Characterization of Physicochemical Environments of Proteins

TAN Kuan Pern

Thesis Supervisors:

Dr. VERMA Chandra, Dr. M. S. Madhusudhan, Dr. Kwoh Chee Keong

School of Computer Science & Engineering

A thesis submitted to Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

January 12, 2017

Acknowledgments

Foremost, I would like to thank my supervisor and co-supervisors, Dr. Chandra Verma, Dr. M. S. Madhusudhan and Dr. Kwoh Chee Keong for their invaluable guidance, support and suggestions. Their knowledge, suggestions, and discussions helped me to become a capable researcher. Especially for Dr. Madhusudhan who has guided me over the past six years, and helped me in developing essential skills of scientific research methods, oral presentations and technical writings.

The thesis would not have been possible without the contributions of the collaborators. I express my utmost thanks to Dr. Raghavan Varadarajan from MBU IISc (India) and his students Patel Siddarth, Shruti Khare for our collaboration work. Especially Dr. Raghavan who has also provided deep and valuable insights into many other topics of my research. I also express my gratitude toward Dorothy Su from Yong Loo Lin School of Medicine, NUS and her co-workers for all the efforts on our joint projects.

Secondly, I would like to acknowledge my teammates in Bioinformatics Institute A* Singapore, Minh Nguyen and Binh Nguyen, and other team members in IISER Pune, Neelesh Soni, Yogendra Ramtirtha, Shipra Gupta and many others for their help and helpful discussions on various projects. I also thank Dr. Mu Yuguang from SBS NTU and Dr. Vladimir Kuznetsov for their valuable suggestions and advices in thesis advisory committee meetings.

I thank the moral support and company of my friends from SBS, especially who have been on the same Ph.D journey with me, my lifelong friend Ng Chee Sheng, Chen Ming Wei, Grace Ngan Jie Ying, Fann Rui Yang, James Ho and See Rui Yin. I also want to thank my friends from SCE, Lee Zhuo Qi and Jaydeep De for their suggestions and guidance through various procedures during the journey.

On research support, I am very grateful to the director Dr. Frank Eisenhaber and Miss Betty Kee from BII for their countless support on many different aspects in my

pursuit of the doctorate degree. I would also like to thank Yong Tai Pang (BII) and Parichit Sharma (IISER) for IT support for the many years; Violet Lin, Bharat Adkar and Anusmita Sahoo for servers design and testing. I also thank the agencies of Bioinformatics Institute A*STAR Singapore, Nanyang Technological University Singapore, Biomedical Research Council (A*STAR) Singapore, Council of Scientific and Industrial Research, Department of Biotechnology and Department of Science and Technology, Government of India for providing the research environment and fundings.

Last but not least, I would like to thank many unnamed researchers who have established and constantly advancing our knowledge frontier, as well as my family for their constant love and support.

Contents

Acknowledgments	i
List of Figures	viii
List of Tables	xi
List of Abbreviations	xiii
List of Notation	xiii
Abstract	xiv
1 Introduction	2
1.1 Background and Motivations	2
1.2 Research Issues	3
1.2.1 Residue Depth and Its Utilities	4
1.2.2 Protein Intramolecular Hydrogen Bonds	4
1.2.3 Protein Multibody Statistical Potential	5
1.2.4 Other Molecular System Modelling Projects	5
1.3 Approaches	5
1.4 Synopsis	7
2 Preliminaries	12
2.1 Introduction	12
2.2 Protein Sequence Analysis	13
2.2.1 Sequence alignment	13
2.2.2 BLAST and PSI-BLAST	14
2.3 Homology modelling	16
2.4 Protein Statistical Potential	17
2.4.1 Physics-based and knowledge-based potential	17

2.4.2	Higher order statistical potential	18
2.5	Solvent Accessible Surface Area	19
2.5.1	Golden spiral approximation	20
2.6	Cell-list Algorithm	23
2.7	Molecular Dynamics	23
2.8	Ramachandran Map	25
2.9	Benchmark Measures	26
2.9.1	Pearson’s correlation coefficient	26
2.9.2	Matthews Correlation Coefficient	27
2.9.3	Spearman’s Rank Correlation Coefficient	27
3	Characterization of Environments in the Protein Interior	28
3.1	Background and Motivations	29
3.1.1	Importance of the protein solvation description	29
3.1.2	Reviews of previous work	30
3.2	Definition and Computation of Depth	30
3.2.1	Removal of non-bulk water	30
3.2.2	Consideration of bulk water dynamics	31
3.3	Properties and distribution of Depth	34
3.3.1	Comparison between depth and SASA	34
3.3.2	Residue depth preference of amino acids	35
3.4	Utilities of Residue Depth	36
3.4.1	Background and Motivations	36
3.4.2	Temperature-sensitive mutant design	37
3.4.3	Protein ligand binding site prediction	48
3.4.4	Other Utilities of Depth Measure	61
3.5	Chapter Summary & Discussions	63
4	Characterization of Intramolecular Hydrogen Bonding in Proteins	66
4.1	Background and Motivations	67
4.1.1	Properties and importance of hydrogen bonding in proteins	67
4.1.2	Previous work	67

4.2	Geometry of Protein Hydrogen Bonds	68
4.2.1	Data sets and parameters	68
4.2.2	Overall distribution of donor-acceptor distance and $\hat{H}$ angle . . .	69
4.2.3	Classification of hydrogen bonds and environments	71
4.2.4	Hydrogen bond geometry in different protein environments	72
4.3	Structural and Functional Implications Hydrogen Bond Geometry	76
4.3.1	Secondary structure formation	76
4.3.2	Donor-acceptor distance in secondary structures	80
4.3.3	Correlation with partial charge	81
4.4	Explanation for hydrogen bond geometry preferences	82
4.4.1	$\hat{H}$ angle distribution in small organic molecules	83
4.4.2	Quantum simulation	83
4.5	Hydrogen bond number in protein models	88
4.5.1	Scalability of hydrogen bond number with chain length	88
4.5.2	Hydrogen bond number in low-resolution X-ray structures	89
4.5.3	Hydrogen bond number in NMR models	90
4.6	Protein 3D Model Refinement with Hydrogen Bonding Restraints	92
4.6.1	Importance of high-resolution protein 3D model	92
4.6.2	Review of previous work	93
4.7	Refinement protocol	96
4.7.1	Hydrogen bond prediction and selection	96
4.7.2	Hydrogen bond restraints	97
4.8	Refinement performance	99
4.8.1	Refinement of low-resolution X-ray structure	99
4.8.2	Refinement of NMR structures	102
4.8.3	Utilities of hydrogen bond refinement protocol	106
4.9	Chapter Summary and Discussion	112

5	Amino Acid Cliques Multibody Statistical Potential	117
5.1	Background and Motivations	117
5.1.1	Reviews of mutational effect prediction methods	119
5.1.2	Residue environment	120
5.2	Cliques of amino acid residues	120
5.2.1	Computation of residue cliques	121
5.2.2	Derivation of <i>Packpred</i> statistical potential	121
5.2.3	Training of the potential	123
5.3	Validation of <i>Packpred</i> statistical potential	124
5.3.1	Recognition of native structure among structural decoys	124
5.3.2	Correlation with structural stability	124
5.4	Web-Server	130
5.5	Chapter Summary	130
6	Discussions and Conclusions	132
6.1	Residue Depth	132
6.1.1	Summary	132
6.1.2	Suggestions and future directions	133
6.2	Hydrogen Bonding in Proteins	134
6.2.1	Summary	134
6.2.2	Discussions and future directions	135
6.3	Multibody Statistical Potential <i>Packpred</i>	136
6.3.1	Summary	136
6.3.2	Suggestion for improvements	137
6.4	Concluding Remarks	138
7	Appendix: Modelling Non-Globular Protein Systems	141
7.1	Modelling of Cohesin Complex	142
7.1.1	Background and motivations	142
7.1.2	Modelling protocol	143
7.1.3	Prediction of critical lysines	147
7.1.4	Summary and conclusions	149

7.2	Modelling of OCTN Transporters	149
7.2.1	Background and motivations	149
7.2.2	Modelling protocol	152
7.2.3	Model interpretation and mechanism hypothesis	153
7.2.4	Validation of mechanism hypothesis	155
7.2.5	Summary and conclusions	156
7.3	Modelling of EZH2-VAV1 Interaction	157
7.3.1	Background and motivations	157
7.3.2	Modelling protocol	158
7.3.3	Design of inhibitory peptides	162
7.3.4	Summary and conclusions	164
7.4	Conclusions	166
	Publication	168
	Bibliography	170
	A Appendix	196
A.1	Data sets	196
A.1.1	Ramachandran map data-set	196
A.1.2	Depth and accessibility data-set	200
A.1.3	Depth binding propensity data-set	205
A.1.4	High-resolution single domain data set	208
A.1.5	Small organic molecules in Open Crystallography Database	210
A.1.6	Relaxed resolution single domain data set	210
A.1.7	Dataset of non-redundant crystal structures	223
A.1.8	Dataset of redundant crystal structures	241
A.1.9	Dataset of NMR structures	241
A.1.10	X-ray density maps data set	243
A.1.11	<i>Packpred</i> potential training set	243

List of Figures

2.1	Illustration of Shrake-Rupley algorithm	20
2.2	Illustrations of Golden Spiral approximation	21
2.3	Illustration of peptide backbone	26
2.4	Ramachandran maps	26
3.1	Degree of solvation in protein environments	30
3.2	Schematic diagram to illustrate detection of non-bulk water in protein. .	31
3.3	Comparison between residue depth and solvent inaccessibility	35
3.4	Correlation of depth and accessibility with H/D exchange rate	35
3.5	Depth distribution of different amino acids	36
3.6	Flowchart of TSpred algorithm	38
3.7	Temperature sensitive mutation prediction on CcdB	42
3.8	Illustration of different pocket identification methods	49
3.9	Flowchart of metaPocket 2.0	51
3.10	Principle of DEPTH residue binding site prediction	52
3.11	Binding site propensity of Cystein	53
3.12	Schematic diagram on detection of cavity water	56
3.13	Comparison between DEPTH and ConCavity	58
3.14	Binding site detection on West Nile Virus NS2B/NS3 protease	60
3.15	Correlation between experimentally measured free energy change on mu- tation and cavity sizes	62
4.1	Illustrative figure of a hydrogen bond	69
4.2	Geometry distribution of hydrogen bond in protein	70
4.3	Geometry distribution of hydrogen bond of different types	73

4.4	Hydrogen bond geometry in different secondary structures	75
4.5	Dihedral angles of simulated α -helices	77
4.6	$\hat{H}$ angle and donor-acceptor distance of α -helix in MD simulation	79
4.7	Dihedral angles of simulated α -helices	79
4.8	Allowed geometrical distribution of hydrogen bond on a plane	80
4.9	MC-MC hydrogen bond length of different amino acids	80
4.10	(A) Melting temperature of all- α and all- β proteins	81
4.11	Correlation between donor-acceptor distances and partial charges	82
4.12	Distribution of $\hat{H}$ angle in small organic molecules	83
4.13	Optimized structure of MC-MC and SC-MC mimetics	85
4.14	Optimized structure of glycine dipeptides	86
4.15	(Top) Electron localization functions of MC-MC (left) and SC-MC (right) near the hydrogen bond (represented in orange dotted line) (Bottom) Schematic representation of electron localization in MC-MC and SC-MC hydrogen bond	87
4.16	Linear scalability of hydrogen bond number and chain length	89
4.17	Number of hydrogen bond at different structure resolutions	90
4.18	Distribution of correlation coefficient values of hydrogen bond numbers and structural resolution	91
4.19	Comparison of hydrogen bond number in X-ray and NMR structures . .	91
4.20	Consensus of hydrogen bonds in NMR models	92
4.21	Accuracy and application of protein structure models	93
4.22	Previous hydrogen bonding potential	94
4.23	Flowchart of hydrogen bond structure refinement	96
4.24	Degree of improvement in electron density map fitting of hydrogen bond refined models	100
4.25	Example of hydrogen bond refinement on human antibody Fc fragment .	100
4.26	Comparison of hydrogen bond number in low-resolution, refined, and high- resolution models	101
4.27	Distribution of recovered hydrogen bonds in NMR models at different depths	102

4.28	Improvement in structural accuracy of NMR models by hydrogen bond refinement	105
4.29	Distribution of recovered hydrogen bonds in X-ray models at different depths	105
4.30	Improvement in structural accuracy of X-ray models by hydrogen bond refinement	106
4.31	Example of recovered hydrogen bond in Hen Egg White Lysozyme	108
4.32	Coincidence between local structure improvement and hydrogen bond recovery	110
4.33	Examples of recovered hydrogen bonds Hen Egg White Lysozyme and Ribonuclease T1	112
4.34	Comparison of $\hat{H}$ -angle in MC-MC to SC-MC hydrogen bond	113
5.1	Illustration of a amino acid residue clique	121
5.2	Snapshots of <i>Packpred</i> web-server	130
7.1	Schematic representation of coiled-coil interaction	144
7.2	Paircoil2 prediction of coiled-coil structure in SMC1 and SMC3.	144
7.3	Bissulfosuccinimidyl suberate cross-linker	145
7.4	3D model of yeast SMC1 and SMC3 proteins	146
7.5	Electrostatics of SMC1-SMC3 complex model	146
7.6	Micrograph analysis on coiled-coil segment lengths	147
7.7	Estimated length for coiled-coil region of SMC1 and SMC3	147
7.8	Homology modelling of OCTNs	154
7.9	Rate of substrate uptake by OCTNs and mutants	154
7.10	Comparison of surface and total expression level of OCTNs	156
7.11	Predicted helices from EZH2 interface	158
7.12	3D model of VAV1 protein	159
7.13	Superposition of Calponin Homology domains and its properties	160
7.14	Sequence alignment between structural templates and predicted EZH2 interacting helix	160
7.15	Comparative model of the VAV1-EZH2 interaction	161
7.16	Pull-down assays with EZH2 mutants	162

7.17 VAV1-EZH2 interface and its designed inhibitor	163
---	-----

List of Tables

3.1	Sequence based rules to predict temperature sensitive mutation positions	41
3.2	Prediction benchmark of TSpred method	44
3.3	TSpred prediction on T4-lysozyme	45
3.4	Performance benchmark of DEPTH binding site prediction vs other methods	57
3.5	Performance benchmark binding site prediction for DEPTH, ConCavity and MetaPocket 2.0.	59
3.6	Performance benchmark of DEPTH pK_a prediction vs other methods . .	63
4.1	Donor, acceptor and acceptor antecedent atoms	69
4.2	List of hydrogen bonds with $\hat{H}$ angle less than 120° in Hen Egg White Lysozyme	72
4.3	Statistical significance of geometry difference of different hydrogen bonds	74
4.4	Statistical significance of geometry difference of different hydrogen bonds	74
4.5	Fisher's exact test on melting temperature of all- α / all- β protein	81
4.6	$\hat{H}$ angles of different types of hydrogen bonds	85
4.7	Chemical properties of hydrogen bonds of simulated Glycine dimers. . . .	86
4.8	Hydrogen bond parameters for MC-MC in α -helix, β -sheet, others	94
4.9	Hydrogen bond parameters for SC-MC and SC-SC	95
4.10	Sequence separation modifier at different sequence separation	97
4.11	Improvement in structure quality, hydrogen bond fulfilment ratio and hy- drogen bond number refinement models	104
4.12	Experimentally measured pK_a and predictions using X-ray, NMR, and refined models	108
4.13	Refinement statistics on HSPC034 protein model refinement	109

4.14	Refinement statistics on Hen Egg White Lysozyme protein and Ribonuclease T1.	111
5.1	Ranking of native structure by <i>Packpred</i> potentials on Melo's decoy set .	124
5.2	Performance of binary classification of destabilizing mutations by <i>Packpred</i> and other methods	128
5.3	Performance of rank-ordering of degree of phenotypical change of <i>Packpred</i> and other methods	129
7.1	Lysine modification sites in human SMC1A and SMC3 collected from literatures	148
7.2	Estimated electrostatic potential on $C\beta$ atom of Lysine residues of SMC1	150
7.3	Estimated electrostatic potential on $C\beta$ atom of Lysine residues of SMC3	151
7.4	Predicted binding affinity of alanine scanning mutagenesis on VAV1 and EZH2.	161
7.5	VAV1-EZH2 Inhibiting peptide designs	165

List of Abbreviation and Notation

PDB	Protein Data Bank
R.M.S.D.	Root Mean Squared Deviation
MC	Main chain
SC	Side chain
SNP	single nucleotide polymorphism
BLAST	Basic Local Alignment Search Tool
PSSM	position specific substitution matrix
PPI	Protein-protein interface
TM	Transmembrane
$\hat{H}$	hydrogen bond donor - acceptor - antecedent angle
SASA	Solvent Accessible Surface Area
pK_a	acid dissociation constant at (negative) logarithmic scale of base 10
pH	hydrogen ions concentration at (negative) logarithmic scale of base 10
NMR	Nuclear magnetic resonance
SAXS	small-angle X-ray scattering
cryo-EM	Cryo-electron microscopy
SPR	Surface plasmon resonance
R_{free}	cross-validated value of crystallographic R-factor
MD	molecular dynamics
MP2	Mller-Plesset perturbation theory
DFT	density functional theory
NMA	N-methylacetamide
ATP	Adenosine triphosphate
TEA	tetraethylammonium
OCTNs	Organic cation transport proteins
HEWL	Hen Egg White Lysozyme
MCC	Matthews correlation coefficient
SCC	Spearman rank correlation coefficient
HMM	Hidden Markov model
$J = S $	Cardinality of a set S
$A \cap B$	Intersection of set A and set B
$A \cup B$	Union of set A and set B
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2

Abstract

Proteins are molecular machines in cells that perform a diverse set of essential biological functions. The functions of proteins are determined by its 3D structure. The structure creates local microenvironments for protein atoms to interact with one another. A detailed understanding of these microenvironments would allow better characterization and engineering of protein functions. For example, this knowledge forms the basis of modern therapeutics innovations such as rational drug and vaccine design, and could have implications in other industries, including bioprocessing, biomimetics, biomaterials among others. This thesis presents my results on the characterization of physicochemical properties of microenvironment in proteins. To investigate the complex nature of protein microenvironment, the characterization effort can be broadly categorized into three interconnected topics, namely (i) residue depth (ii) hydrogen bonding, and (iii) multibody statistical potential.

The first topic aims to quantify protein microenvironment using the biophysical parameter of residue depth. Depth of an amino acid measures the degree of amino acid burial in proteins. I have shown that the energetics of proteins, the spatial distribution and chemical properties of amino acids are dependent on residue depth. To exemplify and utilize the results, I have designed several computational methods for protein engineering and functional characterization. First, a novel method to design temperature sensitive alleles of proteins was proposed by making point mutations of these residues. Next, I have used residue depth to identify small molecule ligand binding site on proteins by supplementing it with solvent accessibility and evolutionary information. Benchmarks have shown that the method has comparable or better than the best available methods, and could reveal unconventional sites unidentifiable with other methods. In addition, I have also shown that residue depth can be used in the estimation of protein cavity

volume using a Monte Carlo sampling approach, and pK_a of amino acid residues using a linear model.

The second topic studies the physicochemical properties of hydrogen bonding in different protein environments. I have performed statistical analysis on databases and classified hydrogen bonds into different types, and characterized the geometrical preference and variations of the different types. By analyzing quantum simulation of the system, I have shown that the geometrical preference of main-chain hydrogen bond is due to electron density arising from the planar nature of the peptide bond. I have also performed empirical simulations that strongly suggest the causal link between this geometrical preference and secondary structure formation. Next, I have discovered that low-resolution protein models in databases are consistently missing hydrogen bonds. To ameliorate the models, I have designed a two-step refinement protocol. First, a simple algorithm was used to predict missing pairs of donor-acceptor to form hydrogen bonds based on their mutual preference and specificity. Second, Gaussian restraints were applied on the geometry distribution of the missing pairs, after which a standard modelling protocol can be implemented to refine the protein model. The refinement protocol was shown capable of re-introducing hydrogen bonds in the local environment as well as improving overall model quality. The refinement has functional implication on the protein chemical properties, as exemplified with the more accurate pK_a prediction.

The third topic is constructing an environmental dependent protein statistical potential Packpred. Here, I have explicitly defined protein microenvironments as a set of tightly packed amino acids, dubbed as "residue cliques". Employing Sippl's formulation, the non-random occurrence of microenvironments is characterized. The non-random occurrence is indicative of the strength of interaction among amino acids, and can be interpreted as an energy potential. I have evaluated the capability of the potential in describing protein energetics on a large number of mutagenesis data. The benchmark has shown that, as compared to all other competing methods, Packpred has the best performance not only in binary classification of destabilizing mutants, but also correctly rank-ordering the degree of phenotypical change associated with different mutations.

Lastly, I also present three biomolecular system modelling studies involving non-globular proteins. These system are (i) Cohesin ring protein with coiled-coil structure

(ii) transmembrane transporters OCTN-1 and -2, (iii) interaction interface between oncogenic proteins VAV1 and EZH2. Modelling of these systems are challenging because conventional tools and framework of comparative modelling are not applicable. Instead, an integrative modelling approach was undertaken pertaining to individual systems. In all the modelling work I have proposed experimentally testable hypotheses to decipher the biological mechanism underlying the systems.

In conclusion, in this thesis I have presented an extensive characterization of physicochemical environments of protein. The complex nature of the environment was elucidated by three interdependent topics of residue depth, hydrogen bonding and amino acid cliques. In addition to novel results, for every investigation I have also explored their biological utilities, and have built open-access tools for them. I hope that the work presented here would facilitate future research into protein structures and their functions.

Chapter 1

Introduction

This chapter introduces background on proteins and their structure and how these could be investigated computationally. The significance, utilities and justification of research topics are elucidated. The synopsis of the thesis, which summarizes the research work and findings, is outlined at the end of this chapter.

1.1 Background and Motivations

Protein molecules are the molecular machines which perform most of biological functions in cells [1]. The functions of proteins are largely determined and modulated by their 3D structures. The 3D structure of a protein is formed by the folding of a linear polymeric chain of amino acids into a compact framework. This 3D architecture, also known as the native structure, is formed spontaneously and precisely place amino acids in their preset spatial orientations. These amino acids constitute local micro-environments, and the functional groups in these micro-environments work cooperatively to perform biological functions. These functions include molecular recognition, scaffolding, catalysis of chemical reactions etc. A careful study of these local environments can provide insights into the principles governing protein structural architecture, which would be useful in annotating functions and manipulation of physicochemical properties of proteins. For instances, in the field of computational biology, such knowledge would be useful in protein 3D structure modeling, refinement, assessment and biomolecular functional annotation. Besides, this study could also provide insights into the basis of molecular recognition. This forms the basis of protein design [2], which includes antibody, vaccine, drug development and

enzyme optimization. It is intended in this study to characterize the local environments of amino acid residues by studying several interrelated features. Such characterizations involve the quantitative descriptions of structure motifs and the distributions of several physical features. The descriptions and importance of the features will be introduced in the following section.

1.2 Research Issues

The physicochemical properties of proteins are determined by the chemical nature and the mutual interactions of its constituent atoms. Most of the relevant components of these inter-atomic interactions are already known. They include hydrophobic effect, hydrogen bonding, electrostatic interactions, van der Waals interactions, and backbone and side chain configurational entropy. With the knowledge of the individual components though, the protein structure problem still remains unresolved. Several factors that are contributing to the problem have been identified. First, some of the interactions (e.g. hydrophobic effect and hydrogen bonding) are not completely understood, and consequently their mathematical treatments are not precise; Second, the description of mutual dependency of these interactions are underdeveloped; and third, the link between biochemical/biological function of protein and its micro-environment properties remains elusive. The goal of this thesis is hence to advance the understanding of physicochemical properties through formulating better descriptors of protein features and motifs, as well as establishing their biological importances. Better understanding of these physicochemical properties also forms the basis of predictive models and biomolecule design methods. Specifically, three interrelated topics, namely

- (i) Residue depth
- (ii) Protein intramolecular hydrogen bonding
- (iii) Protein multibody statistical potentials

are pursued. The significance of the topics are explained as follows.

1.2.1 Residue Depth and Its Utilities

The major stabilizing force of protein native structure is the hydrophobic effect [3, 4]; or in other words, the free energy of transferring non-polar amino acids from water environment into organic solvent environment in protein interior [5]. Traditionally, it is assumed that the hydrophobicity free energy is in proportion to its accessible surface area. With this assumption however, the protein structure is overly simplified as a homogeneous non-polar globule, whereas information about the degree of desolvation of amino acid and their preferred spatial distribution in protein are lost. Solvent accessibility measure however, has a small dynamic range, where a large number of amino acid residues, for example those only one layer below the surface layer, are classified as totally buried. To overcome the drawbacks, the biophysical parameter of “residue depth” is introduced [6]. Depth measures the distance of an amino acid to the solvent waters. With this measure, the interior of protein can be distinctly segregated into layers, and micro-environments located at different layers can be separately characterized.

1.2.2 Protein Intramolecular Hydrogen Bonds

Another major force stabilizing protein structure is the hydrogen bonding interaction. Hydrogen bonding is an attractive force between an electronegative atom (acceptor) and a hydrogen atom covalently bonded to another electronegative atom (donor) [7, 8]. Hydrogen bonding is important in both structural and functional roles of proteins. For instance, hydrogen bonds involving protein backbone atoms could provide structural stability (e.g. by forming secondary structures), while side chain hydrogen bonds are often key players in molecular recognition and substrate catalysis. The physicochemical properties of hydrogen bonds in bulk solvent have been extensively studied, yet their dependence in term of protein environments are unclear. Specifically, the characterization of the variations in strength and geometry of protein hydrogen bonds is of interest to us. Such knowledge about hydrogen bonds is useful to derive spatial structural restraints for further biological utilities for protein, such as its 3D model refinement and function prediction.

1.2.3 Protein Multibody Statistical Potential

The physicochemical properties of proteins are often quantified using a statistical approach by constructing a protein “knowledge-based potential”. Knowledge-based potentials are constructed by quantifying the non-random occurrences of structural motifs, or other physicochemical features of amino acids found in proteins (hence it is also known as “statistical potential”). In this study, we recognize that a protein micro-environment is an entity constituted by multiple amino acid residues. We designed a geometric construct “residue clique”, which is a set tightly packed amino acids in space as a proxy of protein micro-environment. Clique statistical potential can be interpreted as a multibody statistical potential that describes the spatial distribution of amino acids in protein, as well as their contributions to protein energetics.

1.2.4 Other Molecular System Modelling Projects

Apart from theoretical studies on properties of protein micro-environment in general, I have also performed molecular modelling of specific protein systems, in collaboration with several different experimentalist groups. The 3D modelling of these systems are challenging, as conventional comparative modelling approaches are not applicable. Instead, integrative modelling approach is adopted, in which informations from multiple sources are combined on an *ad hoc* basis, pertaining to the need of individual systems and availability of the data. The sources of informations are diverse, including biochemistry, heuristics rules derived from database studies, imaging experiments and chemical kinetics data etc. The modelling work of three different systems, the fibrous protein of Cohesin, the transmembrane proteins OCTNs and protein-protein interface between oncogenic proteins of VAV1 and EZH2 are presented here. With the 3D models, we have been able to propose biological hypotheses to our collaborators who subsequently verified them experimentally.

1.3 Approaches

The computational studies of protein systems could be broadly classified into two general approaches, namely the physics-based methods and the empirical methods. The

former approach uses physics-based potential and attempts to quantify the physicochemical properties through computer simulations. Notable techniques of this approach include molecular dynamics simulation, a numerical solution to Newton’s equations of motion; and the more intensive quantum mechanical simulation. The major challenge to physics-based methods is the enormous search space needed to be sampled to arrive at a meaningful conclusion. However, it is noted that under most physiological conditions proteins usually exhibit a non-chaotic behaviour that is completely determined by its primary sequence of amino acids [9]. This strongly suggests that the evolution of proteins followed certain patterns, which are most likely constrained by biological functioning. In contrast to the physics-based methods, the empirical methods aim to recognize and characterize these patterns found in proteins. Among others, two such patterns are conserved sub-sequences and sub-structures in proteins, which are termed “sequence motifs” and “structure motif” respectively. Empirical methods are often based on searching and comparing with known cases deposited in biological databases. A wide variety of computational techniques have been developed in this regime, for example the popular BLAST algorithms that search for similar protein sequences in a non-redundant database and perform multiple sequence alignment with the hits, SALIGN [10] and CLICK [11, 12] which compare 3D (sub-)structures of proteins, and Pfam [13–15] method that aims to identify and cluster among conserved protein families and domains. These methods seek to draw statistical rules from evidences and utilize many concepts and learning architectures from computational sciences. For instance, BLAST [16] performs a k-mer search on the sequence database and CLICK performs a clique search on the structure database; Pfam/HMMER uses hidden Markov model (HMM) and many secondary structure predictions methods are based on k-nearest neighbors and boosting methods.

Our approach to study of protein environments is largely based on empirical methods. The majority of our research is concerned with the characterization of novel feature descriptors, and to establish their biological utilities. For different descriptors and utilities, different computational techniques are appropriately utilized and developed. The approach of our study follows several guidelines, namely

- (i) Reconciliation with experimental data

The validity of a computational model is established only if its predictions agree with known experimental results, and preferably capable of suggesting a experimentum crucis (critical experiment) for testing.

(ii) Utilities and relevance on biological and biochemical studies

Besides basic characterizations, we also develop predictive or design protocols that could facilitate further biological and/or biochemical characterization of the proteins.

(iii) Clear and intuitive interpretation of models

Designed models should favourably have a clear and intuitive interpretation to provide insights into the biology of the protein system. Such models could be easily tuned and/or modified when applied to systems that are dissimilar to the training cases.

1.4 Synopsis

Protein structure consists of a multitude of dissimilar possible micro-environments in its interior in terms of their physicochemical properties. Chemical groups could exhibit different behaviours in different micro-environments. It is hence essential to characterize these micro-environments to understand protein structures and their biological functions. The associations between physicochemical properties with different micro-environments are complex, and involve multiple factors such as the chemical nature of their constituent amino acids, their spatial arrangements and interactions.

To differentiate between different regions of protein, we investigate and further develop the biophysical parameter of residue depth. Depth measures the degree of burial of a chemical group. It differentiates between different layers in protein interior, and is helpful to quantitate the spatial localization of amino acids of different chemical nature. Combined with solvent accessibility measure, we show that cavities on protein surface and interior can be easily characterized. These cavities represent targets for small molecule binding and control of protein thermal stability.

Subsequently, in the later chapters, we characterize the major determinant of spatial arrangement of amino acids in protein, namely the hydrogen bonding interactions. Specifically, their frequencies of formation, preference and variations of geometrical features at different depths, and links to secondary structure formation are investigated. Different approaches are employed in the investigations, including database survey, empirical and quantum mechanical simulations. We further showcase the usefulness of environmental characterization of hydrogen bonds, specifically in the tasks of low-resolution protein model refinement and amino acids pK_a prediction.

Lastly, to investigate the interactions among amino acids, we quantify the frequencies of them clustered together in 3D space. This clustering frequency can be interpreted as proportional to the strength of their interactions. As these interactions are also expected to be different at dissimilar regions in the protein, our quantification takes residue depth into consideration. To validate the model, we correlate the interacting strength with mutational destabilization in two protein systems, CcdB and T4-lysozyme.

The thesis contains of 9 chapters. A short summary and significant findings for each chapter is presented as follows.

Chapter 1: Introduction

The background, significance and synopsis of the study are provided.

Chapter 2: Preliminaries

This chapter introduces the preliminary knowledge on several topics related to the study of protein structure. Topics include basic biochemistry about protein structure, sequence and structure based methods and protein statistical potentials. The formulae of several parameters used to assess the performance of predictive models in later chapters are also provided.

Chapter 3. Characterization of Protein Interior Environments by Residue Depth

This chapter introduces the concept of depth, its definition, computation and significance in biophysical studies. Depth is compared to solvent accessibility in terms of dynamic range and correlations with biochemical properties

of amino acids. The spatial preference of amino acid in protein interior is quantified.

Chapter 4. Utilities of Residue Depth

This chapter discusses the utilities of residue depth. Depth is used in our designs of two biological utilities (1) the design of temperature-sensitive mutant of protein, and (2) the prediction of small molecule binding site. The prediction protocols are explained in detail. Additionally, two other protocols that are developed in collaborations, namely (3) protein cavity size estimation and (4) ionizable amino acid pK_a prediction, are briefly discussed. All these methods use depth as their primary component to describe physicochemical environment. Extensive benchmarks have shown the performance of all our methods to be at least on par, or significantly better than the other state-of-the-art methods.

Chapter 5. Characterization of Intramolecular Hydrogen Bonding in Proteins

Hydrogen bonding constitutes a major component in protein structure stability. This chapter investigates the geometrical properties of protein hydrogen bonds by large-scale study of high resolution protein structures. From the statistical analysis, we update the definition of protein hydrogen bond by relaxing the threshold value of donor-acceptor-antecedent angle to 100° . Next, we empirically categorize hydrogen bonds into 4 types depending on whether the donor/acceptor atom belongs to the main chain or side chain of amino acids. Feature preferences of different hydrogen bonds are found distinctly different from one another, and vary with the environment. Simulations of model systems strongly suggest that the geometrical preference of hydrogen bonds is linked to the formation of secondary structures. The geometrical preference of protein hydrogen bond is compared to its counterparts in other small molecules. Finally, quantum mechanical simulations were performed to investigate the chemical nature of the geometry of protein hydrogen bond.

CHAPTER 6. Protein 3D Model Refinement Using Protein Hydrogen Bond Restraints

This chapter describes a protein structure refinement protocol we designed using hydrogen bond restraints. We first re-establish and quantify the finding that the number of hydrogen bonds in protein is tightly correlated to the chain length. Analysis on low-resolution X-ray and NMR structure however showed that these models are missing a substantial number of hydrogen bonds. Therefore, we designed a two-step protocol to refine these low-resolution models. First, we predict donor-acceptor in close spatial vicinity that are likely to form hydrogen bonds, based on their mutual preferences and specificity. Second, the geometrical preference of different hydrogen bonds learned from the previous chapter is converted into spatial restraints, coupled with other stereochemical restraints, used to refine low-resolution X-ray structure and NMR models. The protocol achieves the same level of performance with state-of-the-art methods, but is much less computational intensive. The biochemical utility of the protocol is also demonstrated by the more accurate prediction of ionizable residues pK_a using the refined models.

CHAPTER 7. Amino Acid Clique Multibody Statistical Potential

This chapter describes a novel multi-body statistical potential based on cliques of amino acids. The concept of amino acid clique, reference state and construction of the potential is elucidated. The statistical potential is assessed for its capability to predict mutational stability of proteins. Benchmarks using the saturated mutagenesis data of CcdB and T4 lysozyme proteins showed that the potential is markedly superior to other competing methods, in terms of both binary classification and rank-ordering of the destabilization of mutants.

CHAPTER 8. Other Molecular System Modellings

Modelling the 3D structure of non-globular proteins is a major challenge in computational biology. This chapter demonstrates the application of integrative modelling, which utilizes information from various experimental sources, physical laws and statistical rules to model such protein structures. The modelling of three different non-globular protein systems, namely coiled-coil protein, transmembrane protein, protein-protein interface are undertaken. All of

these studies were collaborations with experimental biologists. In each of the system, hypotheses are drawn from the models, and experiment validation of these hypothesis are elucidated.

CHAPTER 9. Future Directions and Conclusions

In this chapter, the important conclusions from this study are summarized. Some possible further developments are also recommended.

Chapter 2

Preliminaries

This chapter provides the preliminaries of our study on characterizing the physico-chemical properties of protein environments. Topics include basic protein biochemistry, introduction to computational methods of protein sequence, structure and energetics. The algorithm and techniques that will be used recurrently throughout the work are also described in this chapter.

2.1 Introduction

Protein is a class of important biomolecules in cells. Protein molecules are involved in virtually all biological processes and perform a vast variety of biological functions [1], including the substrate catalysis, molecular transport, scaffolding support, immune responses, cell signaling, signal transduction, transcription regulation, cell cycle, motility and adhesion etc. Apart from biological functions in living cells, proteins are also of importance in industry applications, for example as catalytic enzymes or drugs and vaccines [2].

Chemically, a protein molecule is a linear chain polymer formed by monomeric unit of 20 types of amino acids. The linear form of protein molecule is also called a polypeptide. To function properly, a polypeptide folds spontaneously into a definitive 3D architecture, known as its native structure. It is widely accepted that under native physiological conditions, the native structure is completely determined by

the protein's amino acid sequence [9]. This is known as the Anfinsen's dogma. Furthermore, the trajectory of protein folding is also likely conserved, as it has been suggested that the energy landscape of protein conformations has funnel-like shape [17] which enables the fast protein folding (usually within millisecond) process. The folding and native structure of protein is driven and stabilized by non-covalent interactions such as hydrophobic effect, hydrogen bonding, salt-bridges, ionic interactions, van der Waals forces etc. Hence, the variation of these interactions (for example, due to mutations or a change of physiological environment) could affect the physicochemical properties, and ultimately the functions of proteins.

The 3D structure of a protein can be determined experimentally by several methods, most commonly through X-ray crystallography and NMR spectroscopy, but also the emerging technique of Cryoelectron microscopy [18, 19]. These experiments are however costly and laborious. Consequently, computational methods are often applied to predict the structure of a protein, as well as its biochemical functions and activities. These computational methods could use either or both of protein sequence and structure as inputs. For examples phylogenetic and evolutionary information about a protein can mostly be analysed using its sequence alone, but structural information would be essential for characterization about its biochemistry, energetics and structural dynamics. Some of the basics about these computational methods will be reviewed in the following sections.

It should also be mentioned that the forms and functions of a protein could be altered with additional biochemical processes, such as post-translational modification, and/or association with metal ions or other prosthetic groups. Nonetheless these auxiliary modifications are more system-specific, and are out of the scope of our study.

2.2 Protein Sequence Analysis

2.2.1 Sequence alignment

All important biomolecules of proteins such as DNA, RNA and proteins are all polymers of linear chains. Correspondingly, a biological sequence is the linear string

representation of monomers (4 types for DNA, RNA and 20 types for protein) in the chains. Our study focuses only on protein molecules. As all biological sequences are evolutionarily connected, rich informations about structures, functions, and phylogenetics can be inferred from comparing their sequences.

Central to biological sequence comparison method is the substitution matrix (also known as similarity matrix). A substitution matrix measures the preference of substitution between two amino acid residues. Two popular classes of substitution matrix for protein sequences are PAM (Point Accepted Mutation) matrix [20] and BLOSUM (BLOcks SUBstitution Matrix) [21]. PAM matrix is a Markov chain model and its construction requires an explicit evolutionary model. In contrast, BLOSUM is constructed from, and could be applied more appropriately on more divergent sequences. In our study, only BLOSUM class of matrix is used. For a pair of amino acids, its value in BLOSUM substitution matrix is the log-odd ratio of its frequency of observed occurrences to a background distribution. The observed occurrences in BLOSUM matrices are calculated from sequence profiles derived from automated alignments of 504 protein families [21]. The automated alignment was effected using the program PROTOMAT [21] and generated only ungapped sequence stretches. Background frequency of occurrence are estimated simply assuming random, non-specific substitution between amino acid pairs.

2.2.2 BLAST and PSI-BLAST

In addition to substitution matrix, another component of sequence comparison is the alignment method. Smith-Waterman algorithm [22] has been proposed as an optimal local alignment method (assuming convexity of the alignment problem). While accurate, the algorithm is comparatively time-consuming and not suited for large-scale comparison task, for example searching homologs for a query sequence in a database. A widely used heuristic algorithm for the purpose is BLAST (Basic Local Alignment Search Tool) [16]. Briefly, the algorithm first seeds the alignment by locating exact matches of k-mer (known as “word”), and then extends the seeds iteratively by local alignments. Only extended sequences above a certain threshold

score are reported as a match. The statistical significance of a match is then assessed by assuming an extreme value distribution of the alignment score. The assessment is typically reported as an “e-value”, which estimates the expected number of alignments with larger score than the match by chance in the database. The alignment with the matched sequence could be further refined with Smith-Waterman algorithm.

A major drawback of the BLAST method is that the values in the similarity matrix is fixed. These values were derived from a diverse set of protein families. However in practice, the family-specific substitution profile of a query protein is often more useful, and the similarity matrix should be customized accordingly. Followingly, an extended variant PSI-BLAST (position-specific iterative BLAST) [23], which is the automated, iterative use of BLAST had been proposed. In PSI-BLAST, an initial multiple sequence alignment of high confidence is first obtained from a BLAST search. For each position in the alignment, its substitution profile is re-calculated using the amino acid frequencies of the position only. The re-calculation takes into consideration of sequence weighting, the expected and observed frequencies of amino acid as well as the original similarity matrix. In the next iteration of search, these position-specific profiles are used instead of the original similarity matrix. This process can be repeated until a convergence is reached. The PSI-BLAST algorithm is often substantially more sensitive than the original BLAST method [23].

With the multiple sequence alignment (also sometimes known as a sequence profile), the degree of conservation of every position i can be quantified, for example using the Jensen-Shannon divergence, calculated as

$$J_i = \frac{1}{2} \sum_{a=1}^{20} (f_{ia}^{obs} \log(\frac{f_{ia}^{obs}}{f_a^{exp}}) + f_a^{exp} \log(\frac{f_a^{exp}}{f_{ia}^{obs}})) \quad (2.1)$$

where f_{ia}^{exp} , f_{ia}^{obs} are the expected and observed frequencies of an amino acid type a at position i .

2.3 Homology modelling

Predicting the 3D structure of protein is one of the most important topics in computational biology. Historically, two general approaches have been undertaken toward the goal, namely the *de novo* and the comparative modelling method. *De novo* methods aim to assemble protein native structure using first principles of physics and efficient sampling of conformational space. This approach requires vast computing powers and is still largely impractical to date. In contrast, the comparative modelling makes use on the observation that small change in the sequence usually also results in only a small change in the 3D structure [24, 25]. Hence, known structures of homologous proteins can be used as “structural templates” for proteins of similar sequences. Homology modelling has been used in many applications, for example protein – protein interaction prediction, molecular docking, and functional annotation among others.

Several approaches to model generation methods are available in homology modelling, for example threading [26, 27], fragment assembly [28] and satisfaction of spatial restraints. In this work, we follow the approach of model generation by satisfying spatial restraints. The approach is proposed and implemented in the comparative modelling software suite Modeller [29]. It usually consists of several steps. First, homologs of query protein in a structural database (usually Protein Data Bank [18]) are detected using sequence search approach (e.g. using BLAST). Next, the sequences of the query protein and the collected homologs are aligned. The spatial restraints between corresponding positions are then learned from the corresponding aligned template structure. Spatial restraints can be derived from a wide range of candidate features, for example $C^\alpha - C^\alpha$ distances, dihedral angles of main chain and side chain atoms among others. A spatial restraint usually takes the form of a probability density function. In the simplest scenario, by assuming independence of feature restraints, the molecular probability density (mol-PDF) function of a protein to be modelled is simply the product of individual feature restraints. Numerically maximizing this objective function by efficiently sampling possible protein conformational space would hence generate a restraints-satisfying

3D model. The initial guess of the solution could be the template structure. A wide variety of optimization techniques, for example simulated annealing, conjugate gradients, or even molecular dynamics could be applied for the maximization. As homology modelling relies mostly only on spatial restraints derived from homologous proteins, the resulting 3D model might not be physically valid. Typically, an independent scoring function is needed to assess the protein model by examining its energetics. The method to evaluate energetics of protein is discussed in the next section.

2.4 Protein Statistical Potential

2.4.1 Physics-based and knowledge-based potential

Two general approaches were commonly used in computational prediction of protein free energy, namely the molecular mechanics force fields [30] and statistical potentials. To compute the free energy, the molecular mechanics approach first quantifies the "partial charge" distribution of molecular fragments (chemical groups or larger moieties) using quantum mechanical simulation in isolation, and subsequently samples the conformational space by numerically solving Newton's equations of motion. This approach has been widely used in studying the dynamics of proteins. However, molecular dynamic approach is computationally intensive. Its accuracy is also limited by rounding error and the point charges approximation of wave-functions. In fact, it has been shown that an energetically minimized structure by a molecular dynamics force-field would not correspond to the native structure of protein solved by high-resolution experiment methods [31], and the use of the force-fields yields no improvement or even leads to degradation of protein model quality [31, 32].

The statistical potentials instead approach the problem with methodology inspired from statistical mechanics. It relies on the postulation that the native structure generally has the lowest free energy of all states under the native conditions [9, 33–35]. By analysing the high-resolution protein structures deposited in the Protein

Data Bank, a statistical scoring function to determine the preferred spatial conformation of amino acids can be derived. Due to its dependence on known protein structures, such a scoring function is often termed a knowledge-based potential.

The construction of a knowledge-based potential usually involves statistical analysis of certain geometrical features in proteins. Examples of such geometrical features include number of neighbour in contact of an amino-acid residue [34] or distance between pairs of protein atoms [35]. Most statistical potentials were constructed following Sippl’s formulation, which is an analogy drawn from Boltzmann distribution. The potential of a feature r is computed as

$$U(r) = -kT \log(N^{obs}(r)/N^{ref}(r)) \quad (2.2)$$

where $N^{obs}(r)$ is the observed number of event r in structure databases, and $N^{ref}(r)$ is the expected number of event r in a ‘reference state’. The reference state refers to an idealized, uniform distribution of protein atoms when no potentials (interatomic forces) were present. In earlier work, the reference state was approximated by an ideal gas model [35], and later work such as DFIRE [36] and DOPE [37] have modified the reference state model in consideration of boundary effects. Recently, reference states which have more carefully modelled the angular components [38, 39] are also introduced.

2.4.2 Higher order statistical potential

Statistical potentials are being widely utilized, and have achieved impressive successes in protein folding, protein – protein docking, and protein design [2, 28, 34, 35, 37, 40–42]. These potentials are however approximations as they assume pairwise interactions to be independent of one another. As a result, they cannot reveal higher order correlation among amino acid residues and are not suited to describe cooperative phenomena, especially in local environments. For instances, it has been shown that a two-body statistical potential alone cannot fold a protein [43–49]. The correlation between potential and model quality also diminishes when the model is close to native conformation.

One possible cause of this failure is the assumption of independence of interaction between pairs of atoms. This might not hold true at near-equilibrium structure, since in spatial vicinity, there exist stronger cooperative and shielding effects [50–53] among residues. These effects could be essential in biological functions and dynamics of proteins, for example catalytic sites and nucleation core of folding protein. In accordance, increased efforts have been made in quest for multibody statistical potentials which are more informative about cooperative interactions among residues [54–57]. Since recent past, as the number of experimentally solved protein structures has grown rapidly, more sophisticated geometric constructs, optimization and refinement on these potentials are possible. To date, several multibody statistical potentials have been proposed and benchmarked, but so far neither 3-body contact potential nor 4-body non-overlapping tessellation method have led to better results than pair-wise potentials [47, 49, 54, 58, 59] in protein model quality assessment and refinement.

Apart from cooperative effect, we argue that environmental dependence also need to be considered when designing a statistical potential. Our efforts and results in such design will be detailed in Chapter 7 of the thesis.

2.5 Solvent Accessible Surface Area

Apart from interatomic interactions among protein atoms, another important class of interaction is the protein-solvent interaction. Most proteins function in aqueous condition in cells, and hence the degree of solvation of its constituent atom should be carefully quantified. A widely used measure of protein solvation measure is the solvent accessible surface area [60]. It measures the fraction of an atom that is accessible to water solvent, commonly normalized with the total surface area of the atom. The fraction can be further calibrated with a reference Ala-X-Ala tripeptide (X is the amino acid of interest) in extended conformation, and reported in a percentage format.

Though several fast approximation methods based on probabilistic models exist [61–64], in our study we chose to implement the numerical method of Shrake-Rupley algorithm [65], as its accuracy is tunable with computational resources. In

Figure 2.1: Illustration of Shrake-Rupley algorithm. The orange circle represents solvent molecule and black circle represents protein atoms. Each dot of protein circle represents a surface area element. The mutually shielded surface area (green) are discarded when calculating the solvent accessible surface area.

Shrake-Rupley algorithm, for every atom in protein, a sphere of radius equals to the sum of atom's van der Waal's radius and water's radius was constructed, with the interested atom placed at the center. A number (default 92) of discrete points are placed uniformly on the sphere. Each of these points represents a surface element of the atom. The sphere of two atoms in vicinity can overlap with one another, representing their mutual shielding from solvent molecule. These shielded points are removed, and only the unshielded (solvent accessible) points are retained. The sum of the surface elements represented by the retained points give the accessible surface area of an atom. As the procedure is equivalent to finding the area rolled by a ball, it is also known as the "rolling-ball algorithm".

I have implemented the algorithm in C++ programming language for all the work covered in this thesis. The source code is released under LGPL(v3.0) license and can be downloaded at <http://mspc.bii.a-star.edu.sg/tankp/download.html>.

2.5.1 Golden spiral approximation

To accurately compute the surface element of an atom, it is essential to achieve a uniform distribution of points on its spherical surface. The exact solution to this problem corresponds to the classical regular polyhedron problem, and has been shown possible only for 5 cases of $n = \{4, 8, 10, 12, 20\}$. However, several other

Figure 2.2: Uniform placement of points on the surface of a sphere using Golden Spiral approximation method. Placement of 50 (left), 92 (center), 200 (right) points are shown in the figure.

numerical approaches can be implemented for the purpose, including position constrained simulation, cylindrical projection of uniform points on rectangular area, as well as a variety of spiral algorithms [66]. In this study, Golden spiral approximation [66], was chosen due to its high accuracy, symmetricity and ease of implementation.

Algorithm 1 Golden Spiral Approximation

Input: $(x_0, y_0, z_0), r, n$ $\triangleright$ origin, radius and the number of points to be place on sphere

Output: $\vec{\theta}, \vec{\phi}$ $\triangleright$ polar coordinates of points on a unit sphere

procedure GOLDENSPIRAL($(x_0, y_0, z_0), r, n$)

for $k \leftarrow 1, N$ **do**

$h \leftarrow -1 + 2(k - 1)/(N - 1)$

$\theta_k = \arccos(h)$

if $k = 1$ **or** $k = N$ **then**

$\phi_k \leftarrow 0$

else

$\phi_k \leftarrow (\phi_{k-1} + 3.6/\sqrt{(N(1 - h^2))}) \bmod (2\pi)$

end if

end for

for $k \leftarrow 1, N$ **do**

$q = 2k/n - 1 + (1/n)$

$p = \text{sqr}(1 - q^2)$

$\phi = 2.4k$

$x_k = r \cdot p \cos(\phi) + x_0$

$y_k = r \cdot q + y_0$

$z_k = r \cdot p \sin(\phi) + z_0$

end for

end procedureGoldenSpiral

Algorithm 2 Shrake-Rupley Algorithm

Input: $\vec{x}, \vec{y}, \vec{z}, \vec{r}, r_w, n$ $\triangleright$ (x,y,z) coordinates, radii of protein atoms, radius of solvent and resolution number

Output: $\vec{A}$ $\triangleright$ Areas of protein atoms

procedure SHRAKE-RUPLEY(N)

for all $atom \in protein$ **do**

$\vec{S}_i = GoldenSpiral((x_i, y_i, z_i), n, r_i)$

end for

$\triangleright$ Removal of screened surface elements

$\vec{E} \leftarrow \vec{S}$

for $i \leftarrow 1, N$ **do**

for $j \leftarrow 1, N$ **do**

if $i \neg j$ **then**

$\triangleright$ intra-atomic screening ignored

for all $e_i \in S_i$ **do**

for all $e_j \in S_j$ **do**

if $Distance(e_i, e_j) \leq r_i + r_w$ **then**

$E_i = E_i \setminus e_i$

end if

end for

end for

end if

end for

end for

$\triangleright$ Calculation of surface area

for $i \leftarrow 1, N$ **do**

$A_i = |E_i| 4\pi r^2 / n$

end for

end procedureShrake-Rupley

Two limitations on accessible surface as a measure for residue burial are however noted. Firstly, the measure is useful only in describing solvation environment on protein surface. Beyond which in the protein interior, all atoms are shielded from solvent. This poses a limit when we want to differentiate between protein atoms lying just below protein surface, and protein atoms in the core. Secondly, by its computation, accessible surface area is the sum of continuous sweep of solvent molecule; while in actuality the solvent molecules are discrete points and cannot simultaneously occupy the same space.

2.6 Cell-list Algorithm

The physicochemical environment of protein is a direct consequence of the spatial arrangement of protein atoms and their interactions. The distances among protein atoms are often prerequisite to the calculation of physical interactions. Enumerating all interatomic distances is $O(n^2)$ in computation. However in many cases, only short-ranged interaction is required (e.g. in the computation of contacts, surface area, van der Waal's force etc) for the computation. In these cases, cell-list algorithm, which is $O(n)$ in computation, could be utilized.

The cell list algorithm consists of several steps, first it divides the 3D space into cells of certain width (corresponds to the cut-off distance). The immediate neighbourhood of any cell is then computed and recorded. Next, for every atom the algorithm determines the cell it belongs to. Finally, only interatomic distances of atoms in the same or neighbouring cells are computed.

A standalone program implementing the algorithm has been written in C++ programming language. The source code of the program is released under LGPL(v3.0) license and can be downloaded at <http://mspc.bii.a-star.edu.sg/tankp/download.html>.

2.7 Molecular Dynamics

Molecular dynamics (MD) is a versatile computational method to investigate the dynamics of a biomolecular system. Typically, the method simulates the trajectory

Algorithm 3 Cell-list Algorithm

Input: $\vec{x}, \vec{y}, \vec{z}, r_c$ ▷ (x,y,z) coordinates, distance cut-off
Output: $\vec{D}$ ▷ short-ranged interatomic distances
▷ Cell assignment

procedure CELL-LIST($\vec{x}, \vec{y}, \vec{z}$)

$n \leftarrow |\vec{x}|$

for $i \leftarrow 1, n$ **do**

$Cells[[x_i/d], [y_i/d], [z_i/d]] \cup i$

end for ▷ Neighbouring cells

for all $cell \in Cells$ **do**

for all $i \in \{-1, 0, 1\}$ **do**

for all $j \in \{-1, 0, 1\}$ **do**

for all $k \in \{-1, 0, 1\}$ **do**

$Neighbours[cell[p, q, r]] \leftarrow Neighbours[cell[p, q, r]] \cup cell[p + i, q + j, r + k]$

end for

end for

end for

end for ▷ Evaluate distances for atoms in neighbouring cells

for all $cell \in Cells$ **do**

for all $cell_{ngh} \in Neighbours[cell]$ **do**

for all $atom_i \in cell$ **do**

for all $atom_j \in cell_{ngh}$ **do**

$D(atom_i, atom_j) = dist(atom_i, atom_j)$

end for

end for

end for

end for

end procedureCell-list

of every atom in the system by numerically solving Newton’s equations of motion subjected to a force-field:

$$\dot{v} = dv/dt = F(x), \dot{x} = dx/dt = v \quad (2.3)$$

where x , v , F represents the position, velocity and force exerted on an atom at time t .

The force-field specifies the strength of interactions among different atoms in the system, and are generally expressed in the form of simple algebraic functions between point masses. The MD force-field is normally composed of covalent (bond length, bond angle, dihedral angle) and noncovalent interactions (electrostatic and van der Waals interaction). It should be noted that MD force-field is only an approximation to the quantum mechanical description of the interatomic interaction, and hence the parameters of the algebraic functions need to be carefully calibrated. The force-fields that are commonly used in biomolecular system include AMBER [67–69], CHARMM [30, 70], GROMOS [71, 72] and OPLS [73, 74].

In this work, molecular dynamics simulations are performed using GROMOS96 [71] force field and SPC216 water model [75, 76].

2.8 Ramachandran Map

In a protein individual amino acids are covalently linked by peptide bonds. Peptide bonds possess double bond characteristics due to the delocalization of lone pair electrons on backbone nitrogen to the oxygen atom. One important consequence of the double-bond characteristics is that the peptide bond exerts a planar constraints on the atoms along the bond (C^α -C-O-NH- C^α). As a result, for a peptide backbone, only dihedral angles (ϕ , ψ) around the peptide bond (Figure 2.3 B) are needed to specify its degree of freedom in space.

Not all values are allowed for the dihedral angles. Certain combinations of (ϕ , ψ) angles are energetically prohibitive as they would result in steric clashes of neighbouring atoms (primarily C^β atoms). The allowed region can be calculated

Figure 2.3: Illustration of dihedral angles (ϕ , ψ) at peptide backbone linked by a peptide bond. Image taken from G.N. Ramachandran, 1963 [77]

and represented in a Ramachandran map. The original Ramachandran map is constructed using hard-sphere potential among atoms (Figure 2.4 A) [77]. With the large number of protein structures deposited in the Protein Data Bank, it is now possible to reproduce the map using statistics from the solved structures. In this study I have reproduced the map using a high resolution (resolution $< 2 \text{ \AA}$, R-factor < 0.25), non-redundant (sequence similarity $< 30\%$), single domain (chain length 100 – 200 residues) training set of 1175 protein structures (Figure 2.4 B).

2.9 Benchmark Measures

This section provides the mathematical preliminaries on three types of correlation coefficients which will be used throughout the thesis, namely (1) Pearson’s correlation coefficient, (2) Matthews correlation coefficient and (3) Spearman’s rank correlation coefficient.

2.9.1 Pearson’s correlation coefficient

Pearson’s correlation coefficient, or simply correlation coefficient, measures the linear dependence of two variables x and y . It is calculated as

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (2.4)$$

Figure 2.4: (A) The original Ramachandran map. Fully allowed region and generously allowed region are represented in solid and dashed line respectively. Image taken from G.N. Ramachandran, 1963 [77] (B) Reconstructed Ramachandran map using high-resolution protein structures in Protein Data Bank represented in a contour plot. The propensity (in logarithmic scale) of different region is covered with lines of different colors.

where n is the sample size.

2.9.2 Matthews Correlation Coefficient

Matthews correlation coefficient (MCC) [78] is normally used to measure the goodness of a binary classification. It is a balanced measure particularly useful when the sizes of the classes are very different. It is calculated as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.5)$$

where TP , TN , FP and FN represents the number of true positive, true negative, false positive and false negative predictions. Another way to interpret MCC is to construct a data set with TP as (1, 1), FP as (1, -1), TN as (-1, -1) and FN as (-1, 1). MCC is simply the Pearson's correlation coefficient of this classifier set.

2.9.3 Spearman's Rank Correlation Coefficient

In some scenarios, the relationship between the predictor and target variable is not linear. Yet the predictors can be assessed with its capability of ranking target variables. Spearman's rank correlation coefficient is used in these cases. Spearman's rank correlation can be defined as the Pearson correlation coefficient between the ranked variable, as is calculated as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2.6)$$

where d is the difference between the actual and the predicted ranks of a mutation, and n is the number of levels.

Chapter 3

Characterization of Environments in the Protein Interior

The physicochemical properties of any chemical group is environment-dependent. In bulk solvent, the characterization of amino acids properties is straightforward, as the environment is simply an homogeneous aqueous solution of water molecules in this case.

However, this is not true in proteins, where the environment is inhomogeneous. A variety of interatomic forces, such as van der Waal's force, hydrogen bonding, salt-bridges etc are interacting among amino acids of different chemical natures. The environmental properties could change drastically across different regions in protein. As an example, the relative permittivity a polar chemical group experiences in bulk solvent is around 80, yet in protein the value could range from 20-30 on protein surface and only 2-4 in protein interior [79]. It is hence not surprising that the same chemical group or amino acid would exhibit different behaviour depending on their locations in proteins.

In this chapter we introduce “residue depth” to quantify the location of an amino acid / chemical group in protein. Depth is a measurement of residue burial based on the distance to the nearest surface bulk water. We first illustrate the concept and computation of depth, and compare it with solvent accessibility, a complementary measure for residue burial, over their dynamic ranges.

Equipped with the depth measure, the spatial frequency distribution of amino acids with different chemical natures in protein are also quantified.

3.1 Background and Motivations

3.1.1 Importance of the protein solvation description

The biophysical characteristics of protein amino acid residue depends on the chemical environment it is in. Solvation is one such important environment descriptor, as it affects the many parameters related to protein biophysics, such as dielectrics, hydrophobicity, solvation, folding dynamics, structural flexibility, kinetically accessible conformations etc [80–87]. Hence, accurately quantifying protein amino acid solvation (or equivalently, its degree of burial) is important in understanding and manipulation of functions of proteins.

A more commonly used measure of burial had been the solvent-accessible surface area (SASA) [65], or solvent accessibility for short. As the name implies, SASA measures the proportion of surface of an amino acid residue accessible to a water molecule. A major drawback of the measure is that it cannot differentiate among solvent inaccessible residues, and its dynamic range is concentrated only on the protein surface. For example, the residues in protein core and residues just below the protein surface would be indistinguishable, as both would have the same zero solvent accessibility. Yet it is likely that these two environments would have very different physicochemical properties (Figure 3.1). Secondly, the numerical value of SASA of flexible side chain on protein surface is unstable. This is because SASA has a strong dependency on the orientation of the atoms, which in turns are affected by solvent dynamics. Hence usually SASA description can only be coarse-grained into 2 (buried / exposed) or 3 (buried, intermediate, exposed) classes.

In accordance, the “depth” of a protein amino acid residue was proposed as another burial measure complementary to solvent accessible surface area. Depth is defined as the distance of an atom of a biomolecule (protein molecules in this study) to its nearest (water) molecule from bulk solvent [6, 88, 89]. As the value of residue

Figure 3.1: (A) Schematic diagram showing two solvent inaccessible ($ASA = 0\%$) amino acids in different environments (orange: just below protein surface, red: protein that core) cannot be differentiated by their solvent accessible surface area (represented by green dots). The structure is rendered using PyMol software suite [90] (B) Schematic diagram showing that different protein environments (cyan: surface, pink: intermediate, purple: deep) can be differentiated by using residue depth measure (distances from surface solvent, represented by arrows in the diagram)

depth increases monotonically from protein surface to protein core, it allows better differentiation among different layers of amino acid residues (Figure 3.1).

3.1.2 Reviews of previous work

As the depth measure allows finer description of residue burial, it has been shown to be a concise descriptor of protein environment. For examples, it has been shown to be better suited than SASA to characterize and predict properties of proteins such as hydrogen/deuterium amide proton exchange rates [6, 91]. Residue depth also correlate with structural stability under mutagenesis [6], sizes of globular domains [6, 92], protein-protein interaction hot spots [6], and post-translational modification [92, 93], such as phosphorylation sites [94]. In addition, it has been shown that depth and a closely related variant ([94]) correlate well with hydrophobicity [6, 92, 93], residue conservation [93] and folding nucleation sites ([93, 94]). Also, it is recently demonstrated that residue depth can discriminate between accuracy of protein 3D model when used together with deep-sequencing techniques [95].

3.2 Definition and Computation of Depth

3.2.1 Removal of non-bulk water

In proteins, amino acid residue depth is defined as the average of the depths of its constituent atoms.

Depth is designed to differentiate the protein environment and the bulk solvent environment, and is defined as the nearest distance of the atom to its nearest water molecule from bulk solvent [6]. Importantly, bulk solvent is distinguished from non-bulk solvent in depth algorithm. Unlike in the bulk solvent, non-bulk waters are surrounded primarily by the protein environments rather other water molecules. These waters could have structural and functional roles, and are best classified as a constituent of protein environment. To differentiate between bulk and structural waters, it is noted that most structural waters can be classified into two groups (a) water molecules residing deeply in protein to provide structural support or involved in chemical interaction (b) water molecules residing in surface cavities. In both cases, the non-bulk solvent is either isolated, or making only limited contact to bulk solvents. This hinted to us that the differentiation of bulk and non-bulk can be achieved by inspecting the immediate neighbourhood of the water molecule in question. Surrounding a bulk water are the other bulk waters, while around a non-bulk water, fewer bulk water and more protein atoms are found. Here, the immediate neighborhood of a water molecule is defined as a spherical volume of a specified radius (default value of 4.2 Å, or 1.5 hydration shell). A water molecule is considered non-bulk, if there are less than a certain number of neighborhood waters (default 4) within this spherical volume (Figure 3.2). These non-bulk waters are not used to calculating depths (i.e. distance to protein atoms).

3.2.2 Consideration of bulk water dynamics

Surface bulk water surrounding a protein is not static but freely diffusing. To accurately compute depth, the dynamics of bulk water has to be considered. While

Figure 3.2: Schematic diagram to illustrate detection of non-bulk water in protein. The center water molecule (oxygen atom coloured in purple) is in consideration. In (A), there are 4 other water molecules within the spherical volume (represented in dashed line), and hence the center water molecule is considered bulk. In (B), there are less than 4 water molecules within the volume, and the center water molecule is considered non-bulk.

it is possible to examine the dynamics of water molecules through molecular dynamic simulation [6], a heuristic rule was made to circumvent the computationally intensive approach. In depth algorithm, the effect of solvent dynamics is approximated using a Monte Carlo approach. Rather than simulating solvent dynamics however, the configurations of protein are sampled instead. In the samplings, a protein molecule is treated as a rigid body, and is repeatedly solved, each time in a different orientation. New orientations are generated by rotating the protein by a random angle about an axis passing through its center of mass. Each solvation of the protein is considered to represent a snapshot of the dynamics of bulk-water. With sufficient number of repeated solvations (default 25), water molecules can explore all regions accessible to bulk solvent water, hence mimicking bulk-water dynamics.

In all our work, a SPC216 water model [75, 76] is used. Solvation is performed with a pre-equilibrated water box, and no energy minimization is performed. Water molecules that clash with protein atoms are removed. For simplicity purpose, an atomic clash between a water molecule and a protein atom is defined when the two are within 2.6 Å of each other. The depth algorithm is hence summarized as follow (Algorithm 3.2.2).

It should be noted that our implementation differs from the original proposal and is improved by (i) adoption of SPC model to better describe the water dynamics (ii) modifying the default number of neighborhood waters to define cavity water is increased from 2 to 4. (iii) iterative removal of cavity waters until equilibrium is reached (only one iteration of cavity water in the original implementation).

The algorithm has been implemented in C++ programming language. The program also implements Open Multi-Processing (OpenMP) interface to support shared memory multiprocessing architecture. Source code of the program is released under LGPL(v3.0) license and can be downloaded at <http://mspc.bii.a-star.edu.sg/tankp/download.html>.

Next we will demonstrate the utility of residue depth by its capability of characterizing and predicting physico-chemical properties of proteins.

3.3 Properties and distribution of Depth

Next we proceed to characterize the properties of depth in proteins. We first examine the correlation between residue depth and SASA. For the purpose we have constructed a dataset with 1457 single domain (chain length between 100 - 200) proteins from the Protein Data Bank. The dataset consists of only high resolution (resolution $\leq 2\text{\AA}$, R-factor $\leq 25\%$) and non-redundant (sequence identity cut-off 30%) protein structures from different protein families. In the dataset, there are a total number of 168456 amino acid residues (ALA: 13121, ARG: 8633, ASN: 6985, ASP: 9960, CYS: 2310, GLN: 6469, GLU: 11952, GLY: 11737, HIS: 4059, ILE: 9765, LEU: 15584, LYS: 10110, MET: 3648, PHE: 7032, PRO: 7319, SER: 9930, THR: 9277, TRP: 2428, TYR: 5624, VAL: 12513).

3.3.1 Comparison between depth and SASA

We first compare the corresponding dynamic ranges between SASA and residue depth. SASA of amino acid residues were computed for the dataset using Shrake-Rupley algorithm (see Chapter 2) with a resolution of 92. As expected the means of

Algorithm 4 Depth algorithm

Input: $Atoms, Waters, N_{min}, r_{probe}, r_{ngh}, cycle$

Output: $\vec{\mu}(depth), \vec{\sigma}(depth)$

```

procedure DEPTH( $Atoms, Waters, N_{min}, r_{probe}, r_{ngh}, cycle$ )
  for  $k \leftarrow 1, cycle$  do
    for all  $doatom \in Atoms$ 
       $depth[atom][k] = \infty$ 
    end for
     $\vec{R} \leftarrow rotate(\vec{R}, (\theta_{random}, \phi_{random}))$ 
     $\vec{R} \leftarrow translate(vecR, r_{random})$ 
    for all  $sol \in Waters$  do
      for all  $atom \in Atoms$  do
        if  $dist(sol, atom) < r_{probe} + r_j$  then
           $Waters \leftarrow Waters \setminus sol$ 
        end if
      end for
    end for
    while  $\neg convergence$  do
       $Neighbour[i] = \{\}$ 
      for all  $sol_i \in Waters$  do
        for all  $sol_j \in Waters$  do
          if  $distance(sol_i, sol_j) < r_{ngh}$  then
             $Neighbour[sol_i] \leftarrow Neighbour[sol_i] \cup sol_j$ 
          end if
        end for
      end for
      for  $i \leftarrow 1, N$  do
        if  $|Neighbour[sol_i]| < N_{min}$  then
           $Waters \leftarrow Waters \setminus sol_i$ 
        end if
      end for
    end while
    for all  $atom \in Atoms$  do
      for all  $sol \in Waters$  do
         $depth[atom][k] = \min(depth[atom][k], dist(atom, sol))$ 
      end for
    end for
    end for
    for all  $atom \in Atoms$  do
       $\mu(depth[atom]) = \sum_{k=1}^{cycle} depth[atom][k] / cycle$ 
       $\sigma(depth[atom]) = (\frac{1}{cycle} \sum_{k=1}^{cycle} depth[atom][k] - \mu(depth[atom]))^{1/2}$ 
    end for

```

Figure 3.3: Boxplot representation of relationship between residue depth and solvent accessible surface area. (A) Solid red line shows the median, and boxes display the quartiles of the distributions. (B) Coloring of dynamic range from surface (blue) to interior (red) on an example protein PDB:1MVT. Dotted lines indicate the boundaries of different protein environments discerned by the two measures.

depth and SASA monotonically anti-correlates with one another - in general solvent accessible residues are on protein surface, and inaccessible residues are deeply buried (Figure 3.3). For a typical single domain protein, the value of depth ranges from ~ 2.8 Å to ~ 9 Å. The depth-SASA relationship can be fitted into a log-log regression equation of

$$\log(ASA) = -4.14 \log(depth) + 9.74 \quad (3.1)$$

It is verified that the dynamic range of SASA is concentrated only on protein surface. Numerically, half of the SASA range (50% to 100%) only describe the outermost 1 Å (depth ~ 2.8 Å - 4 Å) layer of protein. On the other hand, residue depth increases gradually from surface to interior, and is capable of differentiating between the deepest residues (depth > 8 Å).

Second, residue depth has previously been demonstrated to be a better descriptor of protein dynamics than solvent accessibility [6]. The dynamics of a protein can be probed with hydrogen-deuterium exchange experiment. HDX experiment measures the rate of replacement of hydrogen atom in protein when it is immersed in deuterium solution, and hence is a direct indicator of local dynamics.

Figure 3.4: Correlation of depth and accessibility with H/D exchange rate in 8 proteins. Correlation coefficients for residue depth and solvent accessibility are shown in blue and green bars respectively. Figure reproduced from Chakravarty, Suvobrata and Varadarajan 1999 [6].

Our collaborator has collected extensive exchange-rate data for a set of eight proteins (HEW lysozyme PDB:193L, E. coli thioredoxin PDB:3TRX, Calbindin D9K PDB:1CLB, Pseudomonas aeruginosa cytochrome c-551 PDB:351c, Ribonuclease T1 PDB:1BVI, Ribonuclease A PDB:1RUV, Antidigoxin VL PDB:domain 1MAJ, Equine lysozyme PDB:1EQL) and compare the correlations of exchange rate of residue depth with solvent accessibility. Residue depth on average has a correlation coefficient of 0.61 ± 0.10 , while solvent accessibility 0.40 ± 0.10 . Furthermore, residue depth has better correlation than solvent accessibility in all cases.

3.3.2 Residue depth preference of amino acids

The residue depth distribution profiles of biochemically similar amino-acid show similarity to each other (Figure 3.5). All amino acid residues show a peak in the distribution at low depth values (3–4 Å). Hydrophobic amino acids (ALA, ILE, LEU, MET, PHE, VAL, TRP and TYR) and CYS showed a higher tolerance for deep environments than other residues, with a second peak in the distribution at around 7 Å. In contrast, charged and large polar groups of amino acids (ARG, ASN,

ASP, GLU, GLN, HIS and LYS) have a lower propensity to be in deep environments. The same trend was noticed for PRO. Smaller polar amino acids SER and THR along with GLY exhibit a similar trend albeit with a relatively larger tail of the distribution at deep environments.

3.4 Utilities of Residue Depth

3.4.1 Background and Motivations

The functioning of a protein relates critically to the intricate interatomic interactions of its atoms. The environmental dependence of these interatomic interactions is crucial to the correct functioning of many biological reactions carried out by proteins. Hence, a concise protein environment descriptor could markedly improve our understanding and prediction of protein functions and properties. In the preceding chapter we have characterized the measure of depth and established its capability of describing protein environment. Expectedly, the utilities of depth measure can be expanded into other design and predictive algorithms for protein functions. In this chapter we demonstrate the design of two such methods, namely (1) the design of temperature-sensitive mutant of protein and (2) the prediction of small molecule binding sites on proteins. Several other utilities of depth developed in collaboration will also be briefly discussed. The importance of these protein properties and our principles to predicting are detailed in individual section of the chapter. All our methods have been benchmarked, and achieved performance at least on-par, or better than other state-of-the-art methods.

3.4.2 Temperature-sensitive mutant design

Background and Motivations

Temperature sensitive (Ts) mutants of a protein are those whose levels of activity is dependent on its surrounding temperature. Typically, when the temperature rises above a threshold value (restrictive temperature), the activity of the mutant

Figure 3.5: Depth distribution of the 20 different amino acids represented in histograms

protein decreases drastically; while below the threshold, in permissive temperatures, the mutants would have similar activity levels as the wild type. Ts mutants are powerful tools to study protein function *in vivo* and in cell culture [96–98]. Ts mutants have found utility in many areas of biology research. These include the investigation of folding pathways [99], macro-molecular assembly [100], controlling the genotype of a cell *in vivo* [101], nervous system defects [102], phenotypic effects [103],[104], coordination between different genes [104], especially during the cell-division cycle and developmental biology [105–107]. Ts mutants also provide a means for conditional expression of genes, as demonstrated in the model organism of *drosophila* [108, 109].

