2 potential and it is known roughly what weighting is appropriate to use for these relative to the contact potential, the weights were not changed.

For each combination of the previously existing SAINT2 potentials shown, addition of the membrane potential into the model significantly improved the AIC values. Judging by the magnitude of coefficients, the membrane potential appeared to be more important than the orientation, RAPDF and Lennard-Jones potentials, but less important than the contact potential. In models that included both the contact potential and membrane potential, the contact potential had the most negative coefficient, followed by the membrane potential.

During decoy generation by SAINT2, it is not possible to use any kind of normalisation to weight the score components, as the mean and standard deviation for a given potential and target cannot be known until a pool of decoys has been generated. The previously implemented potentials in SAINT2 had already been adjusted to give similar ranges of values, and Section 4.1.2 explains how this was now carried out for the membrane potential. Therefore, as the potentials in SAINT2 now had similar spreads, the relative importance of the potentials in the linear models constructed above could then be used to choose a range of weights for the membrane potential that were likely to be effective. Weights chosen were 0 (control), 0.5, 1, and 10, to cover the range that would be likely to perform best. The weight of 0.5 sits between the weights for RAPDF (0.303 for short peptides, 0.156 for long) and the contact potential (1).

4.2.2.2 Testing the membrane potential during decoy generation

Having established approximately the correct range of weights to use for the membrane potential, and without the solvation potential, 10,000 decoys were generated for each weight as in Section 4.2.1. The boxplots in Figures 4.10 and 4.11 compare the distributions of TM-scores for decoys generated by SAINT2-ScafFold. The TM-scores shown in this section are for the entire protein, including the static segment, against the full native structure, in order to assess the relative position of the second half of the protein to the first half. The left-most box for each target PDB shows the previous results from Section 4.2.1. The remaining four boxes show the results with the range of weights chosen for the membrane potential (MP), in all cases using no solvation potential.

The differences between the different weighting set-ups are small, particularly in terms of the quality of the average decoys, shown by the position of the boxes. Only for one target, 1orsC in Reverse mode, was there a large shift to the whole distribution of TM-scores seen. In this case, the MP weight of 0.5 appeared to generate far more 'correct' answers of TM-score > 0.5 than the other set-ups, and a longer tail of better decoys stretching up to a TM-score of 0.75. Comparing

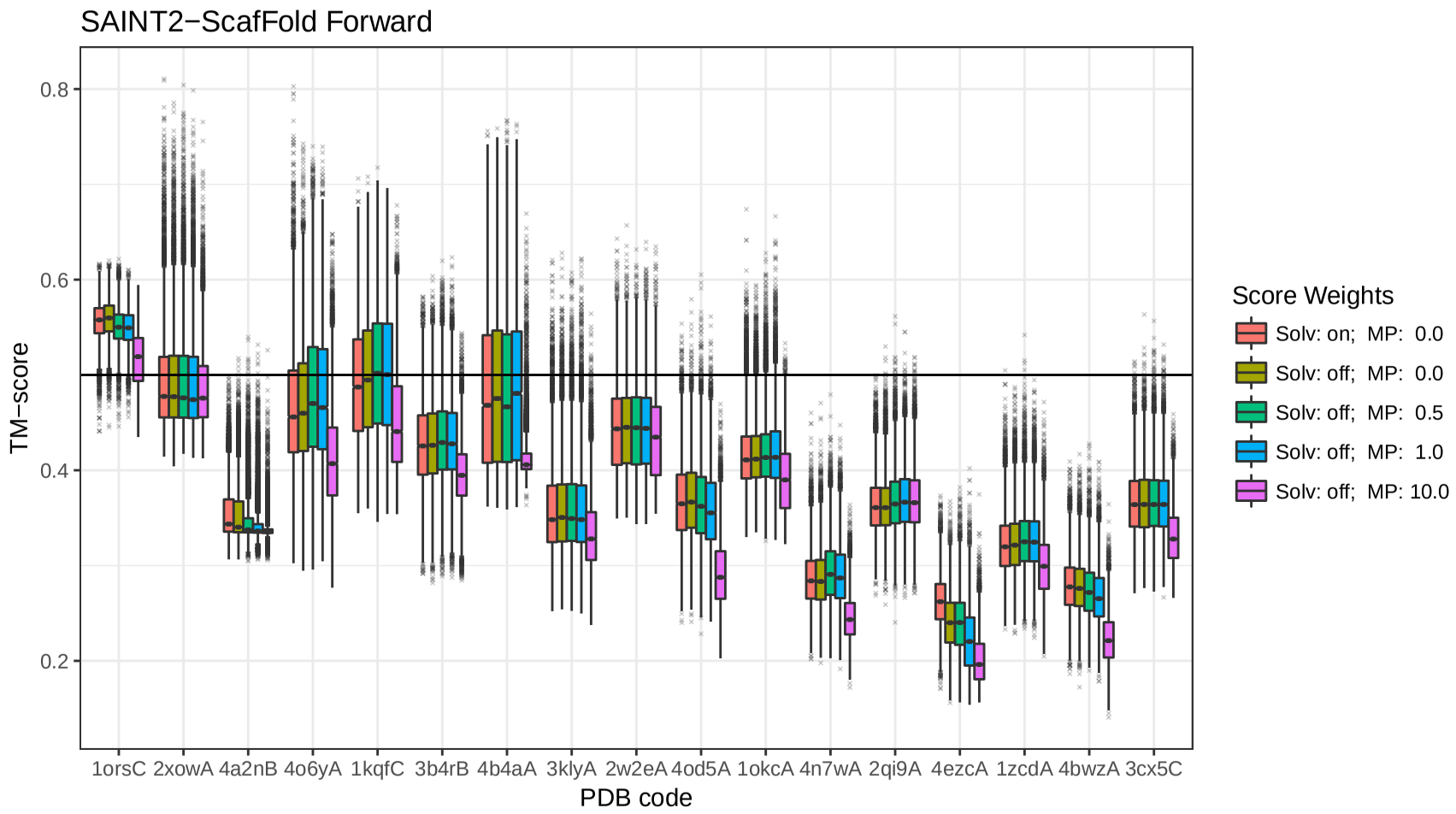


Figure 4.10: Boxplots of the TM-scores of decoys generated by the Forward SAINT2-ScafFold protocol, starting from a segment of two transmembrane spans. Decoys were generated for each target using a range of weights for the membrane potential (MP) component of the score, and with the solvation (Solv) potential either on or off, shown in different colours. Targets are ordered by increasing length.

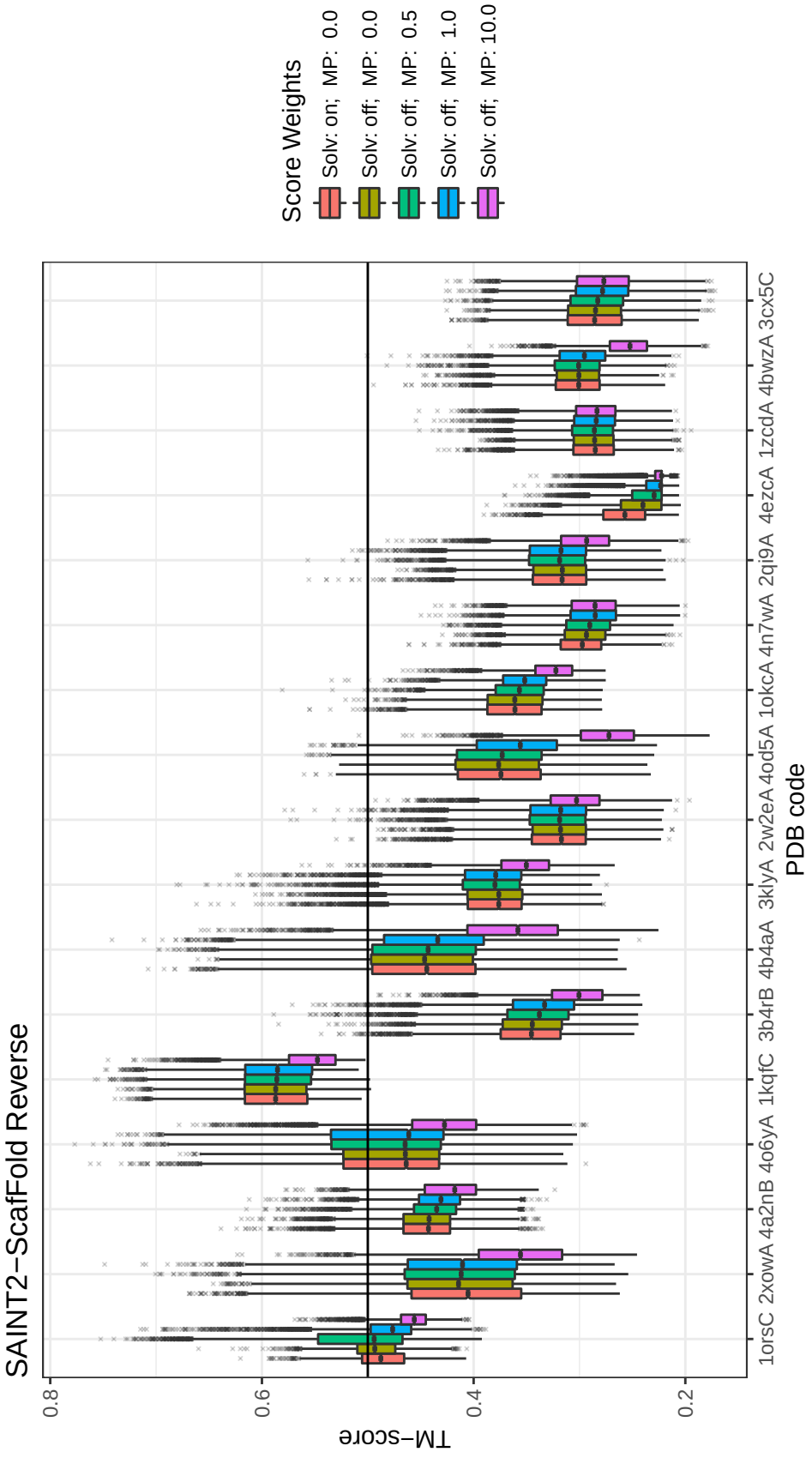


Figure 4.11: Boxplots of the TM-scores of decoys generated by the Reverse SAINT2-ScafFold protocol, starting from a segment of two transmembrane spans. Decoys were generated for each target using a range of weights for the membrane potential (MP) component of the score, and with the solvation (Solv) potential either on or off, shown in different colours. Targets are ordered by increasing length.

the set-ups with zero MP weight, the solvation potential barely influenced the overall distribution. For many targets, the median decoy generated using an MP weight of 10 was at least 0.05 TM-score units worse than all other methods. This suggests that the weights chosen were appropriate, and giving greater weight to the MP is unlikely to produce better results.

While the median and quartiles were similar in the majority of cases, Figures 4.10 and 4.11 also show the outliers. Figure 4.12 shows comparisons of the best TM-score generated for each target by the different set-ups. Comparing the very best TM-score output by each set-up, there appeared to be greater differences.

Figure 4.12A shows the effect of removing the solvation potential, which appears to slightly improve the accuracy of Forward predictions yet most Reverse predictions are worse. It is difficult to understand why such a change would affect the two directions of prediction differently. The results of Chapter 3 indicate that the structures of N- and C-termini of membrane proteins are differently structured. In an artificial situation where two helices are held rigidly, there may be a different effect on the folding of the remainder of the protein if they are not closely interacting, as may be more common in the Reverse set-ups. Overall, removal of the solvation potential does not cause a large effect, therefore it is reasonable to interpret the different membrane weight results with reference to the set-up with no solvation potential.

The MP weight of 0.5 improves the best TM-score obtained for most targets using both Forward and Reverse (Figure 4.12B). In three cases, a decoy was completed to give a correct fold, where no correct answer was generated using no MP. Figures 4.12C and D show that weights of 0.5 and 1 give similar results, but 0.5 appears to be slightly better for both Forward and Reverse. Overall, there is some evidence to support that a moderate MP weight of 0.5 may improve the ability of SAINT2 to generate decoys using the ScafFold method and the embedding estimated using the full native structure.

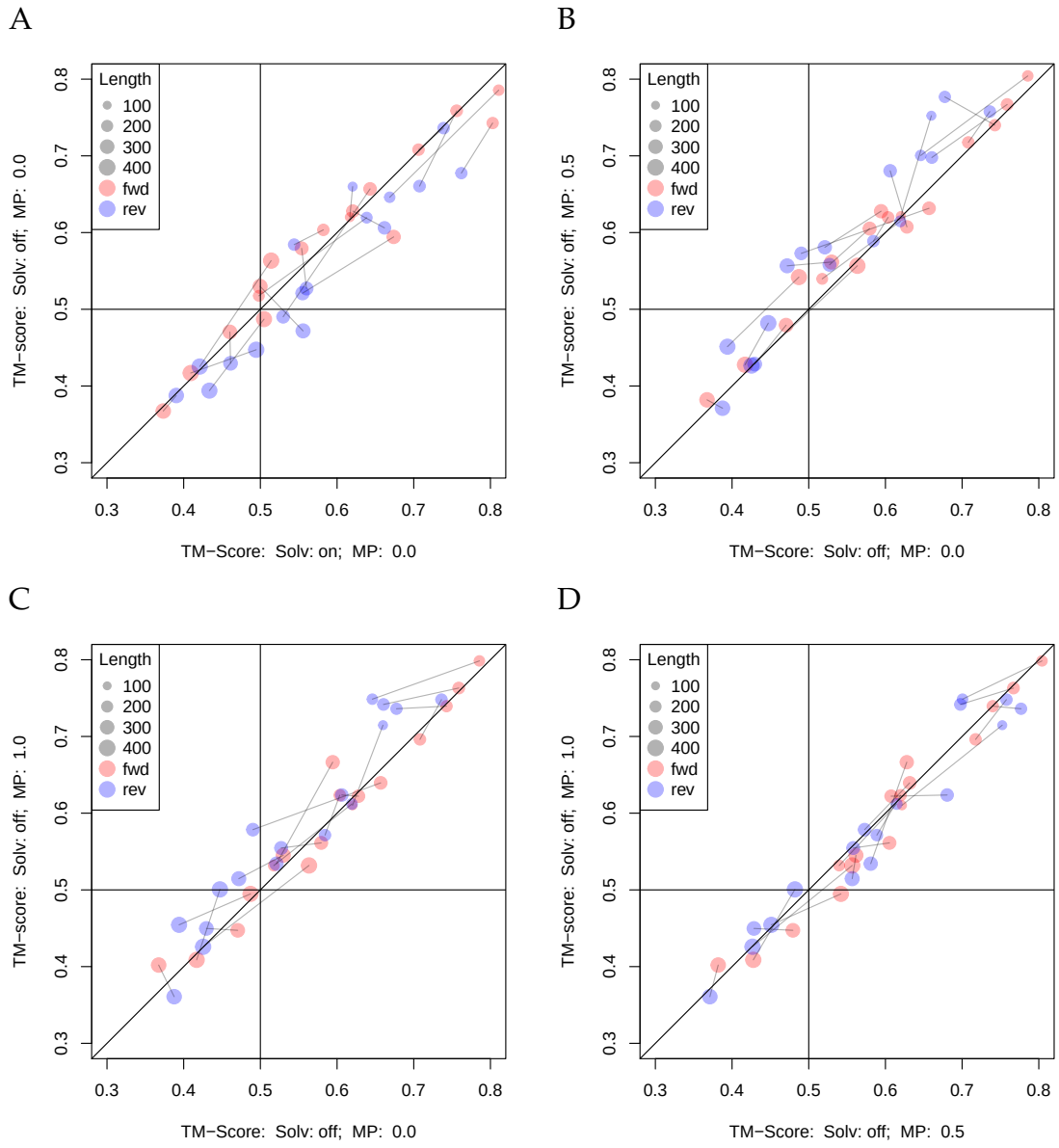


Figure 4.12: Comparisons of the TM-score of the best decoy generated of 10,000 decoys for each target. The *x*- and *y*-axes show the weights used for the membrane potential (MP) component of the score, and whether the solvation (Solv) potential was on or off. Forward (fwd) results are shown in red, with Reverse (rev) in blue, and the Forward and Reverse results for the same PDB target are connected by a grey line.

In a realistic prediction scenario, the embedding of the first segment of the protein could only be estimated using the hydrophobicity of the segment helices or homology to existing proteins using iMembrane (Kelm *et al.*, 2009). This would introduce additional sources of error, and the membrane position should be sampled during decoy generation in order to avoid the MP guiding helices incorrectly. Therefore, the MP may not be as effective in protein structure prediction, but the success of the MP during decoy generation does indicate that it helps to guide folding in a more efficient way. The improvement could even reflect that decoy generation using the MP is closer to *in vivo* folding.

4.2.2.3 Ranking of decoys generated with membrane potential

The results observed in Section 4.2.2.1 were used to estimate the combination of scores that would be most effective for ranking complete decoys. On this basis, new combinations of weights were tested for decoy generation, with some success. The optimal combination of weights for the ranking of decoys generated under the new potential weights may differ from those used before. Therefore, the same method was used to build a linear regression model to predict TM-score based on the scores of decoys generated with membrane potential weight 0.5 in Section 4.2.2.2.

Tables 4.4 and 4.5 show similar correlation and linear model coefficients to those in Tables 4.2 and 4.3. One difference was that for the decoys generated with the membrane potential and no solvation potential, the MP was negatively correlated with the other potentials, except the contact potential. The linear model that included RAPDF, Lennard-Jones, contact and membrane potentials was chosen to be used for decoy ranking, as coefficients for the orientation potential were positive or very small. This model will be referred to as LM1; the model that includes only terms for RAPDF, Lennard-Jones, and membrane potentials is referred to as LM2 and is used in Chapter 5. Additional models (not shown) that included a solvation potential term were also fitted, but the coefficients for the solvation potential were positive, as found previously.

	TM-score	Solvation	Orientation	RAPDF	Lennard-Jones	Contact	Membrane
TM-score	1.000	-0.086	-0.066	-0.171	-0.081	-0.446	-0.100
Solvation	-0.080	1.000	0.364	0.792	0.043	0.321	-0.132
Orientation	-0.073	0.354	1.000	0.325	0.046	0.210	-0.067
RAPDF	-0.171	0.786	0.318	1.000	0.099	0.378	-0.145
Lennard-Jones	-0.106	0.062	0.056	0.127	1.000	0.108	-0.020
Contact	-0.478	0.298	0.208	0.374	0.135	1.000	0.051
Membrane	-0.091	-0.136	-0.069	-0.160	-0.029	0.046	1.000

Table 4.4: Correlation coefficients between the components of the SAINT2 scoring function and TM-score. Data shown for decoys generated with a membrane potential weight of 0.5 and no solvation potential. The Pearson correlation coefficient is given in the upper triangle, and Spearman's rank correlation coefficient is given in the lower triangle. Coefficients are shaded according to a scale from -1 (red) to 1 (blue) through white (0).

		LM2		LM1				
Intercept	0.4967	0.4963	0.4938	0.4936	0.4967	0.4963	0.4937	0.4936
Solvation								
Orientation					-0.0008	-0.0010	0.0023	0.0021
RAPDF	-0.0118	-0.0132	0.0000	-0.0012	-0.0116	-0.0129	-0.0007	-0.0017
Lennard-Jones	-0.0047	-0.0048	-0.0024	-0.0025	-0.0047	-0.0047	-0.0024	-0.0025
Contact			-0.0316	-0.0309			-0.0318	-0.0311
Membrane		-0.0092		-0.0058		-0.0092		-0.0058
AIC	-155987	-157037	-167922	-168421	-155992	-157046	-167991	-168479

Table 4.5: Coefficients for each component of the SAINT2 scoring function in linear models to predict TM-score. Data shown for decoys generated with a membrane potential weight of 0.5 and no solvation potential. Each column represents a separate linear regression model, and where a score component is not included in a given model, the cell is left blank. Coefficients are shaded according to a scale from -0.0318 (red) to 0.03 (blue) through white (0). AIC values are shaded according to a scale from the least negative (white) to the most negative (green).

Ranking each population of decoys using LM1, I selected the top ranked decoy and the top five ranked decoys to test how close these were to the best decoy in the set. Table 4.6 shows the TM-scores achieved for the top ranked, best of top five ranked and best overall decoy. Overall, the targets for which any correct decoy was generated corresponded to a shorter length to be predicted. Some targets were predicted correctly by Forward but not by Reverse, but in these cases the length to be predicted was longer in the Reverse mode. Where the TM-score of the best decoy was > 0.62, ranking by LM1 was able to select a correct structure (TM-score ≥ 0.5) as the top ranked decoy. Out of 34 training set segment build set-ups, a correct decoy was generated for 26. Of these 26, the top ranked decoy was correct in 16 cases, and in one case the top ranked decoy was not correct but there was a correct decoy in the five top ranked decoys.

4.2.2.4 Performance of decoy generation and ranking on test set

The SAINT2-ScafFold protocol was also used to predict the structures of the six protein chains in the test set, in both Forward and Reverse modes, starting from the terminal two helices. Table 4.6 shows the results for these targets, as they were ranked by LM1. The best TM-scores for targets in the test set were mostly lower than those targets of similar length in the training set. One exception was 3rlbA, which was predicted well using the score trained in this chapter for decoy generation, including the membrane potential. It should be noted that the other test set targets were some of the most difficult targets attempted in Section 3.3.4, with few or no correct answers (see Table 3.3). We would expect SAINT2-ScafFold to reduce the complexity of the problem compared to SAINT2-Wholly, and this led to correct answers for more than half of the 12 segment builds tested. Most of these correct "best" scores are in the region where LM1 failed to select a correct decoy for those targets in the training set ($\sim 0.5 - 0.6$). Therefore, it is not surprising that the top ranked decoys were not correct for these set-ups.