Despite its wide utility, Ts mutants are still most commonly generated by random mutagenesis. The process involves the introduction of random mutations to protein coding DNA by the means of chemical mutagens, ultraviolet (UV) radiation or error prone Polymerase chain reaction (PCR) techniques. These mutated proteins are then screened to select Ts mutants (if any) [102, 110]. This method, not only laborious and expensive, but is also not feasible for model organisms with long generation times, where it is impractical to obtain large numbers of progeny [111]. This calls for a faster and more accurate method in designing and generating Ts mutants.

Principle of Design

To design Ts mutants we have made use of the observation that Ts phenotype usually correlates with decreased protein stability [112, 113]. And this reduction in stability generally corresponds with decreased protein activity. A number of strategies could be adopted could destabilize a protein. For example, by comparing mesophilic proteins to their thermophilic counterpart, a list of thermal stabilizing mutations could arise from improved core packing, addition of hydrogen bonds and/or salt bridges, shorter and/or tighter surface loops, enhanced secondary structure propensities or oligomerization [114]. Among the strategies, optimized protein core regions with tight atomic packing of hydrophobic residues are most commonly adopted [115]. It is also known that significant destabilization in protein stability

Figure 3.6: (A) Flowchart of TSpred algorithm (B) Flowchart of optional homology modelling in TSpred.

and activity can be achieved by mutating buried residues as compared to residues on protein surface [116–118]. Accordingly, it is proposed that by mutating hydrophobic residues in the protein core to a diverse range of amino acids would likely generate a Ts mutant of a protein [108, 109, 111, 119].

As established in the preceding chapter, depth has the better capability than solvent accessibility in differentiating between the protein interior and outer layers that are also inaccessible to solvent. Using the depth measure, deeply buried positions can be easily identified. If these positions are occupied by hydrophobic residues, a variety of mutations with diverse stereochemical properties are suggested to generate a Ts mutant. It is assumed that these mutations will destabilize the protein to different extents, and that at least one mutant is likely to be temperature sensitive. The aim of TSpred algorithm is to replace the current method of generating Ts mutant from using laborious large-scale random mutagenesis to testing of only a few highly possible candidate mutations. As a single mutation is sufficient to transform the protein into a TS mutant, we aim to make few predictions instead of covering all temperature sensitive mutants.

Design Protocol

(1) Prediction based on 3D structure

Depth of a residue (or atom) measures its distance to the closest bulk solvent (water). While it directly measures degree of burial, a threshold depth value for different types of amino acid is to be determined to identify positions in protein interior. An average depth value that corresponds to 5% solvent accessibility was found by an earlier study [119] as an optimal value to define buried residues. Correspondingly, we studied a high-resolution (resolution 1.7 Å, R-free 0.2) non-redundant dataset (30% sequence identity cut-off) dataset of 561 single domain (chain length of 120–180 amino acids) proteins extracted from the PDB, and determined the threshold depth values of Val, Ile, Leu, Met, Phe and Trp to be 6.25, 6.75, 6.75, 7.00, 7.00 and 7.00 Å, respectively.

In cases where a high-resolution structure is not available, a homology model of the protein will be attempted in replacement for the depth computation. For homology modelling, we search for suitable structural templates in the PDB database using three iterations of PSI-BLAST (version 2.2.28)[16, 21] and a stringent e-value cut-off of 0.0001. Among the hits, only the one with the best e-value is chosen as template. Next, a sequence alignment between query sequence and the template sequence was built using SALIGN [10]. The resulting alignment is input to the automodel protocol of Modeller [29] to construct a 3D model of the protein. The model is assessed for accuracy/reliability with the GA341 [120, 121] and Discrete Optimized Protein Energy (DOPE) statistical potentials [37]. Residue burial information is only taken from those models that satisfy a stringent GA341 and DOPE cutoff of 0.75 and 1, respectively.

(2) Prediction based on sequence information

In the cases where only sequence information or low-resolution 3D model are available, a sequence-based prediction algorithm had been proposed. It is based on the observation that the likelihood of an amino acid residing at a buried position can be inferred from its hydrophobicity and that of its flanking residues. Accordingly, two parameters of the average hydrophobicity H_{av} [122, 123], and hydrophobic

moment H_{mom} [124] were computed to estimate this likelihood. While average hydrophobicity is simply the moving average of hydrophobicity, hydrophobic moment is introduced because of structural constraint consideration. Both helices and strands often have one solvent-exposed hydrophilic face and one buried hydrophobic face. Buried regions of such sequences are therefore non-continuous and cannot be identified using only H_{av} . H_{av} and H_{mom} are computed as

$$H_{av}(j) = \sum_{n=j-3}^{j+3} H(n)/7 \quad (3.2)$$

$$H_{mom}(j) = \left\{ \left[\sum_{n=j-4}^{j+4} H(n) \sin(\delta \cdot n) \right]^2 + \left[\sum_{n=j-4}^{j+4} H(n) \cos(\delta \cdot n) \right]^2 \right\}^{1/2} \quad (3.3)$$

where the $H(n)$ is the normalized Rose residue hydrophobicities [125], δ is the phase angle and is dependent on the periodicity of the secondary structure that the sequence is assumed to adopt. All three possible values of (i) $\delta(\alpha\text{-helix}) = 100^\circ$, (ii) $\delta(\text{flat } \beta\text{-sheet}) = 180^\circ$ (iii) $\delta(\text{curved } \beta\text{-sheet}) = 160^\circ$ are tried. A series of rules based on the two parameters give the predicted positions and confidence level. These rules are mainly based on the previous work of large scale data analysis (Table 3.1) [111].

(3) Suggested mutations

At the predicted positions, a set of mutations with diverse chemical nature are suggested. These amino acids include (i) Ala (hydrophobic, small) (ii) Trp (hydrophobic, large) (iii) Asn (polar) (iv) Asp (charged) and (v) Pro (secondary structure breaker).

Performance Benchmark of Ts mutant Design

(1) Case study on CcdB protein

We first demonstrate the design protocol using *Escherichia coli* CcdB protein as an example. CcdB is chosen because the availability of saturation mutagenesis data [119] which can be used for prediction validation. The structure-based method identified 14 target positions (V5, F17, V18, V20, M32, V33, L34, L36, M63, M68,

Confidence Level (%)	Prediction criteria
≥ 95	Residue, as well as both flanking residues, are hydrophobic and $H_{av} \geq 75$
≥ 90	(a) Residue is hydrophobic and $H_{av} \geq 75$ or (b) Residue, as well as both flanking residues, are hydrophobic and $H_{av} \geq 65$
≥ 80	Residue is hydrophobic and any of the following conditions are met: (a) $H_{av} \geq 60$ and preceding residue is hydrophobic (b) $H_{av} \geq 65$ and both flanking residues are hydrophobic (c) $H_{av} \geq 70$ 4. $H_{mom} \geq 200$ and residues at either (-3 and +4) or (-4 and +3) relative to the residue are hydrophobic (3) $H_{av} \geq 70$ (d) $H_{mom} \geq 200$ and residues at either (-3 and +4) or (-4 and +3) relative to the residue are hydrophobic

Table 3.1: Sequence based rules to predict temperature sensitive mutation positions

Figure 3.7: A functionally active CcdB dimer with one monomer in surface representation and the other in ribbon representation. The labeled location of the predictions made by the sequence-based, structure-based and both predictions are coloured blue, red and purple. Image taken from Tan et al, 2014 [126].

I90, I94, M97, F98) (Figure 3.7), and the sequence-based method identified 10 target positions (F17, V18, V33, L34, L50, V53, V54, L96, M97, F98). Eight Ts mutant targets (V5, V20, M32, L36, M63, M68, I90, I94) were exclusively predicted by the structure-based method. All these predictions were verified to be correct (100% accuracy). Four target positions were exclusively predicted by the sequence-based method (L50, V53, V54 and L96) and one false positive (V53) was found. For false negatives, V54 (depth 6.18 Å) only marginally missed the depth threshold (6.25 Å) for structure prediction. The other false positives of L50 and L96 are likely involved in active sites (4 Å within catalytic residues) and hence not the intended targets of our algorithm.

In CcdB, both sequence- and structure-based methods predicted ~10-15% of all residues to be mutation target, but the intersection of predictions narrow the candidates to ~6% of all residues. All intersection predictions were verified true positives.

(2) Performance benchmark

To benchmark our prediction method, 36 mutants from 6 proteins for which extensive mutagenesis data are available are used to construct the testing set. These

proteins are gene V (PDB: 1YHA), lambda repressor (PDB: 1LMB), T4 Lysozyme (PDB: 2LZM), CcdB (PDB: 3VUB), Gal4 (PDB: 3CQQ) and Ura3 (PDB: 1DQW)) [108, 119, 127–132] The performance of our predictions was assessed by coverage (the number of predictions) and precision (number of true positives / number of predictions) 3.2. Results from both sequence-based and structure-based methods are reported.

The benchmark result showed that our methods successfully predicted all 36 experimentally validated Ts mutant positions in the six proteins, CcdB, Lysozyme, gene V, gal4, Ura3 and Lamda repressor (Table 3.2). The sequence-based method correctly predicted 22 and the structure-based method 28 of the 36 cases. The two methods were correct together in 14 of the cases. In addition to the correct predictions, the sequence-based method made one confirmed false positive identification, Val53 in CcdB. Our methods also made 71 other predictions in these six proteins that are yet to be experimentally validated.

(3) Effect of model accuracy on prediction

For cases where depth and residue information was inferred from homology model, we need to quantify the effect of model accuracy on Ts mutant design. Using T4 lysozyme as proxy, we built 15 of its homology models using individual templates of varying sequence similarity. The templates were identified by searching DBAli [133] for structures similar to T4 lysozyme (PDB 2LZM). We defined similarity as a minimum MAMMOTH P-value of 10 and at least 100 equivalent positions (61% structure overlap with T4 lysozyme). The selected templates were between 22 and 91% identical in sequence to T4 lysozyme and all identifiable by PSI-BLAST search with an e-value cutoff of 0.0001.

For the five known Ts mutant positions in T4 lysozyme, high accuracy models (template sequence identity >40%) correctly identified four mutant positions (Table 3.2), yet for models built with low target-template sequence identity (<25%) templates, it was found that their DOPE and GA341 scores failed to reach the acceptance threshold values, and only one or two mutant positions were correctly predicted. In these cases, sequence-based method was also shown better in identifying Ts mutation than structure-based method.

Protein	PDB ID	Chain length	Residue position	Residue type	Prediction type
gene V	1YHA	87	35	VAL	Both
			45	VAL	Both
			47	ILE	Structure
			63	VAL	Structure
			81	LEU	Structure
lambda repressor	1LMB	92	78	ILE	Sequence
			51	PHE	Both
			65	LEU	Both
			76	PHE	Both
			84	ILE	Both
			18	LEU	Structure
			36	VAL	Structure
T4 lysozyme	2LZM	164	47	VAL	Structure
			6	MET	Both
			102	MET	Both
			149	VAL	Structure
			153	PHE	Structure
CcdB	3VUB	101	103	VAL	Sequence
			17	PHE	Both
			18	VAL	Both
			33	VAL	Both
			34	ILE	Both
			54	VAL	Both
			5	VAL	Structure
			36	LEU	Structure
			63	MET	Structure
			50	LEU	Sequence
			53*	VAL	Sequence
			96	LEU	Sequence
			97	MET	Sequence
			98	PHE	Sequence
Gal4	3COQ	88	68	PHE	Both
			69	LEU	Sequence
			70	LEU	Sequence
Ura3	1DQW	267	25	MET	Structure
			32	LEU	Structure
			118	ILE	Structure

Table 3.2: The wild-type residue (three-letter amino acid code) at the position is listed under residue type. * The prediction of VAL53 in CcdB as a Ts mutant position is a false positive identification.

	Template quality			predictions			prediction type—					
PDB	ID	DOPE	GA341	S	D	B	M6	M102	V103	V149	F153	
1pqj:A	90.8	-2.08	1	3 (8)	4 (11)	2 (3)	B	B	S	D		D
1d3n:A	86.1	-1.96	1	3 (8)	4 (11)	2 (3)	B	B	S	D		D
1t8a:A	81.6	-1.65	1	3 (8)	4 (13)	2 (3)	B	B	S	D		D
1cx6:A	79.9	-2.03	1	3 (8)	4 (11)	2 (3)	B	B	S	D		D
1lpy:A	78.8	-1.95	1	3 (8)	4 (10)	2 (3)	B	B	S	D		D
1swz:A	77.5	-2.21	1	3 (8)	4 (13)	2 (3)	B	B	S	D		D
1lwk:A	77	-1.74	1	3 (8)	4 (12)	2 (3)	B	B	S	D		D
1swy:A	74.5	-2.22	1	3 (8)	4 (12)	2 (3)	B	B	S	D		D
1sx2:A	72.4	-2.28	1	3 (8)	4 (13)	2 (4)	B	B	S	D		D
1wth:A	43.2	-1.49	1	3 (8)	3 (12)	1 (2)	S	B	S	D		D
1k28:A	43.2	-1.43	1	3 (8)	5 (15)	3 (5)	B	B	B	D		D
2anv:A	24.2	0.54	0.12	3 (8)	1 (11)	1 (2)	S	B	S			
2anx:B	23.9	0.6	0.08	3 (8)	2 (11)	2 (3)	S	B	B			
2anv:B	23.5	0.49	0.13	3 (8)	2 (11)	2 (3)	S	B	B			
2anx:A	22.1	0.64	0.13	3 (8)	1 (10)	1 (2)	S	B	S			

Table 3.3: The number of predictions made by the sequence-based (S), structure-based (D) or both methods (B) are listed for each of the models with the number of experimentally validated predictions within brackets. The performance of the different models on the experimentally validated mutant positions are additionally shown in separate columns.

Web-server description

We have made our method of Ts mutant prediction accessible to public through a web-server at <http://mspc.bii.a-star.edu.sg/TSpred/>. Our server supports both sequence and structure inputs, and several options are provided for each. For input sequences, users could either specify a database, GenBank [134] or UniProt [135, 136], identification number or upload the protein amino acid sequence in FASTA format. If the input consists of multiple sequences, Ts mutant predictions are made for each of the sequences separately. For input structures, the users could either specify the four-letter PDB code with optional additional letters to select specific chains of the protein (the biological unit is used for predictions) or upload a file in PDB format. User uploaded structures are used for prediction without any model assessment.

The web server outputs the residue positions that have been predicted as Ts mutation targets and prediction method (structure-based, sequence-based or both). In the case where a homology model was used, the server only displays the results of the structure-based method if the models satisfy the assessment criteria. The user has the option to review details of the hydrophobic moment, average hydrophobicity and residues flanking of predicted mutant positions. In cases where 3D structural data have been used in making the Ts mutant predictions, the PDB structure or the homology model is displayed using a Jmol plugin. The backbone chain trace of the proteins is coloured according to the residue depth and the mutant positions are shown as spheres centered on the C $^{\alpha}$ atoms. The spheres are also coloured according to residue depth. In the case of homology models, the user is notified about the target-template sequence identity and the alignment used in model construction is provided.

Summary and Conclusions

Generating temperature sensitive mutant of proteins is a useful experimental technique to conditionally express genes and determine the roles of gene and gene products of different cellular processes. In this section, we propose a targeted Ts mutant design method to circumvent the traditional laborious large-scale screening

technique. Our method aims to identify a small number of residue positions in proteins and suggests substitutions that are likely to produce Ts mutants. Our predictions are based on the assumption that mutation of hydrophobic residues in protein interior is likely to destabilize a protein and rendering it temperature sensitive.

For protein with high resolution 3D structure (either solved experimentally or from homology modelling), the protein interior environment is identified using the depth measure. Whereas in cases where 3D structure is not available, a sequence based prediction method is used. This method estimates the interior by using the measure of hydrophobicity and hydrophobic moment. Our method has been experimentally validated in 36 positions in six different proteins. Both the methods are shown capable of accurately identifying Ts mutants, with structure-based method slightly outperforming sequence-based method. Nonetheless, using the common predictions of the two is recommended as this could substantially reduces (halves) the number of predictions and hence increases precision (90–100% in the case study). Also, by using the common predictions, we could reduce the Ts mutant predictions for experimental testing to only 3 positions per protein.

The dependency of structure-based prediction on homology model quality was also benchmarked. Our results show that the use homology model could generate prediction of similar accuracy as long as a high quality template (as measured by DOPE and GA341 score) is used.

As shown in the case in CcdB, Ts mutant position might not necessarily be in protein interior. Mutation at residue positions near an active site are also likely to be temperature-sensitive. These targets are however more difficult to predict and do not fall into the regime of our algorithm. For practical purpose also, usually one or a few Ts mutants are sufficient to facilitate further investigation of a biological system. To make accessible of the design method to experimentalists, our program is made accessible through a user-friendly web-server.

3.4.3 Protein ligand binding site prediction

Background and Motivations

Binding cavities are invagination on protein surface to accommodate its ligands. These ligands could include small molecule chemical substrate (molecular weight <1.5 KDa) to be catalyzed by the protein, protein-protein interface or protein-DNA interface etc. Identification of ligand binding cavities on protein is important, as such knowledge often helps in protein functional annotation and serves as a starting point to guide ligand docking, targeted mutational studies, as well as structure-based drug design [137]. This section focuses on the predicting protein binding cavities of small molecules. These binding sites are often located in pockets on the protein surface and provide micro-environments unlike in solvent and other parts of proteins to facilitate catalytic reactions of the ligand.

Review of previous work

Given the importance, several methods have been proposed to detect protein cavities over the years. These methods can be broadly divided into three classes of (1) pure geometrical methods (2) incorporation of evolutionary information and (3) meta-algorithms. Pure geometrical methods are normally used as a base prediction, for which the other two classes of methods improve upon. These methods are individually reviewed as follows.

Pure geometrical methods

Pure geometrical methods predict ligand binding sites using only spatial coordinates of protein atoms. These methods focuses on the description of geometry or shape of the binding sites. Generally the methods consist of the following steps:

- (a) Identification of potential binding sites using geometrical constructs
- (b) Ranking / Selection of potential binding sites
- (c) (optional) Mapping of prediction to protein residues

Identification of potential binding sites using geometrical constructs

Protein evolves to form characteristic cavities to bind its intended ligands. These cavities can be identified by general computational approaches that define geometrical shapes. Different methods of this class of predictor use different cavity shape definitions. For example, one of the first method, POCKET [138] uses cubic grid points representation. Grid points are labeled as ‘protein’ or ‘solvent’ depending on whether a protein atom is within 3 Å of it. Grid points are scanned along x-, y-, and z- direction, and ‘solvent’ grid points that line between ‘protein’ points are retained as potential binding sites (Figure 3.8A). LIGSITE [139] improved upon this representation by including four more cubic diagonals, and a later version used Connolly surface. SURFNET [140] uses a single point representation. First, a sphere with a large radius was placed in between protein atoms (within 10 Å apart), and potentially containing the atoms (Figure 3.8B). The size of the sphere is dynamically reduced until no protein atoms were contained. Only spheres of biologically relevant size (1 to 4 Å) are kept, and the center of the reduced sphere represents the potential binding site. CAST [141] uses alpha-shape representation by computing triangulation (Figure 3.8C) of the protein’s surface atoms. Small triangles are clustered into larger triangles in the vicinity to form a potential binding site (Figure 3.8D). PASS [142] uses an iterative approach fill protein cavity with probe spheres. Probe spheres in the vicinity are clustered into an active site points, represented by the central probe in the cluster.

Ranking / Selection of potential binding sites

Binding sites identified using pure geometrical approaches often contain large number of false positives. More recent methods attempted to exclude these false predictions using auxiliary filters. For examples, PocketPicker [143] calculates the ‘buriedness-index’ of grid probes by scanning the protein surroundings, and retain clusters of probes within biologically relevant range (16-26). PASS uses a similar approach by calculating ‘burial count’, which is defined as the number of atoms within an 8 Å distance of a probe. Only probes with a count above a threshold are retained. Empirical filters were also used, as in Q-SiteFinder, which uses the van der Waal’s potential (of a methyl probe) and an interaction energy threshold to determine favourable binding clefts. PocketFinder uses the same strategy

Figure 3.8: Illustration of different pocket identification methods (image taken from Huang & Schroeder, 2006 [139])

as Q-SiteFinder, except that the van der Waal's interaction was calculated with Lennard-Jone potential with parameters taken from AutoDock [144]

Incorporation of Evolutionary Information

The residues in protein ligand binding site are usually evolutionarily conserved to bind specific target and/or catalyse biochemical reactions. Binding sites predicted by pure geometrical method can be vastly improved by integrating evolutionary information into their pipelines. ConCavity [145] is the pioneering and representative method in this class. In this method, evolutionary information is expressed in term of sequence conservation, and quantified by Jensen-Shannon divergence. Different empirical functional forms, such as product, the arithmetic mean, the geometric mean, the product of exponentials, and the product of exponentials are calculated. Basal binding site prediction using several pure geometrical methods of Surfnets, PocketFinder, Ligsite were also pre-calculated. All different possible combinations using the geometrical predictions and divergences are tested, and performing functional form is selected.

By integrating evolutionary sequence conservation, ConCavity greatly improved the performance of ligand binding site prediction. However, the method also highlighted the difficulty that multi-chain proteins pose for structure-based methods for identifying ligand binding sites.

Meta-algorithms

Meta-algorithms are primarily jury systems that seek to combine the predictions from the other different ligand binding site prediction methods. MetaPocket [137] and its successor MetaPocket 2.0 [146] are the representatives for this class of methods. MetaPocket uses 4 predictors (LIGSITEcs [139], PASS [142], Q-SiteFinder [147], and SURFNET [140], and MetaPocket 2.0, 8 predictors (Fpocket [148], GHECOM [149], ConCavity [145] and POCASA [150] in addition to previous 4) (Figure 3.9). A normalized z-score was computed for predictions from each of the predictors, and the top 3 predicted sites from each predictor were considered. These sites were clustered based on spatial vicinity and ranked by the total z-score. Finally, the mass centers of these consensus clusters were output as prediction.

Figure 3.9: Illustration of the metaPocket 2.0 procedure. Image taken Huang et al, 2009 ([137])

MetaPockets performed better than its component predictors after extensive training. It should however be noted that, in most meta-algorithms, only metadata (prediction accuracy in this case) were considered, and the assumptions of its component predictors are neglected. As such, the performance assessment of meta-algorithm is valid only when all assumptions from individual predictor are fulfilled.

Principle of Prediction

Depth binding site prediction follows both pure geometrical-based method, and incorporation of evolutionary information. To explain our method of protein ligand binding site prediction, we first recall the concept of residue depth and solvent accessible surface. Depth measures the distance of an atom/residue from bulk water, and solvent accessibility measures the fraction of surface of an atom/residue to solvent molecule. The two measures are usually complementary and anti-correlates with each other. Exposed residue on surface has low value of depth and high solvent accessibility, and buried residues in the protein interior has high depth and low solvent accessibility. This relationship, however, does not hold for amino acid

Figure 3.10: Schematic diagram showing prediction of protein small molecule binding site by identifying residues with simultaneously deep, yet exposed to solvents. Green dashed lines show the mean and one standard deviation of the regression line (coloured in yellow). Outliers (deep yet solvent accessible residues) are circled.

residues lining a cavity on a protein. In the depth algorithm, water molecules residing in cavities are regarded as an integral part of protein structure but not bulk solvent. These structural waters are not used in defining residue depth. This suggested that for residues lining a ligand binding site, the amino acid residues would simultaneously be deep and exposed to solvent (Figure 3.10). In other words, on a function describing depth-SASA relationship, outlier residues that are spatially close would likely form a ligand binding site.

Prediction Protocol

(1) Calibrating binding cavity probabilities for amino acids

Following the prediction principle, the relationship between depth and solvent accessibility is quantified. We recognized that different types of amino acid residues could have different propensities in forming a ligand binding sites. Hence, the depth-SASA relationship is computed specifically each of them. To achieve so, a data set 900 high resolution (X-ray crystallization resolution < 2 Å, R-free $<$

Figure 3.11: (A) Probability of a CYS amino acid residue lining a small ligand binding site represented as a heat-map. The probability of binding is color-coded from black to yellow with ascending value. (B) Server screenshot demonstrating the identification of binding site using predicted residues above a threshold value. Residues colored in red are predicted to form a binding site.

0.25), globular, single chain, ligand-bound proteins between 150 and 200 amino acid residues in length was extracted from the PDB. Residue depth and SASA values were computed for all residues of all structures in the calibration set. Residue depth and SASA values were segmented into discrete bins of size 0.1 Å and 1%, respectively. To compensate for sparseness of data, Gaussian blurring technique was utilized. Every depth and SASA measurement is treated as a Gaussian probability distribution (parameterized by its mean and standard deviation) rather than a single point. This ensures that neighboring bins are also filled. Gaussian blurring with standard deviations of 1 Å and 10% was applied to depth and SASA, respectively. The probability, P_b , of an amino acid R to form part of a binding cavity was then parameterized by the residue depth D and SASA S using the relationship $P_b = R_{S,D}^{bound} / R_{S,D}^{total}$, where the numerator is the number of observed occurrences of residue R bound to a ligand in the training set, and the denominator is the total number of such residues found in the same depth-SASA category (Figure 3.11A).

(2) Identifying binding cavities residues

When predicting binding cavities on a protein, SASA and depth are first computed for all residues of the protein. Probability values are assigned corresponding to its residue depth and SASA categories. All residues with probability values above a

certain threshold value are selected as binding cavity residues. The protein is then resolvated (25 cycles by default). This time only clashing waters are removed. Water molecules that are within 4.2 Å from any of the selected residues are inspected. An inspected water molecule is retained if there is at least one other water molecule within 4.2 Å of it. The rest of the waters are not considered. All protein residues that are within a distance of 4.2 Å from the retained water molecules are also considered as residues lining the binding cavity (Figure 3.11B). A continuous patch of these residues constitutes an independent binding cavity linings prediction.

(3) Parameters optimization

Two parameters still remained undetermined for ligand binding site detection, namely (1) the threshold probability value (P) and (2) the minimum number of neighborhood waters (n) to define a bulk solvent. The second parameters also specifies the size and curvature of the intended cavities. A small n value detects a cragged cavity, while a large n value would classify a flatter region as a cavity. To optimize the parameters, first the dataset of 325 structures of single and multi-chain proteins listed in LigASite v7.0 [151] holo structure (bound to small molecule ligands) were obtained from the Protein Quaternary Structure (PQS) server [152]. These structures were randomly bifurcated into a training set of 100 and a testing set of 225 with average chain lengths of 282 and 245, respectively. The training and testing sets do not overlap with each other. A grid search for n in the range 2 - 5 in steps of one and the binding probability threshold (P) in the range 0.10–0.80 in steps of 0.05 over the training set was performed. Residues were classified binarily as belonging to binding site or non-binding site, and the performance of each combination of (n , P) is measured by the best Matthews correlation coefficient (MCC). The optimized parameters of (n , P) is determined to be (4, 0.50).

(4) Incorporation of evolutionary information

An enhanced version of the predictor was constructed by incorporating evolution information (in term of conservation) of the residues. For every residue an adjusted binding probability score

$$q_i = \alpha P'_i + (1 - \alpha) J'_i \quad (3.4)$$

is assigned, where α is an optimized weighting coefficient, P_i and J_i are the normalized binding probabilities of residue and conservation score respectively. Where binding probabilities are calculated based on depth and solvent accessibilities as previously described, the conservation score J_i is defined as its Jensen-Shannon divergence with respect to a background distribution of amino acids occurrence. To compute the Jensen-Shannon divergence, a multiple sequence alignment of homologues for an query protein was first obtained by running 5 iterations of PSI-BLAST [23] against the uniref90 sequence database [153] with an e-value cutoff of 0.0001. Divergence value at a position i is given by

$$J_i = \frac{1}{2} \sum_{a=1}^{20} (f_{ia} \log(\frac{f_{ia}}{f_a^{bg}}) + f_a^{bg} \log(\frac{f_a^{bg}}{f_{ia}})) \quad (3.5)$$

where f_{ia} is the frequency of residue a at position i and f_a^{bg} is the frequency of the residue a in the background distribution. In addition, pseudo-counts were introduced [154] to account for sparseness of data using the following formulae

$$f_{ia} = \frac{n_{ia} + b_{ia}}{N_i + B_i} \quad (3.6)$$

$$b_{ia} = B_i \sum_{k=1}^{20} \frac{n_{ik}}{N_i} \cdot \frac{t_{ka}}{T_k} \quad (3.7)$$

, where

$$T_k = \sum_{a=1}^{20} t_{ka}; B_i = m \cdot R_i \quad (3.8)$$

where for amino acid a at position i , f_{ia} is its occurring probability, n_{ia} is its frequency and b_{ia} is its pseudo count. N_i and B_i are the total number of residue counts and pseudo counts at position i , respectively. t_{ka} is the probability that amino acid

k would be substituted by amino acid a as estimated from the BLOSUM62 matrix [21]. T_k is the overall probability of substituting amino acid k . m is a parameter that has been set to 5 and R_i is the number of different residue types at position i . As conservation score J and binding probability P differ in magnitude, both measures were normalized respectively to unity using

$$S'_i = \max(\min(\frac{S_i - S_{\min,5}}{S_{\max,5} - S_{\min,5}}, 1), 0) \quad (3.9)$$

where S_i is the normalized measure, $S_{\min,5}$ and $S_{\max,5}$ are the mean values of the smallest and largest five values of the respective measures.

(4) Cavity water detection and binding residue identification

In the method with incorporated evolutionary information, we first predict waters in binding cavities likely to be displaced by ligand, after which the binding residues are predicted.

We estimate for every cavity water, its likelihood to be displaced by a small molecule ligand. The displacement likelihood D is given by

$$D = \prod_e \left(1 - \prod_i (1 - q_i^c) \right) \quad (3.10)$$

where q_i^c is the adjusted binding probabilities of residue i from chain c that is within 5.6 Å of the cavity water. Here we made the implicit assumption that displacement likelihood of a cavity water from two different chains is independent of one another. Additionally, we made the assumption that a ligand must displace at least two water molecules (distance between displaced waters should not exceed 4.2 Å, i.e. 1.5 hydration shells) for a binding event to occur. Hence, for every cavity water, its neighbouring cavity waters within 4.2 Å are identified. The displacement likelihood of a cavity water was assigned as the average of the two highest displacement likelihoods of its and its neighbouring waters.

A cavity water was predicted to be displaced when the averaged displacement likelihood exceeds a threshold value β . All residues with at least one atom within

Figure 3.12: Schematic diagram showing the detection of cavity waters and binding site residues. Water molecules on protein surface are shown as circles, and color-coded with likelihood of being replaced by ligand (red: likely; blue: unlikely). The residues within 6.5 Å of predicted displaced water are represented as grey triangles, and predicted as the binding site.

6.5 Å of this cavity water were listed as binding site residues candidates (Figure 3.12). The candidate residues listed from different solvation iterations could be slightly different owing to differences in cavity water configurations. A residue was predicted to be part of the binding site if it was listed in >60% of the solvation iterations.

The values of weighting coefficient α and threshold value β were similarly optimized for their Matthews Correlation coefficient (MCC) over a training set of 99 ligand bound PDB structures using a grid search of step 0.05 and 0.1, respectively. The optimized values of α and β were 0.7 and 0.8, respectively.

Performance Benchmark

We have compared the performance of our predictors (both geometrical and incorporation of evolutionary information) to other contemporary methods of LIGSITE [155], Pocket-Finder [156], SURFNET [140], ConCavity [145] and MetaPocket 2.0 [146]. In our tests, ConCavity was run using LIGSITE for structural geometry and evolutionary information for the queries were taken from the ConCavity web site.

The probability value threshold for ConCavity for binary classification was set at 0.085. LIGSITE, Pocket-Finder and SURFNET were also run from within the

Single-chain PDBs (92)				Multi-chain PDBs (133)				Entire testing set (225)			
D	L	P	S	D	L	P	S	D	L	P	S
0.44	0.53	0.48	0.47	0.39	0.37	0.37	0.38	0.39	0.4	0.39	0.39

Table 3.4: The Matthew’s correlation values for the DEPTH (D), LIGSITE (L), Pocket-Finder (P), and SURFNET (S) over the testing data set.

ConCavity program, with all parameters set to their default values, and without the use of evolutionary information. MetaPocket 2.0 was run through the web-server, and the number of predicted binding sites was set to 1. MetaPocket 2.0 returns predictions for only a fraction of the LigASite testing set. The comparisons of MetaPocket to DEPTH and ConCavity were done on 110 proteins (70 single- and 40 multi-chain).

In comparison of pure geometrical methods (Table 3.4), DEPTH has an overall similar (MCC ~ 0.39) performance as the other methods (MCC 0.39 - 0.40). Interestingly however, while binding sites from single-chain and multi-chain protein are examined separately, DEPTH only has a smaller discrepancy of $\Delta\text{MCC} \sim 0.05$ in performance, while the other methods have ~ 2 fold larger than this value. This suggests that the other methods could have overfitted their models on single chain proteins. This render the methods less accurate in detecting binding sites formed by multi-chain proteins, which essentially has the same physicochemical properties. The problem with overfitting is exemplified when meta-algorithm that combines these methods is used. For example, in MetaPocket 2.0, the MCC of single chain protein was ~ 0.55 , but for multi-chain protein this dropped drastically to only ~ 0.33 (>4 -fold discrepancy in magnitude as compared to DEPTH).

For methods that incorporate evolutionary information, both DEPTH and ConCavity has an elevated accuracy (DEPTH MCC: 0.50, ConCavity MCC: 0.51) (Table 3.5), and a two-tailed paired t-test showed that the difference is not statistically significant ($P = 0.8$). More detailedly, it was shown that while DEPTH (0.63) is not as sensitive as ConCavity (0.80) and MetaPocket 2.0 (0.71), its predictions are more specific (DEPTH specificity: 0.92, ConCavity specificity: 0.87, MetaPocket 2.0 specificity: 0.89) and more precise (DEPTH precision: 0.48, ConCavity precision: 0.43, MetaPocket precision: 0.43). An interesting finding is that, though both

Figure 3.13: (A) Venn diagram showing the overlaps among DEPTH, ConCavity and the true binding site residues (LigASite) (B) Comparison of MCC values between DEPTH and ConCavity over different protein chain lengths. Depth prediction is coloured in blue, while ConCavity predictions coloured in red.

DEPTH and ConCavity are accurate at the same level, their predictions only partially overlap with each other (Figure 3.13A). Of all predictions made by DEPTH, 67.6% overlap with ConCavity. Of the consensus predictions by the two methods, 61.3% are true binding sites. In all, 49.4% of binding sites were predicted by both methods, and 86.3% of all binding sites were predicted by at least one of the methods. It is noted that the difference between the two methods is more prominent at multi-domain proteins (Figure 3.13B). Two complications could arise in this case, (i) the binding sites could be formed from amino acid residues from different protein domains, (ii) cavities could form at around protein-protein interface as an artefact of the binding event. These artefact cavities are challenging to be differentiated from actual binding site using shape analysis alone. Depth method is better suited in these cases as it takes into consideration of the context of location of the cavity in protein. The partial overlap between the two high performance methods also hinted that a meta-algorithm could be developed to combine DEPTH and ConCavity for more accurate prediction.

Web-server and case study

(1) Server description

	DEPTH	ConCavity	MetaPocket 2.0
Single-chain			
N	80	80	70
MCC	0.55	0.53	0.55
Difference		0.02	0
P-value		0.39	0.47
Multi-chain			
N	120	120	40
MCC	0.47	0.5	0.33
Difference		0.02	0.15
P-value		0.34	0.04
All			
N	200	200	110
MCC	0.5	0.51	0.47
Difference		0	0.03
P-value		0.78	0.04
Sensitivity	0.63	0.8	0.71
Specificity	0.92	0.87	0.89
Accuracy	0.89	0.87	0.87
Precision	0.49	0.43	0.43

Table 3.5: Performance benchmark for DEPTH, ConCavity and MetaPocket 2.0.

A user-friendly web-server which computes depths and predict protein ligand binding sites has been constructed, and is accessible at <http://mspc.bii.a-star.edu.sg/depth>. (This server also calculates cavity volumes, and predicts the pK_a of ionizable amino acid residues, as will be briefly discussed in the following sections). Users can submit a protein structure (in PDB format) or specifying its four-letter PDB code. While optimal parameters are set by default on the server, users have the option to override these values to cater to specific biological systems. The results of the computation/prediction are returned in pictorial representation and/or rendered using the (Jmol viewer, with appropriate accompanying figure legends).

(2) Case study: West Nile Virus NS2B/NS3 protease

We illustrate the use of our method by a case study on the West Nile Virus NS2B/NS3 protease (PDB:2FP7). The protease is a protein complex of two chains (chain A:NS2B and chain B:NS3), 47 and 148 amino acid residues long, respectively. The binding site is made up of 16 residues from both chains. This binding site is unusual as it is flat geometrically, and binds peptide instead of small molecules as ligands. In this case, no binding sites could be predicted using the default values of number of neighborhood waters ($n = 4$) and probability threshold ($P = 0.50$). Only after increasing the minimum number of neighborhood waters to 5 and decreasing the probability threshold to 0.45 can the binding site be detected. Our prediction correctly identifies 9 of the 16 binding site residues (Figure 3.14).

Summary and Discussions

Binding sites are important motifs on protein surface, as they possess chemical environment that has high selectivity for ligand binding and its catalytic reaction. As depth is a concise parameter to describe protein environment, we explored its utility in predicting the location small molecule binding cavities on proteins. To make the prediction, we made use of the observation that, often some of the residues lining a ligand binding cavity are both deep and solvent exposed. Using the depth-SASA pair values for a residue, its likelihood to form part of a small molecule binding cavity is estimated. The parameters of the method, namely the number of neighbouring water to define free solvent and threshold probabilities, were calibrated

Figure 3.14: (A) DEPTH binding site prediction on West Nile Virus NS2B/NS3 protease (PDB:2FP7), using modified parameters of $(n, P) = (5, 0.45)$. Predicted binding cavity residues are colored red while the rest of the protein is colored blue. The inhibitor is shown in white stick representation to highlight the flat geometry of the binding site. (B) Details of the prediction in pictorial and list representations. Image taken from Tan et al, 2011 [88]

over a training set of 900 high-resolution X-ray crystal structures of single-domain proteins bound to small molecules. Residues that have high propensities of ligand binding, together with their neighbouring residues are identified as residue lining a binding site. As compared to other geometrical based approaches, the coarse-grained method of DEPTH has achieved performance of similar level. It is also found that DEPTH does not suffer from overfitting, and can be used to detect binding site formed at the interface of multiple proteins. Further, to reduce false positive rates, evolutionary information is utilized along with structural information for a prediction. The enhanced method performs as well as the other best method of ConCavity. Additionally, predictions made by DEPTH method are more specific and precise than that of ConCavity. This renders DEPTH a better suited method to provide a limited set of targets for mutational experiments. Lastly, we have built a web-server for our method and demonstrated through the server the effects of changing the parameters of the algorithm. Users can follow the principle we laid out in the case study of the West Nile Virus protease to detect binding sites of different sizes and shapes.

Figure 3.15: Correlation between experimentally measured free energy change on mutation (ΔG^0) and cavity sizes computed by DEPTH (a), CASTp (b) and McVol (c). Image taken from Tan et al, 2013 [89]

3.4.4 Other Utilities of Depth Measure

(1) Protein cavity size estimation

Another utility of DEPTH algorithm is to compute the volume of protein cavities. Cavities in protein interior are most likely destabilizing, due to the loss of van der Waals' interactions. In DEPTH algorithm, protein cavities could be identified as the spatial location of non-bulk waters. By clustering non-bulk waters from multiple iterations, the 3D shape of a protein cavity can be accurately described. The volume of the 3D shape could be calculated using a Voronoi procedure [157]. On a benchmark set of 40 proteins with cavity creating mutants (mutations larger residues (VAL, LEU, ILE, MET, PHE and TRP) to ALA) [158], we showed that DEPTH has the most accurate results (error $\sim 0.9 \text{ \AA}^3$) as compared to all other methods (McVol: $1.3 \text{ \AA}^3$, CASTp: $23.2 \text{ \AA}^3$, VOIDOO: $-96.1 \text{ \AA}^3$). Secondly, we also shown that our volume estimate has the best correlation ($R^2 = 0.75$) with the destabilization ($\Delta\Delta G_0$) using 34 mutants from five proteins (RNase S, Barnase, Gene V protein, T4 Lysozyme and Human Lysozyme) as compared with other methods (CASTp: 0.65, McVol: 0.28) (Figure 3.15).

(2) Ionizable amino acid pK_a prediction

Acid dissociation constant, or more commonly termed as pK_a , measures the tendency of protonation of a chemical group. Protonation states of residues modulate

Residue	N	DEPTH	MM-SCP	DEMM	Improvement
ASP	(12)	0.72	0.67	0.53	0.14
		(1.04)	(0.79)	(0.67)	(0.12)
GLU	(12)	0.37	0.45	0.37	0
		(0.48)	(0.58)	(0.5)	(-0.02)
HIS	(15)	1.14	1.15	0.65	0.49
		(1.45)	(1.54)	(0.88)	(0.57)
LYS	(7)	0.43	0.48	0.33	0.1
		(0.53)	(0.66)	(0.4)	(0.13)
Total	(47)	0.71	0.73	0.49	0.22
		(1.03)	(1.04)	(0.67)	(0.36)
P-value	(1-tailed)	0.000*	0.004*		
P-value	(2-tailed)	0.001*	0.012*		

Table 3.6: Performance benchmark of DEPTH, MM-SCP and DEMM. Table shows mean absolute errors and R.M.S.D. (in brackets) in pH units of the prediction for different amino-acids.

many protein properties, such as folding [159], stability [160, 161], solubility [162], dynamics [163], interactions [164] and functions [165, 166]. In proteins, pK_a of an ionizable amino acid residues (ASP, GLU, HIS and LYS) is sensitive to its environment. In this collaboration work, we have built a pK_a value predictor by utilizing depth as a primary feature to concisely describe the protein environment. We also found that our predictor synergize well with another physics-based method of micro-environment modulated screened Coulomb potential approximation (MM-SCP) [167]. Hence a meta-predictor DEMM combining these two methods are constructed. The predictor takes the form of a multivariate linear regression model, with features consisting main-chain atom depth, polar side-chain atom depth, hydrogen bonds and Coulombic interaction. The model was trained using a dataset of 175 residues and tested on a separate set of 47 residues. The testing result showed that while the root mean squared deviation (R.M.S.D.) of DEPTH and MM-SCP predictor are both ~ 1 pH units away from experimentally determined values, in DEMM this error was reduced to ~ 0.67 pH units (Table 3.6). This performance is by far the most accurate computational method (apart from much more intensive quantum simulation) for protein pK_a prediction.

3.5 Chapter Summary & Discussions

In this chapter, we first introduced the biophysical parameter of depth. Depth measures residue burial measure in protein, and is calculated as the distance of an atom / residue to the nearest bulk solvent water. The computation of depth is elucidated, particularly on the differentiation of bulk and non-bulk waters in protein structures.

Conventionally, the measure of solvent accessibility has been used to quantify protein-solvent interaction. The measure quantifies physicochemical environments on protein surface, and has been shown useful in applications of estimating the strength protein-protein interaction and other binding events. However, when describing the physicochemical environment in protein interior, residue depth is a simple yet more informative measure. Depth has a wider dynamic range in protein interior and hence could more concisely describes the physicochemical environment in the protein interior. Several physical properties of proteins have been established to correlate well with depth. In this chapter, we attested the capability of depth by using it to construct several par excellence protein biochemical properties and function predictors.

First, residue depth provides a gradually stratified profile of residue/atomic burial, and is capable of identifying the deepest buried residues. These buried locations are the most sensitive to thermal stability of protein. A temperature-sensitive design protocol was constructed by mutating hydrophobic residues in this region to a biochemically diverse set of amino acids. Currently, the set consists of 5 amino acids (Ala, Trp, Asn, Asp or Pro). While in all our test cases the set worked successfully, these suggestions were disregard of the underlying mechanism for temperature sensitivity. The factors to sensitivity are complex and could include the rate of protein synthesis, the susceptibility to proteolysis, interaction with other regulator proteins such as chaperones and ubiquitins. It is suggested a more detailed investigation into these mechanisms could further reduce the mutations needed to generate a temperature sensitive mutant of a specific protein.

In the second application, we noticed that in general, residue depth values show a trend that is the inverse of SASA values. Exceptions to this trend are the protein ligand binding cavities, which are simultaneously deep and solvent exposed. By quantifying these relationships for different amino acids, a binding site prediction protocol is proposed. The pure geometrical and the evolutionary information incorporated version of our predictors are on par with the best methods of the categories respectively. Additionally, DEPTH method could find its niche applications in detecting binding sites at protein interfaces, as well as geometrically flat binding sites with simple adjustment on parameters. Moreover, we have also noticed that the two best predictors of DEPTH and ConCavity are making substantially different predictions of binding site residues. This is indicative of possible further improvement in the method by synergizing the strengths of the two methods.

Lastly, we briefly discussed two other depth-related prediction protocols developed in collaborations, namely the estimation of cavity volume in protein interior and the pK_a prediction of ionizable residues in proteins. The depth computation protocol has explicitly defined cavity waters from the bulk solvent. The clustering of these waters define a cavity, and through the use of Voronoi method, its volume can be computed. We have demonstrated the cavity volume estimated using DEPTH algorithm correlates the best with experimentally determined destabilization free energy as compared to other methods. One application of this protocol could be to design cavity-filling mutations to structurally stabilize proteins. Lastly it is demonstrated that by incorporating the depth measure and other simple biophysical parameters, an accurate pK_a predictor can be easily constructed in the form of a multivariate regression model. This predictor was found cooperative to another physics-based method of MM-SCP. In accordance, we built a meta-algorithm DEMM combining the two methods. DEMM has a record-breaking accuracy with error at only around half a pH unit. This is close to the accuracy of quantum simulation and should suffice most if not all biological purposes.

All the protocols described in this chapter have been made accessible through a web-server to reach a wider and appropriate group of audiences. Given the ease of computation and the wide utilities of the depth measure, we also proposed that

it should be included as a standard measure in all structural studies related to proteins and their functions.

Chapter 4

Characterization of Intramolecular Hydrogen Bonding in Proteins

Hydrogen bonding is an important molecular interaction in proteins. As a short-ranged and strong non-bonded interaction, hydrogen bonding contributes substantially to the stability of protein local environments and is pivotal for protein functions. Features of hydrogen bonds are environment-dependent. This chapter describes the characterization of hydrogen bond in different protein environments, with focus on its geometrical distribution and variations, as measured by donor-acceptor distance and donor-acceptor-antecedent angle. Hydrogen bonds are empirically classified based on the chemical properties of donor-acceptor pairs, as well as their locations in protein structures. By statistical analysis of high-resolution structural data, we have devised and updated parameters describing hydrogen bonding parameters in proteins. Further, we have also performed first principle quantum simulation and explained the distribution of the hydrogen bond geometrical features. This chapter describes the rationales and methods investigating hydrogen bonding in a large protein database. Knowledge extracted from this database study would have various utilities in protein biological function annotation and predictions. These utilities will be discussed in the next chapter.

4.1 Background and Motivations

4.1.1 Properties and importance of hydrogen bonding in proteins

Hydrogen bonding is an attractive force between an electronegative acceptor atom and a hydrogen atom, which is in turn covalently bound to an electronegative donor atom [7, 8]. It is prominent among non-covalent interactions as it is directional, atom-specific and acts in the length scale of $\sim 2.5 \text{ \AA} - 3.5 \text{ \AA}$. In protein, the strength of hydrogen bonding energy is estimated to be around $\sim 1.3 - 8 \text{ kcal/mol}$ [3, 168] depending on the environment. The physical nature of hydrogen bonding is complex. While it is largely electrostatic in nature, it also possesses some covalent bonding features [169, 170]. It has been established that in proteins, hydrogen bonds contribute to structural integrity and stability [171–174], folding [175–179], pK_a s of ionizable amino acid residues [89, 180], atomic packing [181] etc.

4.1.2 Previous work

Given its importance in proteins, much effort has been spent in characterizing the energetics and stereochemistry of hydrogen bonding [5, 172, 179, 182–188]. Recently, it has been increasingly recognized that the strength of protein hydrogen bonds is context-specific, and could vary with protein environments such as the polarity, contact order of donor and acceptor residues, temperature etc [189–191]. Some experimental evidences also suggested that hydrogen bonding could also indirectly stabilize proteins by enhancing van der Waal’s interaction, explaining the increased packing density in the interior of proteins [192–194]. The parameters describing hydrogen bonds in proteins can be learned from statistical analyses of known protein structures. Such analyses have been extensively documented earlier and have provided crucial insights into protein structural stability and function [195–204]. However, these studies were performed over small datasets (e.g. $n = 15$ in the study by Baker and Hubbard) of known structures available then. The number of protein 3D structure deposited in the protein data bank (PDB) has since

been growing exponentially [18]. With much more data, comprehensive analyses of hydrogen bonding and at a high-resolution resolution is possible. This is in addition to the fact that we have also developed different means of quantifying protein environments as described in previous chapters. Here we aim to revamp the descriptors of various types of hydrogen bonding with statistical analyses of protein environments. Our analyses supplement and update older studies of hydrogen bond parameterization [195–204].

4.2 Geometry of Protein Hydrogen Bonds

4.2.1 Data sets and parameters

561 high-resolution (resolution 1.7 Å or better; R-free 0.2 or better), non-redundant (at 30% sequence identity), single domain protein structures (chain lengths between 80-180 amino acids) were extracted from the PDB to form the data set. The data set consists of a total of 67563 hydrogen bonds.

Hydrogen bonds are defined with three atoms in our study, namely a pair of electronegative donor and acceptor atoms, and closest atom covalently bonded to the acceptor atom (also known as “acceptor-antecedent atom”) (Figure 4.1). The donor atom must covalently bond to a hydrogen atom, which in turn interacts with the acceptor atom. However, in most of the X-ray crystallography structures, the resolution is not high enough to discern the precise position of the hydrogens. Hydrogen atom is hence not used in our method to detecting a hydrogen bond. All nitrogen atoms in the main chain amino group are competent donors, except for Proline which does not have a hydrogen. All oxygen atoms in the main chain carbonyl group are competent acceptors, with the carbon atom in the carbonyl group as their antecedent atoms. For side chain, the competent hydrogen bond donors, acceptors and acceptor antecedents are listed in Table 4.1.

In this study, we limit our scope to conventional hydrogen bonds listed as above. Other polar/dipole interactions, such as salt-bridge, CH–O interaction and cation- π interaction, are much less frequent than the conventional hydrogen bonds in protein, and are not included in this study.

Figure 4.1: Atoms involved in defining the geometry of a hydrogen bond

Amino acid	Donor atom	Acceptor atom
ARG	NE, NH1, NH2	
ASN	ND2	OD1 (CG), OD2 (CG)
CYS	SG	
GLN	NE2	OE1 (CD)
HIS	ND1, NE2	ND1 (CD2), NE2 (CD2)
LYS	NZ	
SER	OG	OG (CB)
THR	OG1	OG1 (CB)
TRP	NE1	
TYR	OH	OH (CZ)
GLU		OE1 (CD), OE2 (CD)

Table 4.1: Donor, acceptor and acceptor antecedent atoms (in parenthesis) involved in side-chain hydrogen bonds.

Two parameters were used to quantify the geometry of a hydrogen bond, namely the donor-acceptor distance and the donor-acceptor-antecedent angle, which will be dubbed as " $\hat{H}$ angle" hereinafter.

4.2.2 Overall distribution of donor-acceptor distance and $\hat{H}$ angle

We first investigated the overall geometrical preferences of hydrogen bonds in proteins. From the high-resolution, single domain data training set, for all donor-acceptors pairs within a cut-off of 4.0 Å, the statistics of their donor-acceptor distances and $\hat{H}$ angles were collected. These observations were compared to a

reference state model where there was no distance or angular preference due to hydrogen bonding. In the reference state, the atoms were randomly distributed as if they were non-interacting entities, akin to molecules of an ideal gas. The reference state can be modelled with:

$$N^{exp}(r, \theta) = N_0 \cdot 4\pi r^2 \cdot dr \quad (4.1)$$

where $N^{exp}(r, \theta)$ is the expected number of donor-acceptor pairs at a distance r and $\hat{H}$ angle θ , and dr is the bin size for collecting statistics. We used a bin size of 0.09 Å here (20 bins spanning the range 2.3 – 4.1 Å). N_0 is a normalizing constant represents the background density of donor-acceptor pairs at a certain cut-off distance. As we assumed a uniform density distribution in the reference state model, the value can be calibrated at any large distance r_{cut} as

$$N_0 = \frac{N^{obs}(r_{cut})}{4\pi r_{cut}^2 \cdot dr} \quad (4.2)$$

where $N^{obs}(r)$ is the observed number of donor-acceptor pairs at the cut-off distance r_{cut} . We set r_{cut} to a generous value of 4 Å (0.5 Å greater than conventional definition of donor-acceptor distance limit).

Comparison between the observations and reference state can be summarized with their log-odd ratios $\log(N^{obs}(r, \theta)/N^{exp}(r, \theta))$ and represented in a 2-D heatmap (Figure 4.2). From the comparison it was shown that the preferred donor-acceptor distance and $\hat{H}$ angle are concentrated at ~ 2.9 Å and $\sim 155^\circ$ respectively. The $\hat{H}$ angle has a heavy-tailed distribution extends to the less obtuse angle region.

The composition of hydrogen bonds at different geometries is further discussed in the following sections.

4.2.3 Classification of hydrogen bonds and environments

4.2.3.1 Hydrogen bond criteria

From the observed preferences, we empirically redefine hydrogen bonding criteria based on a cut-off log-odd value of 4.0 or higher. This value is chosen as it corre-

Figure 4.2: The average geometry ($\hat{H}$ angle – donor-acceptor distance) of hydrogen bond in proteins is depicted with a heat-map of normalized log-odd of occurrence compared to background distribution $\log(N^{obs}/N^{exp})$. A 0.05 Å and 1° standard deviations for bond distance and $\hat{H}$ angle was applied for Gaussian blurring of the frequency counting for plotting.

sponds to a maximum donor-acceptor distance of 3.5 Å, which is the sum of van der Waal’s radii of donor and acceptor atom with a buffer of ~ 0.5 Å. Only additional energy contributed by an interatomic interaction can enable two atoms to come closer than this distance - suggesting hydrogen bonding at play. Based on this cut-off value, we define a hydrogen bond to have the donor-acceptor distance less than 3.5 Å and $\hat{H}$ angle lying between 100 – 180°.

We note that the lower bound of our $\hat{H}$ angle is relaxed from the conventional value of 120° [195]. We believe that this expansion in the definition is a better description of hydrogen bonds. For instance, we have reviewed the hydrogen-deuterium exchange experiments in hen egg white lysozyme (HEWL) [205]. In H/D exchange experiment, amide protons in the protein are continuously being replaced by deuterium atoms in the bulk solvent. The exchange rate for different amide protons are different. Amides in the protein interior is protected from solvent and have slower exchange rates, while the exchange rate on protein surface is fast. The exception to this is when the amide group is involved in a hydrogen bonding. These

Residue	Exchange rate	Depth	Exposed	Bond-type	Bond-length	$\hat{H}$ angle
78	3.62E-04	4.20	Yes	MC-SC	2.99	111.7
42	2.56E-04	4.47	Yes	MC-MC	3.27	118.5
23	3.09E-03	4.59	Yes	MC-MC	3.04	119.3
41	1.28E-02	5.47	Yes	MC-SC	2.97	117.9
39	5.50E-05	5.60	No	MC-SC	3.42	101.5
111	4.07E-04	6.69	No	MC-MC	3.04	118.1
57	1.79E-06	7.03	No	MC-MC	3.36	113.7
28	4.39E-07	8.62	No	MC-MC	3.05	119.3

Table 4.2: List of hydrogen bonds with $\hat{H}$ angle less than 120° in Hen Egg White Lysozyme (HEWL). Depth is measured in ($\text{\AA}$) unit, and $\hat{H}$ angle is measured in ($^\circ$) unit.

amides are also protected, as the exchange process could not proceed without first breaking the hydrogen bonds. In HEWL, the residues 23, 41, 42 and 78 all consist of surface amides with slow exchange rates, hence strongly suggesting the existence of hydrogen bondings. However, no hydrogen bonds could be detected with the conventional $\hat{H}$ angle cut-off of 120° for these residues. Only by relaxing the cut-off value to 100° the relevant bonds can be identified (Table 4.2). Further supporting evidences for expanding the $\hat{H}$ angle definition from small molecule databases study will also be presented in section 4.4.1.

4.2.3.2 Protein environment parameters

Hydrogen bonds are empirically classified into 4 types based on whether the donor/acceptor atoms are from the main-chain/side-chain: (i) donor: main-chain; acceptor: main-chain (MC-MC), (ii) donor: main-chain; acceptor: side-chain (MC-SC), (iii) donor: side-chain; acceptor: main-chain (SC-MC) and (iv) donor: side-chain; acceptor: side-chain (SC-SC). MC-MC hydrogen bonds are further divided into 3 subtypes based on secondary structure: (i) α -helix (MC-MC-A) (ii) β -strand (MC-MC-B) and (iii) coil (MC-MC-C).

Other parameters used include (i) atomic depth (2) secondary structure and (3) partial charge of atoms. The values of partial charge were extracted from the GROMOS96 force field [71].

4.2.4 Hydrogen bond geometry in different protein environments

4.2.4.1 Type-specific donor-acceptor distance and $\hat{H}$ angle

Geometrical preference of hydrogen bond of 4 different types (MC-MC, MC-SC, SC-MC and SC-SC) has been investigated. Different hydrogen bonds have slightly different preferred donor-acceptor distances and $\hat{H}$ angles, and are distinct from one another. On $\hat{H}$ angle, MC-MC bonds typically occur with the highest values of $\sim 155^\circ$. The exception occurs when they are exposed on protein surface, and the mean $\hat{H}$ angle drops to $\sim 115^\circ$. The mean values of $\hat{H}$ angle of other hydrogen bond types, MC-SC, SC-MC and SC-SC are $\sim 130^\circ$, $\sim 135^\circ$ and $\sim 125^\circ$ respectively. These preferences do not change with change in residue depth (Figure 4.3 A). On donor-acceptor distances, while the preferred values of different hydrogen bond types are distinct from one another, all of them showed a tendency to decrease as they occur in increasingly buried environments (Figure 4.3 B). The MC-MC bonds and MC-SC bonds are the longest bonds (ranging from ~ 3.1 Å when exposed to ~ 2.9 Å when buried) while the SC-SC bonds are the shortest (~ 2.9 Å - ~ 2.8 Å). To test the statistical significance of the difference in geometries (both $\hat{H}$ angle and donor-acceptor distance) in the different hydrogen bond types, the hydrogen bonds were first segregated into different depth regions. Depth values ranged from 3.0 Å to 10.0 Å and were binned at intervals of 0.5 Å. Statistical significance was established using a Welch's t-test in each of the bins. Most of the differences in donor-acceptor distances and $\hat{H}$ angles between different types of hydrogen bonds were statistically significant ($p < 0.05$) (Table 4.3). Some exceptions were at the protein surface (depth < 4 Å) where there are large structural/geometrical fluctuations. Statistical differences could also not be detected in the protein core (depth > 8 Å) primarily because of sparseness of data in the region.

4.2.4.2 Donor-acceptor distance and $\hat{H}$ angle in secondary structures

We investigated the environmental variation of MC-MC type hydrogen bonds as they are crucial to the formation of secondary structures. MC-MC hydrogen bonds

Figure 4.3: The geometry ((A) donor-acceptor distance and (B) $\hat{H}$ angle) of different hydrogen bond types at different solvation environments measured by depth. For a description of the different types of hydrogen bonds, MC-MC (blue), MC-SC (green), SC-MC (red), SC-SC (cyan), refer to the methods section. The size of the data points is proportional to the number of observations.

DEPTH	I-II	I-III	I-IV	II-III	II-IV	III-IV
3.25	0.00*	0.00*	0.00*	0.01*	0.60	0.00*
3.75	0.00*	0.00*	0.00*	0.00*	0.03*	0.00*
4.25	0.00*	0.03*	0.00*	0.00*	0.00*	0.00*
4.75	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
5.25	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
5.75	0.00*	0.00*	0.00*	0.00*	0.01*	0.00*
6.25	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
6.75	0.00*	0.00*	0.00*	0.00*	0.01*	0.00*
7.25	0.00*	0.00*	0.00*	0.02*	0.00*	0.00*
7.75	0.00*	0.00*	0.00*	0.01*	0.11	0.00*
8.25	0.00*	0.00*	0.00*	0.00*	0.03*	0.00*
8.75	0.00*	0.00*	0.00*	0.02*	0.04*	0.00*
9.25	0.00*	0.00*	0.00*	0.00*	0.31	0.00*
9.75	0.00*	0.00*	0.00*	0.47	0.01*	0.00*

Table 4.3: Statistical significance of difference as measured in p -value in $\hat{H}$ angle in different classes of H-bonds (I: MC-MC, II: MC-SC, III: SC-MC, IV: SC-SC) at various levels of depth. * indicates the statistical significance at conventional cut-off value $p = 0.05$.

DEPTH	I-II	I-III	I-IV	II-III	II-IV	III-IV
3.25	0.00*	0.00*	0.00*	0.90	0.04*	0.00*
3.75	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
4.25	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
4.75	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
5.25	0.56	0.00*	0.00*	0.00*	0.00*	0.00*
5.75	0.12	0.00*	0.00*	0.00*	0.00*	0.00*
6.25	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
6.75	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
7.25	0.00*	0.47	0.00*	0.00*	0.00*	0.00*
7.75	0.02*	0.73	0.00*	0.05	0.00*	0.00*
8.25	0.01*	0.02*	0.00*	0.00*	0.00*	0.29
8.75	0.00*	0.38	0.35	0.09	0.00*	0.19
9.25	0.02*	0.14	0.00*	0.01*	0.00*	0.00*
9.75	0.06	0.84	0.00*	0.15	0.00*	0.00*

Table 4.4: Statistical significance of difference as measured in p -value in donor-acceptor distance in different classes of H-bonds (I: MC-MC, II: MC-SC, III: SC-MC, IV: SC-SC) at various levels of depth. * indicates the statistical significance at conventional cut-off value $p = 0.05$.

were sub-classified into α -helical, β -strand and others. The 3 sub-classes have slightly different geometry preferences (Figure 4.4 B). Surprisingly, $\hat{H}$ angles of both α -helical and β -strand both have the mean value of $\sim 155^\circ$, and do not vary with residue depth (except variation on protein surface). This is in spite of the very different structural constraints that could be exerted from the secondary structures. Hydrogen $\hat{H}$ angles in regions without secondary structure had a smaller $\hat{H}$ angle, but also tended toward $\sim 155^\circ$ with increasing depth.

Of the 3 sub-classes, donor-acceptor distance in β -strand are the shortest (Figure 4.4 A). On the surface (depth < 6 Å) they are slightly longer than 3.0 Å. This value converges to a little under 2.9 Å at a depth of ~ 6 Å and remains steadily at this value for all higher depths. The donor-acceptor distances of helices and non-secondary structures converge to a value ~ 2.95 Å only at depth of about 8 Å. The donor-acceptor distances of these 2 subclasses at the surface are ~ 3.0 Å and ~ 3.2 Å respectively.

The deviations from the mean hydrogen $\hat{H}$ angle are greater at the surface (depth < 6 Å) than in the interior (depth > 6 Å) for β -sheets ($\sim 15.5^\circ$ on surface, $\sim 9.1^\circ$ in

Figure 4.4: The donor-acceptor distance (A) and $\hat{H}$ angle (B) of different MC-MC hydrogen bond types at different depths. The hydrogen bond of the MC-MC type includes alpha helix (blue), anti-parallel beta-strand (purple), parallel beta-strand (orange), beta-strand excluding GLY and termini (green), non-secondary structure (red), and non-secondary structure excluding (i , $i+3$) hydrogen bond (brown). The size of the symbols are indicative of the data points sampled, normalized by colour.

protein interior). The deviation from the average at lower depths is less pronounced in cases where Glycine residues and terminal residues (1 residue at the N and C termini of strands) are not included in the computation. The standard deviation of hydrogen $\hat{H}$ angles in α -helices and non-secondary structure regions are $\sim 8.8^\circ$ and $\sim 15.5^\circ$ respectively and do not change with depth. Interestingly, hydrogen bonds in non-secondary structures are seldom found at depths > 7 Å.

4.3 Structural and Functional Implications Hydrogen Bond Geometry

4.3.1 Secondary structure formation

It is interesting to note that the $\hat{H}$ angle in both α -helix and β -strand converges to the same value of $\sim 155^\circ$. To investigate the potential causal link between $\hat{H}$ angle variation and secondary structure formation, we have simulated the scenarios of model secondary structures at different $\hat{H}$ angle values. The following sections detail the simulation system and the results obtained.

4.3.1.1 Restraining $\hat{H}$ angles in α -helices

A set of simulations was performed to fold a 24-mer poly-Alanine peptide from an extended conformation to an α -helix. The simulations were carried out using the `refine_slow` option of Modeller [29] that optimizes structure by satisfying spatial restraints. All the $(i, i+4)$ H-bonding pairs of amides and carbonyl oxygens in an α -helix were restrained to donor-acceptor distance of 2.9 Å with a standard deviation of 0.2 Å. In each simulation, the corresponding $\hat{H}$ angles were restrained to one of 11 different values ranging from 100° to 180° in steps of 10°, and the values 125° and 155°. In addition to the H-bond restraints, stereochemical restraints of all amino acids were appropriately added while considering L-J and electrostatics over the entire system.

α -helices are successfully generated in 7 of the 11 restraining attempts. When the helices are formed, the hydrogen $\hat{H}$ angle lies in a narrow range of 145-160°, regardless of the initial restraining angles. In the 4 cases that α -helices were not formed, the restraints caused several atomic clashes in the system (violation of the Ramachandran map) (Figure 4.5). Here, an atomic clash is defined when two atoms falls within a distance of 0.25 Å less than the sum of their van der Waal's radii.

4.3.1.2 Restrained Molecular Dynamics Simulation on α -helix

A 16-mer poly alanine α -helix was constructed using the secondary structure module in Modeller. The average ϕ and ψ angles of all the residues in the starting structure were -64° and -42° respectively. The average value of the $\hat{H}$ angle of the $(i+4, i)$ hydrogen bond in starting structure was $\sim 154^\circ$. The helical peptide was placed at the center of a dodecahedron box such that the distance from the peptide to the walls of the box was at least 10 Å. The simulation set up consists of 7228 atoms, including 7065 SPC216 water molecules.

Three sets of 10 nanosecond (ns) molecular dynamic (MD) simulations were performed on the system described above. The temperature of the system is set by assigning a random velocity to every atom following the Maxwell-Boltzmann distribution. The system was energy minimized until the maximum force on the system

Figure 4.5: The Ramachandran map of the 24-mer polyAlanine with residue positions depicted as red open triangles in cases where α -helices were formed (mean angle $\hat{H} \sim 156^\circ$, mean number of hydrogen bonds formed $N \sim 19.3$) and not formed ($\hat{H} \sim 119^\circ$, $N \sim 3.5$; labelled in red). The orange, green, blue and white colors represent the favored, allowed, generously allowed and disallowed regions on the map.

was less than 1000 kJ/mole/nm. Before the 10 ns production run, the system was subjected to 100 picoseconds (ps) of canonical ensemble (NVT) simulation, followed by another 100 ps of isothermal-isobaric ensemble (NPT) simulation. The production run was 10 ns of simulation in the NPT ensemble at a temperature of 300 K using Berendsen thermostat coupling.

First, as a control, the helix was subjected to free MD for 10 ns without any external constraints. In the second 10 ns MD simulation, a +40 kJ/mol energy barrier at 155° was applied to all $(i+4, i)$ hydrogen bond $\hat{H}$ angles. This energy barrier was applied to force the system to explore $\hat{H}$ angles other than what it was in the initial structure. In the third simulation, the last frame of the second simulation was used as the starting structure. The energy barriers were removed, and instead a -40 kJ/mol energy well at 140° was applied to all $(i+4, i)$ hydrogen bond donor-acceptor pairs. This was done to increase the sampling rate at the $\sim 155^\circ$ region. All simulations and constraint manipulations were performed with the Gromacs (ver 5.1) software suite, using the CHARMM36 force field.

In the free simulation, all $(i+4, i)$ hydrogen bonds remain intact and the $\hat{H}$ angles are $\sim 170^\circ - 175^\circ$. In the second simulation, when there was an energy barrier placed at 155° , the $\hat{H}$ angle sampled the $110^\circ - 140^\circ$ region. But more importantly, the donor acceptor distances all exceeded 3.2 Å. In the third simulation the $\hat{H}$ angles all had values $\sim 155^\circ$ (Figure 4.7). Note that though the well of the potential in the third simulation had its minima at 140° , the $\hat{H}$ angles sampled higher values and mimicked our database observations.

In the first and third set of simulations, the structural integrity of the α -helix was preserved, as evidenced from the Ramachandran map (Figure 4.7). However in the second simulation where $\hat{H}$ angles had values less than 140° (on average 127°), the helix was observed to melt, with around half of all the hydrogen bonds breaking during the simulation. The dihedral angles ϕ and ψ , during most of the simulation time values from all three system fall within the allowed and generously allowed region of Ramachandran map. However, the different systems preferred distinctly different dihedral regions of the map. Dihedral angles in the second simulation are concentrated away from alpha helix region as a result of the melting structure.

Figure 4.6: Figure shows the $\hat{H}$ angle and donor-acceptor distance sampled in all three simulations on a 16-mer α -helix. A snapshot structure was taken every 100 ps along the simulation trajectory, and the average of $\hat{H}$ angle and donor-acceptor distance is represented as a data point in the figure. The data points from the first, second and third simulations are coloured blue, green and red respectively. Snapshots of the α -helix during the (top) first, (center) third and (bottom) second simulation are also shown in the figure. Hydrogen bonds are represented in yellow dashed lines.

Note that the region of the Ramachandran map in the free simulation has a larger spread of ϕ and ψ angle values than the simulations from set 3. This is perhaps indicative of the force field parameters that confine the donor-acceptor-acceptor antecedent atoms to a near-linear geometry. This discrepancy in the force field is perhaps also responsible for the helix to explore regions of the Ramachandran map that are slightly away from the center of the “ideal” helix location (Figure 4.6).

4.3.1.3 Sampling $\hat{H}$ angles in β -sheets

We next examined the range of allowed $\hat{H}$ angles that are allowed in either parallel and antiparallel β -sheets. Our model systems consisted of two 6-residue strands that formed parallel and antiparallel sheets in an almost planar configuration (the dihedral angle formed by the first and last C atoms on both strands $\sim 9.3^\circ$). For the parallel and antiparallel sheets, templates were extracted from PDB:3AUM

Figure 4.7: Dihedral angle propensity aggregated from the (A) second (B) first and (C) third set of simulation. A snapshot structure was taken every 100 ps along the simulation trajectory, and the ϕ and ψ angles of all residues in the structure are used. Gaussian kernel density estimation with bandwidth of 0.1° is applied to the data set for plotting purpose.

(residues 111-115, 121-125) and PDB:2PEC (residues 204-208, 226-230) respectively. To each strand, in both systems, an extra residue was engineered so as to maintain the relative planarity of the system.

To simulate sheets with different $\hat{H}$ angles, we sheared one strand with respect to other by $3.5 \text{ \AA}$, in steps of $0.1 \text{ \AA}$, in two directions so as to maintain the overall planarity of the sheet. In addition to shearing along the length of the sheet, the two strands were also laterally separated (from $-0.75 \text{ \AA}$ to $0.75 \text{ \AA}$ in steps of $0.1 \text{ \AA}$) so as to sample different H-bond distances. After every step, we computed the H-bond donor-acceptor distances, $\hat{H}$ angles and atomic clashes in the 4 central residues of each strand.

In both sets of shearing simulations, clash-free sheets were formed when $\hat{H}$ angle lies within a narrow range of $\sim 10^\circ$ with a mean close to $\sim 155^\circ$ (Figure 4.8), with the parallel system having a slightly more permissive upper bound. Interestingly, the lower bound for the $\hat{H}$ angle was $\sim 150^\circ$ regardless of donor-acceptor distance.

4.3.2 Donor-acceptor distance in secondary structures

(A) Amino acid secondary structure preference

Figure 4.8: Allowed donor-acceptor distance and $\hat{H}$ angle configuration in a planar (A) antiparallel β -sheet and (B) parallel β -sheet generated from simulations. Structures with fully formed H-bonds and steric clashes are indicated with green, and red points respectively.

Our investigation showed that, the mean donor-acceptor distances in α -helix and β -sheet had a small, but statistically significant difference. We further substantiated this result by correlating the donor-acceptor distance length of an amino acid with its secondary structure propensity [206]. We found that helical preference (helix/sheet ratio) has a correlation coefficient of ~ 0.45 (4.9).

(B) Difference in protein melting temperature

Earlier studies have established that shorter donor-acceptor distance implies a stronger hydrogen bonding interaction [8]. We postulate and test the hypothesis that proteins with β -sheet would tend to have higher stability, as measured by their melting temperatures, as compared to α -helical proteins.

Experimentally determined melting temperature data were collected from the Protherm database [207], with only wild-type proteins (i.e. no mutations) were chosen. As the pH value of buffer has a substantial effect on melting temperature, we choose only thermal denaturation data from pH range of 6 - 8. These data corresponds to a total of 326 structures with 3D structure solved and deposited in the PDB. The dataset consists of 53 all- α proteins, 46 all- β proteins and 227 others.

Figure 4.9: MC-MC hydrogen bond length of different amino acid types and their secondary structure preference. Y-axis indicates propensities difference ($P(\text{helix})/P(\text{sheet})$) of a particular amino acid residue.