	PDB code	TMHs	Forward				Reverse			
			length to predict	TM-score			length to predict	TM-score		
				top1	top5	best		top1	top5	best
Training set	1orsC	4	76	0.56	0.56	0.62	81	0.52	0.52	0.75
	2xowA	6	104	0.66	0.76	0.80	136	0.60	0.66	0.70
	4a2nB	5	131	0.34	0.41	0.54	127	0.41	0.46	0.61
	4o6yA	6	154	0.62	0.68	0.74	154	0.64	0.71	0.78
	1kqfC	4	139	0.60	0.60	0.72	110	0.68	0.74	0.76
	3b4rB	7	160	0.52	0.57	0.62	168	0.41	0.43	0.59
	4b4aA	6	141	0.61	0.72	0.77	180	0.52	0.57	0.70
	3klyA	7	194	0.50	0.60	0.61	188	0.62	0.68	0.68
	2w2eA	8	170	0.58	0.58	0.63	215	0.31	0.52	0.57
	4od5A	9	226	0.48	0.48	0.61	231	0.48	0.48	0.56
	1okcA	6	198	0.52	0.58	0.63	208	0.36	0.41	0.58
	4n7wA	10	259	0.40	0.41	0.48	250	0.24	0.33	0.43
	2qi9A	10	246	0.56	0.56	0.56	275	0.41	0.48	0.56
	4ezcA	12	292	0.29	0.29	0.38	271	0.22	0.23	0.39
	1zcdA	12	303	0.31	0.38	0.54	331	0.40	0.40	0.45
	4bwzA	13	344	0.27	0.38	0.43	319	0.33	0.37	0.48
	3cx5C	8	283	0.40	0.45	0.56	322	0.43	0.43	0.43
Test set	3rlbA	6	133	0.69	0.70	0.73	102	0.69	0.76	0.80
	1e12A	7	181	0.40	0.40	0.50	176	0.37	0.37	0.48
	2vpzC	8	186	0.32	0.44	0.53	209	0.29	0.33	0.46
	2dyrC	7	203	0.42	0.42	0.58	197	0.45	0.52	0.62
	3m73A	10	249	0.44	0.44	0.54	255	0.28	0.31	0.49
	1u7gA	11	321	0.38	0.40	0.45	309	0.31	0.32	0.46

Table 4.6: ScafFold results from 10,000 decoys using ranking by LM1. Targets are listed in order of the total length, and TM-scores are calculated over the whole length, including the segment. "Length to predict" is the number of residues not in the segment that were modelled by SAINT2-ScafFold. The TM-score of the decoy ranked highest by LM1 is labelled "top1". The highest TM-score of the top five ranked decoys by LM1 is labelled "top5". The highest TM-score of all decoys generated is labelled "best". TM-scores ≥ 0.5 are shaded grey.

4.2.3 Sampling efficiency

When generating the decoys using a membrane potential in Section 4.2.2.2, I also sought to understand the efficiency of sampling, by observing which proposed moves were accepted, and at what stage of the sampling process. With a large number of trajectories to analyse, a full picture of the acceptance of moves could be built up to show how the sequential algorithm works. For every proposed move, the residue positions included in the proposed fragment were recorded, in addition to the move number in the trajectory, and whether the move was then accepted. For every combination of residue position and move number, the move acceptance ratio was separately calculated, equal to the number of accepted moves at that position divided by the total number of moves proposed at that position.

Figure 4.13 visualises the move acceptance ratio throughout decoy generation for 1kqfC, a typical target. As these decoys were generated by SAINT2-ScafFold in the Forward direction, there are no moves proposed in the first 77 residues, as this is the fixed segment section of the protein. To the left of the plot, all predicted contacts are displayed, in addition to the DSSP secondary structure annotation (Kabsch and Sander, 1983) as calculated by JOY (Mizuguchi *et al.*, 1998). Each step up along the curve is where an extrusion occurs between one move and the next. Extrusions are always accepted and therefore are not included in the calculation of acceptance ratios. The number of moves between each extrusion increases as the trajectory continues, according to the allocation described in Section 4.1.1.1, as there is a greater number of possible positions to make moves in a longer peptide. The additional moves after all extrusions have taken place are the full length moves.

The first trend which can be observed in this and other acceptance plots is the tendency for moves to be accepted more at the growing end of the peptide, shown by the darker colour. Moves outside of the $\sim$ 20 C-terminal residues at a given time are less likely to be accepted. At the growing terminus, any proposed move is less likely to impact the global structure, or favourable interactions and

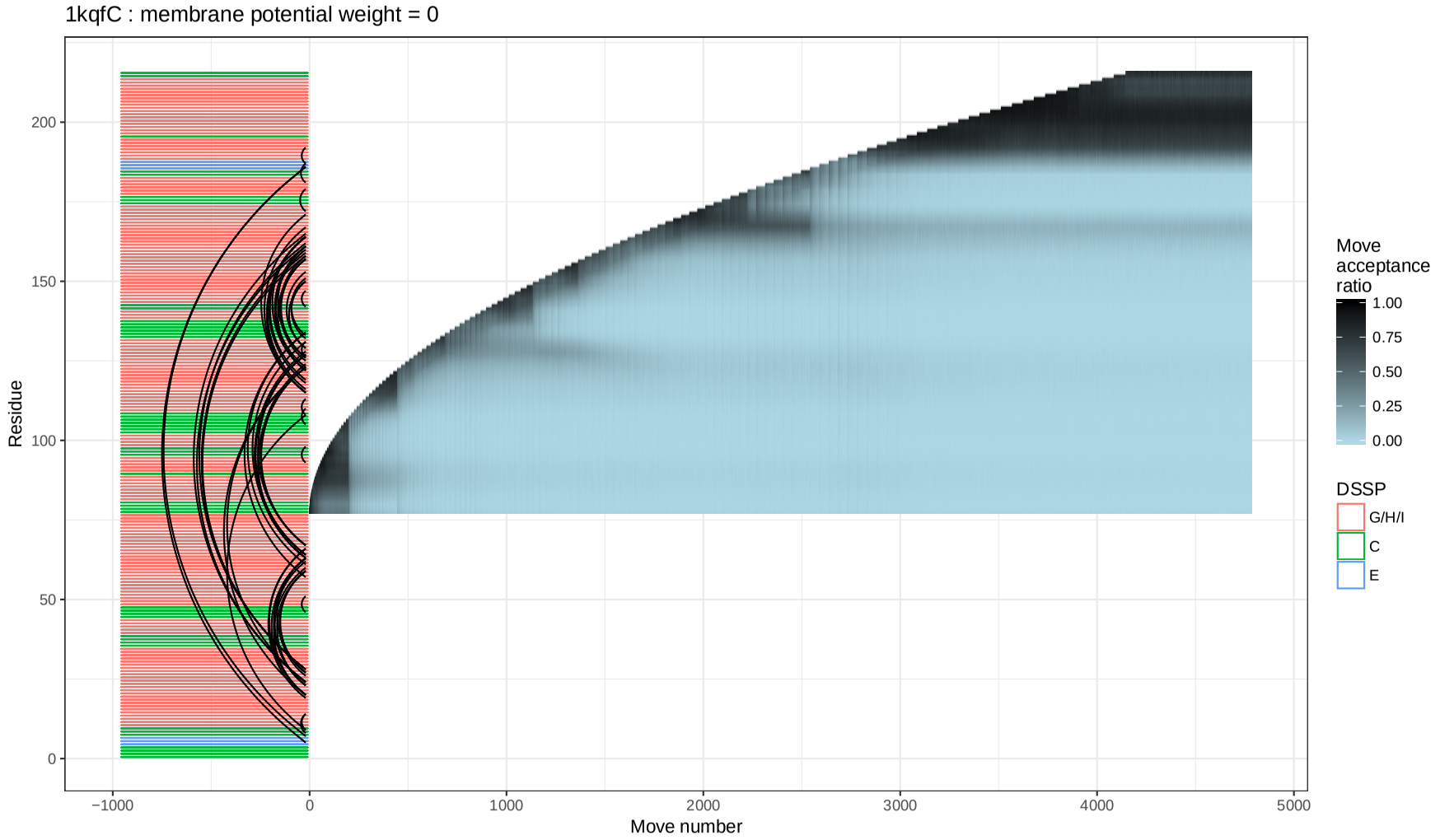


Figure 4.13: Acceptance ratio during Forward decoy generation for 1kqfC. Black arcs on the left represent predicted contacts between residues.

constraints which are holding it in place. This overall trend leads to an effective but greedy exploration of space, where a reasonable conformation is likely to be kept once the peptide has grown past that point.

Moves are also more likely to be accepted in helical regions, which is consistent with previous work based on the low resolution phase of the Rosetta AbinitioRelax protocol and another fragment-based predictor, EdaFoldAA (Kandathil *et al.*, 2016). Moves can be tolerated in helical regions more easily, even in the later stages of decoy generation, as the fragments are far more alike than the fragments in loop regions.

A peculiarity of the sequential methods of SAINT2 is that the extrusion steps themselves may have the effect of momentarily increasing the temperature by forcing moves, which could be unfavourable, to happen. This mainly affects the terminal residues, where there is already a high acceptance ratio, therefore it is unlikely to be a significant advantage for the protocol over the In vitro method. However, particularly after some extrusions in loop residues, the acceptance ratio is elevated throughout the length of the protein in the moves following an extrusion step. The further from the terminus, the quicker the acceptance ratio returns to the baseline low level after an extrusion.

The predicted contacts also impact the acceptance ratio in specific locations. In this example, there are no predicted contacts to tether the nascent peptide to the segment until the chain reaches around 110 residues long (there is one short range contact within this region). Before that point, in the first 200 moves, the acceptance ratio is high, at around 0.5. When the peptide becomes long enough to include the residues for this contact, moves in this region are suppressed, which may be due to the majority of proposed moves failing to satisfy the contact. In this way it is likely to be difficult to move away from the first conformation which satisfies the contact. A similar effect is observed prior to the contacts introduced at roughly 450, 1150 and 2600 moves. Each of these points corresponds to the end of a stretch of residues extruded with no new medium- to long-range contacts, by introducing consideration of a new contact. The

contacts appear to have a huge impact on the folding pathway in the SAINT2 protocol, and the local character of the search may make it possible to satisfy each one in turn in a more systematic way than the In vitro approach.

It is also possible to compare the acceptance plots for a given target using different weights for the membrane potential. Introduction of the membrane potential would be expected to increase the average score difference between a proposed move and the previous state. As the different weight set-ups were run at the same temperature, i.e. a score increase of the same magnitude in either set-up has the same probability of acceptance, an increased membrane potential weight may lead to a lower acceptance ratio. Comparing Figures 4.14, 4.15, and 4.16 to Figure 4.13, it is clear that acceptance ratios become lower as the weight of the membrane potential increases. Figure 4.17 shows the difference in acceptance ratios from subtracting the map for membrane potential 0 from the map for membrane potential 0.5. The largest differences are where ratios were high when there was no membrane potential, where few contacts were present. The membrane potential appears to have a similar, but weaker, restrictive effect to the contact potential. This indicates that in order to achieve the best results with the membrane potential, or any other new score, it might be best to simultaneously test a range of temperatures as well as score weights.

Figure 4.18 shows the average acceptance ratio at each residue averaged over the course of the whole protein for two different targets. Residues further from the growth terminus have much lower average acceptance ratios as they are static for more of the decoy generation. In the case of the Reverse target, these are the higher residue numbers as the chain grows towards the left of the graph. It can still be seen that alpha-helical residues have more accepted moves than their neighbouring loop residues. This insight into the move acceptance during sequential runs of SAINT2 allows us to investigate how the unique sampling strategy affects the efficiency of the protocol.

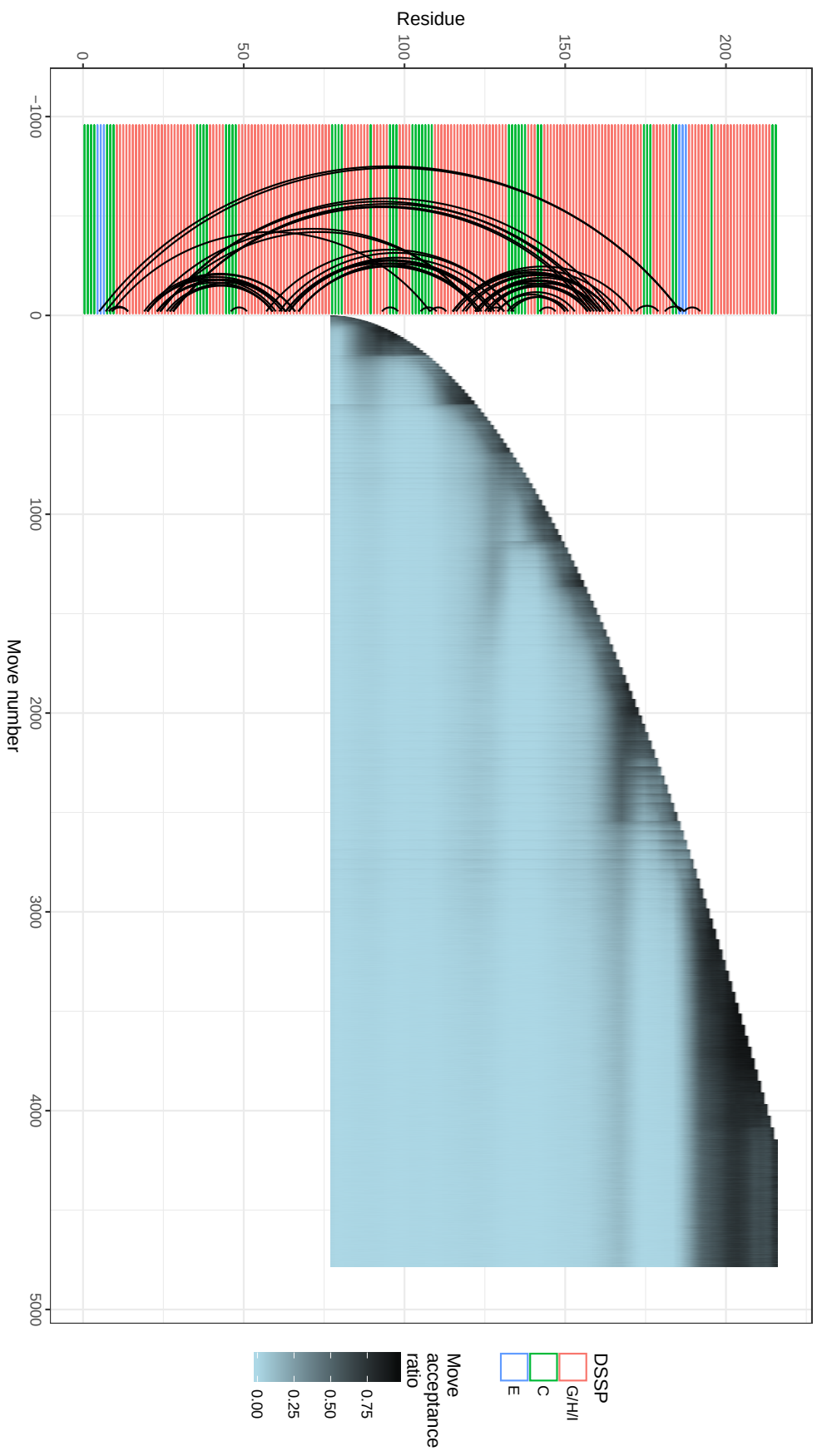


Figure 4.14: Acceptance ratio during Forward decoy generation for 1kqfC.

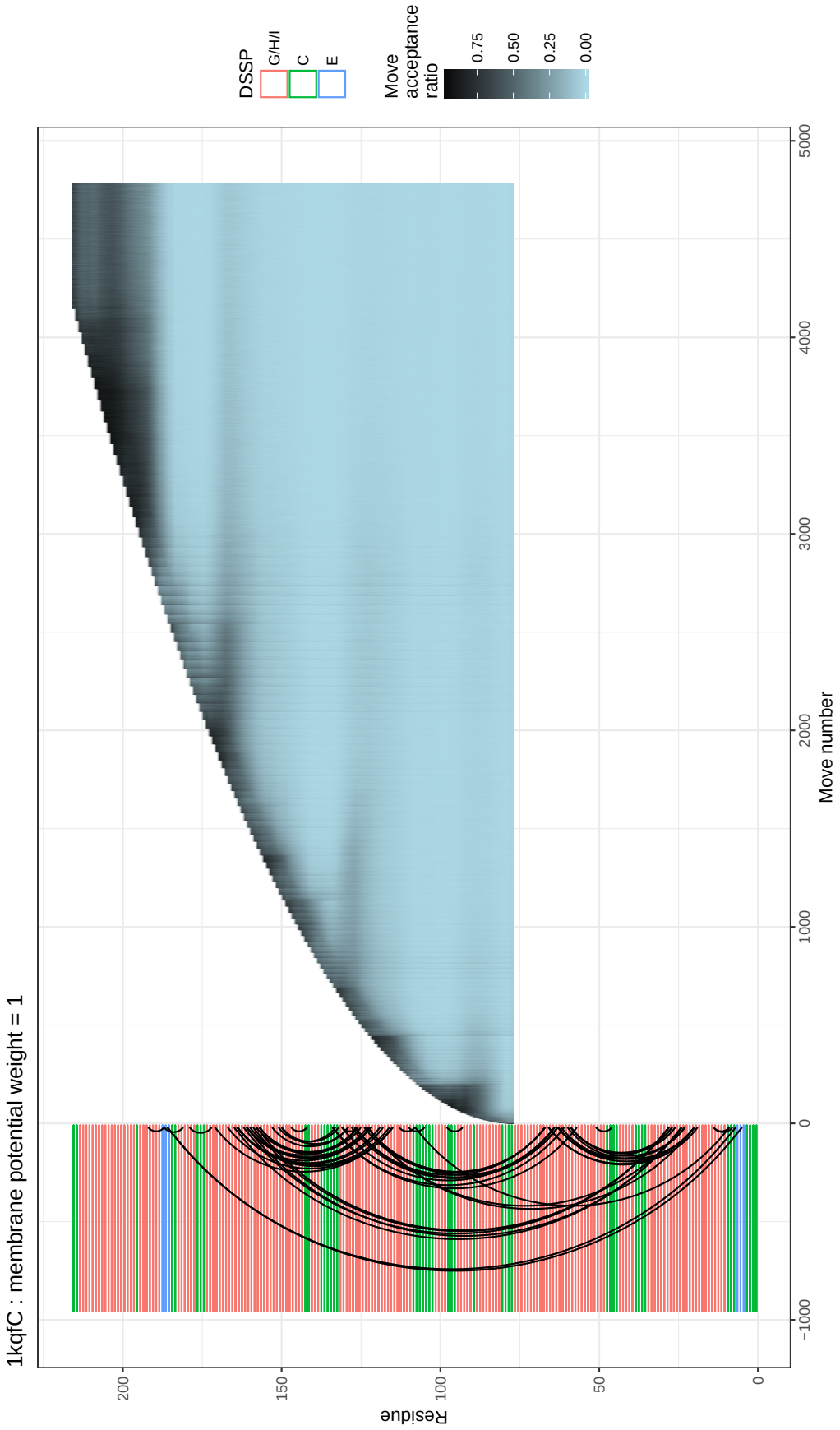


Figure 4.15: Acceptance ratio during Forward decoy generation for 1kqfC.

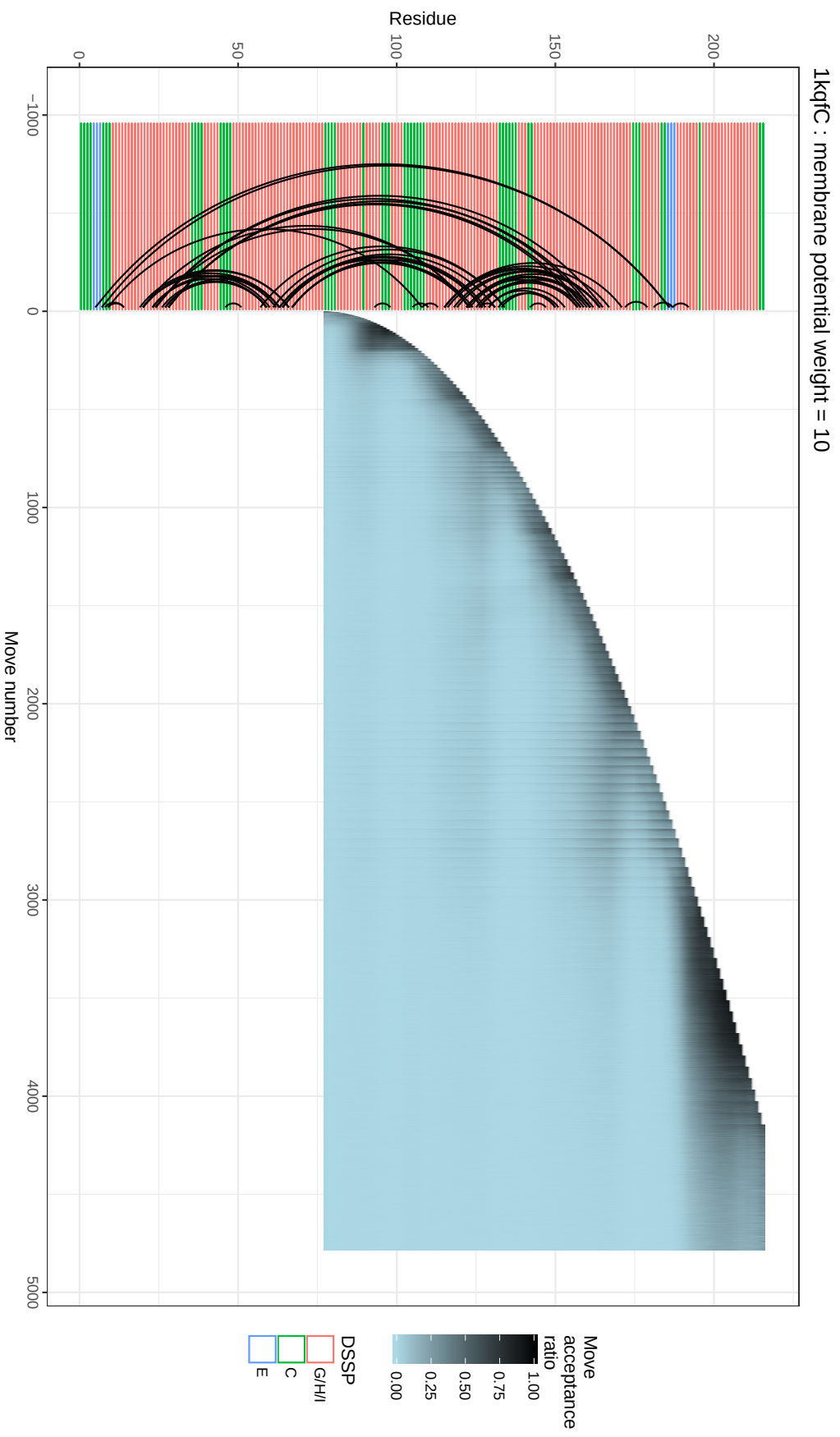


Figure 4.16: Acceptance ratio during Forward decoy generation for 1kqfC.

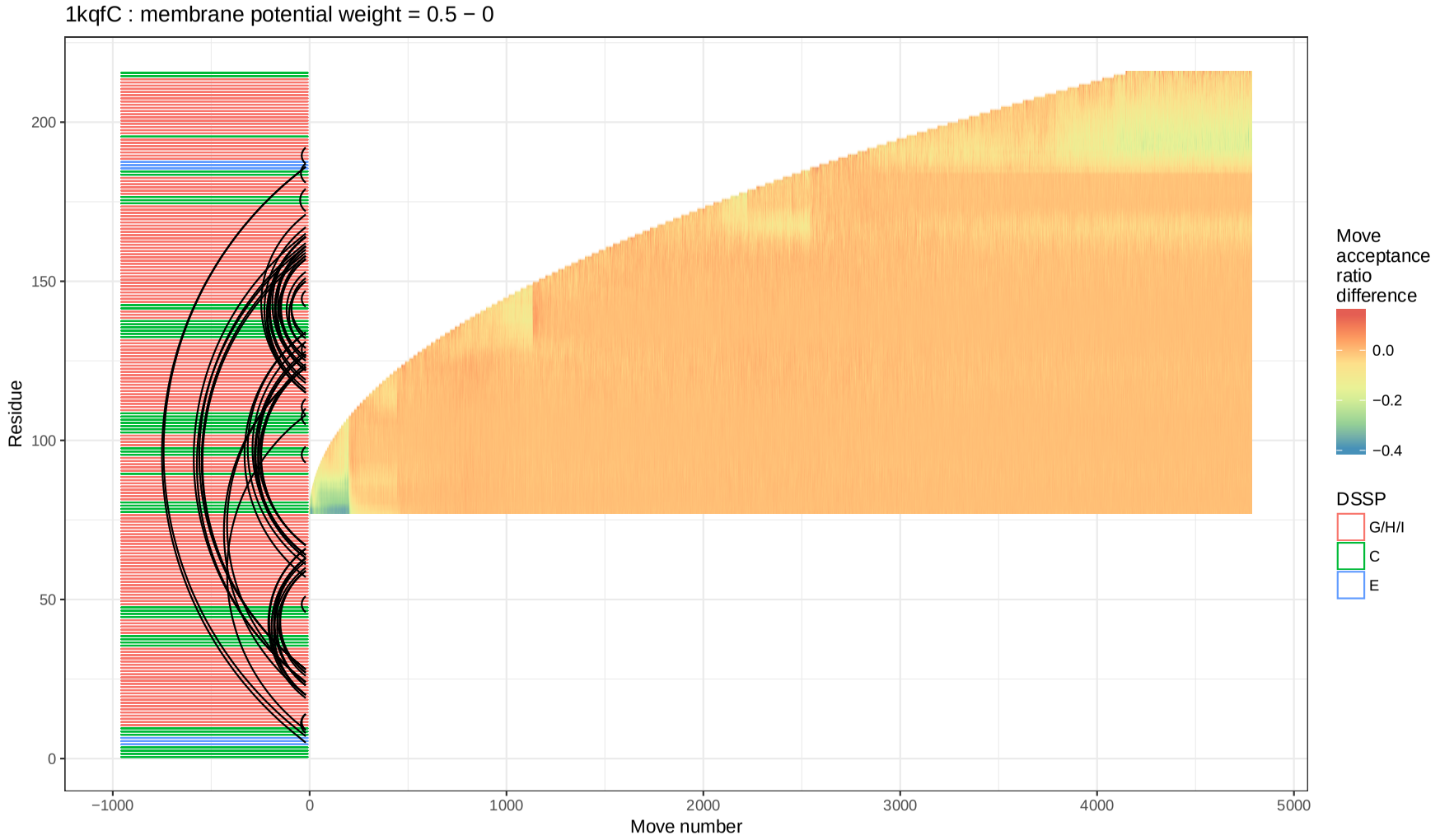


Figure 4.17: Acceptance ratio difference map of Forward decoy generation for 1kqfC to compare the acceptance ratios with (weight = 0.5) and without a membrane potential.

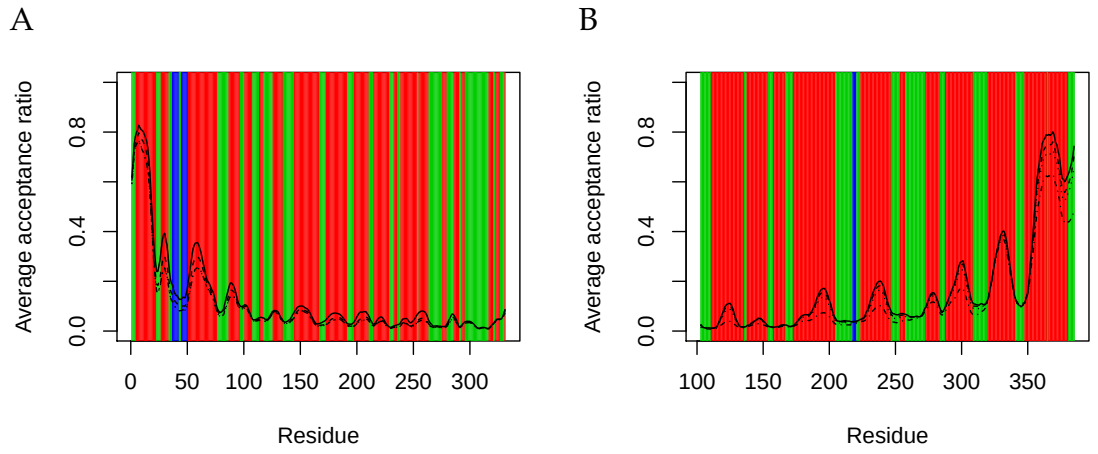


Figure 4.18: Acceptance ratios at each residue averaged over all moves. DSSP annotation uses the same colour scheme as Figure 4.17: G/H/I: red; C: green; E: blue. The solid line is for a membrane potential weight of zero, and dashed lines are for the three non-zero weights. A) 1zcdA, Reverse mode. B) 3cx5C, Forward mode.

4.3 Conclusions

In this chapter, I have described how adaptations were made to improve the performance of SAINT2 on membrane proteins, and to improve the ability of the algorithm to imitate biological folding. In order to simplify prediction of long protein targets, the SAINT2-ScafFold method of building on a segment was developed. This version of SAINT2 represents an attempt to imitate a possible biological scenario in which part of the protein has already been inserted and folded while the rest of the protein is synthesised. The SAINT2-ScafFold protocol did not improve the prediction of the second half of proteins when compared to the full sampling method, except in a small number of cases. The likely causes of this are related to the rigidity of the segment, which *in vivo* would be somewhat flexible to allow for the correct positioning of later helices. To give flexibility, moves could also be allowed in the segment, perhaps at a lower rate.

I then used the SAINT2-ScafFold protocol to test the implementation of a membrane potential, as the starting segment provided a reference point for the membrane position. Scoring of previous sets of decoys was used for three purposes: to confirm that a membrane potential could be useful; to estimate appropriate weights for use in decoy generation; and to assess the merit of

removing the solvation potential. Decoy generation was found to be slightly better for most targets when using a membrane potential weight smaller than that for the contact potential, but greater than the RAPDF or Lennard-Jones weights. Using these weights, correct decoys were generated for 5/6 test set targets, building in the Forward mode from an N-terminal segment, and 2/6 targets building in Reverse. Due to the small size of the test set, there was no indication that the Forward and Reverse modes perform differently (Chi-squared test p-value = 0.24).

The linear model trained on decoys generated for targets in the training set was able to select correct decoys when the best decoy in the set had a TM-score > 0.62. The simple linear model used here is probably not well suited to this ranking task, as we do not expect potentials to be linearly correlated with TM-score. It would also be preferable not to normalise every potential for every population of decoys. As decoys were only scored on a relative and not absolute scale, it was not possible to convert the ranking score of a decoys to a probability of the decoy being correct. Other machine learning methods may achieve more success in ranking decoys, and using clustering methods and ranking by the popularity of clusters may also be effective. In order to find out the best weights to use during decoy generation, it may be better to train a scoring function on a set of perturbed native structures instead of a pool of decoys.

It is clear that many more membrane-specific adaptations could be made to improve prediction. There are membrane-specific methods for inferring predicted contacts from coevolution in multiple sequence alignments (Xiao and Shen, 2015; Zhang *et al.*, 2016). The quality of the fragment libraries could be improved by using membrane-specific secondary structure prediction, such as the consensus method TOPCONS (Tsirigos *et al.*, 2015). In order to best imitate the folding pathway, a membrane solvation potential that reflects the propensity of each residue to be exposed at a particular depth could also be included.

While making small changes to the scoring function of SAINT2, I was also able to observe the impact of this change to the acceptance ratios, at different

residues in the chain, and different move steps within the overall trajectory. The acceptance ratio maps go one step further than previous studies of move acceptance, by showing how the trajectory progresses when using this unique structure prediction algorithm.

While SAINT2 is an exciting prospect for membrane proteins, as there is some evidence that they are adopting tertiary folds cotranslationally, SAINT2 is not yet reaching the accuracy that is achieved by the leading structure prediction methods. The best prediction methods (e.g. Hopf *et al.*, 2012; Ovchinnikov *et al.*, 2017; Teixeira *et al.*, 2017) rely heavily on the best predicted contact information, and use few other scores or constraints to arrive at the correct fold. SAINT2 also relies on contact information to generate good models, but the nature of the contacts, and the ScafFold method implemented in this chapter, take the algorithm further away from an imitator of biological folding. It seems that for membrane protein structure prediction, this attempt to imitate folding is not currently competitive with the leading methods.

In the final results chapter, I will describe the application of the SAINT2-ScafFold method to generate complete homology models for targets for which no complete homologous structure has been solved, in both membrane and soluble proteins.

5 Completion of partial homology models

In this chapter, I use the SAINT2-ScafFold approach described in Chapter 4 to complete partial models of proteins.

5.1 Background

Homology modelling is the most accurate means of generating coordinates for a protein of unknown structure (Forrest *et al.*, 2006). Structural genomics projects have pursued the aim of ensuring as many protein families as possible have at least one solved structure so that the less accurate *de novo* modelling is not required (Montelione, 2012). The growth in available structural data means that 32% of Pfam families have a member with known structure and could be modelled by homology (Ovchinnikov *et al.*, 2017). A further 27% (total 59%) can be modelled from a reliable template found by HHsearch (Soding, 2005).

In the case of membrane proteins, the deposition of structures has proceeded at a much slower rate than that for soluble proteins (Koehler Leman *et al.*, 2015). Yet there are still many families for which at least one structure is available, and membrane specific template modelling approaches have been developed in order to make the most effective use of them (e.g. Kelm *et al.*, 2010; Ebejer *et al.*, 2013;

Werner and Church, 2013; Chen *et al.*, 2014). In some cases, especially where the best available structural homologue is remote, the coverage of the target may be poor. Significant changes in loop regions are common and allow for the diversity of function within protein family, for example in GPCRs. However, in some cases transmembrane helix segments are missing, and there is currently no tool designed to complete such models. An example of a purpose-built membrane protein prediction homology modelling tool is Medeller (Kelm *et al.*, 2010). In the Medeller protocol, a high accuracy core model is constructed from the transmembrane segments. If a complete but lower accuracy structure is required, FREAD (Choi and Deane, 2010) is used to model loops. Medeller, like all other homology tools, is therefore not designed to effectively predict an extra terminal transmembrane segment.

De novo predictors may perform better on targets for which structural homologues are available, as the presence of homologues in the fragment library improves the accuracy of fragments and increases the accuracy of predictions (de Oliveira *et al.*, 2015). However, these methods also depend on contact predictions to produce good models (see Section 1.4.4.1). If a transmembrane helix is not present in all members of the family, there is less information available to inform contact predictions. Therefore it is likely that fewer contact predictions will be made either within such a helix or between it and the rest of the structure.

In this chapter, I first assessed the frequency of the problem of incomplete homology models in membrane protein modelling. I found that around 57% of human membrane protein prediction scenarios that could be attempted by template-based modelling would rely on an incomplete template, based on a recent database (Pieper *et al.*, 2013). I then used the ability to build on a rigid segment of a native structure using SAINT2-ScafFold to provide models for these extra terminal helices, without disrupting the segment. Using this strategy, an accuracy of < 5 Å RMSD was achieved for 29/35 of single helix predictions. For a set of longer targets, I also built homology models, in order to test the

method in a more realistic setting. The quality of the models generated from homology segments was comparable to those generated from native segments.

5.2 Methods

5.2.1 Prevalence of incomplete homology models for human membrane protein targets

Data on the template coverage for membrane proteins in the human genome was gathered from the Survey of the Human Transmembrane Proteome (Pieper *et al.*, 2013). The database was constructed using TMHMM (Krogh *et al.*, 2001) on the human genome to find all transmembrane proteins, and Modbase (Pieper *et al.*, 2011) to catalogue the possible templates for these proteins. Proteins predicted to have at least eight transmembrane helices were extracted, as the problem of structure completion is likely to be less common in shorter proteins. The locations of predicted transmembrane helices were extracted from the ModBase sketches for each Gene ID. The locations were used to establish which of the linked templates covered the greatest number of predicted transmembrane helices, requiring at least half of each helix to be covered. Where two templates covered the same number of helices, the template with higher sequence identity to the target was used. Templates covering a small number of helices were often observed not to be membrane proteins, and also had high recorded E-values from PSI-BLAST (Altschul *et al.*, 1997), therefore only templates with at least six helices were considered.

5.2.2 Native Structures completed by SAINT2-ScafFold

To compile a test set of cases to simulate completion of an incomplete homology model, the data set of membrane proteins used in the previous chapter was used as a starting point. Each structure was individually inspected to determine whether the N-terminal helix, or two helices, were buried, or whether the rest of the transmembrane bundle could conceivably fold and exist as a bundle without them. The same assessment was made at the C-terminus, for the final

two helices. Table 5.1 shows for each terminus which targets were used to test completion of one or two helices at the N- or C-terminus.

The targets were predicted using SAINT2-ScafFold in the Forward direction for incomplete C-termini and the Reverse direction for incomplete N-termini. The segment length used for the 'rigid' version was up to the last residue of the transmembrane span prior to the first helix to be predicted, as annotated by RosettaMP's `mp_span_from_pdb` tool (Alford *et al.*, 2015). An alternative 'flexible' set up was also run, in which the segment ended at the first residue of the transmembrane span prior to the first helix predicted. The same fragment library and predicted contacts were used as in those in Chapters 3 and 4.

The calculation of the number of moves to perform was similar to that used in Section 4.1.1.1. In that case, comparisons were to be made against SAINT2-Wholly, and therefore the length adjustment was relative to the full length of the protein. In this chapter, I aimed to use the most suitable number of moves for the remaining length, regardless of the total length, as a fair comparison to SAINT2-Wholly was not required. SAINT2 has produced correct answers using 10,000 growth moves and 1,000 full length moves on proteins up to $\sim$ 150 residues, using SAINT2-Wholly in Forward or Reverse mode. Therefore, the number of moves for the remaining helices was scaled relative to this as an appropriate benchmark figure. The following calculation was used to determine an appropriate number of moves for the length remaining to be sampled:

$$\text{growth moves} = \frac{M_g \left(L - S\right)^2}{R^2} \tag{5.1}$$

M_g is the standard number of growth moves (10,000) for a reference protein length of R residues, here taken to be 150, L is the length of the full peptide, and S is the length of the segment. Similarly, the number of full length moves was calculated in the following way:

$$\text{full length moves} = \frac{M_f \left(L - S\right)}{R} \tag{5.2}$$

where M_f is the standard number of full length moves (1,000).

Target	Length	TMHs	N1	N2	C1	C2
1orsC	132	4	✓	✓	✓	✓
3rlbA	176	6	✓		✓	✓
2xowA	179	6	✓	✓	✓	✓
4a2nB	192	5	✓	✓	✓	
4o6yA	210	6	✓		✓	✓
1kqfC	216	4	✓	✓	✓	✓
3b4rB	216	7	✓	✓	✓	✓
4b4aA	225	6	✓	✓	✓	✓
1e12A	239	7	✓	✓		
2vpzC	250	8			✓	✓
3klyA	257	7	✓	✓	✓	
2dyrC	259	8	✓	✓		
2w2eA	263	8	✓	✓	✓	
4od5A	274	9			✓	✓
1okcA	292	6		✓	✓	✓
4n7wA	307	10	✓			
3m73A	313	10			✓	
2qi9A	324	10	✓		✓	
4ezcA	345	12				
1zcdA	376	12	✓			
1u7gA	383	11	✓	✓	✓	
4bwzA	384	13	✓	✓		
3cx5C	385	8			✓	✓
4kppA	395	12	✓		✓	✓
3qe7A	407	14			✓	✓
3o7qA	414	12			✓	
1pv6A	417	12			✓	
4ky0C	422	11	✓			
3qnqA	432	10			✓	✓
3giaA	433	13			✓	✓
1otsA	444	15	✓	✓		
2jlnA	463	12			✓	✓
4gc0A	475	13			✓	
4ikvA	492	14			✓	
2wswA	508	13	✓	✓		
4m48A	532	13			✓	✓

Table 5.1: Dataset for testing the completion of the final one or two helices of a protein. N1/N2: N-terminal one/two helices; C1/C2: C-terminal one/two helices. Ticks indicate that these helices were used for prediction as they are not buried in the structure.

The accuracy of prediction was assessed by calculation of the RMSD to the native structure. The RMSD was calculated for the backbone atoms in the transmembrane spans of the final helix or helices which were not part of the ScafFold segment. In SAINT2-ScafFold, the segment section of the protein is not rotated or translated therefore no alignment was performed after decoy generation before calculation of RMSD.

5.2.2.1 Decoy ranking using linear models LM1 and LM2

In order to rank structures to select the most accurate decoy, I used the linear model from Section 4.2.2.3. The scores for each population of decoys were normalised as before, and the predictive model (LM1) included the RAPDF, Lennard-Jones, contact and membrane potentials. An alternative model (LM2) from Section 4.2.2.3, which did not include the contact potential, was used in this chapter to simulate ranking decoys in the absence of predicted contacts between the final transmembrane span and the segment helix bundle.

5.2.3 Homology models completed by SAINT2-ScafFold

The longest targets were also predicted by SAINT2-ScafFold building from an incomplete homology model rather than the crystal structure. PSI-BLAST (Altschul *et al.*, 1997) was used to identify possible templates in the PDB, searching using the default parameters for multiple iterations until no further structures were included in the next iteration. A template was selected with $< 40\%$ sequence identity to the target, as the results in Section 5.3.1 indicated that templates are likely to cover all transmembrane helices if they are more closely related. For one target, 2jlnA, there were no distant homologues, therefore a template of 96% sequence identity but in a different conformation was used to generate a homology model for the segment. All templates used for homology modelling are shown in Table 5.2. For 4m48A, the template 4us3A was chosen as it is actually one helix shorter than its target and therefore the perfect test for building from an incomplete model.

Target							Template							
PDB	Length	TMHs	N1	N2	C1	C2	PDB	Length	TMHs	SID	Sequences	RMSD	1 TMH	2 TMHs
4ky0C	422	11	✓				5lluA	396	11	0.36	84	1.62	28	
3giaA	433	13			✓	✓	3ncyA	422	13	0.19	125	3.61	38	75
1otsA	444	15	✓	✓			5tqqA	603	15	0.21	124	1.70	62	92
2jlnA	463	12			✓	✓	2x79A	465	12	1.00	125	2.68	46	96
4gc0A	475	13			✓		4ldsA	421	13	0.33	68	2.35	50	
4ikvA	492	14			✓		4q65A	437	14	0.30	85	4.36	38	
2wswA	508	13	✓	✓			2witA	531	13	0.27	125	2.23	47	96
4m48A	532	13			✓	✓	4us3A	441	12	0.26	125	2.77	59	107

Table 5.2: Dataset for testing the completion of the final one or two helices of a protein.
TMHs: the number of transmembrane helix spans.
N1/N2: N-terminal one/two helices; C1/C2: C-terminal one/two helices. Ticks indicate that these helices were used for prediction as they are not buried in the structure.
SID: the sequence identity between the template and target in the MP-T alignment.
Sequences: the number of sequences in the MP-T alignment.
RMSD: the root mean squared deviation in Å between the coordinates of the backbone atoms of the core segment helices (i.e. not the 3 terminal helices which are to be predicted or flexible).
1 TMH/2 TMHs: the number of residues to be predicted in the one/two helix prediction scenario.

A homology model was built for each target using the web server Memoir (Ebejer *et al.*, 2013). If no hits were found by iMembrane (Kelm *et al.*, 2009) for a target, it was removed. For every target, the multiple sequence alignment produced by MP-T satisfied the recommended minimum number of homologous sequences for a good model.

The complete model output by Medeller (Kelm *et al.*, 2010) at the final stage of the pipeline was used as the segment input to SAINT2-ScafFold. The transmembrane span annotations were taken from the embedding of the original native structure in Section 5.2.2. This ensured that the segment lengths were the same for the native structure and homology model tests. The only change tested was the effect of inaccuracies in previous helices arising from the homology model, and not varying lengths from differing span definitions. The homology model was embedded in the membrane by using PyMOL to align the model to the already embedded native structure (Schrödinger, 2015).

The accuracy of prediction was assessed using the same RMSD calculation as in Section 5.2.2. The structural alignment used for this calculation was the superposition that minimised the RMSD between the backbone atoms of the transmembrane spans in all but the final three helices. This ensured that the superposition used to evaluate accuracy was identical for the flexible and rigid runs, and for cases with either one or two helices missing.

5.3 Results and discussion

5.3.1 Prevalence of incomplete homology models for human membrane protein targets

In order to set up tests which were representative of the challenge posed by completion of partial homology models, I first assessed what is commonly missing from available templates. Using the procedure described in Section 5.2.1, data was gathered from a survey of the human membrane proteome (Pieper *et al.*, 2013) and analysed to find out the number of helices missing from possible template structures. Figure 5.1 shows the number of transmembrane

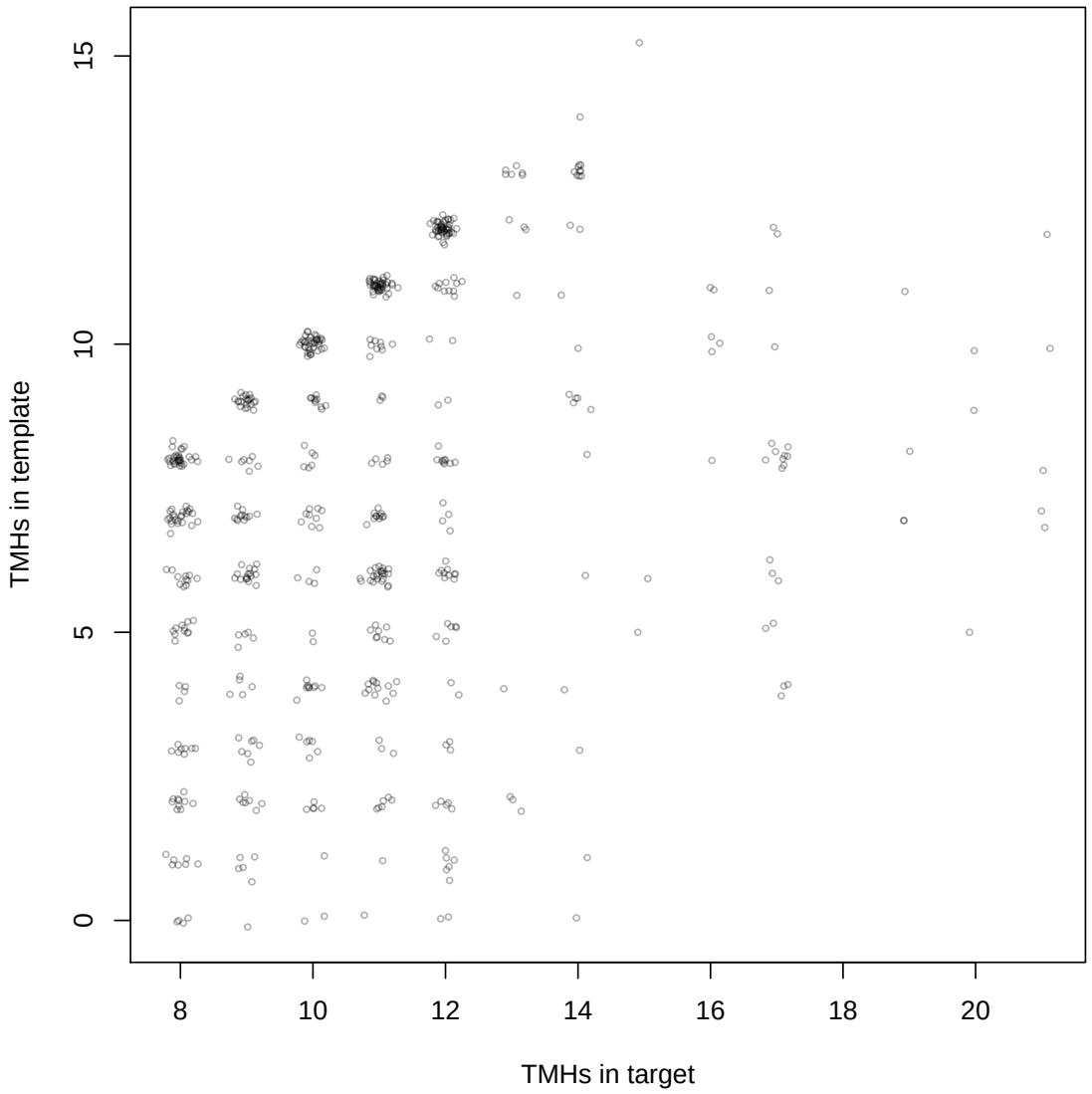


Figure 5.1: The number of transmembrane helices (TMHs) in the "best" available template against the number of TMHs in the target, for targets with at least eight TMHs.

helices in the template with the best coverage against the number of helices in the target. It was common for the template and target to have the same number of transmembrane helices; however, there were many cases where the 'best' templates were shorter than the target. The database catalogues many templates which were short and potentially unreliable, therefore I considered only templates of at least six helices and targets of at least eight helices. Of these targets, only 43% (190/440) had no helices missing from either terminus of the template.