The melting temperature of all- α proteins is 58.1°C ($\pm 13.0^{\circ}\text{C}$) and all- β proteins have a slightly higher mean value at 62.5°C ($\pm 14.2^{\circ}\text{C}$). Primary examination using independent t-test showed that the difference was not significant ($p \sim 0.12$). However we observed that all- β protein has a heavy-tailed distribution and a second peak at $\sim 70^{\circ}\text{C}$ (Figure 4.10A). This suggested the use of a non-parametric statistical is more appropriate. A Fisher's exact test was hence performed, and the result showed the association of all- α /all- β and heat resistance (using threshold temperature of 70°C) is indeed statistically significant ($p < 0.05$) (Table 4.5). In this test, we have also shown that the density of hydrogen bonds (hydrogen bond per residue) is not correlated with melting temperature, for both all- α (corr. coef ~ 0.03) and all- β proteins (corr. coef ~ -0.03) (Figure 4.10B).

4.3.3 Correlation with partial charge

Earlier studies have established that shorter donor-acceptor distances imply stronger binding [8]. As electrostatic interaction is one major energetic component to the hydrogen bond, we suspect that the donor-acceptor distance could be influenced

Figure 4.10: Comparison of distribution of melting temperatures of all- α (red) and all- β (blue) proteins in Protherm database. (B) Correlation between hydrogen bond density (number of hydrogen bonds per residue) and melting temperature. Data points from all- α and all- β proteins are coloured red and blue respectively.

	$< 70^{\circ}C$	$\geq 70^{\circ}C$	Total
all α	45	8	53
all β	29	17	46
Total	74	25	99

Table 4.5: Fisher's exact test on heat-resistant (melting temperature $>70^{\circ}C$) with all- α / all- β protein. p-value of the test ~ 0.02

Figure 4.11: Correlation between donor-acceptor distance and partial charge on donor hydrogen (blue), acceptor oxygen (red) and total partial charge (green).

by partial charges on the atoms. To investigate this relationship in proteins, we collected the statistics on donor-acceptor distances for all 70 types of hydrogen bonds formed with different combinations of donor and acceptor atom types. The mean distances were correlated with partial charges of the donor (represented with hydrogen bond), acceptors as well as total charge (product of donor and acceptor partial charges). The result showed that donor-acceptor distance is best correlated with partial charge on donor (Pearson correlation coefficient $r \sim 0.78$) or total charge ($r \sim 0.76$), but not with acceptor atom ($r \sim 0.11$) (Figure 4.11). While the result is not intuitive, it is in agreement with a previous quantum mechanical study on hydrogen bonds [208].

4.4 Explanation for hydrogen bond geometry preferences

Statistical analyses performed over the Protein Data Bank have provided us with environment-dependent geometrical parameters describing hydrogen bonds in proteins. These parameters, while utilitarian, are however empirical results specific to single domain globular proteins. The validity of the parameters cannot be warranted when extended to other systems, such as protein folding, protein-protein in-

teractions, protein-ligand interactions etc. To gain further insights into the physicochemistry of the geometrical preferences, we have attempted two complementary approaches - (a) by comparing to experiment data of small molecules and (b) by performing quantum chemical calculations on test systems that mimicked hydrogen bonded peptide units. The results of the investigation are detailed in the following sections.

4.4.1 $\hat{H}$ angle distribution in small organic molecules

To investigate the hydrogen bond geometry when it is devoid of structural and environment constraints in proteins, we made an extensive survey on other small organic molecules deposited in the Crystallography Open Database [209, 210] for statistics on donor-acceptor distance and $\hat{H}$ angle. For appropriate comparison, only hydrogen bonds constituted by a nitrogen donor, an oxygen acceptor and a carbon acceptor antecedent were considered. We also required that the hydrogen atom be explicitly present for a hydrogen bond to be detected. 2886 hydrogen bonds from 1922 small molecules were found in the dataset. From the distribution of hydrogen $\hat{H}$ angle (Figure 4.12 A) we made two observations:

- (1) The mean $\hat{H}$ angle in small organic molecule is $\sim 125^\circ (+ - 16.5^\circ)$. This value is similar to hydrogen $\hat{H}$ angle of SC-SC hydrogen bonds in protein.
- (2) More than one third of hydrogen bonds would have been missed if an $\hat{H}$ angle threshold of 120° was used. This is contrast to $\sim 3\%$ if a threshold of 100° should be used.

Next, we similarly surveyed the Cambridge structural database [211] for MC-MC hydrogen bonds in dipeptides and tripeptides. A total of 22 hydrogen bonds were found in 12 structures (Figure 4.12 B). It was observed that hydrogen $\hat{H}$ angle also has a mean value of $155^\circ (+ - 6.5^\circ)$, similar to proteins deposited in the Protein Data Bank. It was observed that $\hat{H}$ angles of hydrogen bonds involving GLY residue typically are $\sim 140^\circ$ or smaller.

Figure 4.12: Histograms of hydrogen $\hat{H}$ angle in (A) small organic molecules (B) dipeptide and tripeptides.

4.4.2 Quantum simulation

Collectively, the results from database study presented in the preceding sections had suggested that (1) the $\hat{H}$ angle preference of hydrogen bond in protein is different from other small organic molecules, and (2) peptide bond between amino acids could play an important role determining the preference. In the next sections, we attempt to quantitatively decipher the preference of $\hat{H}$ angle from first principle quantum mechanical simulations. Specifically, we have performed quantum mechanical calculations on small-molecule models of proteins with two objectives (1) to compare trends of the calculated $\hat{H}$ angle with those obtained from the statistical analysis, and (2) to rationalize the differences in $\hat{H}$ angle among different types of hydrogen bonds.

4.4.2.1 $\hat{H}$ angle preference of different hydrogen bond types

As a first attempt we have used simplified systems to model different types of hydrogen bondings (MC-MC, MC-SC, SC-MC, SC-SC). For main-chain mimetic, we chose the widely used N-methylacetamide (NMA)[212, 213], and for side-chain mimetic, a methanol molecule is used. The simplicity of this system allows for more detailed simulation. We performed geometry optimization of the dimer using second

Figure 4.13: (a) Optimized structure of NMA dimer which is a model for MC-MC type of hydrogen bond and (b) the saturated planar structure which is a model for the SC-MC type of hydrogen bond. Hydrogen bond is represented as a dotted line in both cases.

Type	Simulated ($^{\circ}$)	Observed ($^{\circ}$)
SC-SC	100	~ 125
MC-SC	112.5	~ 130
SC-MC	112.8	~ 135
MC-MC	137.9	~ 155

Table 4.6: Comparison of $\hat{H}$ angles of different types of hydrogen bonds between quantum simulation using N-methyl-acetamides (NMA) and actual observation in Protein Data Bank.

order Mller-Plesset perturbation theory (MP2) with the 6-31G(d, p)++ basis set. All the calculations were performed using the GAMESS program package [214] .

As a common artifact to local optimization techniques, the optimization process could be trapped in local minimum, and hence the optimized structure be dependent on the choice of the initial structure. To ensure adequate sampling of the coordinate space, we have performed six different simulations. In each simulation, $\hat{H}$ angle is first constrained to a different value (from 85° to 160° in steps of 15°), and the rest of the molecule was optimized. All constraints were then released to allow a full optimization. From the simulations, only the final structure that has the most linear donor-hydrogen-acceptor angle and a *trans*-peptide bond was selected for further analysis.

We have collected $\hat{H}$ angle values from the simulations and compared with their corresponding hydrogen bond types in proteins (Table 4.6). The simulated $\hat{H}$ angle

differ from observation by $\sim 10 - 25^\circ$. Nonetheless, it is noteworthy that their ranks matched with that of observed values (MC-MC > SC-MC > MC-SC > SC-SC).

4.4.2.2 $\hat{H}$ angle preference of main-chain main-chain hydrogen bonds

The previous section has suggested the simplified model might not be sufficiently robust and accurate to replicate $\hat{H}$ angles in actuality. Hence, for more detailed analysis of hydrogen bonds between peptides, we adopted the more realistic models of dimers of glycine dipeptide. For these larger systems however, obtaining exact solution with *ab initio* methods is not computationally feasible. Instead, we have used the approximation method of density functional theory (DFT) with the B3LYP exchange-correlation energy functional (with the same 6-31G(d,p)++ basis set) for the simulation. To prepare the initial conformation, the glycine dimers were oriented in an antiparallel manner similar to that in β -sheet. Dihedral (ψ , ϕ) angles at (140° , -120°) and (140° , -100°) were sampled.

For SC-MC hydrogen bond, we noticed that the major difference between main-chain and side-chain donor/acceptor is the adjacent peptide bond. A peptide bond is sometimes considered as a partial double bond, due to the resonance of electron between the nitrogen and oxygen atoms. To model the side-chain donor, we simply added two hydrogen atoms to saturate the C=O bond adjacent to the donor amino group. To prevent overly perturbing the system, the system was relaxed in three steps after the hydrogenation: (1) first, only the added hydrogen atoms were relaxed, and all other atoms were frozen; (2) the atoms on the donor molecule were then relaxed, while acceptor atoms were still kept frozen; (3) lastly, all atoms were allowed to relax.

Using the more realistic models, the quantum simulation has successfully reproduced the $\hat{H}$ as observed in proteins for both MC-MC (simulated $\hat{H} \sim 149.0^\circ$, observed $\hat{H} \sim 155^\circ$) and SC-MC (simulated $\hat{H} \sim 132.5^\circ$, observed $\hat{H} \sim 135^\circ$) hydrogen bonds (Table 4.7). This suggested our system is sufficiently detailed for further explanatory analyses.

Figure 4.14: Optimized structures of dimers of glycine dipeptide in anti-parallel β -sheet orientation optimized with two different sets of Ramachandran angles as models of MC-MC type of hydrogen bonds. (ψ, ϕ) in (a) are $(140^\circ, -120^\circ)$ and in (b) are $(140^\circ, -100^\circ)$. The positions where the hydrogen atoms are added to obtain models of SC-MC type of hydrogen bonds are marked with arrows.

	$\psi = 140^\circ, \phi = -120^\circ$		$\psi = 140^\circ, \phi = -100^\circ$	
	MC-MC	SC-MC	MC-MC	SC-MC
	Hydrogen bond parameters			
H-bond angle	170.2	174.1	175.7	170.6
$\hat{H}$	150.3	131.2	147.8	133.7
NO distance	2.93	3.04	2.93	3.13
NC distance	4.05	3.96	4.03	4.08
H-bond length	1.92	2.02	1.91	2.12
	Mulliken partial charges			
N of donor	-0.27	-0.33	-0.29	-0.29
H of donor	0.45	0.32	0.46	0.34
O of acceptor	-0.61	-0.57	-0.6	-0.56
C of acceptor	0.19	0.25	0.19	0.17

Table 4.7: Chemical properties of hydrogen bonds in the simulated system of pair of Glycine dimers.

4.4.2.3 Electron density distribution around MC-MC hydrogen bond

Of the hydrogen bonding parameters from the two systems (MC-MC and SC-MC), bond length and $\hat{H}$ angle are almost the same. The $\hat{H}$ angle however, underwent a marked decrease of 14°-19° from MC-MC to SC-MC case. To rationalize the decrease of $\hat{H}$ angle, we have computed the electron localization (often also known as “partial charge” distribution) of atoms around the hydrogen bond using the quantum chemical software suite MOLDEN [215]. Inspecting the electron density map, it is found that while the total partial charge on nitrogen does not change much from MC-MC to SC-MC case (Table 4.7), its spatial distribution is distinct in the two systems.

In MC-MC hydrogen bond, due to the resonance on the peptide bond, the three chemical bonds with N (-H, -C, -C α) are planar as in the case of sp^2 hybridization. Whereas in SC-MC bonding, all chemical bonds are saturated, in sp^3 hybridization. It is notable that in the former case, the electron density at p-orbital is almost equally dispersed below and above the plane (perpendicular to hydrogen bond and peptide bond), whereas in the saturated molecule, the p -orbital density is in the form of a lone pair and is more concentrated around the N (Figure 4.15).

We rationalize that the decrease $\hat{H}$ angle is due to the interaction between the concentrated negative charge and the acceptor C=O group. Here O is the partially negatively charged hydrogen bond acceptor and C is the partially positively charged acceptor antecedent atom. The two atoms hence formed an electric dipole. Coulombic interactions with the concentrated charge on N caused the dipole to rotate in such a way that the C moves towards the N due to attraction (4.04 Å to 4.02 Å), and O moves away from the N due to repulsion (2.93 Å to 3.09 Å). With hydrogen bond angle constrained to be linear (most energetically favoured orientation in optimized structure), the dipole rotation begets a decrease in $\hat{H}$ angle.

As a control, we have also prepared a system of SC-MC hydrogen bonding, but with an artificial planarity constraint exerted on the peptide bond. This system has the same number of atoms and electrons as the original SC-MC system, but the planarity constraint introduces a symmetry to the electron distribution.

Figure 4.15: Schematic representation showing the decrease in $\hat{H}$ -angle going from a MC-MC (left) to SC-MC (right) type of hydrogen bond.

In effect, the uniform distribution of electron density around N as in the case of MC-MC hydrogen bond is reproduced. The optimization result showed that the optimized $\hat{H}$ angle ($\sim 153^\circ$) is again close to $\sim 155^\circ$.

4.5 Hydrogen bond number in protein models

4.5.1 Scalability of hydrogen bond number with chain length

Several studies that looked at hydrogen bonding in proteins led to the “hydrogen bonding hypothesis”. The hypothesis postulates that all hydrogen bond donors and acceptors in proteins are either paired with one another or with water [196, 197, 216, 217]. One conjecture from the hypothesis is that the number of hydrogen bond should scale approximately linearly with the protein chain length. In fact, it has been reported that the total number of hydrogen bond scales with number of amino acid residues at a constant factor of 0.71 [197]. We proceed to verify this postulation by performing linear regression on number of hydrogen bonds and chain length with two datasets:

1. High-resolution single domain data set

561 high-resolution (resolution 1.7 Å or better; R_{free} 0.2 or better), non-redundant (at 30% sequence identity), single domain protein structures (chain lengths between 80-180 amino acids) were extracted from the PDB to form the data set (this is the data set as described in Chapter 5).

2. Relaxed resolution single domain data set

A lower resolution data set (2.2 Å or better; R_{free} of 0.25 or better) of non-redundant (at 30% sequence identity) protein chains was similarly constructed. This data set contains 3400 proteins with a lower- and upper-limit on chain length of 80 and 1290 amino acids.

In the first single domain dataset, a fitted relationship of $y = 0.83x - 12$, with goodness of fit 0.71 was obtained, while in the second dataset, a slightly modest number of $y = 0.80x - 8$ was found (Figure 4.16). The two equations agree closely for

Figure 4.16: The relationship between hydrogen bond numbers and protein size. Hydrogen bonds were computed in 3400 protein structures whose sizes ranged from 80 to ~ 1300 residues. The same computation was repeated over a higher resolution dataset of high-resolution single domain proteins of sizes 80-180 residues (inset). The data were fit to straight lines (shown in red).

single domain protein (for example, for a standard 150 a.a. protein, first equation gives $150 \times 0.83 - 12 = 112.5$ and second equation gives $150 \times 0.80 - 8 = 112.0$). The slight decrease in number for larger proteins could be because the second dataset contains protein structures of lower resolution and longer chain lengths which could form multiple domains. The linear regressions were also attempted on different subtypes of hydrogen bonds, but none has shown better fitting (MC-MC: 0.53, MC-SC: 0.35, SC-MC: 0.25, SC-SC: 0.20) than total number of hydrogen bonds ($R^2 = 0.71$). This implies that the linear scaling relationship only applies when considering all possible hydrogen bond including main-chain and side-chain donors/acceptors.

In conclusion, we have verified the linear scalability of hydrogen bond with protein chain length. And by including proteins of long chain length, we have ascertained that there is no apparent limit of the range of validity to this relationship.

Figure 4.17: The number of hydrogen bonds per residue in an X-ray protein structure as a function of its resolution. The plot, consisting of 5440 points (one per protein) was smoothed using a moving average window of size 500. The curve was fit to a line (shown as red dashes) at resolutions less than 2.1 Å. The line $y = -0.09x + 1.0$ has a goodness of fit of 0.995.

4.5.2 Hydrogen bond number in low-resolution X-ray structures

The linear scalability relationship is violated, when we examine protein structures of low-resolution. Specifically we have tested the relationship with the following dataset:

3. Dataset of non-redundant crystal structures

A non-redundant set (30% sequence identity cut-off) of 5440 single domain (chain length 120-220 amino-acid) X-ray crystallography structures with resolutions range between 0 Å – 5 Å were obtained from PDB.

Until 2.1 Å resolution, the average number of hydrogen bonds per residue is almost a constant at around 0.82. At resolutions lower than 2.1 Å, the average decreases monotonically (Figure 4.17). For every 1 Å lowering of the resolution there are ~12% fewer hydrogen bonds in the protein structures.

To reinforce this finding, we made another dataset of protein structures with redundant models solved at different resolutions:

4. Dataset of redundant crystal structures

Figure 4.18: (A) The distribution of correlation coefficient values of hydrogen bond numbers to structural resolution for a set of 50 proteins. Each of the protein structure was experimentally determined at least 30 times and at varying resolutions. The distribution has a median of ~ 0.4 and a spread of ~ 0.2 . (B) A specific example of calmodulin with 109 structures. Red dotted lines indicate best fitted lines for resolution lower and higher than 2.1 Å

This dataset consists of 50 proteins whose structures have been solved and deposited in the PDB more than 30 times. All PDB entries of a particular protein are 100% identical in sequence. Each protein has at least one structure at a resolution of 2.1 Å or better and at least one structure with resolution lower than 3.2 Å.

A Pearson correlation coefficient between resolution and hydrogen bond per residue was calculated for each of the protein in the data set. In 44 of the 50 cases (88%) the number of hydrogen bonds is positively correlated to resolution (Figure 4.18). The median value of the correlation coefficients is -0.37, an interquartile range value of 0.23. A two-tailed one-sample *t*-test showed that this correlation is significant (p -value $< 10^{-4}$).

4.5.3 Hydrogen bond number in NMR models

4.5.3.1 Comparison between X-ray and NMR structure

Next we examined the protein model solved with NMR method with the following data set:

Figure 4.19: A comparison of the number of hydrogen bonds (normalized per residue) in crystal structures and the corresponding number in NMR structures of the same protein. The diagonal red line indicates when the number of hydrogen bonds found in the 2 experimental sources are the same.

5. Dataset of NMR structures

This dataset consists of 145 protein structures solved independently by NMR and X-ray crystallography. Chain lengths were greater than 120 amino acids, and the crystal structures were of resolution 2 Å or better. For comparison purposes, we required the X-ray and NMR structures to be in the same conformation. To ensure this, we chose NMR-X-ray pairs whose main-chain root mean squared deviation (R.M.S.D.) was less than 5 Å.

Of the same protein, models solved by the two experimental techniques apparently yielded different numbers of hydrogen bonds. NMR structures consistently had fewer hydrogen bonds per residue (mean of $\sim 16\%$ with a standard deviation of $\sim 13\%$) (Figure 4.19). In only 17 cases was this trend reversed.

4.5.3.2 Consensus of Hydrogen bonds in different NMR models of the same protein

Generally, NMR structures are solved with satisfying spatial restraints obtained from NMR signals (e.g. COSY or NOESY peaks). Often multiple models are

Figure 4.20: (A) Distribution of Jaccard Index for hydrogen bonds for NMR ensembles in the data-set. (B) hydrogen bond consensus among NMR models as a function as hydrogen bond per residue.

valid for a set of NMR signals. In fact, 139 of the 145 proteins in the NMR dataset are represented by an ensemble of multiple models, and only in 6 cases, the ensemble consisted of one solitary model. We used the Jaccard index to quantify the consensus between hydrogen bonds in different models within an ensemble. The Jaccard index $J(A, B)$ of two sets A and B is given by

$$J = \frac{|A \cap B|}{|A \cup B|} \quad (4.3)$$

The consensus of hydrogen bonds in an ensemble is defined as the mean of the Jaccard indexes of all possible pairs in the ensemble. For example, for a NMR ensemble with 10 models, a Jaccard index is computed for every pair of the models (a total of $_{10}C_2 = 45$ pairs). The consensus value is taken as the mean of the Jaccard indexes. On average, in the data set, this mean consensus value was found to be $\sim 64\%$, with standard deviation of $\sim 5\%$ (Figure 4.20 A). This figure is similar to what has earlier been reported as the consensus between NMR and X-ray structures [218]. The consensus tends to increase with the degree of saturation of donor-acceptor pairs in protein with a correlation coefficient of ~ 0.56 (Figure 4.20 B).

4.6 Protein 3D Model Refinement with Hydrogen Bonding Restraints

4.6.1 Importance of high-resolution protein 3D model

Rich information about biological and chemical properties about a protein can be obtained by analysing and comparing its 3D structure. The limit of our understanding are however limited by the resolution of the structure [219] (Figure 4.21). High-resolution protein structure is a prerequisite to a wide variety of computational methods, such as molecular dynamic (MD) simulation, molecular motifs comparison, binding site prediction, protein-protein interaction prediction, ionizable amino acid pK_a prediction etc. The accuracy of these methods critically relies on the detailed description of spatial location of protein atoms and their interatomic interaction. With low-resolution structure however, only coarse-grained information like radius of gyration and volume could be obtained.

It is possible to refine a low-resolution structure to its high-resolution counterpart by utilizing the knowledge about hydrogen bonding in proteins. In the preceding section we have detailedly characterized the properties of protein hydrogen bonds. Particularly, (i) the parameters of ideal donor-acceptor distance and $\hat{H}$ angle in different environments have been characterized. (ii) hydrogen bond number is highly predictable given the protein chain length, except for poor quality models such as low-resolution X-ray, NMR models and homology modelling. In these models often there are unpaired donors and acceptors, even in regions where they could not form hydrogen bond with the solvent.

In following sections, we will utilize the knowledge about hydrogen bond in the form of spatial restraints to refine low-resolution 3D models.

4.6.2 Review of previous work

Refining 3D structures of proteins using hydrogen bond restraints has been previously attempted both in the context of X-ray structures [198, 220] and computational models [201, 221–223]. In fact, the degree of hydrogen bonding satisfaction

Figure 4.21: Accuracy and application of protein structure models. Image taken from Baker & Sali, 2001 [219]

Figure 4.22: Image taken from Fabiola et al 2002. [198] showing the proposed hydrogen bonding potential on (A) donor-acceptor distance (B) $\hat{H}$ angle.

is used as a measure of local accuracy in protein models [224, 225]. The number of unsatisfied hydrogen bond donors and acceptors has also been shown to be a good predictor of molecular replacement performance [226] in protein X-ray crystallography. Besides, hydrogen bonding is included in force-fields used in molecular mechanics or homology modelling, either implicitly or explicitly [30, 69, 73, 227–233]. The parametrization of hydrogen bond in these force-fields, is however inaccurate based on our extensive analyses on high-resolution protein data set in Chapter 5. For example in the widely used X-ray structure determination and refinement software suite Phenix [198, 220], its hydrogen bonding potential had included a double-well restraint on $\hat{H}$ angle, but without considering the chemical nature of donor-acceptor atoms (at hindsight, the minima of the double-well corresponds to MC-MC hydrogen bonds and others) (Figure 4.22). In donor-acceptor distance also, an arbitrary switching function was implemented to prune the long-range effect of the potential.

Further, to the best of our knowledge, no current MD force-fields have included hydrogen bonding potentials that considered the effect of environment dependency. Correspondingly, it is observed that hydrogen bonds are often broken or missing in MD simulations [234]. Nonetheless, this artefact has been shown repairable with protein-specific polarization potentials [216, 234] obtained from customized quantum mechanical calculation. These *ab initio* quantum calculations, while accurate,

	Depth < 6Å				Depth > 6Å			
	$\mu(\text{D})$	$\sigma(\text{D})$	$\mu(\hat{H})$	$\sigma(\hat{H})$	$\mu(\text{D})$	$\sigma(\text{D})$	$\mu(\hat{H})$	$\sigma(\hat{H})$
MCMC-A	2.97	0.12	153.85	8.8	2.97	0.12	153.85	8.8
MCMC-B	2.93	0.12	150.21	15.49	2.89	0.09	155.75	9.15
MCMC-L	2.93	0.12	150.21	15.49	2.89	0.09	155.75	9.15
MCSC	2.98	0.17	129.63	16.36	2.96	0.15	129.25	15.83

Table 4.8: Hydrogen bond parameters for MC-MC in α -helix (MCMC-A), β -sheet (MCMC-B), other (MCMC-L) and MC-SC

Figure 4.23: Flowchart illustrating the hydrogen bond structure refinement protocol. For NMR structures the last step is effected by Modeller while for low-resolution X-ray structures it is PHENIX.

are computationally intensive and not feasible in most scenarios. Instead, in the following sections, we would describe our efforts and results in refining protein hydrogen bonds using an empirical method derived from our statistical analysis.

4.7 Refinement protocol

4.7.1 Hydrogen bond prediction and selection

To refine a low-resolution protein model, hydrogen bond pairs in the proteins were predicted and then restrained to their ideal geometries. To predict the correct

	D		$\hat{H}$			D		$\hat{H}$	
	μ	σ	μ	σ		μ	σ	μ	σ
OH-OE1	2.66	0.21	126.64	13.27	NE2-OE1	2.94	0.18	129.61	15.03
OH-OE2	2.66	0.21	126.64	13.27	NE2-OE2	2.94	0.18	129.61	15.03
OH-OD1	2.67	0.16	125.18	12.08	NE1-OD1	2.95	0.17	125.87	14.60
OH-OD2	2.67	0.16	125.18	12.08	NE1-OD2	2.95	0.17	125.87	14.60
OG-OD1	2.72	0.20	127.17	14.98	NE2-OD1	2.95	0.17	125.87	14.60
OG-OD2	2.72	0.20	127.17	14.98	NE2-OD2	2.95	0.17	125.87	14.60
OG1-OD1	2.77	0.22	126.32	15.52	N-OE1	2.95	0.18	131.17	15.23
OG1-OD2	2.77	0.22	126.32	15.52	N-OE2	2.95	0.18	131.17	15.23
OH-O	2.77	0.23	132.36	13.37	ND1-OG	2.95	0.23	125.93	18.22
OH-NE2	2.77	0.15	117.32	10.26	ND2-OG	2.95	0.23	125.93	18.22
OG-OE1	2.78	0.24	125.09	13.68	NH1-O	2.96	0.20	138.67	15.76
OG-OE2	2.78	0.24	125.09	13.68	NH2-O	2.96	0.20	138.67	15.76
OH-OH	2.78	0.28	119.89	10.58	NZ-NE2	2.97	0.23	123.01	14.28
OH-OG1	2.79	0.19	121.84	10.43	NE1-OG	2.97	0.22	126.20	17.07
OH-OG	2.79	0.20	124.06	15.99	NE2-OG	2.97	0.22	126.20	17.07
OG1-OE1	2.80	0.24	127.66	13.66	OG1-NE2	2.97	0.28	137.31	23.58
OG1-OE2	2.80	0.24	127.66	13.66	NZ-OG	2.97	0.25	125.71	16.86
OG-NE2	2.82	0.16	126.29	12.93	N-OD1	2.98	0.16	133.02	15.62
OG1-OH	2.82	0.21	125.82	16.93	N-OD2	2.98	0.16	133.02	15.62
OG-OH	2.83	0.26	120.36	12.00	NE1-OG1	2.98	0.22	125.67	17.00
NZ-OD1	2.85	0.23	128.07	14.82	NE2-OG1	2.98	0.22	125.67	17.00
NZ-OD2	2.85	0.23	128.07	14.82	NE-O	2.98	0.22	141.58	15.88
NZ-OE1	2.86	0.25	126.51	14.34	N-O	2.98	0.15	146.14	17.78
NZ-OE2	2.86	0.25	126.51	14.34	NE1-O	2.98	0.18	134.68	15.96
OG1-OG1	2.87	0.21	121.18	12.79	NE2-O	2.98	0.18	134.68	15.96
NZ-OH	2.88	0.47	123.32	14.26	NE-OG	2.99	0.20	134.47	16.89
NZ-O	2.90	0.22	142.34	16.22	ND1-NE2	2.99	0.25	124.40	13.54
OG-O	2.90	0.26	134.81	14.76	ND2-NE2	2.99	0.25	124.40	13.54
OG-OG	2.90	0.27	128.17	18.95	NH1-OG1	2.99	0.18	127.32	16.31
NE-OD1	2.90	0.17	123.70	15.71	NH2-OG1	2.99	0.18	127.32	16.31
NE-OD2	2.90	0.17	123.70	15.71	NE-OG1	3.00	0.22	133.15	16.03
N-NE2	2.90	0.29	124.60	9.99	NE1-OH	3.00	0.21	121.06	11.96
NE-OE1	2.90	0.20	126.93	16.82	NE2-OH	3.00	0.21	121.06	11.96
NE-OE2	2.90	0.20	126.93	16.82	ND1-OG1	3.00	0.25	124.93	14.92
ND1-OE1	2.91	0.20	127.72	15.15	ND2-OG1	3.00	0.25	124.93	14.92
ND2-OE1	2.91	0.20	127.72	15.15	ND1-O	3.02	0.22	134.54	15.11
ND1-OE2	2.91	0.20	127.72	15.15	ND2-O	3.02	0.22	134.54	15.11
ND2-OE2	2.91	0.20	127.72	15.15	NH1-OG	3.02	0.22	126.00	15.50
OG1-O	2.92	0.25	136.08	13.57	NH2-OG	3.02	0.22	126.00	15.50
ND1-OH	2.92	0.23	123.26	14.99	NE-OH	3.06	0.18	117.30	12.37
ND2-OH	2.92	0.23	123.26	14.99	N-OH	3.09	0.20	128.55	16.02
NZ-OG1	2.92	0.22	131.02	14.16	NE1-NE2	3.09	0.18	115.68	12.42
OG-OG1	2.92	0.25	121.95	12.75	NE2-NE2	3.09	0.18	115.68	12.42
NH1-OD1	2.93	0.19	126.13	15.62	N-OG	3.10	0.18	125.42	13.49
NH1-OD2	2.93	0.19	126.13	15.62	N-OG1	3.10	0.18	129.08	15.32
NH2-OD1	2.93	0.19	126.13	15.62	NH1-OH	3.10	0.25	128.12	16.68
NH2-OD2	2.93	0.19	126.13	15.62	NH2-OH	3.10	0.25	128.12	16.68

donor-acceptor pair for hydrogen bonding, two empirical scores of (a) preference score and (b) specificity score are devised. The first score measures the feasibility of hydrogen bond formation, while the second score ensures a one-to-one pairing among donors and acceptors. The computations of the scores are described as follows.

First, for every donor (d) and acceptor (a) atom in the input protein structure, all potential hydrogen bonding partners candidate within 6.0 Å were listed. A preference score, $S(d_i, a_j)$ is computed for each of these possible hydrogen bonds (d_i, a_j) as

$$S(d_i, a_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{D(d_i, a_j) - \mu}{\sigma}\right)^2\right) \cdot M(d_i, a_j) \cdot H(d_i, a_j) \quad (4.4)$$

where $D(d, a)$, $M(d, a)$ and $H(d, a)$ are the distance, hydrogen bond type and sequence separation modifiers respectively. The distance modifier $D(d, a)$ is the difference between donor-acceptor distance $D(d_i, a_j)$ and the ideal hydrogen bond length μ . The mean (μ) and standard deviation (σ) values were learned from the statistical study (see Chapter 5) and set to 2.9 Å and 0.3 Å respectively. The Gaussian form of $D(d, a)$ makes its values comparable to the other modifiers.

The value of the hydrogen bond type modifier $M(d, a)$ reflects the relative abundance of the different types of hydrogen bonds. The values of MC-MC, MC-SC, SC-MC and SC-SC types were set to 0.55, 0.11, 0.20 and 0.14 respectively.

The value of the sequence separation modifier $H(d, a)$ is determined by counting the number of hydrogen bonds with the sequence separation in the high-resolution dataset. The value is normalized as relative frequency 4.10. Generally, the modifier has smaller values when the sequence separation is small and increases with increasing sequence separation. However, there is a peak in the values at a sequence separation of 4 due to hydrogen bonds in α helices. The values of this modifier range from 0.37 to 0.92.

A specificity score $F(d, a)$ is then computed using the preference score $S(d, a)$. $F(d, a)$ ensures that each donor is uniquely paired with one acceptor.

$$F(d_i, a_j) = \frac{1}{2} S(d_i, a_j) \left(\frac{S(d_i, a_j)}{\sum_a S(d_i, a^*)} + \frac{S(d_i, a_j)}{\sum_d S(d^*, a_j)} \right) \quad (4.5)$$

S	0	1	2	3	4	5	6	7	8	9	≥ 10
H	0	0	0.1	0.37	0.93	0.87	0.85	0.91	0.91	0.91	0.92

Table 4.10: Sequence separation modifier H at different sequence separation

where the i and j in the summation represent all possible acceptors (a_j) and donors (d_j). Pairing between donor and acceptor is straightforward if they are unique in the search radius of 6 Å. In cases where donor or acceptors had various potential partners, the one with the largest specificity score was chosen. Once paired, the particular donor-acceptor combinations were disregarded in subsequent computations of specificity score to ensure the one-to-one donor-acceptor mapping. The process was iterated until a sufficient number of hydrogen bonds were predicted, as determined by in section 4.5.1 .

4.7.2 Hydrogen bond restraints

After the prediction phase, geometrical restraints are applied to predicted donor-acceptor pairs. These hydrogen bond geometrical restraints are all in Gaussian form. Subsequently homology restraints and stereochemical restraints are combined with hydrogen bond restraints, and structure optimization is performed based on them to give the refined model.

A. Structure refinement of low-resolution X-ray models

To test the refinement protocol on low-resolution X-ray models, we prepare the following data set:

6. X-ray density maps data set

This dataset consists of 57 pairs of structures solved by X-ray crystallography and have their corresponding electron density maps deposited in the PDB. Each pair consists of a structure of the same protein or a very close homologue (95-100% identical in sequence). Pairs consist of one high (2.1 Å or better) and one low-resolution (2.1 Å or worse) structure such that the difference in resolution is 1 Å or greater and their main chain R.M.S.D. ≤ 1 Å. At the level of these pairs, the entire dataset is non-redundant at 50% sequence identity.

Refinement of the low-resolution X-ray structure considering its diffraction electron density map and the Gaussian restraints of the predicted hydrogen bonds was performed using the PHENIX software suite [220]. Our newly introduced hydrogen bond restraints counter the effect of default hydrogen bonding potential in PHENIX (c.f. 4.6.2). All parameters were set to their default values, except the number of macro-cycle, which was set to 10 (instead of default value of 3) to allow for better optimization convergence. For an unbiased evaluation of the performance of this protocol, the refinement procedure was fully automated and devoid of any manual intervention.

B. Structure refinement of NMR structures

To test the protocol on NMR models, we used the previously mentioned dataset of 145 NMR structures. Structure refinement was implemented through the Modeller (v9.11) [29] automodel function. The input NMR structure was taken as initial configuration. If an ensemble of models were deposited in the PDB for a particular NMR structure, the template was chosen as that structure which had the least R.M.S.D. to the average structure. All stereochemical restraints were added by default. Homology restraints were learned from a sequence alignment of the protein with itself. One sequence in the alignment was taken as the target while the other was assigned as the template, corresponding to the input low-resolution structure. Homology restraints were introduced as distance restraints between all C $^{\alpha}$ atoms only. This was done to allow movement of protein atoms during the refinement process. The standard deviations of the distance restraints were set to a relaxed value of 1 Å, roughly equivalent to the magnitude of atomic fluctuations between different models of the same NMR structure. Hydrogen bond restraints were introduced as Gaussian restraints on bond length and bond angle. To refine the structure, both sets of hydrogen bond restraints and homology derived structural restraints are simultaneously minimized. 20 refined models were built for each input structure.

4.8 Refinement performance

4.8.1 Refinement of low-resolution X-ray structure

A. Structure quality improvement assessment

We refined 57 structures of the X-ray density maps dataset by first predicting the plausible hydrogen bonding partners of all donor/acceptor atoms. The predictions were utilized as restraints in refinement process. To benchmark the prediction protocol, the actual hydrogen bond partners from the high-resolution structure were used as restraints. We fitted the initial low-resolution and refined structures into the electron density map of the high-resolution structure of the same protein. The crystallographic R -factor computed for each fitting was used to measure structure improvement on refinement.

To facilitate the fitting, the main chains of the low-, high- and refined models were first structurally aligned using CLICK [12]. R values were computed after 5 macro-cycles of rigid body refinement with all parameters set to their default values using the PHENIX software suite.

A paired t-test of the R values of 57 initial low-resolution structures and refined models in the X-ray density maps dataset showed that the overall improvement on refinement is statistically significant (p -value = 0.05). However, only 8 of the 57 structures improved beyond an R value threshold of 0.005. Importantly, none of the structure had significantly poorer R values after refinement. In the benchmark, where the hydrogen bonded donor-acceptor pairs in the high-resolution structure were restrained, 10 of the 57 structures improved beyond this threshold (Figure 4.24).

We illustrated the improvement obtained by the refinement protocol when it was used on the human antibody Fc fragment (PDB:4HAG). Starting from a low-resolution X-ray structure, our protocol improves the structure by 0.57 Å, which is an improvement of ~ 0.025 in R-factor. The R.M.S.D. and R-factor were computed in comparison to an 1.9 Å resolution structure of the same protein and its

Figure 4.24: Distribution of improvement in electron density map fitting measured by R values using predicted hydrogen bonds (blue) or native sets of hydrogen bonds (black). The regions outside the two parallel green dotted lines indicate improvements that are significant at 95% confidence level.

Figure 4.25: Superposition of low-resolution (grey) and refined (white) structures of human antibody Fc fragment at residue 248 and 252. The structures are rendered using UCSF Chimera software suite [235]

structure factors (PDB:4J12). The refinement successfully reintroduced several hydrogen bonds in secondary structures that were initially missing (Figure 4.25).

B. Hydrogen bond recovery assessment

Another indicator of successfully refined models is the recovery of missing hydrogen bonds from the initial low-resolution models. Two measures were studied to quantify the degree of recovery, viz., the total number of intramolecular hydrogen bond and overlap ratio of native hydrogen bonds in refined models. The total number of hydrogen bonds were counted and compared to the expected number. The overlap ratio of native hydrogen bonds was calculated as the fraction of hydrogen bonds in the refined models that are also present in the high-resolution structure.

As expected, the numbers of intramolecular hydrogen bonds in low-resolution models were slightly less than their high-resolution counterparts ($\sim 5\%$). Refinements on low-resolution structure, using either predicted hydrogen bonds or known native sets of hydrogen bonds, were able to increase the number to the expected levels (Figure 4.26A). The magnitude of increase in hydrogen bond numbers using these two sets was indistinguishable. Including the native hydrogen bonds as input to the refinement protocol resulted in a higher overlap ratio (88.3 vs 85.9%) (Figure 4.26B).

C. Correlation of structural improvement and hydrogen bond recovery

We proceed to confirm that the structural improvements are attributed to the recovery of hydrogen bonds. Correlation between recovery rate of hydrogen bonds and improvement in electron density map fitting was quantified by a Pearson correlation coefficient measure.

$$\rho(\Delta HB, \Delta R_{free}) = \frac{cov(\Delta HB, \Delta R_{free})}{\sigma(\Delta HB) \cdot \sigma(\Delta R_{free})} \quad (4.6)$$

where ΔHB is the recovery rate of hydrogen bonds, i.e. the difference between number of hydrogen bond in refined models and initial model; ΔR_{free} is the difference in structural accuracy of refined models and initial model measured by R_{free} .

The correlations by using predicted hydrogen bonds ($\rho=0.236$) and native hydrogen bonds ($\rho=0.294$) were similar, with the latter correlates being marginally but

Figure 4.26: (A) Comparison of number of intramolecular hydrogen bonds of low-resolution structures (red), refined models using predicted (blue) and native sets (purple) of hydrogen bonds and high-resolution structures (green). The green diagonal dotted line represents the expected number of hydrogen bonds. (B) Comparison of recovered native intramolecular hydrogen bonds in refined models using predicted (red) and native sets (blue) of hydrogen bonds. The green diagonal line represents cases where the initial and refined models have the same hydrogen bond overlap ratio.

insignificantly better. These correlation values though small are statistically significant as confirmed by a one-tailed probability test (refinement using predicted hydrogen bond, $p = 0.04$; refinement using native hydrogen bond, $p = 0.01$) when compared to the initial low-resolution structure.

D. Distribution of missing and recovered hydrogen bonds

Most of the hydrogen bonds recovered from the refinements are of the MC-MC type, and SC- related hydrogen bonds were not successfully retrieved. In the high-resolution structure datasets, the MC-MC type constitutes $\sim 70\%$ of all hydrogen bonds. In low-resolution models, the share of MC-MC hydrogen bonds is $\sim 60\%$. The refinement recovers the ratio to 62% and 64% using predicted or native hydrogen bonds respectively (Figure 4.27A). To check if this recovery was context/environment dependent, we correlated recovery to residue depth. The recovery rate was uniform at different levels of depth (Figure 4.27B).

Figure 4.27: (A) Box-and-whisker plots showing the distributions of the normalized frequency of different types of intramolecular hydrogen bond in X-ray structures (yellow), NMR structures (orange), refined models using predicted hydrogen bonds (blue) and refined models using native hydrogen bonds (green). (B) The frequency distributions of intramolecular hydrogen bond at different depths in NMR structures (red), refined models using predicted hydrogen bonds (blue) and refined models using native hydrogen bonds (green).

4.8.2 Refinement of NMR structures

NMR structures are constructed from experimentally determined spectroscopy data. Typically, these data are considered to be of a lower resolution than X-ray structures [236]. While NMR spectroscopy could be used to directly detect hydrogen bonds [237–239], these experiments are difficult and expensive and seldom used in conjunction with structure determination [238, 240, 241]. Earlier studies have also noted the inaccuracy of details such as packing in NMR structures [236, 242]. Similarly, as we established in the section 4.5.3.1, NMR structures have a low hydrogen bond overlap ratio as in other low-resolution models. We have considered NMR structures as good representatives of low-resolution structures that retain several crucial protein features such as the overall fold and topology. Our aim in this part of the study was to determine whether these NMR structures could be refined to be similar to their X-ray counterparts by introducing missing hydrogen bonds. Apart from testing the predicted and native hydrogen bonds as restraints, two other protocols were tested for NMR structures. The first additional protocol

restrains a union set of all hydrogen bonds found in the ensemble of NMR models. The second additional protocol is similar to the hydrogen bond prediction protocol but with the crucial difference that it does not assign a one to one pairing of donor and acceptors. All possible hydrogen bonds partners detected within a distance of 6.0 Å are paired with one another. In all, 145 NMR structures were subjected to these 4 refinement protocols. We assessed the overall accuracy of the refinement protocols in terms of its R.M.S.D., total number of hydrogen bonds and number of hydrogen bonds that overlapped with the native structure (Table 4.11).

As in the case of the X-ray structures, restraining the native hydrogen bonds resulted in the best R.M.S.D. improvement of ~ 0.3 Å. The predicted hydrogen bonds improved the initial NMR model by ~ 0.1 Å. The two additional protocols were ~ 0.1 Å worse in term of R.M.S.D. than the initial NMR structure. A reason for this deterioration in these additional protocols is probably because the excessive restraints result in a 10 – 30% increase in non-native hydrogen bonds and hence incorrect conformations. Predicting excessive hydrogen bonds in itself may not be detrimental to structure refinement. For instance, the protocol that makes use of native hydrogen bonds over-predicts the number of hydrogen bonds by $\sim 18\%$. However, the overlap it has with native MC-MC hydrogen bonds is over 95% with only $\sim 2\%$ over prediction. Though the MC-MC overlapping for the three other protocols are close to one another, the predicted hydrogen bond protocol has the fewest over predictions ($< 1\%$). Both additional protocols underperformed in comparison with the predicted hydrogen bond protocol and are hence not discussed here. In the sections to follow, we only report and discuss the results from the predicted and native hydrogen bond protocols.

A. Improvement in R.M.S.D. with respect to the high-resolution X-ray structure.

In general, the magnitude of improvement on refinement is inversely related to the accuracy of the initial models. As in the case of the X-ray structures, restraining the native hydrogen bonds resulted in the best R.M.S.D. improvement of ~ 0.3 Å. The predicted hydrogen bonds improved the initial NMR model by ~ 0.1 Å (Table 4.11). Starting structures with R.M.S.D.s of 1 Å or greater usually improved.

	R.M.S.D.	OVERLAP	OVERLAP (MC-MC)	HB	HB(MC- MC)
NMR	2.00 Å	62.5%	77.6%	-17.43%	-18.29%
ALL	2.06 Å	72.3%	88.8%	27.55%	4.61%
UNION- NMR	2.13 Å	73.2%	87.2%	8.73%	-9.48%
PREDICT	1.90 Å	71.7%	89.2%	-0.24%	-6.30%
CORRECT	1.73 Å	83.5%	95.7%	18.94%	1.86%

Table 4.11: Comparison of structure quality (measured as R.M.S.D. to high-resolution X-ray structure), hydrogen bond fulfilment ratio and hydrogen bond number (as compared to high-resolution X-ray structure) of initial NMR models and refinement models using different hydrogen bond prediction protocols.

Greater the R.M.S.D. value to the high-resolution X-ray structure, greater was the improvement in accuracy. However, NMR structures with R.M.S.D. < 1 Å almost always deteriorated (Figure 4.28). In 127 (88%) out of 145 refined NMR models accuracy improved in terms of main-chain R.M.S.D. after hydrogen bond refinement. A paired t -test showed that the improvement is statistically significance (p -value $< 10^{-4}$). In cases where structures were less accurate than the starting structure, R.M.S.D. was never worse than 0.3 Å R.M.S.D.. In contrast, when refinement improved accuracy, 18 cases were better than 0.3 Å in R.M.S.D.. Using predicted hydrogen bonds, the refinement plateaus at around 0.3 Å, when the structural difference is larger than 2 Å. Nonetheless, if native hydrogen bonds were known, the refinement increases linearly with no effective limit found. On average, the resolution of a initial low-resolution structure of R.M.S.D. 4 Å to native structure is improved by about 1 Å.

B. Hydrogen bond recovery assessment

The motivation for refinement was the fact that NMR models had $\sim 20\%$ fewer hydrogen bonds and an overlap ratio of $\sim 63\%$ (78% in MC-MC) with respect to their X-ray counterpart. A measure of improved accuracy is hence the degree of restoration of this ratio to native, high-resolution X-ray structure levels. Refinement successfully increased the number of hydrogen bonds in 143 of the 145 models. When using predicted hydrogen bonds in refinement, the number of hydrogen bonds

Figure 4.28: Improvement in structural accuracy of NMR models measured by main-chain R.M.S.D. difference between refined models and initial models, using high-resolution X-ray structures as reference. Blue dots represent improvements when predicted hydrogen bonds are used while red dots represent improvements when native sets of hydrogen bonds are used. A moving average line (orange: predicted hydrogen bonds; green: native hydrogen bonds) was plotted for each case. A dotted pink line at 1 Å was added to indicate the effective range of the refinement protocol.

Figure 4.29: (left) Comparison of number of intramolecular hydrogen bonds of low-resolution structures (red), refined models using predicted (blue) or native sets (green) of hydrogen bonds with their high-resolution structures (yellow). (right) hydrogen bond overlap ratio in refined models using predicted (blue) and native sets (red) of hydrogen bonds.

increased to levels similar to the X-ray counterpart (difference of $\sim 0.25\%$) with an overlap of 72% (89% in MC-MC). When refining with native hydrogen bonds, the number of hydrogen bonds overshoot the number in the X-ray structure by 19%. This is compensated by an overlap of 84% (96% in MC-MC). On the other hand, more native hydrogen bonds (18% vs 12%) were recovered in refined models if they were used as input (Figure 4.29). Overall, the H-bond refinement algorithm has a false positive rate of 67.1% when introducing H-bonds with predicted donor-acceptor pairs. This value decreases to 17.7% when using donor-acceptor pairs inferred from high-resolution structure.

C. Correlation of structural improvement and hydrogen bond recovery

Similar to the assessment in low-resolution X-ray structure refinement, Pearson correlation coefficient was calculated to quantify the relationship between the recovery of hydrogen bonds and structural accuracy improvement. Using predicted hydrogen bonds, a correlation coefficient of 0.43 was achieved, and using native hydrogen bonds the value rose to 0.48. Both coefficients were statistically significant with p -value $< 10^{-4}$.

D. Distribution of missing and recovered hydrogen bonds

Figure 4.30: (left) Box-plots showing the distributions of normalized frequency of different types of intramolecular hydrogen bond in X-ray models (yellow), NMR models (orange), refined models using predicted hydrogen bonds (blue) and refined models using native hydrogen bonds (green). (right) Frequency distributions of intramolecular hydrogen bond at different depths in NMR models (red), refined models using predicted hydrogen bonds (blue) and refined models using native hydrogen bonds (green).

The refinement protocols made a substantial improvement in hydrogen bond recovery, especially in the MC-MC category (Figure 4.30). The number of hydrogen bonds increased from 50% to 57% (as compared to $\sim 69\%$ in high-resolution native structures). The refinement protocol failed to predict SC-related hydrogen bonds, and hence no increase in number of hydrogen bonds was detected in these categories (MC-SC, SC-MC, SC-SC). If native set of hydrogen bonds could be correctly predicted, recoveries would be detected in all categories (MC-MC: 60%, MC-SC 3%, SC-MC 6%, SC-SC 3%).

In contrast to the low-resolution X-ray structure, the missing hydrogen bonds in NMR models are more concentrated on the protein surface (< 5.5 Å depth). These missing hydrogen bonds were difficult to predict, and refinement using such predictions only marginally ($\sim 0.5\%$) increased the number of hydrogen bonds. Nonetheless, a dramatic recovery ($\sim 3\%$) of these hydrogen bonds could be achieved, if the native hydrogen bonds were correctly identified.

4.8.3 Utilities of hydrogen bond refinement protocol

4.8.3.1 Improved pK_a prediction

In this section we demonstrate the utility of our hydrogen bond refined models. Particularly, we are interested in showing how the refined structures improve our detection and prediction of physicochemical properties of protein environment. For this purpose we show that the refined models have improved predicted pK_a values over their unrefined counterparts. The pK_a of a residue is sensitive to its immediate neighborhood environment and our refined structures better model these regions of the protein.

From a dataset of 69 proteins in which experimentally measured pK_a s are available, 10 residues from 4 proteins that had both NMR and X-ray structures deposited in the PDB were used as a benchmark. As with the tests earlier, we used the NMR structures as the initial model and refined them using the two protocols. Using DEPTH [89], the pK_a s of the 10 residues were computationally predicted in the X-ray, NMR and refined structures. We had earlier established that pK_a predictions from DEPTH were as accurate as the best currently available prediction methods [89]. In the case of the 10 residues, the pK_a predictions on the X-ray structure were on average ~ 0.57 pH units off from their experimentally determined values (Table 4.12). We consider this difference to be within tolerable limits and close to or even smaller than measurement/prediction errors. Predictions on the NMR structures however were on average ~ 1.55 pH unit away from the experimental values, and statistically significantly different from X-ray structures (p value < 0.01 by a Wilcoxon signed-rank test). With the refined models, the average errors were reduced to 1.11 and 0.97 pH units by using predicted and native sets of hydrogen bonds partners respectively.

We investigate further for one example, Hen Egg White Lysozyme (PDB:4LZT) in the list, where the pK_a predictions of Asp (Asp-66 and Asp-52) have improved considerably. The pK_a value of Asp-66 was measured to be 0.9, which is substantially reduced from its model value of 3.8 in free solution. Our refinement protocol successfully reintroduced the missing hydrogen bond between the side chain atoms

PDB code				pK_a prediction				Improvement	
X-RAY	NMR	residue	pK_a	X-ray	NMR	model (pre-dict)	model (na-tive)	predict	native
1BEO	1BEG	ASP:21:A	2.50	2.24	3.97	3.96	3.40	-0.01	-0.57
1BEO	1BEG	ASP:30:A	2.51	2.72	3.96	3.36	3.05	-0.60	-0.91
2RN2	1RCH	ASP:94:A	3.20	2.63	3.77	3.87	3.87	0.10	0.10
2RN2	1RCH	ASP:134:A	4.10	4.45	3.17	3.79	3.96	-0.62	-0.79
2RN2	1RCH	ASP:148:A	2.00	1.31	4.32	4.28	3.83	-0.04	-0.49
3RN3	2AAS	ASP:14:A	1.80	1.54	4.92	4.28	3.66	-0.64	-1.26
3RN3	2AAS	ASP:53:A	3.70	4.08	2.71	3.45	3.26	-0.74	-0.55
4LZT	1E8L	GLU:35:A	6.20	4.20	6.75	6.64	4.62	-0.11	1.03
4LZT	1E8L	ASP:52:A	3.70	3.57	4.93	4.46	3.52	-0.47	-1.05
4LZT	1E8L	ASP:66:A	0.90	1.73	3.81	2.53	2.48	-1.28	-1.33

Table 4.12: Table showing the experimentally measured pK_a and predictions using X-ray, NMR, and refined models using predicted and native set of hydrogen bonds. The last two columns showed the improvements ($pK_a^{refine} - pK_a^{NMR}$) on pK_a prediction in refined models.

of Asp-66 and Tyr-53 (Figure 4.31). This hydrogen bonding interaction stabilizes the deprotonated form of Asp and leads to its reduced pK_a . For Asp-52, its pK_a was measured to be 3.7. In the X-ray structure, the residue was shown to form a hydrogen bond with Asn-59 and the pK_a was estimated to be 3.6. This estimated value changes to 4.9 if the hydrogen bonding interaction was missing, as in the case of the NMR structure. Our refinement protocol was capable of recovering on average ~ 0.5 hydrogen bond in the NMR structure and a better pK_a estimation of 4.5 is obtained. Using native set of hydrogen bonds, however, improves the estimation to an accurate value of 3.5. Accurate determination of the pK_a (and hence charge state) on Asp 52 is important, as it is one of the active site residues and acts as a nucleophile in the hydrolysis reaction of the lysozyme.

4.8.3.2 Case studies

(A) HSPC034

We applied our refinement protocol to the NMR structure of human protein HSPC034 (PDB: 1XPW). This structure has a main chain R.M.S.D. of 1.44 Å to a 1.60 Å

Figure 4.31: Restoration of hydrogen bonds between residue pairs of (A) Asn 59 – Asp 52 and (B) Asp 66 – Tyr 53 from NMR model (grey) in refined model (white). The structures are rendered using UCSF Chimera software suite [235].

resolution X-ray structure (PDB:1TVG) of the same protein. An earlier study has attempted to refine the NMR structure using the Rosetta software [226]. Our protocol that used predicted hydrogen bonds reported the same magnitude in R.M.S.D. improvement of ~ 0.15 Å (Table 4.13) with the previous study. The protocol that uses native hydrogen bonds results in a larger improvement of ~ 0.3 Å. 11 and 15 of the missing 17.6 of MC-MC hydrogen bonds were recovered by the predicted and native hydrogen bond protocols respectively. Even though the overall recovery of the 36% missing hydrogen bonds was a modest 10% in the predicted hydrogen bond protocol (a notable 27% in the native hydrogen bond protocol), it resulted in an improvement of 0.15 Å in main chain R.M.S.D.. We believe that the significant number here is the overlap of the MC-MC hydrogen bonds that would determine the overall fold of the protein. Our protocol has performed lesser sampling, and all refined models were considered without any filtering with statistical potential. In the refined models, the improvement in local regions coincides with its native hydrogen bonds recovery (Figure 4.32). Majority ($\sim 75\%$) of the residues have R.M.S.D. < 1 Å, and the larger overall R.M.S.D. is contributed from unstructured loop regions and terminus in the structure. Importantly, the protocol was shown to be capable of discarding incorrect hydrogen bonds that are incompatible with the rest of the high-confidence hydrogen bonds. For example, a hydrogen bond between PHE-119 and VAL-66 was observed in 7 out of 20 NMR models. This MC-MC hydrogen bond occurs at a region with depth of 7 Å, and should have

	R.M.S.D.	OVERLAP	OVERLAP(MC-MC)	HB	HB(MC-MC)
NMR	1.44 Å	64%	79%	90%	86%
PREDICT	1.27 Å	74%	93%	108%	107%
CORRECT	1.16 Å	91%	98%	102%	99%

Table 4.13: Summary of refinement statistics on HSPC034 in terms of R.M.S.D. to high-resolution X-ray structure, hydrogen bonds overlap (OVERLAP) and hydrogen bond number (HB).

Figure 4.32: Coincidence between local structure improvement and hydrogen bond recovery. Solid blue bars showed the difference in residue-wise R.M.S.D. between refined and initial model of HSPC034, while the red bars showed the change in native hydrogen bonds occupancy for each residue of the protein. In both cases negative values indicate closer measure to high-resolution native structure.

a bond- angle at around 155° . Nonetheless, the observed hydrogen bonds in NMR models had an average bond angle of $\sim 135^\circ$, and was thus rejected.

(B) HEWL and Ribonuclease T1

In this section we present two case studies, Hen Egg White Lysozyme (HEWL) and Ribonuclease T1. For both proteins a high-resolution X-ray structure and an NMR structure are deposited in the PDB (PDB:193L and PDB:1YGW respectively). Data from hydrogen exchange experiments under physiological conditions is also available for both these proteins [205, 243]. We refined the NMR structures of RNase T1 (PDB:1IYY) and HEWL (PDB:1E8L), using the native and predicted hydrogen bond protocols.

We tested whether the hydrogen bond protected amide donors in refined models correlate better with experiment data than the low-resolution structure. Low-resolution and refined structures were represented by an ensemble of models. For fair comparison, an intramolecular hydrogen bond was counted for the ensemble only if it was observed in half of all models (e.g. hydrogen bonds were observed in at least 5 models if 10 NMR models were deposited in PDB). We made the assumption that all exposed donors are susceptible to hydrogen exchange, except for those that are hydrogen-bonded.

There are 30 and 20 exposed amide groups in HEWL and ribonuclease T1 respectively (Table 4.14). For HEWL, high-resolution X-ray model correctly predicted the protection states of 25 amides, as compared to 22 for NMR models. Refined models also made similar correct number of predictions with high-resolution structure (25 if using predicted hydrogen bonds, 24 if using native hydrogen bonds). Prediction results were similar in ribonuclease T1 protein using different classes of models (14, 14, 13 and 14 for high-resolution model, low-resolution model, refined model with predicted hydrogen bonds, refined model with native hydrogen bonds respectively).

Two examples of protected amides explained by recovered hydrogen bond are illustrated here. In HEWL a small α -helix from residue 106 to 112 was recovered in refined model. As a result, amide group of PHE-108 at the N-terminal of the helix

Protein Name	X-ray	NMR	Refined Model (predict)	Refined Model (native)
	Hydrogen Bond Assessment*			
HEW Lysozyme	106 (100%)	68.4 (48.4%)	92.0 (51.7%)	119.1 (65.7%)
Ribonuclease T1	72 (100%)	45.5 (37.2%)	69.6 (46.6%)	87.5 (70.0%)
	Structure Accuracy Assessment**			
HEW Lysozyme	0 Å	1.55 Å	1.59 Å	1.48 Å
Ribonuclease T1	0 Å	2.10 Å	1.91 Å	1.61 Å
	Hydrogen/Deuterium Exchange Assessment			
HEW Lysozyme	25/30	22/30	25/30	24/30
Ribonuclease T1	14/20	14/20	13/20	14/20

Table 4.14: Summary of the assessment of protein model refinement on Hen Egg White Lysozyme (HEWL) protein and Ribonuclease T1. * number of hydrogen as compared to native structure is represented in percentage term in the parenthesis. ** structural accuracy is measured as main-chain R.M.S.D. to the native structure.

was able to form hydrogen bond with carboxylic group of ARG-112 (Figure 4.33A). Also in ribonuclease T1, a MC-SC hydrogen bond was formed between GLN-85 and ASN-81 in refined model which protected GLN-85 from hydrogen exchange (Figure 4.33B). This hydrogen bond was not detected in low-resolution structure, as the orientations of residues were not favorable.

Overall, for the two proteins their hydrogen bond refined models correlated better with HD exchange experiment data than their low-resolution structure, and the accuracies were similar to that of a high-resolution X-ray structure. These case studies validated the usefulness of our refinement protocol.

4.9 Chapter Summary and Discussion

In this chapter, we have studied in details various physical attributes of hydrogen bonding in proteins. Though H-bonds have been characterized extensively in the

Figure 4.33: Examples of recovered intramolecular hydrogen bonds in refined model (yellow) from low-resolution model (red) that protects amide proton from hydrogen exchange in (i) Hen Egg White Lysozyme (PDB:1E8L) and (ii) Ribonuclease T1 (PDB:1IYY). The structures are rendered using UCSF Chimera software suite [235].

past, here we have re-examined the analyses with more experimental data of higher resolution. Further the characterization has been extended to include environment dependency (quantified by depth measure) and hydrogen bond classification.

First, we have revisited the definition of hydrogen bond. With abundance of data, we were able to quantitatively examine the preferred geometrical distribution of donor, acceptor and acceptor-antecedent due to hydrogen bonding. The investigation suggested that the lower bound of antecedent ($\hat{H}$) angle should be to $\sim 100^\circ$ from the conventional 120° , while the latter was proposed based on guesswork with very small database analysis. The new definition is supported by hydrogen/deuterium exchange experiment as well as small molecule database study.

We then empirically categorized protein hydrogen bond into four types based on whether the donor and acceptor atom belong to main chain or side chain. We studied the geometrical distribution of hydrogen bond, as measured by donor-acceptor distance and donor-acceptor-antecedent angle, and their variations with environment as measured by depth. With a sufficiently large database, we are able to establish the statistical difference in geometry of different types of hydrogen bonds.

These geometrical difference are then further characterized. On donor-acceptor distance, the length decreases with depth for all types of hydrogen bond. This is presumably due to lower permittivity in protein interior and hence stronger hydrogen bond strength. In $\hat{H}$ angle, we found that SC-SC hydrogen bond has a preferred angle at $\sim 120^\circ$, a value also indicated by previous studies [188]. This is

Figure 4.34: $\hat{H}$ angle of MC-MC H-bonds without secondary structures at different sequence separation shown in moving average (standard deviation shown in red dotted line) of window size 300.

also the preferred angle for non-peptidic small organic molecule hydrogen bonds in the Crystallography Open Database. However, the $\hat{H}$ angle increases when main chain atom is involved, with MC-MC hydrogen bond having the largest angle at $\sim 155^\circ$. This result is also reproduced for small peptides (2 - 3 mers) deposited in the Cambridge crystallography database. As the small peptides are devoid of protein context, it suggests the preference is due to innate chemical nature of main chain atom rather than environment. We noted however, there are two exceptions of where MC-MC $\hat{H}$ angle could deviate from 155° . First, hydrogen bonds in β -strands close to the protein surface tend to have smaller $\hat{H}$ value. However when the β -sheet termini were removed and when Glycine were excluded, the $\hat{H}$ angle values were back to $\sim 155^\circ$. This suggests that functional group of amino acid could also plays a role in $\hat{H}$ angle preference in these environments. Second, $\hat{H}$ angle of hydrogen bond with residue contact order less than 4 is predominantly smaller (Figure 4.34). For instance, 3_{10} helices have a preferred $\hat{H}$ angle $\sim 125^\circ$.

Next, we further show that the only $\hat{H}$ angle allowed for secondary structure is tightly constrained around $\sim 155^\circ$ by attempting a range of donor-acceptor distances

and $\hat{H}$ angles. All other angles would introduce steric clashes incompatible to other structural restraints of secondary structure. Taken together, these studies suggest a possible causal connection between peptide bond planarity and the formation of secondary structures.

An additional discovery is that hydrogen bond length is slightly shorter (~ 0.1 Å) in β -sheets than in α -helices. This is exemplified by the (anti-)correlation ($r \sim -0.55$) between β -sheets preference and the donor-acceptor distance of different amino acids. As donor-acceptor distance is correlated with partial charge distribution on donor group, we postulated that amino acids with high occurrence propensity in β -sheets could have more polarized amide groups. One evidence supporting this postulation is the observation that some strong β -sheet former amino acids are β -branched (e.g. isoleucine, leucine, valine). β -branched carbons could increase the dipole moment of the adjacent atoms (in this case, the amide group) through the effect of hyperconjugation [244]. As shorter donor-acceptor distances are symptomatic of stronger hydrogen bonds, we postulated that β -sheets could be energetically more stable in α -helices. Our preliminary study on Protherm database agrees with this hypothesis, where the mean melting temperature of all- β proteins is higher than all- α proteins.

To rationalize the chemical nature of $\hat{H}$ angle preference of hydrogen bond involving main chain atom, we conducted a series of quantum mechanical simulations. Using simplified system of N-methyl-acetamide and methanol, *ab initio* simulation has reproduced the trend of increasing value of preferred $\hat{H}$ angle with main chain atoms (MC-MC > SC-MC > MC-SC > SC-SC). However, the simulated $\hat{H}$ angles are consistently smaller than observed values. The discrepancy could be attributed to several factors, for example oversimplification of the models, absence of any secondary interactions and adequate environment description. To rectify this, we proceed the simulation with more realistic system (though with less powerful simulation method) of Glycine dimers. In this simulation, the optimized $\hat{H}$ angles of MC-MC and SC-MC ($\sim 155^\circ$ and $\sim 135^\circ$ respectively) were accurately reproduced. We subsequently visualized the electron localization around hydrogen

bonding atoms. We found that the nonsymmetrical distribution of charge in SC-MC donor interacts with the dipole of acceptor group, and consequently rotates the acceptor and decreases the optimal $\hat{H}$ angle. We further showed that the symmetrical distribution of charge is sufficient to turn the optimal $\hat{H}$ angle back to $\sim 155^\circ$, as demonstrated in a the SC-MC system with artificial planarity constraint applied on the main chain backbone.

As compared to previous studies, we have shown that the electron density distribution about the donor amide group is the key determinant of the $\hat{H}$ angle, in contrast to the hybridization state of the acceptor as suggested previously [201]. Secondly, our results also show that donor-acceptor distance correlates with partial charge on donor, but not acceptor. This is in agreement with a previous *ab initio* quantum mechanical computations [208].

To illustrate the importance of these subtle differences in hydrogen bond geometries, we have demonstrated the means to refining low-resolution structures by utilizing these geometrical parameters. To achieve so, we have additionally revisited the “hydrogen bonding hypothesis” and verified that all donors and acceptors are either paired with one another or with water. We updated the scalability formula and showed that the new formula is accurate for all ranges of amino-acid chain length.

However, we discovered that the hydrogen bond number in low-resolution models is consistently lesser than predicted. In addition, the hydrogen bonds from these low-resolution models might not agree with one another. This strongly suggested inaccuracies and inconsistencies of hydrogen bonding in low-resolution structure.

With knowledge about the optimal number and geometry of different types of hydrogen bonds, we attempt to refine these low-resolution structures. Our refinement protocol involves 2 steps, (a) to accurately predict pairing of donors and acceptors by considering its feasibility and one-to-one pairing constraint (b) to optimize the structure using stereochemical, homology and the newly introduced hydrogen bond geometrical restraints. 145 NMR and 57 low-resolution X-ray structures have been chosen to test the refinement protocol. As a control study, hydrogen bonds from high-resolution structure of the same protein are used in parallel to the predicted hydrogen bonds.

The benchmark results showed that our protocol is capable of refining low-resolution model, with an improvement of 0.10 Å and 0.27 Å in R.M.S.D by using predicted and native hydrogen bond pairs respectively. In the case study of HSPC034, the degree of our refinement is shown equivalent to the state-of-the-art Rosetta method. The latter however has a much more complicated potential function and computational demand ($\sim 10,000$ samplings is required). In addition, our studies also show that to resolve structures accurately, the number of intra-protein hydrogen bond should scale ~ 0.8 times the number of amino acid residues. Accuracy of the structure decreases when the number of hydrogen bonds in the system deviates from this number (either higher or lower).

Finally, we have estimated the potential gains to functional annotation by our refinement methods by correlating our studies to pK_a estimations and data on H-X protection. Using the contemporary depth pK_a prediction method, our results show that the low-resolution structures and the NMR structures have computed pK_a s that are farther away from the experimental value than the higher resolution X-ray structures. The pK_a s computed using the refined models, while not quite as accurate as the higher resolution X-ray structures, are nevertheless statistically significantly better than the low-resolution/NMR structures.

In conclusion, in this chapter we have demonstrated that hydrogen bonds in proteins do not belong to one homogeneous class. Different hydrogen bonds have distinctively different geometry and strength under different environmental conditions. We have also shown that the knowledge of correct hydrogen bonding could be used to refine local inaccuracies in protein 3D models. In achieving so, we have devised a versatile refinement protocol using optimal hydrogen bond parameters learned from high-resolution structures. The protocol can generally be applied to different theoretical and experimental 3D models/structure of proteins. These refinements though numerically small in terms of R.M.S.D. measurement, could have functional implication. More accurate inference on physicochemical properties of protein, as evidenced in the ionizable amino acid pK_a prediction, are shown achievable from the refined models.

Chapter 5

Amino Acid Cliques Multibody Statistical Potential

5.1 Background and Motivations

Quantification of free energy of a protein at different conformations is indispensable in predicting its structure, dynamics and activity [17, 245–248] . And such knowledge is crucial to the annotation, and possibly design of the function of a protein [249–251]. Though it is possible to experimentally measure the free energy of protein, this approach is generally restrictive in terms of labors and cost. In the post Human Genome Project era, growing number of protein sequences and structures are being solved and deposited with high-throughput technologies. In accordance, efficient and accurate methods of computational prediction of protein free energy are in need.

One important use of protein free energy quantification is to characterize the effect of single amino acid mutations in protein. Such mutations can potentially destabilize a protein, leading to its malfunction and possibly diseased conditions. It is estimated that $\sim 80\%$ of Mendelian disease-associated single mutations are caused by protein destabilization effects [252]. This is the case in many common diseases, such as cystic fibrosis, and complex diseases, including diabetes and cancer [252, 253]. With large-scale projects such as Human Genome Project [253], HapMap Project [254], Exome Sequencing Project and the 1000 Genomes Project [255],

large number of single amino acid mutation (also known as a missense mutation) are being identified. It would be instrumental to have a fast and automated computational method to accurately predicted phenotypical effect of these mutations in order to identify association between genetics and diseases, as well as to understand phenotypical differences between human individuals.

Identifying the effect of non-synonymous single nucleotide polymorphism (nsSNP) could also provide valuable insights into the development of personalized medicine. One example is hOCTN1, a polyspecific organic cation transporters known to regulate the cellular uptake of the anticancer agents doxorubicin and mitoxantrone [256].

It has been shown that the transport function of protein was markedly impaired in some of the naturally occurring ns SNPs (Toh et al, 2013) (c.f. appendix chapter section 7.2). Cellular uptake of cationic drugs could be decreased to less than 10% in patients with these mutations. If sequencing data are available, an accurate prediction of SNP effect could suggest alternative class of medicines for the affected population.

Further, accurate prediction of mutational stability is also useful in protein design work. Recently, a protein design principle has been demonstrated in term of vaccine design [257]. Broadly speaking, immune system (antibodies, T-cell and B-cell) recognizes invading antigen proteins through "epitopes", a segment usually made up of five or six amino acids. To design a specific antigen to the epitope and without eliciting other unwanted immune response, the strategy is to use human proteins as scaffolds, where epitope residues are "grafted" onto. Technically, the grafting involves mutations of human proteins to the corresponding epitope's amino acid type. A key consideration of the method is to determine if the mutations are tolerated by the protein. A mutation could introduce destabilization and disrupt the protein scaffold. An accurate protein free energy quantification would be essential to assess the feasibility of such a design.

5.1.1 Reviews of mutational effect prediction methods

The approaches to predicting mutational effect can be broadly categorized into sequence-based and structure-based methods. The former methods usually perform a multiple sequence alignment of the protein, and utilize the sequence profiles to extract relevant information from different databases. These methods rely on previously known protein sequences and their characterizations deposited in databases. For examples, in the SIFT [258] method, mutational effect predictions were made based on a customized position specific substitution matrix (PSSM), constructed with PSI-BLAST [23] and MOTIF [259] finder to identify conserved local sequence regions. Polyphen [260] similarly used an improved version of PSSM, as well as examined the location of the mutation in terms of protein domain collected in the Pfam database [14]. The method also includes simple structural features such as accessible surface area and volume of an amino acid to make predictions. SuSPect [261] represents state-of-the-art in the sequence-based approach. This elaborated method had examined 77 features, before which was reduced to 9. Improvements made by SuSPect include better information extraction using PSSMs and Pfam domain profiles, position-specific known mutants and its conservation. Notably, the method also includes information from protein-protein interaction network, and would search the UniProt database for known functional annotations of a mutated position.

The second approach uses only structural information of protein, without exploiting sequence and evolutionary information. The majority of structure-based methods are based on machine learning methods, by training a predictor with known experimental thermodynamic data from the ProTherm database [207]. Different methods differ from each other in terms of the feature sets and machine learning architecture used. For example, I-mutant [262] was trained on pH , temperature and mutation type features using support vector machine with RBF kernel. Automute [263] constructed a statistical contact potential with Delaunay tessellation method, and trained their models with additional attributes such as ordered identities of amino acids, pH and temperature. PoPMuSiC [264] calculated 26

different statistical energy functions, and linearly combined them in an artificial neural network architecture. mCSM [265] utilized a graph metric to summarize physicochemical interactions within a cut-off distance as pattern signatures, and trained them with Gaussian process regression model. SDM [266] is unique in that it does not rely on machine learning but is a statistical method. The method constructs an environment-specific amino acid substitution matrix based on observed substitutions during the evolutionary time. DUET [267] is a meta-algorithm which consolidates the methods of mCSM and SDM.

5.1.2 Residue environment

This chapter describes a mutational effect predictor *Packpred*, based on an environment-dependent multibody statistical potential. The effect of a mutation critically depends on the protein environment. In this potential, protein environment is described using a set of spatially close amino acids, dubbed a ‘residue clique’. Stabilization effect associated with a residue clique could be estimated by its natural occurring frequencies in proteins. Commonly occurring cliques are assumed to be stabilizing, while rare or unobserved cliques represent unfavourable and potentially structurally disruptive environments.