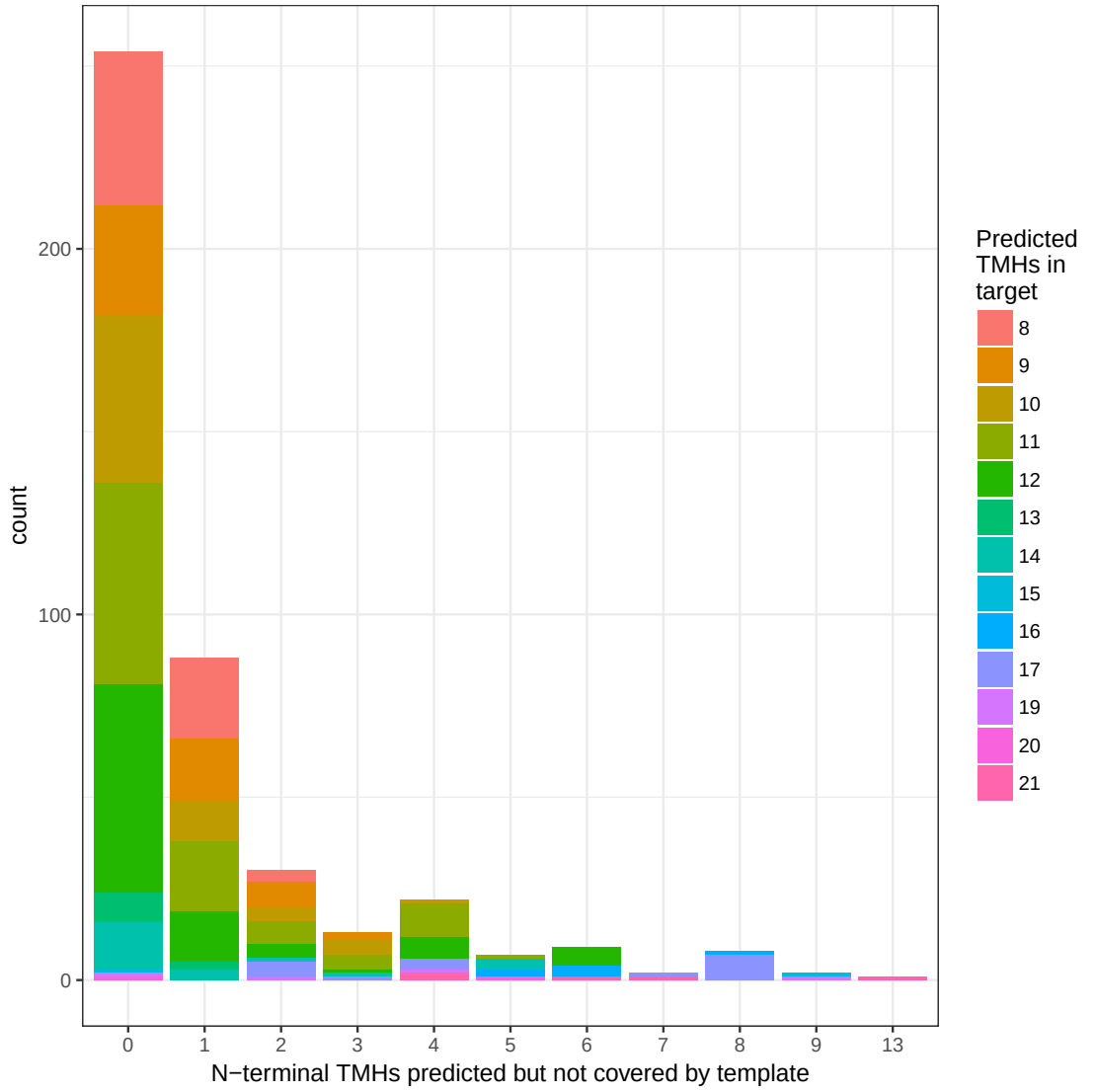


Figure 5.2: The number of N-terminal transmembrane helices (TMHs) not covered by the best available template of at least six TMHs.

Figures 5.2 and 5.3 show the number of helices missing from templates at each terminus. Over half, 64% (118/182) of incomplete templates are missing no more than two helices at the N-terminus, and 62% (94/152) are missing no more than two helices at the C-terminus. It is interesting that the numbers of templates that are incomplete at the N- and C-termini are similar, having observed the asymmetry which is present in membrane protein structures in Chapter 3. My results in Section 3.3.1 suggest that the N-terminal helix tends to be more centrally located in the helical bundle than the C-terminal helix. I would

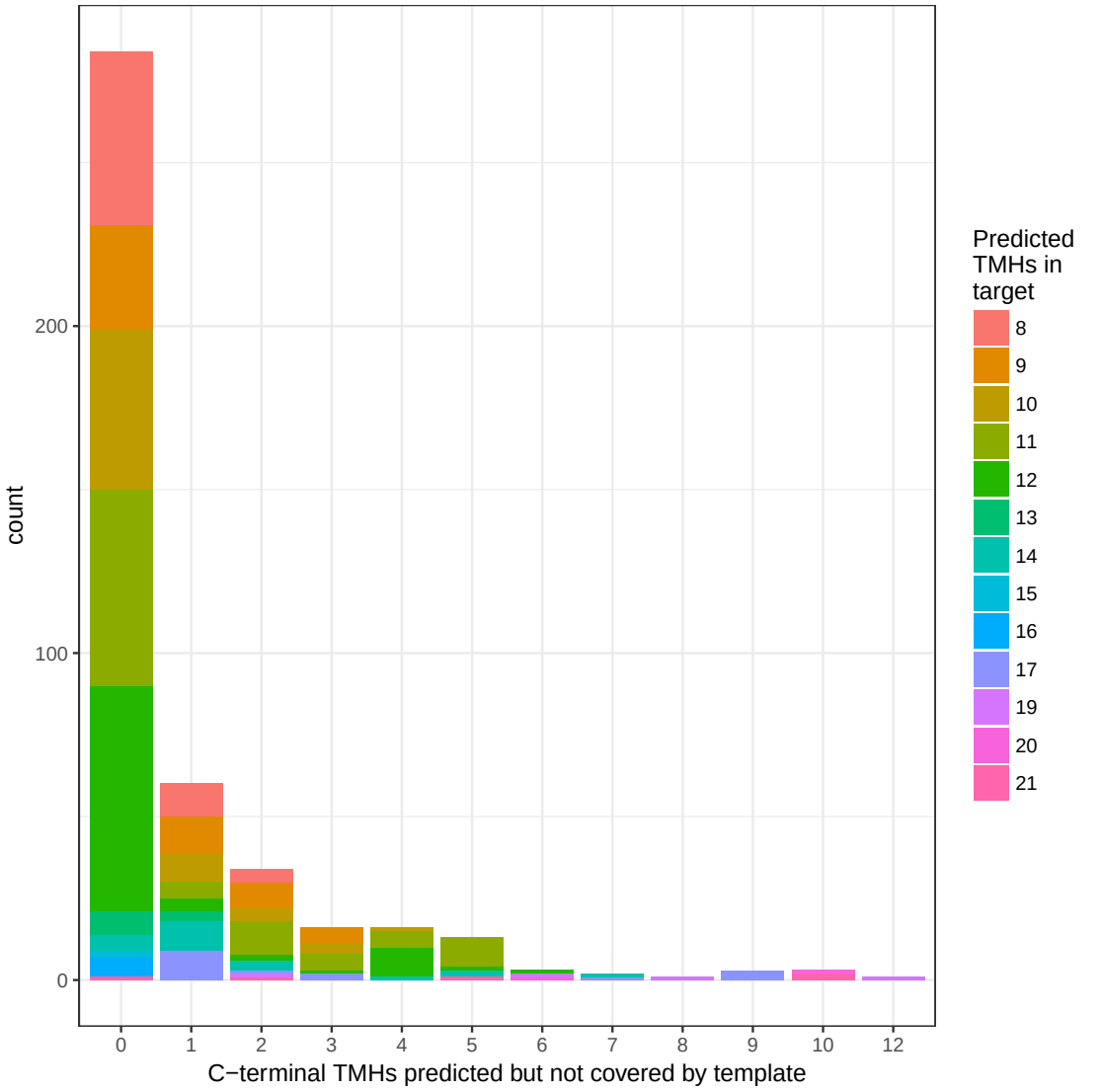


Figure 5.3: The number of C-terminal transmembrane helices (TMHs) not covered by the best available template of at least six TMHs.

expect that it is less likely for a helix to be added or removed at a terminus which is buried, and therefore I would expect to find more C-terminal helices missing from templates. However, the C-terminal helix is more frequently missing from a template than the N-terminal helix. There may be other factors affecting which terminal helices are not covered by templates, for example the truncation of proteins in order to generate a construct that can be crystallised. There may be other cases where the predicted N-terminal helix is actually a signal peptide that is cleaved, and therefore not present in the full length functional protein.

Figure 5.4 shows the expected relationship between sequence identity and the number of transmembrane helices missing from the template. Almost all templates with at least one helix missing have a sequence identity to the target of $< 40\%$, and those with two helices missing are mostly below 20%. From the coverage shown by these templates, despite using a simple and not recent transmembrane helix identification program, it is clear that only an incomplete template is available for many targets. In the majority of these cases, only one or two helices are not present in the best template. Therefore, when investigating the performance of SAINT2-ScafFold in such scenarios, I concentrated on completing just one or two terminal helices.

5.3.2 Native structures completed by SAINT2-ScafFold

To establish the accuracy of prediction that SAINT2-ScafFold could attain in completing the final helix of a structure, 4,000 decoys were generated in a variety of set ups for the targets listed in Table 5.1. The accuracy was measured by calculation of RMSD over the ScafFold modelled transmembrane helix residues with the scaffold segments aligned. To give an impression of the quality of a model at several different RMSDs, Figure 5.5 shows typical models for two set ups where the span has been scored as 3, 5 or 7 Å RMSD. An RMSD of $\sim 3-5$ Å would not provide reliable input for a molecular dynamics simulation, for example, but could potentially be used as a starting point for experimental work to confirm the proposed conformation.

When building from a partial model, or remodelling loops in a homology model, it is not always clear at which point to change the structure and where to keep it exactly the same. SAINT2-ScafFold can be started from any length segment, therefore I set up comparisons between a rigid version, with the segment ending at the end of a helix, and a flexible version, where the segment is shortened and ends at the start of the previous helix (see Figure 5.6 and Section 5.2.2). Allowing moves in the helix prior to the loop and helix to be predicted may help to allow more conformations if the fragment library is

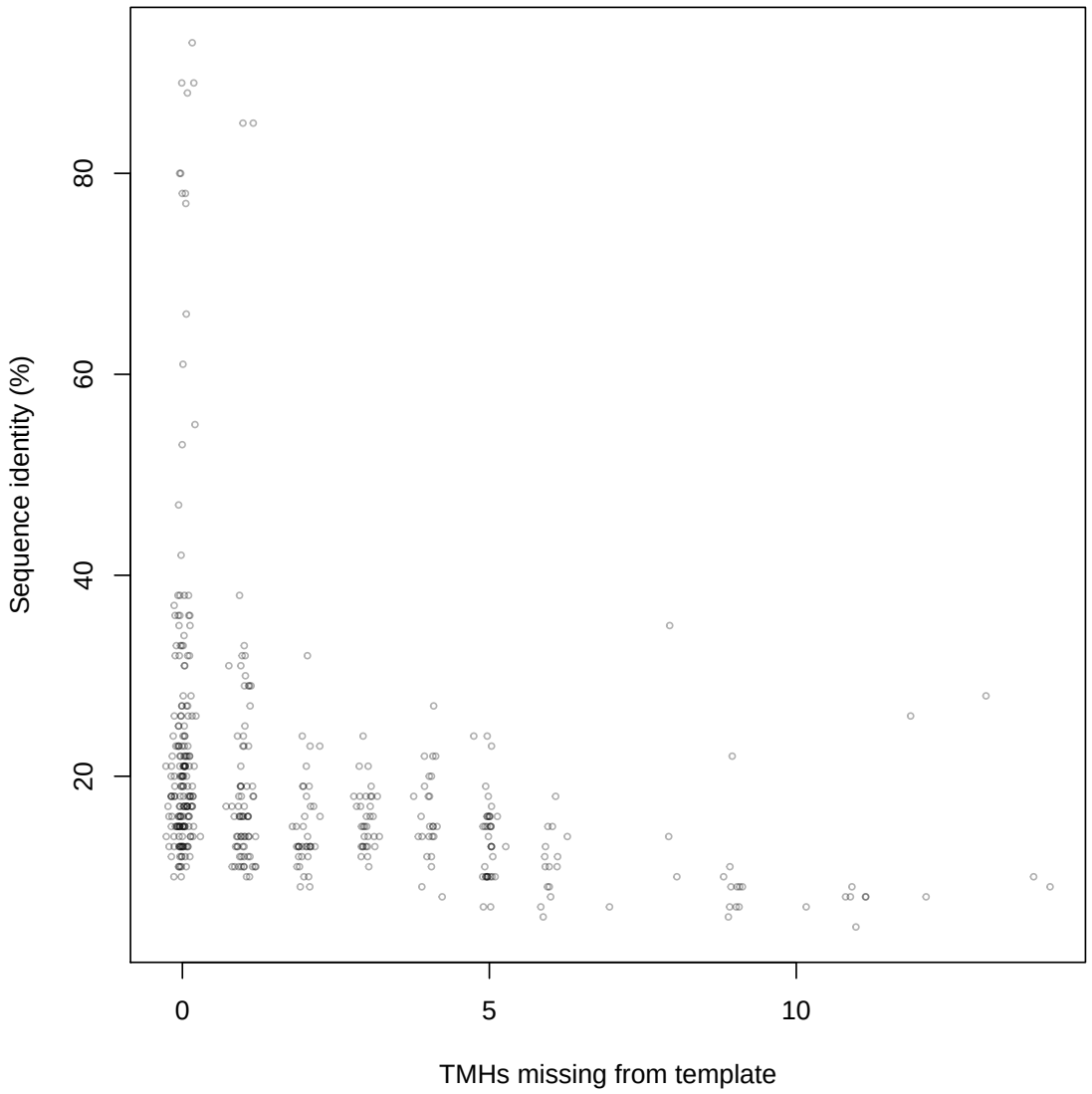


Figure 5.4: The relationship between transmembrane helices (TMHs) missing from the template structure and sequence identity between the template and target.

limiting exploration of the conformational space, providing a better "take-off point" for the remaining helix. Figure 5.7 shows the accuracy of prediction of an N- or C-terminal helix under the two different modes: 'rigid', sampling only the helix to be predicted, and 'flexible', which allows moves in the previous helix.

For a few targets, it was encouraging that the majority of decoys had an RMSD of $\sim 5 - 10$ Å, which is not far from the accuracy of a typical template (see Table 5.2). The rigid and flexible modes generated comparable distributions of RMSDs for most targets, with no clear preference for one over the other. For

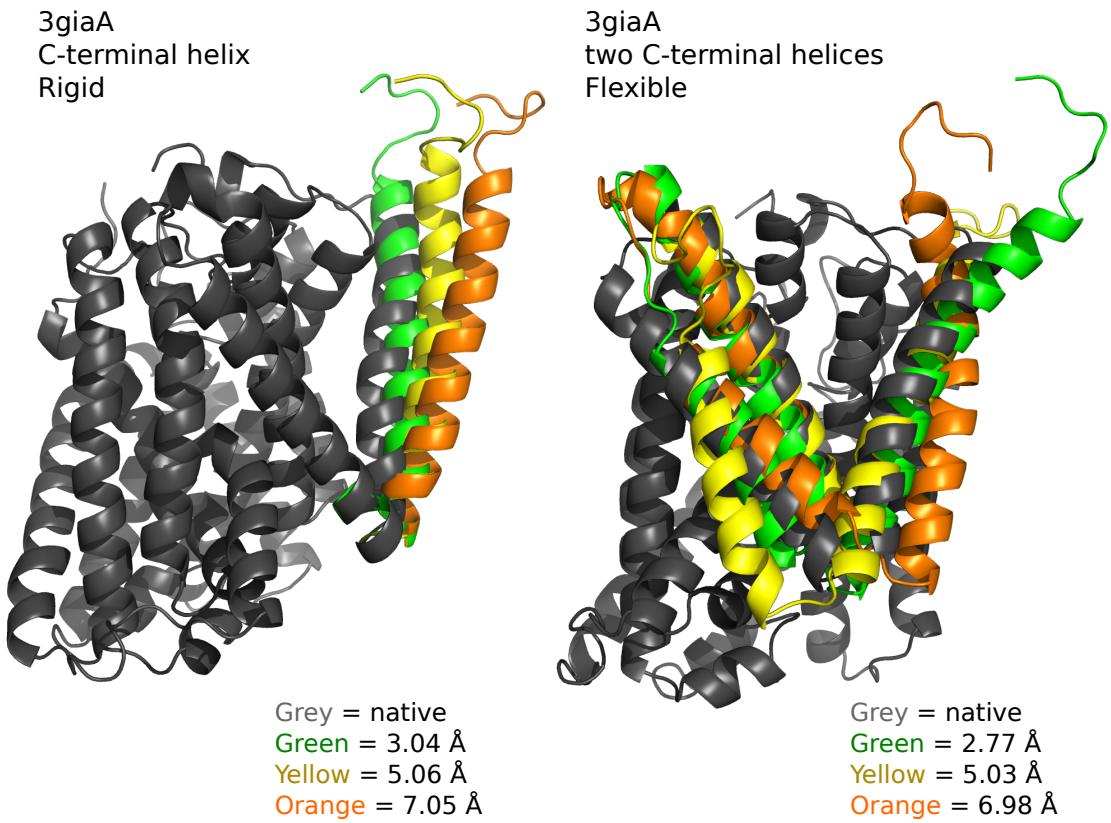


Figure 5.5: Examples of decoys of different RMSDs generated for the C-terminal helices of 3giaA. On the left, the decoys were generated for the last helix in rigid mode, and on the right, the last two helices were generated in flexible mode.

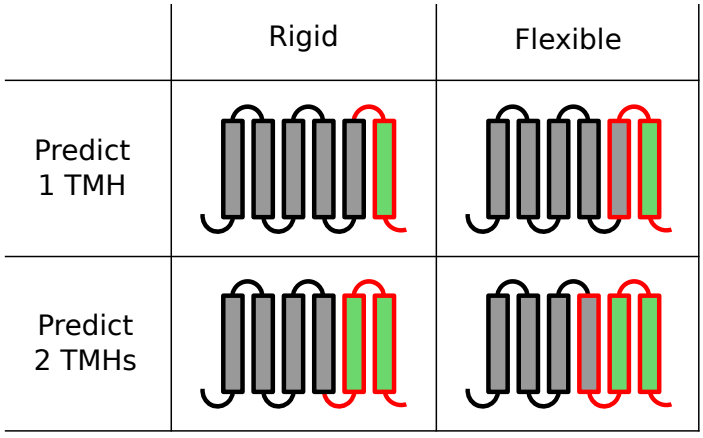


Figure 5.6: Rigid and flexible versions of SAINT2-ScafFold. Regions which are sampled in decoy generation are shown with a red outline. The region over which RMSD is calculated is shaded green. The loop before the transmembrane helices (TMHs) to be predicted is always sampled, and for the flexible version, the residues in one extra TMH are also sampled.

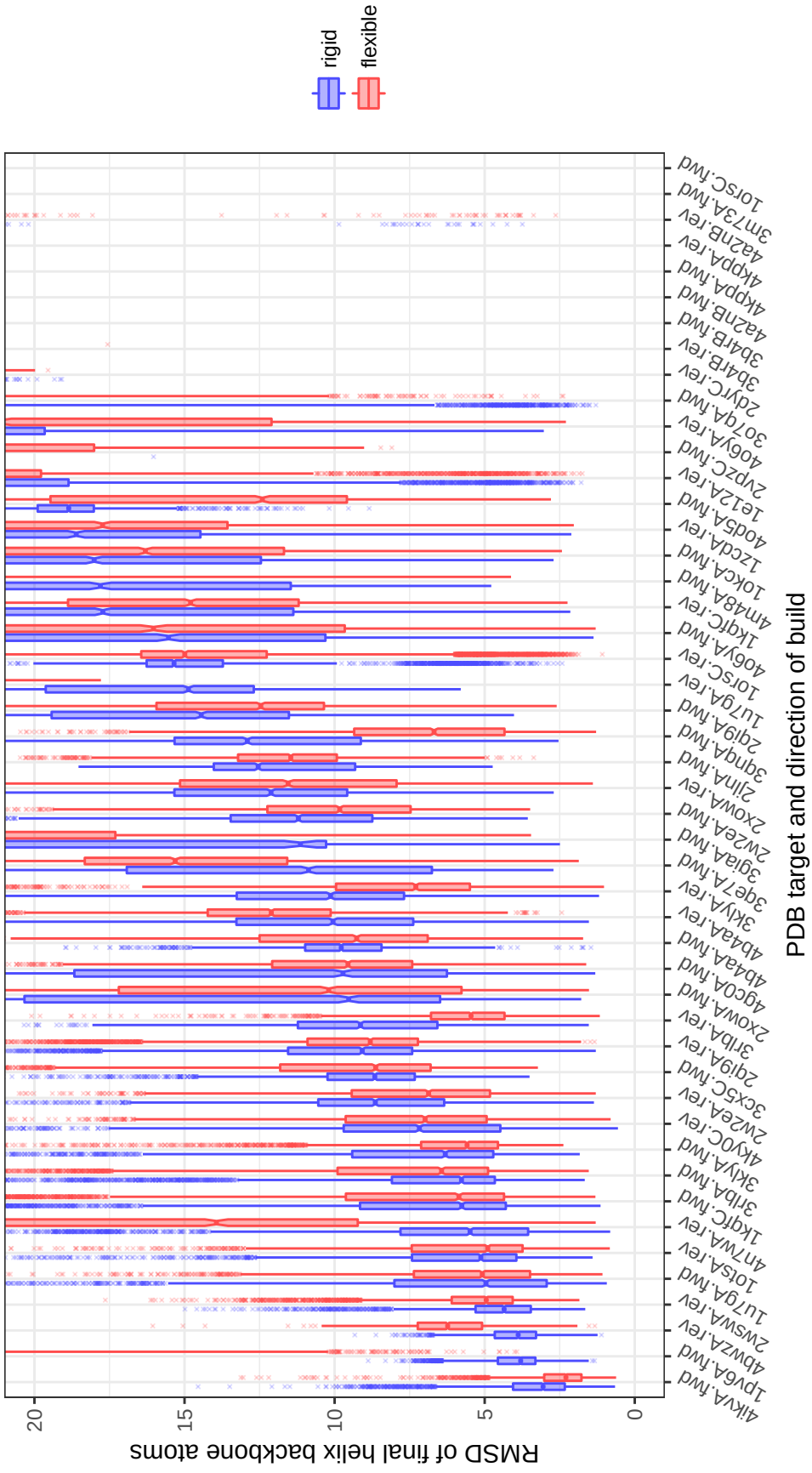


Figure 5.7: SAINT2-ScafFold performance on completing the final helix from a native structure. RMSD was calculated over the backbone atoms of the residues in the terminal transmembrane span of the protein. In the case of forward predictions, the helix to be predicted is C-terminal, and in reverse predictions it is N-terminal.

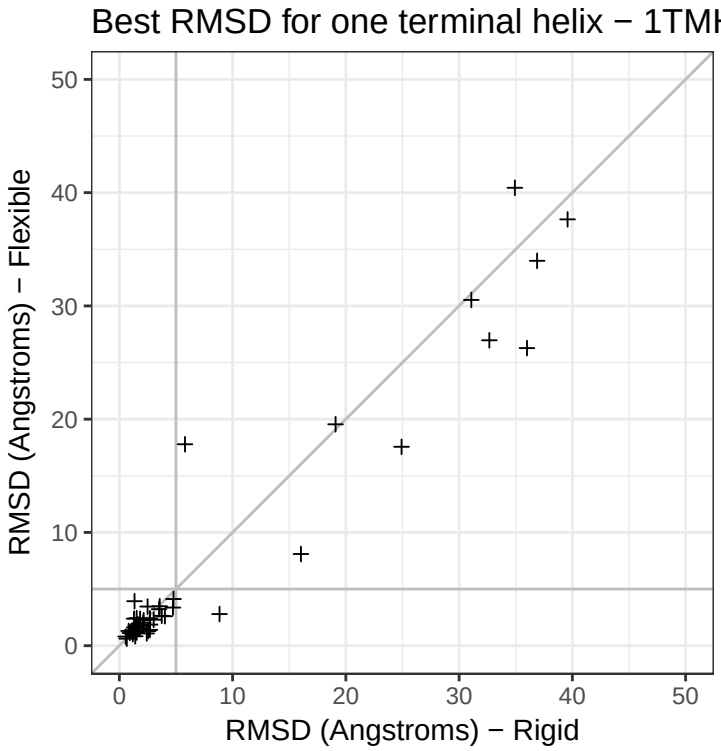


Figure 5.8: Best RMSD for predicting one terminal helix by the rigid or flexible mode.

several target helices, no model better than 20 Å was generated. In decoys that were inspected for these targets, it was common for the final two helices to be modelled as one continuous long helix. This may have been due to poor secondary structure prediction and a lack of appropriate fragments.

Our primary interest is the accuracy of the best model generated in the population, and whether it can be reliably selected using a decoy ranking score. Figure 5.8 compares the best out of 4,000 decoys produced by the rigid and flexible modes when predicting the terminal helix of the 48 target prediction scenarios.

For 38/48 target terminal helices (including prediction of both termini of some proteins), SAINT2-ScafFold was able to generate a model with RMSD < 5 Å in either rigid or flexible mode. The flexible mode generated a decoy of this quality for one target more than the rigid mode, but generally the accuracy of the two modes was very similar, with most points close to the line $y = x$.

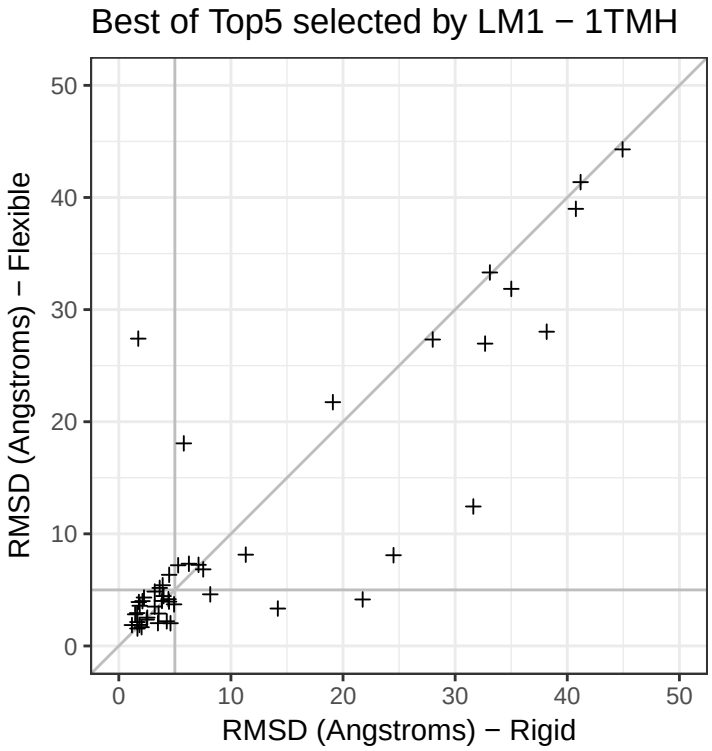


Figure 5.9: Best RMSD of the top five ranked decoys by LM1 when predicting one terminal helix by the rigid or flexible mode.

The decoy population for each target was also scored and ranked using the linear model developed in Chapter 4, LM1. Figure 5.9 shows the best RMSD of the top five ranked decoys for each target, comparing the flexible and rigid modes. By allowing a choice of the five top ranked models, without any form of clustering, I was able to select a model < 5 Å for 28/48 (rigid) or 27/48 (flexible) targets. If selection was restricted to just one decoy (Figure 5.10), the RMSD increased greatly, to the extent that many models are not accurate enough to be useful. It is common to submit five decoys for evaluation in assessments such as CASP, and in a practical structure prediction example, experimental data and human judgement may aid in choosing the most appropriate model of five. Therefore I used the best of the top five decoys (Best of Top5) to compare other aspects of model generation.

When extending the challenge to predicting two consecutive terminal helices, the performance of the flexible and rigid modes was similar, shown in Figure 5.11.

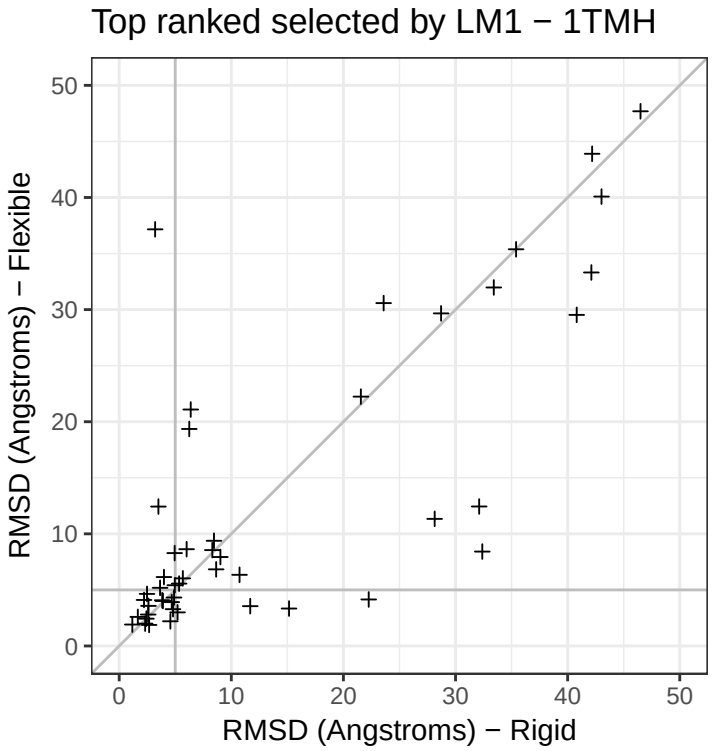


Figure 5.10: RMSD of the top ranked decoy by LM1 when predicting one terminal helix by the rigid or flexible mode.

At the most relevant region of RMSD < 10 Å, the rigid and flexible modes perform similarly, but achieve an RMSD < 5 Å for only $9/32$ and $10/32$ targets respectively. As the flexible mode samples a greater number of positions in the peptide chain, there is a computational cost, depending on the length of loops involved. For many targets, the flexible mode requires ~ 2 times the number of moves of the rigid mode, therefore it may not be worth the trade-off in computational cost if more decoys could otherwise be generated rigidly using equivalent computational power.

While more decoys could be generated by the rigid mode than the flexible mode in a given time, greater variation in a population of decoys may be needed in order for a larger number of decoys to be beneficial. Unsurprisingly, the flexible mode generated a greater diversity of decoys than the rigid mode. Figure 5.12 shows the standard deviation of RMSD by the two modes, when predicting either one or two helices. The difference is more pronounced for

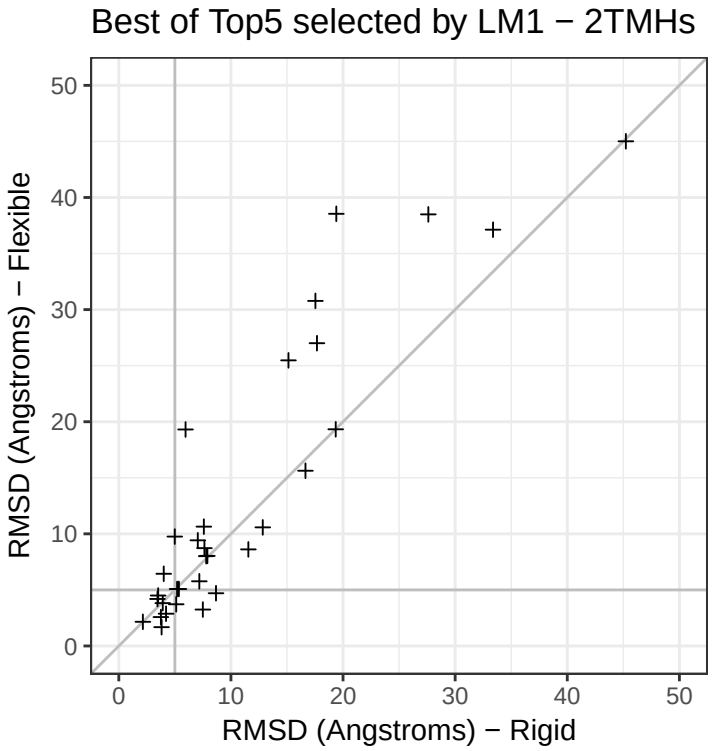


Figure 5.11: Best RMSD of the top five ranked decoys by LM1 when predicting two terminal helices by the rigid or flexible mode.

the two-helix predictions. Therefore, in order to improve the accuracy of the best decoys in a population, it may be necessary to invest extra computational time to generate many decoys in the flexible mode.

In a true incomplete template homology modelling scenario, there may be no predicted contacts between the template helices and the extra helix or helices to be predicted. I therefore tested a second decoy ranking method, LM2, which was trained without the inclusion of the contact potential in order to give appropriate weight to the other terms. Figure 5.13 shows that for prediction of either one or two helices, LM1 was much more successful, but LM2 still achieved < 5 Å RMSD for 16/48 one-helix targets and 5/32 two-helix targets. Having inspected the correlation between the LM scores and RMSD (not shown), the strength varied greatly from one target to another, but LM1, including the contact potential, was more consistent.

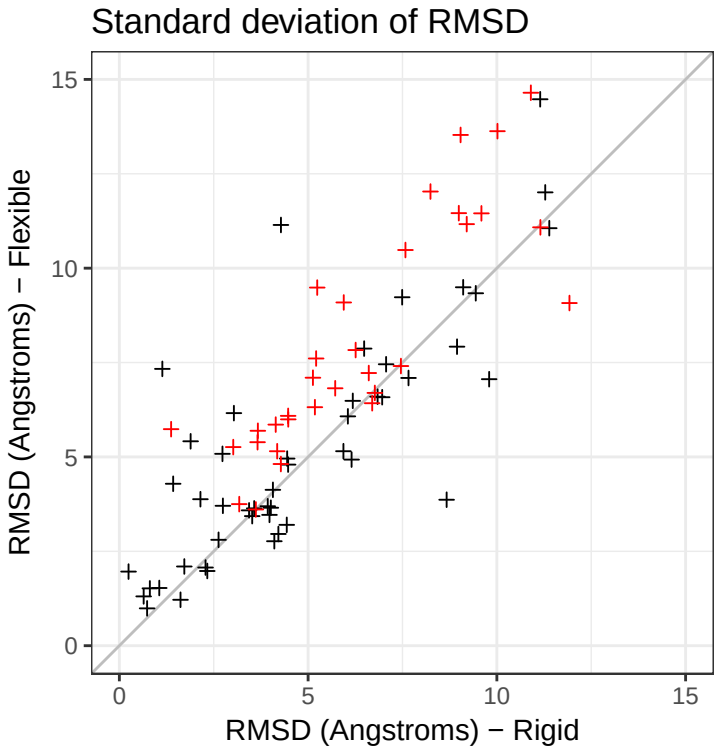


Figure 5.12: Standard deviation of the RMSD of all decoys when predicted by rigid or flexible. Predictions of one terminal helix are shown in black, two terminal helices in red.

5.3.3 Homology models completed by SAINT2-ScafFold

The above tests gave an indication of the likely performance of SAINT2-ScafFold in a homology modelling scenario starting from an incomplete template, however there were several simplifications. In order to more closely simulate such a challenge, I built homology models for eight of the longest protein targets, to observe how this might affect SAINT2-ScafFold (see Section 5.2.3). Figure 5.14 compares the accuracy of SAINT2-ScafFold starting from a homology model segment to starting from a native structure segment.

For this small set of test cases, the homology and native set ups achieved similar results. There were two targets where the homology segment performs better by > 1 Å: one was the two C-terminal helices of 2jlnA, for which the 'homology model' is actually the same protein in a different conformation. In the homology set up, the flexible mode was again similar to the rigid mode,

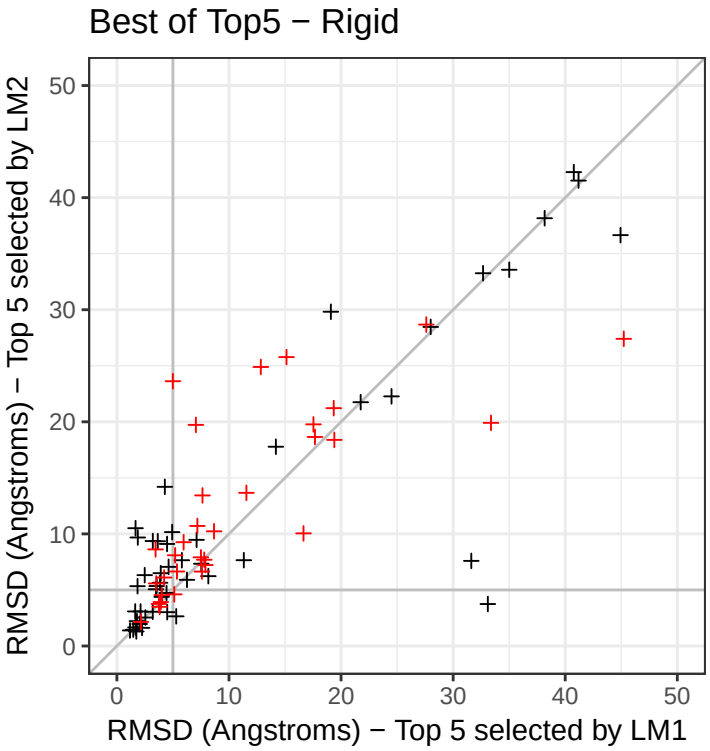


Figure 5.13: Best of the top five ranked decoys when ranked by LM2, compared to LM1, both using the rigid mode. Predictions of one terminal helix are shown in black, two terminal helices in red.

except for 2jlnA (Figure 5.15). There was very little difference in the ability of LM1 and LM2 to select a good decoy, except once again in 2jlnA, where LM1 was much more effective (Figure 5.16).

It is encouraging that these final helix completions can be relatively successful, and even moderately accurate in the case where the original template structure did not include the C-terminal helix to be predicted. However, there were a number of simplifications to the homology modelling pipeline for convenience, and a more rigorous test would be very useful to indicate the true performance of SAINT2-ScafFold in a real prediction scenario.

One main limitation was that the majority of the homology models were built from complete templates, which were then truncated to form segments for SAINT2-ScafFold. As a result, the loops and neighbouring helices are likely to accommodate the additional helix much more readily than an incomplete template from a family of proteins which has diverged and evolved so that

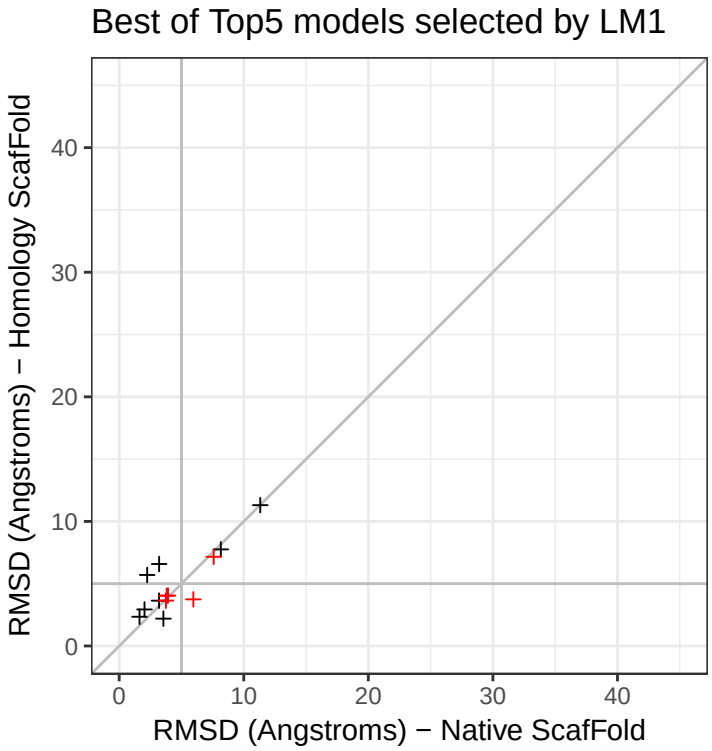


Figure 5.14: Comparison of Best of Top5 RMSD by rigid mode, selected by LM1, for the targets found in both the Native and Homology ScafFold sets. Predictions of one terminal helix are shown in black, two terminal helices in red.

some members have one more or fewer helices. In addition, the iMembrane annotation used in construction of the homology model was in some cases the annotation from the target itself, due to the web server selecting the closest match. Artificially accurate annotation of the transmembrane spans was also used in choosing the segment end point, in order to facilitate the comparison with native segments by ensuring the same segment length. Normally, the transmembrane helices would be predicted by one of the common machine learning methods (e.g. Tsirigos *et al.*, 2015), and therefore there would be less certainty about where the loop begins and how long the segment should be. Even if this is not known accurately, SAINT2-ScafFold has performed similarly here on a shorter segment to allow more flexibility. The Memoir pipeline also uses FREAD (Choi and Deane, 2010) to complete the core models output by Medeller, and the loop library it uses may contain the target loops themselves or similar loops from homologues. I also used the embedding of the native structure to easily

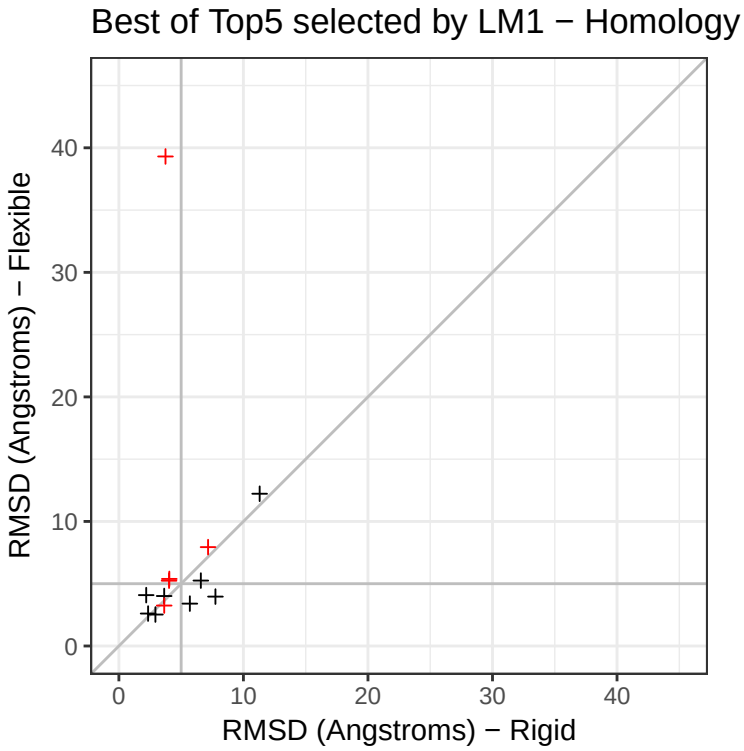


Figure 5.15: Best of Top5, selected by LM1, for rigid and flexible using the Homology set up. Predictions of one terminal helix are shown in black, two terminal helices in red.

embed the template, by aligning to it. This step could simply use the embedding predicted for the template instead, e.g. from MemProtMD (Stansfeld *et al.*, 2015), the OPM (Lomize *et al.*, 2011), or PDBTM (Kozma *et al.*, 2013).