We also utilized the measure of “residue depth” as characterized in the previous chapters. Depth measures the degree of burial (and hence solvation effect) on an amino acid residue, and has been shown to correlate well with structural stability [6] and free energy change of cavity-creating mutations in globular proteins [6, 89]. The parameter differentiates between different regions in protein. This is an important feature, as a hydrophobic clique occurring in the protein interior or on protein surface is likely to have different physicochemical environments as well as evolutionary constraints.

5.2 Cliques of amino acid residues

A clique is defined as a sub-graph in which all possible pairs of vertices are linked. A linkage distance $d(r_i, r_j)$ between two amino acids (r_i, r_j) is defined when at least

Figure 5.1: (A). A 5-residue clique (P9,W28,H89,Q96,L98) of cut-off 7.5 Å (shown in red stick and enveloped with molecular surface) from human recombinant MTCP-1 protein (PDB:1A1X). (B) zoom view of the residue clique. At least 4 of side-chain non-hydrogen atoms of the residues are within 7.5 Å (shown in dotted line) of each other. The structure is rendered using PyMol Chimera software suite [90]

four, or more than half of side chain non-hydrogen atoms (whichever is smaller) is within a cut-off distance (Figure 5.1) (for Glycine, C^α atom is used). Mathematically, given amino acid residues $\{r_i, r_j, \dots\}$ and their linkage definition distance $d(r_i, r_j)$ and a cut-off clique size d_{cut} , a clique C is defined when

$$C = \{r_1, r_2, \dots, r_i, r_j, \dots\} \mid \forall_{i,j} d(r_i, r_j) \leq d_{cut}\} \quad (5.1)$$

We define a (k, d_{cut}) “residue clique” to be a clique of k amino acids in spatial vicinity. Residue cliques defined with different combinations of k and d_{cut} (N ranges from 2 to 4, d_{cut} ranges from 7.0 Å to 10.5 Å in step of 0.5 Å) have been computed and investigated in this study.

5.2.1 Computation of residue cliques

Brute-force computation of cliques with fixed vertex parameter k is in polynomial time complexity there are $O(n^k)$ subgraphs in a graph with n vertexes to check. A slight improvement is applied on the brute-force computation to accelerate the computation. It is noticed that all k -th order subsets of a $(k+1)$ -th order clique must themselves form individual cliques. Hence we can test whether a $(k+1)$ -th can be formed by ‘growing’ from any of a k -th cliques. This process is repeated

from the lowest number of k ($= 2$, which is the linkage definition hash table) to highest desired number ($= 4$ in this study).

Algorithm 5 Clique algorithm

Input: *Clique, Residues, Neighbour*

Output: D

```

procedure CLIQUE(Clique, Residues, Neighbours)
   $D \leftarrow \emptyset$ 
  for all  $r_i \in \text{Residues}$  do
     $det \leftarrow true$ 
    for all  $r_j \in \text{Clique}$  do
      if  $\{r_i, r_j\} \notin \text{Neighbours}$  then
         $det \leftarrow false$ 
        break
      end if
    end for
    if  $det = true$  then
       $C' \leftarrow C \cup \{r_i\}$ 
       $D \leftarrow D \cup \{C'\}$ 
    end if
  end for

```

5.2.2 Derivation of *Packpred* statistical potential

A residue clique statistical potential was constructed by adopting the formulation of Sippl's potential of mean force [35],

$$E_c = -kT \log((P_c^{obs} + \alpha P_c^{exp}) / (P_c^{exp} + \alpha P_c^{exp})) \quad (5.2)$$

where c is some residue clique type $\{r_1, r_2, \dots\}$; P_c^{obs} is the observed number of residue clique c ; P_c^{exp} is its expected number in a hypothetical reference state without energetic interactions; α is the ratio of pseudo-count introduced to account for sparse statistics [35], and is taken as zero in our study. $-kT$ is assumed constant and will be omitted from further discussion.

In this study, the non-interacting reference state of a protein is modelled by reshuffling its constituent amino acids. In practice, P_c^{exp} approximated using the following formula

$$P_c^{exp} = \frac{z!}{\prod_{r \in R} n(r)!} \frac{\prod_{r \in R} \prod_{i=0}^{n(r)-1} (n(r) - i)}{\prod_{j=0}^{z-1} (N - j)}; \quad (5.3)$$

$$z = \sum_r n(r) \quad (5.4)$$

$$P_c^{exp} = \sum_{c \in C} P_c^{exp} / \sum_{c \in C} P_c^{obs} \quad (5.5)$$

where $n(r)$ is the number of residue of amino acid type r in the protein, N is the total number of amino acid in protein. For example, in a protein (or a protein region) of 100-amino acids that consists of 5 alanines, 10 isoleucine, 3 glycines (and 82 other amino acids), the un-normalized probability of random occurrence for residue clique (A,A,G,I,I) is $5!/(2!1!2!) \cdot (5/100)(4/99)(3/98)(10/97)(9/96) = 3.586 \cdot 10^{-5}$. The first term accounts for multiplicity of choosing from identical elements, and the rest of the terms account for sampling without replacement. Two variants of the potentials, $Packpred_0$ and $Packpred_d$ have been tested. The first variant uses all amino acids from the entire protein, while the second variant is depth-dependent. For each (N, d_{cut}) type of cliques, the statistical potential is built at 5 different levels of depth (2.80 Å - 5.25 Å, 4.25 Å - 6.25 Å, 5.25 Å - 7.25 Å, 6.25 Å - 8.25 Å, 7.25 Å - ∞).

5.2.3 Training of the potential

5.2.3.1 Training data set

A set of non-redundant (30% sequence identity cutoff), high quality (resolution < 2.5 Å, R-free < 0.25), single chain protein structures with chain length 100 - 200 amino acids were downloaded from Protein Data Bank to train the statistical potential. The training set consists of 3753 proteins.

5.2.3.2 Training procedure and parameters

It is noticed that protein structure could change with molecular vibration even at near-equilibrium position. To capture these dynamics, clique profile of a protein is taken from multiple models of its structure. For each protein, 10 homology models were built using Modeller [29] (CHARMM22 Force Field, default Modeller option: very-slow) with the native structure as the structural template. The maximum R.M.S.D. among models is ~ 0.5 Å. These homology models were used together with the native structure (i.e. 11 structures for each protein) to train the potential. Cliques of order $k = 2, 3, 4, 5$ with threshold sizes of 7.5 Å to 11.0 Å in step of 0.5 Å were computed for every protein structure in the data set.

5.3 Validation of *Packpred* statistical potential

5.3.1 Recognition of native structure among structural decoys

Recognition of native structure among decoys is a weak test of statistical potential [31]. In this test, the statistical potential scores a number of models with modelling error artificially introduced. The *Packpred* potentials are tested on Melo’s decoy set [120]. The decoy set consists of 20 proteins, each with a native model and 300 decoys. The result (Table 5.1) showed that both variants are reasonably well at recognizing native model (average ranking of ~ 3.65). Interestingly, *Packpred* is capable of recognizing the native structure of PDB:2PNA. The structure is solved by NMR, and other potentials had not been able to distinguish between the native state of the protein from the decoys [37].

5.3.2 Correlation with structural stability

Ultimately, the purpose of protein 3D modelling is to explain and predict biochemical activity of the molecule. In regard to this, we have also tested our potential on functional stability of two protein systems with saturated mutagenesis data, the T4-lysozyme and Controller of cell division or death B (CcdB).

	1bbh	1c2r	1cau	1cew	1cid	1dxt	1eaf	1gky	1lga	1mdc	1mup	1onc	2afn	2cmd	2fbj	2mta	2pna	2sim	4sbv	8i1b	correct	average
S_d	4	1	1	3	0	2	1	12	3	6	4	12	2	1	1	1	2	24	1	1	10	3.6
S_0	2	1	1	1	0	0	1	2	1	1	1	2	1	1	1	1	52	1	1	1	16	3.7

Table 5.1: Ranking of native structure by *Packpred* potentials on Melo’s decoy set. S_d and S_0 refer to the versions of Packpred scores with and without residue depth component, respectively.

5.3.2.1 Experiment setup

(A) Saturated mutagenesis data sets

Previously published saturation mutagenesis of two proteins, T4-lysozyme [268] and Controller of cell division or death B (CcdB) protein [95] were analyzed in this study. T4 Lysozyme is 164 amino acid residues protein. The native structure of the protein (PDB: 2LZM) had been solved crystallographically at a resolution of 1.7 Å [269, 270]. Each position except the first had been mutated to 13 types of amino acids (A, C, E, F, G, H, K, L, P, Q, R, S, T). The data set consists of 1966 mutations after excluding key catalytic site residues (D10, E11, R145, R148P). CcdB is a cytotoxin (an inhibitor of DNA gyrase) of length 101 amino acids. Its native structure has been solved at 1.4 Å resolution level and deposited in the Protein Data Bank (PDB: 3VUB) [271]. Full saturated mutagenesis (all 20 amino acid types) had been performed on all positions of the protein. A final set of 1561 mutations, after removal of active site residues (I24, I25, N95, F98, W99, G100, I101), was obtained.

(B) Mutational sensitivity score

In both saturated mutagenesis data sets of T4-lysozyme and CcdB, an assessment was made by the experimentalists on the phenotypical effect for every mutation. We refer to their assessment scores collectively as “mutational sensitive (MS) score” in this study. For T4-lysozyme data set, the score is assigned based on phenotypical effect based on the plaque-forming ability of the mutant. Subject to the same experiment condition, a mutant is assigned to one of the four levels of sensitivity if

the size of the plaque are (i) similar to native control, (ii) significantly smaller, (iii) with hazy morphology or difficulty in discerning plaques and (iv) no plaque formation [268]. For CcdB data set, the mutational sensitivity score was quantitatively defined as the titer number at which the protein activity (in this case, inducing cell death) decreases by 5-fold or more relative to its previous dilution [95]. A total of 8 levels of sensitivity is defined for the CcdB data set.

(C) Scoring protein and its mutants

Scores of all residue cliques are needed in computation of the total score of a protein structure. To score a residue clique, the mean μ and standard deviation σ of its depth was first computed. The depth of a residue clique is defined as the average depths of its constituent atoms. A Gaussian probability density function $N(x|\mu, \sigma^2)$ was then accordingly built. The clique score was computed as the weighted sum of integrand at every depth level, namely

$$S_{\mu, \sigma}^c = \sum_{d \in D} \frac{1}{d_f - d_i} \int_{x=d_i}^{x=d_f} E_d^c \cdot N(x|\mu, \sigma^2) dx \quad (5.6)$$

where d is one depth level, d_i and d_f is the lower and upper bound of the level.

Most residue cliques in a protein are overlapping with one another, and an amino acid residue can participate in multiple cliques. The score of a residue is taken as the average of the clique scores of all the cliques it participates in. The score of a protein is further taken as the average of all its residue scores.

To score a mutant, we simply modify the clique type of the cliques comprising the mutated residue accordingly. The mutant protein score is now computed using the mutated cliques. Here we have implicitly assumed that the neighbouring context of residues in the protein would not change upon a single point mutation. No explicit 3D modelling, clique computation or depth calculation were performed for the mutant protein.

5.3.2.2 Results

We validate the utility of prediction of *Packpred* potential by two measures, (i) its ability to discriminate destabilizing mutations from neutral mutations, and (ii) rank-ordering the degree of phenotypical change among mutations.

(A) Binary classification of neutral and destabilizing mutation

A mutation is regarded as neutral if there is with no perceptible phenotypical difference as compared to native sequence (MS score = 2 in CcdB; MS score = 2, 3 in T4-lysozyme), and regarded as destabilizing otherwise. For T4-lysozyme, there are 1363 ($\sim 69\%$) neutral and 616 (31%) destabilizing mutations; whereas in CcdB data set, 1259 ($\sim 80\%$) of mutations are neutral and 303 ($\sim 20\%$) are destabilizing.

Packpred potential only reports a log-odd score of likelihood with no absolute scale. To perform binary classification, a threshold value is to be determined. To achieve so, we have used the T4-lysozyme mutational data as the calibration set, and the threshold value was obtained by optimizing its MCC value. The optimization first estimated the value using brute-force method and subsequently refined it using conjugate gradient and Nelder-Mead method. For (3, 10.0) variant, the value was determined to be ~ -0.01 .

The different methods we have compared ourselves to have provided their own default threshold values. However benchmark testing showed that by using these default thresholds, all other competing methods would have low predictive power ($\text{MCC} < \sim 0.2$). Hence, we have also performed the same recalibration for these other methods using the T4-lysozyme set.

To measure the efficacy of the methods, for both T4-lysozyme and CcdB, the data-set was randomly divided into ten equally-sized subgroups, each consists of 90% of all data. Each mutation was predicted to be either neutral or destabilizing depending on whether its score is above and below the threshold values. The same recalibrated threshold values were used for both T4-lysozyme and CcdB data sets for all methods without further adjustments.

Methods	CcdB				T4-Lysozyme			
	Default		Recalibrated		Default		Recalibrated	
	μ	σ	μ	σ	μ	σ	μ	σ
mCSM	0.154	0.008	0.389	0.01	0.048	0.008	0.223	0.006
SDM	0.234	0.012	0.329	0.011	0.203	0.008	0.244	0.004
DUET	0.192	0.008	0.399	0.009	0.087	0.006	0.240	0.007
SIFT	N/A	N/A	N/A	N/A			0.291	0.005
AutoMute 2.0	0.093	0.007	0.212	0.009	0.054	0.007	0.226	0.008
I-mutant 2.0	0.148	0.003	0.361	0.007	0.100	0.007	0.302	0.006
SuSPect	0.212	0.004	0.403	0.005	0.266	0.007	0.376	0.004
Packpred (2, 10)			0.459	0.009			0.372	0.007
Packpred (3, 10)			0.471	0.010			0.373	0.006
Packpred (4, 10)			0.400	0.010			0.374	0.005

Table 5.2: Performance of binary classification of neutral and destabilizing mutations of different methods as measured by Matthews correlation coefficient (MCC).

The results (Table 5.2) indicated that *Packpred* had similar discriminating power on T4-lysozyme data set ($MCC \sim 0.374$) as SuSPect ($MCC \sim 0.376$), and is appreciably better than the other methods ($MCC \sim 0.24$). On CcdB dataset, *Packpred* ($MCC \sim 0.47$) is significantly better than all other methods compared ($MCC \sim 0.40$).

(B) Rank ordering the degree of phenotypical change by mutations

Apart from binary classification, it is also favourable for a statistical potential to be capable of rank ordering the degree of phenotypical change caused by single amino acid mutations. The degree of change is measured by the mutational sensitivity score, which categorizes each mutation into one of 4 and 8 levels in T4-lysozyme and CcdB data sets. We chose to use Spearman’s rank correlation coefficient (SCC) to measure the performance of rank-ordering, as we made no assumption on linear relationship between the scores and the phenotypical change.

To calculate the predicted ranks, all mutations were first categorized by their MS levels. For each level, the mean of mutational scores were computed. The mean scores were then compared with one another and ranked accordingly. As number of mutations at different MS levels are considerably different from one another

(e.g. there are large number of neutral mutations at MS=2), another robust and iterative version of measurement, with equal sampling at different MS levels was also proposed. In this approach, 10 mutations from each MS level are randomly sampled to compute the SCC value between mutational scores and MS levels. This sampling process is iterated until the mean of SCC value converges (~ 150 and ~ 200 samplings for T4-lysozyme and CcdB respectively).

The result (Table 5.3) revealed that quantifying mutational strength in T4-lysozyme data set is straightforward, with all but the SuSPect method achieving perfect correlation ($SCC = 1.0$) using all data. This is in contrast to the more challenging CcdB data set, where no other method except *Packpred* has achieved a satisfactory score. *Packpred* reported a near perfect correlation ($SCC = 0.976$) in the data set. In addition, *Packpred* also had the best performance ($SCC \sim 0.740$) when SCC was calculated using the equal sampling approach ($SCC \sim 0.285$ for SDM (the best other method)).

It is noted that the performance of the 4-residue clique potential is worse than 2-residue clique and 3-residue clique. This is counter-intuitive as higher-order cliques are assumed to better capture cooperative effect among residues. We presume that the reason for the weak performance is due to sparseness of data. As the order of clique increases, the number of different types of cliques grows super-exponentially. While a 2-residue clique potential consists of 210 terms, 3-residue clique potential consists of 1540 terms, and 4- and 5-residue cliques would consist of 8855 and 42504 terms. With our dataset, the average number of observation per bin is ~ 106.8 , ~ 322.7 , ~ 56.2 and ~ 5.8 for 2-, 3-, 4-, and 5-residue clique respectively. Furthermore, this is disregard of the fact that cliques with high depth value and rare amino acid types would have a far lesser number than the average.

5.4 Web-Server

The *Packpred* has been made accessible through a web-server at <http://cospi.iiserpune.ac.in/Packpred> (Figure 5.2). Users submit input by specifying the four-letter PDB code or upload

METHODS	CcdB			T4-Lysozyme		
	ρ	μ	σ	ρ	μ	σ
mCSM	-0.143	-0.104	0.148	1.00	0.591	0.392
SDM	0.310	0.285	0.272	1.00	0.707	0.296
DUET	-0.119	-0.098	0.154	1.00	0.643	0.361
SIFT	N/A	N/A	N/A	N/A	N/A	N/A
AutoMute 2.0	-0.048	-0.008	0.226	1.00	0.551	0.424
I-mutant 2.0	-0.239	-0.116	0.166	1.00	0.731	0.284
SuSPect	0.119	0.207	0.216	0.8	0.805	0.212
Packpred (2, 10)	0.976	0.701	0.190	1.00	0.808	0.187
Packpred (3, 10)	0.976	0.74 0	0.164	1.00	0.811	0.167
Packpred (4, 10)	0.833	0.659	0.187	1.00	0.808	0.168

Table 5.3: Performance of rank-ordering of degree of phenotypical change as measured by Spearman’s rank correlation coefficient. ρ represents values calculated using all data, and μ and σ represent the mean and standard deviation values calculated with equal sampling approach.

a protein structure in PDB format. User can also choose to perform specific single point mutations, alanine scanning mutagenesis or saturated mutagenesis. The results of the prediction are returned in both pictorial representation and tabular formats. Users can download the full results, or view the top 20 stabilizing, destabilizing and native-like mutants. Standalone version of the program is also available for download.

5.5 Chapter Summary

An environment-dependent multi-body statistical potential, *Packpred*, was constructed to predict the effect and strength on a single amino acid change. On a validation set of ~ 3500 mutations from saturated mutagenesis experiments, we established that *Packpred* has comparable or better performance in predicting neutral from functionally destabilizing mutants, and is the only method capable of rank-ordering the strength of mutational effects.

Packpred is distinct from other methods as it made no use of machine learning architectures, but is a pure statistical potential, calibrated only with the naturally

Figure 5.2: Snapshots of input page and sample output page the web server (<http://cospi.iiserpune.ac.in/Packpred>).

occurring frequencies of residue environments. The predictor does not have any free parameters for data fitting, and made no use of thermodynamic or mutational data. Hence it does not suffer from over-fitting errors as is the case of many other predictors, which have been shown to have drastically lower performance in blind tests [266]. Also, *Packpred* uses only structural information, and does not rely on evolutionary information or previously annotated mutations.

Lastly, the predictor has been made accessible through a web-server as a convenient tool for studying and designing the function of proteins.

Chapter 6

Discussions and Conclusions

This chapter summarizes and concludes our study of the physicochemical environment of proteins. The study is divided into three main topics of (1) the characterization of the burial measure of residue depth, (2) the characterization of hydrogen bonding in proteins, (3) the derivation and validation of environment-dependent, multibody statistical potential *Packpred*. Individual section in this chapter discusses the conclusions and possible future directions for each of the topics.

6.1 Residue Depth

6.1.1 Summary

In this section we investigated the biophysical parameter of depth, which measures the distance of an atom/residue to its nearest surface bulk water. We have established the utility of depth as a concise protein interior environment descriptor. Besides the correlations with physical properties previously known, we characterized several other distributions of amino acids. For example their spatial distribution in the protein interior, and subtle conservation pattern that cannot be described with the conventional methods, such as solvent accessibility or substitution matrices. Aided with a better description of the protein environment, we devised two biophysical utilities of (1) the design of temperature-sensitive mutants of protein, and (2) prediction of amino acid residues likely to form small molecule binding sites.

These utilities have found a wide variety of applications, for example to accelerate the mutant selection and experiment design, protein engineering, functional prediction of SNPs and the design of drug or vaccine molecules. The correlation of depth with protein energetics is also further attested with two other work developed in collaboration, namely (3) the prediction of amino acid pK_a and (4) volume estimation of protein internal cavities. The measure is also used throughout our work in characterization of other physicochemical properties of protein environments. Given its usefulness, It is proposed that depth be used concurrently with solvent accessibility to gain a more comprehensive view on the burial of a residue, and the utilities we developed should be adopted into the routine analysis of protein structure.

6.1.2 Suggestions and future directions

One consideration for further development of the depth measure is in regard of the dynamics of protein system. In its current implementation, the dynamics of solvent molecules are approximated with the iterative resolutions of the protein in different orientations. In these resolutions protein is treated as a rigid-body with no internal dynamics. This is likely an oversimplification given that the fact that dynamics of protein and interfacial waters are known to regulate many of its functions [272]. Also, depth is defined using only one single water molecule (the nearest). This could compromise the robustness of the depth computation in cases where the placement of waters are substantially different between resolutions.

In regards to this, we suggested that further development of depth would need to incorporate a more realistic sampling method of dynamics. At the first level, fast approximation method such as anisotropic network models [273, 274] could be employed; and more specifically, depth analysis could be incorporated into molecular dynamics trajectories analysis. The advantages of this approach is that the movement of atoms follows physics-based rules, and that the distances between every protein atom and water are already computed in the MD simulation - only a simple comparison is needed to find the smallest value. An added function of

this incorporation is the detection of novel binding sites on protein. Recent works have suggested that in some of the previously thought undruggable proteins, temporarily accessible binding sites could be found along its MD trajectory potential to be targeted by small molecule inhibitors. With the rapid binding site detection method of depth, these unconventional binding sites could be retrieved easily for further characterizations and to guide the development of their inhibitors.

Lastly, we note that the interaction between solvent molecule and protein amino acids is reciprocal. While out of the scope of our study, it is suggested that water molecules in close contact with protein could exhibit different behaviours than the bulk solvents. This is also evidenced with the recent development of grid inhomogeneous solvation theory (GIST) [275]. An ‘inverse’ depth value could be assigned to these waters for their further analysis.

6.2 Hydrogen Bonding in Proteins

6.2.1 Summary

We next characterize the physicochemical environmental dependence for the other important interatomic interaction of hydrogen bonding in proteins. By empirical statistical analysis, we relaxed the lower bound of donor-acceptor-antecedent angle (dubbed $\hat{H}$ angle) from the conventional 120° to 100° . We then classified hydrogen bonds in protein based on whether their donor and acceptor atoms are from main chain (MC) or side chain (SC). The geometry of different subtypes of hydrogen bonds are separately investigated. We discovered that all bond lengths monotonically decrease with depth, and $\hat{H}$ angle of hydrogen bond types are different from one another. Particularly, we were intrigued by the MC-MC $\hat{H}$ angle of 155° , which is distinct from 120° observed in other small organic molecules. We also showed that this angle is essential for the formation of regular secondary structure of α -helix and β -sheet. Quantum mechanical calculation showed that this preferred $\hat{H}$ angle is the result of the symmetrical distribution of charge around the donor atom, itself the consequence of adjacent planar peptide bond.

Equipped with the knowledge about geometrical distribution of hydrogen bond in proteins, we attempted to utilize them in the task of refining low-resolution protein models. First a simple yet accurate predictive model of total hydrogen bond number in protein is built. We discovered that the accuracy of a protein 3D model inversely relates to the number of missing of hydrogen bond. To recover these missing hydrogen bonds, a second model is built by predicting likely donor-acceptor pairs. Predicted pairs are restrained to form a hydrogen bond with the corresponding ideal geometry. Benchmarks on low-resolution X-ray and NMR models showed that this hydrogen bonding based refinement protocol has a performance on par with the state-of-the-arts Rosetta method, yet is much less computational intensive. The refined models are also shown to better agree with experiments such as hydrogen-deuterium exchange ($\sim 10\%$ accuracy improvement) and pK_a prediction (~ 0.6 pH units).

While the degree of improvement could vary (-0.005 - 0.025 in R-factor, -0.5 - 1.5 in R.M.S.D.), our protocol almost never deteriorates the structural accuracy of input protein structure. Given the speed and benefits of the protocol, we suggest that hydrogen bond refinements should be mandatory features of all X-ray structure refinement when the resolution of the structure is $2.1 \text{ \AA}$ or poorer. NMR structure refinement could also benefit from similar refinement schemes.

6.2.2 Discussions and future directions

Additional environment for characterization

While our work in characterizing protein hydrogen bond is extensive, not all protein environments have been investigated. Two such areas are (1) interfacial hydrogen bonds, which include protein-solvent, protein-protein, protein-ligands (e.g. small molecule, DNA or RNA) etc. The characterization of these hydrogen bonds could provide crucial information in the understanding and engineering of molecular recognition; (2) cooperativity among the hydrogen bonds interaction. In our study hydrogen bonds are mostly studied in isolation, while in proteins hydrogen bonds often form extensive, large scale networks. A typical example is the secondary

structure. The properties of these structures are hypothesized to be dependent on the interactions among the constituent hydrogen bonds. Many of these properties, for example the dipole moment (or higher order terms) of α -helices, the formation of super-secondary structures, are known to essentially regulate biological processes.

Possible improvements in refinement protocol

In our protocol of protein structure refinement, we have included a simple hydrogen bonding partner predictor. The predictor is based on the distance and preference of donor and acceptor candidates. In the benchmark result, we have shown that the knowledge of actual donor-acceptor pairs could substantially improve refinement as compared to the prediction. This means that our protocol can be further improved if additional information was available, for instance the candidate pairs could be learned from experimental data or homologous structures, should such structures exist for the proteins of our interest.

Incorporation of hydrogen bonding in force-fields

Our work has revealed the intricate geometrical distribution of hydrogen bonds and their dependence on protein environment. Several other applications that implemented the use of chemical force-fields, for example the solving of X-ray or NMR structures and molecular dynamics simulation, could benefit by incorporating a hydrogen bonding term derived from our analysis. In particular, in MD simulation atomic depths of donors and acceptors can be calculated to determine the strength / length of hydrogen bonds. We also conjecture that the inclusion of $\hat{H}$ angle restraint would facilitate the modelling of secondary structure formation. The incorporation of the hydrogen bonding term is not only empirical, but also physically justified as we have established its quantum mechanical nature.

6.3 Multibody Statistical Potential *Packpred*

6.3.1 Summary

In this chapter, we have explicitly defined micro-environments composed of tightly packed amino acids in proteins. We dubbed the micro-environments ‘residue cliques’

as its definition is inspired by the concept of cliques (complete sub-graph) in graph theory. The natural frequency of occurrences of different types of residue cliques in different environments (quantified by depth) are counted. Following Sippl’s formulation, the statistics are summarized in the form of a multibody statistical potential, which we named *Packpred*. The correctness of the statistical potential is benchmarked by correlating with energetic/functional destabilization caused by protein mutation. We tested the potential with CcdB and T4-lysozyme systems where the saturated mutagenesis data are available. As compared to other sophisticated, purpose-specific machine learning algorithms, our method is superior in both the tasks of binary classification of destabilizing mutants, as well as rank-ordering the degree of destabilization. It is also noteworthy that our model has a straightforward free energy interpretation by assuming Boltzmann distribution. The model does not have free parameters to train and hence does not suffer from overfitting problem like the other black-box models.

6.3.2 Suggestion for improvements

Notwithstanding the superior performance of *Packpred*, we have identified possible further improvements to the potential. First, the current potential primarily focused only on the packing of amino acids, while other features and interactions have not been carefully incorporated. These interactions include hydrogen bonding, disulphide bonds, long-ranged interaction etc. We have also observed that the more elaborate 4-body variant potential had weaker performance than 3-body potential. The reduced performance could be due to insufficient data to calibrate the higher order cliques, as the number of clique types increases exponentially with its number of residues. While 3-body potential consists of 1540 terms, 4- and 5-body clique would have 8855 and 42504 terms respectively. Lastly, in both scenarios of estimating the reference state of the potential and mutant scoring, we have assumed that the residue clique profile of a protein does not change under the permutations of amino acid residues. In actuality however, premutation of a large residue (e.g. phenylalanine, 7 side-chain atoms) to smaller amino acid (e.g. valine, 3 side-chain

atoms) could possibly introduce a void in the packing, and hence invalidate the clique definition (a densely cluster of atoms in space). In the opposite case, the new cliques could form when a smaller amino acid is permutated with a larger amino acid and the local packing of atom is increased.

A more general suggestion is on the choice of the reference state. In *Packpred* and many other statistical potentials (e.g. DFIRE, DOPE), an ideal gas model is used. The model assumes each amino acid in the protein could localize itself freely in the protein volume, independent of other amino acids. In reality however, the maximum distance between two amino acids in space is bounded – it cannot exceed the sum of peptide backbone lengths between them. Therefore, amino acids that are sequentially close are also expected to be close in 3D space [276, 277]. In other words, ideal gas model and reshuffling methods overestimate the background frequency of interactions between sequentially close residues. A better model for reference state for polymer would be ideal chain model or worm-like model. These models intrinsically capture the sequence-spatial dependence by describing a polymer as uncorrelated, self-avoidance random walk of sequence of amino acids. It is suggested that these models be considered in future development of force-field, especially when continuous sequence motifs are involved in the modelling.

6.4 Concluding Remarks

In this study we have extensively characterized the physicochemical environment of proteins. The characterizations were studied through the proposal and analysis of several novel parameters. We first characterized 'depth', which is a measure of burial of an atom/amino acid from the bulk solvent. It is shown that geometrical and chemical environments of protein are accurately described by residue depth. On the protein surface, used together with the solvent accessibility, depth could identify protein pockets which are often ideal targets for its regulatory control with small molecule binding (e.g. drug / vaccine molecules). Using depth, we devised a fast yet accurate binding site prediction method. The method performs on par with the other state of the art methods, but identifies different binding sites. Given

these properties, it is well suited to be integrated with newer sampling methods (e.g. MD simulation) to identify unconventional binding sites.

In the protein interior, internal cavities could be located using residue depth by applying a grid-based Monte Carlo procedure with it. The thermal stability of proteins could be modulated with these internal cavities. For instance, to further stabilize the protein, one can fill up these cavities through mutations that substitute small amino acids with larger counterparts, as this would introduce more van der Waal's interactions as well as better packing of atoms in the protein interior. On the contrary, the stability of protein could be lowered by introducing cavity-creating mutation. Some of these destabilized proteins could be temperature sensitive. The activities of temperature sensitive proteins can be easily and reversibly controlled by adjusting environmental temperature, rendering a powerful tool for experimentalists to study the mechanisms of many cellular processes.

After establishing the utilities of depth, we used it to measure the environmental dependency of the other important interactions of hydrogen bond. We showed that while hydrogen bond length (parametrized by donor-acceptor distance) decreases with increasing depth, bond $\hat{H}$ angles (donor-acceptor-antecedent angle) mostly remain constant at different environments. We subsequently rationalized that this preferred $\hat{H}$ angle of hydrogen bond between in main chain atoms is a consequence of quantum chemical nature of peptidic bonds and spatial distribution of electron orbitals. We further showed that the $\hat{H}$ angle is consequential to the formation of regular secondary structure of α -helix and β -strands. Given the knowledge of hydrogen bond geometry, we attempted to refine low-resolution protein model with it. Apart from geometry, we also rediscovered the linear scalability of the number of hydrogen bond with protein chain length. A survey however revealed that low-resolution protein models very often missed the number predicted by the scalability rule. We hence designed a simple protocol to identify the most likely donor-acceptor pairs to recover the missing hydrogen bonds. Our protein structure refinement protocol involves the introduction of these predicted hydrogen bonds, restrained to their ideal geometry at different (depth) environments. The refinement protocol was tested on low-resolution models solved by X-ray crystallography and NMR

methods. At overall structure similarity with known high-resolution structure, the protocol achieved a modest improvement. More importantly however, the recovered hydrogen bonds now better describe (1) protein energetics and (2) protein dynamics. The former is verified with the higher accuracy when the refined models are used in ionizable amino acid pK_a prediction; the latter was shown to have a better correlation with hydrogen-deuterium exchange experiment. Given its versatility we suggest that the refinement protocol can be included in different modelling packages, for both comparative modelling and refinement of low resolution-experimental structures.

While the previous chapters have focused on the effect of localization, in chapter 7 we turned our focus on biochemistry, and attempted to explicitly define local micro-environments of amino acid residues. We borrowed the concept of clique from computational geometry to describe a set of tightly packed amino acids as a micro-environment. The stabilization effect of different residue cliques were estimated from their observed occurrences in protein data bank, using Sippl's formulation of statistical potential. The *Packpred* potential is also constructed in a depth-dependent manner. When tested with protein systems for which saturated mutagenesis data (namely CcdB and T4-lysozyme) was available, we showed that the potential has the best agreement with protein activity (in both classification and rank-ordering tasks) as compared to other state of the art methods. The capability of predicting mutational destabilization renders the *Packpred* potential a useful tool in protein design.

In summary, we have extensively characterized the physicochemical environment of different features in proteins. For each feature studied, we have also devised either predictive protocols or testable hypothesis based on their characterizations. The biological utilities of our work are established as benchmark tests on the protocols showed their equivalent or superior performance over conventional methods; and also when our hypotheses corroborated with experimental validations. Apart from theoretical investigations, we also developed tools to assists other researchers, especially experimentalists. We believe that the work presented here makes a positive impact on the current and future research of protein structures and beyond.

Chapter 7

Appendix: Modelling Non-Globular Protein Systems

Apart from globular proteins, two other important classes of structural problem are the modelling of the 3D structure of protein-protein interfaces and non-globular proteins (e.g. fibrous, intrinsically disordered and membrane proteins). Despite their important roles in molecular recognition, metabolism regulation, structural integrity etc, the properties of these structures are much less understood as compared to globular proteins. The 3D structures of non-globular proteins are scarce due experimental limitations. For example, membrane proteins constitute of $\sim 25\%$ of all proteins and more than 40% of drug targets, yet less than 600 unique structures have been solved (compared to the size of entire PDB at $\sim 116,000$) [278]. Insufficient number of known structures poses a challenge for extracting statistical rules for these proteins. In contrast to large scale database study, characterization of physicochemical environment of non-globular proteins would involve focused study of a few examples, the use and design of miscellaneous techniques and importantly, the circulation between theoretical and experimental work. In this chapter, I will describe several studies we have undertaken involving the modelling of non-globular proteins. All the work presented here are in collaboration with experimentalists. The theoretical developments provide suggestions to guide and/or to be tested by experiments. These work include (1) modelling of cell division regulating protein complex of cohesin ring (collaboration with Dr. Dmitry Ivanov (Bioinformatics

Institute A*STAR, Singapore), (2) modelling of cationic molecule transporters of hOCTN-1 and -2 (collaboration with Dr. Dorothy Toh (Yong Loo Lin School of Medicine, National University of Singapore), (3) modelling of protein-protein interface of oncogenic proteins of VAV1-EZH2 (collaboration with Dr. Su I-Hsin, School of Biological Sciences, Nanyang Technological University).

7.1 Modelling of Cohesin Complex

7.1.1 Background and motivations

Cohesin is a protein complex that regulates the separation of sister chromatids during cell division [279]. It has been shown that cohesin complex forms a ring-like structure that encircles the sister chromatids during metaphase [280]. The mechanisms of cohesin functions, for example the initial encircling of chromatids, its role in cell cycle checkpoints and the final release of the chromatids etc, are however unclear. One of the first steps to probe into these mechanisms is to identify critical residues in cohesin that interacts with other ligands (e.g. histones, DNA and/or other regulatory proteins).

Several mutation experiments that involve multiple lysines to arginines mutations had been performed by our collaborators. A lysine to arginine mutation would preserve the positive charge state and thus Coulombic interaction of a residue with the negatively charged histone proteins. However the work of our collaborator (Dr. Dmitry Ivanov) have shown that, mutating all lysines to arginines in the protein is lethal to the organism. This shows that preserving Coulombic interaction is not sufficient to maintain the wild-type interaction. Further experiments mutating fractions of lysines to arginines at different regions of the protein have different effects in the resultant mutants, ranging from lethal, sterile to near wild-type. These results suggested that certain lysines are more critical in maintaining the functions of the proteins. While in principal mutation experiments could be used identify functionally important residues or regions in the protein, a full mutation is however practically impossible, considering the large number of combinations

possible from ~ 260 lysines in the protein. It is therefore instructive to narrow down the lysines candidates for further mutation experiments. To achieve so, a 3D model of the protein is attempted.

Cohesin ring complex is formed by SMC1 and SMC3 components. The proteins (in the model organism *S. cerevisiae*) are 1227 and 1232 amino acids long respectively. The sequence of SMC proteins is known to consists of two helices, intercalated between an ATPase domain and a hinge domain. The two helices, dubbed ‘forward’ helix, and a ‘backward’ helix from were expected to form antiparallel coiled-coil interaction with each other. As a large protein with coiled-coil architecture, no current experimental techniques (NMR, X-ray crystallography) are capable of solving its structure in high-resolution. Further, conventional comparative modeling technique is also ineffective as no full template is available. Instead, a multi-level modeling approach is attempted here. Modelling of the complex structure is done in three steps, including (1) modelling of globular domains (2) modelling of coiled-coil region (3) modelling of SMC1/SMC3 interaction.

7.1.2 Modelling protocol

(1) Modelling of globular domains

A homology model can be built for the globular domains of ATPase and hinge domain, as the 3D structures of their close homologs have been experimentally solved. The hinge region was modelled using known structure of mouse SMC hinge heterodimer (PDB:2WD5) as template, while the ATPase domain was modelled with template from *Pyrococcus furiosus* (PDB:1XEW). The sequence alignments of hinge region and ATPase domain were performed separately using BLAST [16], after which they are joined together to form the final alignment. The template-target sequence similarity is $\sim 34\%$ (SMC3 - hinge 35%, SMC3 - ATPase 34.5%; SMC1 - hinge 31%, SMC1 - ATPase 34.5%). The assumed coiled-coil region (S175-L495; Q695-R1060) has no restraints implemented during this phase of modelling. An initial model was built using Modeller with the default “slow” optimization schedule.

Figure 7.1: Schematic representation of coiled-coil interaction

Figure 7.2: Paircoil2 prediction of coiled-coil structure in SMC1 (left) and SMC3 (right). P-value thresholds of 0.2, 0.1 and 0.05 are represented in purple, blue and green lines respectively.

(2) Modelling coiled-coil interactions

(A) Coiled-coil register assignment

A coiled-coil interaction is a structural motif involving the tight binding of two α -helices. Helices involved in coiled-coil interactions consist of a signature sequence motif of heptad repeat (Figure 7.1). The sequence motif had been studied extensively, and a high accuracy predictor of paircoil2 [281] was used to predict the region and registers (a, b, c, d, e, f, g) of SMC1 and SMC3 sequences. The prediction was run through the paircoil server at <http://paircoil2.csail.mit.edu/paircoil2.html>. From the prediction the forward helix spans from position 175 to 495 for SMC1, 174 to 481 for SMC3, and the backward helix from 695 to 1060 for SMC1, 706 to 1038 for SMC3.

(B) Coiled-coil interaction assignment

Figure 7.3: Bissulfosuccinimidyl suberate cross-linker

To identify the mode of coiled-coil interaction, our collaborators have performed random chemical cross-linking experiments on the protein complex. In the experiment, Bissulfosuccinimidyl suberate (BS3) cross-linker was used. BS3 has 8-atom space arm, which could link two lysines residues within 11.4 Å away by covalently interact with their side chain amine groups. The experiment is of low-resolution, with only 3 useful cross-linkings, i.e. K414-K764, K432-K764 in SMC1, and K180-K1034 in SMC3 have been identified.

A manual assignment of coiled-coil interactions between the forward and backward helices was performed separately for SMC1 and SMC3. The assignment was performed according the following constraints:

- (a) Cross-linkings identified from experiments are fulfilled
- (b) Interactions from the helices arms follow antiparallel coiled-coil interaction pattern
- (c) The preference of assigning a position is based according to paircoil2 prediction.
- (d) The number of non-interacting coiled-coil heptamers are minimized.

In the final assignments there are ~ 49 heptamer repeats in SMC1 and ~ 46 repeats in SMC3. In SMC1, a 67 amino-acid residues loop was predicted.

(3) coiled-coil region modelling

The 3D model of coiled-coil regions can be modelled after the register assignment. Coiled-coil region has a signature heptad repeat pattern, hence the spatial distances among specific atoms (C^α atom are used as representatives here) are fixed. These spatial restraints among representative atoms are learned from a template structure

Figure 7.4: 3D model of yeast SMC1 (top) and SMC3 (bottom). The structures are rendered using PyMol software suite [90].

of Rad50 protein (with 4 complete and 1 incomplete heptad repeat). Independent test showed that the following seven Gaussian restraints: $D(a, a) \sim N(6.3, 0.27)$; $D(a, d) \sim N(6.5, 0.18)$; $D(d, d) \sim N(6.6, 0.08)$; $D(a, e) \sim N(7.4, 0.13)$; $D(a, g) \sim N(8.7, 0.31)$; $D(d, e) \sim N(8.3, 0.19)$; $D(d, g) \sim N(7.9, 0.27)$ are sufficient to reproduce a coiled-coil regions (<1 Å in R.M.S.D.). Apart from these restraints, α -helix restraints were also added to predicted residues involved in coiled-coil interaction. A model was built by using these spatial restraints, homology restraints and using the initial model from previous section as starting configuration.

(4) SMC1-SMC3 complex modelling

SMC1 and SMC3 are known to interact with each other on their hinge regions (SMC1 residue 1-170, 1071-1225; SMC3 residue 1-161, 1051-1230). To model this interaction, distance restraints were learnt from the known structure of homologous SMC hinge region from mouse (PDB:2WD5) as a template. Only $C\alpha$ and $C\beta$ atom that are within 8 Å of each other were considered.

The starting configuration of SMC1-SMC3 complex was prepared by structurally aligned SMC1 and SMC3 models to their corresponding chain in the template structure (SMC1-2WD5:A; SMC3-2WD5:B). SMC1 and SMC3 were then pulled apart for 4 Å along x-axis to reduce the chance of atomic clashing in the starting configuration. Gaussian distance restraints were applied on SMC1-SMC3 interface, with the mean values equal to the learnt distances in the template structure and standard deviations set as 0.5 Å. SMC1 and SMC3 were treated as rigid bodies

Figure 7.5: (A) Electrostatic potential of SMC1-SMC3 complex model. The electrostatic potential surface was calculated using the program Adaptive Poisson-Boltzmann Solver (APBS) [282]. Accessible lysine residues with unneutralized charge on SMC1 and SMC3 are represented as black sticks. The structures are rendered using PyMol software suite [90]. (B): Coulombic surface coloring of SMC1 (top) and SMC3 (bottom) models on their accessible surface. Negatively charged surfaces are colored red and positively charged surfaces are colored blue. The structure is rendered using UCSF Chimera software suite [235].

in the modelling process. To satisfy the spatial restraints mentioned, 3 cycles of standard conjugate gradient and molecular dynamics minimization (at 700K, 500K and 300K) were performed to generate the final complex model.

(5) Validation of model

27 low-resolution electron micrographs of the Cohesin complex are available from the literature [283]. From the micrographs it can be estimated that the average length of the SMC arms is ~ 430 Å, and the maximum observed arm length is ~ 480 Å (Figure 7.6). This is in agreement with the comparative models built, with coiled-coil region about 480-500 Å long. Further, from histogram analysis of the coiled-coil segment length (Figure 7.6), it was suggested that there could be two

Figure 7.6: (left) Annotation of SMC micrograph (performed manually). Two coiled-coils were annotated for each micrograph, colored red and blue respectively (SMC1 and SMC3 cannot be distinguished from micrograph). (note: micrograph at (4,6) was not annotated because image is too fuzzy). (right) Distribution of coiled-coil segment lengths.

Figure 7.7: Estimated length for coiled-coil region of SMC1 (top) and SMC3 (bottom). The structures are rendered using PyMol software suite [90].

kink points along arms, one located at around $1/3 - 1/4$ of the arm and another close to the center of the arm. The first observed kink point coincides with a ~ 65 residues loop insertion in the model. This loop insertion breaks the continuity of the rigid coiled-coil structure or the protein, and could manifest as a kink point. Also the micrograph image is in support of a all- or majority-helical models because of the small number of kinks.

7.1.3 Prediction of critical lysines

(1) Accessibility of Lysines

In a coiled-coil configuration, residues in registers ‘a’ and ‘d’ are buried and residues in ‘g’ and ‘e’ register positions are often involved in salt-bridge formation. Only residues in ‘b’, ‘c’, and ‘f’ are exposed and free to interact. This hypothesis was affirmed by studying homologs of cohesin ring. 24 lysine modification sites (by acetylation and ubiquitination) are compiled from literatures [284, 285] (Table 7.1). Of these sites, 13 were found to be in loop regions, while 11 are in coiled-coil configuration. Further, 9 out of 11 coiled-coil sites are located in one of exposed positions of ‘b’, ‘c’ and ‘f’, while 2 sites are in ‘g’ positions. No residues in buried positions of ‘a’, ‘d’ (and ‘e’) positions were observed to participate in amino-acid modifications. From this consideration, only the lysines at accessible positions are listed as suggestions for mutagenesis studies.

(2) Electrostatic profiling

Local electrostatics was proposed as a second criteria to prioritize among different lysines. This is assuming functionally important interactions between Cohesin complex and histone are mainly electrostatic in nature. The lysine residues which their positive charges are neutralized by negatively charged residues (glutamates and aspartates) in the surrounding environment are considered less important. To roughly estimate the effective electrostatic potential of a Lysine residue j , the sum of electrostatic interaction acting on its C^β atom was computed with $E_j = \sum_i c_i/d_{ij}$, where d_{ij} is the distance between charged atom i and the C^β atom. Lysines with unneutralized positive charge are ranked more highly for mutation suggestions.

7.1.4 Summary and conclusions

In summary, in this section we have described the modelling of cell division regulator protein complex Cohesin, in the aim of identifying critical residues for specific interactions. The protein complex consists of two protein components of SMC1 and SMC3. Both SMCs contains a large region made up of coiled-coil interaction. First,

index	protein	modification	residue	P-value	Coiled-coil	register
1	SMC1A	acetylation	282	0.0086	TRUE	f
2	SMC1A	acetylation	437	0.0035	TRUE	c
3	SMC1A	acetylation	536	0.2999	FALSE	N.A.
4	SMC1A	acetylation	648	0.2995	FALSE	N.A.
5	SMC1A	acetylation	713	0.0243	TRUE	b
6	SMC1A	ubiquitinylation	177	0.0627	FALSE	N.A.
7	SMC1A	ubiquitinylation	660	0.099	FALSE	N.A.
8	SMC1A	ubiquitinylation	736	0.0518	FALSE	N.A.
9	SMC1A	ubiquitinylation	1037	0.0118	TRUE	g
10	SMC3	acetylation	114	0.3363	FALSE	N.A.
11	SMC3	acetylation	140	0.5242	FALSE	N.A.
12	SMC3	acetylation	215	0.02	TRUE	f
13	SMC3	acetylation	336	0.0038	TRUE	g
14	SMC3	acetylation	1190	0.4257	FALSE	N.A.
15	SMC3	ubiquitinylation	188	0.0111	TRUE	c
16	SMC3	ubiquitinylation	194	0.0111	TRUE	b
17	SMC3	ubiquitinylation	215	0.0199	TRUE	f
18	SMC3	ubiquitinylation	956	0.1481	FALSE	N.A.
19	SMC3	ubiquitinylation	963	0.0578	FALSE	N.A.
20	SMC3	ubiquitinylation	985	0.0345	TRUE	c
21	SMC3	ubiquitinylation	1012	0.0345	TRUE	b
22	SMC3	ubiquitinylation	1025	0.0545	FALSE	N.A.
23	SMC3	ubiquitinylation	1034	0.2149	FALSE	N.A.
24	SMC3	ubiquitinylation	1038	0.2149	FALSE	N.A.

Table 7.1: Listing of lysine modification sites in human SMC1A and SMC3 collected from literatures. The configuration (whether in coiled-coil) and register entry was assigned using Paircoil2.

res no.	Pot.	res no.	Pot.	res no.	Pot.	res no.	Pot.	res no.	Pot.
13	0.23	255	-0.41	432	0.04	735	-0.02	932	-0.31
20	0.27	267	-0.40	445	-0.40	746	-0.41	942	0.10
39	-0.25	269	-0.29	460	-0.26	754	-0.38	945	0.10
63	-0.33	275	0.29	472	-0.11	759	-0.03	983	-0.10
95	0.30	281	0.52	473	-0.22	764	-0.07	987	0.27
100	-0.20	286	0.16	491	-0.08	765	-0.21	988	0.21
103	-0.74	293	0.30	498	-0.17	766	-0.41	990	0.01
120	-0.05	295	0.20	511	-0.06	774	-0.32	999	-0.13
124	-0.28	297	0.27	514	0.06	779	-0.19	1004	-0.34
129	-0.28	303	-0.13	523	0.13	781	-0.33	1043	-0.35
143	-0.07	305	-0.45	530	-0.24	790	-0.17	1047	0.10
145	-0.10	307	-0.13	540	0.04	795	-0.35	1048	0.20
179	0.09	316	-0.05	541	0.14	801	-0.50	1055	0.63
180	0.32	324	-0.01	543	0.07	816	-0.03	1057	0.52
186	-0.33	331	0.19	554	-0.16	823	-0.01	1058	0.45
188	-0.12	338	0.26	575	0.23	831	-0.36	1059	0.54
191	-0.16	344	-0.04	576	0.23	848	0.39	1061	0.14
194	-0.24	355	0.00	620	-0.59	851	0.04	1066	-0.35
201	0.29	361	-0.29	639	0.20	861	-0.65	1084	-0.03
211	0.15	369	-0.50	642	0.56	873	-0.26	1113	-0.60
214	-0.08	377	-0.17	644	0.26	880	-0.15	1121	0.14
219	-0.17	379	-0.31	645	0.44	885	-0.10	1124	0.04
225	-0.23	386	-0.29	650	0.47	893	0.04	1134	-0.62
229	-0.05	394	-0.43	661	0.53	894	0.11	1191	-0.56
230	-0.06	406	-0.31	679	-0.60	898	0.24	1197	0.16
235	0.06	414	-0.32	689	0.09	920	0.08	1214	-0.25
249	-0.42	426	0.11	691	-0.23	927	-0.50		

Table 7.2: Estimated electrostatic potential (Pot.) on $C\beta$ atom of Lysine residues of SMC1

res no.	Pot.	res no.	Pot.	res no.	Pot.	res no.	Pot.	res no.	Pot.
4	-0.40	291	-0.05	530	-0.1	767	0.4	1032	0.29
9	-0.02	300	-0.32	545	0.24	770	0.31	1034	0.41
12	0.34	309	-0.51	549	-0.1	781	-0.39	1037	0.31
38	-0.03	316	-0.14	551	-0.03	799	-0.39	1047	-0.37
57	0.00	318	-0.14	584	0.05	802	-0.64	1066	-0.36
112	-0.04	339	-0.03	603	-0.04	810	-0.25	1072	-0.16
113	-0.06	342	-0.12	619	0.37	818	0.15	1098	-0.54
125	-0.50	349	0.06	620	0.52	843	-0.44	1116	0.09
150	-0.43	351	0.34	622	0.27	854	-0.3	1131	-0.22
158	-0.34	354	0.13	629	-0.14	871	-0.4	1162	-0.12
160	-0.39	364	-0.31	632	-0.02	877	-0.47	1173	-0.28
173	-0.45	370	-0.16	637	0.21	882	-0.5	1177	-0.37
178	0.42	379	0.18	642	-0.28	909	-0.06	1196	-0.18
180	0.24	386	0.14	649	0.32	910	-0.03	1201	0.21
184	0.02	387	0.22	652	0.58	914	-0.03	1205	-0.1
185	0.38	394	-0.22	653	0.57	924	0.23		
192	0.00	396	-0.04	655	0.55	925	0.3		
193	-0.29	411	-0.63	668	0.57	931	-0.14		
198	-0.15	433	0.09	681	0.06	935	0.1		
206	-0.39	455	-0.49	689	0.07	939	0.22		
215	-0.41	469	-0.05	699	-0.18	940	0.36		
219	-0.39	471	0.18	700	-0.21	992	0.26		
228	0.05	480	-0.34	718	-0.41	996	0.28		
269	-0.55	485	-0.17	728	0.17	1003	0.35		
279	-0.15	488	0.03	750	-0.24	1004	0.14		
280	0.08	523	-0.32	751	-0.08	1010	0.09		
289	0.27	528	-0.17	754	-0.87	1023.00	-0.63		

Table 7.3: Estimated electrostatic potential (Pot.) on $C\beta$ atom of Lysine residues of SMC3

we predict the registers of the coiled-coil region using the paircoil algorithm. Next, with few hints from cross-linking experiments and a manual assignment of coiled-coil interaction, the 3D model of Cohesin is built. The model shows good agreement with low-resolution micrograph data. From the 3D model, the accessibility of lysine residues and their electrostatic profiling are computed. The mutation suggestions is limited to only lysine residues with accessible registers or in loop region. Finally, to accommodate with the experiment capacity, the feasibility of the mutations are rank-ordered by the strength of local electrical field of the residue.

7.2 Modelling of OCTN Transporters

7.2.1 Background and motivations

Genes in SLC22A family are known to mediate the transport of organic cations, zwitterions, and organic anions across cell membranes [286]. Two proteins in the family, human organic cation/carnitine transporter 1 and 2 (hOCTN1, hOCTN2), encoded respectively by SLC22A4 and SLC22A5 are being studied. Both the proteins are multispecific, bidirectional, and *pH*-dependent transporters for organic cations.

The ligands and expressions of hOCTN1 and hOCTN2 are different. hOCTN1, is known to regulate the transport of the antioxidant L-ergothioneine and involved in antioxidant defense system of the skin and other organs [287–290]. hOCTN1 also influences the cellular uptake of drug molecules, including the anticancer agents doxorubicin and mitoxantrone [256]. hOCTN1 is expressed in the kidney, small intestine, ocular epithelium, erythrocytes, as well as in mitochondria from cultured skin fibroblasts. hOCTN2 on the other hand, is known to transport carnitine, an essential amino acid for the β -oxidation of fatty acid metabolism. Impaired hOCTN2 function can cause carnitine deficiency [288, 291], fasting-induced hypoketotic hypoglycaemia, hepatic encephalopathy and cardiac myopathy [292, 293]. In addition, hOCTN2 also transport a number of xenobiotics, such as tetraethylammonium (TEA), verapamil, pyrilamine and the β -lactam antibiotics [294]. hOCTN2 is expressed in numerous tissues, including kidney, heart, and liver [295].

Both OCTN1 and OCTN2 are of pharmacological importance due to their functions of regulating cellular uptake of drug and association with the pathogenesis of several diseases, including inflammatory bowel disease, primary carnitine deficiency, diabetes, neurological disorders, and cancer [296]. Our collaborators have previously surveyed the population genetics of Singapore and identified several non-synonymous variants in the Chinese and Indian populations [297] (hOCTN1: R63H, R83P, L134M, M258I, M344K, G482D, M487T, I500N; hOCTN2: E109Q, D122Y, V175M, K191N, A214V, and K302E). Some of the variants are impaired in their functions. To investigate the mechanism of the transporter and provide explanation to dysfunctional variants, I was sought by the collaborator (Dr. Dorothy Su) to provide insights from bioinformatics study and structure modelling approach.

7.2.2 Modelling protocol

OCTNs are known to consists of twelve transmembrane domains (TMDs) with a large extracellular loop between TMDs 1 and 2 and a large intracellular loop located between TMDs 6 and 7 [293, 298]. hOCTN1 and hOCTN2 consists of 557 and 551 amino acids respectively.

The structure of membrane proteins is difficult to solve crystallographically, and in the current Protein Data Bank no close homologs (sequence identity $< 30\%$) can be found for either hOCTN-1 and -2. Therefore in this study, only coarse-grained modelling was attempted. A BLAST search [16] of the non-redundant protein database confirmed that OCTN2 belongs to the Major Facilitator Superfamily. Of the four non-redundant structures (PDB: 1pw4, 2cfp, 2gfp, 3o7q) of the superfamily deposited in the Orientations of Proteins in Membranes (OPM) database [299], the *Escherichia coli* Glycerol-3-Phosphate Transporter (GlpT) (PDB:1PW4) was used as the template for comparative modelling in consideration of substrate similarity. This template structure has sequence identity of $\sim 15\%$ and similarity of $\sim 20\% - 30\%$ to the OCTNs. Different regions (transmembrane, intracellular and extracellular) of the OCTNs and GlpT were predicted using TransMembrane prediction TMHMM [300, 301]. Corresponding regions of the OCTNs and GlpT

were subsequently aligned separately using CLUSTALW [302]. These regions were then reassembled to produce the final alignment. In 3D model building, additional α -helix constraints were applied to predicted transmembrane regions. Comparative modelling was performed using MODELER v9.9. The resulting model (excluding extracellular region) has a Z-score of -1.7 from Protein Structural Analysis [303], which is barely within the limit of a high quality model.

The extracellular domains of the transporters (hOCTN1: 43-141; hOCTN2: 44-142) were modeled separately. The domain of hOCTN2 was modeled using three templates (PDB: 2vq3, 2h0a and 1ndh), while for the domain of hOCTN1 no suitable can be found and *ab initio* modelling was attempted using the Rosetta method [304].

7.2.3 Model interpretation and mechanism hypothesis

By inspecting the model, we deduced that for OCTN1, the mutations M344 and I500 are likely located within α -helices that comprise TMD 7 and 12 respectively, while M258 (TM domain 6), G482 and M487 (extracellular loop) are likely located at the cell membrane boundary. R63, R83, and L134 are in the large extracellular loop; For OCTN2, V175M is in TMD3, K191N in a smaller extracellular loop and A214V in the intracellular loop located between TMD4 and TMD5. K302E is located at the interface between TM6 and intracellular region. The mutations E109K, D122Y are in the large extracellular loop.

From experiments it is known that some of the mutants of OCTN1 and OCTN2 are dysfunctional. Specifically, for OCTN1, the transport function of the R83P variant was only 10% of wild-type, while those of the R63H, G482D, and I500N variants were significantly decreased to 50% of control ($p < 10^{-4}$); for OCTN2, D122Y and K302E have been shown significantly impaired of their substrates (L-Carnitine and tetraethylammonium (TEA) as representatives) - D122Y has only less than $\sim 10\%$ activity as compared to wild-type, and K302E less than $\sim 50\%$. All the other variants have similar activity as wild-type (OCTN1: L134M, M258I, M344K, and M487T; OCTN2: E109Q, D122Y, V175M, K191N, A214V, and K302E).

Figure 7.8: (Left) Homology modelling of OCTN1. Top panel shows its sequence alignment with GlpT (*Escherichia coli* glycerol 3-phosphate transporter, PDB:1PW4) and the bottom panel shows the final 3D model. Insets show the extracellular regions (residue 43-141). Image taken from Toh et al, 2011 [305]. (Right) Homology modelling of OCTN2. Top panel shows its sequence alignment with the same transporter template, and the bottom panel shows the final 3D model. Insets show the extracellular regions (residue 44-142). Image taken from Toh et al, 2013 [306]

Figure 7.9: (A). Transport of 5 M L-ergothioneine in HEK-293 cells expressing wild type and variants of OCTN1. (B) Uptake of L-carnitine by HEK293 cells expressing wild type and variants of hOCTN2. Different from wild-type ** $p < 0.01$; *** $p < 0.001$. **** $p < 0.0001$.

We observed that many of the disruptive mutations occur at, or close to extracellular regions, but without apparent role in ligand interaction. These mutations include G482, M487, R63, R83, L134 for hOCTN1 and E109K, D122Y, K302E in hOCTN2. Interestingly, most of these mutations involve either the removal of positive charge extracellularly or introduction of negatively charged residue intracellularly. We are reminded of the ‘positive-inside rule’ [307] of membrane protein topology. The rule states that positively charged amino acids are manifold more abundant in cytoplasmic, as compared to periplasmic or segments of integral membrane proteins. These positively charged residues could determine the topology of membrane proteins. We suggested to our collaborators that these dysfunctional mutations could impair the protein functions by disrupting their interaction with the cell membrane.

7.2.4 Validation of mechanism hypothesis

Our hypothesis was subsequently tested experimentally. Biotinylation and immunoblot analyses were used to quantify plasma membrane expression of hOCTN2 and its variants. This technique preserves the integrity of cell structure and could

distinguish between membrane and cytosolic proteins. The total expression and surface expression of hOCTN2 were separately measured. The results of the measurement showed that on cell membrane the expression of the D122Y and K302E variants was decreased to $\sim 10\%$ and $\sim 40\%$ of wild type. This is in very good agreement with the degree of their activity decrease. On the other hand, their total protein expressions were not significantly different from the wild-type. This result strongly suggested that decrease in protein activity is mainly due to impaired trafficking of the transporter to cell membrane. In fact, subsequent investigation with immuno-colocalization technique revealed that the transporter proteins were retained in Golgi apparatus and the endoplasmic reticulum. Our result was also reinforced with further chemical kinetic analysis experiments. It was confirmed from the analyses that the binding affinity of K302E variant to L-carnitine remains intact (decreased V_{max} but similar K_m value as compared to wild-type) and hence does not contribute to the decrease in activity.

The same set of experiments were repeated for dysfunctional OCTN1 variants of R63H, R83P, G482D and I500N. Except for G482D, the transporter cell surface expressions of the mutants are affected. The functional impairment of G482D (and I500N) was likely due to decrease binding affinity with substrate, as validated by kinetic experiments ($\sim 50\%$ decrease in K_m).

7.2.5 Summary and conclusions

In this work, we have collaborated with experimentalists to explore the mechanism of nonsynonymous mutations on OCTN1 and OCTN2 occurring naturally in Singapore population. OCTNs are membrane proteins, which are difficult to model accurately due to the lack of structural templates in the Protein Data Bank. In this study, we demonstrate however that even with coarse-grained modelling, by supplementing with other bioinformatics techniques, abundance of information about the protein functions could be obtained.

To model the 3D structure of OCTNs, we first used the structure of GlpT of the same superfamily as structural template. Secondly, we relied on a hidden markov

Figure 7.10: Comparison of surface and total expression level of (A) OCTN1 and (B) OCTN2 of different variants. For OCTN1, 70kDa bands correspond to native protein, while 130kDa bands correspond to glycosylated form of the protein.

model (TMHMM) to assign the topology of the OCTNs as well as the template. The corresponding regions between OCTNs and template were separately aligned and reassembled as final alignment for homology modelling. The locations of mutations were then inferred from the models. By inspecting the location and the charge state of mutations, we discovered that most disruptive mutations are in violation of ‘positive-inside rule’ of membrane protein topology. This prompts the hypothesis that the decrease in activity is associated in membrane insertion of the transporters.

This hypothesis is tested by comparing the total cellular expression levels of transporters and the portion found on cell surface. In all but one (hOCTN1:G482D) of the mutations studied here, the hypothesis is validated when the altered function of the protein is consistent with decreased expression on cell surface. The total cellular expression of the most protein variants, as well as their binding affinities with ligands, remain approximately unchanged by the mutations.

Our study has unveiled a common mechanism of impaired function in hOCTNs, and probably other membrane-associated proteins. The result of our study has impact on the pharmacokinetic profiles of drugs, particularly on the absorption and elimination of drugs that are substrates of OCTNs transporter. Our findings

here could help in the development of personalised medicine that involves hOCTN transporters, especially in Asian population. As a simple example, patients with impaired hOCTNs allele should be prescribed with alternative medicines or have an adjusted level of drug administration.

7.3 Modelling of EZH2-VAV1 Interaction

7.3.1 Background and motivations

EZH2 (Enhancer of zeste homolog 2) is an oncogene, whose overexpression is known to associate with several cancer types, including prostate cancer, breast cancer, and lymphomas [308–310]. The protein has previously been shown to represses tumor suppressor genes through trimethylation of histone 3 (at lysine 27) [311–318]. However, the work of our collaborator (Dr. Su I-Hsin) has shown that in addition to its action in cell nucleus, EZH2 was also found abundantly in cytoplasm and associates with another group of proto-oncogenes, the VAV family of proteins [319–321]. VAV proteins are known to regulate the adhesion, migration and proliferation of cells. Overexpression or mutation of VAV could cause pathogenesis of cancer cells, particularly the metastasis of tumors [322–325]. To decipher the tumorigenesis pathway involving the two proteins, detailed understanding of their interactions is indispensable. The interacting regions of VAV1 and EZH2 have previously been narrowed down to approximately 50 residues on each proteins (VAV1 residues 115–166; EZH2 residues 201–252) through yeast two-hybrid screening [321]. Surface plasmon resonance (SPR) experiment also determined the stoichiometry of the interaction to be 1:1. With this information, we attempted to model the protein-protein interaction, as described in the following sections.

7.3.2 Modelling protocol

(1) Modelling the 3D structure of VAV1

The modelling of VAV1 3D structure was straightforward, as the known structure of a close homolog, VAV-3 (sequence identity 71%) has been solved and deposited

in PDB (PDB:2D86). The 3D model was built using the homology modelling suite Modeller v9.9 [29] with the sequence alignment between VAV1 and VAV-3 computed using SALIGN [10]. The model was assessed using ProSA [303, 326] and resulted in a Z-score of -5.7, which is indicative of a high accuracy model.

(2) Modelling 3D structure of EZH2

Unfortunately for EZH2 no similar sequences were found in protein data bank. Furthermore, no conserved domain or short sequences with biological inference could be identified, and domain predictions from multiple servers [327–329], HMMER [330, 331] do not agree with one another. Hence, both comparative and *ab initio* techniques cannot be employed for the modelling. However, multiple secondary structure prediction algorithms [331–334] gave the consensus result of interface residues with high confidence. The predicted secondary structure consists of two helices with charged residues in-phase along the helices (Figure 7.11). These charged residues are presumably exposed to solvent and involved in interaction with VAV1.

(3) Modelling of VAV1/EZH2 interface

As the 3D model of EZH2 cannot be obtained, conventional protein-protein interaction prediction methods are not applicable. Nonetheless, the mode of interaction could still be inferred from the complementary interface of VAV1. The interacting site of VAV1 (residues 66-115) consists of three helices and one loop, in which a short helix (7-residue) and the loop are sandwiched by two other longer helices (~20 residues each) (Figure 7.12).

A BLAST [16] search on the non-redundant protein database revealed that this interacting site belongs to the Calponin Homology (CH) Domain [335]. Text search of protein complexes with Calponin Domain in PDB returned 26 candidates. Additionally, a search of similar substructure with VAV1 binding site (PDB:2d86:A, residues 72-138) using CLICK returned an additional 106 candidates. A manual inspection of the candidates was performed to exclude false-positives and non-interacting structures. 13 structures (from 9 PDB entries of 1dxx, 1pxy, 1qag, 1sh5, 1wku, 2r0o, 2vzc, 3hoc, 3hor) were selected. This set consists of 8 heterodimers and

Figure 7.11: (Top) Secondary structure prediction (H: helix, E: strand, C: coil) and residue charge on EZH2 interface (Bottom) Predicted helices from EZH2 interface. Negatively charged residues are colored red and positively charged residues are colored blue. It is clear that charged residues are in phase (left) sequence: PADKIFEAISSM (right) sequence: AEELKEKYKELTEQQ.

Figure 7.12: 3D model of VAV1, facing out is the interaction interface with EZH2. The structure is highly electrostatically polarized into a positively charged and a negatively charged region. Electrostatics profile is computed using APBS [282]

5 homodimers. The sequence similarities among the known interacting partners are between 13% $\sim$ 82% and their structural similarities are between 0.77Å - 2.01Å. From these structures, the following relevant rules on calponin domain interaction are found (Figure 7.13):

- (a) All calponin domains are highly electrolytically polarized with approximately net zero charge.
- (b) In all cases, the middle helices are conserved with 2-4 negatively charged residues.
- (c) In most of the cases (12 out of 13), the interaction acts primarily through the α -helix.
- (d) The interaction is primarily electrostatic in nature.
- (e) In most cases the interacting helix has an alternate pattern of charged residues (negative-(neutral)-positive).
- (f) All interaction partners consist of at least an α -helix, and in most of the cases (11 out of 13) the helix aligns parallel with CH domain helices.
- (g) In non-dimerization cases, interaction is coordinated by middle loop and right-sided helix (viewed from solvent) with middle helix placed on top.

Since the orientation of all known Calponin domain interactions is similar, their consensus alignment could be used as a structure template for VAV1-EZH2 interface modelling. To proceed, SALIGN was used to align sequences of the 8 helices with the two predicted helices of EZH2. The 2nd helix was selected as it achieved a better charge complementarity of the interface, after the alignment was fine-tuned (by shifting to C-terminal by one position) (Figure 7.14). A 3D model of interface was built based on this alignment using Modeller (Figure 7.15). The model was assessed and indicated as good quality by using the statistical potential MODTIE [336].

Our model suggested that the complex is mediated by two sets of electrostatic interactions (a) (K234, K241 and K245) on EZH2 and (E92, E95, D98, D101 and

Figure 7.13: (left) Illustrative figure on the properties of Calponin Homology domain interaction by superposition known structures. Positively and negatively charged residues are colored blue and red respectively. (right) Superposition of Calponin Homology (CH) domains. Calponin Homology domain is colored wheat and interacting partners are colored red (helix) and green (loop).

Figure 7.14: Sequence alignment between structural templates and predicted EZH2 interacting helix using SALIGN structure-sequence alignment. The histogram represents the conservation of the residue. The alignment is rendered with ClustalX [337]

Figure 7.15: Comparative model of the VAV1 - EZH2 interaction. Vav1 is shown in surface representation while the EZH2 interacting helix is shown as ribbon. The surface is colored by its electrostatic potential (positive and negative regions are shaded blue and red respectively) and some of the residues contributing to the charged patches are labeled. Key interacting residues from the EZH2 helix are shown in stick representation and labelled. The figure is rendered using UCSF Chimera software suite [235].

D104) on VAV1; (b) (E249) on EZH2 and (K89 and R90) on VAV1. To validate the model, alanine mutations were suggested that would abolish one or more of these charged patches resulting in a decrease in binding affinity (Table 7.4).

(4) Validation of the model

Our mutation suggestions were validated experimentally by our collaborators. Mutant proteins of EZH2 were synthesized and their affinities with VAV1 were mea-

Mutation	Prediction	Reciprocal Mutation
D104A+D101A+E98A	Greatly reduced affinity	K234A
D104A+D101A	Reduced affinity	K234A
D104A+D98A	Reduced affinity	K234A
E95A+E92A	Reduced affinity	K241A+K245A
D104A+E98A+E92A	Completely abolished interaction	K234A+K241A+K245A
K89A+R90A	Reduced affinity	E249A

Table 7.4: Predicted binding affinity result with alanine scanning mutagenesis on critical interacting residues on VAV1 and their reciprocal mutations on EZH2.

sured using GST (Glutathione S-transferase tag) and His-tag (poly-histidine tag) pull-down assays (Figure 7.16). The experiments also tested several other neighbouring residues as control references. In some cases, mutations from lysine (K) to arginine (R) were also made.

From the experiment result, it can be seen that:

- (a) Neighbouring negatively charged residues of E238, E239, E246, E249 when mutated to Alanine, did not weaken the interaction strength with VAV1.
- (b) Lysine residues K234, K241 and K245, when mutated to Alanine, show decreased interaction (>40-50%) with VAV1; when mutated to Arginine, showed only minor decrease (~25%), or even higher affinity (~150%) with VAV1.

Collectively the experimental results verified our postulation that the interactions are predominantly electrostatic in nature, and our predictions on the critical interacting residues are correct.

The double mutant K234/K245A was utilized further in the experiment by our collaborators to disrupt the interaction between VAV1 and EZH2. Comparing the experiment results with native EZH2, we are able to show that the interaction between EZH2-VAV1 is critical for cellular transformation and metastasis. In particular, the role of EZH2 to promote the disassembly of adhesion structures and reducing cell spreading is demonstrated.

7.3.3 Design of inhibitory peptides

With the mechanical understanding of the protein-protein interaction, It is possible for us to design inhibitory peptides that compete with EZH2 to interact with VAV1. A strongly competitive peptide would deplete cytosolic VAV1 and hence terminate VAV1-dependent downstream signalling. In principle such a peptide could be used as anticancer drug that prevents metastasis of tumors. The design of inhibitory peptide should take into consideration two requirements, namely (a) to maintain the helical conformation of the peptide and (b) to enhance its interaction with VAV1.

Figure 7.16: Relative interaction strength between VAV1 and EZH2 mutants as measured by (A) *in vitro* GST pull-down assay (B): *in vitro* His pull-down assay (C) *in vivo* pull-down assay with full-length proteins

(1) Stapled peptide design

To maintain helical conformation of the design peptide, we suggest the technique of stapled peptides [338]. In this protein engineering technique, non-interacting residues in the peptide are identified, and chemically cross-linked (stapled) with hydrocarbons. With the hydrocarbon staple, the peptide is locked in a helical conformation. Our model suggested that T236, L240, Y244 and T248 (Thr) are suitable to be staple targets. These residues are located at the opposite side of the interface, and assumed to contribute minimally to the interaction with VAV1.

(2) Identifying mutations that enhances interactions

To design peptide with higher affinity to VAV1, we re-examined the pull-down experiments for insights. Firstly, the critical lysine residues (K234, K241 and K245) would be preserved. Secondly, C-terminal residues of the peptide (Q250 - P253) are not in close contact with VAV1. These residues should be removed to promote cell membrane permeability and decrease non-specific interactions. Lastly, it was shown in the experiments that mutations at K243, E246 and E249 could potentially enhance the interaction. To find the optimal mutations, we examined the immediate environment of K243 and E246. The two residues form a 4-body residue clique with two other residues, Y134 on VAV1 and E242 from EZH2 (Figure 7.17A).