The most thorough way to carry out a fairer and more realistic test would be to extract pairs of related alpha-helical membrane proteins from the PDB with differing numbers of transmembrane helices. One could find out the expected number of predicted contacts to an extra transmembrane helix by predicting contacts for families where members without solved structures are predicted to have additional helices by transmembrane helix prediction methods. It is expected that few contacts would be predicted, due to such prediction operating at a whole family level and depending on a very large number of sequences. It would be important to ensure that homologues are removed from every stage of annotation and loop modelling in addition to the removal from fragment libraries which is already carried out.

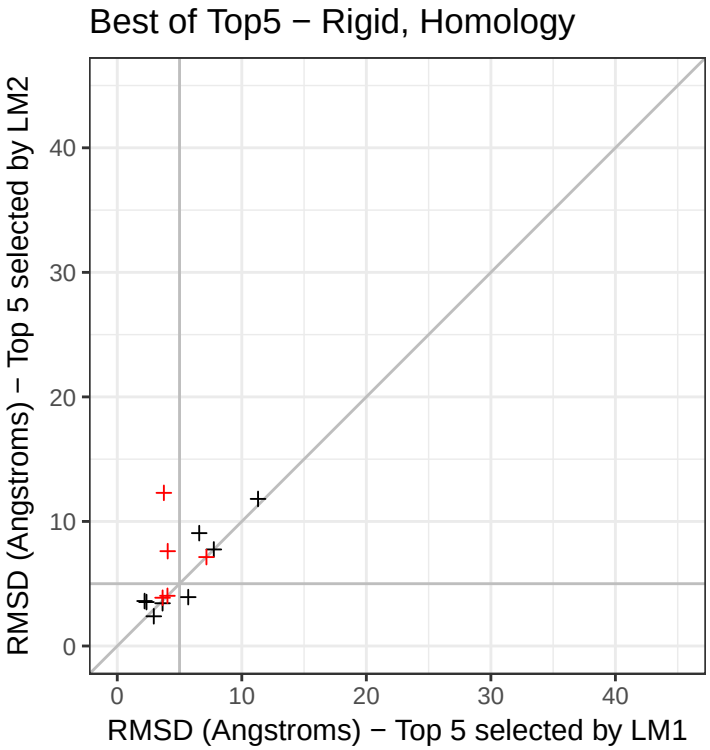


Figure 5.16: Best of Top5 ranked by LM1 and LM2 using the rigid Homology set up. Predictions of one terminal helix are shown in black, two terminal helices in red.

5.4 Conclusions

In this chapter I have shown that completion of homology models is a task which may often be required, but there are not yet tools specifically designed for this purpose. During the development of SAINT2-ScafFold, we were approached by a group seeking to model the final helix of a rhomboid type protease with seven transmembrane helices, GluP. A homology model had been generated using a template with six transmembrane helices, and I provided models of possible conformations for the seventh helix which are awaiting experimental validation.

Recognising the need for a reliable method to complete homology models with partial coverage, I tested the performance of SAINT2-ScafFold for this purpose. I demonstrate that by holding rigid the segment part of a homology model, the remaining section can be completed with an accuracy approaching that of the rest of the template. Selection of the best decoy is not always possible, but I show that even when predicted contacts are not considered

in decoy selection, some good models can be identified. One weakness of the normalised ranking method is that there is no indication of whether a predicted best ranked model is correct, with no objective measure of model quality. Use of a homology model to begin SAINT2-ScafFold did not greatly affect the RMSD of the predicted final helix or helices, which is encouraging for genuine prediction problems.

I also analysed the improvement in the accuracy of the best decoy as the number of decoys was increased by taking a samples from the population of decoys. It would be useful to determine the number of decoys to obtain a model close to the the best possible model, or the number required to obtain "good" model. For most targets the accuracy of the best decoy did not converge in the 4,000 decoys generated, therefore a greater number would be required to ascertain this. More thorough testing is required in order to better imitate the conditions in which such homology models would be built, but the indication of the preliminary results shown here is promising for the future.

6 Conclusions and future work

In this chapter, I summarise the themes that have been discussed in the previous chapters, and suggest possible directions for future research in each area.

6.1 Kink evolution and flexibility

In the work on kink conservation presented in Chapter 2, I studied angle differences, in addition to absolute values. This helped to move towards an understanding of helix kinks as points of conformational flexibility in a helix, and away from a binary classification. I found that there were many examples of kinks which differ between homologous structures, and the use of families helped to show the range of angles found.

Not presented here, I carried out some preliminary work towards constructing phylogenetic trees of kink angles in membrane proteins, but there were very few families with enough members. From this set, I was not able to draw any conclusions about the coevolution of kinks and sequence from the structure of trees. In order to gain a greater understanding of kink evolution, it would be beneficial to analyse the sequences of homologues for which no structure is available. An analysis of the soluble families, together with additional sequence data, could provide insights into the evolution of the sequence and

structure of kinks. A statistical model of kink angle changes over time may be a good description of the evolutionary process. With sufficient data to train on, parameters in a fitted model could indicate whether kinks more often evolve in abrupt steps, or through cumulative gradual mutations in surrounding residues. Knowledge of trends in kink evolution could help to inform homology modelling of kinks, and reliably predicted kink positions and sizes would be a valuable input for the *de novo* methods used in later chapters.

The error estimation method described could also be applied to trajectories of molecular dynamics simulations, to indicate when significant angle changes are taking place. It would be interesting to study examples of bound and unbound states of the same protein, in order to analyse the distributions of kink angles observed in each case.

6.2 Cotranslational folding in membrane proteins

Understanding of the insertion process of membrane proteins is important as expression of membrane proteins is one of the largest barriers to our understanding of their structure and function. The structural analyses in Chapter 3 provided a complementary line of evidence to the experimental studies which have suggested co-translational folding in membrane proteins. If it is true that transmembrane helices are sometimes inserted in pairs, this raises further questions about the mechanism of the translocon and the role of other insertases and chaperones such as YidC that may assist it. One could also ask how many membrane protein folds would not be topologically possible to obtain by insertion in a pairwise fashion, if the pairs do not dissociate after insertion. This mechanism is not likely to be the only method of insertion of membrane proteins, particularly as we know that some membrane proteins consist of inverted topology repeats (Forrest, 2015). Therefore, the original unit which was duplicated would likely require insertion to be possible in either orientation (Nasie *et al.*, 2010).

Computationally predicted coevolutionary contacts may give an indication of the most important interactions during folding as well as for stability of the folded structure. It may be possible to analyse patterns of these contacts in known structures, including those that are not satisfied in the native structure but may be correlated because they have roles in folding.

When observing solved membrane protein structures alone, the conclusions that can be drawn were limited due to a relatively small number of dissimilar experimental structures. However, as experimental methods improve and work continues, the power of statistical analysis techniques is always increasing and may yield more significant results.

The result seen in soluble proteins of cotranslational structure prediction (SAINT2 Forward) outperforming the reverse mode was not replicated in membrane proteins. This makes it difficult to draw any firm link between biological folding and the improvement seen in membrane proteins of SAINT2 Forward over SAINT2 In vitro. As the performance was not direction-dependent, the difference from In vitro may have resulted from the efficiency of the algorithm in general, without imitating biology. As no specific direction was favoured, the bi-directional sequential growth in the *ab initio* protocol of RosettaMembrane may be appropriate. However, in general, *de novo* conformational sampling techniques appear to be less important in the current age of protein structure prediction, where the quality of predicted contacts is the dominating factor determining which targets can be predicted.

6.3 SAINT2-ScafFold as a model for cotranslational folding

In Chapter 4, the ScafFold version of SAINT2 allowed us to investigate other possible aspects of cotranslational folding. We were surprised that the accuracy of prediction of the remainder of a protein was not reliably improved by providing a rigid segment as a scaffold against which to fold. There were clearly limitations of using a completely rigid segment, as this did not accurately

represent how an N-terminal section of a structure would exist while the remaining structure is translated. Allowing moves in the segment would imitate *in vivo* folding better, and these moves could compensate for a poor take-off point or a lack of accurate fragments. In order to retain the information in the correct first half, it may be appropriate to propose fewer moves in the "segment" part of the structure than in the growing region.

Fragment-based structure prediction is analogous to protein folding, but it may not produce the results that could be found with molecular dynamics using a full-atom or coarse-grained representation. As forcefields and computational power continue to improve, it may be possible to study the tertiary folding of larger numbers of membrane proteins by these methods. The interaction between a growing segment of a membrane protein and the translocon is not well understood, particularly how the translocon might stabilise inserted transmembrane helices. This, and other chaperone-type molecules in the membrane, could also be incorporated into coarse-grained models.

6.4 Adaptation of SAINT2 for membrane proteins

In Chapter 4, I described just one of many possible improvements and adaptations to the membrane environment. The Flib pipeline could be improved by using secondary structure prediction specific to membrane proteins (e.g. Leman *et al.*, 2013). I confirmed that the solvation potential trained in soluble proteins is not helpful for membrane protein structure prediction. A knowledge-based, membrane-layer specific solvent accessibility potential would likely help to orient helices correctly in the membrane layers, preferentially exposing hydrophobic residues to lipid tails. The membrane could also be modelled in an asymmetric way, using the predicted topology of helices and asymmetric versions of the potentials. When using these potentials, it would also be necessary to sample alternative positions of the protein in the membrane.

While attempting to imitate folding, my approach has not generated the accuracy of models produced by the leading methods. It has recently been suggested that membrane proteins may not depend on the membrane environment to adopt the correct fold (Popot and Engelman, 2016). The leading approaches rely more on the high accuracy of contacts predicted, rather than attempting to imitate folding or the membrane environment. Imitating the membrane environment requires additional computational time for calculation of the score at each step, which may be better spent generating more decoys. Using a recent membrane-specific method for prediction of coevolutionary contacts (e.g. Teixeira *et al.*, 2017) may improve the results of prediction, as could the use of metagenomic sequences (Ovchinnikov *et al.*, 2017).

6.5 Sampling efficiency

Through analysing acceptance rates in Section 4.2.3, we learned that moves in loop residues are less frequently accepted, as has been seen in other prediction programs (Kandathil *et al.*, 2016). This knowledge could be used within SAINT2 to bias move steps to take place more frequently in these regions. It may also be necessary to try running SAINT2 at different temperatures in order to improve sampling efficiency. It is also interesting that the extrusion moves affect the ability of the growing terminus to explore higher energy conformations. This is a difference between the sequential and In vitro modes of SAINT2, which could be investigated by accepting a small number of moves regardless of score in the In vitro mode. It would be useful to find out how the In vitro mode is affected by this, as this algorithm seems more likely to become trapped in a local minimum.

6.6 Completion of partial homology models

Chapter 5 introduced the application of SAINT2-ScafFold to the task of completing homology models, but with many simplifications. To determine its usefulness for real prediction cases, it will be necessary to extract a representative

test set of related proteins of different lengths. The prediction of the non-overlapping regions can be tested by using the shorter protein as template for homology modelling, then using SAINT2 to complete the structure.

As we know that predicted contacts are important for the accuracy of models, different sized samples could be taken from the available predicted contacts to the unmodelled region. In this way, someone seeking to complete a partial model for a membrane protein could estimate the likely accuracy of the resulting model based on the number of contacts both between the segment and non-segment regions, and within the non-segment region.

I found that accuracy in the non-segment region was not improved by flexibility in the last part of the structure in common between the models. It may be worth testing whether flexibility in other regions of the template, for example loops predicted by Medeller with low confidence, should be remodelled. This could be carried out by fragment replacement in SAINT2, by a hybrid *ab intitio*/knowledge-based loop modelling method (Marks *et al.*, 2017), or by a method similar to the local Monte-Carlo search used in I-TASSER (Yang *et al.*, 2015).

The SAINT2-ScafFold protocol could also be useful for completion of homology models of soluble proteins. While SAINT2-ScafFold is currently designed for growth of a protein chain from one terminus, it could be changed to allow fragment substitutions at any location in the chain. For example, this could introduce flexibility between two domains for which homology models are available, or to allow both termini to be extruded and sampled simultaneously. SAINT2 offers a framework with potential to tackle many structure prediction problems in both membrane and soluble proteins, using ideas inspired by biology.

Bibliography

Alam, A. and Jiang, Y. (2009). High-resolution structure of the open NaK channel. *Nat. Struct. Mol. Biol.*, **16**(1), 30–34. (Cited on pages 35 and 39.)

Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A., and Gray, J. J. (2015). An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.*, **11**(9), e1004398. (Cited on pages 17, 80, 83, 84, 107, 118, and 154.)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–410. (Cited on page 23.)

Altschul, S. F., Madden, T. L., Schäffer, 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 Res.*, **25**(17), 3389–3402. (Cited on pages 23, 153, and 156.)

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. *Proc. Natl. Acad. Sci. U. S. A.*, **47**(9), 1309–14. (Cited on page 8.)

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**(5481), 905–20. (Cited on page 11.)

Bansal, M., Kumar, S., and Velavan, R. (2000). HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, **17**(5), 811–9. (Cited on page 36.)

Barrett, P. J., Song, Y., Van Horn, W. D., Hustedt, E. J., Schafer, J. M., Hadziselimovic, A., Beel, A. J., and Sanders, C. R. (2012). The amyloid precursor protein has a flexible transmembrane domain and binds cholesterol. *Science (80-.).*, **336**(6085), 1168–71. (Cited on pages 35 and 39.)

Barth, P., Schonbrun, J., and Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl. Acad. Sci.*, **104**(40), 15682–15687. (Cited on page 84.)

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl. Acad. Sci. U. S. A.*, **106**(5), 1409–14. (Cited on page 26.)

Ben-Shem, A., Jenner, L., Yusupova, G., and Yusupov, M. (2010). Crystal structure of the eukaryotic ribosome. *Science*, **330**(6008), 1203–9. (Cited on page 11.)

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Res.*, **41**(D1), D36–D42. (Cited on page 19.)

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**(12), 980. (Cited on page 43.)

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235–42. (Cited on page 7.)

Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2012). The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure*, **20**(3), 391–6. (Cited on page 19.)

Betancourt, M. R. and Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, **59**(5), 305–309. (Cited on page 21.)

Bettinelli, I., Graziani, D., Marconi, C., Pedretti, A., and Vistoli, G. (2011). The approach of conformational chimeras to model the role of proline-containing helices on GPCR mobility: the fertile case of Cys-LTR1. *ChemMedChem*, **6**(7), 1217–1227. (Cited on pages 18, 35, and 40.)

Bhattacharya, D., Cao, R., and Cheng, J. (2016). UniCon3D: De novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics*, **32**(18), 2791–2799. (Cited on page 30.)

Bischoff, L., Wickles, S., Berninghausen, O., van der Sluis, E. O., and Beckmann, R. (2014). Visualization of a polytopic membrane protein during SecY-mediated membrane insertion. *Nat. Commun.*, **5**, 4103. (Cited on page 75.)

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., and Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell*, **61**(3). (Cited on page 11.)

Caffrey, M., Li, D., and Dukkipati, A. (2012). Membrane protein structure determination using crystallography and lipidic mesophases: recent advances and successes. *Biochemistry*, **51**(32), 6266–88. (Cited on page 15.)

Cao, Z., Hutchison, J. M., Sanders, C. R., and Bowie, J. U. (2017). Backbone Hydrogen Bond Strengths Can Vary Widely in Transmembrane Helices. *J. Am. Chem. Soc.*, **139**(31), 10742–10749. (Cited on page 39.)

Chen, K. Y. M., Sun, J., Salvo, J. S., Baker, D., and Barth, P. (2014). High-Resolution Modeling of Transmembrane Helical Protein Structures from Distant Homologues. *PLoS Comput. Biol.*, **10**(5), e1003636. (Cited on pages 25, 68, and 152.)

Choe, S. and Grabe, M. (2009). Conformational dynamics of the inner pore helix of voltage-gated potassium channels. *J Chem Phys*, **130**(21), 215103. (Cited on pages 18 and 40.)

Choi, Y. and Deane, C. M. (2010). FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins Struct. Funct. Bioinforma.*, **78**(6), 1431–1440. (Cited on pages 22, 24, 152, and 172.)

Cong, Q., Kinch, L. N., Pei, J., Shi, S., Grishin, V. N., Li, W., and Grishin, N. V. (2011). An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics*, **27**(24), 3371–3378. (Cited on page 22.)

Cordes, F. S., Bright, J. N., and Sansom, M. S. (2002). Proline-induced distortions of transmembrane helices. *J. Mol. Biol.*, **323**(5), 951–960. (Cited on page 38.)

Cymer, F. and von Heijne, G. (2013). Cotranslational folding of membrane proteins probed by arrest-peptide-mediated force measurements. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(36), 14640–5. (Cited on pages 71 and 72.)

Cymer, F., Ismail, N., and von Heijne, G. (2014). Weak pulling forces exerted on Nin-orientated transmembrane segments during co-translational insertion into the inner membrane of Escherichia coli. *FEBS Lett.*, **588**(10), 1930–4. (Cited on page 72.)

Cymer, F., von Heijne, G., and White, S. H. (2015). Mechanisms of Integral Membrane Protein Insertion and Folding. *J. Mol. Biol.*, **427**(5), 999–1022. (Cited on pages 74, 75, and 107.)

Dalbey, R. E., Kuhn, A., Zhu, L., and Kiefer, D. (2014). The membrane insertase YidC. *Biochim. Biophys. Acta*, **1843**(8), 1489–96. (Cited on page 72.)

Dang, H., England, P. M., Farivar, S. S., Dougherty, D. A., and Lester, H. A. (2000). Probing the role of a conserved M1 proline residue in 5-hydroxytryptamine(3) receptor gating. *Mol. Pharmacol.*, **57**(6), 1114–1122. (Cited on page 39.)

De Marothy, M. T. and Elofsson, A. (2015). Marginally hydrophobic transmembrane α-helices shaping membrane protein folding. *Protein Sci.*, **24**(7), 1057–74. (Cited on page 74.)

de Oliveira, S., Law, E., Shi, J., and Deane, C. (2017a). Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. *Bioinformatics*, page btx722. (Cited on pages 30, 33, and 71.)

de Oliveira, S. H. P., Shi, J., and Deane, C. M. (2015). Building a better fragment library for de novo protein structure prediction. *PLoS One*, **10**(4), e0123998. (Cited on pages 30, 31, 76, 87, and 152.)

de Oliveira, S. H. P., Shi, J., and Deane, C. M. (2017b). Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics*, **33**(3), 373–381. (Cited on page 28.)

Deane, C. M., Dong, M., Huard, F. P. E., Lance, B. K., and Wood, G. R. (2007). Cotranslational protein folding–fact or fiction? *Bioinformatics*, **23**(13), i142–8. (Cited on pages 12, 76, 81, 83, 88, and 89.)

Deupi, X. (2012). Quantification of structural distortions in the transmembrane helices of GPCRs. *Methods Mol. Biol.*, **914**, 219–35. (Cited on pages 18, 35, and 40.)

Dill, K. A. and Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.*, **4**(1), 10–19. (Cited on page 9.)

Dill, K. A. and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, **338**(6110), 1042–6. (Cited on page 9.)

D'Rozario, R. S. G. and Sansom, M. S. P. (2008). Helix dynamics in a membrane transport protein: comparative simulations of the glycerol-3-phosphate transporter and its constituent helices. *Mol. Membr. Biol.*, **25**(6-7), 571–83. (Cited on pages 35 and 39.)

Ebejer, J.-P., Hill, J. R., Kelm, S., Shi, J., and Deane, C. M. (2013). Memoir: template-based structure prediction for membrane proteins. *Nucleic Acids Res.*, **41**(Web Server issue), W379—-W383. (Cited on pages 151 and 158.)

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**(10), e1002195. (Cited on page 23.)

Ellis, J. J., Huard, F. P. E., Deane, C. M., Srivastava, S., and Wood, G. R. (2010). Directionality in protein fold prediction. *BMC Bioinformatics*, **11**(1), 172. (Cited on pages 10 and 76.)

England, P. M., Zhang, Y., Dougherty, D. A., and Lester, H. A. (1999). Backbone mutations in transmembrane domains of a ligand-gated ion channel: Implications for the mechanism of gating. *Cell*, **96**(1), 89–98. (Cited on pages 35 and 39.)

Faure, G., Ogurtsov, A. Y., Shabalina, S. A., and Koonin, E. V. (2016). Role of mRNA structure in the control of protein folding. *Nucleic Acids Res.*, page gkw671. (Cited on page 11.)

Fedorov, A. N. and Baldwin, T. O. (1999). Process of biosynthetic protein folding determines the rapid formation of native structure. *J. Mol. Biol.*, **294**(2), 579–86. (Cited on page 10.)

Fischer, A. W., Alexander, N. S., Woetzel, N., Karakas, M., Weiner, B. E., and Meiler, J. (2015). BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints. *Proteins*. (Cited on page 26.)

Forrest, L. R. (2015). Structural Symmetry in Membrane Proteins. *Annu. Rev. Biophys.*, **44**(1), 311–337. (Cited on page 178.)

Forrest, L. R., Tang, C. L., and Honig, B. (2006). On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins. *Biophys. J.*, **91**(2), 508–517. (Cited on pages 25 and 151.)

Gimpelev, M., Forrest, L. R., Murray, D., and Honig, B. (2004). Helical packing patterns in membrane and soluble proteins. *Biophys. J.*, **87**(6), 4075–86. (Cited on page 98.)

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Bioinforma.*, **18**(4), 309–317. (Cited on page 27.)

Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M., and Baker, D. (2011). Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS One*, **6**(8), e23294. (Cited on page 30.)

Hall, S. E., Roberts, K., and Vaidehi, N. (2009). Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *J. Mol. Graph. Model.*, **27**(8), 944–950. (Cited on pages 18 and 38.)

Hartl, F. U. and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, **295**(5561), 1852–8. (Cited on page 10.)

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H., and von Heijne, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **433**(7024), 377–81. (Cited on page 16.)

Hilger, D., Polyhach, Y., Jung, H., and Jeschke, G. (2009). Backbone structure of transmembrane domain IX of the Na+/proline transporter PutP of Escherichia coli. *Biophys. J.*, **96**(1), 217–225. (Cited on pages 35 and 39.)

Hill, J. R. and Deane, C. M. (2013). MP-T: improving membrane protein alignment for structure prediction. *Bioinformatics*, **29**(1), 54–61. (Cited on page 23.)

Hill, J. R., Kelm, S., Shi, J., and Deane, C. M. (2011). Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, **27**(13), i15–23. (Cited on page 24.)

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**(7), 1607–21. (Cited on pages 27, 28, 77, 109, and 150.)

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**(4), 789–802. (Cited on page 10.)

Isberg, V., Vroling, B., van der Kant, R., Li, K., Vriend, G., and Gloriam, D. (2014). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.*, **42**(Database issue), D422–5. (Cited on pages 43, 45, 59, and 62.)

Ismail, N., Hedman, R., Schiller, N., and von Heijne, G. (2012). A biphasic pulling force acts on transmembrane helices during translocon-mediated membrane integration. *Nat. Struct. Mol. Biol.*, **19**(10), 1018–22. (Cited on pages 71 and 72.)

Ismail, N., Hedman, R., Lindén, M., and von Heijne, G. (2015). Charge-driven dynamics of nascent-chain movement through the SecYEG translocon. *Nat. Struct. Mol. Biol.*, **22**(2), 145–9. (Cited on page 72.)

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**(2), 195–202. (Cited on page 30.)

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–90. (Cited on page 27.)

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**(7), 999–1006. (Cited on pages 22, 28, and 87.)

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–637. (Cited on pages 5, 80, and 140.)

Kahsay, R. Y., Gao, G., and Liao, L. (2005). An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics*, **21**(9), 1853–8. (Cited on page 20.)

Kandathil, S. M., Handl, J., and Lovell, S. C. (2016). Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. *Proteins Struct. Funct. Bioinforma.*, **84**(4), 411–426. (Cited on pages 30, 142, and 181.)

Kang, H. J., Lee, C., and Drew, D. (2013). Breaking the barriers in membrane protein crystallography. *Int. J. Biochem. Cell Biol.*, **45**(3), 636–44. (Cited on page 14.)

Katritch, V., Cherezov, V., and Stevens, R. C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.*, **53**, 531–56. (Cited on pages 18, 35, 40, and 59.)

Kauko, A., Illergård, K., and Elofsson, A. (2008). Coils in the membrane core are conserved and functionally important. *J Mol Biol*, **380**(1), 170–180. (Cited on page 39.)

Kelm, S., Shi, J., and Deane, C. M. (2009). iMembrane: homology-based membrane-insertion of proteins. *Bioinformatics*, **25**(8), 1086–8. (Cited on pages 16, 17, 24, 44, 78, 80, 88, 136, and 158.)

Kelm, S., Shi, J., and Deane, C. M. (2010). MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics*, **26**(22), 2833–40. (Cited on pages 24, 151, 152, and 158.)

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**(4610), 662–6. (Cited on page 7.)

Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y., and Grishin, N. V. (2011). CASP9 assessment of free modeling target predictions. *Proteins Struct. Funct. Bioinforma.*, **79**(S10), 59–73. (Cited on page 22.)

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins Struct. Funct. Bioinforma.*, **84**(S1), 51–66. (Cited on pages 25 and 27.)

Kneissl, B., Mueller, S. C., Tautermann, C. S., and Hildebrandt, A. (2011). String kernels and high-quality data set for improved prediction of kinked helices in α-helical membrane proteins. *J. Chem. Inf. Model.*, **51**(11), 3017–3025. (Cited on page 37.)

Koehler Leman, J., Ulmschneider, M. B., and Gray, J. J. (2015). Computational modeling of membrane proteins. *Proteins*, **83**(1), 1–24. (Cited on pages 19, 109, and 151.)

Komar, A. A., Kommer, A., Krashennikov, I. A., and Spirin, A. s. (1993). Cotranslational heme binding to nascent globin chains. *FEBS Lett.*, **326**(1-3), 261–263. (Cited on page 11.)

Kosciolek, T. and Jones, D. T. (2014). De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One*, **9**(3), e92197. (Cited on pages 22 and 76.)

Kozma, D., Simon, I., and Tusnády, G. E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**(Database issue), D524–9. (Cited on pages 17, 43, 80, 83, and 173.)

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**(3), 567–80. (Cited on pages 20 and 153.)

Kumar, P. and Bansal, M. (2012). HELANAL-Plus: a web server for analysis of helix geometry in protein structures. *J. Biomol. Struct. Dyn.*, **30**(6), 773–783. (Cited on page 37.)

Kumazaki, K., Chiba, S., Takemoto, M., Furukawa, A., Nishiyama, K.-i., Sugano, Y., Mori, T., Dohmae, N., Hirata, K., Nakada-Nakura, Y., Maturana, A. D., Tanaka, Y., Mori, H., Sugita, Y., Arisaka, F., Ito, K., Ishitani, R., Tsukazaki, T.,

and Nureki, O. (2014). Structural basis of Sec-independent membrane protein insertion by YidC. *Nature*, **509**(7501), 516–20. (Cited on page 72.)

Kupitz, C., Basu, S., Grotjohann, I., Fromme, R., Zatsepin, N. A., Rendek, K. N., Hunter, M. S., Shoeman, R. L., White, T. A., Wang, D., James, D., Yang, J.-H., Cobb, D. E., Reeder, B., Sierra, R. G., Liu, H., Barty, A., Aquila, A. L., Deponte, D., Kirian, R. A., Bari, S., Bergkamp, J. J., Beyerlein, K. R., Bogan, M. J., Caleman, C., Chao, T.-C., Conrad, C. E., Davis, K. M., Fleckenstein, H., Galli, L., Hau-Riege, S. P., Kassemeyer, S., Laksmono, H., Liang, M., Lomb, L., Marchesini, S., Martin, A. V., Messerschmidt, M., Milathianaki, D., Nass, K., Ros, A., Roy-Chowdhury, S., Schmidt, K., Seibert, M., Steinbrener, J., Stellato, F., Yan, L., Yoon, C., Moore, T. A., Moore, A. L., Pushkar, Y., Williams, G. J., Boutet, S., Doak, R. B., Weierstall, U., Frank, M., Chapman, H. N., Spence, J. C. H., and Fromme, P. (2014). Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature*, **513**(7517), 261–265. (Cited on page 15.)

Langelaan, D. N., Wieczorek, M., Blouin, C., and Rainey, J. K. (2010). Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J. Chem. Inf. Model.*, **50**(12), 2213–20. (Cited on pages 37, 38, 44, and 68.)

Langelaan, D. N., Reddy, T., Banks, A. W., Dellaire, G., Dupré, D. J., and Rainey, J. K. (2013). Structural features of the apelin receptor N-terminal tail and first transmembrane segment implicated in ligand binding and receptor trafficking. *Biochim. Biophys. Acta*, **1828**(6), 1471–83. (Cited on page 63.)

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. Crystallogr.*, **26**(2), 283–291. (Cited on page 4.)

Law, E. C., Wilman, H. R., Kelm, S., Shi, J., and Deane, C. M. (2016). Examining the Conservation of Kinks in Alpha Helices. *PLoS One*, **11**(6), e0157553. (Cited on pages 35, 43, 48, and 67.)

Lebon, G., Warne, T., Edwards, P. C., Bennett, K., Langmead, C. J., Leslie, A. G. W., and Tate, C. G. (2011). Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature*, **474**(7352), 521–5. (Cited on page 68.)

Leman, J. K., Mueller, R., Karakas, M., Woetzel, N., and Meiler, J. (2013). Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins*, **81**(7), 1127–40. (Cited on pages 20 and 180.)

Leman, J. K., Mueller, B. K., and Gray, J. J. (2017). Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics*, **33**(5), 754–756. (Cited on page 17.)

Levinthal, C. (1969). How to fold graciously. *Mossbauer Spectrosc. Biol. Syst.*, **67**, 22–24. (Cited on page 9.)

Li, E., Wimley, W. C., and Hristova, K. (2012). Transmembrane helix dimerization: Beyond the search for sequence motifs. *Biochim. Biophys. Acta - Biomembr.*, **1818**(2), 183–193. (Cited on page 20.)

Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., Agard, D. A., and Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods*, **10**(6), 584–90. (Cited on page 8.)

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science (80-.).*, **334**(6055), 517–520. (Cited on page 19.)

Liu, W., Chun, E., Thompson, A. A., Chubukov, P., Xu, F., Katritch, V., Han, G. W., Roth, C. B., Heitman, L. H., IJzerman, A. P., Cherezov, V., and Stevens,

R. C. (2012). Structural basis for allosteric regulation of GPCRs by sodium ions. *Science*, **337**(6091), 232–6. (Cited on pages 64 and 68.)

Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (2011). Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *J. Chem. Inf. Model.*, **51**(4), 930–46. (Cited on page 173.)

Lomize, M. A., Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**(5), 623–5. (Cited on pages 17, 43, 78, 80, and 88.)

Lu, Y., Turnbull, I. R., Bragin, A., Carveth, K., Verkman, A., and Skach, W. R. (2000). Reorientation of Aquaporin-1 Topology during Maturation in the Endoplasmic Reticulum. *Mol. Biol. Cell*, **11**(9), 2973–2985. (Cited on page 75.)

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**(15), 3255–63. (Cited on page 45.)

Mahlab, S. and Linial, M. (2014). Speed controls in translating secretory proteins in eukaryotes–an evolutionary perspective. *PLoS Comput. Biol.*, **10**(1), e1003294. (Cited on page 11.)

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722–2728. (Cited on page 22.)

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., and Deane, C. M. (2017). Sphinx: Merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, **33**(9), 1346–1353. (Cited on page 182.)

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**(12), e28766. (Cited on page 27.)

McCaughan, L. and Krimm, S. (1982). Biochemical profiles of membranes from x-ray and neutron diffraction. *Biophys. J.*, **37**(2), 417–26. (Cited on page 13.)

McDermott, A. (2009). Structure and dynamics of membrane proteins by magic angle spinning solid-state NMR. *Annu. Rev. Biophys.*, **38**, 385–403. (Cited on page 14.)

Meruelo, A. D., Samish, I., and Bowie, J. U. (2011). TMKink: a method to predict transmembrane helix kinks. *Protein Sci.*, **20**(7), 1256–64. (Cited on page 68.)

Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., and Elofsson, A. (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics*, **30**(17), i482–i488. (Cited on page 22.)

Michel, M., Skwark, M. J., Menéndez Hurtado, D., Ekeberg, M., and Elofsson, A. (2017). Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics*, **33**(18), 2859–2866. (Cited on pages 28 and 29.)

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**(7), 617–23. (Cited on pages 24, 43, 46, and 140.)

Montelione, G. T. (2012). The Protein Structure Initiative: achievements and visions for the future. *F1000 Biol. Rep.*, **4**(7), 7. (Cited on pages 19 and 151.)

Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)-round IX. *Proteins Struct. Funct. Bioinforma.*, **79**(S10), 1–5. (Cited on page 21.)

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, **82 Suppl 2**(0 2), 1–6. (Cited on page 21.)

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins*, **84 Suppl 1**(Suppl 1), 4–14. (Cited on page 21.)

Narunsky, A., Nepomnyachiy, S., Ashkenazy, H., Kolodny, R., and Ben-Tal, N. (2015). ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure. *Structure*, **23**(11), 2162–2170. (Cited on page 8.)

Nasie, I., Steiner-Mordoch, S., Gold, A., and Schuldiner, S. (2010). Topologically random insertion of EmrE supports a pathway for evolution of inverted repeats in ion-coupled transporters. *J. Biol. Chem.*, **285**(20), 15234–15244. (Cited on page 178.)

Ni, Z., Bikadi, Z., Shuster, D. L., Zhao, C., Rosenberg, M. F., and Mao, Q. (2011). Identification of proline residues in or near the transmembrane helices of the human breast cancer resistance protein (BCRP/ABCG2) that are important for transport activity and substrate specificity. *Biochemistry*, **50**(37), 8057–8066. (Cited on pages 35 and 39.)

Nicola, A. V., Chen, W., and Helenius, A. (1999). Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat. Cell Biol.*, **1**(6), 341–5. (Cited on page 11.)

Nugent, T. and Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**(1), 159. (Cited on page 20.)

Nugent, T. and Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(24), E1540–7. (Cited on pages 27 and 77.)

Nugent, T. and Jones, D. T. (2013). Membrane protein orientation and refinement using a knowledge-based statistical potential. *BMC Bioinformatics*, **14**(1), 276. (Cited on pages 17, 114, 115, and 125.)

Oberai, A., Joh, N. H., Pettit, F. K., and Bowie, J. U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **106**(42), 17747–50. (Cited on page 14.)

Ott, C. M. and Lingappa, V. R. (2002). Integral membrane protein biosynthesis: why topology is hard to predict. *J. Cell Sci.*, **115**(10), 2003–2009. (Cited on page 18.)

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N. V., and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*, **4**, e09248. (Cited on page 109.)

Ovchinnikov, S., Kim, D. E., Wang, R. Y. R., Liu, Y., Dimaio, F., and Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins Struct. Funct. Bioinforma.*, **84**(S1), 67–75. (Cited on pages 25, 29, and 77.)

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science (80-.).*, **355**(6322), 294–298. (Cited on pages 22, 29, 109, 150, 151, and 181.)

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**(12), 993–6. (Cited on page 14.)

Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trézéguet, V., Lauquin, G. J.-M., and Brandolin, G. (2003). Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature*, **426**(6962), 39–44. (Cited on pages 35 and 39.)

Pedretti, A., Mazzolari, A., Ricci, C., and Vistoli, G. (2015). Enhancing the Reliability of GPCR Models by Accounting for Flexibility of Their Pro-Containing Helices: the Case of the Human mAChR1 Receptor. *Mol. Inform.*, **34**(4), 216–227. (Cited on page 40.)

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E. C., Pettersen, E. F., Huang, C. C., Datta, R. S., Sampathkumar, P., Madhusudhan, M. S., Sjolander, K., Ferrin, T. E., Burley, S. K., and Sali, A. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **39**(Database), D465–D474. (Cited on page 153.)

Pieper, U., Schlessinger, A., Kloppmann, E., Chang, G. A., Chou, J. J., Dumont, M. E., Fox, B. G., Fromme, P., Hendrickson, W. A., Malkowski, M. G., Rees, D. C., Stokes, D. L., Stowell, M. H. B., Wiener, M. C., Rost, B., Stroud, R. M., Stevens, R. C., and Sali, A. (2013). Coordinating the impact of structural genomics on the human α-helical transmembrane proteome. *Nat. Struct. Mol. Biol.*, **20**(2), 135–138. (Cited on pages 152, 153, and 158.)

Popot, J. L. and Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, **29**(17), 4031–4037. (Cited on page 73.)

Popot, J. L. and Engelman, D. M. (2016). Membranes Do Not Tell Proteins How to Fold. *Biochemistry*, **55**(1), 5–18. (Cited on page 181.)

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B.-H., Das, R., Grishin, N. V., and Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, **77 Suppl 9**, 89–99. (Cited on page 76.)

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, **450**(7170), 663–9. (Cited on pages 71 and 72.)

Reis, M. d., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, **32**(17), 5036–5044. (Cited on page 11.)

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**(2), 173–175. (Cited on page 23.)

Ri, Y., Ballesteros, J. A., Abrams, C. K., Oh, S., Verselis, V. K., Weinstein, H., and Bargiello, T. A. (1999). The role of a conserved proline residue in mediating conformational changes associated with voltage gating of Cx32 gap junctions. *Biophys. J.*, **76**(6), 2887–2898. (Cited on pages 35 and 39.)

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93. (Cited on pages 26 and 100.)

Rollauer, S. E., Tarry, M. J., Graham, J. E., Jääskeläinen, M., Jäger, F., Johnson, S., Krehenbrink, M., Liu, S.-M., Lukey, M. J., Marcoux, J., McDowell, M. A., Rodriguez, F., Roversi, P., Stansfeld, P. J., Robinson, C. V., Sansom, M. S. P., Palmer, T., Högbom, M., Berks, B. C., and Lea, S. M. (2012). Structure of the TatC core of the twin-arginine protein transport system. *Nature*, **492**(7428), 210–4. (Cited on page 101.)

Rosato, A., Tejero, R., and Montelione, G. T. (2013). Quality assessment of protein NMR structures. *Curr. Opin. Struct. Biol.*, **23**(5), 715–24. (Cited on page 7.)

Rosenberg, M. F., Bikadi, Z., Hazai, E., Starborg, T., Kelley, L., Chayen, N. E., Ford, R. C., and Mao, Q. (2015). Three-dimensional structure of the human breast cancer resistance protein (BCRP/ABCG2) in an inward-facing conformation. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **71**(8), 1725–1735. (Cited on page 39.)

Sadlish, H., Pitonzo, D., Johnson, A. E., and Skach, W. R. (2005). Sequential triage of transmembrane segments by Sec61alpha during biogenesis of a native multispanning membrane protein. *Nat. Struct. Mol. Biol.*, **12**(10), 870–8. (Cited on page 75.)

Šali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.*, **234**(3), 779–815. (Cited on page 25.)

Samudrala, R. and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction 1 1Edited by F. Cohen. *J. Mol. Biol.*, **275**(5), 895–916. (Cited on page 25.)

Sander, I. M., Chaney, J. L., and Clark, P. L. (2014). Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.*, **136**(3), 858–61. (Cited on pages 10 and 11.)

Sansom, M. S. and Weinstein, H. (2000). Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. *Trends Pharmacol. Sci.*, **21**(11), 445–451. (Cited on page 39.)

Sansom, M. S. P., Scott, K. A., and Bond, P. J. (2008). Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem. Soc. Trans.*, **36**(Pt 1), 27–32. (Cited on pages 17 and 80.)

Sato, Y., Sakaguchi, M., Goshima, S., Nakamura, T., and Uozumi, N. (2003). Molecular dissection of the contribution of negatively and positively charged residues in S2, S3, and S4 to the final membrane topology of the voltage sensor in the K+ channel, KAT1. *J. Biol. Chem.*, **278**(15), 13227–34. (Cited on page 74.)

Saunders, R. and Deane, C. M. (2010a). Protein structure prediction begins well but ends badly. *Proteins*, **78**(5), 1282–1290. (Cited on page 12.)

Saunders, R. and Deane, C. M. (2010b). Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.*, **38**(19), 6719–6728. (Cited on page 11.)

Saunders, R., Mann, M., and Deane, C. M. (2011). Signatures of co-translational folding. *Biotechnol. J.*, **6**(6), 742–751. (Cited on pages 76, 81, 82, 88, and 89.)

Schrödinger, L. (2015). The PyMOL Molecular Graphics System, Version 1.8. (Cited on pages 13 and 158.)

Seemayer, S., Gruber, M., and Soding, J. (2014). CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**(21), 3128–3130. (Cited on page 28.)

Seppälä, S., Slusky, J. S., Lloris-Garcerá, P., Rapp, M., and von Heijne, G. (2010). Control of membrane protein topology by a single C-terminal residue. *Science*, **328**(5986), 1698–700. (Cited on page 75.)

Shi, S., Pei, J., Sadreyev, R. I., Kinch, L. N., Majumdar, I., Tong, J., Cheng, H., Kim, B.-H., and Grishin, N. V. (2009). Analysis of CASP8 targets, predictions and assessment methods. *Database*, **2009**(0), bap003–bap003. (Cited on page 22.)

Simoncini, D. and Zhang, K. Y. J. (2013). Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm. *PLoS One*, **8**(7), e68954. (Cited on page 30.)

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. (Cited on page 25.)

Singh, H., Singh, S., Raghava, G. P. S., Raghava, G., and Zhou, Y. (2014). Evaluation of Protein Dihedral Angle Prediction Methods. *PLoS One*, **9**(8), e105667. (Cited on page 30.)

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7), 951–960. (Cited on page 151.)

Stansfeld, P. J., Goose, J. E., Caffrey, M., Carpenter, E. P., Parker, J. L., Newstead, S., and Sansom, M. S. P. (2015). MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure*, **23**(7), 1350–61. (Cited on pages 17, 84, and 173.)

Tate, C. G. (2012). A crystal clear solution for determining G-protein-coupled receptor structures. *Trends Biochem. Sci.*, **37**(9), 343–52. (Cited on page 14.)

Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P., and Mason, J. S. (2014). Unifying family A GPCR theories of activation. *Pharmacol. Ther.*, **143**(1), 51–60. (Cited on pages 18, 35, and 40.)

Teixeira, P. L., Mendenhall, J. L., Heinze, S., Weiner, B., Skwark, M. J., and Meiler, J. (2017). Membrane protein contact and structure prediction using co-evolution in conjunction with machine learning. *PLoS One*, **12**(5), e0177866. (Cited on pages 29, 109, 150, and 181.)

Tieleman, D. P., Shrivastava, I. H., Ulmschneider, M. R., and Sansom, M. S. (2001). Proline-induced hinges in transmembrane helices: possible roles in ion channel gating. *Proteins*, **44**(2), 63–72. (Cited on pages 18, 35, and 39.)

Traag, V. A., Van Dooren, P., and Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*, **84**(1), 016114. (Cited on page 45.)

Trovato, F. and O'Brien, E. P. (2016). Insights into Cotranslational Nascent Protein Behavior from Computer Simulations. *Annu. Rev. Biophys.*, **45**(1), 345–369. (Cited on page 12.)

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., and Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**(W1), W401–407. (Cited on pages 20, 149, and 172.)

Tsukazaki, T., Mori, H., Fukai, S., Ishitani, R., Mori, T., Dohmae, N., Perederina, A., Sugita, Y., Vassylyev, D. G., Ito, K., and Nureki, O. (2008). Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature*, **455**(7215), 988–91. (Cited on page 72.)

Tu, L., Khanna, P., and Deutsch, C. (2014). Transmembrane segments form tertiary hairpins in the folding vestibule of the ribosome. *J. Mol. Biol.*, **426**(1), 185–98. (Cited on page 75.)

Tyka, M. D., Keedy, D. A., André, I., DiMaio, F., Song, Y., Richardson, D. C., Richardson, J. S., and Baker, D. (2011). Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *J. Mol. Biol.*, **405**(2), 607–618. (Cited on page 84.)

Ulmschneider, M. B., Sansom, M. S. P., and Di Nola, A. (2005). Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins*, **59**(2), 252–65. (Cited on page 16.)

Ulmschneider, M. B., Koehler Leman, J., Fennell, H., and Beckstein, O. (2015). Peptide Folding in Translocon-Like Pores. *J. Membr. Biol.*, **248**(3), 407–17. (Cited on page 12.)

van der Kant, R. and Vriend, G. (2014). Alpha-bulges in G protein-coupled receptors. *Int. J. Mol. Sci.*, **15**(5), 7841–64. (Cited on page 40.)

van Meer, G., Voelker, D. R., and Feigenson, G. W. (2008). Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.*, **9**(2), 112–24. (Cited on page 13.)

Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013). Molecular signatures of G-protein-coupled receptors. *Nature*, **494**(7436), 185–94. (Cited on page 59.)

Viklund, H. and Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, **24**(15), 1662–8. (Cited on page 20.)

Viklund, H., Granseth, E., and Elofsson, A. (2006). Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J. Mol. Biol.*, **361**(3), 591–603. (Cited on page 15.)

Visiers, I., Braunheim, B. B., and Weinstein, H. (2000). Prokink: a protocol for numerical evaluation of helix distortions by proline. *Protein Eng. Des. Sel.*, **13**(9), 603–606. (Cited on pages 37 and 38.)

Voorhees, R. M., Fernández, I. S., Scheres, S. H. W., and Hegde, R. S. (2014). Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell*, **157**(7), 1632–43. (Cited on pages 71, 72, and 73.)

Walters, R. F. S. and DeGrado, W. F. (2006). Helix-packing motifs in membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **103**(37), 13658–63. (Cited on page 20.)

Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**(12), 1589–1591. (Cited on pages 43, 78, and 79.)

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.*, **13**(1), 1005324. (Cited on pages 28 and 29.)

Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G. W., Tate, C. G., and Schertler, G. F. X. (2008). Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature*, **454**(7203), 486–91. (Cited on page 14.)

Waudby, C. A., Launay, H., Cabrita, L. D., and Christodoulou, J. (2013). Protein folding on the ribosome studied using NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.*, **74**, 57–75. (Cited on page 11.)