We queried a high resolution protein dataset (see Chapter 7) for cliques of similar size (i.e. 6 – 7 Å) with at least one glutamate and one tyrosine. 3445 of such cliques with 171 different amino acid compositions were found. For each type of clique with different amino acid composition, its observed frequency was normalized against the background frequencies of the amino acids in proteins. The clique compositions with highest normalized frequencies represent candidates for substitutions.

K243 is spatially close to Y134, which is a side chain hydrogen bond acceptor. Therefore it is preferable to introduce an amino acid with side chain hydrogen bond donor in the position. Among top 15 clique candidates, only a histidine (H) and a glutamine (Q) satisfy this criterion. A histidine substitution is preferred as it is found that the hydrogen bonding between H:ND1 and Y:OH is particularly strong from our hydrogen bond study.

Figure 7.17: (left) interatomic distances among interacting residues on VAV1 and EZH2 interface. (right) Interface between VAV1 and inhibiting peptide (design 2). 3 mutations K243H, E246W and E249Y are shown. W246 and Y249 are displayed in sphere representation showing their fitting into respective clefts. Figure is rendered using PyMol [90].

At the bottom of the peptide, it was also noticed that a hydrophobic cleft on VAV1 was formed by residues Y134, S137, A138 and W141. As this cleft is spatially close to position 246, the side chain should be designed to slot into the cleft. In consideration of size, hydrophobicity and π - π stacking interaction, a tryptophan (W) residues was suggested.

K249A was shown to have increased binding affinity. This suggested that negative charge density is not compatible with the micro-environment in vicinity. In consideration of its surrounding of charged (R113, E116), hydrophobic (L112, W141) and aromatic (W141) residues, amino acid tyrosine (Y) which has a aromatic group and a polar hydroxyl group was chosen.

In conclusion, 5 staple peptide designs are suggested to inhibit interaction between VAV1 and EZH2 (Table 7.5). While the first design is just a collection of all known mutations that up-regulate binding affinities, the other 4 designs were arrived by carefully mimicking and optimizing the (proposed) native interface (Figure 7.17B).

7.3.4 Summary and conclusions

In this collaboration work, we have modelled the 3D structure of protein-protein interacting between two oncogenic proteins, VAV1 and EZH2. Conventional methods

Res no.	native	<i>in vitro</i>	<i>in vivo</i>	Ip-1	Ip-2	Ip-3	Ip-4	Ip-5
230	M			M	M	M	M	M
231	F			F	F	F	F	F
232	P			P	P	P	P	P
233	D			D	D	D	D	D
234	K		D	K	K	K	K	K
235	G			G	G	G	G	G
236	T			T	T	T	T	T
237	A			A	A	A	A	A
238	E	X		E	E	E	E	E
239	E	X		E	E	E	E	E
240	L			L	L	L	L	L
241	K	D	X	R	K	K	R	R
242	E			E	E	E	E	E
243	K	U	X	A	H	Q	H	Q
244	Y			Y	Y	Y	Y	Y
245	K	D	D	K	K	K	K	K
246	E	U		A	W	W	W	W
247	L			L	L	L	L	L
248	T			T	T	T	T	T
249	E	U		A	Y	Y	Y	Y

Table 7.5: Experiment data from *in vivo* and *in vitro* and (down) inhibiting peptide designs (Ip-1 - Ip-5). In the left table, labels ‘U’, ‘D’, ‘X’ indicates ‘up-regulated’, ‘down-regulated’ and ‘no-effect’ on on interatomic interaction upon alanine scanning mutagenesis. In the right table, positions with essential lysines (234, 241 and 245) are shown with green background, positions suggested for cross-linkings (236, 240, 244, 248) are shown with grey background. Mutated positions on inhibiting peptide designs are colored with red.

are not applicable in this modelling due to the novelty of EZH2 sequence. Instead, its interface was reciprocally modeled by considering its charge state, secondary structure prediction and complementarity with VAV1. The interface on VAV1 belongs to Calponin domain. By inspecting all known Calponin domains, rules and structural templates of its interaction were derived, and 3D model of the interface was built accordingly. Based on the model, mutation suggestions on EZH2 were made to reduce or abolish its interaction with VAV1. Alanine and arginine mutations were performed for lysines at positions 234, 241 and 245, with several other positions as controls. The mutation experiments validated our model and predictions. Mutant EZH2 from the experiment was subsequently used to investigate the effect of abolishing its interaction with VAV1. It is shown that tumorigenesis effect of EZH2 is dependent on the interaction by comparing with the native protein, particularly on the metastasis of tumor.

This discovery has suggested to us that disrupting EZH2-VAV1 interaction could be a strategy of treatment for EZH2-related cancer types. Further, the detailed understanding of VAV1-EZH2 interacting interface has allowed us to design anticancer drugs that mimic the interactions of EZH2. In particular, short helix inhibitory peptides are proposed. Further, we suggested hydrocarbon cross-linking technique (stapled peptides) to maintain the helicity of the peptides. Using the EZH2 model we also identified residue positions suitable for cross-linkings. Lastly, by drawing insights from our previous study on residue clique preference and protein hydrogen bonds, we have made proposals to mutate certain residues on the inhibitory peptides, which could enhance its interaction with VAV1 and outcompete native EZH2 proteins.

7.4 Conclusions

In this chapter we have described the modelling of three separate non-globular protein systems, the fibrous Cohesin, protein-protein interface between oncogenic proteins of VAV1 and EZH2, and the transmembrane proteins OCTNs. Conventional homology modelling are not applicable in any of the three cases as no suitable structural templates are available. Instead, we have demonstrated the use

of integrative modelling, where spatial restraints are system specific and derived from different sources. Importantly, theoretical results from the previous chapters are also appropriately applied in these real case studies. With the 3D model of the non-globular proteins, we gained biological insights on different systems, and accordingly proposed experimentally testable hypotheses. Thus far most of our proposals have been validated correct by experiments, while a few experiments in collaboration are still ongoing. We expect that through reciprocity between theoretical and experimental research our understanding and potential manipulation of these biological systems could be further enhanced.

Publication

- (a) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. **Kuan Pern Tan**, Thanh Binh Nguyen, Siddharth Patel, Raghavan Varadarajan, and M. S. Madhusudhan. Nucl. Acids Res. (1 July 2013) 41 (W1): W314-W321.
- (b) Functional Analysis of Novel Variants in the Organic Cation/Ergothioneine Transporter 1 Identified in Singapore Populations. Dorothy Su Lin Toh, Florence Shin Gee Cheung, Michael Murray, **Tan Kuan Pern**, Edmund Jon Deoon Lee, and Fanfan Zhou. Molecular Pharmaceutics 2013 10 (7), 2509-2516
- (c) TSPred: A web server for rational design of temperature-sensitive mutants. **Kuan Pern Tan**, Shruti V. Khare, R. Varadarajan, M.S. Madhusudhan. Nucleic Acids Research 2014
- (d) DEMM: a meta-algorithm to predict the pKa of ionizable amino acids in proteins. Nguyen T.B., Tan K.P., Madhusudhan M.S. IFMBE Proceedings, Vol. 46, 343-346 (2015).
- (e) EZH2VAV Interaction is Crucial for EZH2-Mediated Neoplastic Transformation. Nandini Venkatesan, Jong Fu Wong, **Kuan Pern Tan**, Hwa Hwa Chung, Yin Hoe Yau, Engin Cukuroglu, Merry Gunawan, Chuan Fa Liu, Abdollah Allahverdi, Lars Nordenskiold, Jonathan Gke, Susana Geifman Shochat, Valerie Chun Ling Lin, M.S. Madhusudhan, I-hsin Su (in review for publication)

- (f) Packpred: Predicting the functional strengths of point mutations in proteins. **Kuan Pern Tan**, Parichit Sharma, Shruti Khare, Chee Keong Kwoh, Raghavan Varadarajan, M.S. Madhusudhan (submitted for preparation)
- (g) Protein 3D Model Refinement Using Hydrogen Bond Restraints. **Tan Kuan Pern** International Conference of Biomolecular Forms and Functions (ICBFF) 2013.
- (h) Geometric variation of Intramolecular Hydrogen Bonds in Proteins. **Tan Kuan Pern** MGMS conference "Exploring Mechanisms in Biology: Theory and Experiment" 2015

Bibliography

1. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. & Ofran, Y. Automatic prediction of protein function. eng. *Cell Mol Life Sci* **60**, 2637–2650 (2003).
2. Russ, W. P. & Ranganathan, R. Knowledge-based potential functions in protein design. eng. *Curr Opin Struct Biol* **12**, 447–452 (2002).
3. Nick Pace, C., Scholtz, J. M. & Grimsley, G. R. Forces stabilizing proteins. eng. *FEBS letters* **588**, 2177–2184 (2014).
4. Pace, C. N. *et al.* Contribution of hydrophobic interactions to protein stability. *Journal of Molecular Biology* **408**, 514–528. ISSN: 00222836 (2011).
5. Honig, B. & Yang, A. S. Free energy balance in protein folding. eng. *Advances in protein chemistry* **46**, 27–58 (1995).
6. Chakravarty, S. & Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure (London, England : 1993)* **7**, 723–32. ISSN: 0969-2126 (1999).
7. Arunan, E. *et al.* Defining the hydrogen bond: An account (IUPAC Technical Report). English. *Pure and Applied Chemistry* **83**, 1619–1636. ISSN: 1365-3075 (Jan. 2011).
8. Arunan, E. *et al.* Definition of the hydrogen bond (IUPAC Recommendations 2011). English. *Pure and Applied Chemistry* **83**, 1637–1641. ISSN: 0033-4545 (2011).
9. Anfinsen, C. B. Principles that govern the folding of protein chains. eng. *Science* **181**, 223–230 (1973).
10. Braberg, H. *et al.* SALIGN: a web server for alignment of multiple protein sequences and structures. eng. *Bioinformatics* **28**, 2072–2073 (2012).
11. Nguyen, M. N. & Madhusudhan, M. S. Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Research* **39**. ISSN: 03051048. doi:10.1093/nar/gkr348 (2011).
12. Nguyen, M. N., Tan, K. P. & Madhusudhan, M. S. CLICK–topology-independent comparison of biomolecular 3D structures. *Nucleic acids research* **39**, W24–8. ISSN: 1362-4962 (July 2011).

13. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. eng. *Nucleic Acids Res* **39**, W29–37 (2011).
14. Finn, R. D. *et al.* *Pfam: The protein families database* 2014. doi:10.1093/nar/gkt1223.
15. Finn, R. D. *et al.* HMMER web server: 2015 update. eng. *Nucleic Acids Res* **43**, W30–8 (2015).
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–10. ISSN: 0022-2836 (1990).
17. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function and Genetics* **21**, 167–195. ISSN: 08873585 (1995).
18. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242. ISSN: 13624962 (2000).
19. Adrian, M., Dubochet, J., Lepault, J. & McDowell, a. W. Cryo-electron microscopy of viruses. *Nature* **308**, 32–36. ISSN: 0028-0836 (1984).
20. Schwartz, M. O. D. & M, R. *Chapter 22: A model of evolutionary change in proteins* in *in Atlas of Protein Sequence and Structure* (1978).
21. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).
22. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. eng. *J Mol Biol* **147**, 195–197 (1981).
23. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. eng. *Nucleic Acids Res* **25**, 3389–3402 (1997).
24. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. eng. *EMBO J* **5**, 823–826 (1986).
25. Hubbard, T. J. & Blundell, T. L. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. eng. *Protein engineering* **1**, 159–171 (1987).
26. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (New York, N.Y.)* **253**, 164–170. ISSN: 0036-8075 (1991).
27. Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358**, 86–89. ISSN: 0028-0836 (1992).

28. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology* **268**, 209–225. ISSN: 0022-2836 (1997).
29. Sali, A. *et al.* Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**, 779–815. ISSN: 0022-2836 (1993).
30. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**, 187–217. ISSN: 1096-987X (1983).
31. Summa, C. M. & Levitt, M. Near-native structure refinement using in vacuo energy minimization. eng. *Proc Natl Acad Sci U S A* **104**, 3177–3182 (2007).
32. Ye, J. C. W. & Yuzhen. in *Computational Methods for Protein Structure Prediction and Modeling* (2007). ISBN: 978-0-387-33319-9.
33. Tanaka, S. & Scheraga, H. A. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. eng. *Macromolecules* **9**, 945–950 (1976).
34. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).
35. Sippl, M. J. *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.* 1990. doi:10.1016/S0022-2836(05)80269-4.
36. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. eng. *Protein Sci* **11**, 2714–2726 (2002).
37. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein science : a publication of the Protein Society* **15**, 2507–2524. ISSN: 0961-8368 (2006).
38. Zhou, H. & Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical Journal* **101**, 2043–2052. ISSN: 00063495 (2011).
39. Park, J. & Saitou, K. ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC bioinformatics* **15**, 307. ISSN: 1471-2105 (2014).
40. Tobi, D., Shafran, G., Linial, N. & Elber, R. On the design and analysis of protein folding potentials. eng. *Proteins* **40**, 71–85 (2000).

41. Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. eng. *Proteins* **52**, 51–67 (2003).
42. Kryshtafovych, A., Venclovas, C., Fidelis, K. & Moult, J. Progress over the first decade of CASP experiments. eng. *Proteins* **61 Suppl 7**, 225–236 (2005).
43. Liwo, A. *et al.* A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry* **18**, 849–873 (1997).
44. Betancourt, M. R. & Thirumalai, D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. eng. *Protein Sci* **8**, 361–369 (1999).
45. Vendruscolo, M. & Domany, E. Pairwise contact potentials are unsuitable for protein folding. *The Journal of Chemical Physics* **109**, 11101–11108 (1998).
46. Vendruscolo, M., Najmanovich, R. & Domany, E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? eng. *Proteins* **38**, 134–148 (2000).
47. Carter Jr., C. W., LeFebvre, B. C., Cammer, S. A., Tropsha, A. & Edgell, M. H. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. eng. *J Mol Biol* **311**, 625–638 (2001).
48. Fernández, A. Conformation-dependent environments in folding proteins. *The Journal of Chemical Physics* **114**, 2489–2502 (2001).
49. Khatun, J., Khare, S. D. & Dokholyan, N. V. Can contact potentials reliably predict stability of proteins? eng. *J Mol Biol* **336**, 1223–1238 (2004).
50. Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics* **107**, 3698–3706 (1997).
51. Thomas, P. D. & Dill, K. A. Statistical potentials extracted from protein structures: how accurate are they? eng. *J Mol Biol* **257**, 457–469 (1996).
52. Furuichi, E. & Koehl, P. Influence of protein structure databases on the predictive power of statistical pair potentials. eng. *Proteins* **31**, 139–149 (1998).
53. Summa, C. M., Levitt, M. & Degrado, W. F. An atomic environment potential for use in protein structure prediction. eng. *J Mol Biol* **352**, 986–1001 (2005).

- 54. Munson, P. J. & Singh, R. K. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. eng. *Protein Sci* **6**, 1467–1481 (1997).
- 55. Godzik, A., Kolinski, A. & Skolnick, J. Topology fingerprint approach to the inverse protein folding problem. eng. *J Mol Biol* **227**, 227–238 (1992).
- 56. Godzik, A., Kolinski, A. & Skolnick, J. De novo and inverse folding predictions of protein structure and dynamics. eng. *J Comput Aided Mol Des* **7**, 397–438 (1993).
- 57. Mayewski, S. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. eng. *Proteins* **59**, 152–169 (2005).
- 58. Gan, H. H., Tropsha, A. & Schlick, T. Lattice protein folding with two and four-body statistical potentials. eng. *Proteins* **43**, 161–174 (2001).
- 59. Singh, R. K., Tropsha, A. & Vaisman, I. I. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *Journal of computational biology : a journal of computational molecular cell biology* **3**, 213–221. ISSN: 1066-5277 (1996).
- 60. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. eng. *J Mol Biol* **55**, 379–400 (1971).
- 61. Wodak, S. J. & Janin, J. Analytical approximation to the accessible surface area of proteins. eng. *Proc Natl Acad Sci U S A* **77**, 1736–1740 (1980).
- 62. Fraternali, F. & Cavallo, L. Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome. eng. *Nucleic Acids Res* **30**, 2950–2960 (2002).
- 63. Cavallo, L., Kleinjung, J. & Fraternali, F. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. eng. *Nucleic Acids Res* **31**, 3364–3366 (2003).
- 64. Weiser, J., Shenkin, P. S. & Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *Journal of Computational Chemistry* **20**, 217–230 (1999).
- 65. Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of molecular biology* **79**, 351–371. ISSN: 00222836 (1973).
- 66. Saff, E. B. & Kuijlaars, A. B. J. Distributing many points on a sphere. *The Mathematical Intelligencer* **19**, 5–11.
- 67. Weiner, S. J. *et al.* A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society* **106**, 765–784. ISSN: 0002-7863 (1984).

68. Weiner, S. J., Kollman, P. a., Nguyen, D. T. & Case, D. a. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **7**, 230–252. ISSN: 0192-8651 (1986).
69. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179–5197 (1995).
70. MacKerel Jr., A. D. *et al.* in *The Encyclopedia of Computational Chemistry* 10, Sp. Iss. SI, 271–277 (1998). ISBN: 9780470845011. doi:10.1002/jcc.21287.
71. Van Gunsteren, W. F. *et al.* *Biomolecular Simulation: The GROMOS96 manual and userguide* 1–1042. ISBN: 9783728124227. doi:citeulike-article-id:8062915. arXiv: 978-3-7281-2422-7. <<http://amzn.com/3728124222>> (Hochschuleverlag AG an der ETH Zurich, 1996).
72. Schuler, L. D., Daura, X. & Van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry* **22**, 1205–1218. ISSN: 01928651 (2001).
73. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **110**, 1657–1666 (1988).
74. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236. ISSN: 0002-7863 (1996).
75. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. *In Intermolecular Forces* (ed Pullman, B.) chap. 331-342 (D. Reidel Publishing Company, Dordrecht, 1981).
76. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *Journal of Physical Chemistry* **91**, 6269–6271 (1987).
77. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology* **7**, 95–99. ISSN: 00222836 (1963).
78. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *eng. Biochim Biophys Acta* **405**, 442–451 (1975).
79. Li, L., Li, C., Zhang, Z. & Alexov, E. On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *Journal of Chemical Theory and Computation* **9**, 2126–2136. ISSN: 15499618 (2013).

80. Kauzmann, W. Some factors in the interpretation of protein denaturation. eng. *Adv Protein Chem* **14**, 1–63 (1959).
81. Richards, F. M. Areas, volumes, packing and protein structure. eng. *Annu Rev Biophys Bioeng* **6**, 151–176 (1977).
82. Janin, J. & Chothia, C. The structure of protein-protein recognition sites. eng. *J Biol Chem* **265**, 16027–16030 (1990).
83. Serrano, L., Kellis Jr., J. T., Cann, P., Matouschek, A. & Fersht, A. R. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. eng. *J Mol Biol* **224**, 783–804 (1992).
84. Jackson, S. E., ElMasry, N. & Fersht, A. R. Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. eng. *Biochemistry* **32**, 11270–11278 (1993).
85. Otzen, D. E., Rheinhecker, M. & Fersht, A. R. Structural factors contributing to the hydrophobic effect: the partly exposed hydrophobic minicore in chymotrypsin inhibitor 2. eng. *Biochemistry* **34**, 13051–13058 (1995).
86. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. eng. *Proc Natl Acad Sci U S A* **93**, 13–20 (1996).
87. Bogan, A. A., Cohen, F. E. & Scanlan, T. S. Natural ligands of nuclear receptors have conserved volumes. eng. *Nature structural biology* **5**, 679–681 (1998).
88. Tan, K. P., Varadarajan, R. & Madhusudhan, M. S. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. eng. *Nucleic Acids Research* **39**, W242–8 (2011).
89. Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R. & Madhusudhan, M. S. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. eng. *Nucleic Acids Res* **41**, W314–21 (2013).
90. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8 (Nov. 2015).
91. Pedersen, T. G. *et al.* A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. eng. *J Mol Biol* **218**, 413–426 (1991).
92. Pintar, A., Carugo, O. & Pongor, S. Atom depth in protein structure and function. eng. *Trends in biochemical sciences* **28**, 593–597 (2003).
93. Pintar, A. & Pongor, S. The "first in-last out" hypothesis on protein folding revisited. eng. *Proteins* **60**, 584–590 (2005).

94. Pintar, A., Carugo, O. & Pongor, S. Atom depth as a descriptor of the protein interior. eng. *Biophys J* **84**, 2553–2561 (2003).
95. Adkar, B. V. *et al.* Protein model discrimination using mutational sensitivity derived from deep sequencing. eng. *Structure* **20**, 371–381 (2012).
96. Horowitz, N. H. BIOCHEMICAL GENETICS OF NEUROSPORA. *Advances in Genetics Incorporating Molecular Genetic Medicine* **3**, 33–71 (1950).
97. Fried, M. CELL-TRANSFORMING ABILITY OF A TEMPERATURE-SENSITIVE MUTANT OF POLYOMA VIRUS. *Proceedings of the National Academy of Sciences* **53**, 486–491 (1965).
98. Couso, J. P., Bate, M. & Martinez-Arias, A. A wingless-dependent polar coordinate system in *Drosophila* imaginal discs. *Science* **259**, 484–489 (1993).
99. Goldenberg, D. P., Smith, D. H. & King, J. Genetic analysis of the folding pathway for the tail spike protein of phage P22. eng. *Proc Natl Acad Sci U S A* **80**, 7060–7064 (1983).
100. Warner, J. R. & Udem, S. A. Temperature sensitive mutations affecting ribosome synthesis in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **65**, 243–257 (1972).
101. Balakrishnan, R. & Backman, K. Controllable alteration of cell genotype in bacterial cultures using an excision vector. eng. *Gene* **67**, 97–103 (1988).
102. Suzuki, D. T., Grigliatti, T. & Williamson, R. Temperature-Sensitive Mutations in *Drosophila melanogaster*, VII. A Mutation (parats) Causing Reversible Adult Paralysis. *Proceedings of the National Academy of Sciences* **68**, 890–893 (1971).
103. Novick, P. & Botstein, D. Phenotypic analysis of temperature-sensitive yeast actin mutants. eng. *Cell* **40**, 405–416 (1985).
104. Hereford, L. M. & Hartwell, L. H. Sequential gene function in the initiation of *Saccharomyces cerevisiae* DNA synthesis. *Journal of Molecular Biology* **84**, 445–461 (1974).
105. Hartwell, L. H., Culotti, J. & Reid, B. Genetic control of the cell-division cycle in yeast. I. Detection of mutants. eng. *Proc Natl Acad Sci U S A* **66**, 352–359 (1970).
106. Bucking-Throm, E., Duntze, W., Hartwell, L. H. & Manney, T. R. Reversible arrest of haploid yeast cells in the initiation of DNA synthesis by a diffusible sex factor. eng. *Exp Cell Res* **76**, 99–110 (1973).
107. Wilkinson, L. E. & Pringle, J. R. Transient G1 arrest of *S. cerevisiae* cells of mating type alpha by a factor produced by cells of mating type a. eng. *Exp Cell Res* **89**, 175–187 (1974).

108. Mondal, K. *et al.* Design and Isolation of Temperature-sensitive Mutants of Gal4 in Yeast and Drosophila. *Journal of Molecular Biology* **370**, 939–950 (2007).
109. Mondal, K., VijayRaghavan, K. & Varadarajan, R. Design and Utility of Temperature-Sensitive Gal4 Mutants for Conditional Gene Expression in Drosophila. *Fly* **1**, 282–286 (2007).
110. Patrick, M., Duncan, R. & Coombs, K. M. Generation and genetic characterization of avian reovirus temperature-sensitive mutants. *eng. Virology* **284**, 113–122 (2001).
111. Varadarajan, R., Nagarajaram, H. & Ramakrishnan, C. A procedure for the prediction of temperature-sensitive mutants of a globular protein based solely on the amino acid sequence. *Proceedings of the National Academy of Sciences* **93**, 13908–13913 (1996).
112. Parsell, D. A. & Sauer, R. T. The structural stability of a protein is an important determinant of its proteolytic susceptibility in Escherichia coli. *Journal of Biological Chemistry* **264**, 7590–7595 (1989).
113. Pakula, A. A. & Sauer, R. T. Genetic Analysis of Protein Stability and Function. *Annual Review of Genetics* **23**, 289–310 (1989).
114. Razvi, A. & Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein science : a publication of the Protein Society* **15**, 1569–1578 (2006).
115. Sammond, D. W. *et al.* Comparing residue clusters from thermophilic and mesophilic enzymes reveals adaptive mechanisms. *PLoS ONE* **11** (Jan. 2016).
116. Pakula, A. A., Young, V. B. & Sauer, R. T. Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proceedings of the National Academy of Sciences* **83**, 8829–8833 (1986).
117. Kellis, J. T., Nyberg, K. & Fersht, A. R. Energetics of complementary side chain packing in a protein hydrophobic core. *Biochemistry* **28**, 4914–4922 (1989).
118. Shortle, D., Stites, W. E. & Meeker, A. K. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29**, 8033–8041 (1990).
119. Bajaj, K. *et al.* Structural Correlates of the Temperature Sensitive Phenotype Derived from Saturation Mutagenesis Studies of CcdB. *Biochemistry* **47**, 12964–12973 (2008).
120. Melo, F., Sanchez, R. & Sali, A. Statistical potentials for fold assessment. *eng. Protein Sci* **11**, 430–448 (2002).

121. John, B. & Sali, A. Comparative protein structure modeling by iterative alignment, model building and model assessment. eng. *Nucleic Acids Res* **31**, 3982–3992 (2003).
122. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences* **78**, 3824–3828 (1981).
123. Rose, G. D. & Roy, S. Hydrophobic basis of packing in globular proteins. *Proceedings of the National Academy of Sciences* **77**, 4643–4647 (1980).
124. Eisenberg, D. Three-Dimensional Structure of Membrane and Surface Proteins. *Annual Review of Biochemistry* **53**, 595–623 (1984).
125. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. Hydrophobicity of amino acid residues in globular proteins. eng. *Science* **229**, 834–838 (1985).
126. Tan, K. P., Khare, S., Varadarajan, R. & Madhusudhan, M. S. TSpred: A web server for the rational design of temperature-sensitive mutants. *Nucleic Acids Research* **42** (2014).
127. Sandberg, W. S., Schlunk, P. M., Zabin, H. B. & Terwilliger, T. C. Relationship between in Vivo Activity and in Vitro Measures of Function and Stability of a Protein. *Biochemistry* **34**, 11970–11978 (1995).
128. Alber, T., Sun, D. P., Nye, J. A., Muchmore, D. C. & Matthews, B. W. Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry* **26**, 3754–3758 (1987).
129. Hecht, M. H., Nelson, H. C. & Sauer, R. T. Mutations in lambda repressor's amino-terminal domain: implications for protein stability and DNA binding. eng. *Proc Natl Acad Sci U S A* **80**, 2676–2680 (1983).
130. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of lambda repressor. eng. *Nature* **339**, 31–36 (1989).
131. Jakubowska, A. & Korona, R. Lack of evolutionary conservation at positions important for thermal stability in the yeast ODCase protein. eng. *Mol Biol Evol* **26**, 1431–1434 (2009).
132. Chakshusmathi, G. *et al.* Design of temperature-sensitive mutants solely from amino acid sequence. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7925–7930 (2004).
133. Marti-Renom, M. A. *et al.* DBAli tools: mining the protein structure space. eng. *Nucleic Acids Res* **35**, W393–7 (2007).
134. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. eng. *Nucleic Acids Res* **37**, D26–31 (2009).

135. Leinonen, R. *et al.* UniProt archive. eng. *Bioinformatics* **20**, 3236–3237 (2004).
136. Consortium, T. U. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* **39**, D214–D219 (2011).
137. Huang, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. eng. *OMICS* **13**, 325–330 (2009).
138. Levitt, D. G. & Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. eng. *J Mol Graph* **10**, 229–234 (1992).
139. Huang, B. & Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural biology* **6**, 19. ISSN: 1472-6807 (2006).
140. Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. eng. *J Mol Graph* **13**, 307–308, 323–330. ISSN: 02637855 (1995).
141. Dundas, J. *et al.* CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research* **34**. ISSN: 03051048. doi:10.1093/nar/gkl282 (2006).
142. Brady Jr., G. P., Stouten, P. F. W., Brady, G. P. & Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. eng. *J Comput Aided Mol Des* **14**, 383–401. ISSN: 0920654X (2000).
143. Weisel, M., Proschak, E. & Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. eng. *Chem Cent J* **1**, 7 (2007).
144. Goodsell, D. S., Morris, G. M. & Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *Journal of molecular recognition : JMR* **9**, 1–5. ISSN: 0952-3499 (1998).
145. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology* **5**. ISSN: 1553734X. doi:10.1371/journal.pcbi.1000585 (2009).
146. Zhang, Z., Li, Y., Lin, B., Schroeder, M. & Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics (Oxford, England)* **27**, 2083–8. ISSN: 1367-4811 (2011).
147. Laurie, A. T. R. & Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. eng. *Bioinformatics* **21**, 1908–1916. ISSN: 13674803 (2005).

148. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **10**, 168. ISSN: 1471-2105 (2009).
149. Kawabata, T. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **78**, 1195–1211. ISSN: 1097-0134 (2010).
150. Yu, J., Zhou, Y., Tanaka, I. & Yao, M. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **26**, 46–52 (2010).
151. Dessailly, B. H., Lensink, M. F., Orengo, C. A. & Wodak, S. J. LigASite - A database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Research* **36**. ISSN: 03051048. doi:10.1093/nar/gkm839 (2008).
152. Henrick, K. & Thornton, J. M. PQS: A protein quaternary structure file server. *Trends in Biochemical Sciences* **23**, 358–361. ISSN: 09680004 (1998).
153. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288. ISSN: 13674803 (2007).
154. Marti-Renom, M. A., Madhusudhan, M. S. & Sali, A. Alignment of protein sequences by their profiles. *Protein Sci* **13**, 1071–1087. ISSN: 0961-8368 (2004).
155. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. eng. *J Mol Graph Model* **15**, 359–363, 389 (1997).
156. An, J., Totrov, M. & Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP* **4**, 752–61. ISSN: 1535-9476 (2005).
157. Gerstein, M., Tsai, J. & Levitt, M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *Journal of molecular biology* **249**, 955–966. ISSN: 0022-2836 (1995).
158. Chakravarty, S., Bhinge, A. & Varadarajan, R. *A Procedure for Detection and Quantitation of Cavity Volumes in Proteins* 2002. doi:10.1074/jbc.M201373200. <<http://eprints.iisc.ernet.in/12358/>>.
159. García-Moreno E, B. & Fitch, C. A. Structural interpretation of pH and salt-dependent processes in proteins with computational methods. *Methods in enzymology* **380**, 20–51. ISSN: 0076-6879 (Jan. 2004).
160. Hendsch, Z. S., Jonsson, T., Sauer, R. T. & Tidor, B. Protein stabilization by removal of unsatisfied polar groups: Computational approaches and experimental tests. *Biochemistry* **35**, 7621–7625. ISSN: 00062960 (1996).

161. Schaefer, M. *et al.* pH-Dependence of Protein Stability : Absolute Electrostatic Free Energy Differences. *Journal of Physical Chemistry B* **5647**, 1663–1683. ISSN: 1520-6106 (1997).
162. Tjong, H. & Zhou, H.-X. Prediction of protein solubility from calculation of transfer free energy. *Biophysical journal* **95**, 2601–9. ISSN: 1542-0086 (2008).
163. Antosiewicz, M. D. & M, J. The impact of protonation equilibria on protein structure. *Journal of Physics: Condensed Matter* **17**, S1607–S1616. ISSN: 0953-8984 (2005).
164. Warshel, A. Calculations of enzymatic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry* **20**, 3167–3177. ISSN: 0006-2960 (1981).
165. Bartik, K., Redfield, C. & Dobson, C. M. Measurement of the Individual pKa Values of Acidic Residues of Hen and Turkey Lysozymes by Two-Dimensional ¹H NMR. *Biophysical Journal* **66**, 1180–1184. ISSN: 0006-3495 (1994).
166. Oda, Y. *et al.* Individual ionization constants of all the carboxyl groups in ribonuclease HI from Escherichia coli determined by NMR. *Biochemistry* **33**, 5275–84. ISSN: 0006-2960 (1994).
167. Mehler, E. L. & Guarnieri, F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophysical journal* **77**, 3–22. ISSN: 00063495 (1999).
168. Pace, C. N. *et al.* Contribution of hydrogen bonds to protein stability. eng. *Protein science : a publication of the Protein Society* **23**, 652–661 (2014).
169. Cordier, F., Rogowski, M., Grzesiek, S. & Bax, A. Observation of Through-Hydrogen-Bond 2hJHC in a Perdeuterated Protein. *Journal of Magnetic Resonance* **140**, 510–512. ISSN: 1090-7807 (1999).
170. Isaacs, E. D. *et al.* Covalency of the Hydrogen Bond in Ice: A Direct X-Ray Measurement. *Physical Review Letters* **82**, 600–603 (Jan. 1999).
171. Eswar, N. & Ramakrishnan, C. Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. eng. *Protein engineering* **13**, 227–238 (2000).
172. Myers, J. K. & Pace, C. N. Hydrogen bonding stabilizes globular proteins. *Biophysical journal* **71**, 2033–2039. ISSN: 00063495 (1996).
173. Bordo, D. & Argos, P. The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. eng. *Journal of Molecular Biology* **243**, 504–519 (1994).

174. Eswar, N. & Ramakrishnan, C. Secondary structures without backbone: an analysis of backbone mimicry by polar side chains in protein structures. eng. *Protein engineering* **12**, 447–455 (1999).
175. Bennaim, A. The Role of Hydrogen-Bonds in Protein Folding and Protein Association. English. *Journal of Physical Chemistry* **95**, 1437–1444 (1991).
176. Deechongkit, S. *et al.* Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. *Nature* **430**, 101–105. ISSN: 0028-0836 (2004).
177. Rose, G. D., Fleming, P. J., Banavar, J. R. & Maritan, A. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 16623–16633. ISSN: 0027-8424 (2006).
178. Chatterjee, S. *et al.* Expanding the Peptide beta-Turn in alpha gamma Hybrid Sequences: 12 Atom Hydrogen Bonded Helical and Hairpin Turns. English. *Journal of the American Chemical Society* **131**, 5956–5965 (2009).
179. Fersht, A. R. *The hydrogen bond in molecular recognition* 1987. doi:10.1016/0968-0004(87)90146-0.
180. Thurlkill, R. L., Grimsley, G. R., Scholtz, J. M. & Pace, C. N. Hydrogen Bonding Markedly Reduces the pK of Buried Carboxyl Groups in Proteins. *Journal of Molecular Biology* **362**, 594–604. ISSN: 00222836 (2006).
181. Kurochkina, N. & Privalov, G. Heterogeneity of packing: structural approach. eng. *Protein science : a publication of the Protein Society* **7**, 897–905 (1998).
182. Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* **31**, 725–732. ISSN: 0006-2960 (1992).
183. Connelly, P. R. *et al.* Enthalpy of hydrogen bond formation in a protein-ligand binding reaction. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 1964–1968. ISSN: 0027-8424 (1994).
184. Eberhardt, E. S. & Raines, R. T. Amide-Amide and Amide-Water Hydrogen Bonds: Implications for Protein Folding and Stability. *Journal of the American Chemical Society* **116**, 2149–2150 (1994).
185. Byrne, M. P., Manuel, R. L., Lowe, L. G. & Stites, W. E. Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry* **34**, 13949–13960. ISSN: 0006-2960 (1995).
186. Habermann, S. M. & Murphy, K. P. Energetics of hydrogen bonding in proteins: a model compound study. *Protein science : a publication of the Protein Society* **5**, 1229–1239. ISSN: 0961-8368 (1996).

187. Sheu, S.-Y. Y., Yang, D.-Y. Y., Selzle, H. L. & Schlag, E. W. Energetics of hydrogen bonds in peptides. eng. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12683–12687. ISSN: 0027-8424 (2003).
188. Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6946–6951. ISSN: 0027-8424 (2004).
189. Cordier, F. & Grzesiek, S. Temperature-dependence properties as studied by of protein hydrogen bond high-resolution NMR. English. *Journal of Molecular Biology* **317**, 739–752 (2002).
190. Nisius, L. & Grzesiek, S. Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network. eng. *Nature chemistry* **4**, 711–717. ISSN: 1755-4349 (Sept. 2012).
191. Pace, C. N. Energetics of protein hydrogen bonds. eng. *Nature structural & molecular biology* **16**, 681–682 (2009).
192. Honig, B. Protein folding: from the levinthal paradox to structure prediction. eng. *Journal of Molecular Biology* **293**, 283–293 (1999).
193. Lazaridis, T., Archontis, G. & Karplus, M. Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. eng. *Advances in protein chemistry* **47**, 231–306 (1995).
194. Schell, D., Tsai, J., Scholtz, J. M. & Pace, C. N. Hydrogen bonding increases packing density in the protein interior. eng. *Proteins* **63**, 278–282 (2006).
195. Baker, E. N. & Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Progress in Biophysics and Molecular Biology* **44**, 97–179. ISSN: 00796107 (Jan. 1984).
196. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* **238**, 777–793. ISSN: 0022-2836 (1994).
197. Sticke, D. F. *et al.* Hydrogen bonding in globular proteins. eng. *Journal of Molecular Biology* **226**, 1143–1159. ISSN: 00222836 (Aug. 1992).
198. Fabiola, F., Bertram, R., Korostelev, A. & Chapman, M. S. An improved hydrogen bond potential: impact on medium resolution protein structures. *Protein science : a publication of the Protein Society* **11**, 1415–1423. ISSN: 0961-8368 (June 2002).

199. Grzybowski, B. A., Ishchenko, A. V., DeWitte, R. S., Whitesides, G. M. & Shakhnovich, E. I. Development of a knowledge-based potential for crystals of small organic molecules: Calculation of energy surfaces for C=O center dot center dot center dot H-N hydrogen bonds. English. *Journal of Physical Chemistry B* **104**, 7293–7298 (2000).
200. Ippolito, J. A., Alexander, R. S. & Christianson, D. W. Hydrogen bond stereochemistry in protein structure and function. eng. *Journal of Molecular Biology* **215**, 457–471. ISSN: 0022-2836 (1990).
201. Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology* **326**, 1239–1259. ISSN: 00222836 (Feb. 2003).
202. Lommerse, J. P. M., Price, S. L. & Taylor, R. Hydrogen bonding of carbonyl, ether, and ester oxygen atoms with alkanol hydroxyl groups. *Journal of Computational Chemistry* **18**, 757–774. ISSN: 1096-987X (Apr. 1997).
203. Taylor, R. & Kennard, O. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research* **17**, 320–326. ISSN: 0001-4842 (Sept. 1984).
204. Taylor, R., Kennard, O. & Versichel, W. Geometry of the N-H=O=C Hydrogen-Bond .1. Lone-Pair Directionality. English. *Journal of the American Chemical Society* **105**, 5761–5766. ISSN: 0002-7863 (Sept. 1983).
205. Radford, S. E., Buck, M., Topping, K. D., Dobson, C. M. & Evans, P. A. Hydrogen exchange in native and denatured states of hen egg-white lysozyme. *Proteins: Structure, Function, and Bioinformatics* **14**, 237–248 (1992).
206. Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* **13**, 222–245. ISSN: 0961-8368 (1974).
207. Kumar, M. D. *et al.* ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. eng. *Nucleic Acids Res* **34**, D204–6 (2006).
208. Raghavendra, B., Mandal, P. K. & Arunan, E. Ab initio and AIM theoretical analysis of hydrogen-bond radius of HD (D = F, Cl, Br, CN, HO, HS and CCH) donors and some acceptors. eng. *Physical chemistry chemical physics : PCCP* **8**, 5276–5286. ISSN: 1463-9076 (2006).
209. Grazulis, S. *et al.* Crystallography Open Database - an open-access collection of crystal structures. Eng. *Journal of applied crystallography* **42**, 726–729 (2009).
210. Grazulis, S. *et al.* Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. eng. *Nucleic Acids Research* **40**, D420–7 (2012).

- 211. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. eng. *Acta crystallographica. Section B, Structural science* **58**, 380–388 (2002).
- 212. Buck, M. & Karplus, M. Hydrogen bond energetics: A simulation and statistical analysis of N-methyl acetamide (NMA), water, and human lysozyme. *Journal of Physical Chemistry B* **105**, 11000–11015. ISSN: 10895647 (2001).
- 213. Miller, D. J. & Lisy, J. M. Modeling competitive interactions in proteins: vibrational spectroscopy of $M+(n\text{-methylacetamide})_1(\text{H}_2\text{O})_{n=0-3}$, $M=\text{Na}$ and K , in the 3 microm region. *The journal of physical chemistry. A* **111**, 12409–16. ISSN: 1089-5639 (2007).
- 214. Schmidt, M. W. *et al.* General atomic and molecular electronic structure system. *Journal of Computational Chemistry* **14**, 1347–1363. ISSN: 0192-8651 (1993).
- 215. Schaftenaar, G. & Noordik, J. H. Molden: A pre- and post-processing program for molecular and electronic structures. *Journal of Computer-Aided Molecular Design* **14**, 123–134. ISSN: 0920654X (2000).
- 216. Fleming, P. J. & Rose, G. D. Do all backbone polar groups in proteins form hydrogen bonds? English. *Protein Science* **14**, 1911–1917. ISSN: 0961-8368 (July 2005).
- 217. Savage, H. J., Elliott, C. J., Freeman, C. M. & Finney, J. L. Lost Hydrogen-Bonds and Buried Surface-Area - Rationalizing Stability in Globular-Proteins. English. *Journal of the Chemical Society-Faraday Transactions* **89**, 2609–2617 (1993).
- 218. Garbuzynskiy, S. O., Melnik, B. S., Lobanov, M. Y., Finkelstein, A. V. & Galzitskaya, O. V. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? eng. *Proteins* **60**, 139–147 (2005).
- 219. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–6. ISSN: 0036-8075 (2001).
- 220. Headd, J. J. *et al.* Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D: Biological Crystallography* **68**, 381–390. ISSN: 09074449 (2012).
- 221. Li, Y. & Zhang, Y. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Structure, Function and Bioinformatics* **76**, 665–676. ISSN: 08873585 (2009).
- 222. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. *Protein Structure Prediction Using Rosetta* 2004. doi:10.1016/S0076-6879(04)83004-0.

- 223. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **5**, 725–738. ISSN: 1754-2189 (2010).
- 224. Kota, P., Ding, F., Ramachandran, S. & Dokholyan, N. V. Gaia: automated quality assessment of protein structure models. eng. *Bioinformatics* **27**, 2209–2215. ISSN: 1367-4811 (Aug. 2011).
- 225. Shmymgelska, A. & Levitt, M. Generalized ensemble methods for de novo structure prediction. eng. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1415–1420 (2009).
- 226. Ramelot, T. A. *et al.* Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. eng. *Proteins* **75**, 147–167. ISSN: 1097-0134 (Apr. 2009).
- 227. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **30**, 1545–1614. ISSN: 01928651 (2009).
- 228. Ewig, C. S., Thacher, T. S. & Hagler, A. T. Derivation of class II force fields. 7. Nonbonded force field parameters for organic compounds. English. *Journal of Physical Chemistry B* **103**, 6998–7014 (1999).
- 229. Hagler, A. T., Lifson, S. & Dauber, P. Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 2. A benchmark for the objective comparison of alternative force fields. *Journal of the American Chemical Society* **101**, 5122–5130 (1979).
- 230. Lii, J. H. & Allinger, N. L. Directional hydrogen bonding in the MM3 force field: II. *Journal of Computational Chemistry* **19**, 1001–1016. ISSN: 1096-987X (1998).
- 231. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* **102**, 3586–3616 (1998).
- 232. Mayo, S. L., Olafson, B. D., Iii, W. A. G., Eb, E. & El, E. A. E. T. DREIDING: A Generic Force Field for Molecular Simulations. *Journal of Physical chemistry* **101**, 8897–8909. ISSN: 0022-3654 (1990).
- 233. Weiner, P. K. & Kollman, P. A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **2**, 287–303 (1981).
- 234. Duan, L. L., Mei, Y., Zhang, Q. G. & Zhang, J. Z. H. Intra-protein hydrogen bonding is dynamically stabilized by electronic polarization. English. *Journal of Chemical Physics* **130**, 115102. ISSN: 1089-7690 (Mar. 2009).
- 235. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. eng. *J Comput Chem* **25**, 1605–1612 (2004).

236. Ratnaparkhi, G. S., Ramachandran, S., Udgaonkar, J. B. & Varadarajan, R. Discrepancies between the NMR and x-ray structures of uncomplexed barstar: Analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **37**, 6958–6966. ISSN: 00062960 (1998).
237. Cordier, F., Barfield, M. & Grzesiek, S. Direct observation of C-alpha-H-alpha center dot center dot center dot O=C hydrogen bonds in proteins by interresidue ($^3J(C\alpha-C')$) scalar couplings. English. *Journal of the American Chemical Society* **125**, 15750–15751 (2003).
238. Cordier, F., Nisius, L., Dingley, A. J. & Grzesiek, S. Direct detection of N-H center dot center dot center dot O=C hydrogen bonds in biomolecules by NMR spectroscopy. English. *Nature protocols* **3**, 235–241 (2008).
239. Meier, S., Strohmeier, M., Blackledge, M. & Grzesiek, S. Direct observation of dipolar couplings and hydrogen bonds across a beta-hairpin in 8 M urea. English. *Journal of the American Chemical Society* **129**, 754–755 (2007).
240. Cordier, F., Wang, C., Grzesiek, S. & Nicholson, L. K. Ligand-induced strain in hydrogen bonds of the c-Src SH3 domain detected by NMR. eng. *Journal of Molecular Biology* **304**, 497–505 (2000).
241. Grzesiek, S., Cordier, F., Jaravine, V. & Barfield, M. Insights into biomolecular hydrogen bonds from hydrogen bond scalar couplings. English. *Progress in Nuclear Magnetic Resonance Spectroscopy* **45**, 275–300. ISSN: 00796565 (Dec. 2004).
242. Jensen, P. R., Axelsen, J. B., Lerche, M. H. & Poulsen, F. M. Improvement of hydrogen bond geometry in protein NMR structures by residual dipolar couplings—an assessment of the interrelation of NMR restraints. eng. *Journal of biomolecular NMR* **28**, 31–41 (2004).
243. Mullins, L. S., Raushel, F. M. & Pace, C. N. Conformational stability of ribonuclease T1 determined by hydrogen-deuterium exchange. *Protein Science* **6**, 1387–1395 (1997).
244. Alabugin, I. V., Gilmore, K. M. & Peterson, P. W. Hyperconjugation. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 109–141. ISSN: 17590876 (2011).
245. Dill, K. A. Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501–1509 (1985).
246. Dill, K. A. Additivity principles in biochemistry. *Journal of Biological Chemistry* **272**, 701–704. ISSN: 00219258 (1997).
247. Dobson, C. M., Šali, A. & Karplus, M. Protein folding: A perspective from theory and experiment. *Angewandte Chemie - International Edition* **37**, 868–893. ISSN: 14337851 (1998).

- 248. Shakhnovich, E. Protein Folding Thermodynamics and Dynamics : Where Physics , Chemistry , and Biology Meet Fundamental Model of Protein Folding. *Chemical Reviews* **106**, 1559–1588. ISSN: 0009-2665 (2006).
- 249. Sterner, R., Merkl, R. & Raushel, F. M. Computational Design of Enzymes. *Chemistry and Biology* **15**, 421–423. ISSN: 10745521 (2008).
- 250. Wu, X. *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science (New York, N.Y.)* **329**, 856–861. ISSN: 0036-8075 (2010).
- 251. Correia, B. E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–6. ISSN: 1476-4687 (2014).
- 252. Wang, Z. & Moulton, J. SNPs, protein structure, and disease. eng. *Hum Mutat* **17**, 263–270 (2001).
- 253. Venter, J. C. *et al.* The sequence of the human genome. eng. *Science* **291**, 1304–1351 (2001).
- 254. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. eng. *Nature* **449**, 851–861 (2007).
- 255. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. eng. *Nature* **491**, 56–65 (2012).
- 256. Okabe, M. *et al.* Profiling SLCO and SLC22 genes in the NCI-60 cancer cell lines to identify drug uptake transporters. eng. *Mol Cancer Ther* **7**, 3081–3091 (2008).
- 257. Azoitei, M. L. *et al.* Computational design of protein antigens that interact with the CDR H3 loop of HIV broadly neutralizing antibody 2F5. eng. *Proteins* **82**, 2770–2782 (2014).
- 258. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. eng. *Nucleic Acids Res* **31**, 3812–3814 (2003).
- 259. Smith, H. O., Annau, T. M. & Chandrasegaran, S. Finding sequence motifs in groups of functionally related proteins. eng. *Proc Natl Acad Sci U S A* **87**, 826–830 (1990).
- 260. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. eng. *Nature methods* **7**, 248–249 (2010).
- 261. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. eng. *J Mol Biol* **426**, 2692–2701 (2014).
- 262. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. eng. *Nucleic Acids Res* **33**, W306–10 (2005).

- 263. Masso, M. & Vaisman II. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. eng. *Adv Bioinformatics* **2014**, 278385 (2014).
- 264. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. eng. *Bioinformatics* **25**, 2537–2543 (2009).
- 265. Pires, D. E., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. eng. *Bioinformatics* **30**, 335–342 (2014).
- 266. Worth, C. L., Preissner, R. & Blundell, T. L. SDM—a server for predicting effects of mutations on protein stability and malfunction. eng. *Nucleic Acids Res* **39**, W215–22 (2011).
- 267. Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. eng. *Nucleic Acids Res* **42**, W314–9 (2014).
- 268. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, a. R. Systematic mutation of bacteriophage T4 lysozyme. eng. *J Mol Biol* **222**, 67–88. ISSN: 0022-2836 (Nov. 1991).
- 269. Remington, S. J. *et al.* Structure of the lysozyme from bacteriophage T4: an electron density map at 2.4 Å resolution. eng. *J Mol Biol* **118**, 81–98 (1978).
- 270. Weaver, L. H. & Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. eng. *J Mol Biol* **193**, 189–199 (1987).
- 271. Loris, R. *et al.* Crystal structure of CcdB, a topoisomerase poison from *E. coli*. eng. *J Mol Biol* **285**, 1667–1677 (1999).
- 272. Tompa, K. *et al.* Interfacial water at protein surfaces: Wide-line NMR and DSC characterization of hydration in ubiquitin solutions. *Biophysical Journal* **96**, 2789–2798. ISSN: 00063495 (2009).
- 273. Atilgan, a. R. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal* **80**, 505–15. ISSN: 0006-3495 (2001).
- 274. Doruker, P., Atilgan, a. R. & Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins* **40**, 512–524. ISSN: 0887-3585 (2000).
- 275. Nguyen, C. N., Kurtzman Young, T. & Gilson, M. K. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *The Journal of Chemical Physics* **137**. doi:<http://dx.doi.org/10.1063/1.4733951>. <<http://scitation.aip.org/content/aip/journal/jcp/137/4/10.1063/1.4733951>> (2012).

- 276. Betancourt, M. R. Empirical model of residue contact probabilities for polypeptides. *Journal of Chemical Physics* **132**. ISSN: 00219606. doi:10.1063/1.3328613 (2010).
- 277. Ferrada, E. & Melo, F. Effective knowledge-based potentials. *Protein Science* **18**, 1469–1485. ISSN: 09618368 (2009).
- 278. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Current opinion in structural biology* **18**, 581–586. ISSN: 0959440X (2008).
- 279. Nasmyth, K. & Haering, C. H. Cohesin: Its Roles and Mechanisms. *Annual Review of Genetics* **43**, 525–558. ISSN: 0066-4197 (2009).
- 280. Gruber, S., Haering, C. H. & Nasmyth, K. Chromosomal cohesin forms a ring. *Cell* **112**, 765–777. ISSN: 00928674 (2003).
- 281. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: Improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356–358. ISSN: 13674803 (2006).
- 282. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037–41. ISSN: 0027-8424 (2001).
- 283. Haering, C. H., Löwe, J., Hochwagen, A. & Nasmyth, K. Molecular architecture of SMC proteins and the yeast cohesin complex. *Molecular Cell* **9**, 773–788. ISSN: 10972765 (2002).
- 284. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science (New York, N.Y.)* **325**, 834–840. ISSN: 0036-8075 (2009).
- 285. Wagner, S. a. *et al.* A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles. *Molecular & Cellular Proteomics* **10**, M111.013284–M111.013284. ISSN: 1535-9476 (2011).
- 286. Roth, M., Obaidat, A. & Hagenbuch, B. OATPs, OATs and OCTs: The organic anion and cation transporters of the SLCO and SLC22A gene super-families. *British Journal of Pharmacology* **165**, 1260–1287. ISSN: 00071188 (2012).
- 287. Koepsell, H., Lips, K. & Volk, C. *Polyspecific organic cation transporters: Structure, function, physiological roles, and biopharmaceutical implications* 2007. doi:10.1007/s11095-007-9254-z.
- 288. Ohashi, R. *et al.* Molecular and physiological evidence for multifunctionality of carnitine/organic cation transporter OCTN2. *Molecular pharmacology* **59**, 358–366. ISSN: 0026-895X (2001).

- 289. Schimmenti, L. A. *et al.* Expanded newborn screening identifies maternal primary carnitine deficiency. *Molecular Genetics and Metabolism* **90**, 441–445. ISSN: 10967192 (2007).
- 290. Martínez, A. *et al.* Association of the organic cation transporter OCTN genes with Crohn's disease in the Spanish population. *European journal of human genetics : EJHG* **14**, 222–226. ISSN: 1018-4813 (2006).
- 291. Tamai, I. *et al.* Molecular and functional identification of sodium ion-dependent, high affinity human carnitine transporter OCTN2. *Journal of Biological Chemistry* **273**, 20378–20382. ISSN: 00219258 (1998).
- 292. Longo, N., Amat Di San Filippo, C. & Pasquali, M. Disorders of carnitine transport and the carnitine cycle. *American Journal of Medical Genetics - Seminars in Medical Genetics* **142 C**, 77–85. ISSN: 15524868 (2006).
- 293. Amat Di San Filippo, C., Pasquali, M. & Longo, N. Pharmacological rescue of carnitine transport in primary carnitine deficiency. *Human Mutation* **27**, 513–523. ISSN: 10597794 (2006).
- 294. Ohashi, R. *et al.* Na(+)-dependent carnitine transport by organic cation transporter (OCTN2): its pharmacological and toxicological relevance. *The Journal of pharmacology and experimental therapeutics* **291**, 778–784. ISSN: 0022-3565 (1999).
- 295. Shoji, Y. *et al.* Evidence for linkage of human primary systemic carnitine deficiency with D5S436: a novel gene locus on chromosome 5q. *American journal of human genetics* **63**, 101–8. ISSN: 0002-9297 (1998).
- 296. Pochini, L., Scalise, M., Galluccio, M. & Indiveri, C. OCTN cation transporters in health and disease: role as drug targets and assay development. *Journal of biomolecular screening* **18**, 851–867. ISSN: 1552-454X (2013).
- 297. Toh, D. S. L., Yee, J. Y., Koo, S. H., Murray, M. & Lee, E. J. D. Genetic variations of the SLC22A5 gene in the Chinese and Indian populations of Singapore. *Drug metabolism and pharmacokinetics* **25**, 112–9. ISSN: 1880-0920 (2010).
- 298. Pao, S. S., Paulsen, I. T. & Saier, M. H. Major facilitator superfamily. *Microbiology and molecular biology reviews : MMBR* **62**, 1–34. ISSN: 1092-2172 (1998).
- 299. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625. ISSN: 13674803 (2006).
- 300. Krogh, a., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567–580. ISSN: 0022-2836 (2001).

- 301. Sonnhammer, E. L., von Heijne, G. & Krogh, a. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **6**, 175–182. ISSN: 1553-0833 (1998).
- 302. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680. ISSN: 0305-1048 (1994).
- 303. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* **35**. ISSN: 03051048. doi:10.1093/nar/gkm290 (2007).
- 304. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* **32**. ISSN: 03051048. doi:10.1093/nar/gkh468 (2004).
- 305. Toh, D. S. L. *et al.* Functional analysis of pharmacogenetic variants of human organic cation/carnitine transporter 2 (hOCTN2) identified in Singaporean populations. *Biochemical Pharmacology* **82**, 1692–1699 (2011).
- 306. Toh, D. S. *et al.* Functional analysis of novel variants in the organic cation/ergothioneine transporter 1 identified in Singapore populations. *eng. Mol Pharm* **10**, 2509–2516 (2013).
- 307. Von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology* **225**, 487–494. ISSN: 00222836 (1992).
- 308. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629. ISSN: 0028-0836 (2002).
- 309. Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11606–11. ISSN: 0027-8424 (2003).
- 310. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nature genetics* **42**, 181–5. ISSN: 1546-1718 (2010).
- 311. Cao, R. *et al.* Role of hPHF1 in H3K27 methylation and Hox gene silencing. *Molecular and cellular biology* **28**, 1862–72. ISSN: 1098-5549 (2008).
- 312. Chen, Y.-H., Hung, M.-C. & Li, L.-Y. EZH2: a pivotal regulator in controlling cell differentiation. *American journal of translational research* **4**, 364–75. ISSN: 1943-8141 (2012).

- 313. Fujii, S., Ito, K., Ito, Y. & Ochiai, A. Enhancer of zeste homologue 2 (EZH2) down-regulates RUNX3 by increasing histone H3 methylation. *Journal of Biological Chemistry* **283**, 17324–17332. ISSN: 00219258 (2008).
- 314. Fujii, S. & Ochiai, A. Enhancer of zeste homolog 2 downregulates E-cadherin by mediating histone H3 methylation in gastric cancer cells. *Cancer Science* **99**, 738–746. ISSN: 13479032 (2008).
- 315. Ezhkova, E. *et al.* Ezh2 Orchestrates Gene Expression for the Stepwise Differentiation of Tissue-Specific Stem Cells. *Cell* **136**, 1122–1135. ISSN: 00928674 (2009).
- 316. Chou, R. H., Yu, Y. L. & Hung, M. C. The roles of EZH2 in cell lineage commitment. *American Journal of Translational Research* **3**, 243–250. ISSN: 19438141 (2011).
- 317. Min, J. *et al.* An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-kappaB. *Nature medicine* **16**, 286–94. ISSN: 1546-170X (2010).
- 318. Rao, Z.-Y. *et al.* EZH2 supports ovarian carcinoma cell invasion and/or metastasis via regulation of TGF-beta1 and is a predictor of outcome in ovarian carcinoma patients. *Carcinogenesis* **31**, 1576–83. ISSN: 1460-2180 (2010).
- 319. Su, I. H. *et al.* Polycomb group protein Ezh2 controls actin polymerization and cell signaling. *Cell* **121**, 425–436. ISSN: 00928674 (2005).
- 320. Su, I. H. & Tarakhovsky, A. Lysine methylation and 'signaling memory'. *Current Opinion in Immunology* **18**, 152–157. ISSN: 09527915 (2006).
- 321. Hobert, O., Jallal, B. & Ullrich, a. Interaction of Vav with ENX-1, a putative transcriptional regulator of homeobox gene expression. *Molecular and cellular biology* **16**, 3066–3073. ISSN: 0270-7306 (1996).
- 322. Fernandez-Zapico, M. E. *et al.* Ectopic expression of VAV1 reveals an unexpected role in pancreatic cancer tumorigenesis. *Cancer Cell* **7**, 39–49. ISSN: 15356108 (2005).
- 323. Groysman, M., Nagano, M., Shaanan, B. & Katzav, S. Mutagenic analysis of Vav reveals that an intact SH3 domain is required for transformation. *Oncogene* **17**, 1597–1606. ISSN: 0950-9232 (1998).
- 324. Hornstein, I., Alcover, A. & Katzav, S. Vav proteins, masters of the world of cytoskeleton organization. *Cellular Signalling* **16**, 1–11. ISSN: 08986568 (2004).
- 325. Lazer, G., Pe'er, L., Schapira, V., Richard, S. & Katzav, S. The association of Sam68 with Vav1 contributes to tumorigenesis. *Cellular Signalling* **19**, 2479–2486. ISSN: 08986568 (2007).

- 326. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function and Genetics* **17**, 355–362. ISSN: 08873585 (1993).
- 327. Cheng, J. DOMAC: An accurate, hybrid protein domain prediction server. *Nucleic Acids Research* **35**. ISSN: 03051048. doi:10.1093/nar/gkm390 (2007).
- 328. Ebina, T., Toh, H. & Kuroda, Y. Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics. *Biopolymers* **92**, 1–8. ISSN: 00063525 (2009).
- 329. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research* **31**, 3701–3708. ISSN: 03051048 (2003).
- 330. Bystroff, C. & Shao, Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S54–S61. ISSN: 1367-4803 (2002).
- 331. Bystroff, C., Thorsson, V. & Baker, D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *Journal of molecular biology* **301**, 173–190. ISSN: 0022-2836 (2000).
- 332. Garnier, J., Gibrat, J. F. & Robson, B. GOR secondary structure prediction method version IV. *Methods Enzymol* **266**, 540–553 (1996).
- 333. Bryson, K. *et al.* Protein structure prediction servers at University College London. *Nucleic Acids Research* **33**. ISSN: 03051048. doi:10.1093/nar/gki410 (2005).
- 334. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic acids research* **36**. ISSN: 13624962. doi:10.1093/nar/gkn238 (2008).
- 335. Castresana, J. & Saraste, M. Does Vav bind to F-actin through a CH domain? *FEBS Letters* **374**, 149–151. ISSN: 00145793 (1995).
- 336. Davis, F. P. *et al.* Protein complex compositions predicted by structural similarity. *eng. Nucleic Acids Res* **34**, 2943–2952 (2006).
- 337. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876–4882. ISSN: 03051048 (1997).
- 338. Verdine, G. L. & Hilinski, G. J. Stapled peptides for intracellular drug targets. *Methods in Enzymology* **503**, 3–33. ISSN: 00766879 (2012).

Appendix A

Appendix

A.1 Data sets

This section lists PDB identifiers for all the protein structures used in the thesis.

A.1.1 Ramachandran map data-set

The following protein structures are used to generate the high-resolution Ramachandran map used in the thesis (c.f. Section 2.8).

1A1XA, 1A3AA, 1A3CA, 1A62A, 1AE9A, 1ALYA, 1AT0A, 1ATZA, 1AYOA, 1BBHA, 1BEAA, 1BEBA, 1BGCA, 1BGFA, 1BKRA, 1BYFA, 1BYRA, 1C7KA, 1CCWA, 1CEWI, 1CO6A, 1CPQA, 1CV8A, 1CXQA, 1D0QA, 1D2OA, 1D2SA, 1D4OA, 1D4TA, 1D7PM, 1DBWA, 1DDWA, 1DLWA, 1DQGA, 1DY5A, 1DZKA, 1E29A, 1E87A, 1EAQA, 1EARA, 1EB6A, 1ECAA, 1EJ2A, 1EJ8A, 1EKGA, 1ELKA, 1EW4A, 1EX2A, 1EY4A, 1EYHA, 1F1EA, 1F1MA, 1F32A, 1F3VA, 1F46A, 1F5MA, 1F6BA, 1F7LA, 1F86A, 1F9ZA, 1FC3A, 1FITA, 1FL0A, 1FLMA, 1FM0E, 1FNLA, 1FQIA, 1G12A, 1G3KA, 1G5TA, 1G8IA, 1G8KB, 1GAKA, 1GBSA, 1GMIA, 1GMUA, 1GMXA, 1GNYA, 1GP0A, 1GPRA, 1GR3A, 1GU2A, 1GU9A, 1GUIA, 1GWYA, 1GXJA, 1GY7A, 1H05A, 1H4AX, 1H4XA, 1H6FA, 1H97A, 1HD2A, 1HE1A, 1HTWA, 1HUFA, 1HXRA, 1HYVA, 1HZTA, 1I0VA, 1I1QB, 1I4JA, 1I4UA, 1I58A, 1I8AA, 1ID0A, 1IDPA, 1IFRA, 1IIBA, 1IJYA, 1IKTA, 1IO0A, 1IOOA, 1IQ4A, 1IS3A, 1ISPA, 1IT2A, 1ITVA, 1IWLA, 1IWMA, 1IZMA, 1J24A, 1J2RA, 1J3AA, 1J3MA,

1J3WA, 1J83A, 1J98A, 1JATB, 1JBEA, 1JF3A, 1JF8A, 1JH6A, 1JHJA, 1JHSA,
1JIDA, 1JIWI, 1JKEA, 1JL1A, 1JMVA, 1JR8A, 1JUVA, 1JWQA, 1JYAA, 1JYHA,
1JYOA, 1K2XB, 1K3SA, 1K6KA, 1K8KE, 1K94A, 1KAFA, 1KGDA, 1KHCA,
1KHIA, 1KHYA, 1KMTA, 1KMVA, 1KNGA, 1KOEa, 1KPFA, 1KPTA, 1KQRA,
1KSHB, 1KT6A, 1KTGA, 1KXOA, 1L3KA, 1L3PA, 1L6PA, 1L7LA, 1L8RA, 1LB6A,
1LF7A, 1LKIA, 1LKKA, 1LMIA, 1LO7A, 1LQVA, 1LR5A, 1LU4A, 1LWBA, 1LYQA,
1M1FA, 1M2DA, 1M48A, 1M4IA, 1M4JA, 1M55A, 1M70A, 1M7BA, 1M9ZA, 1MAIA,
1MBAA, 1MBMA, 1MC2A, 1MDCA, 1MF7A, 1MG4A, 1MGTA, 1MI8A, 1MK4A,
1MKAA, 1MWQA, 1MXIA, 1MY7A, 1N08A, 1N1FA, 1N62A, 1N71A, 1N8VA,
1NBCA, 1NBUB, 1NEPA, 1NIGA, 1NJHA, 1NLQA, 1NQZA, 1NRVA, 1NRZA,
1NS5A, 1NTVA, 1NWWA, 1NWZA, 1NXMA, 1NZ0A, 1O13A, 1O6DA, 1O7IA,
1O8XA, 1O9RA, 1OA8A, 1OBOA, 1OD3A, 1OGDA, 1OH0A, 1OH4A, 1OI0A,
1OO0A, 1OOHA, 1OQVA, 1ORUA, 1OU8A, 1OW1A, 1OW4A, 1P5VB, 1P6OA,
1P90A, 1PBJA, 1PK6A, 1PL3A, 1PM4A, 1PMHX, 1POCA, 1PQ1A, 1PVMA,
1PZ4A, 1Q1FA, 1Q4UA, 1Q5ZA, 1Q8DA, 1Q9UA, 1QF8A, 1QFOA, 1QGVA,
1QKRA, 1QMYA, 1QU9A, 1QV1A, 1R4VA, 1R8SA, 1R9FA, 1R9WA, 1RG8A,
1RKIA, 1RLIA, 1RLJA, 1ROCA, 1ROWA, 1RTTA, 1RW1A, 1RXDA, 1RY9A,
1RYLA, 1S14A, 1S1QA, 1S3CA, 1S4KA, 1S5AA, 1S5UA, 1S7ZA, 1SAUA, 1SBZA,
1SEIA, 1SENA, 1SFPA, 1SHUX, 1SJWA, 1SMBA, 1SRAA, 1SVYA, 1SZHA, 1T1JA,
1T2WA, 1T3YA, 1T6UA, 1T82A, 1T9FA, 1T9IA, 1TC5A, 1THQA, 1TJXA, 1TP6A,
1TT8A, 1TU9A, 1TUHA, 1TVGA, 1U14A, 1U2BA, 1U53A, 1U5DA, 1U5XA, 1U7KA,
1UCDA, 1UFYA, 1UGXA, 1UJCA, 1UKFA, 1UNQA, 1UPQA, 1UUYA, 1UUZA,
1UWFA, 1UWWA, 1UXZA, 1UZKA, 1V2BA, 1V2XA, 1V2ZA, 1V30A, 1V4PA,
1V74A, 1V8CA, 1V8HA, 1V96A, 1V9TA, 1VCHA, 1VETA, 1VH5A, 1VI4A, 1VKFA,
1VKIA, 1VKKA, 1VL7A, 1VMBA, 1VMHA, 1VPMA, 1VQSA, 1VSRA, 1VYIA,
1VZIA, 1W0NA, 1W1HA, 1W4SA, 1WBIA, 1WC2A, 1WDJA, 1WHIA, 1WJXA,
1WKAA, 1WKCA, 1WKOA, 1WL8A, 1WLUA, 1WN2A, 1WNAA, 1WNYA, 1WO8A,
1WOLA, 1WPNA, 1WPUA, 1WS8A, 1WUBA, 1WVHA, 1WWIA, 1WWZA, 1X3KA,
1X6OA, 1X6ZA, 1X8QA, 1X91A, 1XAUa, 1XBIA, 1XE0A, 1XEOA, 1XFSA, 1XHNA,
1XIZA, 1XJUA, 1XKIA, 1XKPB, 1XKPC, 1XLQA, 1XODA, 1XS0A, 1XT5A,
1XWVA, 1XWWA, 1Y0BA, 1Y0HA, 1Y43B, 1Y5HA, 1Y71A, 1Y93A, 1Y9LA,

1YB0A, 1YB3A, 1YD9A, 1YFUA, 1YGHA, 1YHFA, 1YKUA, 1YLMa, 1YLXA,
1YN9A, 1YOZA, 1YPQA, 1YPYA, 1YQ5A, 1YQHA, 1YSQA, 1YSRA, 1YTLA,
1YV4A, 1YW5A, 1YZ1A, 1YZFA, 1Z0JA, 1Z2UA, 1Z2WA, 1Z6MA, 1Z6NA, 1Z9LA,
1Z9NA, 1ZBFA, 1ZC3B, 1ZCEA, 1ZD7A, 1ZE3H, 1ZHSA, 1ZHVA, 1ZK5A, 1ZMAA,
1ZTDA, 1ZUOA, 1ZX8A, 1ZZWA, 256BA, 2A0BA, 2A15A, 2A2KA, 2A4VA, 2A67A,
2A7BA, 2A9IA, 2A9SA, 2AI4A, 2AJ6A, 2ANXA, 2AQ6A, 2ASFA, 2AU7A, 2AVKA,
2AZWA, 2B06A, 2B0VA, 2B3GA, 2BBRA, 2BCMA, 2BEMA, 2BK9A, 2BKMA,
2BNLA, 2BRFA, 2BRJA, 2BWQA, 2BYCA, 2BZ1A, 2BZVA, 2C2IA, 2C2QA,
2C8SA, 2C92A, 2CARA, 2CAYA, 2CCMA, 2CE2X, 2CF7A, 2CHCA, 2CHHA,
2CJ4A, 2CJTA, 2CKKA, 2CM4A, 2CVDA, 2CW9A, 2CX1A, 2CX7A, 2CXXA,
2CY5A, 2CYJA, 2D48A, 2DC4A, 2DJFA, 2DP9A, 2DQAA, 2DREA, 2DTCA,
2DTJA, 2DVKA, 2DXAA, 2DY0A, 2E5YA, 2E7VA, 2E85A, 2EBEA, 2EH3A, 2EHGA,
2EHPA, 2EJ8A, 2EJNA, 2EJXA, 2ENDA, 2EO4A, 2ERBA, 2ERVA, 2ES9A, 2F01A,
2F22A, 2F46A, 2F4WA, 2F9HA, 2FA1A, 2FB6A, 2FCJA, 2FCKA, 2FCLA, 2FD5A,
2FDXA, 2FEXA, 2FG1A, 2FHPA, 2FHQA, 2FHZA, 2FI1A, 2FIPA, 2FKBA, 2FLHA,
2FOJA, 2FP1A, 2FP7B, 2FPHX, 2FQ4A, 2FR5A, 2FRGP, 2FSRA, 2FSXA, 2FTBA,
2FTRA, 2FUJA, 2FULA, 2FWHA, 2FWTA, 2FYGA, 2FYQA, 2G2CA, 2G3WA,
2G5RA, 2G64A, 2G7BA, 2G7SA, 2GDMA, 2GF6A, 2GHTA, 2GIAB, 2GIYA,
2GKGA, 2GKPA, 2GLZA, 2GMYA, 2GRCA, 2GRRB, 2GS5A, 2GU3A, 2GU9A,
2GUIA, 2GUKA, 2GX5A, 2GZBA, 2H1CA, 2H29A, 2H30A, 2H5CA, 2HAZA,
2HBOA, 2HC8A, 2HD9A, 2HDVA, 2HEWF, 2HFNA, 2HHZA, 2HIYA, 2HKVA,
2HL0A, 2HMHA, 2HNGA, 2HP7A, 2HQ4A, 2HQ7A, 2HQ9A, 2HQSC, 2HSBA,
2HU9A, 2HUHA, 2HW2A, 2HX5A, 2HY5A, 2HY5B, 2HY5C, 2HZQA, 2I02A,
2I2QA, 2I4AA, 2I6CA, 2I74A, 2I7DA, 2I8GA, 2I8TA, 2I9CA, 2IABA, 2IAYA,
2IC2A, 2ICGA, 2IDLA, 2IF6A, 2IGIA, 2IGPA, 2IIHA, 2IJQA, 2IKBA, 2IM8A,
2IMGA, 2IMJA, 2IMLA, 2IMSA, 2IN0A, 2IQJA, 2IQYA, 2IT9A, 2ITEA, 2IU5A,
2J12A, 2J1AA, 2J1VA, 2J2JA, 2J6AA, 2J6BA, 2J73A, 2J8WA, 2JCQA, 2JDAA,
2JDCA, 2JEKA, 2JGBA, 2JJUA, 2JKGA, 2MCMA, 2MHRA, 2MSBA, 2NL9A,
2NLVA, 2NR7A, 2NRKA, 2NRLA, 2NRRA, 2NUHA, 2NUJA, 2NVHA, 2NW0A,
2NWFA, 2NX2A, 2NX4A, 2NYUA, 2O0QA, 2O5FA, 2O5HA, 2O5UA, 2O6FA,
2O6LA, 2O70A, 2O7AA, 2O90A, 2O99A, 2OA2A, 2OAFa, 2OB5A, 2OD0A, 2ODAA,

2OEBA, 2OFCA, 2OFKA, 2OGBA, 2OGFA, 2OH1A, 2OHWA, 2OIKA, 2OIWA, 2OIXA, 2OIZD, 2OKFA, 2OKMA, 2OKUA, 2OLMA, 2OLPA, 2OMDA, 2OMLA, 2OMZB, 2OOCA, 2OOKA, 2OPCA, 2OQ8A, 2ORWA, 2OS0A, 2OTMA, 2OU5A, 2OU6A, 2OV0A, 2OVSA, 2OWPA, 2OX6A, 2OXGB, 2OYOA, 2OZJA, 2OZNA, 2P08A, 2P0DA, 2P0NA, 2P12A, 2P14A, 2P2OA, 2P39A, 2P45B, 2P58C, 2P65A, 2P75A, 2P8GA, 2PA7A, 2PAGA, 2PD1A, 2PH0A, 2PI2E, 2PIEA, 2PJSA, 2PLNA, 2PMAA, 2PN0A, 2PNDA, 2PPNA, 2PQVA, 2PR7A, 2PRXA, 2PTHA, 2PTTA, 2PTTB, 2PU9A, 2PV2A, 2PVBA, 2PWWA, 2PYQA, 2Q03A, 2Q3TA, 2Q3WA, 2Q3XA, 2Q4FA, 2Q4NA, 2Q5CA, 2Q5XA, 2Q82A, 2QEBA, 2QF4A, 2QFEA, 2QH9A, 2QHKA, 2QHLA, 2QIYA, 2QJWA, 2QJZA, 2QLWA, 2QMLA, 2QMMA, 2QNGA, 2QQ4A, 2QQYA, 2QS9A, 2QTDA, 2QUOA, 2QUPA, 2QV8A, 2QVGA, 2QYCA, 2QZQA, 2R16A, 2R25A, 2R2AA, 2R2YA, 2R4IA, 2R4QA, 2RA6A, 2RA9A, 2RBBA, 2RBGA, 2RC3A, 2RDGA, 2RDKA, 2RDMA, 2REMA, 2RFFA, 2RG8A, 2RGQA, 2RH3A, 2RJ2A, 2RK3A, 2RKQA, 2RL8A, 2RLDA, 2SAKA, 2SCPA, 2SICI, 2TNFA, 2UV4A, 2V0UA, 2V1MA, 2V2PA, 2V4XA, 2V9VA, 2VB1A, 2VBUA, 2VEBA, 2VJWA, 2VKPA, 2VLQB, 2VMHA, 2VNGA, 2VOFA, 2VOKA, 2VQ8A, 2VSMB, 2VUVA, 2VVWA, 2VXGA, 2VXTI, 2VZCA, 2VZPA, 2W0GA, 2W0IA, 2W1RA, 2W1SA, 2W31A, 2W3GA, 2W3XA, 2W72B, 2W9HA, 2W9YA, 2WCJA, 2WCWA, 2WFIA, 2WFOA, 2YQYA, 2YSKA, 2YVAA, 2YVQA, 2YWNA, 2YWWA, 2YXBA, 2YYOA, 2YZJA, 2YZKA, 2Z08A, 2Z0BA, 2Z0TA, 2Z0XA, 2Z10A, 2Z3HA, 2Z3QA, 2Z5BA, 2Z5WA, 2Z6FA, 2Z6OA, 2Z8AA, 2ZAYA, 2ZCAA, 2ZCMA, 2ZDBA, 2ZF9A, 2ZGLA, 2ZHNA, 2ZK9X, 2ZOUA, 2ZQOA, 2ZS0B, 2ZS0D, 2ZU1A, 2ZVYA, 3A07A, 3A0YA, 3A35A, 3A4CA, 3B33A, 3B44A, 3B6IA, 3B79A, 3B7CA, 3B8FA, 3BA3A, 3BB9A, 3BBBA, 3BDVA, 3BFQG, 3BJKA, 3BJNA, 3BKRA, 3BLNA, 3BODA, 3BQAA, 3BRCA, 3BS1A, 3BS2A, 3BT5A, 3BVFA, 3BWUD, 3BWVA, 3BWZA, 3BX6A, 3BY8A, 3BYQA, 3C5KA, 3C71A, 3C7MA, 3C7XA, 3C8IA, 3C8LA, 3C9PA, 3CANA, 3CBNA, 3CCGA, 3CEXA, 3CG4A, 3CG6A, 3CHBD, 3CHMA, 3CI6A, 3CITA, 3CK1A, 3CNBA, 3CNEA, 3CNVA, 3CP5A, 3CPTB, 3CQNA, 3CRYA, 3CT5A, 3CT6A, 3CTGA, 3CYPB, 3CZXA, 3D1BA, 3D1MA, 3D1PA, 3D32A, 3D3BA, 3D3MA, 3D4EA, 3D6RA, 3D79A, 3D7AA, 3D7JA, 3D85C, 3D9NA, 3DAIA, 3DB7A, 3DCMX, 3DD7A, 3DFGA, 3DJ9A, 3DLQI, 3DM8A,

3DMNA, 3DO8A, 3DOUA, 3DPJA, 3DQYA, 3DR6A, 3DR9A, 3DRZA, 3DSZA, 3DT5A, 3DXEA, 3E0HA, 3E0ZA, 3E1EA, 3E4GA, 3E6SA, 3E8MA, 3E99A, 3E9VA, 3EBQA, 3EBTA, 3EBYA, 3EC6A, 3ECFA, 3EDOA, 3EEAA, 3EF4A, 3EJGA, 3EJKA, 3ELLA, 3ELSA, 3ENUA, 3ER7A, 3ESMA, 3EULA, 3EURA, 3EWOA, 3EXNA, 3EYEA, 3EYTA, 3EZIA, 3F0DA, 3F14A, 3F2VA, 3F2ZA, 3F40A, 3F43A, 3F4AA, 3F4MA, 3F59A, 3F5BA, 3F5OA, 3F62A, 3F6GA, 3F7EA, 3F7LA, 3F7XA, 3F8XA, 3F9SA, 3FBUA, 3FCDA, 3FDXA, 3FF9A, 3FFVA, 3FFYA, 3FG8A, 3FGYA, 3FH1A, 3FH2A, 3FHVA, 3FIQA, 3FIVA, 3FJSA, 3FKAA, 3FKEA, 3FL2A, 3FM2A, 3FN5A, 3FNCA, 3FOVA, 3FPNA, 3FQMA, 3FSAA, 3FSOA, 3FW2A, 3FZ4A, 3FZEA, 3G0MA, 3G14A, 3G7GA, 3G98A, 3GA3A, 3GA4A, 3GBWA, 3GBYA, 3GDWA, 3GFPA, 3GGNA, 3GGQA, 3GHJA, 3GHXA, 3GLVA, 3GMXA, 3GPGA, 3GQQA, 3GRDA, 3GWNA, 3GXBA, 3GXHA, 3GXSA, 3GY9A, 3GYKA, 3H05A, 3H0NA, 3H51A, 3H6QA, 3H6RA, 3H79A, 3H87A, 3H8TA, 3H9WA, 3HCJA, 3HH1A, 3HHVA, 3HIUA, 3HM4A, 3HOIA, 3HP4A, 3HQXA, 3HT1A, 3HTNA, 3HVWA, 3HWUA, 3HX8A, 3HXIA, 3HYNA, 3HZPA, 3I24A, 3I2VA, 3I7MA, 3IA1A, 3IBMA, 3IDFA, 3IDUA, 3IE4A, 3IGNA, 3IGRA, 3IHTA, 3IJMA, 3IKBA, 3IMKA, 3INOA, 3IP0A, 3IQ1A, 3IQTA, 3IRBA, 3IU5A, 3IU6A, 3IV4A, 3IVVA, 3IWTA, 3IX3A, 3IXCA, 3JRNA, 3JRVA, 3JSRA, 3JTWA, 3JUIA, 3JX9A, 3JXSA, 3K0ZA, 3K7CA, 3K9WA, 3KDHA, 3KE7A, 3NULA, 3TSSA, 3VUBA, 4UBPA

A.1.2 Depth and accessibility data-set

The following protein structures are used to study the depth distribution and accessibility of different amino acids (c.f. Section 3.3).

1A1XA, 1A3AA, 1A3CA, 1A62A, 1AE9A, 1ALYA, 1AOCA, 1APYA, 1AQZA, 1AT0A, 1ATZA, 1AYOA, 1BBHA, 1BEAA, 1BEBA, 1BGCA, 1BGFA, 1BKRA, 1BYFA, 1BYRA, 1C7KA, 1CCWA, 1CCZA, 1CEWI, 1CMCA, 1CO6A, 1CPQA, 1CS3A, 1CV8A, 1CXQA, 1D0QA, 1D2OA, 1D2SA, 1D2ZB, 1D4OA, 1D4TA, 1D7PM, 1DBWA, 1DDWA, 1DG6A, 1DLWA, 1DQGA, 1DY5A, 1DZ3A, 1DZKA, 1E29A, 1E87A, 1EAQA, 1EARA, 1EB6A, 1ECAA, 1EERA, 1EJ2A, 1EJ8A, 1EKGA, 1ELKA, 1EUWA, 1EW0A, 1EW4A, 1EX2A, 1EY4A, 1EYHA, 1F1EA, 1F1MA, 1F32A,

1F39A, 1F3VA, 1F46A, 1F5MA, 1F6BA, 1F7DA, 1F7LA, 1F86A, 1F9ZA, 1FC3A, 1FITA, 1FL0A, 1FLMA, 1FM0E, 1FNLA, 1FQIA, 1FXLA, 1G12A, 1G1TA, 1G3KA, 1G5TA, 1G8KB, 1GAKA, 1GBSA, 1GMIA, 1GMUA, 1GMXA, 1GNYA, 1GP0A, 1GPRA, 1GR3A, 1GU2A, 1GU9A, 1GUIA, 1GVJA, 1GWMA, 1GWYA, 1GXJA, 1GY7A, 1H03P, 1H05A, 1H2CA, 1H4AX, 1H4XA, 1H6FA, 1H7CA, 1H8UA, 1H97A, 1HD2A, 1HE1A, 1HTWA, 1HUFA, 1HUWA, 1HXIA, 1HXR, 1HYVA, 1HZTA, 1I0RA, 1I0VA, 1I12A, 1I1JA, 1I1QB, 1I4JA, 1I4UA, 1I58A, 1I8AA, 1ID0A, 1IDPA, 1IFRA, 1IIBA, 1IJYA, 1IKTA, 1IO0A, 1IOOA, 1IQ4A, 1IS3A, 1ISPA, 1IT2A, 1ITVA, 1IUJA, 1IWLA, 1IWMA, 1IZMA, 1J0PA, 1J24A, 1J2RA, 1J3AA, 1J3MA, 1J3WA, 1J5UA, 1J83A, 1J8RA, 1J98A, 1JATB, 1JBEA, 1JBOA, 1JD5A, 1JF3A, 1JF8A, 1JH6A, 1JHJA, 1JHSA, 1JIDA, 1JIWI, 1JKEA, 1JKGA, 1JL1A, 1JMVA, 1JOSA, 1JR8A, 1JUVA, 1JWQA, 1JYAA, 1JYHA, 1JYOA, 1K2XB, 1K3SA, 1K4NA, 1K6KA, 1K8KE, 1K8KG, 1K94A, 1KAFA, 1KGDA, 1KHCA, 1KHIA, 1KHYA, 1KMTA, 1KMVA, 1KNGA, 1KOE, 1KPFA, 1KPTA, 1KQRA, 1KSHB, 1KT6A, 1KTGA, 1KXOA, 1L3KA, 1L3PA, 1L6PA, 1L7LA, 1L8RA, 1LB6A, 1LF7A, 1LKIA, 1LKKA, 1LM8B, 1LMIA, 1LO7A, 1LQ9A, 1LQVA, 1LR5A, 1LU4A, 1LWBA, 1LXJA, 1LYQA, 1M1FA, 1M2DA, 1M48A, 1M4IA, 1M4JA, 1M55A, 1M70A, 1M7BA, 1M9ZA, 1MAIA, 1MBAA, 1MBMA, 1MC2A, 1MDCA, 1MF7A, 1MG4A, 1MGTA, 1MI8A, 1MK4A, 1MKAA, 1MWQA, 1MXIA, 1MY7A, 1N08A, 1N13B, 1N1FA, 1N62A, 1N71A, 1N8VA, 1NBCA, 1NBUB, 1NEPA, 1NFVA, 1NIGA, 1NJHA, 1NLQA, 1NQZA, 1NRVA, 1NRZA, 1NS5A, 1NTVA, 1NU0A, 1NWWA, 1NWZA, 1NXMA, 1NYCA, 1NZ0A, 1O13A, 1O4VA, 1O6DA, 1O7IA, 1O8XA, 1O9RA, 1OA8A, 1OBOA, 1OD3A, 1OGDA, 1OH0A, 1OH4A, 1OI0A, 1OO0A, 1OOHA, 1OQVA, 1ORUA, 1OSYA, 1OU8A, 1OW1A, 1OW4A, 1P0ZA, 1P57A, 1P5VB, 1P6OA, 1P90A, 1PBJA, 1PDOA, 1PK6A, 1PKOA, 1PL3A, 1PM4A, 1PMHX, 1POCA, 1PQ1A, 1PQHA, 1PVMA, 1PZ4A, 1Q1FA, 1Q4UA, 1Q5ZA, 1Q8DA, 1Q9UA, 1QDDA, 1QF8A, 1QFOA, 1QFTA, 1QGVA, 1QJ8A, 1QJPA, 1QKRA, 1QMYA, 1QU9A, 1QV1A, 1QVEA, 1QW2A, 1R0UA, 1R29A, 1R4VA, 1R75A, 1R7LA, 1R8SA, 1R9FA, 1R9WA, 1REGX, 1RG8A, 1RKIA, 1RLIA, 1RLJA, 1ROCA, 1ROWA, 1RSSA, 1RTTA, 1RW1A, 1RXDA, 1RXQA, 1RY9A, 1RYLA, 1S14A, 1S1QA, 1S3CA, 1S4KA, 1S5AA, 1S5UA, 1S7KA, 1S7ZA, 1SAUA, 1SBXA, 1SBZA, 1SEIA, 1SENA, 1SFPA, 1SH8A,

1SHUX, 1SJWA, 1SJYA, 1SKZA, 1SMBA, 1SQWA, 1SR4C, 1SRAA, 1STMA, 1SVPA, 1SVYA, 1SZHA, 1T1JA, 1T2WA, 1T3YA, 1T6UA, 1T82A, 1T92A, 1T9FA, 1T9IA, 1TC5A, 1TFEA, 1THQA, 1TJXA, 1TP6A, 1TQGA, 1TR0A, 1TT8A, 1TU9A, 1TUAA, 1TUHA, 1TVGA, 1TWUA, 1U14A, 1U2BA, 1U2WA, 1U53A, 1U5DA, 1U5XA, 1U7IA, 1U7KA, 1UCDA, 1UFYA, 1UGNA, 1UGXA, 1UJCA, 1UKFA, 1UNNC, 1UNQA, 1UPQA, 1UUYA, 1UUZA, 1UWFA, 1UWWA, 1UXZA, 1UZ3A, 1UZKA, 1V0AA, 1V2BA, 1V2XA, 1V2ZA, 1V30A, 1V4PA, 1V54E, 1V74A, 1V8CA, 1V8HA, 1V8YA, 1V96A, 1V9TA, 1V9YA, 1VCHA, 1VETA, 1VH5A, 1VI4A, 1VJLA, 1VKFA, 1VKIA, 1VKKA, 1VL7A, 1VLSA, 1VMBA, 1VMHA, 1VPMA, 1VQSA, 1VR7A, 1VSRA, 1VYIA, 1VYKA, 1VZIA, 1W0NA, 1W1HA, 1W4SA, 1WBIA, 1WC2A, 1WDJA, 1WHIA, 1WHSB, 1WJXA, 1WKAA, 1WKCA, 1WKOA, 1WL8A, 1WLUA, 1WN2A, 1WNAA, 1WNYA, 1WO8A, 1WOLA, 1WPNA, 1WPUA, 1WS8A, 1WVHA, 1WWIA, 1WWZA, 1X0TA, 1X3KA, 1X6OA, 1X6ZA, 1X82A, 1X8QA, 1X91A, 1XAUA, 1XBIA, 1XE0A, 1XEOA, 1XFSA, 1XG0C, 1XHNA, 1XIZA, 1XJUA, 1XKIA, 1XKPB, 1XKPC, 1XLQA, 1XODA, 1XQAA, 1XRKA, 1XS0A, 1XSVA, 1XT5A, 1XTEA, 1XVHA, 1XW3A, 1XWVA, 1XWWA, 1Y0BA, 1Y0HA, 1Y12A, 1Y43B, 1Y5HA, 1Y60A, 1Y71A, 1Y7RA, 1Y93A, 1Y9LA, 1YB0A, 1YB3A, 1YD9A, 1YFUA, 1YGHA, 1YGTA, 1YHFA, 1YKUA, 1YLMA, 1YLXA, 1YN9A, 1YNBA, 1YOAA, 1YOZA, 1YPQA, 1YPYA, 1YQ5A, 1YQHA, 1YSQA, 1YSRA, 1YTLA, 1YV4A, 1YW5A, 1YZ1A, 1YZFA, 1Z0JA, 1Z2UA, 1Z2WA, 1Z3EA, 1Z67A, 1Z6MA, 1Z6NA, 1Z9LA, 1Z9NA, 1ZBFA, 1ZC3B, 1ZCEA, 1ZD7A, 1ZE3H, 1ZHSA, 1ZHVA, 1ZK5A, 1ZKOA, 1ZLDA, 1ZMAA, 1ZPSA, 1ZTDA, 1ZUOA, 1ZX8A, 1ZZWA, 256BA, 2A0BA, 2A15A, 2A1IA, 2A2KA, 2A4VA, 2A67A, 2A7BA, 2A9IA, 2A9SA, 2AI4A, 2AJ6A, 2ANXA, 2AQ6A, 2AR5A, 2ASFA, 2ASKA, 2AU7A, 2AVKA, 2AZWA, 2B06A, 2B0AA, 2B0VA, 2B3GA, 2BBEA, 2BBRA, 2BCMA, 2BDRA, 2BEMA, 2BK9A, 2BKMA, 2BM5A, 2BNLA, 2BRFA, 2BRJA, 2BSJA, 2BWQA, 2BYCA, 2BZ1A, 2BZVA, 2C2IA, 2C2QA, 2C8SA, 2C92A, 2CARA, 2CAYA, 2CCMA, 2CE2X, 2CF7A, 2CHHA, 2CJ4A, 2CJTA, 2CKKA, 2CM4A, 2CO3A, 2CVDA, 2CVEA, 2CWZA, 2CX1A, 2CX7A, 2CXXA, 2CXYA, 2CY5A, 2CYJA, 2CZVC, 2D28C, 2D48A, 2D4PA, 2D5MA, 2DC4A, 2DJFA, 2DKOA, 2DM9A, 2DP9A, 2DPFA, 2DQAA, 2DQLA, 2DREA, 2DT4A, 2DTCA, 2DTJA, 2DVKA, 2DWKA,

2DXAA, 2DY0A, 2E2DC, 2E56A, 2E5YA, 2E7VA, 2E85A, 2EBEA, 2ECUA, 2EGJA, 2EHGA, 2EHPA, 2EJ8A, 2EJNA, 2EJXA, 2ENDA, 2EO4A, 2ERBA, 2ERVA, 2ES9A, 2F01A, 2F22A, 2F46A, 2F4WA, 2F5GA, 2F62A, 2F6EA, 2F9HA, 2FA1A, 2FB6A, 2FCJA, 2FCKA, 2FCLA, 2FD4A, 2FD5A, 2FDXA, 2FEXA, 2FG1A, 2FHQA, 2FHZA, 2FI1A, 2FIPA, 2FJ8A, 2FKBA, 2FLHA, 2FOJA, 2FP1A, 2FP7B, 2FPHX, 2FQ4A, 2FR5A, 2FRGP, 2FSRA, 2FSXA, 2FTBA, 2FTRA, 2FUFA, 2FUJA, 2FULA, 2FWHA, 2FWTA, 2FYGA, 2FYQA, 2FZSA, 2G1UA, 2G2CA, 2G3RA, 2G3WA, 2G5RA, 2G64A, 2G7BA, 2G7IA, 2G7SA, 2GDMA, 2GF6A, 2GHYA, 2GIAB, 2GIYA, 2GKGA, 2GKPA, 2GLZA, 2GMYA, 2GRCA, 2GRRB, 2GS5A, 2GU3A, 2GU9A, 2GUDA, 2GUIA, 2GUKA, 2GX5A, 2GYQA, 2GZBA, 2H1CA, 2H29A, 2H2BA, 2H30A, 2H5CA, 2H88D, 2H8EA, 2HA8A, 2HAZA, 2HBOA, 2HC8A, 2HD9A, 2HDVA, 2HEWF, 2HEYR, 2HFNA, 2HHZA, 2HIYA, 2HKVA, 2HL0A, 2HMHA, 2HNGA, 2HP7A, 2HQ4A, 2HQ7A, 2HQ9A, 2HQLA, 2HQSC, 2HQTA, 2HSBA, 2HU9A, 2HUHA, 2HVWA, 2HW2A, 2HX0A, 2HX5A, 2HY5A, 2HY5B, 2HY5C, 2HZQA, 2I02A, 2I2QA, 2I4AA, 2I6CA, 2I74A, 2I7AA, 2I7DA, 2I8GA, 2I8TA, 2I9CA, 2IA7A, 2IABA, 2IAYA, 2IB0A, 2IC2A, 2ICGA, 2IDLA, 2IF6A, 2IGIA, 2IGPA, 2IIHA, 2IJQA, 2IKBA, 2IM8A, 2IMGA, 2IMJA, 2IMLA, 2IMSA, 2IN0A, 2INWA, 2IQJA, 2IQYA, 2ISBA, 2IT9A, 2ITEA, 2IU5A, 2IWRA, 2IYVA, 2J12A, 2J1AA, 2J1VA, 2J2JA, 2J6AA, 2J6BA, 2J73A, 2J8WA, 2J9WA, 2JCQA, 2JDAA, 2JDCA, 2JDJA, 2JE3A, 2JEKA, 2JGBA, 2JJUA, 2JKGA, 2JLIA, 2LISA, 2MCMA, 2MHRA, 2MSBA, 2NL9A, 2NLVA, 2NMLA, 2NPTA, 2NR7A, 2NRKA, 2NRLA, 2NRRA, 2NSZA, 2NUHA, 2NUJA, 2NVHA, 2NW0A, 2NWFA, 2NX2A, 2NX4A, 2NYUA, 2O0QA, 2O1QA, 2O5FA, 2O5HA, 2O5UA, 2O6FA, 2O6LA, 2O6PA, 2O70A, 2O7AA, 2O90A, 2O95A, 2O99A, 2OA2A, 2OAFA, 2OB5A, 2OD0A, 2OD4A, 2ODAA, 2OEBA, 2OFCA, 2OFKA, 2OFZA, 2OGBA, 2OGFA, 2OHWA, 2OIKA, 2OIWA, 2OIXA, 2OIZD, 2OJ5A, 2OKFA, 2OKMA, 2OKUA, 2OLMA, 2OLPA, 2OMDA, 2OMLA, 2OMZB, 2OOCA, 2OOKA, 2OPCA, 2OQ8A, 2OQBA, 2ORWA, 2OS0A, 2OTMA, 2OU6A, 2OV0A, 2OVSA, 2OWPA, 2OX6A, 2OXGB, 2OYAA, 2OYOA, 2OZHA, 2OZJA, 2OZNA, 2P08A, 2P0DA, 2P0NA, 2P12A, 2P14A, 2P25A, 2P2OA, 2P39A, 2P45B, 2P58C, 2P65A, 2P75A, 2P8GA, 2P8IA, 2PA7A, 2PAGA, 2PD1A, 2PI2E, 2PIEA, 2PJSA, 2PKHA, 2PLNA, 2PMAA,

2PN0A, 2PNDA, 2PPNA, 2PQQA, 2PQVA, 2PR7A, 2PRVA, 2PRXA, 2PTHA, 2PTTA, 2PTTB, 2PU9A, 2PV2A, 2PVBA, 2PWWA, 2PYQA, 2Q03A, 2Q3TA, 2Q3WA, 2Q3XA, 2Q48A, 2Q4FA, 2Q4NA, 2Q5CA, 2Q5XA, 2Q82A, 2Q8OA, 2QACA, 2QEBA, 2QEUA, 2QF4A, 2QFEA, 2QH9A, 2QHKA, 2QHLA, 2QHQA, 2QIMA, 2QIPA, 2QIYA, 2QJWA, 2QJZA, 2QLWA, 2QMLA, 2QMMA, 2QNGA, 2QNTA, 2QQ4A, 2QQYA, 2QQZA, 2QS9A, 2QSAA, 2QTDA, 2QUOA, 2QUPA, 2QV8A, 2QVGA, 2QVKA, 2QYCA, 2QZQA, 2R0XA, 2R16A, 2R25A, 2R2AA, 2R2CA, 2R2YA, 2R47A, 2R4IA, 2R4QA, 2R5OA, 2R78A, 2RA6A, 2RA9A, 2RBBA, 2RBGA, 2RC3A, 2RDKA, 2RDMA, 2REMA, 2RFFA, 2RG8A, 2RGQA, 2RH3A, 2RJ2A, 2RK3A, 2RK9A, 2RKQA, 2RL8A, 2RLDA, 2SAKA, 2SCPA, 2SICI, 2TGIA, 2TNFA, 2UV4A, 2UVOA, 2V0UA, 2V1MA, 2V2PA, 2V4XA, 2V7FA, 2VB1A, 2VBUA, 2VEBA, 2VGXA, 2VH3A, 2VJWA, 2VKJA, 2VKPA, 2VLQB, 2VMHA, 2VN6A, 2VNGA, 2VOKA, 2VQ4A, 2VQ8A, 2VSMB, 2VUVA, 2VVPA, 2VVWA, 2VXGA, 2VXTI, 2VY8A, 2VZCA, 2VZPA, 2W0GA, 2W0IA, 2W1RA, 2W1SA, 2W31A, 2W3GA, 2W3XA, 2W6KA, 2W72B, 2W9HA, 2W9YA, 2WCJA, 2WCWA, 2WEWA, 2WFIA, 2WFOA, 2WFWA, 2WKBA, 2WLVA, 2YQYA, 2YSKA, 2YVAA, 2YVQA, 2YWIA, 2YWNA, 2YWWA, 2YXBA, 2YYOA, 2YZJA, 2YZKA, 2Z08A, 2Z0BA, 2Z0TA, 2Z0XA, 2Z10A, 2Z14A, 2Z3HA, 2Z3QA, 2Z5BA, 2Z5BB, 2Z5EA, 2Z5WA, 2Z6FA, 2Z6OA, 2Z8AA, 2ZAYA, 2ZCAA, 2ZDBA, 2ZDPA, 2ZEXA, 2ZF9A, 2Zfdb, 2ZGLA, 2ZHNA, 2ZIBA, 2ZK9X, 2ZNRA, 2ZOUA, 2ZQOA, 2ZS0B, 2ZS0D, 2ZU1A, 2ZVYA, 2ZX2A, 2ZZDA, 3A07A, 3A0YA, 3A35A, 3A4CA, 3A6SA, 3B33A, 3B44A, 3B64A, 3B6IA, 3B79A, 3B7CA, 3B8FA, 3BA3A, 3BB9A, 3BBBA, 3BCWA, 3BCYA, 3BDVA, 3BEDA, 3BFPA, 3BFQG, 3BJKA, 3BJNA, 3BKRA, 3BLNA, 3BM1A, 3BODA, 3BQAA, 3BQXA, 3BRCA, 3BS1A, 3BS2A, 3BT5A, 3BUTA, 3BVFA, 3BWUD, 3BWUF, 3BWVA, 3BWZA, 3BY8A, 3BYQA, 3C1QA, 3C5KA, 3C71A, 3C7MA, 3C7XA, 3C85A, 3C8IA, 3C8LA, 3C9PA, 3CANA, 3CBNA, 3CCGA, 3CEXA, 3CG4A, 3CG6A, 3CGIA, 3CHBD, 3CHMA, 3CI6A, 3CITA, 3CK1A, 3CM3A, 3CNBA, 3CNEA, 3CNUA, 3CNVA, 3CP3A, 3CP5A, 3CPTB, 3CQBA, 3CQNA, 3CRNA, 3CRYA, 3CT5A, 3CT6A, 3CTGA, 3CYPB, 3CZ1A, 3CZ6A, 3CZXA, 3D0JA, 3D1BA, 3D1MA, 3D1PA, 3D32A, 3D3BA, 3D3MA, 3D5PA, 3D6RA, 3D79A, 3D7AA, 3D7JA, 3D85C, 3D9NA, 3DAIA, 3DB7A, 3DCMX, 3DD7A, 3DDCB,

3DF8A, 3DFGA, 3DJ9A, 3DJHA, 3DLQI, 3DM8A, 3DMNA, 3DN7A, 3DO8A, 3DOUA, 3DPJA, 3DQGA, 3DQYA, 3DR6A, 3DR9A, 3DRZA, 3DSZA, 3DT5A, 3DXEA, 3E0HA, 3E0ZA, 3E11A, 3E1EA, 3E4GA, 3E4VA, 3E6SA, 3E7HA, 3E8MA, 3E8OA, 3E99A, 3E9VA, 3EBQA, 3EBTA, 3EBYA, 3EC6A, 3ECFA, 3EDOA, 3EEAA, 3EEHA, 3EF4A, 3EFYA, 3EJGA, 3EJKA, 3EK3A, 3ELKA, 3ELLA, 3ELSA, 3EN8A, 3ENUA, 3EO6A, 3EOIA, 3EPVA, 3ER7A, 3ES4A, 3ESMA, 3EULA, 3EURA, 3EWOA, 3EXNA, 3EYEA, 3EYTA, 3EZIA, 3F0DA, 3F14A, 3F1PB, 3F2VA, 3F2ZA, 3F40A, 3F43A, 3F4AA, 3F4MA, 3F59A, 3F5BA, 3F5OA, 3F62A, 3F6CA, 3F6GA, 3F7EA, 3F7LA, 3F7XA, 3F8BA, 3F8XA, 3F9SA, 3FBUA, 3FCDA, 3FDQA, 3FDXA, 3FF9A, 3FFVA, 3FFYA, 3FG8A, 3FG9A, 3FGYA, 3FH1A, 3FH2A, 3FHVA, 3FIQA, 3FIVA, 3FJSA, 3FKAA, 3FKCA, 3FKEA, 3FL2A, 3FM2A, 3FN5A, 3FNCA, 3FOVA, 3FPNA, 3FPWA, 3FQMA, 3FSAA, 3FSOA, 3FW2A, 3FZ4A, 3FZEA, 3G0MA, 3G14A, 3G16A, 3G48A, 3G7GA, 3G98A, 3GA3A, 3GA4A, 3GBWA, 3GBYA, 3GDWA, 3GE3E, 3GFPA, 3GGNA, 3GGQA, 3GGYA, 3GHJA, 3GHXA, 3GKNA, 3GLVA, 3GMGA, 3GMXA, 3GPGA, 3GPKA, 3GQQA, 3GRDA, 3GWNA, 3GXBA, 3GXHA, 3GXSA, 3GY9A, 3GYKA, 3GZBA, 3GZRA, 3H05A, 3H0NA, 3H51A, 3H6QA, 3H6RA, 3H79A, 3H87A, 3H8TA, 3H9WA, 3HCJA, 3HH1A, 3HHVA, 3HIUA, 3HM4A, 3HMZA, 3HOIA, 3HP4A, 3HQXA, 3HSAA, 3HT1A, 3HTNA, 3HUHA, 3HV2A, 3HVWA, 3HWUA, 3HX8A, 3HXIA, 3HYNA, 3HZPA, 3I24A, 3I2VA, 3I3FA, 3I7MA, 3IA1A, 3IBMA, 3IBZA, 3IDFA, 3IDUA, 3IE4A, 3IGNA, 3IHTA, 3IJMA, 3IKBA, 3IMKA, 3INOA, 3IP0A, 3IQ1A, 3IQTA, 3IRBA, 3IU5A, 3IU6A, 3IUOA, 3IV4A, 3IVVA, 3IWTA, 3IX3A, 3IXCA, 3JRNA, 3JRVA, 3JSRA, 3JTWA, 3JUIA, 3JVL, 3JX9A, 3JXSA, 3K0ZA, 3K67A, 3K7CA, 3K9WA, 3KDHA, 3KE7A, 3NULA, 3TSSA, 3VUBA, 4UBPA

A.1.3 Depth binding propensity data-set

The following protein structures are used to generate the propensity of small-molecule ligand binding on depth-accessibility map (c.f. Section 4.2.4).

102M, 104M, 106M, 109M, 110M, 111M, 121P, 148L, 154L, 181L, 182L, 183L, 184L, 186L, 187L, 188L, 1A5V, 1A5W, 1A5X, 1A6G, 1A6K, 1A6M, 1A6N, 1A85,

1ABS, 1AJG, 1AJH, 1ALU, 1AQB, 1ASW, 1AZI, 1B1C, 1B2V, 1B8Y, 1BJE, 1BVC, 1BVD, 1BVV, 1BZ6, 1BZP, 1BZR, 1BZS, 1C5I, 1C8S, 1CAQ, 1CH1, 1CH2, 1CH3, 1CH7, 1CH9, 1CIO, 1CIZ, 1CJW, 1CO8, 1CO9, 1CP0, 1CQQ, 1CTQ, 1CV8, 1CXQ, 1CZH, 1CZK, 1CZL, 1CZN, 1CZO, 1CZR, 1CZS, 1CZU, 1D03, 1D04, 1D2S, 1D2U, 1D3S, 1D4O, 1DF7, 1DG5, 1DG7, 1DG8, 1DG9, 1DLY, 1DNL, 1DO1, 1DO4, 1DO7, 1DRF, 1DTI, 1DUO, 1DWR, 1DWS, 1DWT, 1DXC, 1DXD, 1DY3, 1E3B, 1EIZ, 1EJ0, 1EJ2, 1EK0, 1EMY, 1EPZ, 1EQD, 1EQM, 1ERB, 1ERX, 1EUO, 1EX7, 1EX8, 1F63, 1F9H, 1FCS, 1FEL, 1FEM, 1FEN, 1FJJ, 1FLV, 1FM4, 1FRW, 1FVG, 1FW9, 1FZQ, 1G1T, 1G5T, 1G81, 1GDI, 1GDJ, 1GDK, 1GDL, 1GJN, 1GKY, 1GNR, 1GNY, 1GR3, 1GUI, 1GVC, 1GWL, 1GWM, 1H0A, 1H0P, 1H1X, 1HBP, 1HD2, 1HFC, 1HJT, 1HQ2, 1HRM, 1HSY, 1HUQ, 1HYB, 1HYV, 1I05, 1I06, 1I6T, 1I76, 1I82, 1I8A, 1ID0, 1IKE, 1IKJ, 1IOP, 1IOZ, 1IQQ, 1J1F, 1J1G, 1J3F, 1J52, 1J53, 1J8R, 1JAH, 1JAI, 1JDO, 1JF0, 1JF2, 1JJ9, 1JK3, 1JOM, 1JP9, 1JPB, 1JRL, 1JUV, 1JV4, 1K12, 1KAO, 1KDK, 1KMQ, 1KMV, 1KOI, 1KQR, 1KT3, 1KT4, 1KT5, 1KT7, 1KY2, 1KY3, 1L1Q, 1L1R, 1L4U, 1L4Y, 1L83, 1L84, 1L8J, 1LF0, 1LF5, 1LGW, 1LGX, 1LH1, 1LH2, 1LH3, 1LH5, 1LH6, 1LH7, 1LHN, 1LHO, 1LHS, 1LHT, 1LHU, 1LHV, 1LHW, 1LI3, 1LI6, 1LKE, 1LQY, 1LTW, 1LUE, 1M7B, 1M8G, 1MBC, 1MBD, 1MBI, 1MBN, 1MBO, 1MCY, 1MGN, 1MH1, 1ML7, 1MLF, 1MLG, 1MLH, 1MLJ, 1MLL, 1MLN, 1MLQ, 1MLR, 1MLS, 1MLU, 1MMQ, 1MOA, 1MOC, 1MOD, 1MR3, 1MTI, 1MTJ, 1MTK, 1MVS, 1MYM, 1MYZ, 1MZ0, 1N6H, 1N6I, 1N6K, 1N6L, 1N6N, 1N6O, 1N6P, 1N6R, 1N9H, 1N9I, 1N9X, 1NDC, 1NHB, 1NP4, 1NPF, 1NPG, 1NZ2, 1NZ3, 1NZ4, 1NZ5, 1O16, 1O1X, 1O8S, 1O9V, 1O9W, 1OD6, 1OFJ, 1OFK, 1OFV, 1OH4, 1OIX, 1OV7, 1OVH, 1OVJ, 1OWY, 1OWZ, 1P2T, 1PEE, 1PMH, 1Q0N, 1Q1F, 1Q1Y, 1Q91, 1Q92, 1QF9, 1QJP, 1QRA, 1QSR, 1QUD, 1QUH, 1QV0, 1QV1, 1QY1, 1QY2, 1R2Q, 1RA1, 1RA3, 1RA8, 1RA9, 1RAO, 1RB0, 1RBP, 1RCD, 1RCF, 1RCI, 1RF7, 1RG7, 1RM8, 1RMZ, 1RN8, 1RNJ, 1RU2, 1RVD, 1RX1, 1RX2, 1RX6, 1RX9, 1S36, 1S3V, 1S3W, 1S7F, 1SEH, 1SGC, 1SIX, 1SK7, 1SL4, 1SL9, 1SPE, 1SPV, 1STA, 1STB, 1STY, 1SVI, 1SXU, 1SXW, 1SXX, 1SXY, 1SY0, 1SY1, 1SY2, 1SY3, 1SYB, 1SYL, 1SZ7, 1T68, 1TAL, 1TES, 1THQ, 1U0X, 1U72, 1U7R, 1U7S, 1UCA, 1UCC, 1UCD, 1UFJ, 1URM, 1USN, 1UUX, 1UUY, 1UWF, 1V2G, 1V2X, 1V4B,

1V8M, 1V8R, 1V8V, 1V8Y, 1V9H, 1V9Q, 1VG1, 1VK5, 1VSD, 1VSH, 1VXA, 1VXB, 1VXC, 1VXD, 1VXF, 1VXG, 1VXH, 1W1D, 1W1G, 1W8L, 1W8M, 1W92, 1WC2, 1WC8, 1WC9, 1WLA, 1WLJ, 1WNZ, 1WOZ, 1WRM, 1WS1, 1WUB, 1WVP, 1X1R, 1X3K, 1X3S, 1X46, 1X6O, 1X8O, 1X8P, 1X8Q, 1XCM, 1XEP, 1XJ0, 1XLR, 1XTQ, 1XXN, 1XZ1, 1XZ3, 1Y63, 1Y93, 1Y9Q, 1YCL, 1YFW, 1YFX, 1YMA, 1YMC, 1YOA, 1YOG, 1YOH, 1YOI, 1YP6, 1YVD, 1YWA, 1YWB, 1YWC, 1YWM, 1YZG, 1YZQ, 1Z06, 1Z07, 1Z1S, 1Z2A, 1Z4I, 1Z4J, 1Z4K, 1Z4L, 1Z4M, 1Z4P, 1Z4R, 1ZCN, 1ZD9, 1ZJ6, 1ZND, 1ZNE, 1ZNG, 1ZNK, 1ZNL, 1ZP5, 1ZV9, 1ZVX, 1ZW6, 220L, 225L, 227L, 2A3F, 2A5J, 2ACP, 2AF9, 2AH7, 2AL0, 2AL7, 2ALL, 2AMM, 2ASN, 2AT0, 2AT3, 2AT6, 2AT8, 2ATE, 2ATV, 2ATZ, 2AU9, 2AUU, 2B45, 2B58, 2BFQ, 2BLI, 2BLJ, 2BMD, 2BSC, 2BW9, 2BWH, 2C8S, 2CE2, 2CIG, 2CL0, 2CL6, 2CL7, 2CLC, 2CLD, 2CM4, 2CMM, 2CU9, 2CUT, 2CVE, 2CVJ, 2CW9, 2CY2, 2D4N, 2D5M, 2D6O, 2DFN, 2DM5, 2E2Y, 2E5X, 2EB8, 2EB9, 2EF2, 2EKT, 2EKU, 2ESB, 2EU7, 2EVK, 2EVW, 2F21, 2F32, 2F8P, 2F9L, 2FCR, 2FFQ, 2FN4, 2FQO, 2FQT, 2FRF, 2FRI, 2FRJ, 2FRK, 2FSR, 2FUP, 2FVV, 2FZJ, 2G0R, 2G0V, 2G0X, 2G0Z, 2G10, 2G11, 2G12, 2G14, 2G1K, 2G1U, 2G3H, 2GDM, 2GTF, 2GUI, 2GWL, 2HBO, 2HLV, 2HPS, 2HR6, 2HRM, 2HU6, 2HX5, 2HXS, 2HYS, 2HZQ, 2I6C, 2ICA, 2IEK, 2IT6, 2IYQ, 2IYU, 2IYV, 2IYW, 2IYX, 2IYY, 2J1A, 2JAO, 2JAR, 2JCQ, 2JCR, 2JE2, 2JE3, 2JHO, 2LH1, 2LH2, 2LH3, 2LH5, 2LH6, 2LH7, 2MBW, 2MGB, 2MGC, 2MGD, 2MGE, 2MGF, 2MGG, 2MGH, 2MGK, 2MGL, 2MGM, 2MYA, 2MYB, 2MYC, 2MYD, 2MYE, 2NSH, 2NSL, 2NSR, 2NSS, 2O58, 2O5B, 2O5L, 2O5M, 2O5O, 2O5Q, 2O5S, 2O5T, 2O7N, 2OFI, 2OFR, 2OH3, 2OH8, 2OH9, 2OHA, 2OHB, 2OS1, 2OT9, 2OTY, 2OU0, 2OVD, 2OZY, 2P0B, 2P39, 2P3I, 2P3J, 2P3K, 2PRB, 2PTM, 2PY4, 2Q9P, 2Q9R, 2QCB, 2QG6, 2QHK, 2QYQ, 2R6V, 2RAY, 2RAZ, 2RB0, 2RB1, 2RB2, 2RBN, 2RBO, 2RBP, 2RBQ, 2RBR, 2RBS, 2RDG, 2RGA, 2RGB, 2RGC, 2RGE, 2SPL, 2SPN, 2ULL, 2UV4, 2UV5, 2UV6, 2UV7, 2V1E, 2V1F, 2V1G, 2V1I, 2V1J, 2V1K, 2V5T, 2VEB, 2VEZ, 2VLX, 2VLY, 2VLZ, 2VM0, 2VMG, 2VRY, 2VXK, 2W2T, 2W3W, 2W5Y, 2W6W, 2W6X, 2W6Y, 2W9G, 2W9H, 2WEW, 2WEX, 2WFO, 2WGR, 2X1V, 2X1Z, 2X20, 2X21, 2YVO, 2YVP, 2YZY, 2Z0X, 2Z6S, 2Z6T, 2Z98, 2ZND, 2ZSN, 2ZSO, 2ZSP, 2ZSQ, 2ZSR, 2ZSS, 2ZST, 2ZSX, 2ZSY, 2ZSZ, 2ZT0,

2ZT1, 2ZT2, 2ZT3, 2ZT4, 2ZU3, 2ZUR, 2ZWJ, 3A0T, 3A2G, 3A5A, 3A5B, 3A9M, 3ACD, 3B45, 3BA2, 3C76, 3C77, 3C78, 3CI1, 3CI3, 3CI4, 3CQV, 3CSZ, 3CT1, 3CT5, 3CU3, 3CZ9, 3CZA, 3D80, 3D84, 3DAU, 3DCM, 3DFR, 3DKE, 3DMX, 3DMZ, 3DN0, 3DN1, 3DN2, 3DN3, 3DN4, 3DN6, 3DN8, 3DNA, 3DPE, 3E3U, 3E3Z, 3E4N, 3E55, 3E5O, 3ECL, 3ECX, 3ECZ, 3ED9, 3EDA, 3EDB, 3EHX, 3EIG, 3EJ0, 3ELN, 3EMM, 3EXN, 3F0U, 3F15, 3F16, 3F17, 3F18, 3F19, 3F1A, 3F34, 3F35, 3F36, 3F37, 3F38, 3F39, 3F5B, 3F64, 3F6J, 3F8K, 3F8Y, 3F91, 3FLL, 3FQ0, 3FQF, 3FQV, 3FQZ, 3FRB, 3FS6, 3FZG, 3G7P, 3GA4, 3GAH, 3GAI, 3GAJ, 3GAN, 3GHV, 3GHW, 3GI2, 3GK6, 3GK9, 3GKT, 3GMV, 3GUI, 3GVF, 3GY9, 3GYA, 3GYF, 3H4O, 3H57, 3H58, 3H6D, 3HC9, 3HEN, 3HEO, 3HEP, 3HH3, 3HH4, 3HH5, 3HH6, 3HNB, 3HT6, 3HT7, 3HT8, 3HTB, 3HTD, 3HTF, 3HTG, 3HU8, 3HUA, 3HUK, 3HUQ, 3HYF, 3HYN, 3HZA, 3I3S, 3IGN, 3IIS, 3IIU, 3IMK, 3IP0, 3JUC, 3K21, 3K45, 3K8Y, 3K9N, 3K9Z, 3KAD, 3KAG, 3KAI, 3KB0, 3KLB, 3KQ0, 3L1N, 3LBH, 3LBN, 3LX3, 3RAB, 3UKD, 4MBN, 4Q21, 4UKD, 5MBN, 5P21, 5UKD, 721P, 821P, 8DFR, 966C

A.1.4 High-resolution single domain data set

The following protein structures are used in the parametrization of Hydrogen bond geometry (donor-acceptor distance and $\hat{H}$ angle) (c.f. Section 5.2.1).

1A6MA, 1AQZA, 1B0BA, 1B9OA, 1BQKA, 1C7KA, 1CCWA, 1CXQA, 1DK8A, 1DWKA, 1DY5A, 1E7LA, 1EAJA, 1EAQA, 1EB6A, 1EUWA, 1EXRA, 1F4PA, 1FJJA, 1FM0E, 1G8KB, 1GK8I, 1GNYA, 1GQVA, 1GU2A, 1GWMA, 1H4AX, 1H97A, 1HFES, 1I76A, 1J98A, 1J9BA, 1JBEA, 1JLTA, 1JLTB, 1K2XA, 1K2XB, 1KMTA, 1KNGA, 1KNMA, 1KQ6A, 1KQRA, 1L2HA, 1LO7A, 1MC2A, 1MQKL, 1N62A, 1NC7A, 1NKIA, 1NWWA, 1NWZA, 1OALA, 1OAQH, 1OAQL, 1OD3A, 1OH4A, 1OKOA, 1OOHA, 1OUWA, 1P0ZA, 1P28A, 1P6OA, 1PM1X, 1PQHA, 1PWAA, 1QD9A, 1QDDA, 1QJPA, 1QTNA, 1QVEA, 1R26A, 1R29A, 1R2QA, 1R8SA, 1RG8A, 1RM6C, 1S69A, 1SENA, 1SJWA, 1SMBA, 1SX7A, 1SXVA, 1T0AA, 1T3YA, 1TO4A, 1TT8A, 1TU9A, 1TYJA, 1U69A, 1U7IA, 1UFYA, 1UGWA, 1UNQA, 1UOWA, 1UUYA, 1UXZA, 1UY4A, 1VJFA, 1VKEA, 1VKKA, 1VL7A, 1VL9A,

1VMFA, 1VMHA, 1VMJA, 1VPMA, 1VR7A, 1W0NA, 1W9SA, 1WCUA, 1WDDS,
1WKQA, 1WLUA, 1WMSA, 1WMZA, 1WWIA, 1X6ZA, 1XBIA, 1XG0C, 1XJUA,
1XO7A, 1XPHA, 1XRKA, 1XSOA, 1XT5A, 1Y55X, 1Y93A, 1YN9A, 1YPHC,
1YPQA, 1Z2UA, 1Z6MA, 1ZKIA, 1ZKKA, 1ZLBA, 1ZV9A, 2A2KA, 2A3MA,
2A50B, 2AENA, 2AJ7A, 2ANXA, 2AQ6A, 2AQMA, 2AQPA, 2AU7A, 2AXWA,
2BEMA, 2BK9A, 2BKMA, 2BWKA, 2C4BA, 2C4WA, 2C6UA, 2CALA, 2CE2X,
2CHCA, 2CJ4A, 2CJTA, 2CKKA, 2CM5A, 2D37A, 2D4PA, 2DKOA, 2DT4A,
2DXEA, 2ERXA, 2F01A, 2F22A, 2F62A, 2F9SA, 2FCLA, 2FG1A, 2FLHA, 2FR5A,
2FUFA, 2FWHA, 2G1UA, 2G2SB, 2G7BA, 2GJ3A, 2GKGA, 2GLZA, 2GMYA,
2GUDA, 2GYQA, 2H2RA, 2H8EA, 2H8FA, 2H8FB, 2HEWF, 2HUHA, 2HW2A,
2HX5A, 2HXS, 2I3HA, 2I6CA, 2I8DA, 2I8TA, 2IA1A, 2IBLA, 2IMJA, 2IONA,
2J1SA, 2J23A, 2J8WA, 2JCQA, 2JDAA, 2JDCA, 2JJUA, 2JLIA, 2LISA, 2NRLA,
2NSZA, 2NWFA, 2O7AA, 2O8QA, 2O90A, 2OA2A, 2OFCA, 2OFZA, 2OH1A,
2OIZD, 2OKFA, 2OLMA, 2OSSA, 2OXGY, 2OZHA, 2P45B, 2P6OA, 2P7OA,
2P8GA, 2PBDP, 2PIEA, 2PR7A, 2PRVA, 2Q2HA, 2Q3TA, 2Q4NA, 2QCBA, 2QEUA,
2QF4A, 2QF9A, 2QIMA, 2QIPA, 2QJZA, 2QL8A, 2QLWA, 2QNLA, 2QNTA,
2QUDA, 2R0XA, 2R4IA, 2R6QA, 2R99A, 2RA9A, 2RBDA, 2RE2A, 2RFRA, 2RI7A,
2RKQA, 2UYKA, 2UYWA, 2UYZA, 2V0UA, 2V1MA, 2V4XA, 2V7FA, 2V9VA,
2VB1A, 2VIFA, 2VLQB, 2VMHA, 2VUVA, 2VVPA, 2VXTI, 2VY7A, 2VZCA,
2VZPA, 2W1WA, 2W3JA, 2W72A, 2W72B, 2W87A, 2WCWA, 2WFIA, 2WKKA,
2WLVA, 2WNXA, 2WQ4A, 2WRAA, 2WTPA, 2WVFA, 2WWEA, 2WYTA, 2WZ8A,
2WZ9A, 2X2SA, 2X32A, 2X46A, 2X52A, 2X5PA, 2X5YA, 2X7KA, 2XFGB, 2XG3A,
2XHFA, 2XOLA, 2XOMA, 2XR6A, 2XS4A, 2XSTA, 2Y0OA, 2Y1QA, 2Y71A,
2Y78A, 2Y8EA, 2Y8GA, 2Y9FA, 2YFUA, 2YH5A, 2YKZA, 2YZ1A, 2Z0XA, 2Z3HA,
2Z51A, 2Z6WA, 2ZEXA, 2ZFDB, 2ZHNA, 2ZIBA, 2ZNRA, 2ZS0A, 2ZS0B, 2ZS0C,
2ZS0D, 3A57A, 3A6RA, 3AP9A, 3AQJA, 3AWUB, 3B79A, 3BA3A, 3BEDA, 3BFQG,
3BLNA, 3BS2A, 3BT5A, 3C3MA, 3C8LA, 3CE1A, 3CHMA, 3CI6A, 3CIPG, 3CJEA,
3CM3A, 3CP5A, 3CT6A, 3CTGA, 3CXKA, 3D0JA, 3D1PA, 3D22A, 3D5PA, 3D9XA,
3DAUA, 3DB7A, 3DD7A, 3E10A, 3E4GA, 3E8MA, 3EBTA, 3EC9A, 3EDOA,
3EERA, 3EF4A, 3EF8A, 3EJVA, 3EURA, 3EVOA, 3EY6A, 3EY8A, 3F1PB, 3F2ZA,
3F37A, 3F6QA, 3F6VA, 3F7EA, 3F7LA, 3F7XA, 3F8XA, 3FAJA, 3FCNA, 3FIAA,

3FIQA, 3FK8A, 3FQ4A, 3FSAA, 3FYMA, 3FZWA, 3G0KA, 3G16A, 3G39A, 3G46A, 3G9AB, 3GA4A, 3GAXA, 3GBWA, 3GHJA, 3GIXA, 3GKMA, 3GMXB, 3GPGA, 3GRDA, 3GXHA, 3GY9A, 3GZBA, 3GZRA, 3H08A, 3H5JA, 3H6QA, 3H87A, 3HM4A, 3HT1A, 3HTNA, 3HUPA, 3HV2A, 3HWUA, 3HX8A, 3HZAA, 3HZPA, 3I3FA, 3I7MA, 3IA4A, 3IISM, 3IJMA, 3IMKA, 3IP0A, 3IQ1A, 3IQ2A, 3IQTA, 3ITFA, 3IVVA, 3JRVA, 3JUDA, 3JVL A, 3K67A, 3KA8A, 3KBEA, 3KE7A, 3KFFA, 3KHQA, 3KKGA, 3KLRA, 3KUUA, 3KWKA, 3KWUA, 3KYZA, 3L0FA, 3L0FB, 3L4RA, 3L51A, 3L51B, 3L8HA, 3L91A, 3LASA, 3LB2A, 3LBEA, 3LHNA, 3LJDA, 3LJWA, 3LLOA, 3LQWA, 3LWZA, 3LX3A, 3LYDA, 3LYGA, 3LZLA, 3M1HA, 3M1OA, 3M1XA, 3M2OA, 3M7AA, 3MC3A, 3MDXA, 3ML1B, 3MMHA, 3MNGA, 3MPZA, 3MR0A, 3N08A, 3N1FA, 3N4IB, 3N4JA, 3N5AA, 3NBCA, 3NBKA, 3NECA, 3NJKA, 3NKVA, 3NOHA, 3O1CA, 3O2RA, 3O5QA, 3OBLA, 3OD9A, 3OGNA, 3OHEA, 3OJ0A, 3OJ6A, 3OMDA, 3ON9A, 3ONHA, 3OR5A, 3OSEA, 3OSTA, 3OXPA, 3P1GA, 3P3VA, 3P48A, 3P6DA, 3P6IA, 3PG6A, 3PIWA, 3PJYA, 3PNAA, 3PP2A, 3PWFA, 3Q62A, 3Q64A, 3QDSA, 3QHBA, 3QL9A, 3QOOA, 3QU3A, 3QZBA, 3QZMA, 3R0LD, 3R0NA, 3R2RA, 3R3QA, 3RDOA, 3ROBA, 3ROFA, 3RRIA, 3RWNA, 3S4KA, 3S6FA, 3S6MA, 3S8IA, 3S9XA, 3SB1A, 3SK2A, 3SM1A, 3SXHA, 3T3LA, 3TQ5A, 3ZW5A, 3ZXHA, 4A2VA, 4UBPB

A.1.5 Small organic molecules in Open Crystallography Database

The following protein structures are used to study the hydrogen bond geometry in small organic molecules (c.f. Section 5.4.1).

1₀0₀1₀000001, 1₀0₀1₀000001, 1₁0₀1₁100181, 1₁0₀2₁100269, 1₁0₀2₁100296, 1₁0₀3₁100369, 1₁0₀3₁10037

1AOPA, 1ATGA, 1AUNA, 1AXNA, 1AYEA, 1B0BA, 1B1CA, 1B2VA, 1B5LA, 1B6AA, 1B6GA, 1B7VA, 1B8OA, 1B9OA, 1BF2A, 1BHEA, 1BJ7A, 1BKRA, 1BQBA, 1BQCA, 1BS0A, 1BS1A, 1BU8A, 1BUEA, 1BX4A, 1BX7A, 1BXEA, 1BXOA, 1BXVA, 1BY2A, 1BZ4A, 1C3DA, 1C3PA, 1C44A, 1C52A, 1C7KA, 1CC8A, 1CCZA, 1CEOA, 1CEXA, 1CHDA, 1CI3M, 1CIZA, 1CJCA, 1CJWA, 1CQYA, 1CRBA, 1CRZA, 1CS8A, 1CT5A, 1CTJA, 1CUOA, 1CVZA, 1CXYA, 1CY5A, 1CZNA, 1CZTA, 1D0BA, 1D2KA, 1D3GA, 1D4OA, 1D5TA, 1D7LA, 1DBIA, 1DC9A, 1DL2A, 1DLWA, 1DLYA, 1DM1A, 1DMGA, 1DMTA, 1DQ3A, 1DQGA, 1DQYA, 1DS1A, 1DUSA, 1DUWA, 1DVOA, 1DXJA, 1DY2A, 1DYMA, 1DYPA, 1DYQA, 1E0CA, 1E25A, 1E29A, 1E43A, 1E4FT, 1E5KA, 1E9MA, 1E9WA, 1EA7A, 1EANA, 1EAZA, 1EB6A, 1EDGA, 1EDQA, 1EFDN, 1EG2A, 1EH3A, 1EH6A, 1EJ0A, 1EJ2A, 1EJGA, 1EK0A, 1ELJA, 1ELTA, 1ELVA, 1EN2A, 1ENFA, 1EOEA, 1EOKA, 1EPWA, 1EQFA, 1ES5A, 1ES6A, 1ESWA, 1ET9A, 1EU1A, 1EU8A, 1EW4A, 1EWFA, 1EXFA, 1EXRA, 1EYHA, 1F0IA, 1F20A, 1F2FA, 1F4PA, 1F7SA, 1F94A, 1F9VA, 1FAOA, 1FAZA, 1FBNA, 1FC6A, 1FCQA, 1FCYA, 1FDRA, 1FGYA, 1FK5A, 1FKMA, 1FL0A, 1FL2A, 1FLJA, 1FM4A, 1FNLA, 1FO3A, 1FO8A, 1FQIA, 1FT5A, 1FVGA, 1FW4A, 1FY7A, 1FYEA, 1FZQA, 1G0CA, 1G12A, 1G1TA, 1G2BA, 1G2NA, 1G2RA, 1G33A, 1G3PA, 1G3QA, 1G55A, 1G5AA, 1G66A, 1G6AA, 1G6XA, 1G8AA, 1G8PA, 1G94A, 1G9GA, 1G9KA, 1GA8A, 1GAIA, 1GAKA, 1GBGA, 1GCIA, 1GHMA, 1GIRA, 1GISA, 1GK7A, 1GMXA, 1GNTA, 1GNUA, 1GNYA, 1GP0A, 1GP6A, 1GPIA, 1GPPA, 1GQ8A, 1GQEA, 1GQVA, 1GS9A, 1GSMA, 1GTKA, 1GUIA, 1GUUA, 1GV2A, 1GV8A, 1GV9A, 1GVDA, 1GVZA, 1GWMA, 1GWUA, 1GXNA, 1GYDB, 1GYUA, 1GZ2A, 1GZ8A, 1H12A, 1H1HA, 1H2EA, 1H4AX, 1H6HA, 1H6TA, 1H70A, 1H75A, 1H76A, 1H98A, 1HBKA, 1HCZA, 1HD2A, 1HDKA, 1HDOA, 1HFUA, 1HG5A, 1HG8A, 1HH5A, 1HH8A, 1HJ8A, 1HKKA, 1HN0A, 1HP1A, 1HQ0A, 1HT6A, 1HVXA, 1HX0A, 1HXIA, 1HZTA, 1I0VA, 1I1NA, 1I1WA, 1I27A, 1I2AA, 1I2HA, 1I2TA, 1I5GA, 1I5PA, 1I60A, 1I71A, 1I76A, 1I8AA, 1I8OA, 1I9GA, 1I9SA, 1I9ZA, 1IA5A, 1IA6A, 1ICXA, 1ID0A, 1IDKA, 1IEJA, 1IIRA, 1IJBA, 1IJTA, 1IKOP, 1IKPA, 1IKTA, 1IM5A, 1IMJA, 1IO0A, 1IO3A, 1IOWA, 1IPCA, 1IQQA, 1IQVA, 1IQZA, 1ISPA, 1ITXA, 1IUQA, 1IUZA, 1IWDA, 1IWNA, 1IXHA, 1IXKA, 1IXLA, 1IXZA, 1IYNA, 1IZDA, 1J02A, 1J09A, 1J0PA, 1J1LA,

1J1TA, 1J27A, 1J2AA, 1J2LA, 1J3AA, 1J5XA, 1J6OA, 1J74A, 1J7KA, 1J7XA, 1J8EA, 1J8RA, 1J9BA, 1JAEA, 1JAKA, 1JB3A, 1JBEA, 1JBKA, 1JBWA, 1JC9A, 1JCFA, 1JDLA, 1JEOA, 1JF8A, 1JFBA, 1JFXA, 1JG1A, 1JHJA, 1JHSA, 1JIXA, 1JJFA, 1JJVA, 1JK7A, 1JL7A, 1JLNA, 1JM1A, 1JMLA, 1JNDA, 1JNIA, 1JO8A, 1JOIA, 1JP4A, 1JRLA, 1JT1A, 1JTAA, 1JU3A, 1JUGA, 1JVWA, 1JWQA, 1JX6A, 1JY1A, 1JYHA, 1JYKA, 1K12A, 1K1BA, 1K2I1, 1K3IA, 1K4NA, 1K5CA, 1K77A, 1K7CA, 1K7IA, 1K7JA, 1K7KA, 1K87A, 1KA1A, 1KB0A, 1KCQA, 1KEAA, 1KFRA, 1KFWA, 1KG2A, 1KGDA, 1KGSa, 1KHCA, 1KHIA, 1KIDA, 1KJWA, 1KK1A, 1KL9A, 1KLXA, 1KM8A, 1KMOA, 1KMQA, 1KMVA, 1KNGA, 1KNMA, 1KOEa, 1KP6A, 1KQ3A, 1KQ6A, 1KQRA, 1KQUA, 1KS5A, 1KS8A, 1KT6A, 1KT9A, 1KTHA, 1KUFA, 1KUXA, 1KV7A, 1KV9A, 1KW4A, 1KWFA, 1KY3A, 1KYHA, 1KZFA, 1KZLA, 1L2HA, 1L3KA, 1L6PA, 1L8JA, 1L9LA, 1LC0A, 1LF7A, 1LFKA, 1LG7A, 1LJ5A, 1LJ8A, 1LKEA, 1LLNA, 1LMHA, 1LMLA, 1LN4A, 1LO6A, 1LOUA, 1LQXA, 1LQYA, 1LR0A, 1LR7A, 1LRIA, 1LRZA, 1LS1A, 1LS6A, 1LS9A, 1LTZA, 1LU4A, 1LUFA, 1LV7A, 1LVGA, 1LY2A, 1LYVA, 1LZJA, 1LZLA, 1M13A, 1M15A, 1M1QA, 1M2KA, 1M33A, 1M40A, 1M47A, 1M4LA, 1M6IA, 1M6TA, 1M7BA, 1M9ZA, 1MB3A, 1MC2A, 1MCXA, 1MD6A, 1MDXA, 1MEOA, 1MF7A, 1MG4A, 1MGPA, 1MGTA, 1MH1A, 1MHNA, 1MIDA, 1MIXA, 1MJ0A, 1MJ4A, 1MJ5A, 1MJNA, 1MJSA, 1MK0A, 1MKYA, 1MLWA, 1MP8A, 1MR3F, 1MS6A, 1MTZA, 1MUWA, 1MW9X, 1MWPA, 1MXGA, 1MZ4A, 1MZSA, 1N0RA, 1N1FA, 1N3YA, 1N40A, 1N4KA, 1N4WA, 1N6AA, 1N7OA, 1N83A, 1N8IA, 1N8UA, 1N9BA, 1N9LA, 1NB9A, 1NC5A, 1NEPA, 1NG2A, 1NG6A, 1NKGA, 1NM2A, 1NM8A, 1NMLA, 1NMYA, 1NNFA, 1NNHA, 1NOTA, 1NP1A, 1NPSA, 1NPUA, 1NQ6A, 1NR0A, 1NRWA, 1NTYA, 1NUYA, 1NWAA, 1NWZA, 1NXCA, 1NZJA, 1O06A, 1O13A, 1O1YA, 1O1ZA, 1O20A, 1O4YA, 1O5ZA, 1O88A, 1O8PA, 1O8VA, 1O8XA, 1O98A, 1O9GA, 1OALA, 1OAPA, 1OB4A, 1OB7A, 1OC7A, 1OCSA, 1OD3A, 1ODMA, 1ODOA, 1OEEA, 1OEWA, 1OFLA, 1OG1A, 1OGOx, 1OGQA, 1OH4A, 1OHTA, 1OI7A, 1OIXA, 1OJQA, 1OK0A, 1OKSA, 1OLRA, 1OLTA, 1OM0A, 1OMRA, 1OOIX, 1OOTA, 1OP0A, 1OPDA, 1OPKA, 1OS6A, 1OS8A, 1OSHA, 1OUVA, 1OW1A, 1OWLA, 1OXJA, 1OXXK, 1OYGA, 1OYWA, 1OZ2A, 1OZ9A, 1OZAA, 1OZNA, 1P0HA, 1P1LA, 1P1MA, 1P2FA, 1P3CA, 1P5DX, 1P77A, 1P99A, 1P9AG,

1P9GA, 1PA2A, 1PA7A, 1PB7A, 1PBJA, 1PEFA, 1PENA, 1PG6A, 1PGVA, 1PI1A, 1PJXA, 1PKOA, 1PM1X, 1PMHX, 1PRZA, 1PS9A, 1PSWA, 1PUJA, 1PV5A, 1PVLA, 1PVXA, 1PWAA, 1PYFA, 1PZ4A, 1Q0RA, 1Q1CA, 1Q1FA, 1Q1UA, 1Q2LA, 1Q2UA, 1Q33A, 1Q35A, 1Q44A, 1Q5ZA, 1Q79A, 1Q7HA, 1Q7RA, 1Q8DA, 1Q8IA, 1Q92A, 1Q9BA, 1QADA, 1QAZA, 1QBAA, 1QCSA, 1QCXA, 1QDDA, 1QE3A, 1QF9A, 1QFMA, 1QGIA, 1QGQA, 1QHOA, 1QHQA, 1QJ8A, 1QJPA, 1QM7A, 1QNRA, 1QNXA, 1QOIA, 1QPCA, 1QQQA, 1QQYA, 1QSAA, 1QT9A, 1QTSA, 1QTWA, 1QUSA, 1QV1A, 1QWKA, 1QWOA, 1QWYA, 1QWZA, 1QXYA, 1QY5A, 1R0UA, 1R18A, 1R1MA, 1R26A, 1R2QA, 1R3DA, 1R4VA, 1R4XA, 1R55A, 1R5LA, 1R5MA, 1R5YA, 1R62A, 1R6CX, 1R6JA, 1R6VA, 1R6XA, 1R75A, 1R85A, 1R8ME, 1R9HA, 1R9LA, 1R9OA, 1R9WA, 1RC9A, 1RFSA, 1RGZA, 1RH9A, 1RHSA, 1RI6A, 1RIEA, 1RJBA, 1RJUV, 1RK6A, 1RL0A, 1RLHA, 1RLJA, 1RLKA, 1RM8A, 1RMGA, 1RO0A, 1ROCA, 1RPJA, 1RRKA, 1RTQA, 1RTTA, 1RU4A, 1RUTX, 1RUWA, 1RV9A, 1RVKA, 1RW1A, 1RWHA, 1RWJA, 1RWRA, 1RY6A, 1RYBA, 1RYOA, 1RYQA, 1RZ3A, 1RZ4A, 1S1FA, 1S1PA, 1S21A, 1S29A, 1S2BA, 1S2MA, 1S2OA, 1S2WA, 1S2ZA, 1S4QA, 1S68A, 1S69A, 1S8NA, 1S9UA, 1SAUA, 1SBXA, 1SDIA, 1SEFA, 1SENA, 1SF9A, 1SFSA, 1SGLA, 1SGWA, 1SHUX, 1SIFA, 1SK4A, 1SL8A, 1SMBA, 1SMLA, 1SNGA, 1SO7A, 1SPJA, 1SPVA, 1SQ9A, 1SRAA, 1SRPA, 1SUUA, 1SVBA, 1SVIA, 1SVSA, 1SVYA, 1SX7A, 1SZNA, 1T00A, 1T07A, 1T1EA, 1T1GA, 1T1UA, 1T27A, 1T2IA, 1T2JA, 1T32A, 1T3TA, 1T3YA, 1T46A, 1T4WA, 1T6AA, 1T6CA, 1T6EX, 1T71A, 1T7VA, 1T8HA, 1T8KA, 1T95A, 1T9HA, 1TA0A, 1TA8A, 1TAZA, 1TBFA, 1TDHA, 1TG0A, 1TG7A, 1THXA, 1TIFA, 1TJYA, 1TK1A, 1TKEA, 1TKJA, 1TL2A, 1TP6A, 1TPYA, 1TQ4A, 1TQGA, 1TQHA, 1TS9A, 1TT8A, 1TTZA, 1TU9A, 1TUAA, 1TUKA, 1TUOA, 1TVGA, 1TXJA, 1TYJA, 1TZVA, 1U02A, 1U09A, 1U0JA, 1U14A, 1U2KA, 1U2PA, 1U4GA, 1U53A, 1U5FA, 1U5PA, 1U61A, 1U6TA, 1U7LA, 1U7OA, 1U83A, 1U8XX, 1U9AA, 1U9CA, 1UA4A, 1UAIA, 1UARA, 1UASA, 1UCDA, 1UCSA, 1UCTA, 1UD2A, 1UEKA, 1UETA, 1UG6A, 1UGMA, 1UGNA, 1UH4A, 1UH9A, 1UHAA, 1UHNA, 1UI0A, 1UJ6A, 1UJ8A, 1UJBA, 1UJPA, 1UKFA, 1ULRA, 1UMKA, 1UNFX, 1UNQA, 1UNSA, 1UOKA, 1UOWA, 1UOYA, 1UOZA, 1UP9A, 1UPQA, 1UR1A, 1URGA, 1URRA, 1US0A, 1US3A, 1US5A, 1UT9A, 1UU3A, 1UUQA, 1UV0A, 1UV4A,

1UWFA, 1UWVA, 1UX6A, 1UXOA, 1UXXX, 1UXYA, 1UY4A, 1UYLA, 1UZ0A, 1UZ5A, 1UZE, 1V04A, 1V0AA, 1V0LA, 1V30A, 1V33A, 1V5CA, 1V93A, 1V9FA, 1V9MA, 1VAJA, 1VB0A, 1VBLA, 1VBWA, 1VCKA, 1VCTA, 1VE4A, 1VEMA, 1VF8A, 1VFYA, 1VG1A, 1VH2A, 1VHEA, 1VHFA, 1VHNA, 1VHOA, 1VHUA, 1VI3A, 1VIPA, 1VJFA, 1VJKA, 1VJVA, 1VJZA, 1VK1A, 1VK2A, 1VK4A, 1VK5A, 1VKBA, 1VKKA, 1VKWA, 1VL1A, 1VL9A, 1VLBA, 1VLOA, 1VLYA, 1VMBA, 1VPDA, 1VPRA, 1VQZA, 1VR3A, 1VR8A, 1VRMA, 1VSRA, 1VYFA, 1VYIA, 1VYKA, 1VYRA, 1VZOA, 1VZWA, 1W0HA, 1W0NA, 1W0OA, 1W1DA, 1W1OA, 1W24A, 1W2LA, 1W3UA, 1W3WA, 1W41A, 1W4SA, 1W4TA, 1W4XA, 1W66A, 1W6KA, 1W78A, 1W7OA, 1W8GA, 1W8OA, 1W9DM, 1W9HA, 1W9IA, 1WADA, 1WAKA, 1WB0A, 1WBEA, 1WC2A, 1WCHA, 1WCUA, 1WCWA, 1WD3A, 1WD5A, 1WDEA, 1WDPA, 1WDYA, 1WF3A, 1WGBA, 1WHZA, 1WJ9A, 1WJXA, 1WKAA, 1WKBA, 1WKYA, 1WL8A, 1WLJA, 1WLYA, 1WM1A, 1WMAA, 1WMDA, 1WNAA, 1WNHA, 1WOJA, 1WOLA, 1WOSA, 1WOUA, 1WP5A, 1WQGA, 1WRIA, 1WRJA, 1WS0A, 1WU3I, 1WU4A, 1WV3A, 1WVHA, 1WVNA, 1WWIA, 1WXIA, 1WXRA, 1WY3A, 1WZAA, 1WZZA, 1X0TA, 1X1IA, 1X1NA, 1X38A, 1X3KA, 1X3LA, 1X3SA, 1X42A, 1X46A, 1X6JA, 1X6OA, 1X6ZA, 1X8BA, 1X8QA, 1X91A, 1X9DA, 1X9QA, 1XBBA, 1XBIA, 1XC3A, 1XD6A, 1XDNA, 1XDWA, 1XDZA, 1XE1A, 1XEOA, 1XEUA, 1XFJA, 1XFKA, 1XGKA, 1XHDA, 1XJCA, 1XJDA, 1XJTA, 1XKGA, 1XKNA, 1XKRA, 1XKWA, 1XL0A, 1XMKA, 1XMTA, 1XOVA, 1XPHA, 1XQOA, 1XRFA, 1XS5A, 1XT5A, 1XTEA, 1XTPA, 1XV5A, 1XVQA, 1XVXA, 1XWWA, 1XWYA, 1XXNA, 1XZZA, 1Y02A, 1Y08A, 1Y0KA, 1Y0MA, 1Y0PA, 1Y2FA, 1Y4CA, 1Y4WA, 1Y57A, 1Y63A, 1Y6BA, 1Y80A, 1Y81A, 1Y8AA, 1Y93A, 1Y9TA, 1Y9UA, 1YA9A, 1YB3A, 1YBZA, 1YCKA, 1YD0A, 1YE8A, 1YFQA, 1YG9A, 1YH2A, 1YHFA, 1YHTA, 1YI9A, 1YIIA, 1YIQA, 1YKSA, 1YLEA, 1YN4A, 1YONA, 1YP5A, 1YQBA, 1YQEA, 1YQSA, 1YQTA, 1YR2A, 1YRWA, 1YS1X, 1YSQA, 1YT3A, 1YT8A, 1YTQA, 1YU5X, 1YULA, 1YW5A, 1YW9A, 1YWFA, 1YWMA, 1YZFA, 1YZLA, 1YZMA, 1Z06A, 1Z16A, 1Z1LA, 1Z1SA, 1Z21A, 1Z22A, 1Z2NX, 1Z2UA, 1Z3QA, 1Z3XA, 1Z4RA, 1Z57A, 1Z67A, 1Z6MA, 1Z6NA, 1Z70X, 1Z7WA, 1Z8OA, 1Z9TA, 1ZA0A, 1ZARA, 1ZBFA, 1ZC0A, 1ZCEA, 1ZCJA, 1ZD8A, 1ZD9A, 1ZDEA, 1ZDYA, 1ZGHA, 1ZGKA, 1ZHVA, 1ZHXA, 1ZI8A, 1ZINA, 1ZIWA,