Werner, T. and Church, W. B. (2013). Kink characterization and modeling in transmembrane protein structures. *J. Chem. Inf. Model.*, **53**(11), 2926–36. (Cited on pages 24, 68, and 152.)

White, S. H. and Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 319–65. (Cited on page 43.)

Wilman, H. R., Ebejer, J.-P., Shi, J., Deane, C. M., and Knapp, B. (2014a). Crowdsourcing Yields a New Standard for Kinks in Protein Helices. *J. Chem. Inf. Model.*, **54**(9), 2585–2593. (Cited on page 37.)

Wilman, H. R., Shi, J., and Deane, C. M. (2014b). Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins*, **82**(9), 1960–70. (Cited on pages 18, 36, 37, 38, 41, 42, 48, 51, and 67.)

Wimley, W. C., Creamer, T. P., and White, S. H. (1996). Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host-Guest Pentapeptides. *Biochemistry*, **35**(16), 5109–5124. (Cited on page 16.)

Xiao, F. and Shen, H.-B. (2015). Prediction Enhancement of Residue Real-Value Relative Accessible Surface Area in Transmembrane Helical Proteins by Solving the Output Preference Problem of Machine Learning-Based Predictors. *J. Chem. Inf. Model.*, **55**(11), 2464–2474. (Cited on page 149.)

Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma.*, **80**(7), n/a–n/a. (Cited on page 25.)

Xu, F., Wu, H., Katritch, V., Han, G. W., Jacobson, K. A., Gao, Z.-G., Cherezov, V., and Stevens, R. C. (2011). Structure of an agonist-bound human A2A adenosine receptor. *Science*, **332**(6027), 322–7. (Cited on page 68.)

Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**(7), 889–95. (Cited on page 88.)

Yang, J., Jang, R., Zhang, Y., and Shen, H.-B. (2013). High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics*, **29**(20), 2579–87. (Cited on page 29.)

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**(1), 7–8. (Cited on pages 25 and 182.)

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, **62**(4), 1010–25. (Cited on pages 26 and 100.)

Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P., and Bowie, J. U. (2004a). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(4), 959–963. (Cited on page 38.)

Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P., and Bowie, J. U. (2004b). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.*, **101**(4), 959–63. (Cited on page 52.)

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**(13), 3370–3374. (Cited on page 22.)

Zhang, H., Huang, Q., Bei, Z., Wei, Y., and Floudas, C. A. (2016). COMSAT: Residue contact prediction of transmembrane proteins based on support vector machines and mixed integer linear programming. *Proteins Struct. Funct. Bioinforma.*, **84**(3), n/a–n/a. (Cited on page 149.)

Zhang, L., Sato, Y., Hessa, T., von Heijne, G., Lee, J.-K., Kodama, I., Sakaguchi, M., and Uozumi, N. (2007). Contribution of hydrophobic and electrostatic interactions to the membrane integration of the Shaker K+ channel voltage sensor domain. *Proc. Natl. Acad. Sci. U. S. A.*, **104**(20), 8263–8. (Cited on page 74.)

Zhang, S.-Q., Kulp, D. W., Schramm, C. A., Mravic, M., Samish, I., and DeGrado, W. F. (2015). The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions. *Structure*, **23**(3), 527–41. (Cited on pages 20, 86, 87, and 97.)

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4), 702–710. (Cited on page 22.)

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**(7), 2302–9. (Cited on pages 22 and 44.)

Appendices

A

Estimation of error in kink angle measurement

A method was developed by Henry Wilman that estimates the error in each angle measured by Kink Finder. This heuristic method is based on observed errors in a set of 'ideal' kinks with good cylinder fits.

Kink Finder fits a cylinder to every 6-residue segment of a helix by minimising r, where

$$r = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(d_i - \bar{d}\right)^2} \tag{A.1}$$

m is the number of backbone atoms in the segment (24), d_i is the shortest distance from backbone atom i to the fitted helix axis, and $\bar{d}$ is the mean of all distances. The angle is measured between the axes of the cylinders fitted to adjacent segments, and this angle is assigned to the final residue of the first segment.

Each cylinder fit has a value of r (Equation A.1), which measures the distance of the atoms from the cylinder surface. The calculation of a kink angle requires two cylinder fits, one for the set of six residues N-terminal of the kink position, and one for the set of six residues C-terminal to the kink (see Figure A.1). The goodness of these two fits (r_n and r_c, the r for the N- and C-terminal cylinder fits) are assumed to have an equal effect on the angle error. For each measured

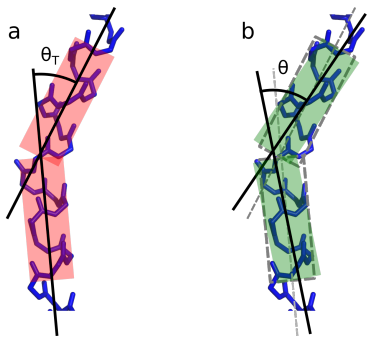


Figure A.1: Relating angle error to goodness of fit. (a) Example 'ideal' kink, with low r_n and r_c. The true angle (θ_T) is the angle between the two fitted axes. (b) Cylinders are rotated (in green) from their fitted positions (dashed lines), and a measured angle (θ) is calculated. r_n and r_c are calculated from the rotated cylinders (green). Carrying out this rotation many times provides the data for Figure A.2.

angle, θ, the 'goodness of fit' is approximated by the sum of the r of the two cylinder fits. Although there is no way to directly measure the 'true' angle, 18 of the best fitted ('ideal') kinks were used to estimate the effect, assuming that the fitted axes for these provided the 'true' angle. These 18 helices (with the lowest $r_n + r_c$, of the kinks in the membrane protein set) have a range of true angles between 0° and 50°. The $r_n + r_c$ for all of these 18 is below 0.6 Å. Even for an ideal helix, $r_n + r_c$ cannot be less than 0.27 Å.

Taking these 'ideal' kinks, we simulated the relationship between $r_n + r_c$, α (the error), and the true angle. For each measurement in each kink, both cylinders were rotated about their midpoint by an angle and direction, using a randomly generated rotation matrix (Figure A.1). This provided a series of measured angles, θ, based on non-optimised cylinder fits. The $r_n + r_c$ and θ were recorded for each, and used to characterise the relationship between the two (Figure A.2).

For a given range of $r_n + r_c$, α (and similarly, θ), has a distribution that is close to normal. To find out how this distribution is related to θ, we assumed that it is normally distributed with mean zero and variance σ_α^2:

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

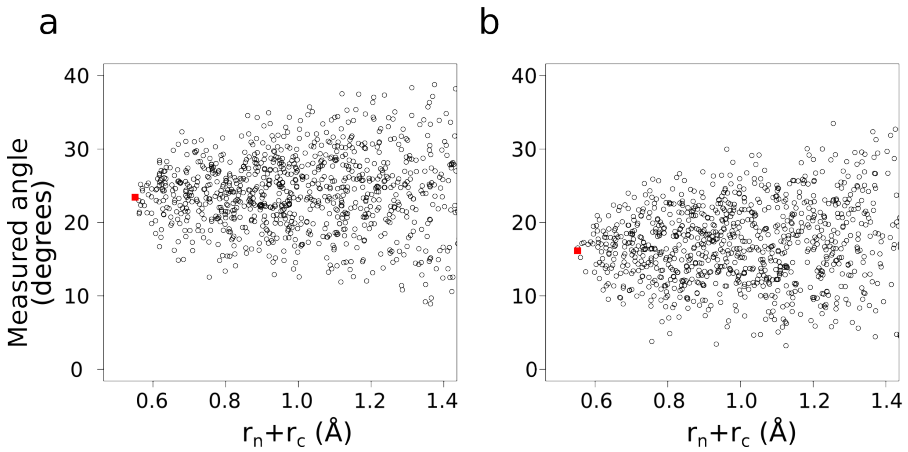


Figure A.2: Measured angle against goodness of fit ($r_n + r_c$) for two kinks. (a) At residue 255 in chain A of protein 1PB2. (b) At residue 259 in chain A of protein 1Y2L. The red squares indicate the angle and $r_n + r_c$ for the optimum cylinder fits.

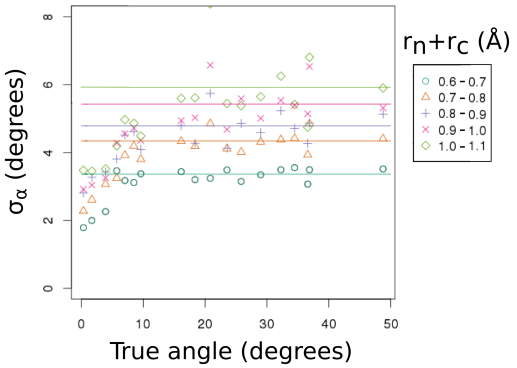


Figure A.3: The standard deviation, σ_α, of α (measured angle - true angle) for bins of $r_n + r_c$ (y-axis) are shown for 18 ideal kinks, plotted against their true angle as determined by the optimised cylinder fits. The standard deviation of α for a given range of $r_n + r_c$ is constant for angles above 10°. Horizontal lines are fitted to the points where true angle $> 10°$ for each range of $r_n + r_c$.

The data was binned based on $r_n + r_c$, and we made this assumption for each bin. In each bin for each kink, σ_α was calculated. Excluding kinks under 10°, the size of α can be considered to depend only on the value of $r_n + r_c$ (Figure A.3). Therefore, the errors for all of the kinks with angles above 10° were combined to calculate the relationship between $r_n + r_c$ and angle error.

From this point, a statistical confidence interval was used, rather than using the standard deviation, as this does not rely on the assumption of normality.

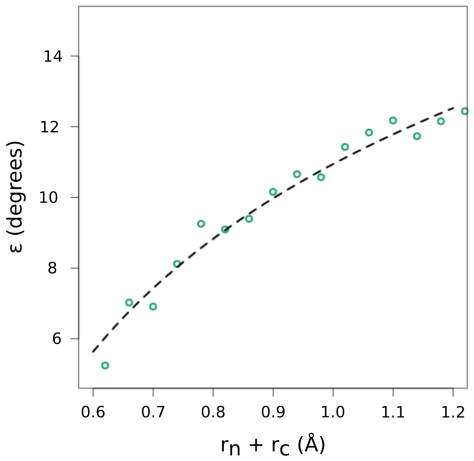


Figure A.4: The error, ε, for a range of values of $r_n + r_c$ (quality of fit), where ε represents the size of the 95% confidence interval of angle error. For the combined data from all 12 kinks with angles $\geq 12°$, the angle errors are binned by their $r_n + r_c$ values. The value at the 95^{th} percentile of $|\alpha|$ (where α is measured angle - true angle) is taken as the value of ε for each $r_n + r_c$ bin (green points). A log plot is fitted to the values between 0.6 and 1.0 (dashed black line).

From the 12 kinks over 10°, all values of α were binned into ranges of $r_n + r_c$. The distribution is symmetric, and is assumed to be so in the rest of this work. The 95th percentile of $|\alpha|$ for each bin was taken to give ε, deriving the size of the statistical confidence interval:

$$95\% \text{ confidence interval for true angle} = \theta \pm \varepsilon \tag{A.3}$$

where θ is the measured angle, and we will refer to ε as the error. We use a log fit to quantify the relationship between $r_n + r_c$ and ε (Figure A.4), which gives:

$$\varepsilon = (6.349 \times \ln(r_n + r_c - 0.2937) + 13.15) \tag{A.4}$$

Thus ε is an estimate of the uncertainty in the kink angles measured by Kink Finder and provides a simple way to compare the angles in two helices. This version of Kink Finder is available online at http://www.stats.ox.ac.uk/research/proteins/resources.

B

Tables of proline in kink pairs

Non-redundant datasets, resolution < 5 Å and R < 0.4

	Membrane					Soluble				
	CK	CS	NC	other	total	CK	CS	NC	other	total
All	1189	1806	789	320	4104	40190	481388	88390	19556	629524
%	29.0	44.0	19.2	7.8	100.0	6.4	76.5	14.0	3.1	100.0
a) PP	592	36	78	74	780	8457	157	2351	486	11451
P-	167	112	315	46	640	5642	3180	35662	2605	47089
-P	179	10	52	32	273	5040	390	1930	285	7645
- -	251	1648	344	168	2411	21051	477661	48447	16180	563339
b) PP	49.8	2.0	9.9	23.1		21.0	0.0	2.7	2.5	
P-	14.0	6.2	39.9	14.4		14.0	0.7	40.3	13.3	
-P	15.1	0.6	6.6	10.0		12.5	0.1	2.2	1.5	
- -	21.1	91.3	43.6	52.5		52.4	99.2	54.8	82.7	
c) PP	14.4	0.9	1.9	1.8	19.0	1.3	0.0	0.4	0.1	1.8
P-	4.1	2.7	7.7	1.1	15.6	0.9	0.5	5.7	0.4	7.5
-P	4.4	0.2	1.3	0.8	6.7	0.8	0.1	0.3	0.0	1.2
- -	6.1	40.2	8.4	4.1	58.7	3.3	75.9	7.7	2.6	89.5

Table B.1: The number of aligned helix pairs in each class, and occurrence of proline at the position with the largest angle or in the four following residues. The helix pair classes CK, CS, NC and other are defined in the Results. PP: proline in both helices; P-: proline in the helix with the larger kink angle; -P: proline in the helix with the smaller kink angle; - -: proline in neither helix. a) the frequency of each type in each class b) the frequency of each type as a percentage of the pairs in that class c) the frequency of each type as a percentage of the total number of pairs.

High quality non-redundant dataset, resolution < 2 Å and R < 0.2

	Soluble				
	CK	CS	NC	other	total
All	12622	136141	27894	5972	182629
%	6.9	74.5	15.3	3.3	100.0
a) PP	2647	36	734	112	3529
P-	1685	1156	11258	796	14895
-P	1483	94	612	81	2270
- -	6807	134855	15290	4983	161935
b) PP	21.0	0.0	2.6	1.9	
P-	13.3	0.8	40.4	13.3	
-P	11.7	0.1	2.2	1.4	
- -	53.9	99.1	54.8	83.4	
c) PP	1.4	0.0	0.4	0.1	1.9
P-	0.9	0.6	6.2	0.4	8.2
-P	0.8	0.1	0.3	0.0	1.2
- -	3.7	73.8	8.4	2.7	88.7

Table B.2: The number of aligned helix pairs in each class, and occurrence of proline at the position with the largest angle or in the four following residues. The helix pair classes CK, CS, NC and other are defined in the Results. PP: proline in both helices; P-: proline in the helix with the larger kink angle; -P: proline in the helix with the smaller kink angle; - -: proline in neither helix. a) the frequency of each type in each class b) the frequency of each type as a percentage of the pairs in that class c) the frequency of each type as a percentage of the total number of pairs.

C Membrane protein datasets

C.1 PDB codes in Set 1 only, total 72

PDB code and chain identifier	Length (residues)	Experimental method	Resolution (Å)	R-factor	Free R-factor
1C17M	177	NMR			
1KF6D	119	XRAY	2.700	0.23	0.28
1M0KA	262	XRAY	1.430	0.13	0.18
1PW4A	451	XRAY	3.300	0.30	0.33
1RZHM	307	XRAY	1.800	0.23	0.23
2A65A	519	XRAY	1.650	0.20	0.22
2BL2A	156	XRAY	2.100	0.19	0.20
2BS2C	256	XRAY	1.780	0.23	0.24
2CFQA	417	XRAY	2.950	0.26	0.30
2KSFA	107	NMR			
2LOSA	121	NMR			
2MGYA	169	NMR			
2NPKA	424	XRAY	2.000	0.17	0.19
2NQ2A	337	XRAY	2.400	0.22	0.26
2NWLA	422	XRAY	2.960	0.24	0.27
2R9RB	514	XRAY	2.400	0.21	0.24
2WSSW	190	XRAY	3.200	0.22	0.27
2XOVA	181	XRAY	1.650	0.19	0.22
2ZUPA	189	XRAY	3.700	0.30	0.33
2ZUQA	176	XRAY	3.300	0.28	0.35
2ZW3A	226	XRAY	3.500	0.34	0.35

2ZY9A	473	XRAY	2.940	0.26	0.29
3AG3A	514	XRAY	1.800	0.17	0.20
3AG3C	261	XRAY	1.800	0.17	0.20
3AONA	217	XRAY	2.000	0.20	0.25
3AQPA	741	XRAY	3.300	0.30	0.32
3AR4A	995	XRAY	2.150	0.24	0.28
3B4RA	224	XRAY	3.300	0.25	0.32
3B9WA	407	XRAY	1.300	0.15	0.17
3D31C	295	XRAY	3.000	0.25	0.28
3DH4A	530	XRAY	2.700	0.27	0.29
3EGWC	225	XRAY	1.900	0.17	0.20
3HZQA	114	XRAY	3.820	0.29	0.31
3L1LA	445	XRAY	3.000	0.22	0.28
3LEOA	155	XRAY	2.100	0.17	0.20
3LW54	166	XRAY	3.300	0.35	0.38
3M76A	314	XRAY	1.500	0.13	0.15
3MP7A	482	XRAY	2.900	0.28	0.32
3RQWA	322	XRAY	2.910	0.21	0.23
3SOXB	237	XRAY	3.600	0.28	0.33
3TDSA	268	XRAY	1.980	0.18	0.20
3UDCA	285	XRAY	3.350	0.25	0.27
3UX4A	201	XRAY	3.260	0.24	0.30
3WFDB	465	XRAY	2.300	0.19	0.23
3ZE3A	130	XRAY	2.050	0.20	0.22
3ZOJA	279	XRAY	0.880	0.10	0.11
4AL0A	152	XRAY	1.160	0.12	0.13
4BBJA	736	XRAY	2.750	0.20	0.25
4BEMJ	182	XRAY	2.100	0.18	0.22
4DVEA	198	XRAY	2.090	0.19	0.20
4ENEA	446	XRAY	2.400	0.24	0.28
4G7VS	185	XRAY	2.500	0.20	0.24
4GD3A	235	XRAY	3.300	0.20	0.24
4HE8D	176	XRAY	3.300	0.21	0.26
4HFIA	317	XRAY	2.400	0.20	0.22
4HKRA	214	XRAY	3.350	0.28	0.28
4J05A	530	XRAY	2.900	0.22	0.26
4JR9A	466	XRAY	2.600	0.23	0.26
4M64A	486	XRAY	3.350	0.31	0.36
4MESA	182	XRAY	2.000	0.18	0.21
4MS2A	237	XRAY	2.750	0.23	0.26
4NV5A	291	XRAY	2.790	0.23	0.24
4O9PB	283	XRAY	2.890	0.23	0.29

4OH3A	599	XRAY	3.250	0.24	0.31
4OO9A	444	XRAY	2.600	0.24	0.28
4P79A	198	XRAY	2.400	0.22	0.25
4PGRA	217	XRAY	1.950	0.21	0.23
4PHZB	252	XRAY	2.590	0.24	0.29
4PHZC	256	XRAY	2.590	0.24	0.29
4PIRA	456	XRAY	3.500	0.22	0.26
4TQ3A	303	XRAY	2.410	0.22	0.26
4WGVA	415	XRAY	3.100	0.25	0.29

C.2 PDB codes in both Set 1 and Set 2, total 21

An "A" next to a PDB code indicates that the chain is in Set2A, while a "T" indicates it is part of the test set of 9 chains.

PDB code and chain identifier	Length (residues)	Experimental method	Resolution (Å)	R-factor	Free R-factor
1KQFC A	217	XRAY	1.600	0.18	0.20
1OKCA A	297	XRAY	2.200	0.22	0.27
1ORSC A	132	XRAY	1.900	0.23	0.25
1ZCDA A	388	XRAY	3.450	0.30	0.32
2JLNA	501	XRAY	2.850	0.24	0.28
2WSWA T	509	XRAY	2.290	0.21	0.23
3CX5C A	385	XRAY	1.900	0.24	0.26
3GIAA	444	XRAY	2.320	0.20	0.23
3H90A	283	XRAY	2.900	0.26	0.28
3O7QA	438	XRAY	3.140	0.22	0.27
3WDOA	453	XRAY	3.150	0.27	0.29
4A2NB A	194	XRAY	3.400	0.24	0.28
4B4AA A	249	XRAY	3.500	0.25	0.29
4BWZA A	394	XRAY	2.980	0.22	0.25
4GC0A	491	XRAY	2.600	0.23	0.25
4HUQS A	174	XRAY	3.000	0.22	0.26
4IKVA T	507	XRAY	1.900	0.18	0.22
4MRSA	614	XRAY	2.350	0.20	0.22
4N6HA	414	XRAY	1.800	0.17	0.19
4N7WA A	307	XRAY	1.950	0.19	0.22
4O6YA A	230	XRAY	1.700	0.20	0.22

C.3 PDB codes in Set 2 but not in Set1, total 34

An "A" next to a PDB code indicates that the chain is in Set2A, while a "T" indicates it is part of the test set of 9 chains.

PDB code and chain identifier	Length (residues)	Experimental method	Resolution (Å)	R-factor	Free R-factor
1E12A A T	253	XRAY	1.800	0.24	0.26
1JB0B	740	XRAY	2.500	0.20	0.22
1OTSA	465	XRAY	2.510	0.26	0.30
1PV6A	417	XRAY	3.500	0.29	0.34
1U7GA A T	385	XRAY	1.400	0.14	0.17
1YEWB	247	XRAY	2.800	0.27	0.30
2DYRA	514	XRAY	1.800	0.20	0.23
2DYRC A T	261	XRAY	1.800	0.20	0.23
2J8CM	307	XRAY	1.870	0.18	0.20
2QI9A A	326	XRAY	2.600	0.26	0.28
2VPZC A	253	XRAY	2.400	0.25	0.25
2W2EA A	279	XRAY	1.150	0.14	0.17
2XOWA A	179	XRAY	2.090	0.20	0.24
2YVXA	473	XRAY	3.500	0.29	0.34
3B4RB A	224	XRAY	3.300	0.25	0.32
3KLYA A	280	XRAY	2.100	0.18	0.21
3KZIB	510	XRAY	3.600	0.30	0.31
3KZIC	461	XRAY	3.600	0.30	0.31
3M73A A T	314	XRAY	1.150	0.14	0.15
3N5KA	994	XRAY	2.200	0.19	0.22
3PUWG	296	XRAY	2.300	0.22	0.26
3QE7A	429	XRAY	2.780	0.25	0.30
3QNQA	442	XRAY	3.290	0.23	0.26
3RKOJ	184	XRAY	3.000	0.23	0.28
3RKOL	613	XRAY	3.000	0.23	0.28
3RLBA A T	192	XRAY	2.000	0.21	0.23
3TLWA	321	XRAY	2.600	0.21	0.23
4DX5A	1057	XRAY	1.900	0.20	0.23
4EZCA A	384	XRAY	2.360	0.20	0.23
4KPPA T	405	XRAY	2.300	0.20	0.24
4KYOC T	431	XRAY	3.000	0.21	0.27
4M48A	543	XRAY	2.960	0.22	0.26
4MLBC	492	XRAY	2.350	0.20	0.23
4OD5A A	303	XRAY	3.560	0.28	0.30