1ZJRA, 1ZK5A, 1ZKJA, 1ZKLA, 1ZLBA, 1ZLDA, 1ZLMA, 1ZMAA, 1ZN6A, 1ZODA, 1ZPWX, 1ZR6A, 1ZRHA, 1ZS9A, 1ZSQA, 1ZSWA, 1ZTFA, 1ZU0A, 1ZUUA, 1ZVDA, 1ZWXA, 1ZX6A, 1ZZKA, 1ZZMA, 2A0BA, 2A0MA, 2A0NA, 2A14A, 2A1IA, 2A1VA, 2A2CA, 2A2KA, 2A3MA, 2A4DA, 2A4VA, 2A5JA, 2A5SA, 2A65A, 2A6YA, 2A6ZA, 2A7BA, 2A90A, 2A9IA, 2A9OA, 2AANA, 2ABKA, 2ABSA, 2AC1A, 2AE0X, 2AG6A, 2AGKA, 2AH2A, 2AH5A, 2AHEA, 2AHNA, 2AIQA, 2AJ6A, 2ALEA, 2AMHA, 2ANYA, 2AP1A, 2APLA, 2APXA, 2AQ5A, 2AQMA, 2AR1A, 2AR5A, 2ARSA, 2ASCA, 2ATMA, 2ATVA, 2AU3A, 2AVKA, 2AXCA, 2AXNA, 2AYDA, 2AZPA, 2AZWA, 2B02A, 2B0TA, 2B0VA, 2B1EA, 2B1KA, 2B1MA, 2B3HA, 2B49A, 2B4VA, 2B4WA, 2B4ZA, 2B7UA, 2B8IA, 2B9EA, 2B9LA, 2B9WA, 2BBHA, 2BBRA, 2BCEA, 2BEPA, 2BF6A, 2BFWA, 2BG1A, 2BGIA, 2BGSA, 2BH4X, 2BHUA, 2BI7A, 2BIBA, 2BJQA, 2BJUA, 2BJVA, 2BK8A, 2BK9A, 2BKFA, 2BL1A, 2BMDA, 2BMWA, 2BOGX, 2BQ8X, 2BRFA, 2BSYA, 2BUEA, 2BVLA, 2BW0A, 2BWKA, 2BWMA, 2BWQA, 2BX6A, 2BYOA, 2BZ7A, 2BZGA, 2BZLA, 2C0HA, 2C0YA, 2C1IA, 2C1XA, 2C2PA, 2C43A, 2C4XA, 2C53A, 2C5GA, 2C60A, 2C6UA, 2C71A, 2C78A, 2C7SA, 2C7XA, 2C8BX, 2C8SA, 2C9EA, 2C9RA, 2CALA, 2CB9A, 2CBOA, 2CBZA, 2CC1A, 2CDNA, 2CE0A, 2CE2X, 2CF5A, 2CFMA, 2CG7A, 2CGQA, 2CI1A, 2CISA, 2CIUA, 2CIWA, 2CJ7A, 2CJJA, 2CJZA, 2CKKA, 2CKXA, 2CM1A, 2CM5A, 2CMTA, 2CNQA, 2CU9A, 2CULA, 2CVBA, 2CVCA, 2CVEA, 2CVJA, 2CW4A, 2CW9A, 2CWRA, 2CWSA, 2CWYA, 2CX1A, 2CXAA, 2CXCA, 2CXHA, 2CXVA, 2CXYA, 2CY2A, 2CY7A, 2CYGA, 2CYJA, 2CZLA, 2D0SA, 2D1SA, 2D2EA, 2D3IA, 2D3YA, 2D48A, 2D4PA, 2D58A, 2D59A, 2D5BA, 2D81A, 2D9RA, 2DBNA, 2DBYA, 2DCBA, 2DCFA, 2DCHX, 2DCKA, 2DDXA, 2DE2A, 2DEJA, 2DH2A, 2DHOA, 2DJHA, 2DKVA, 2DSXA, 2DT8A, 2DULA, 2DUYA, 2DWKA, 2DXAA, 2DY1A, 2E0QA, 2E0TA, 2E26A, 2E3BA, 2E3HA, 2E3NA, 2E3WA, 2E4TA, 2E56A, 2E6MA, 2E7VA, 2E7ZA, 2E9LA, 2EA3A, 2EAEA, 2EAQA, 2EBFX, 2ECEA, 2EF0A, 2EGUA, 2EHGA, 2EHS A, 2EI9A, 2EIFA, 2EJ9A, 2EK8A, 2ELFA, 2ENGA, 2ERFA, 2ERLA, 2ERWA, 2ESAA, 2ESBA, 2ESUA, 2ETBA, 2ETJA, 2EU9A, 2EVAA, 2EVEA, 2EW0A, 2EW6A, 2EWFA, 2EX2A, 2EYIA, 2F1NA, 2F1WA, 2F21A, 2F4QA, 2F60K, 2F6EA, 2F6RA, 2F71A, 2F82A, 2F8LA, 2F9FA, 2F9LA, 2FB6A, 2FBAA, 2FC3A, 2FCFA, 2FCLA, 2FDJA, 2FDRA, 2FE5A,

2FEMA, 2FEZA, 2FFQA, 2FG1A, 2FH7A, 2FHFA, 2FI0A, 2FI1A, 2FI9A, 2FJ8A, 2FJ9A, 2FK8A, 2FK9A, 2FL7A, 2FLSA, 2FMAA, 2FN8A, 2FNCA, 2FOZA, 2FPQA, 2FQ3A, 2FQXA, 2FR2A, 2FR7A, 2FRGP, 2FSJA, 2FSQA, 2FSRA, 2FSTX, 2FSUA, 2FSXA, 2FU0A, 2FUPA, 2FUZA, 2FVGA, 2FVVA, 2FVYA, 2FW5A, 2FWHA, 2FWTA, 2FWVA, 2FXUA, 2FY6A, 2FY7A, 2FYGA, 2FYOA, 2FYTA, 2FZPA, 2G03A, 2G29A, 2G2CA, 2G2OA, 2G3RA, 2G52A, 2G5RA, 2G5XA, 2G62A, 2G6FX, 2G7BA, 2G9FA, 2GA8A, 2GF9A, 2GFOA, 2GG6A, 2GGCA, 2GHSA, 2GJLA, 2GKEA, 2GKGA, 2GKPA, 2GKRI, 2GM6A, 2GNPA, 2GO2A, 2GOUA, 2GPIA, 2GPJA, 2GQTA, 2GS5A, 2GS8A, 2GSJA, 2GTQA, 2GU3A, 2GUFA, 2GUIA, 2GUYA, 2GWMA, 2GWNA, 2GWRA, 2GX6A, 2GZQA, 2GZSA, 2GZVA, 2H14A, 2H17A, 2H1VA, 2H2TB, 2H30A, 2H58A, 2H5CA, 2H6DA, 2H7MA, 2H7OA, 2HAIA, 2HBTA, 2HBWA, 2HC1A, 2HC8A, 2HCFA, 2HCUA, 2HD9A, 2HDOA, 2HE4A, 2HE7A, 2HHCA, 2HHGA, 2HHPA, 2HIAA, 2HJEA, 2HKJA, 2HL7A, 2HL9A, 2HLRA, 2HLVA, 2HLYA, 2HM7A, 2HP7A, 2HPJA, 2HPSA, 2HPWA, 2HQ0A, 2HQ2A, 2HQ6A, 2HQKA, 2HS5A, 2HSHA, 2HTSA, 2HUJA, 2HVFA, 2HVMA, 2HW2A, 2HW4A, 2HWVA, 2HWXA, 2HXP, 2HXS, 2HY7A, 2HYKA, 2HYVA, 2HZCA, 2HZQA, 2I0KA, 2I0OA, 2I0ZA, 2I1JA, 2I1QA, 2I1UA, 2I24N, 2I49A, 2I4AA, 2I4IA, 2I53A, 2I5FA, 2I5UA, 2I5VO, 2I61A, 2I6CA, 2I6DA, 2I6JA, 2I6VA, 2I9CA, 2I9WA, 2IA7A, 2IAFA, 2IAYA, 2IBJA, 2ICGA, 2ICIA, 2ICTA, 2IDCA, 2IGDA, 2IGPA, 2IGVA, 2IGXA, 2IHDA, 2II0A, 2II2A, 2II6A, 2IJAA, 2IJES, 2ILLA, 2ILRA, 2IM9A, 2IMGA, 2IMQX, 2IMRA, 2IN3A, 2IONA, 2IORA, 2IP1A, 2IP6A, 2IPXA, 2IQYA, 2IRXA, 2IS9A, 2ISBA, 2IT6A, 2IU1A, 2IVNA, 2IW0A, 2IW1A, 2IWAA, 2IWCA, 2IWNA, 2IWRA, 2IY9A, 2IYVA, 2IZRA, 2J07A, 2J0AA, 2J0IA, 2J0PA, 2J13A, 2J1AA, 2J2CA, 2J3XA, 2J47A, 2J4XA, 2J6AA, 2J7UA, 2J89A, 2J8BA, 2J8GA, 2J8KA, 2JC4A, 2JC5A, 2JCKA, 2JCNA, 2JCQA, 2JD9A, 2JDCA, 2JEKA, 2JEWA, 2JFNA, 2JFRA, 2JG0A, 2JG6A, 2JGPA, 2JH1A, 2JHQA, 2JHSA, 2JINA, 2JJNA, 2JJQA, 2JKBA, 2JKUA, 2JLIA, 2JLQA, 2LISA, 2NLRA, 2NLSA, 2NN5A, 2NOOA, 2NPNA, 2NQ3A, 2NQCA, 2NQWA, 2NR7A, 2NRLA, 2NRRA, 2NRTA, 2NS0A, 2NSFA, 2NSMA, 2NSQA, 2NSZA, 2NUHA, 2NWFA, 2NWH, 2NX2A, 2NXCA, 2NXFA, 2NYKA, 2NZLA, 2O04A, 2O0BA, 2O0JA, 2O0MA, 2O0PA, 2O14A, 2O1AA, 2O2GA, 2O2IA, 2O2XA, 2O36A, 2O37A, 2O3HA,

2O5VA, 2O71A, 2O7AA, 2O7IA, 2O7RA, 2O7SA, 2O85A, 2O8LA, 2O9SA, 2O9UX,
2OBIA, 2OBLA, 2OC3A, 2OC5A, 2OCGA, 2OCHA, 2OCSA, 2OD5A, 2ODHA,
2ODLA, 2ODPA, 2OEBA, 2OF3A, 2OFZA, 2OG2A, 2OG4A, 2OG5A, 2OGTA,
2OITA, 2OJHA, 2OKMA, 2OKTA, 2OLGA, 2OLNA, 2OLRA, 2OMLA, 2ONSA,
2OO3A, 2OP6A, 2OPCA, 2OPJA, 2OQ5A, 2OQKA, 2OQZA, 2OROA, 2OS0A,
2OSSA, 2OSXA, 2OT9A, 2OV0A, 2OVJA, 2OVUA, 2OXNA, 2OY3A, 2OY7A,
2OYPA, 2OYRA, 2OZFA, 2OZPA, 2OZTA, 2P0BA, 2P0DA, 2P0KA, 2P0OA,
2P14A, 2P17A, 2P18A, 2P39A, 2P3HA, 2P3KA, 2P41A, 2P4DA, 2P4FA, 2P4HX,
2P4OA, 2P51A, 2P57A, 2P5DA, 2P5KA, 2P5YA, 2P65A, 2P6WA, 2P70A, 2P7SA,
2P8RA, 2P9BA, 2P9WA, 2PAGA, 2PB7A, 2PBNA, 2PC1A, 2PE4A, 2PE8A, 2PESA,
2PETA, 2PFZA, 2PGEA, 2PGFA, 2PHZA, 2PI6A, 2PJZA, 2PKPA, 2PKTA, 2PL3A,
2PLCA, 2PLUA, 2PLWA, 2PM1A, 2PMKA, 2PNDA, 2PNEA, 2PNUA, 2PNWA,
2PNYA, 2PO4A, 2PPLA, 2PPNA, 2PQ3A, 2PQ6A, 2PQ8A, 2PQXA, 2PSDA,
2PSTX, 2PT2A, 2PTHA, 2PU3A, 2PV4A, 2PVBA, 2PVUA, 2PWAA, 2PWWA,
2PX4A, 2PXRC, 2PXXA, 2Q09A, 2Q0IA, 2Q0ZX, 2Q1HA, 2Q3HA, 2Q3WA, 2Q40A,
2Q43A, 2Q4MA, 2Q4UA, 2Q4WA, 2Q5TA, 2Q5XA, 2Q88A, 2Q8KA, 2Q99A, 2Q9VA,
2QC5A, 2QCPX, 2QDXA, 2QEDA, 2QEVA, 2QFEA, 2QFFA, 2QG1A, 2QG6A,
2QGUA, 2QHPA, 2QHSA, 2QHTA, 2QIKA, 2QIMA, 2QJLA, 2QJXA, 2QK0A,
2QK1A, 2QK2A, 2QKDA, 2QLTA, 2QMLA, 2QMQA, 2QN0A, 2QNGA, 2QNKA,
2QNWA, 2QO4A, 2QOLA, 2QORA, 2QP2A, 2QPWA, 2QPZA, 2QQ8A, 2QQIA,
2QQJA, 2QR3A, 2QR7A, 2QRUA, 2QSAA, 2QSBA, 2QSKA, 2QSVA, 2QTFA,
2QTZA, 2QUAA, 2QUOA, 2QUPA, 2QVKA, 2QX2A, 2QXIA, 2QY9A, 2QYUA,
2QZQA, 2QZSA, 2QZUA, 2R0CA, 2R0JA, 2R0SA, 2R16A, 2R1NA, 2R2YA, 2R2ZA,
2R31A, 2R39A, 2R3AA, 2R44A, 2R48A, 2R4GA, 2R4QA, 2R51A, 2R58A, 2R5GA,
2R5TA, 2R751, 2R99A, 2R9FA, 2R9VA, 2RA7A, 2RA8A, 2RA9A, 2RAUA, 2RB8A,
2RBKA, 2RCIA, 2RCTA, 2RDGA, 2RDIA, 2RDQA, 2REIA, 2RERA, 2REUA,
2RFAA, 2RFFA, 2RGXA, 2RH3A, 2RHDA, 2RHFA, 2RHJA, 2RIKA, 2RJ2A,
2RJNA, 2RKMA, 2RKQA, 2RKUA, 2RKVA, 2SLIA, 2UURA, 2UUXA, 2UVJA,
2UWFA, 2UX7A, 2UY2A, 2UYOA, 2UYTA, 2V05A, 2V0CA, 2V0SA, 2V0UA,
2V1MA, 2V26A, 2V3GA, 2V3IA, 2V3KA, 2V3UA, 2V3VA, 2V4VA, 2V4XA, 2V5TA,
2V75A, 2V7FA, 2V7KA, 2V8IA, 2V9KA, 2V9VA, 2VACA, 2VANA, 2VAPA, 2VB1A,

2VBUA, 2VC8A, 2VCAA, 2VCHA, 2VCLA, 2VCNA, 2VECA, 2VFBA, 2VFKA, 2VFRA, 2VG9A, 2VGAA, 2VGDA, 2VGQA, 2VH7A, 2VHKA, 2VIMA, 2VJWA, 2VK2A, 2VK3A, 2VM5A, 2VMHA, 2VN4A, 2VNRA, 2VOSA, 2VOVA, 2VPJA, 2VQ4A, 2VQ8A, 2VQPA, 2VQXA, 2VRIA, 2VU4A, 2VU9A, 2VUVA, 2VUWA, 2VVKA, 2VWRA, 2VX5A, 2VXRA, 2VXZA, 2VY8A, 2VYOA, 2VYWA, 2VZ5A, 2W0BA, 2W0GA, 2W0IA, 2W15A, 2W18A, 2W22A, 2W2JA, 2W2RA, 2W39A, 2W3JA, 2W3TA, 2W3WA, 2W3ZA, 2W47A, 2W4FA, 2W4OA, 2W5AA, 2W5NA, 2W5QA, 2W61A, 2W6KA, 2W86A, 2W91A, 2W9HA, 2WA7A, 2WAAA, 2WANA, 2WAOA, 2WAWA, 2WB0X, 2WBFX, 2WBNA, 2WBQA, 2WBXA, 2WCJA, 2WCOA, 2WDCA, 2WEIA, 2WF1A, 2WF7A, 2WFBA, 2WFIA, 2WFOA, 2WFPA, 2WGYA, 2WHKA, 2WHLA, 2WHMA, 2WICA, 2WIYA, 2WJ5A, 2WJEA, 2WJRA, 2WKQA, 2WL1A, 2WLRA, 2WM3A, 2WM5A, 2WM8A, 2WMFA, 2WN4A, 2WNFA, 2WNKA, 2WNPF, 2WO4A, 2WOLA, 2WOYA, 2WPGA, 2WQMA, 2WRYA, 2WSDA, 2WSIA, 2WT0A, 2WTAA, 2WTJA, 2WURA, 2WUUA, 2WV3A, 2WW5A, 2WW8A, 2WWEA, 2WXFA, 2WXUA, 2WY4A, 2WYQA, 2WZ8A, 2WZ9A, 2WZBA, 2WZKA, 2X1BA, 2X25B, 2X3MA, 2X3NA, 2X42A, 2X46A, 2X47A, 2X49A, 2X4LA, 2X55A, 2X5NA, 2X5OA, 2X5PA, 2X5YA, 2X7MA, 2X7QA, 2X8XX, 2X97A, 2X9KA, 2X9OA, 2X9XA, 2X9ZA, 2XA5A, 2XB0X, 2XBGA, 2XBKA, 2XBTa, 2XC2A, 2XDYA, 2XFDA, 2XFRA, 2XG3A, 2XGVA, 2XIOA, 2XIRA, 2XJKA, 2XJPA, 2XKIA, 2XKRA, 2XM5A, 2XMBA, 2XMZA, 2XNQA, 2XODA, 2XOMA, 2XOVA, 2XQ0A, 2XQOA, 2XRIA, 2XSEA, 2XSTA, 2XTPA, 2XU0A, 2XU3A, 2XVSA, 2XW9A, 2XY2A, 2XY4A, 2XZVA, 2Y1QA, 2Y24A, 2Y4SA, 2Y88A, 2Y8BA, 2Y8KA, 2Y9RT, 2YA0A, 2YASA, 2YB1A, 2YB6A, 2YD1A, 2YD4A, 2YFUA, 2YH5A, 2YHAA, 2YHCA, 2YHSA, 2YIHA, 2YQCA, 2YRXA, 2YV1A, 2YV2A, 2YV5A, 2YVIA, 2YVTA, 2YVWA, 2YWDA, 2YWJA, 2YWKA, 2YWNA, 2YWRA, 2YXBA, 2YXFA, 2YXMA, 2YXXA, 2YZLA, 2YZQA, 2YZVA, 2Z0MA, 2Z0QA, 2Z0XA, 2Z14A, 2Z1AA, 2Z1EA, 2Z2NA, 2Z2XA, 2Z2ZA, 2Z38A, 2Z4UA, 2Z5YA, 2Z62A, 2Z6HA, 2Z6OA, 2Z72A, 2Z81A, 2Z83A, 2Z84A, 2Z8LA, 2Z8XA, 2ZB4A, 2ZBXA, 2ZC1A, 2ZEQA, 2ZFIA, 2ZFYA, 2ZGDA, 2ZGGA, 2ZHJA, 2ZHNA, 2ZIBA, 2ZICA, 2ZK9X, 2ZKMX, 2ZMUA, 2ZNRA, 2ZNTA, 2ZPMA, 2ZPOA, 2ZPTX, 2ZQ5A, 2ZQ7A, 2ZQEA, 2ZTXA, 2ZU5A, 2ZUMA, 2ZUYA, 2ZWSA, 2ZWUA, 2ZXQA, 2ZYCA, 2ZYOA, 2ZZJA,

2ZZRA, 3A02A, 3A03A, 3A09A, 3A0NA, 3A0SA, 3A0TA, 3A0UA, 3A1KA, 3A1TA, 3A1WA, 3A2QA, 3A2ZA, 3A38A, 3A40X, 3A4CA, 3A4GA, 3A6TA, 3A72A, 3A7IA, 3A7LA, 3A7NA, 3A7RA, 3A99A, 3A9BA, 3AA4A, 3AALA, 3AAMA, 3AAPA, 3AB6A, 3ABAA, 3ACHA, 3ACXA, 3ADYA, 3AFHA, 3AFVA, 3AG7A, 3AGNA, 3AHZA, 3AI4A, 3AI5A, 3AIIA, 3AJ7A, 3AJDA, 3AJEA, 3AJZA, 3AKBA, 3AKHA, 3AKSA, 3AL2A, 3ALFA, 3AMRA, 3AOXA, 3AP9A, 3APAA, 3ARXA, 3ATRA, 3AW5A, 3AWMA, 3B2NA, 3B34A, 3B40A, 3B49A, 3B4XA, 3B5OA, 3B6EA, 3B72A, 3B79A, 3B7FA, 3B7OA, 3B84A, 3B8BA, 3B9UA, 3BB0A, 3BB7A, 3BC9A, 3BCHA, 3BCIA, 3BCYA, 3BDIA, 3BDLA, 3BE4A, 3BERA, 3BG8A, 3BHYA, 3BI7A, 3BIAX, 3BIQA, 3BIYA, 3BJ91, 3BJCA, 3BJHA, 3BJNA, 3BJOA, 3BK5A, 3BKBA, 3BKPA, 3BKRA, 3BMVA, 3BN6A, 3BO5A, 3BODA, 3BOEA, 3BONA, 3BORA, 3BOVA, 3BPTA, 3BQEA, 3BQVA, 3BQWA, 3BR8A, 3BS2A, 3BT5A, 3BV8A, 3BVSA, 3BVUA, 3BWHA, 3BWXA, 3BWZA, 3BY8A, 3BYVA, 3BZGA, 3BZTA, 3C0CA, 3C0IA, 3C1OA, 3C1RA, 3C4HA, 3C5AA, 3C5EA, 3C5HA, 3C5KA, 3C5VA, 3C65A, 3C6AA, 3C71A, 3C7FA, 3C7XA, 3C8UA, 3C8YA, 3C96A, 3C9PA, 3C9XA, 3C9ZA, 3CA7A, 3CAOA, 3CB6A, 3CBNA, 3CBQA, 3CBZA, 3CC6A, 3CDUA, 3CE7A, 3CECA, 3CFWA, 3CFZA, 3CG3A, 3CG4A, 3CGXA, 3CH0A, 3CH4B, 3CHJA, 3CHMA, 3CIVA, 3CJ1A, 3CJJA, 3CJMA, 3CJWA, 3CKKA, 3CKMA, 3CLJA, 3CLMA, 3CLYA, 3CM0A, 3CMQA, 3COIA, 3COUA, 3CP0A, 3CP5A, 3CPGA, 3CQTA, 3CQVA, 3CRRA, 3CSGA, 3CSKA, 3CSSA, 3CT5A, 3CTGA, 3CTKA, 3CTRA, 3CU4A, 3CU9A, 3CUXA, 3CUZA, 3CW3A, 3CX2A, 3CXMA, 3CYIA, 3CZ7A, 3CZCA, 3D02A, 3D06A, 3D1PA, 3D1RA, 3D22A, 3D30A, 3D3YA, 3D3ZA, 3D4EA, 3D4MA, 3D6JA, 3D79A, 3D89A, 3D8RA, 3D9HA, 3DAIA, 3DANA, 3DASA, 3DAUA, 3DB7A, 3DCYA, 3DCZA, 3DDJA, 3DEFA, 3DEOA, 3DF7A, 3DFFA, 3DFGA, 3DGLA, 3DGTA, 3DHAA, 3DHNA, 3DJ5A, 3DJ9A, 3DKCA, 3DKMA, 3DKPA, 3DKRA, 3DLCA, 3DLMA, 3DLNA, 3DM9B, 3DMGA, 3DMIA, 3DMLA, 3DMNA, 3DMSA, 3DNUA, 3DOUA, 3DP5A, 3DPKA, 3DQYA, 3DS8A, 3DSMA, 3DSOA, 3DT5A, 3DT8A, 3DU1X, 3DV9A, 3DX5A, 3DXLA, 3DXSX, 3DXTA, 3DXYA, 3E0HA, 3E13X, 3E21A, 3E23A, 3E2OA, 3E2QA, 3E3UA, 3E46A, 3E4GA, 3E4HA, 3E6JA, 3E7RL, 3E8TA, 3E8YX, 3E9FA, 3E9LA, 3E9OA, 3E9VA, 3EA6A, 3EAFA, 3EAZA, 3EBHA, 3EBVA, 3ED5A, 3EDHA, 3EDYA, 3EEHA, 3EF1A,

3EF6A, 3EG3A, 3EG4A, 3EGAA, 3EH1A, 3EHGA, 3EHIX, 3EIFA, 3EJ2A, 3EJFA, 3EJGA, 3EJNA, 3EKIA, 3ELBA, 3ELJA, 3ELNA, 3ELSA, 3ELWA, 3ELXA, 3EMCA, 3EMMA, 3EMRA, 3EO5A, 3EO7A, 3EQCA, 3ERBA, 3ESSA, 3ESUF, 3EU3A, 3EURA, 3EVFA, 3EVPA, 3EVRA, 3EW8A, 3EWNA, 3EXMA, 3EXNA, 3EXQA, 3EXYA, 3EY6A, 3EYEA, 3EYGA, 3EZMA, 3EZOA, 3F04A, 3F0ZA, 3F2EA, 3F2ZA, 3F3QA, 3F44A, 3F45A, 3F49S, 3F4MA, 3F4SA, 3F5BA, 3F67A, 3F6YA, 3F7RA, 3F7WA, 3F8KA, 3F9MA, 3FAJA, 3FANA, 3FAWA, 3FBLA, 3FCIA, 3FDHA, 3FDJA, 3FEUA, 3FF4A, 3FGDA, 3FGHA, 3FH2A, 3FHDA, 3FHGA, 3FIAA, 3FIBA, 3FJ0A, 3FJ2A, 3FK5A, 3FK8A, 3FKCA, 3FL2A, 3FLBA, 3FM0A, 3FMUA, 3FN7A, 3FNDA, 3FO8D, 3FOJA, 3FOTA, 3FP3A, 3FP5A, 3FQGA, 3FR1A, 3FRGA, 3FRHA, 3FRRA, 3FS5A, 3FSAA, 3FSSA, 3FT9A, 3FTDA, 3FTJA, 3FTKA, 3FUNA, 3FUQA, 3FUWA, 3FVYA, 3FWAA, 3FWKA, 3FXKA, 3FXZA, 3FYDA, 3FYMA, 3FZ4A, 3FZEA, 3FZGA, 3G0LA, 3G0MA, 3G0VA, 3G39A, 3G3LA, 3G3OA, 3G3TA, 3G40A, 3G4PA, 3G51A, 3G5BA, 3G5SA, 3G5TA, 3G63A, 3G6MA, 3G7UA, 3G85A, 3G91A, 3G9TA, 3G9XA, 3GA2A, 3GA3A, 3GA4A, 3GA7A, 3GA8A, 3GBEA, 3GBGA, 3GBSA, 3GBWA, 3GBZA, 3GC2A, 3GCEA, 3GCZA, 3GD0A, 3GDBA, 3GENA, 3GF8A, 3GFPA, 3GG6A, 3GG7A, 3GGDA, 3GH5A, 3GIRA, 3GIWA, 3GJYA, 3GK6A, 3GKJA, 3GKMA, 3GKRA, 3GKZA, 3GLJA, 3GLKA, 3GMFA, 3GMIA, 3GMSA, 3GMUB, 3GMVX, 3GNRA, 3GNZP, 3GO5A, 3GO9A, 3GOEA, 3GOHA, 3GONA, 3GP6A, 3GPIA, 3GQVA, 3GREA, 3GRHA, 3GRLA, 3GRUA, 3GT2A, 3GT5A, 3GVOA, 3GVRA, 3GWFA, 3GWIA, 3GX8A, 3GY9A, 3GYLB, 3GZ9A, 3GZKA, 3H04A, 3H0HA, 3H0OA, 3H0XA, 3H14A, 3H1DA, 3H1GA, 3H20A, 3H2GA, 3H2XA, 3H2YA, 3H31A, 3H34A, 3H41A, 3H4GA, 3H4TA, 3H4XA, 3H6DA, 3H6QA, 3H74A, 3H75A, 3H79A, 3H7CX, 3H7IA, 3H7RA, 3H7UA, 3H8HA, 3H8ZA, 3H93A, 3H9CA, 3H9MA, 3HA9A, 3HAKA, 3HAPA, 3HBEX, 3HBFA, 3HBNA, 3HBZA, 3HC1A, 3HC7A, 3HCZA, 3HD4A, 3HDLA, 3HDXA, 3HF0A, 3HFWA, 3HGBA, 3HGLA, 3HGPA, 3HH8A, 3HICA, 3HJ7A, 3HJEA, 3HJRA, 3HJZA, 3HKOA, 3HL8A, 3HLGA, 3HMCA, 3HMFA, 3HMIA, 3HMMA, 3HMYA, 3HN7A, 3HNXA, 3HNYM, 3HOLA, 3HP4A, 3HP7A, 3HPCX, 3HQCA, 3HR1A, 3HR6A, 3HR8A, 3HRAA, 3HRGA, 3HRLA, 3HRPA, 3HUIA, 3HUTA, 3HV8A, 3HVIA, 3HVWA, 3HX3A, 3HXLA, 3HXWA, 3HY3A, 3HYNA, 3HZ6A, 3HZ7A, 3HZ8A,

3I06A, 3I10A, 3I11A, 3I18A, 3I2VA, 3I31A, 3I33A, 3I35A, 3I36A, 3I45A, 3I47A,
3I4GA, 3I65A, 3I6IA, 3I6VA, 3I7AA, 3I7MA, 3I8BA, 3I94A, 3I99A, 3IANA, 3IARA,
3IB0A, 3IBDA, 3IBZA, 3IC4A, 3ICHA, 3ID1A, 3IDVA, 3IDWA, 3IEEA, 3IESA,
3IG0A, 3IG3A, 3IG5A, 3IGNA, 3IGZB, 3II2A, 3II7A, 3IIDA, 3IIIA, 3IIJA, 3IIXA,
3IJCA, 3IK2A, 3IKWA, 3ILCA, 3ILFA, 3ILSA, 3ILZA, 3IM1A, 3IM6A, 3IM9A,
3IMKA, 3IO3A, 3IOFA, 3IOQA, 3IOSA, 3IOXA, 3IP0A, 3IPCA, 3IPWA, 3IQTA,
3IR4A, 3IRAA, 3IRPX, 3ISUA, 3IU0A, 3IU5A, 3IU6A, 3IUFA, 3IUJA, 3IUUA,
3IUZA, 3IV0A, 3IV4A, 3IVEA, 3IVYA, 3IWLA, 3IXLA, 3IXRA, 3JQ0A, 3JRNA,
3JS6A, 3JS8A, 3JSNA, 3JSRA, 3JSZA, 3JTEA, 3JTMA, 3JU0A, 3JUDA, 3JUGA,
3JVEA, 3JVIA, 3JVL A, 3JWGA, 3JXHC, 3JXPA, 3JYSA, 3JYZA, 3JZ9A, 3JZYA,
3K01A, 3K0BA, 3K0XA, 3K0YA, 3K1HA, 3K1UA, 3K1ZA, 3K3OA, 3K3VA, 3K4ZA,
3K50A, 3K5JA, 3K6IA, 3K6PA, 3K6VA, 3K7IB, 3K7XA, 3K82A, 3K89A, 3K8UA,
3K8WA, 3KA2A, 3KA7A, 3KB5A, 3KB9A, 3KBBA, 3KBGA, 3KC6A, 3KCIA,
3KEVA, 3KEYA, 3KFFA, 3KFJA, 3KFLA, 3KFOA, 3KG4A, 3KH6A, 3KH7A,
3KHIA, 3KHQA, 3KK8A, 3KKQA, 3KKWA, 3KLBA, 3KLKA, 3KLRA, 3KLUA,
3KNVA, 3KOGA, 3KP8A, 3KPKA, 3KQ0A, 3KR6A, 3KR7A, 3KREA, 3KRRA,
3KS3A, 3KSNA, 3KSXA, 3KT9A, 3KV0A, 3KVDD, 3KWLA, 3KWUA, 3KXWA,
3KYFA, 3KZ7A, 3KZDA, 3KZXA, 3L00A, 3L11A, 3L1NA, 3L22A, 3L23A, 3L2AA,
3L3FX, 3L42A, 3L4AA, 3L4EA, 3L4HA, 3L4NA, 3L4RA, 3L4YA, 3L5AA, 3L5IA,
3L5KA, 3L6AA, 3L6GA, 3L6NA, 3L78A, 3L7YA, 3L8FA, 3L9BA, 3L9SA, 3L9UA,
3LA8A, 3LAXA, 3LBJE, 3LCCA, 3LDGA, 3LDUA, 3LE0A, 3LE2A, 3LECA, 3LEQA,
3LEWA, 3LEZA, 3LFUA, 3LHEA, 3LHIA, 3LHOA, 3LHSA, 3LI9A, 3LIGA, 3LIJA,
3LJEA, 3LJTA, 3LJUX, 3LK7A, 3LKMA, 3LL2A, 3LL7A, 3LLBA, 3LLCA, 3LLOA,
3LLVA, 3LM3A, 3LN3A, 3LNBA, 3LO8A, 3LOPA, 3LOVA, 3LP5A, 3LP8A, 3LPCA,
3LPZA, 3LQ0A, 3LQBA, 3LQHA, 3LQYA, 3LS0A, 3LTIA, 3LTJA, 3LUUA, 3LW6A,
3LWAA, 3LWTX, 3LWXA, 3LX5A, 3LXLA, 3LXPA, 3M07A, 3M0BA, 3M1TA,
3M31A, 3M3GA, 3M3MA, 3M3PA, 3M4RA, 3M5LA, 3M5QA, 3M66A, 3M6CA,
3M6WA, 3M70A, 3M7DA, 3M8UA, 3M97X, 3MAMA, 3MAOA, 3MBRX, 3MCXA,
3MDMA, 3MDQA, 3MDUA, 3MEWA, 3MF6A, 3MGWA, 3MGZA, 3MHWU, 3MHZA,
3MI4A, 3MJ6A, 3MJFA, 3MK9A, 3MKCA, 3ML3A, 3ML5A, 3MM4A, 3MOEA,
3MOKA, 3MP8A, 3MPKA, 3MPPG, 3MQZA, 3MSRA, 3MSTA, 3MSWA, 3MT0A,

3MU7A, 3MVNA, 3MVSA, 3MW8A, 3MWFA, 3MWQA, 3MWUA, 3MWZA, 3MX7A, 3MXUA, 3MXZA, 3MY2A, 3MYIA, 3MZFA, 3MZZA, 3N0AA, 3N0KA, 3N0RA, 3N0XA, 3N17A, 3N1DA, 3N2TA, 3N35A, 3N3UA, 3N5AA, 3N6ZA, 3N7OA, 3N7TA, 3N94A, 3N98A, 3N9KA, 3NBMA, 3NCLA, 3NDIA, 3NDQA, 3NE0A, 3NEDA, 3NGPA, 3NH6A, 3NHIA, 3NI2A, 3NIAA, 3NIWA, 3NJUA, 3NKQA, 3NKSA, 3NNBA, 3NO2A, 3NO3A, 3NOHA, 3NOLA, 3NPDA, 3NPHB, 3NQ8A, 3NQHA, 3NQXA, 3NR5A, 3NRSA, 3NRWA, 3NSZA, 3NULA, 3NUQA, 3NV1A, 3NVBA, 3NVSA, 3NW7A, 3NWOA, 3NXPA, 3NY4A, 3NYCA, 3NYKA, 3NYQA, 3NYVA, 3NZ1A, 3NZLA, 3NZTA, 3O04A, 3O0FA, 3O0PA, 3O12A, 3O1GA, 3O22A, 3O26A, 3O2EA, 3O2TA, 3O3UN, 3O49A, 3O5EA, 3O5QA, 3O5XA, 3O6NA, 3O6PA, 3O70A, 3O7JA, 3O8MA, 3O8ZA, 3OBSA, 3OC6A, 3OC8A, 3OC9A, 3OCJA, 3OEBA, 3OEPA, 3OF7A, 3OFFA, 3OG2A, 3OH3A, 3OH8A, 3OHGA, 3OIOA, 3OIZA, 3OJ0A, 3OKPA, 3OMBA, 3ON1A, 3ONHA, 3ONSA, 3OO8A, 3OOIA, 3OOSA, 3OOUA, 3OPNA, 3OPTA, 3OQVA, 3OR5A, 3ORJA, 3ORKA, 3ORYA, 3OSDA, 3OSEA, 3OSTA, 3OTGA, 3OU2A, 3OUIA, 3OV5A, 3OVCA, 3OYVA, 3OZ2A, 3OZUA, 3P02A, 3P06A, 3P0BA, 3P1AA, 3P1GA, 3P1WA, 3P2TA, 3P2WA, 3P3CA, 3P3EA, 3P3GA, 3P3OA, 3P4HA, 3P4LA, 3P5PA, 3P5UA, 3P6DA, 3P6LA, 3P9NA, 3P9ZA, 3PA3A, 3PB6X, 3PBTA, 3PC3A, 3PDTA, 3PDWA, 3PE6A, 3PE7A, 3PEBA, 3PF0A, 3PF2A, 3PFYA, 3PGUA, 3PH2B, 3PHNA, 3PHSA, 3PJXA, 3PKBA, 3PKVA, 3PL1A, 3PLWA, 3PM2A, 3PMEA, 3PMMA, 3PMSA, 3PMTA, 3PN1A, 3PN4A, 3PNNA, 3PO0A, 3PO8A, 3POHA, 3POWA, 3POZA, 3PPVA, 3PQSA, 3PT1A, 3PT5A, 3PU5A, 3PUAA, 3PVFA, 3PX8X, 3PXLA, 3PYCA, 3PYWA, 3PZFA, 3PZWA, 3Q0IA, 3Q13A, 3Q1CA, 3Q1NA, 3Q27A, 3Q2BA, 3Q2EA, 3Q3UA, 3Q60A, 3Q6LA, 3Q71A, 3Q77A, 3Q7CA, 3Q7MA, 3Q8GA, 3QFKA, 3QH4A, 3QIOA, 3QJAA, 3QK2A, 3QKYA, 3QL6A, 3QLEA, 3QNMA, 3QOUA, 3QOWA, 3QSJA, 3QT9A, 3QUTA, 3QVPA, 3QWPA, 3QY7A, 3R0VA, 3R0XA, 3R20A, 3R2GA, 3R3QA, 3R3RA, 3R68A, 3R6DA, 3R6UA, 3R70A, 3R7ZA, 3R9RA, 3RABA, 3RARA, 3RAVA, 3RAYA, 3RC1A, 3RC3A, 3RCMA, 3RCQA, 3RD2A, 3RD5A, 3RFIA, 3RHBA, 3RHEA, 3RJUA, 3RJVA, 3RLEA, 3RLHA, 3RMQA, 3RNJA, 3RNLA, 3RNSA, 3RPWA, 3RQ0A, 3RQ4A, 3RQTA, 3RR2A, 3RR6A, 3RT2A, 3S01A, 3S1YA, 3S25A, 3S2JA, 3S4MA, 3S5QA, 3S5VA, 3S6BA, 3S6MA, 3S83A, 3S8SA, 3S8YA,

3S99A, 3S9JA, 3SC0A, 3SD7A, 3SG0A, 3SGGA, 3SHLA, 3SILA, 3ZRIA, 4PAHA, 5FIVA, 6FD1A, 7A3HA

A.1.7 Dataset of non-redundant crystal structures

The following protein structures are used to study the number of hydrogen bonds in low-resolution protein X-ray structure (c.f. Section 6.2.2).

1A041, 1A171, 1A251, 1A2Z1, 1A3A1, 1A3C1, 1A581, 1A621, 1A6J1, 1A6M1, 1A781, 1A791, 1AB81, 1AD61, 1AE91, 1AF51, 1AG91, 1AIH1, 1AJK1, 1AJO1, 1ALU1, 1ALV1, 1ALY1, 1AM21, 1AM71, 1AMX1, 1AN71, 1AN81, 1ANK1, 1ANU1, 1AOC1, 1APY1, 1APY2, 1AQC1, 1AQZ1, 1ARO2, 1ASS1, 1AT01, 1ATZ1, 1AU11, 1AUG1, 1AUN1, 1AUO1, 1AUY1, 1AV11, 1AVG3, 1AVO2, 1AVW2, 1AX81, 1AXI1, 1AYO1, 1AYZ1, 1AZS1, 1B061, 1B091, 1B0B1, 1B1C1, 1B1Z1, 1B241, 1B4A1, 1B561, 1B5L1, 1B661, 1B6E1, 1B781, 1B8D1, 1B931, 1B9O1, 1BAI1, 1BD71, 1BD81, 1BEB1, 1BED1, 1BGF1, 1BI72, 1BIN1, 1BJ71, 1BJF1, 1BKB1, 1BKZ1, 1BLE1, 1BM91, 1BO41, 1BOU1, 1BOW1, 1BQU1, 1BSM1, 1BTK1, 1BUN1, 1BUO1, 1BUU1, 1BVY2, 1BWV2, 1BX91, 1BXN2, 1BYE1, 1BYF1, 1BYR1, 1BZ41, 1BZY1, 1C1K1, 1C1L1, 1C251, 1C2Y1, 1C3A1, 1C3A2, 1C3G1, 1C3H1, 1C3M1, 1C411, 1C441, 1C521, 1C6V1, 1C7K1, 1C8S1, 1CBK1, 1CCW1, 1CCZ1, 1CD91, 1CD92, 1CEX1, 1CFZ1, 1CG51, 1CG52, 1CHD1, 1CJW1, 1CL72, 1CNT1, 1CNU1, 1COJ1, 1COZ1, 1CP91, 1CR51, 1CRB1, 1CUN1, 1CUO1, 1CV81, 1CXQ1, 1CYX1, 1CZT1, 1CZY1, 1D0B1, 1D1G1, 1D1Q1, 1D2O1, 1D2S1, 1D2Z2, 1D4O1, 1D8L1, 1D9C1, 1DBF1, 1DBW1, 1DBX1, 1DC91, 1DCF1, 1DD31, 1DD51, 1DDK1, 1DG61, 1DGW1, 1DHN1, 1DI01, 1DI61, 1DIV1, 1DJL1, 1DK01, 1DK71, 1DK81, 1DKG1, 1DKZ1, 1DLY1, 1DM11, 1DNL1, 1DO51, 1DOA2, 1DOV1, 1DOW1, 1DQE1, 1DQG1, 1DQI1, 1DS62, 1DU51, 1DUN1, 1DUS1, 1DV81, 1DVK1, 1DVO1, 1DVP1, 1DWK1, 1DXM1, 1DY21, 1DY51, 1DYO1, 1DYT1, 1DZ31, 1DZF1, 1DZK1, 1DZR1, 1E0R1, 1E1H2, 1E291, 1E2Y1, 1E4C1, 1E502, 1E571, 1E5K1, 1E5P1, 1E6C1, 1E6J3, 1E7L1, 1EAJ1, 1EAQ1, 1EAR1, 1EB61, 1EDY1, 1EE61, 1EEJ1, 1EER1, 1EEX3, 1EGI1, 1EH11, 1EH61, 1EHW1, 1EI71, 1EIA1, 1EJ01, 1EJ21, 1EJ81, 1EJB1, 1EJE1, 1EK01, 1EKR1, 1EL11, 1EL61, 1ELK1, 1ELR1, 1EM81,

1EM91, 1EMU1, 1ENF1, 1EP01, 1EP51, 1EPF1, 1ERN1, 1ESO1, 1ET91, 1ETE1,
1EU31, 1EUM1, 1EUW1, 1EV21, 1EV22, 1EVS1, 1EVX1, 1EW01, 1EW31, 1EX21,
1EX71, 1EXR1, 1EXS1, 1EXT1, 1EYH1, 1EYV1, 1EZ31, 1F081, 1F1C1, 1F1E1,
1F1M1, 1F2E1, 1F2K1, 1F2Q1, 1F2T1, 1F2T2, 1F2V1, 1F321, 1F351, 1F3U2,
1F3V1, 1F461, 1F4P1, 1F5J1, 1F5M1, 1F6B1, 1F6F1, 1F7S1, 1F8Y1, 1F971,
1F9A1, 1F9Z1, 1FAA1, 1FAZ1, 1FB11, 1FBT1, 1FD91, 1FDQ1, 1FEW1, 1FF31,
1FFV1, 1FGY1, 1FIO1, 1FIT1, 1FJJ1, 1FJR1, 1FL01, 1FLC2, 1FM02, 1FM41,
1FNL1, 1FPO1, 1FPZ1, 1FQI1, 1FRS1, 1FS01, 1FT51, 1FUK1, 1FUX1, 1FVG1,
1FVU1, 1FVU2, 1FW11, 1FX31, 1FXK3, 1FYV1, 1FYX1, 1FZD1, 1FZQ1, 1FZY1,
1G0S1, 1G121, 1G161, 1G1S1, 1G1T1, 1G2Q1, 1G3K1, 1G3P1, 1G431, 1G571,
1G5C1, 1G5Q1, 1G5T1, 1G5Z1, 1G661, 1G6G1, 1G731, 1G732, 1G831, 1G8I1,
1G8K2, 1GAK1, 1GCV2, 1GEC1, 1GEF1, 1GEN1, 1GGL1, 1GHE1, 1GJY1, 1GKA2,
1GM61, 1GM71, 1GME1, 1GMI1, 1GML1, 1GMU1, 1GN41, 1GNY1, 1GO31,
1GP01, 1GP71, 1GPQ1, 1GQA1, 1GQO1, 1GQV1, 1GR31, 1GRJ1, 1GS91, 1GSM1,
1GSU1, 1GTV1, 1GTZ1, 1GU21, 1GU31, 1GUI1, 1GUX1, 1GUX2, 1GV81, 1GVJ1,
1GWM1, 1GWY1, 1GXJ1, 1GXL1, 1GXS2, 1GY61, 1GY71, 1GYU1, 1GZ21,
1GZS2, 1H031, 1H051, 1H0H2, 1H1O1, 1H2E1, 1H2I1, 1H322, 1H4A1, 1H4G1,
1H6H1, 1H6P1, 1H7Z1, 1H971, 1H9M1, 1H9S1, 1HAR1, 1HDK1, 1HDM1, 1HDM2,
1HDO1, 1HE11, 1HF21, 1HGX1, 1HH11, 1HH81, 1HIA2, 1HJP1, 1HKQ1, 1HQ81,
1HQZ1, 1HRU1, 1HTN1, 1HTW1, 1HUF1, 1HUS1, 1HV51, 1HW51, 1HY51, 1HY71,
1HZT1, 1I0R1, 1I121, 1I1G1, 1I1R2, 1I2A1, 1I2H1, 1I2M1, 1I3C1, 1I3O1, 1I3O3,
1I3U1, 1I4U1, 1I511, 1I581, 1I5E1, 1I5G1, 1I5N1, 1I6A1, 1I6P1, 1I761, 1I7K1,
1I7Q2, 1I8A1, 1I8D1, 1I8L1, 1I9S1, 1IAM1, 1IAP1, 1IAR2, 1ICX1, 1ID01, 1ID11,
1IDP1, 1IFQ1, 1IHB1, 1IHK1, 1IIZ1, 1IJB1, 1IJT1, 1IJX1, 1IJY1, 1IK91, 1IKO1,
1IM51, 1IMJ1, 1IO01, 1IO21, 1IOO1, 1IPC1, 1IQ41, 1IQ61, 1IQQ1, 1IQV1, 1IRA1,
1IRU10, 1IRU11, 1IRU12, 1IRU13, 1IRU14, 1IRU8, 1IS11, 1IS31, 1ISP1, 1IT21,
1ITV1, 1IU11, 1IU81, 1IUH1, 1IUK1, 1IV21, 1IWD1, 1IWM1, 1IWP2, 1IWP3,
1IX91, 1IXL1, 1IXM1, 1IXR1, 1IYB1, 1IZ61, 1IZM1, 1J052, 1J241, 1J2R1, 1J301,
1J341, 1J342, 1J3A1, 1J3L1, 1J3M1, 1J3W1, 1J5U1, 1J5Y1, 1J6R1, 1J6W1, 1J6X1,
1J771, 1J7D1, 1J7G1, 1J831, 1J8H4, 1J8R1, 1J971, 1J981, 1J9B1, 1JAT1, 1JAT2,
1JAY1, 1JB010, 1JB04, 1JB06, 1JB31, 1JBE1, 1JBK1, 1JC41, 1JD11, 1JE51,

1JEB1, 1JEO1, 1JF81, 1JFM1, 1JFU1, 1JFX1, 1JFZ1, 1JGC1, 1JGJ1, 1JGS1,
1JH61, 1JHC1, 1JHF1, 1JHJ1, 1JHS1, 1JI41, 1JI51, 1JIA1, 1JIG1, 1JJV1, 1JKE1,
1JKG1, 1JKX1, 1JL11, 1JL31, 1JL71, 1JLJ1, 1JLT1, 1JLT2, 1JM11, 1JMV1,
1JMW1, 1JNR2, 1JOC1, 1JOG1, 1JPY1, 1JQL2, 1JRL1, 1JS91, 1JSD2, 1JSX1,
1JUG1, 1JUO1, 1JUV1, 1JVW1, 1JWI1, 1JWI2, 1JWQ1, 1JY51, 1JYH1, 1JYO1,
1JZN1, 1K041, 1K0Z1, 1K121, 1K1E1, 1K2E1, 1K2X1, 1K2X2, 1K3E1, 1K441,
1K4M1, 1K4N1, 1K4Z1, 1K5D2, 1K661, 1K681, 1K6D1, 1K6K1, 1K6Y1, 1K7J1,
1K8F1, 1K8K5, 1K8K6, 1K8K7, 1K941, 1KA91, 1KAC1, 1KAG1, 1KAM1, 1KCG2,
1KEX1, 1KF63, 1KG04, 1KGC1, 1KGD1, 1KHC1, 1KHI1, 1KHY1, 1KID1, 1KJN1,
1KKE1, 1KL91, 1KLL1, 1KLX1, 1KMI2, 1KMT1, 1KMV1, 1KN11, 1KN12, 1KNC1,
1KNG1, 1KNM1, 1KNQ1, 1KO61, 1KOE1, 1KQ61, 1KQF3, 1KQR1, 1KRQ1,
1KRR1, 1KSH1, 1KSH2, 1KT61, 1KTG1, 1KU91, 1KUG1, 1KUU1, 1KUX1, 1KWW1,
1KXG1, 1KXO1, 1KY31, 1KZ11, 1KZL1, 1L1N1, 1L1O3, 1L1Q1, 1L2H1, 1L3I1,
1L3K1, 1L4I1, 1L4Z2, 1L6P1, 1L6W1, 1L6X1, 1L6Z1, 1L7L1, 1L8B1, 1L9B4,
1LB31, 1LB61, 1LCS1, 1LE61, 1LF71, 1LFO1, 1LG71, 1LH01, 1LJ91, 1LK31,
1LKP1, 1LM41, 1LM51, 1LM61, 1LM83, 1LME1, 1LMI1, 1LN11, 1LO71, 1LP94,
1LPJ1, 1LQL1, 1LQS2, 1LQV1, 1LQY1, 1LR01, 1LR51, 1LSG1, 1LSS1, 1LU41,
1LVG1, 1LY11, 1LY21, 1M0D1, 1M0S1, 1M2O2, 1M3S1, 1M451, 1M481, 1M4I1,
1M4J1, 1M4U2, 1M4Y1, 1M551, 1M5Q1, 1M6D1, 1M701, 1M7B1, 1M7G1, 1M8N1,
1MA11, 1MBM1, 1MBQ1, 1MC21, 1MD61, 1MEO1, 1MEX2, 1MF71, 1MGT1,
1MH11, 1MHQ1, 1MHW1, 1MI81, 1MIJ1, 1MIX1, 1MJH1, 1MJN1, 1MJS1, 1MJU1,
1MK41, 1MKZ1, 1ML81, 1MOZ1, 1MP91, 1MQE1, 1MQV1, 1MUG1, 1MVL1,
1MVO1, 1MW51, 1MWW1, 1MY61, 1MZ41, 1MZB1, 1N081, 1N0E1, 1N0R1,
1N121, 1N1A1, 1N1C1, 1N1F1, 1N1Q1, 1N281, 1N2A1, 1N2F1, 1N2M1, 1N3Y1,
1N4X2, 1N5N1, 1N621, 1N711, 1N811, 1N8J1, 1N9P1, 1NAF1, 1NB21, 1NB91,
1NBC1, 1NBQ1, 1NCJ1, 1NE21, 1NEP1, 1NEZ3, 1NF32, 1NF91, 1NFD3, 1NFN1,
1NFV1, 1NG21, 1NG51, 1NG61, 1NGN1, 1NHK1, 1NHY1, 1NIG1, 1NJK1, 1NKI1,
1NKO1, 1NKR1, 1NKS1, 1NL11, 1NLN1, 1NN51, 1NNQ1, 1NOG1, 1NOX1, 1NP61,
1NP81, 1NPB1, 1NPE2, 1NQU1, 1NQZ1, 1NRJ1, 1NRJ2, 1NRZ1, 1NS51, 1NSJ1,
1NSL1, 1NT21, 1NTG1, 1NTV1, 1NU01, 1NUL1, 1NUN1, 1NWA1, 1NWP1, 1NWW1,
1NWZ1, 1NXH1, 1NXJ1, 1NXM1, 1NYE1, 1NYK1, 1NZI1, 1NZN1, 1O1X1, 1O221,

1O3U1, 1O4V1, 1O4W1, 1O501, 1O5H1, 1O5L1, 1O631, 1O6B1, 1O6D1, 1O731, 1O7D3, 1O8B1, 1O8P1, 1O8V1, 1O8X1, 1O911, 1O9R1, 1OA21, 1OA81, 1OAL1, 1O AQ1, 1OBO1, 1OCS1, 1OCX1, 1OCY1, 1OD31, 1OD61, 1OE81, 1OEJ1, 1OF52, 1OFS1, 1OGD1, 1OGY2, 1OH01, 1OH41, 1OHP1, 1OHT1, 1OHU1, 1OI41, 1OI61, 1OIX1, 1OJQ1, 1OKI1, 1OKR1, 1OLL1, 1OMR1, 1ON01, 1ON21, 1ONL1, 1OO01, 1OOH1, 1OQV1, 1OQW1, 1OR01, 1OR41, 1OR71, 1ORJ1, 1ORS3, 1ORU1, 1OSG1, 1OU01, 1OU91, 1OUT1, 1OUT2, 1OUW1, 1OV31, 1OW11, 1OW41, 1OXJ1, 1OXZ1, 1OY31, 1OZ61, 1OZ71, 1OZ72, 1OZ91, 1P0Z1, 1P1C1, 1P2F1, 1P2X1, 1P321, 1P3C1, 1P3R1, 1P3Y1, 1P4P1, 1P4T1, 1P4U1, 1P5S1, 1P5V2, 1P6O1, 1P7N1, 1P7O1, 1P901, 1P9M3, 1PAQ1, 1PBJ1, 1PBW1, 1PC61, 1PCZ1, 1PDK2, 1PDO1, 1PF51, 1PG52, 1PGV1, 1PI11, 1PIN1, 1PKH1, 1PKO1, 1PKU1, 1PL31, 1PM11, 1PMH1, 1PMT1, 1PNO1, 1PP01, 1PQW1, 1PSQ1, 1PT61, 1PU61, 1PUI1, 1PUO1, 1PVH2, 1PVM1, 1PVX1, 1PWA1, 1PWB1, 1PXV1, 1PYO1, 1PZ71, 1PZM1, 1PZS1, 1Q061, 1Q0U1, 1Q0X1, 1Q1F1, 1Q1U1, 1Q231, 1Q2Y1, 1Q3I1, 1Q402, 1Q421, 1Q4U1, 1Q5H1, 1Q5X1, 1Q5Z1, 1Q6O1, 1Q6W1, 1Q771, 1Q7H1, 1Q7L1, 1Q7R1, 1Q8C1, 1Q8M1, 1Q903, 1Q921, 1Q981, 1Q9C1, 1Q9U1, 1QAH1, 1QCS1, 1QD91, 1QDD1, 1QDL2, 1QEZ1, 1QF81, 1QF91, 1QFH1, 1QG31, 1QGV1, 1QHH1, 1QHQ1, 1QHV1, 1QHX1, 1QIB1, 1QIP1, 1QJ81, 1QJC1, 1QJP1, 1QKK1, 1QKR1, 1QMJ1, 1QMV1, 1QMY1, 1QNG1, 1QNX1, 1QOI1, 1QOZ1, 1QPV1, 1QQH1, 1QQL1, 1QQR1, 1QR41, 1QRF1, 1QSM1, 1QST1, 1Q TJ1, 1QTN1, 1QTO1, 1QU01, 1QU11, 1QUA1, 1QV11, 1QVC1, 1QVE1, 1QWD1, 1QWI1, 1QYN1, 1QZU1, 1R0D1, 1R0U1, 1R1M1, 1R291, 1R2Q1, 1R441, 1R451, 1R4V1, 1R551, 1R5A1, 1R5T1, 1R611, 1R621, 1R6N1, 1R8I1, 1R8N1, 1R8S1, 1R8S2, 1R9C1, 1R9W1, 1RCD1, 1RCU1, 1REG1, 1RFE1, 1RFS1, 1RFZ1, 1RG81, 1RHF1, 1RHY1, 1RIE1, 1RIW2, 1RJ81, 1RKT1, 1RKU1, 1RL21, 1RL41, 1RL61, 1RLH1, 1RLI1, 1RLJ1, 1RLW1, 1RM63, 1RM81, 1RNF1, 1RO21, 1RO51, 1ROC1, 1RRA1, 1RRE1, 1RRP2, 1RSS1, 1RTT1, 1RTW1, 1RTY1, 1RUT1, 1RW61, 1RXD1, 1RXQ1, 1RY91, 1RYB1, 1RYL1, 1RYP10, 1RYP11, 1RYP12, 1RYP8, 1RZ11, 1RZ31, 1RZN1, 1S0P1, 1S211, 1S281, 1S2D1, 1S2J1, 1S2X1, 1S351, 1S3J1, 1S3M1, 1S3Q1, 1S3Z1, 1S4C1, 1S4K1, 1S551, 1S571, 1S5A1, 1S5U1, 1S691, 1S6C1, 1S7I1, 1S7K1, 1S7M1, 1S8N1, 1S961, 1S991, 1S9U1, 1SB21, 1SB22, 1SBQ1, 1SBZ1, 1SC01, 1SCT1,

1SCT2, 1SD41, 1SDI1, 1SDO1, 1SEI1, 1SEN1, 1SGL1, 1SGM1, 1SGW1, 1SH81,
1SHU1, 1SHX1, 1SJ71, 1SJN1, 1SJW1, 1SJY1, 1SK41, 1SK71, 1SL41, 1SL61,
1SLU1, 1SMB1, 1SPG1, 1SPV1, 1SQL1, 1SQU1, 1SQW1, 1SR41, 1SR43, 1SRA1,
1SRV1, 1SS41, 1SU01, 1SU11, 1SUR1, 1SVI1, 1SVP1, 1SX71, 1SXR1, 1SXV1,
1SY63, 1SZ71, 1SZ91, 1SZB1, 1SZH1, 1T0A1, 1T0I1, 1T151, 1T2W1, 1T351,
1T3B1, 1T3G1, 1T3Q1, 1T3W1, 1T3Y1, 1T441, 1T4W1, 1T561, 1T571, 1T5I1,
1T621, 1T6A1, 1T6N1, 1T6S1, 1T7S1, 1T821, 1T9F1, 1T9M1, 1TC11, 1TC51,
1TCR1, 1TDQ2, 1TE12, 1TEV1, 1TF11, 1TFE1, 1TFU1, 1TH81, 1TIQ1, 1TJL1,
1TJN1, 1TJO1, 1TK91, 1TKS1, 1TLJ1, 1TLQ1, 1TMK1, 1TN41, 1TO01, 1TO41,
1TP61, 1TP91, 1TQ81, 1TT81, 1TU11, 1TU71, 1TU91, 1TUA1, 1TUE1, 1TUE2,
1TUH1, 1TVG1, 1TW01, 1TW91, 1TWF6, 1TWF7, 1TWU1, 1TX41, 1TX91,
1TXJ1, 1TXL1, 1TY01, 1TYJ1, 1TZA1, 1TZV1, 1U0A1, 1U111, 1U141, 1U1Z1,
1U201, 1U2P1, 1U531, 1U551, 1U5P1, 1U5T2, 1U611, 1U691, 1U6L1, 1U6M1,
1U6T1, 1U791, 1U7I1, 1U7K1, 1U7P1, 1U8S1, 1U8W1, 1U8Z1, 1U9D1, 1UAX1,
1UB31, 1UC31, 1UCD1, 1UCT1, 1UEB1, 1UES1, 1UF91, 1UFB1, 1UFH1, 1UFR1,
1UFY1, 1UGN1, 1UGP1, 1UGX1, 1UHB1, 1UHK1, 1UHN1, 1UI01, 1UI51, 1UJC1,
1UKF1, 1UKK1, 1ULI2, 1ULK1, 1ULY1, 1UMR1, 1UMR2, 1UN21, 1UNA1, 1UOW1,
1UPT1, 1UQR1, 1USC1, 1USP1, 1USU2, 1UT11, 1UT71, 1UTY1, 1UUL1, 1UUN1,
1UUY1, 1UUZ1, 1UV01, 1UVQ1, 1UVQ2, 1UVX1, 1UW61, 1UW71, 1UWF1,
1UWW1, 1UWZ1, 1UXO1, 1UXX1, 1UXZ1, 1UY41, 1UZK1, 1UZX1, 1V0A1,
1V1O1, 1V2A1, 1V2B1, 1V2X1, 1V371, 1V3W1, 1V3Y1, 1V4P1, 1V4X2, 1V5X1,
1V771, 1V7L1, 1V7M3, 1V7P1, 1V7P2, 1V7P3, 1V7R1, 1V8C1, 1V8P1, 1V8Y1,
1V961, 1V9C1, 1V9S1, 1V9T1, 1VAJ1, 1VCD1, 1VCH1, 1VCT1, 1VD31, 1VDM1,
1VDR1, 1VE01, 1VE41, 1VEC1, 1VET1, 1VFR1, 1VG11, 1VG81, 1VGG1, 1VGJ1,
1VH01, 1VH51, 1VH91, 1VHB1, 1VHM1, 1VHS1, 1VHT1, 1VHU1, 1VHX1, 1VHZ1,
1VI01, 1VI41, 1VI61, 1VI71, 1VIA1, 1VIM1, 1VIP1, 1VIU1, 1VIW2, 1VJE1,
1VJF1, 1VJL1, 1VJN1, 1VJX1, 1VK21, 1VK91, 1VKA1, 1VKB1, 1VKC1, 1VKF1,
1VKI1, 1VKK1, 1VKW1, 1VL71, 1VL91, 1VLA1, 1VLG1, 1VLH1, 1VLM1, 1VMD1,
1VMF1, 1VMH1, 1VMJ1, 1VP21, 1VP61, 1VP81, 1VPH1, 1VPM1, 1VQ21, 1VQT1,
1VR31, 1VR71, 1VR81, 1VR91, 1VRA1, 1VRA2, 1VSC1, 1VSR1, 1VYF1, 1VYK1,
1VYQ1, 1W0H1, 1W0N1, 1W151, 1W1H1, 1W2W1, 1W2W2, 1W301, 1W4R1,

1W4S1, 1W6N1, 1W741, 1W7J2, 1W7W1, 1W8A1, 1W8I1, 1W941, 1W9A1, 1W9E1, 1W9S1, 1WA31, 1WA51, 1WBE1, 1WC21, 1WC31, 1WCK1, 1WCM4, 1WCU1, 1WD51, 1WDC2, 1WDC3, 1WDD2, 1WDJ1, 1WDV1, 1WE01, 1WEH1, 1WEK1, 1WFX1, 1WHI1, 1WJ91, 1WKA1, 1WKC1, 1WKJ1, 1WKO1, 1WKP1, 1WKQ1, 1WL81, 1WLF1, 1WLJ1, 1WLT1, 1WMS1, 1WMU1, 1WMX1, 1WMZ1, 1WNA1, 1WNY1, 1WO81, 1WOJ1, 1WOL1, 1WOZ1, 1WPB1, 1WPN1, 1WPV1, 1WPX2, 1WQG1, 1WQL2, 1WRJ1, 1WRM1, 1WS01, 1WS61, 1WU31, 1WUB1, 1WUR1, 1WVH1, 1WVQ1, 1WVT1, 1WWI1, 1WWM1, 1WWR1, 1WWZ1, 1WY11, 1WY61, 1WY71, 1WYY1, 1WZD1, 1WZV1, 1X0P1, 1X121, 1X251, 1X271, 1X3K1, 1X3S1, 1X461, 1X6M1, 1X6O1, 1X751, 1X821, 1X8Q1, 1X911, 1X921, 1X941, 1X9F1, 1X9F2, 1X9F3, 1X9F4, 1X9G1, 1X9Z1, 1XB41, 1XCC1, 1XCV1, 1XE71, 1XEB1, 1XEO1, 1XFS1, 1XG03, 1XG22, 1XGW1, 1XHD1, 1XHF1, 1XHK1, 1XHN1, 1XI31, 1XIQ1, 1XIY1, 1XJC1, 1XJT1, 1XJU1, 1XK51, 1XK91, 1XKI1, 1XKP3, 1XKR1, 1XL31, 1XM51, 1XM71, 1XME2, 1XMP1, 1XNK1, 1XO51, 1XPH1, 1XPJ1, 1XQ41, 1XQI1, 1XRE1, 1XRG1, 1XRK1, 1XS11, 1XSO1, 1XSQ1, 1XT01, 1XT51, 1XTM1, 1XTQ1, 1XTT1, 1XTY1, 1XU11, 1XUV1, 1XVH1, 1XVQ1, 1XVS1, 1XVW1, 1XWM1, 1XWV1, 1XWW1, 1XX61, 1XX71, 1XXQ1, 1XZO1, 1XZP2, 1Y071, 1Y0B1, 1Y0G1, 1Y0K1, 1Y121, 1Y131, 1Y141, 1Y142, 1Y1L1, 1Y1X1, 1Y231, 1Y2T1, 1Y432, 1Y551, 1Y5E1, 1Y5H1, 1Y601, 1Y631, 1Y6H1, 1Y6K2, 1Y6L1, 1Y7M1, 1Y7O1, 1Y7R1, 1Y7U1, 1Y801, 1Y881, 1Y8X1, 1Y931, 1Y9I1, 1Y9K1, 1Y9Q1, 1Y9W1, 1YAC1, 1YAV1, 1YAZ1, 1YB01, 1YB31, 1YBT1, 1YBY1, 1YC61, 1YCK1, 1YDG1, 1YDM1, 1YE81, 1YEM1, 1YF91, 1YFU1, 1YFZ1, 1YG21, 1YGH1, 1YH21, 1YHU1, 1YHU2, 1YHU3, 1YHU4, 1YIO1, 1YJ71, 1YJG1, 1YK31, 1YK91, 1YKI1, 1YLK1, 1YLL1, 1YLM1, 1YM31, 1YMP1, 1YN91, 1YNB1, 1YO71, 1YOA1, 1YOB1, 1YOC1, 1YP11, 1YPH2, 1YPQ1, 1YPR1, 1YPY1, 1YQ11, 1YQ51, 1YQ81, 1YQF1, 1YR01, 1YRE1, 1YRK1, 1YRO1, 1YRV1, 1YSP1, 1YSQ1, 1YSR1, 1YTL1, 1YTQ1, 1Y TZ2, 1YUD1, 1YUK2, 1YUZ1, 1YVK1, 1YW51, 1YWM1, 1YWQ1, 1YY71, 1YZ11, 1YZ41, 1YZ71, 1YZF1, 1YZH1, 1YZV1, 1Z061, 1Z081, 1Z0A1, 1Z0F1, 1Z0J1, 1Z0W1, 1Z0X1, 1Z1S1, 1Z1Y1, 1Z2A1, 1Z2U1, 1Z2W1, 1Z3A1, 1Z3Q1, 1Z3U1, 1Z4E1, 1Z4R1, 1Z541, 1Z5Y2, 1Z671, 1Z6B1, 1Z6G1, 1Z6M1, 1Z6N1, 1Z6O1, 1Z6O2, 1Z8H1, 1Z911, 1Z941, 1Z9L1, 1ZAV1, 1ZB91, 1ZBD1,

1ZBD2, 1ZBF1, 1ZBO1, 1ZBX1, 1ZBX2, 1ZCE1, 1ZD01, 1ZD71, 1ZD91, 1ZDE1, 1ZDN1, 1ZDR1, 1ZE32, 1ZGZ1, 1ZH21, 1ZHV1, 1ZIN1, 1ZIS1, 1ZJR1, 1ZK51, 1ZK81, 1ZKC1, 1ZKI1, 1ZKO1, 1ZL91, 1ZLB1, 1ZMS1, 1ZN81, 1ZNP1, 1ZNW1, 1ZO21, 1ZOF1, 1ZOQ1, 1ZP61, 1ZPS1, 1ZQ71, 1ZR31, 1ZS31, 1ZS61, 1ZSO1, 1ZTD1, 1ZUH1, 1ZUJ1, 1ZUO1, 1ZV41, 1ZVF1, 1ZVP1, 1ZWY1, 1ZWZ1, 1ZX81, 1ZXJ1, 1ZXQ1, 1ZXZ1, 1ZY21, 1ZYE1, 1ZYN1, 1ZYO1, 1ZZW1, 2A065, 2A0J1, 2A151, 2A191, 2A1I1, 2A1V1, 2A221, 2A251, 2A2K1, 2A2L1, 2A2N1, 2A2Q1, 2A2R1, 2A2U1, 2A351, 2A381, 2A4D1, 2A4E1, 2A4V1, 2A502, 2A5D2, 2A5J1, 2A5L1, 2A5Y1, 2A611, 2A671, 2A6A1, 2A6M1, 2A6P1, 2A721, 2A7B1, 2A7L1, 2A8E1, 2A8N1, 2A9S1, 2A9U1, 2A9V1, 2AAL1, 2AAN1, 2AAO1, 2AB01, 2ABK1, 2ABL1, 2ACA1, 2ACF1, 2AE61, 2AEE1, 2AEN1, 2AEW1, 2AF01, 2AG41, 2AH51, 2AH61, 2AHD1, 2AJ61, 2AJ71, 2AJG1, 2AK44, 2AKP1, 2AMH1, 2AMJ1, 2ANX1, 2AO71, 2AP31, 2APL1, 2AQM1, 2AQP1, 2AR11, 2ARC1, 2ARH1, 2ARK1, 2ARP2, 2AS91, 2ASF1, 2ATR1, 2ATV1, 2ATZ1, 2AU51, 2AU71, 2AUK1, 2AUW1, 2AV41, 2AWF1, 2AXW1, 2AZ31, 2AZE1, 2AZN1, 2AZW1, 2B061, 2B0A1, 2B0C1, 2B0P1, 2B0R1, 2B0V1, 2B181, 2B1K1, 2B1L1, 2B1X2, 2B2A1, 2B2Y1, 2B331, 2B341, 2B3R1, 2B592, 2B5G1, 2B5I2, 2B5I3, 2B5I4, 2B671, 2B6C1, 2B711, 2B7K1, 2B821, 2B981, 2B9C1, 2B9U1, 2BA11, 2BBA1, 2BBR1, 2BC31, 2BCG2, 2BCM1, 2BDD1, 2BDR1, 2BDT1, 2BDZ1, 2BEC1, 2BEM1, 2BEP1, 2BFW1, 2BH41, 2BHG1, 2BHM1, 2BHN1, 2BJ01, 2BJ71, 2BJN1, 2BJO1, 2BK91, 2BKG1, 2BKM1, 2BKR1, 2BL02, 2BL03, 2BL21, 2BM31, 2BM51, 2BME1, 2BML1, 2BMM1, 2BMO2, 2BMV1, 2BMX1, 2BNG1, 2BNL1, 2BNM1, 2BOU1, 2BOV1, 2BPD1, 2BQX1, 2BRJ1, 2BSF1, 2BSJ1, 2BUE1, 2BUR1, 2BV11, 2BV61, 2BVU1, 2BWJ1, 2BWQ1, 2BYC1, 2BYO1, 2BZ11, 2BZV1, 2C081, 2C0A1, 2C0J2, 2C0K1, 2C0N1, 2C0Z1, 2C1D2, 2C1S1, 2C211, 2C2I1, 2C2P1, 2C2U1, 2C351, 2C352, 2C3B1, 2C411, 2C4B1, 2C4J1, 2C4W1, 2C6U1, 2C711, 2C7J2, 2C8E1, 2C8S1, 2C921, 2C951, 2C9W1, 2CAL1, 2CAR1, 2CAY1, 2CB31, 2CBY1, 2CC01, 2CC31, 2CCJ1, 2CCL1, 2CCM1, 2CDN1, 2CDO1, 2CE21, 2CF71, 2CFA1, 2CFX1, 2CG41, 2CH81, 2CH91, 2CHC1, 2CHD1, 2CHP1, 2CIH1, 2CIO1, 2CIU1, 2CJ41, 2CJL1, 2CJS1, 2CJT1, 2CJW1, 2CKK1, 2CKZ2, 2CLB1, 2CLY2, 2CM51, 2CMT1, 2CNT1, 2CO31, 2CQZ1, 2CT91, 2CU51, 2CU91, 2CUA1, 2CV81, 2CVB1, 2CVD1, 2CVE1, 2CVH1, 2CVL1, 2CW91,

2CWD1, 2CWO1, 2CWQ1, 2CWZ1, 2CX11, 2CX41, 2CX71, 2CXC1, 2CXH1, 2CXX1, 2CY21, 2CY51, 2CYE1, 2CYY1, 2CZD1, 2CZQ1, 2CZV2, 2D0W1, 2D161, 2D1P1, 2D281, 2D301, 2D371, 2D3E1, 2D3V1, 2D3Y1, 2D481, 2D4G1, 2D4P1, 2D4R1, 2D4U1, 2D591, 2D5K1, 2D5M1, 2D6M1, 2D6Y1, 2D742, 2D7V1, 2DBB1, 2DBO1, 2DC31, 2DC41, 2DCH1, 2DD71, 2DDB1, 2DDZ1, 2DFB1, 2DG52, 2DG71, 2DG81, 2DJF2, 2DKO1, 2DOK1, 2DP91, 2DPX1, 2DQA1, 2DRE1, 2DRU1, 2DSC1, 2DST1, 2DT41, 2DT51, 2DT91, 2DTJ1, 2DUM1, 2DVK1, 2DWK1, 2DWY1, 2DX61, 2DXA1, 2DXQ1, 2DY01, 2DYA1, 2DYC1, 2DYI1, 2DYK1, 2DYQ1, 2DYY1, 2E0T1, 2E1B1, 2E1M3, 2E2C1, 2E2D2, 2E2E1, 2E3U1, 2E3X2, 2E3X3, 2E471, 2E551, 2E561, 2E5Y1, 2E6M1, 2E741, 2E742, 2E744, 2E851, 2E8B1, 2E8E1, 2E9X1, 2E9X2, 2E9X3, 2EA31, 2EB11, 2EBJ1, 2ECU1, 2ED61, 2EDM1, 2EEN1, 2EEY1, 2EF71, 2EFH2, 2EFR1, 2EGJ1, 2EGZ1, 2EH31, 2EHB1, 2EHB2, 2EHG1, 2EHJ1, 2EHO1, 2EHP1, 2EIF1, 2EIS1, 2EJ81, 2EJB1, 2EJN1, 2EJX1, 2EKD1, 2EKM1, 2EKN1, 2ELA1, 2EMQ1, 2ENG1, 2EO01, 2EO41, 2ERF1, 2ERV1, 2ERX1, 2ERY1, 2ESB1, 2ESR1, 2ESU1, 2ET11, 2ETD1, 2ETH1, 2EUI1, 2EVE1, 2EVV1, 2EW01, 2EW11, 2EW51, 2EWC1, 2EXU1, 2F011, 2F061, 2F071, 2F0C1, 2F1D1, 2F1F1, 2F1L1, 2F1R1, 2F1V1, 2F1W1, 2F221, 2F231, 2F2E1, 2F2F3, 2F2L1, 2F2L2, 2F2O1, 2F3L1, 2F3X1, 2F421, 2F461, 2F4E1, 2F4I1, 2F4P1, 2F4W1, 2F4Z1, 2F591, 2F5G1, 2F5J1, 2F621, 2F6E1, 2F6I1, 2F6S1, 2F731, 2F861, 2F8A1, 2F8V2, 2F991, 2F9F1, 2F9H1, 2F9L1, 2F9S1, 2F9Z2, 2FA11, 2FA51, 2FB51, 2FBD1, 2FBE1, 2FBH1, 2FBI1, 2FBK1, 2FBL1, 2FBN1, 2FC31, 2FCA1, 2FCB1, 2FCK1, 2FCL1, 2FCO1, 2FD51, 2FDB1, 2FE11, 2FE31, 2FE71, 2FEX1, 2FF31, 2FFQ1, 2FFS1, 2FG11, 2FG51, 2FG91, 2FGC1, 2FGS1, 2FH51, 2FH52, 2FHD1, 2FHE1, 2FHP1, 2FHQ1, 2FI11, 2FIA1, 2FIW1, 2FJ81, 2FJC1, 2FJR1, 2FK51, 2FK91, 2FKB1, 2FKK1, 2FL41, 2FLH1, 2FLI1, 2FM81, 2FM91, 2FMY1, 2FOM2, 2FP11, 2FP72, 2FPH1, 2FPN1, 2FPO1, 2FPR1, 2FQ41, 2FR21, 2FR51, 2FRE1, 2FRH1, 2FS21, 2FSR1, 2FSU1, 2FSX1, 2FU01, 2FUF1, 2FUK1, 2FUL1, 2FUP1, 2FUR1, 2FVV1, 2FW51, 2FXA1, 2FXM1, 2FXT1, 2FY61, 2FYG1, 2FYQ1, 2FYX1, 2FZC2, 2FZF1, 2FZL1, 2FZP1, 2FZS1, 2G031, 2G0B1, 2G1U1, 2G2C1, 2G2D1, 2G2S2, 2G382, 2G3A1, 2G3B1, 2G3V1, 2G3W1, 2G3Y1, 2G4D1, 2G641, 2G6B1, 2G6Y1, 2G6Z1, 2G7B1, 2G7G1, 2G7S1, 2G841, 2GA01, 2GAG3, 2GAN1, 2GAX1,

2GBB1, 2GBW2, 2GBZ1, 2GC74, 2GCO1, 2GD51, 2GD91, 2GDR1, 2GE31, 2GEB1, 2GEC1, 2GED1, 2GEE1, 2GEF1, 2GEN1, 2GEX1, 2GEY1, 2GEZ1, 2GEZ2, 2GF01, 2GF61, 2GF91, 2GFG1, 2GFN1, 2GGM1, 2GGT1, 2GGV2, 2GGZ1, 2GHT1, 2GHV1, 2GI71, 2GIA1, 2GIA2, 2GIX1, 2GIY1, 2GJ81, 2GJD1, 2GJV1, 2GKG1, 2GKM1, 2GKP1, 2GL91, 2GL92, 2GLL1, 2GLZ1, 2GM31, 2GM51, 2GM61, 2GMW1, 2GMY1, 2GO71, 2GOI1, 2GOJ1, 2GPC1, 2GQQ1, 2GRC1, 2GRJ1, 2GRR2, 2GS41, 2GS51, 2GS91, 2GSL1, 2GTL5, 2GTL6, 2GTL7, 2GU31, 2GUD1, 2GUH1, 2GUI1, 2GUJ1, 2GUX1, 2GVI1, 2GW42, 2GWF1, 2GWM1, 2GXC1, 2GXQ1, 2GYQ1, 2GZ41, 2GZB1, 2GZQ1, 2H011, 2H091, 2H0B1, 2H0R1, 2H0U1, 2H171, 2H1C1, 2H1E1, 2H1T1, 2H291, 2H2T1, 2H2Y1, 2H301, 2H4C2, 2H4U1, 2H571, 2H5C1, 2H5N1, 2H661, 2H6L1, 2H6R1, 2H7J1, 2H883, 2H8E1, 2H8Q1, 2H921, 2HA81, 2HAF1, 2HAL1, 2HBO1, 2HCM1, 2HCU1, 2HD01, 2HD91, 2HDO1, 2HEV1, 2HEW1, 2HEY2, 2HFN1, 2HFT1, 2HHG1, 2HHZ1, 2HI21, 2HIY1, 2HJV1, 2HKU1, 2HKV1, 2HL01, 2HLJ1, 2HLV1, 2HLY1, 2HMOV1, 2HNF1, 2HNG1, 2HP71, 2HPK1, 2HPS1, 2HPV1, 2HQ41, 2HQ61, 2HQ71, 2HQ91, 2HQV1, 2HR21, 2HR31, 2HRA1, 2HRV1, 2HSB1, 2HSJ1, 2HSN1, 2HT61, 2HTD1, 2HTH2, 2HTI1, 2HU91, 2HUE1, 2HUH1, 2HUI1, 2HUP1, 2HUR1, 2HVW1, 2HW21, 2HWJ1, 2HX01, 2HX51, 2HXP1, 2HXS1, 2HY51, 2HY52, 2HYJ1, 2HYT1, 2HYZ1, 2HZA1, 2HZM1, 2HZQ1, 2I021, 2I0M1, 2I101, 2I151, 2I1S1, 2I2L1, 2I2Q1, 2I2R1, 2I2W1, 2I321, 2I3Y1, 2I461, 2I511, 2I5E1, 2I5H1, 2I6C1, 2I6H1, 2I6J1, 2I6X1, 2I741, 2I791, 2I7A1, 2I7D1, 2I7H1, 2I881, 2I8B1, 2I8D1, 2I8G1, 2I8T1, 2I9A1, 2I9C1, 2I9D1, 2I9W1, 2IA01, 2IA11, 2IAA3, 2IAB1, 2IAF1, 2IAL1, 2IB01, 2IBD1, 2IBG1, 2ICG1, 2ICW4, 2ID11, 2ID61, 2IDC1, 2IDG1, 2IDR1, 2IDX1, 2IEL1, 2IEP1, 2IF21, 2IF61, 2IF71, 2IFA1, 2IFR1, 2IFT1, 2IG31, 2IG61, 2IG81, 2IGI1, 2IGS1, 2IGV1, 2IHD1, 2IIH1, 2III1, 2IJ01, 2IJ91, 2IJC1, 2IJQ1, 2IKB1, 2IKK1, 2IL11, 2IL51, 2IM81, 2IMF1, 2IMG1, 2IMJ1, 2IML1, 2IMS1, 2IMZ1, 2IN01, 2IN31, 2IN51, 2IO51, 2IOJ1, 2ION1, 2IPQ1, 2IQC1, 2IQI1, 2IQY1, 2IRP1, 2IS31, 2IS81, 2IS91, 2ISB1, 2ISY1, 2IT21, 2IT91, 2ITB1, 2ITE1, 2IU11, 2IU51, 2IVF3, 2IWL1, 2IWR1, 2IWX1, 2IXC1, 2IXK1, 2IY41, 2IYK1, 2IYV1, 2IZ61, 2IZV1, 2J0V1, 2J121, 2J161, 2J1A1, 2J1L1, 2J1V1, 2J221, 2J2J1, 2J3R2, 2J3T3, 2J3T4, 2J3W1, 2J3W2, 2J411, 2J431, 2J441, 2J491, 2J4B1, 2J4X1, 2J592, 2J671, 2J6A1, 2J6P1, 2J871, 2J882, 2J8H1, 2J8K1,

2J8M1, 2J8W1, 2J961, 2J9L1, 2J9R1, 2JA91, 2JAA1, 2JAB1, 2JAQ1, 2JBA1, 2JBX1, 2JCB1, 2JCQ1, 2JD61, 2JDA1, 2JDC1, 2JE01, 2JE31, 2JEK1, 2JEX1, 2JG61, 2JGB1, 2JGN1, 2JH21, 2JHE1, 2JI21, 2JJ71, 2JK11, 2JK91, 2JKD1, 2JKG1, 2JKY1, 2JLM1, 2LIS1, 2NL91, 2NN51, 2NN69, 2NNN1, 2NNU1, 2NOG1, 2NP31, 2NP51, 2NQ31, 2NR41, 2NR71, 2NR91, 2NRH1, 2NRK1, 2NRL1, 2NRQ1, 2NRR1, 2NRT1, 2NS61, 2NS91, 2NSA1, 2NSZ1, 2NT21, 2NTE1, 2NUJ1, 2NUT3, 2NV01, 2NV41, 2NVN1, 2NVU3, 2NW21, 2NWF1, 2NWI1, 2NWU1, 2NX21, 2NX41, 2NX81, 2NXP1, 2NXY2, 2NYB1, 2NYC1, 2NYI1, 2NYT1, 2NYU1, 2NYX1, 2NZJ1, 2O0115, 2O031, 2O081, 2O091, 2O161, 2O1A1, 2O1C1, 2O1Q1, 2O271, 2O281, 2O2A1, 2O2X1, 2O392, 2O3A1, 2O3B2, 2O4D1, 2O4X1, 2O521, 2O5F1, 2O5H1, 2O5U1, 2O6F1, 2O6L1, 2O6P1, 2O6R1, 2O6S1, 2O6W1, 2O701, 2O7M1, 2O7T1, 2O8I1, 2O8Q1, 2O951, 2O991, 2O9X1, 2OA21, 2OAF1, 2OAR1, 2OB01, 2OB41, 2OB51, 2OBA1, 2OBB1, 2OBH1, 2OBI1, 2OBX1, 2OC61, 2OCY1, 2ODA1, 2ODE2, 2OEB1, 2OER1, 2OFC1, 2OFK1, 2OFZ1, 2OGF1, 2OGG1, 2OGI1, 2OGK1, 2OH11, 2OH31, 2OHD1, 2OHW1, 2OI81, 2OIF1, 2OIK1, 2OIW1, 2OIX1, 2OJ41, 2OJ61, 2OK31, 2OKF1, 2OKM1, 2OKQ1, 2OKU1, 2OKV1, 2OL11, 2OL51, 2OLM1, 2OLP1, 2OLW1, 2OMD1, 2OML1, 2ON51, 2ON71, 2ONF1, 2ONU1, 2OOI1, 2OOJ1, 2OOK1, 2OP61, 2OPE1, 2OPI1, 2OPL1, 2OQ01, 2OQM1, 2OQO1, 2OQT1, 2ORW1, 2OS01, 2OSA1, 2OSO1, 2OSS1, 2OT91, 2OTE1, 2OTM1, 2OTO1, 2OTX1, 2OU31, 2OU51, 2OU61, 2OUA1, 2OUC1, 2OUD1, 2OUG1, 2OUW1, 2OV71, 2OV91, 2OVI1, 2OVJ1, 2OVX1, 2OWA1, 2OWP1, 2OWR1, 2OX11, 2OX61, 2OX71, 2OX91, 2OYO1, 2OZ61, 2OZH1, 2OZN1, 2OZN2, 2P0B1, 2P0K1, 2P0N1, 2P0S1, 2P0T1, 2P121, 2P141, 2P191, 2P1A1, 2P1J1, 2P1Z1, 2P221, 2P223, 2P231, 2P251, 2P282, 2P2O1, 2P2U1, 2P311, 2P381, 2P391, 2P3K1, 2P3P1, 2P452, 2P4K1, 2P4V1, 2P4W1, 2P571, 2P5D1, 2P5Q1, 2P5S1, 2P5V1, 2P651, 2P6C1, 2P6H1, 2P6N1, 2P6W1, 2P6Y1, 2P7O1, 2P7V1, 2P841, 2P8G1, 2P8J1, 2P8T1, 2P971, 2P9J1, 2P9M1, 2P9O1, 2PA71, 2PAG1, 2PAQ1, 2PB91, 2PBD2, 2PBR1, 2PC11, 2PC61, 2PCN1, 2PCS1, 2PDO1, 2PDR1, 2PEZ1, 2PFB1, 2PFC1, 2PFE1, 2PFI1, 2PFS1, 2PFX1, 2PGC1, 2PH01, 2PHP1, 2PI22, 2PIE1, 2PIH1, 2PIM1, 2PKH1, 2PKP1, 2PL11, 2PLG1, 2PLR1, 2PLU1, 2PLW1, 2PMA1, 2PMP1, 2PN01, 2PN21, 2PN61, 2PN81, 2PNL1, 2PNS1, 2POE1, 2PPT1, 2PPW1, 2PQ51, 2PQ71, 2PQF1, 2PQQ1,

2PQV1, 2PR11, 2PR51, 2PR71, 2PRR1, 2PRV1, 2PT51, 2PTH1, 2PTM1, 2PU31, 2PV41, 2PVQ1, 2PWJ1, 2PWQ1, 2PXX1, 2PY21, 2PY81, 2PYB1, 2PYT1, 2PZH1, 2PZZ1, 2Q001, 2Q031, 2Q041, 2Q051, 2Q0S1, 2Q0V1, 2Q0Y1, 2Q161, 2Q1Z1, 2Q1Z2, 2Q221, 2Q241, 2Q2B1, 2Q2G1, 2Q2L1, 2Q371, 2Q3H1, 2Q3L1, 2Q3Q1, 2Q3S1, 2Q3T1, 2Q3X1, 2Q471, 2Q481, 2Q4D1, 2Q4F1, 2Q4M1, 2Q4N1, 2Q4O1, 2Q4Q1, 2Q4V1, 2Q5C1, 2Q5W1, 2Q5X1, 2Q6M1, 2Q781, 2Q7A1, 2Q7B1, 2Q7M1, 2Q8O1, 2Q9K1, 2Q9Q3, 2Q9R1, 2QA91, 2QAC1, 2QAS1, 2QBW1, 2QBX1, 2QC12, 2QC91, 2QCK1, 2QDL1, 2QE91, 2QEA1, 2QEB1, 2QEC1, 2QEJ1, 2QEU1, 2QF41, 2QF91, 2QFA1, 2QFD1, 2QFE1, 2QG31, 2QG81, 2QGG1, 2QGU1, 2QGV1, 2QGX1, 2QH91, 2QHE1, 2QHK1, 2QHR2, 2QHT1, 2QIH1, 2QIM1, 2QIP1, 2QIY1, 2QJ21, 2QJW1, 2QJY3, 2QJZ1, 2QKI2, 2QKL1, 2QKO1, 2QKP1, 2QL31, 2QL81, 2QL91, 2QLC1, 2QLK1, 2QLP1, 2QLV1, 2QM62, 2QMH1, 2QML1, 2QMM1, 2QMO1, 2QN41, 2QND1, 2QNG1, 2QNI1, 2QNL1, 2QNT1, 2QO41, 2QOG1, 2QOR1, 2QPV1, 2QPW1, 2QQ21, 2QQ41, 2QQY1, 2QQZ1, 2QR31, 2QRW1, 2QS71, 2QS91, 2QSD1, 2QSJ1, 2QSR1, 2QSV1, 2QSX1, 2QT11, 2QT51, 2QTP1, 2QTTQ1, 2QTV2, 2QUD1, 2QUO1, 2QV01, 2QV81, 2QVG1, 2QVK1, 2QWT1, 2QWU1, 2QWV1, 2QWW1, 2QWZ1, 2QX01, 2QXX1, 2QY01, 2QY71, 2QYB1, 2QYZ1, 2QZ71, 2QZB1, 2QZC1, 2QZJ1, 2QZQ1, 2R011, 2R0B1, 2R0H1, 2R0J1, 2R0X1, 2R151, 2R161, 2R181, 2R191, 2R1I1, 2R251, 2R252, 2R2A1, 2R2C1, 2R2I1, 2R321, 2R371, 2R3U1, 2R471, 2R4I1, 2R5O1, 2R5S1, 2R5U1, 2R6M1, 2R6Q1, 2R6U1, 2R6V1, 2R761, 2R7H1, 2R8E1, 2R991, 2R9G1, 2R9U1, 2RA61, 2RA71, 2RA91, 2RAA1, 2RAE1, 2RAF1, 2RAS1, 2RB41, 2RBB1, 2RBD1, 2RBG1, 2RC31, 2RCD1, 2RCI1, 2RCT1, 2RCV1, 2RD91, 2RDC1, 2RDG1, 2RDM1, 2RDP1, 2RE11, 2RE31, 2RE71, 2RE91, 2REK1, 2REM1, 2RER1, 2RET2, 2RF41, 2RFL1, 2RFM1, 2RFR1, 2RFT2, 2RG41, 2RG81, 2RGQ1, 2RGS1, 2RGT1, 2RGX1, 2RH01, 2RH31, 2RHM1, 2RI71, 2RIE1, 2RIH1, 2RIQ1, 2RJ21, 2RJN1, 2RJZ1, 2RK01, 2RK31, 2RKH1, 2RKK1, 2RKQ1, 2RL81, 2SAK1, 2TNF1, 2TRC3, 2UCZ1, 2UV41, 2UVO1, 2UVP1, 2UWI1, 2UWN1, 2UX01, 2UX71, 2UY62, 2UYG1, 2UYK1, 2UYZ1, 2UZ81, 2UZH1, 2UZP1, 2UZQ1, 2V0U1, 2V141, 2V1C1, 2V1L1, 2V1M1, 2V1O1, 2V362, 2V4I1, 2V4I2, 2V4V1, 2V4X1, 2V541, 2V571, 2V5E1, 2V5T1, 2V6K1, 2V6V1, 2V701, 2V711, 2V721, 2V731, 2V7F1, 2V7S1, 2V821, 2V9R1, 2V9T2, 2V9U1, 2V9V1, 2VAC1, 2VB11,

2VBU1, 2VD21, 2VEB1, 2VEZ1, 2VF91, 2VFK1, 2VFX1, 2VG91, 2VGA1, 2VGD1, 2VGX1, 2VH11, 2VHK1, 2VI71, 2VID1, 2VIF1, 2VJ21, 2VJ31, 2VJW1, 2VK31, 2VKV1, 2VKW1, 2VLI1, 2VLQ2, 2VMH1, 2VML1, 2VML2, 2VN61, 2VNG1, 2VNL1, 2VO41, 2VO81, 2VO91, 2VOF1, 2VOK1, 2VOL1, 2VPA1, 2VPI1, 2VPR1, 2VPT1, 2VPZ2, 2VQ31, 2VQ51, 2VQ81, 2VQ91, 2VQZ1, 2VR91, 2VRI1, 2VRN1, 2VRP2, 2VRS1, 2VSM2, 2VSW1, 2VSZ1, 2VT31, 2VT81, 2VTW1, 2VTY1, 2VU51, 2VUV1, 2VVD1, 2VVP1, 2VVR1, 2V VW1, 2VVY1, 2VX81, 2VXG1, 2VXP1, 2VXT2, 2VXX1, 2VXZ1, 2VY81, 2VYI1, 2VYW1, 2VZB1, 2VZC1, 2VZF1, 2VZN1, 2VZP1, 2VZY1, 2W011, 2W072, 2W0G1, 2W0I1, 2W151, 2W1J1, 2W1S1, 2W1T1, 2W251, 2W2A1, 2W2S1, 2W2X2, 2W311, 2W3G1, 2W3J1, 2W3L1, 2W3T1, 2W3W1, 2W3X1, 2W431, 2W471, 2W4E1, 2W4L1, 2W4R1, 2W511, 2W531, 2W561, 2W571, 2W581, 2W5E1, 2W5Y1, 2W681, 2W691, 2W6K1, 2W721, 2W722, 2W7Q1, 2W7S1, 2W7Z1, 2W801, 2W821, 2W861, 2W871, 2W8M1, 2W9H1, 2W9N1, 2W9Y1, 2WAC1, 2WAG1, 2WAW1, 2WAX1, 2WB31, 2WB91, 2WBN1, 2WBT1, 2WC11, 2WCJ1, 2WCR1, 2WCU1, 2WCV1, 2WCW1, 2WD52, 2WDQ3, 2WDS1, 2WEE1, 2WEW1, 2WFB1, 2WFC1, 2W FH1, 2WFI1, 2WFO1, 2WFW1, 2WG71, 2WGP1, 2WH61, 2WH71, 2WJ91, 2WJA1, 2WJG1, 2WJI1, 2WJR1, 2WJZ2, 2WKK1, 2WL11, 2WLC1, 2WLP1, 2WLU1, 2WLV1, 2WM81, 2WMM1, 2WMY1, 2WNP1, 2WNS1, 2WNV1, 2WNV2, 2WNV3, 2WNX1, 2WNY1, 2WO11, 2WO32, 2WP41, 2WP71, 2WQ41, 2WQF1, 2WQL1, 2WRA1, 2WRY1, 2WS21, 2WSC16, 2WSC8, 2WSH1, 2WST1, 2WTG1, 2WTL1, 2WUH1, 2WUI1, 2WV31, 2WVF1, 2WVI1, 2WWF1, 2WWO1, 2WWX1, 2WWX2, 2WY32, 2WY41, 2WYT1, 2WZ11, 2WZ81, 2WZ91, 2WZO1, 2X0L2, 2X112, 2X271, 2X291, 2X2S1, 2X321, 2X361, 2X461, 2X4H1, 2X5N1, 2X5Q1, 2X5R1, 2X5Y1, 2X641, 2X6U1, 2X771, 2X781, 2X7B1, 2X7K1, 2X7M1, 2X8R1, 2XA01, 2XAT1, 2XB31, 2XBL1, 2XBT1, 2XBZ1, 2XC81, 2XCB1, 2XD71, 2XDE1, 2XDH1, 2XDN1, 2XDY1, 2XEM1, 2XEV1, 2XF41, 2XG31, 2XG52, 2XGG1, 2XGU1, 2XGV1, 2XGW1, 2XHZ1, 2XI71, 2XIG1, 2XJY1, 2XL41, 2XL61, 2XOM1, 2XOV1, 2XP11, 2XPP1, 2XQN3, 2XQX1, 2XS41, 2XS61, 2XSE1, 2XSQ1, 2XST1, 2XT21, 2XTJ4, 2XTS2, 2XTY1, 2XU31, 2XVM1, 2XVO1, 2XVS1, 2XW71, 2XWC1, 2XWS1, 2XX61, 2XY11, 2XY21, 2XYA1, 2XYK1, 2XYS1, 2XZ41, 2XZ81, 2Y0N1, 2Y1Q1, 2Y3M1, 2Y3Q1, 2Y3V1, 2Y3W1, 2Y441, 2Y482, 2Y694, 2Y7B1,

2Y8G1, 2Y8P1, 2Y961, 2Y9I1, 2YB61, 2YC21, 2YC22, 2YD11, 2YD61, 2YQY1, 2YSK1, 2YUT1, 2YVA1, 2YVE1, 2YVO1, 2YVQ1, 2YVS1, 2YVU1, 2YW31, 2YW61, 2YWD1, 2YWI1, 2YWJ1, 2YWL1, 2YWN1, 2YWR1, 2YWW1, 2YWX1, 2YXB1, 2YXD1, 2YXE1, 2YY61, 2YY81, 2YYH1, 2YYO1, 2YZH1, 2YZI1, 2YZJ1, 2YZK1, 2YZY1, 2Z081, 2Z0U1, 2Z0X1, 2Z101, 2Z161, 2Z2I1, 2Z3B1, 2Z3H1, 2Z511, 2Z5B1, 2Z5B2, 2Z5D1, 2Z5E1, 2Z642, 2Z691, 2Z6O1, 2Z6W1, 2Z791, 2Z7E1, 2Z7F1, 2Z7J1, 2Z841, 2Z8H1, 2Z8L1, 2Z8U1, 2Z901, 2Z981, 2Z991, 2ZAY1, 2ZB91, 2ZC31, 2ZC33, 2ZCA1, 2ZCM1, 2ZCW1, 2ZDB1, 2ZEJ1, 2ZET2, 2ZEX1, 2ZEZ1, 2ZF91, 2ZFU1, 2ZG11, 2ZG61, 2ZGL1, 2ZHH1, 2ZHN1, 2ZHZ1, 2ZIB1, 2ZJD1, 2ZK91, 2ZKI1, 2ZMF1, 2ZND1, 2ZNM1, 2ZNR1, 2ZNZ1, 2ZOU1, 2ZP21, 2ZQO1, 2ZS01, 2ZS02, 2ZS03, 2ZS04, 2ZS61, 2ZSC1, 2ZTD1, 2ZU11, 2ZUQ1, 2ZVF1, 2ZVY1, 2ZWM1, 2ZWR1, 2ZX21, 2ZYC1, 2ZYZ2, 2ZZD2, 3A0Y1, 3A1A1, 3A1B1, 3A1F1, 3A1W1, 3A2Z1, 3A351, 3A431, 3A571, 3A5Z2, 3A681, 3A6M1, 3A6R1, 3A6S1, 3A761, 3A7L1, 3A8G1, 3A8G2, 3A8R1, 3A981, 3A982, 3A9L1, 3AB61, 3ACD1, 3ACH1, 3AG34, 3AGT1, 3AGW1, 3AGY1, 3AH51, 3AIA1, 3AJ11, 3AJM1, 3AJV1, 3AJV2, 3AJW1, 3AJX1, 3AK21, 3AK81, 3AKB1, 3ALQ2, 3ALU1, 3ANP1, 3AP91, 3APA1, 3APQ1, 3AQ21, 3AQB1, 3AQG1, 3B2N1, 3B2U3, 3B421, 3B471, 3B481, 3B491, 3B541, 3B571, 3B5H1, 3B5M1, 3B6E1, 3B6I1, 3B6N1, 3B6Y1, 3B721, 3B771, 3B791, 3B7C1, 3B7M1, 3B7Y1, 3B811, 3B851, 3B8F1, 3B8L1, 3B8Z1, 3B931, 3B9C1, 3B9J1, 3BA31, 3BAL1, 3BB91, 3BBB1, 3BBD1, 3BBP1, 3BBY1, 3BC11, 3BCI1, 3BCY1, 3BDB1, 3BDD1, 3BDI1, 3BDR1, 3BDV1, 3BE41, 3BED1, 3BEM1, 3BES1, 3BF21, 3BFQ1, 3BGE1, 3BH71, 3BHO1, 3BHQ1, 3BHW1, 3BI71, 3BJ51, 3BJ61, 3BJA1, 3BJB1, 3BJK1, 3BJN1, 3BJV1, 3BK61, 3BL21, 3BL41, 3BL51, 3BLN1, 3BLZ1, 3BM11, 3BMB1, 3BMZ1, 3BN32, 3BN61, 3BNK1, 3BNV1, 3BNW1, 3BO61, 3BO71, 3BOD1, 3BOE1, 3BOM1, 3BOM2, 3BOQ1, 3BP62, 3BPK1, 3BPV1, 3BPZ1, 3BQA1, 3BQE1, 3BQO1, 3BQX1, 3BQY1, 3BQZ1, 3BRC1, 3BRJ1, 3BRN1, 3BRO1, 3BS21, 3BSW1, 3BT31, 3BT51, 3BU21, 3BUT1, 3BVF1, 3BVO1, 3BVP1, 3BW61, 3BW81, 3BWD1, 3BWL1, 3BWU1, 3BWU2, 3BWU3, 3BWV1, 3BWY1, 3BWZ1, 3BX41, 3BX42, 3BY41, 3BY51, 3BY81, 3BYI1, 3BYQ1, 3BYW1, 3BZ61, 3BZH1, 3C0T1, 3C0U1, 3C121, 3C191, 3C1D1, 3C1L1, 3C221, 3C3I1, 3C3M1, 3C3P1, 3C4R1, 3C5C1, 3C5J1, 3C5O1, 3C5P1, 3C5R1,

3C641, 3C6F1, 3C6V1, 3C752, 3C7X1, 3C851, 3C8G1, 3C8I1, 3C8J1, 3C8K4, 3C8L1, 3C8O1, 3C8U1, 3C8X1, 3C971, 3C9G1, 3C9P1, 3C9Q1, 3CA91, 3CAN1, 3CB01, 3CBN1, 3CBQ1, 3CBU1, 3CCG1, 3CCY1, 3CDH1, 3CDK2, 3CDL1, 3CE11, 3CEB1, 3CEI1, 3CEU1, 3CEX1, 3CF42, 3CFU1, 3CFW1, 3CFY1, 3CG01, 3CG41, 3CG61, 3CGG1, 3CGM1, 3CH41, 3CHM1, 3CI02, 3CI31, 3CI61, 3CIP2, 3CIT1, 3CJD1, 3CJE1, 3CJJ1, 3CJN1, 3CJX1, 3CK11, 3CK21, 3CKI2, 3CL31, 3CL32, 3CLJ1, 3CLV1, 3CM01, 3CM11, 3CM31, 3CMI1, 3CNB1, 3CNC1, 3CNE1, 3CNG1, 3CNH1, 3CNI1, 3CNV1, 3CNX1, 3CO51, 3COL1, 3COO1, 3CP31, 3CP71, 3CPF1, 3CPH2, 3CQN1, 3CQV1, 3CR31, 3CR71, 3CRJ1, 3CRN1, 3CRT1, 3CRY1, 3CS11, 3CSY4, 3CT51, 3CT61, 3CT81, 3CTD1, 3CTW1, 3CU31, 3CU51, 3CUE1, 3CUE3, 3CUE4, 3CUQ2, 3CUQ3, 3CVO1, 3CWQ1, 3CWR1, 3CWX1, 3CX510, 3CX55, 3CX57, 3CX82, 3CXG1, 3CXJ1, 3CXK1, 3CYP1, 3CZ51, 3CZ61, 3CZU2, 3CZX1, 3D001, 3D011, 3D061, 3D0J1, 3D122, 3D1K1, 3D1K2, 3D1P1, 3D221, 3D301, 3D3B1, 3D3M1, 3D3O1, 3D3S1, 3D4E1, 3D4R1, 3D531, 3D5N1, 3D5P1, 3D6L1, 3D6M1, 3D6R1, 3D6X1, 3D791, 3D7A1, 3D7J1, 3D7L1, 3D7N1, 3D853, 3D891, 3D8D1, 3D8M1, 3D8P1, 3D9N1, 3D9R1, 3D9Y1, 3DA01, 3DA51, 3DA81, 3DAI1, 3DAL1, 3DAU1, 3DAW2, 3DB31, 3DB51, 3DB71, 3DBA1, 3DBO2, 3DC51, 3DCF1, 3DCM1, 3DCN1, 3DCZ1, 3DD71, 3DDC2, 3DDV1, 3DDY1, 3DEO1, 3DEU1, 3DEW1, 3DFG1, 3DH01, 3DH11, 3DHO1, 3DHW1, 3DI51, 3DKA1, 3DKU1, 3DKX1, 3DKZ1, 3DLC1, 3DLD1, 3DLM1, 3DLO1, 3DLQ1, 3DLQ2, 3DM81, 3DMB1, 3DMC1, 3DMM5, 3DMN1, 3DMO1, 3DMP1, 3DN71, 3DNS1, 3DNX1, 3DO81, 3DO91, 3DOU1, 3DPJ1, 3DQG1, 3DQP1, 3DR61, 3DRN1, 3DRX1, 3DSB1, 3DSZ1, 3DTD1, 3DUI1, 3DUK1, 3DV81, 3DWL5, 3DWV1, 3DXE1, 3DXY1, 3DZA1, 3DZM1, 3E051, 3E0H1, 3E0J2, 3E0K1, 3E0O1, 3E0U1, 3E0Y1, 3E101, 3E1E1, 3E201, 3E231, 3E291, 3E2I1, 3E391, 3E3R1, 3E3U1, 3E3V1, 3E4G1, 3E4V1, 3E571, 3E581, 3E5D1, 3E5Y1, 3E6M1, 3E6P1, 3E6Q1, 3E6S1, 3E7D1, 3E7N1, 3E7Q1, 3E8L1, 3E8M1, 3E8P1, 3E8T1, 3E991, 3E9U1, 3E9V1, 3EA61, 3EAA1, 3EAQ1, 3EBE1, 3EBK1, 3EBR1, 3EBT1, 3EBW1, 3EBY1, 3EC21, 3EC61, 3EC81, 3EC91, 3ECF1, 3ECH1, 3ECO1, 3EDH1, 3EDO1, 3EDS1, 3EDU1, 3EEA1, 3EEB1, 3EEF1, 3EER1, 3EES1, 3EEV1, 3EEY1, 3EF41, 3EF81, 3EFA1, 3EFE1, 3EFF3, 3EFY1, 3EG71, 3EGD3, 3EGI1, 3EGN1, 3EGQ1, 3EHC1, 3EHD1, 3EHF1, 3EHG1,

3EHH1, 3EHW1, 3EIK1, 3EIN1, 3EJA1, 3EJC1, 3EJF1, 3EJG1, 3EJJ1, 3EJK1,
3EJV1, 3EK31, 3ELG1, 3ELI1, 3ELN1, 3ELS1, 3ELV1, 3ELX1, 3EMO1, 3EMU1,
3EMW1, 3EMX1, 3EN81, 3ENP1, 3EO41, 3EO51, 3EO81, 3EOI1, 3EOP1, 3EOZ1,
3EP11, 3EPU1, 3EQE1, 3EQZ1, 3ER01, 3ER61, 3ERW1, 3ERX1, 3ERY1, 3ES11,
3ESG1, 3ESI1, 3ESL1, 3ESM1, 3ET61, 3ETI1, 3ETN1, 3ETP1, 3EUL1, 3EUP1,
3EUR1, 3EVO1, 3EW11, 3EWI1, 3EWL1, 3EWR1, 3EXN1, 3EXQ1, 3EXV1,
3EXZ1, 3EY51, 3EY81, 3EYB1, 3EYE1, 3EYK2, 3EYT1, 3EYX1, 3EYY1, 3EZX1,
3EZZ1, 3F041, 3F081, 3F091, 3F0C1, 3F0D1, 3F0P1, 3F0W1, 3F131, 3F1B1,
3F1J1, 3F1R1, 3F1T1, 3F2I1, 3F2V1, 3F2Z1, 3F311, 3F3M1, 3F3X1, 3F441,
3F4A1, 3F4F1, 3F4M1, 3F4W1, 3F591, 3F5B1, 3F5D1, 3F5O1, 3F5T1, 3F6A1,
3F6C1, 3F6D1, 3F6Q1, 3F6R1, 3F7C1, 3F7E1, 3F7L1, 3F7S1, 3F7X1, 3F811,
3F8H1, 3F8K1, 3F8L1, 3F8X1, 3F951, 3F9K1, 3F9S1, 3F9U1, 3F9X1, 3FAN1,
3FB21, 3FBK1, 3FBU1, 3FBZ1, 3FCM1, 3FCN1, 3FD31, 3FD41, 3FDI1, 3FDL1,
3FDW1, 3FDX1, 3FES1, 3FET1, 3FEU1, 3FF01, 3FF41, 3FFL1, 3FFV1, 3FG91,
3FGE1, 3FGR1, 3FGY1, 3FH11, 3FH21, 3FH31, 3FHF1, 3FHG1, 3FHK1, 3FHM1,
3FHV1, 3FI71, 3FIQ1, 3FIW1, 3FIX1, 3FJ21, 3FJV1, 3FK31, 3FK81, 3FK91,
3FKB1, 3FKE1, 3FKF1, 3FKH1, 3FKM1, 3FLD1, 3FLJ1, 3FLO3, 3FLP1, 3FM21,
3FM51, 3FMS1, 3FMZ1, 3FN12, 3FN51, 3FNC1, 3FNI1, 3FNK1, 3FOS1, 3FPW1,
3FQ31, 3FQM1, 3FR31, 3FRQ1, 3FRR1, 3FRT1, 3FRV1, 3FS51, 3FSA1, 3FSD1,
3FTT1, 3FUY1, 3FV51, 3FV61, 3FVB1, 3FVW1, 3FW21, 3FWB1, 3FWZ1, 3FXA1,
3FYF1, 3FYN1, 3FYQ1, 3FZ21, 3FZ51, 3FZE1, 3FZG1, 3FZX1, 3G0K1, 3G0M1,
3G131, 3G141, 3G161, 3G2E1, 3G2S1, 3G391, 3G3Z1, 3G461, 3G5J1, 3G5P1,
3G661, 3G671, 3G6I1, 3G7B1, 3G7E1, 3G7G1, 3G7M1, 3G7P1, 3G7X1, 3G8K1,
3G8L1, 3G8W1, 3G8Z1, 3G9A2, 3G9K2, 3G9V1, 3GA31, 3GA41, 3GAG1, 3GBH1,
3GBS1, 3GBW1, 3GBY1, 3GCG2, 3GDW1, 3GE51, 3GE61, 3GEK1, 3GEU1,
3GFA1, 3GFP1, 3GFS1, 3GG61, 3GGN1, 3GGQ1, 3GGY1, 3GHA1, 3GIO1, 3GIU1,
3GIX1, 3GK61, 3GKM1, 3GL31, 3GLV1, 3GM31, 3GM51, 3GMF1, 3GMG1, 3GMV1,
3GMX1, 3GN31, 3GN41, 3GN51, 3GNZ1, 3GOR1, 3GOV1, 3GOY1, 3GP41, 3GP61,
3GPG1, 3GPR3, 3GPR4, 3GQE1, 3GQH1, 3GQQ1, 3GR01, 3GR51, 3GRA1,
3GRC1, 3GRD1, 3GRN1, 3GRZ1, 3GSL1, 3GT21, 3GT71, 3GTY2, 3GUV1, 3GUZ1,
3GV11, 3GVF1, 3GVY1, 3GW41, 3GWI1, 3GWM1, 3GWR1, 3GWY1, 3GX01,

3GX11, 3GXB1, 3GXH1, 3GXR1, 3GXV1, 3GY91, 3GYD1, 3GYK1, 3GYX2, 3GYZ1, 3GZ81, 3GZB1, 3GZI1, 3GZR1, 3GZX2, 3H051, 3H081, 3H0G4, 3H0G5, 3H0G7, 3H0G8, 3H0N1, 3H0X1, 3H161, 3H1G1, 3H1S1, 3H2B1, 3H2D1, 3H351, 3H3B1, 3H3I1, 3H4O1, 3H4Q1, 3H511, 3H5I1, 3H5J1, 3H6Q1, 3H6R1, 3H6X1, 3H861, 3H871, 3H8D1, 3H8K1, 3H8T1, 3H8U1, 3H931, 3H951, 3H961, 3H9N1, 3HA21, 3HA41, 3HA91, 3HB71, 3HBE1, 3HBW1, 3HCG1, 3HCJ1, 3HCS1, 3HCY1, 3HCZ1, 3HD41, 3HD51, 3HDC1, 3HDE1, 3HDF1, 3HDG1, 3HDP1, 3H DU1, 3HDV1, 3HE01, 3HE11, 3HE81, 3HEB1, 3HEF1, 3HEI1, 3HEI2, 3HF71, 3HFH1, 3HFI1, 3HGB1, 3HGM1, 3HH01, 3HHC1, 3HHF1, 3HHJ1, 3HHL1, 3HHO1, 3HHT1, 3HIE1, 3HIM1, 3HIU1, 3HJ71, 3HJG1, 3HJJ1, 3HJN1, 3HKL1, 3HKS1, 3HKV1, 3HKZ5, 3HLY1, 3HM01, 3HM21, 3HM41, 3HMA1, 3HMC1, 3HMH1, 3HMX2, 3HMZ1, 3HN51, 3HNM1, 3HNR1, 3H NW1, 3HNY1, 3HOG1, 3HOI1, 3HP41, 3HPC1, 3HPE1, 3HPH1, 3HQC1, 3HR41, 3HRA1, 3HRD4, 3HRS1, 3HSR1, 3HST2, 3HT11, 3HTA1, 3HTM1, 3HTN1, 3HU51, 3HUH1, 3HUP1, 3HV21, 3HVA1, 3H VU1, 3HVV1, 3H VW1, 3HWJ1, 3HWU1, 3HX81, 3HXI1, 3HY31, 3HYJ1, 3HYN1, 3HZ41, 3HZ81, 3HZA1, 3HZH1, 3HZH2, 3HZP1, 3H ZS1, 3I061, 3I081, 3I0U1, 3I0Y1, 3I241, 3I2V1, 3I3F1, 3I3G1, 3I3Q1, 3I4P1, 3I4S1, 3I571, 3I5B1, 3I5C1, 3I5G2, 3I5G3, 3I6X1, 3I7D1, 3I7M1, 3I7U1, 3I861, 3I8N1, 3I8O1, 3I8T1, 3I981, 3I9F1, 3I9S1, 3I9X1, 3IA11, 3IA41, 3IA81, 3IAM2, 3IAM5, 3IAM6, 3IAM7, 3IAM8, 3IB61, 3IBM1, 3IBS1, 3IBZ1, 3ICA1, 3ICH1, 3ICL1, 3ICU1, 3ID91, 3IDB2, 3IDF1, 3IE51, 3IEY1, 3IEY2, 3IFU1, 3IG21, 3IGN1, 3IGQ1, 3IGR1, 3IH51, 3IH61, 3IHK1, 3IHO1, 3IHT1, 3IHW1, 3IHX1, 3II21, 3I IJ1, 3IIR1, 3IIS1, 3IJ51, 3IJF1, 3IJM1, 3IJT1, 3IKB1, 3ILO1, 3ILC1, 3ILH1, 3ILX1, 3IM61, 3IMI1, 3IMK1, 3IMM1, 3IN61, 3INO1, 3IOQ1, 3IOS1, 3IP01, 3IPR1, 3IQ11, 3IQ21, 3IQT1, 3IR21, 3IR31, 3IR41, 3IR91, 3IRA1, 3IRB1, 3ISO1, 3IT41, 3IT42, 3IT71, 3ITQ1, 3ITW1, 3IU61, 3IVV1, 3IWT1, 3IX31, 3IX71, 3IX91, 3IXC1, 3IXR1, 3JQO2, 3JR01, 3JR21, 3JRN1, 3JRT1, 3JRV1, 3JS91, 3JSB1, 3JSJ1, 3JSY1, 3JTE1, 3JTF1, 3JTW1, 3JUD1, 3JUI1, 3JUM1, 3JUW1, 3JV11, 3JV91, 3JVI1, 3JVN1, 3JW41, 3JWG1, 3JWH1, 3JWK1, 3JX91, 3JXS1, 3JYB1, 3JYG1, 3JYZ1, 3JZ71, 3JZ91, 3JZV1, 3K051, 3K0L1, 3K0T1, 3K0W1, 3K0Z1, 3K1F13, 3K1R1, 3K1Y1, 3K211, 3K291, 3K2C1, 3K2J1, 3K2N1, 3K2V1, 3K2Z1, 3K321, 3K3C1, 3K3K1, 3K3Q1, 3K3W1, 3K441,

3K512, 3K5B2, 3K5J1, 3K5R1, 3K631, 3K671, 3K691, 3K6E1, 3K6H1, 3K6Q1,
3K6Z1, 3K7I1, 3K7P1, 3K861, 3K8U1, 3K9O1, 3K9W1, 3KA21, 3KA81, 3KAS2,
3KB21, 3KB51, 3KBB1, 3KBE1, 3KBG1, 3KBQ1, 3KBY1, 3KCM1, 3KCN1,
3KCP2, 3KCQ1, 3KD31, 3KDG1, 3Kdq1, 3KDV1, 3KE41, 3KE71, 3KEG1, 3KEO1,
3KEP1, 3KEV1, 3KEY1, 3KF61, 3KF81, 3KF82, 3KFF1, 3KGZ1, 3KH11, 3KH61,
3KH71, 3KHE1, 3KHN1, 3KHT1, 3KIJ1, 3KIP1, 3KK41, 3KKC1, 3KKG1, 3KKQ1,
3KKS1, 3KKW1, 3KKY1, 3KL11, 3KLB1, 3KLQ1, 3KLR1, 3KLX1, 3KM11,
3KM31, 3KM51, 3KMI1, 3KML1, 3KMOV1, 3KNV1, 3KNW1, 3KNY1, 3KO51,
3KOL1, 3KOP1, 3KOS1, 3KP71, 3KPA1, 3KQ01, 3KQG1, 3KQR1, 3KRM1, 3KS41,
3KSH1, 3KSN1, 3KSP1, 3KSV1, 3KTA1, 3KTA2, 3KTM1, 3KTO1, 3KTS1, 3KTY1,
3KU32, 3KU81, 3KUC1, 3KUU1, 3KUV1, 3KV01, 3KVH1, 3KW01, 3KWE1,
3KWK1, 3KWO1, 3KWU1, 3KX01, 3KXA1, 3KXR1, 3KXS1, 3KXY1, 3KY81,
3KYE1, 3KYJ1, 3KYJ2, 3KYP1, 3KYQ1, 3KYS1, 3KZ91, 3KZJ1, 3KZQ1, 3KZT1,
3L001, 3L0F1, 3L0F2, 3L0W2, 3L181, 3L191, 3L1N1, 3L201, 3L2H1, 3L341, 3L3U1,
3L411, 3L421, 3L4A1, 3L4B2, 3L4C1, 3L4E1, 3L4Q1, 3L4Q2, 3L4R1, 3L501,
3L511, 3L512, 3L6I1, 3L6N1, 3L7Q1, 3L7T1, 3L7U1, 3L7X1, 3L8E1, 3L8H1, 3L8M1,
3L8U1, 3L911, 3L9B1, 3L9F1, 3L9J1, 3L9Q1, 3L9S1, 3L9U1, 3L9V1, 3L9Y1,
3L9Z1, 3LAA1, 3LAB1, 3LAC1, 3LAS1, 3LAT1, 3LB21, 3LB51, 3LB91, 3LBE1,
3LBF1, 3LBX1, 3LBX2, 3LBY1, 3LCM1, 3LD21, 3LD31, 3LDT1, 3LDZ2, 3LE01,
3LEV1, 3LF42, 3LF61, 3LF91, 3LFH1, 3LFJ1, 3LFK1, 3LFR1, 3LGB1, 3LGO1,
3LHE1, 3LHH1, 3LHK1, 3LHQ1, 3LIM1, 3LJD1, 3LJE1, 3LJL1, 3LKT1, 3LL83,
3LLO1, 3LLR1, 3LLU1, 3LLV1, 3LMB1, 3LMC1, 3LME1, 3LMN1, 3LND1, 3LNE1,
3LNR1, 3LOC1, 3LOD1, 3LOI1, 3LOR1, 3LP61, 3LPW1, 3LQ91, 3LQB1, 3LQC1,
3LQC2, 3LQH1, 3LQK1, 3LQM1, 3LQN1, 3LQW1, 3LQY1, 3LQZ1, 3LQZ2, 3LR21,
3LRU1, 3LSJ1, 3LSU1, 3LTJ1, 3LTL1, 3LTM1, 3LUA1, 3LUC1, 3LUR1, 3LV91,
3LVL2, 3LVY1, 3LW31, 3LW512, 3LW515, 3LW516, 3LW517, 3LW518, 3LW54,
3LW56, 3LW71, 3LW91, 3LWA1, 3LWF1, 3LWJ1, 3LWK1, 3LWX1, 3LWZ1, 3LX01,
3LX31, 3LXR2, 3LYB1, 3LYD1, 3LYG1, 3LYH1, 3LYI1, 3LYK1, 3LYN1, 3LYP1,
3LYU1, 3LZA1, 3LZL1, 3M0F1, 3M1C2, 3M1H1, 3M1I2, 3M1P1, 3M1X1, 3M1Y1,
3M3G1, 3M3M1, 3M4F1, 3M5K1, 3M5L1, 3M5R1, 3M651, 3M6C1, 3M6J1, 3M6P1,
3M7A1, 3M7F2, 3M7O1, 3M9G1, 3M9L1, 3M9Z1, 3MA21, 3MA22, 3MA91, 3MAH1,

3MAL1, 3MBT1, 3MCF1, 3MCI1, 3MCR1, 3MCS1, 3MCU1, 3MCW1, 3MD31, 3MDP1, 3MDX1, 3ME01, 3ME02, 3ME41, 3ME71, 3MEA1, 3MEP1, 3MER1, 3MEW1, 3MF61, 3MFB1, 3MFF1, 3MFN1, 3MGD1, 3MGK1, 3MGL1, 3MGM1, 3MGW1, 3MH91, 3MHA1, 3MIO1, 3MIT1, 3MJ92, 3MJK1, 3MK31, 3MK72, 3MK91, 3MKB1, 3MKB2, 3MKQ2, 3MKS1, 3ML31, 3MLA1, 3MLC1, 3MLG1, 3MLQ1, 3MM41, 3MMH1, 3MMZ1, 3MN11, 3MNG1, 3MNL1, 3MOL1, 3MOQ1, 3MP21, 3MPD1, 3MPZ1, 3MQ21, 3MQH1, 3MQL1, 3MQM1, 3MQZ1, 3MR01, 3MR71, 3MSE1, 3MSO1, 3MSW1, 3MSX2, 3MTK1, 3MTQ1, 3MTR1, 3MTT1, 3MTV1, 3MTX1, 3MUD1, 3MUJ1, 3MVC1, 3MVK1, 3MVN1, 3MVP1, 3MVS1, 3MW61, 3MWM1, 3MWN1, 3MXN1, 3MXN2, 3MXO1, 3MXQ1, 3MXU1, 3MY21, 3MZ01, 3MZY1, 3N071, 3N081, 3N0U1, 3N101, 3N1E1, 3N1H1, 3N1U1, 3N2N1, 3N3A2, 3N4I2, 3N4J1, 3N4X1, 3N501, 3N5A1, 3N6Y1, 3N701, 3N721, 3N771, 3N791, 3N8E1, 3N8K1, 3N9B1, 3N9G2, 3N9U2, 3NA21, 3NAT1, 3NBI1, 3NBK1, 3NCE1, 3NCE2, 3NCT1, 3NCV1, 3NCX1, 3NE01, 3NE71, 3NEC1, 3NEY1, 3NF51, 3NFD1, 3NFG2, 3NFQ1, 3NFW1, 3NGS1, 3NGW1, 3NHN1, 3NHV1, 3NI61, 3NI71, 3NI81, 3NJ21, 3NJC1, 3NJE1, 3NJK1, 3NKE1, 3NKG1, 3NKL1, 3NKU1, 3NL91, 3NNG1, 3NO81, 3NOH1, 3NQN1, 3NQO1, 3NQR1, 3NQW1, 3NQZ1, 3NR11, 3NR31, 3NR51, 3NR61, 3NRD1, 3NRG1, 3NRH1, 3NRV1, 3NRX1, 3NTK1, 3NUL1, 3NV11, 3NVW1, 3NW01, 3NWC1, 3NWG1, 3NWS1, 3NWZ1, 3NX21, 3NXC1, 3NYK1, 3NZ21, 3NZB1, 3O0A1, 3O0G2, 3O0M1, 3O0P1, 3O0R4, 3O0Z1, 3O101, 3O1G1, 3O221, 3O2Q2, 3O2R1, 3O2U1, 3O361, 3O4A1, 3O5V1, 3O5Y1, 3O601, 3O651, 3O6Q1, 3O6U1, 3O7I1, 3O7M1, 3O851, 3O8S1, 3O8Z1, 3O941, 3OA41, 3OA71, 3OAD1, 3OAD2, 3OAO1, 3OBF1, 3OBL1, 3OBQ1, 3OC81, 3OCA1, 3OCM1, 3OCO1, 3OCQ1, 3OD91, 3OES1, 3OF41, 3OFK1, 3OG51, 3OG61, 3OG91, 3OGA1, 3OGH1, 3OGN1, 3OHE1, 3OHP1, 3OHW1, 3OI81, 3OIR1, 3OJ01, 3OJ61, 3OKX1, 3OMD1, 3OMS1, 3ON21, 3ON31, 3ON41, 3ONO1, 3OOO1, 3OOP1, 3OOV1, 3OOW1, 3OP31, 3OP61, 3OPC1, 3OQP1, 3OR51, 3OSJ1, 3OSK1, 3OSY1, 3OT11, 3OT21, 3OU21, 3OUE1, 3OUI1, 3OUQ1, 3OVK1, 3OWC1, 3OWR1, 3OX61, 3OXP1, 3OYY1, 3OZI1, 3P011, 3P061, 3P0R1, 3P0T1, 3P1G1, 3P1I1, 3P1J1, 3P2A1, 3P2L1, 3P2T1, 3P3V1, 3P431, 3P481, 3P4L1, 3P511, 3P5N1, 3P5U1, 3P6D1, 3P7G1, 3P8B2, 3P8C5, 3P8N1, 3P8S1, 3P941, 3P9N1, 3P9V1, 3P9X1,

3P9Y1, 3PAS1, 3PBF1, 3PCV1, 3PDY1, 3PFS1, 3PFY1, 3PG61, 3PGE1, 3PGE2, 3PH11, 3PH41, 3PHU1, 3PHX1, 3PIM1, 3PJP1, 3PJV1, 3PJY1, 3PK11, 3PL11, 3PM81, 3PMC1, 3PMD1, 3PMI1, 3PN91, 3PNA1, 3PNL2, 3PNT2, 3PNX1, 3PP91, 3PPB1, 3PPE1, 3PR91, 3PRO2, 3PRU1, 3PSE2, 3PSQ1, 3PSS1, 3PU21, 3PU61, 3PUF3, 3PUT1, 3PVE1, 3PXV1, 3PYC1, 3PYI1, 3PYW1, 3PZJ1, 3PZY1, 3Q0Y1, 3Q341, 3Q3J2, 3Q3W1, 3Q491, 3Q621, 3Q631, 3Q641, 3Q6A1, 3Q6L1, 3Q6Z1, 3Q901, 3Q911, 3Q931, 3QB21, 3QBM1, 3QBQ2, 3QBR1, 3QBT2, 3QC71, 3QD51, 3QD62, 3QD81, 3QDS1, 3QEC1, 3QIO1, 3QJG1, 3QK31, 3QKS1, 3QKS2, 3QKX1, 3QLE1, 3QMN1, 3QOC1, 3QOO1, 3QP41, 3QQA1, 3QQI1, 3QSZ1, 3QT14, 3QTH1, 3QU11, 3QUA1, 3QVE1, 3QW91, 3QWM1, 3QWN1, 3QY31, 3QYH2, 3QZ31, 3QZ51, 3QZB1, 3R0A1, 3R0N1, 3R0P1, 3R1W1, 3R2V1, 3R3R1, 3R4Q1, 3R6F1, 3R901, 3R9F1, 3RAB1, 3RAP1, 3RBB1, 3RCZ2, 3RE31, 3RF01, 3RJT1, 3SXL1, 4UBP2, 5TMP1, 7CEI2, 830C1, 8PCH1, 966C1

A.1.8 Dataset of redundant crystal structures

The following protein structures are used to illustrate the correlation between structure resolution and number of hydrogen bonds in protein structures (c.f. Section 6.2.2).

1A0HB, 1A8JH, 1AKJA, 1BJ1V, 1BMBA, 1C1MA, 1C5MD, 1DXTB, 1EZXC, 1FNTA, 1FNTB, 1FNTE, 1FNTH, 1FNTI, 1FNTJ, 1FNTH, 1FNTN, 1FP6A, 1FYZE, 1G0UL, 1G3IG, 1G6NA, 1HL4A, 1K4CA, 1K4CB, 1KD7A, 1KJ3H, 1OTSC, 1OTSD, 1QABA, 1RMLA, 1S9QA, 1U8HA, 1VYWA, 1VYWB, 1Z7QC, 2C1JA, 2FHG2, 2FHH1, 2JE6A, 2P1MA, 2PGZA, 2R2KA, 2WBJA, 2WQAA, 2Z5CC, 3EWTA, 3KNFA, 3M04A, 3PDOB

A.1.9 Dataset of NMR structures

The following pairs of protein structures are used to benchmark the hydrogen bond structure refinement protocol. High-resolution X-ray structure and low-resolution NMR structure are listed as the first and second items in the pair (c.f. Section 6.2.3).

1A3AA-1J6TA, 1A6MA-1MYFA, 1BEB A-1DV9A, 1CPQA-1EKYA, 1DQEA-1XFRA,
1DY5A-2AASA, 1EAJA-1RSFA, 1EAQA-1CMOA, 1EZ3A-1BR0A, 1F3ZA-1O2FA,
1FVKA-1A24A, 1G4IA-1BVMA, 1GP0A-2CNJD, 1GU2A-1E8EA, 1GWYA-2L38A,
1H4AX-2KFBA, 1HCZA-2PCFB, 1HFSA-2SR TA, 1HN4A-1PISA, 1HW5A-2WC2B,
1IAZA-1KD6A, 1ICMA-1UREA, 1IHBA-1BU9A, 1J7DA-1ZGUA, 1JBEA-1CEYA,
1JL3A-2IPAB, 1KMTA-1AJWA, 1KMVA-1YHOA, 1KQ6A-1GD5A, 1M48A-1IRL,
1N1XA-1QWQA, 1NFNA-2KC3A, 1NH2A-1TBAB, 1NWZA-3PHYA, 1O8XA-1OKDA,
1OMRA-1JSAA, 1PAZA-2P80D, 1PDOA-1VSQB, 1QJ8A-1Q9GA, 1QMYA-2JQFS,
1QYZA-2FWLA, 1RG8A-1RMLA, 1S69A-1MWBA, 1SENA-2K8VA, 1SNCA-1JORA,
1SVNA-1AH2A, 1SX7A-2LCBA, 1T2WA-1IJAA, 1T3YA-1WNJA, 1TVGA-1XPWA,
1UOWA-1K5WA, 1UV0A-2GO0A, 1VKBA-2KL2A, 1VYFA-2POAA, 1W6NB-2KM2B,
1WVHA-2GJYA, 1X13A-2BRUB, 1XWVA-1WRFA, 1Y93A-2POJA, 1YARA-2KU2G,
1Z9LA-2RR3A, 1ZKOA-2KA7A, 1ZNKA-1DF3A, 1ZX8A-2KA0A, 2BCNA-2GB8A,
2BEMA-2LHSA, 2CALA-1CURA, 2CDNA-1P4SA, 2CE2X-2LCFA, 2CHDA-2K3HA,
2CUAA-2FWLB, 2CWIA-1I56, 2CX7A-1WFRA, 2D48A-1ITIA, 2E0LA-2K11A,
2E85A-2I8LA, 2ESKA-2FUHA, 2FUFA-2TBDA, 2FY6A-2JZSA, 2G7BA-1BM5A,
2GRRA-2PX9B, 2GUIA-2XY8A, 2ICAA-1DGQA, 2NSZA-2KZTB, 2NTWA-1R84A,
2NVHA-2KH2A, 2NWDX-1IY4, 2NXBB-2L5EA, 2O5GA-2K0FA, 2OZNA-2O4EA,
2PWOA-2JPRA, 2RG8A-2KZTA, 2RHKA-2KKZA, 2UZJA-2JTCA, 2V6VA-2CZOA,
2VB1A-1E8LA, 2VBUA-2P3MA, 2VLQB-2K5X, 2VZCA-2K2RA, 2W0GA-2K5BB,
2WH6A-1Q59, 2WUHA-2Z4FA, 2WYTA-1BA9A, 2X7KA-1XWNA, 2Y1QA-2K77A,
2Y6DA-2DDYA, 2YB6A-2Y4WA, 2YGNA-2D3JA, 2YK9A-2K5BA, 2Z2IA-2JRCA,
2Z6OA-2K07A, 2ZJDA-2K6QA, 2ZK9X-2KSVA, 2ZWUA-2L8M, 3A6RA-1AXJA,
3A6SA-1PUSA, 3AA0B-2KZ7B, 3AA4A-1RCHA, 3AKMA-1KZXA, 3BY8A-1OJGA,
3D06A-2FEJA, 3D3BA-1EY1A, 3E0HA-2KCUA, 3E2OA-2GB8A, 3GFPA-2KBFA,
3H8KA-2KLYA, 3HQCA-2DKQA, 3IP0A-2F65A, 3IVVA-2CR2A, 3JRVB-2K36A,
3JXSA-1K42A, 3LJWB-2KTBB, 3LS6A-1IEZA, 3NI6A-2KI3A, 3OGNA-2L2CA,
3OSKA-1AH1A, 3P3CA-2JT2A, 3P8NA-2K1QA, 3QAPA-2KSYA, 3R62A-2BZMA,
3STKA-2PY1A, 3T3LA-1LY7A, 3ZR0A-1IRYA, 3ZXHB-1FM1A, 4FGFA-1BLDA

A.1.10 X-ray density maps data set

The following pairs of protein structures are used to benchmark the hydrogen bond structure refinement protocol. High-resolution X-ray structure and its low-resolution counterpart are listed as the first and second items in the pair (c.f. Section 6.4.2).

1CXQA-1VSLA, 1E4CP-3FUAA, 1E4MM-1DWJM, 1ESWA-1FP9A, 1F4PA-1J9GA, 1H05A-2Y76A, 1J02A-2DY5A, 1LO7A-1LO9A, 1LR7A-1LR9A, 1M15A-4GVZA, 1MTZA-1XRQA, 1N40A-4G2GA, 1ODMA-2WO7A, 1OYGA-2VDTA, 1P4CA-2A85A, 1PINA-2ZR6A, 1QJ4A-1SCQA, 1RA0A-3RN6A, 1T7VA-1T7XA, 1V7RA-2ZTIA, 1WD3A-2D43A, 1YE3A-1QLJA, 1YFUA-1YFYA, 1YWMA-2O0I1, 1ZODA-1Z3ZA, 2CALA-1A8ZA, 2D48A-1HIJA, 2E1ZA-4FWRA, 2E4TA-2EQDA, 2H6DA-2YZAA, 2H7IA-3FNHA, 2R60A-2R66A, 2R8NA-3U71A, 2W8NA-2W8OA, 2XKIA-2XKHA, 2ZFGA-3FYXA, 2ZGRA-2ZGKA, 2ZNDA-3AAKA, 3A99A-2XJ0A, 3ACXA-3TFVA, 3B4XA-1UE8A, 3BJHA-2H8VA, 3C7AA-3C7CB, 3DAUA-2ANOA, 3ESUF-3ET9F, 3FKCA-3M8VA, 3FRHA-3B89A, 3HGPA-2FOFA, 3I28A-3OTQA, 3PJXA-3PJWA, 3PRNA-1H6S1, 3PVHA-3PTJA, 3Q0WA-3QPLA, 3RZNA-4GVTA, 3U3PA-3U3TA, 3UGJA-3UMMA, 4GV2A-3CE0A, 4J12A-4HAGA

A.1.11 *Packpred* potential training set

The following protein structures are used to generate the *Packpred* potential (c.f. Section 7.2.3).

1A2PA, 1A3AA, 1A62A, 1A7DA, 1AD6A, 1AE9A, 1AM7A, 1AMXA, 1APYA, 1AQEA, 1AQZA, 1ATZA, 1AYOA, 1B0BA, 1B0NA, 1B1CA, 1B2PA, 1B93A, 1BGFA, 1BKBA, 1BKRA, 1BOUA, 1BXEA, 1BYFA, 1BYRA, 1BZ4A, 1C44A, 1C52A, 1C7KA, 1C8SA, 1CCWA, 1CCZA, 1CJWA, 1CNT1, 1CV8A, 1CXQA, 1D0QA, 1D1GA, 1D2ZA, 1D2ZB, 1D4OA, 1D4TA, 1DANU, 1DD3A, 1DG6A, 1DI6A, 1DK7A, 1DK8A, 1DLYA, 1DNLA, 1DQGA, 1DUSA, 1DVOA, 1DW0A, 1DWKA, 1DYOA, 1E1HB, 1E29A, 1E6IA, 1E7LA, 1EAJA, 1EAQA, 1EARA, 1EAZA, 1EB6A, 1ED1A, 1EEEXE, 1EI7A, 1ELKA, 1ELWA, 1EUWA, 1EVXA, 1EW4A, 1EX2A, 1EXRA,

1EYHA, 1EYVA, 1EZSA, 1F1EA, 1F1MA, 1F2TA, 1F2TB, 1F35A, 1F46A, 1F7DA,
1F7LA, 1F86A, 1FAOA, 1FAZA, 1FGYA, 1FITA, 1FJRA, 1FL0A, 1FM0E, 1FPOA,
1FR9A, 1FVGA, 1FZVA, 1G12A, 1G1TA, 1G3KA, 1G6GA, 1G8KB, 1GAKA,
1GHEA, 1GMLA, 1GMXA, 1GNYA, 1GO3E, 1GO3F, 1GP0A, 1GPQA, 1GR3A,
1GS9A, 1GU2A, 1GUIA, 1GV2A, 1GV8A, 1GVJA, 1GWMA, 1GXSB, 1GYUA,
1H32B, 1H3QA, 1H4AX, 1H4XA, 1H6HA, 1H97A, 1H9MA, 1HE1A, 1HKFA, 1HRUA,
1HSSA, 1HTWA, 1HXIA, 1HY5A, 1HZTA, 1I0RA, 1I12A, 1I1JA, 1I2HA, 1I4UA,
1I5NA, 1I7KA, 1I8AA, 1IBYA, 1ID0A, 1ID1A, 1IDPA, 1IHKA, 1IIBA, 1IJYA,
1IKOP, 1IKTA, 1IM5A, 1IOOA, 1IQ6A, 1IU1A, 1IWPB, 1IWPE, 1IX2A, 1IXLA,
1J0PA, 1J24A, 1J2RA, 1J30A, 1J3AA, 1J3WA, 1J5YA, 1J83A, 1J8RA, 1J98A,
1J9BA, 1JB0D, 1JB0F, 1JB0L, 1JB3A, 1JD5A, 1JERA, 1JF8A, 1JHGA, 1JHJA,
1JHSA, 1JIGA, 1JIWI, 1JKEA, 1JL7A, 1JMVA, 1JNPA, 1JNRB, 1JR8A, 1JVWA,
1JYHA, 1K0ZA, 1K12A, 1K1EA, 1K2EA, 1K2XB, 1K4NA, 1K4ZA, 1K5NB, 1K6ZA,
1KB9E, 1KB9G, 1KCQA, 1KDKA, 1KGDA, 1KHCA, 1KHIA, 1KJNA, 1KL9A,
1KLLA, 1KLXA, 1KMTA, 1KNCA, 1KNGA, 1KNMA, 1KOEa, 1KPTA, 1KQ6A,
1KR4A, 1KT6A, 1KTGA, 1KXGA, 1L2HA, 1L3KA, 1L3PA, 1L6PA, 1LF7A, 1LG7A,
1LK3A, 1LO7A, 1LQ9A, 1LQVA, 1LR0A, 1LR5A, 1LU4A, 1LXIA, 1LXJA, 1LY1A,
1M0DA, 1M1FA, 1M2DA, 1M4IA, 1M55A, 1M5Q1, 1M70A, 1M9ZA, 1MC2A,
1MF7A, 1MG4A, 1MGTA, 1MHWa, 1MJKA, 1MJSA, 1MK4A, 1MK7B, 1MKZA,
1MW5A, 1MWQA, 1MY7A, 1N12A, 1N13B, 1N1FA, 1N2MA, 1N62A, 1N71A,
1N8VA, 1NB9A, 1NC7A, 1NE2A, 1NEPA, 1NG2A, 1NG6A, 1NJKA, 1NKIA, 1NP8A,
1NRJA, 1NRZA, 1NS5A, 1NTVA, 1NU0A, 1NULA, 1NWWA, 1NWZA, 1NXJA,
1NXMA, 1NYCA, 1NYKA, 1NZ0A, 1NZIA, 1O13A, 1O22A, 1O3UA, 1O4WA,
1O7IA, 1O8XA, 1O91A, 1OB8A, 1OBOA, 1OCSA, 1OCYA, 1OD3A, 1OD6A,
1ODVA, 1OFSA, 1OH0A, 1OH4A, 1OHUA, 1OI0A, 1OI4A, 1OJ5A, 1OQVA, 1OQWA,
1OR0A, 1ORJA, 1ORUA, 1OSYA, 1OU8A, 1OW1A, 1OW4A, 1OX3A, 1OXJA,
1OY3B, 1OZ9A, 1P0ZA, 1P57A, 1P5VB, 1P6OA, 1P9YA, 1PBJA, 1PDOA, 1PF5A,
1PJUA, 1PKOA, 1PL3A, 1PM4A, 1PMHX, 1PPYA, 1PQHA, 1PSRA, 1PUCA,
1PUOA, 1PVMA, 1PWAA, 1PWBA, 1PYOA, 1PZ4A, 1Q4UA, 1Q5ZA, 1Q7HA,
1Q7LA, 1Q8DA, 1Q9UA, 1QBZA, 1QF8A, 1QFOA, 1QHQA, 1QHXA, 1QJPA,
1QKRA, 1QQRA, 1QSMA, 1QU1A, 1QV1A, 1QW2A, 1QWDA, 1QWIA, 1R0UA,

1R1MA, 1R29A, 1R4VA, 1R62A, 1R6CX, 1R75A, 1R9HA, 1R9WA, 1RDO1, 1REGX,
1REWB, 1RFSA, 1RG8A, 1RIWC, 1RJ8A, 1RK8B, 1RKIA, 1RLHA, 1RLIA, 1RLJA,
1ROCA, 1RTTA, 1RUTX, 1RW1A, 1RXDA, 1RXQA, 1RY9A, 1S0PA, 1S3MA,
1S4CA, 1S4KA, 1S5AA, 1S5UA, 1S67L, 1S7ZA, 1S99A, 1SAUA, 1SBXA, 1SBZA,
1SD4A, 1SEDA, 1SENA, 1SF9A, 1SFXA, 1SH8A, 1SHUX, 1SJWA, 1SJYA, 1SMOA,
1SR4C, 1SRAA, 1SS4A, 1SVYA, 1SX7A, 1SZ7A, 1SZHA, 1T1JA, 1T2WA, 1T3YA,
1T4WA, 1T6AA, 1T6T1, 1T6UA, 1T82A, 1T92A, 1TC5A, 1TD4A, 1TIQA, 1TJOA,
1TK9A, 1TP5A, 1TP6A, 1TQGA, 1TR0A, 1TT8A, 1TU1A, 1TU9A, 1TUAA,
1TUHA, 1TUVA, 1TUWA, 1TVGA, 1TWUA, 1TX4A, 1U14A, 1U53A, 1U5DA,
1U5FA, 1U61A, 1U69A, 1U6TA, 1U79A, 1U7IA, 1U7KA, 1U7PA, 1U9DA, 1UB4A,
1UB9A, 1UCDA, 1UEBA, 1UFYA, 1UJBA, 1UKFA, 1UKKA, 1UKUA, 1ULKA,
1UNQA, 1USCA, 1UT1A, 1UT7A, 1UUNA, 1UUYA, 1UWWA, 1UXOA, 1UZ3A,
1V0AA, 1V1HA, 1V30A, 1V37A, 1V4PA, 1V70A, 1V74A, 1V8CA, 1V8HA, 1V8YA,
1V96A, 1V9YA, 1VC1A, 1VCDA, 1VDMA, 1VH5A, 1VHZA, 1VIMA, 1VJ2A,
1VJEA, 1VKAA, 1VKCA, 1VKEA, 1VKFA, 1VKIA, 1VKKA, 1VL7A, 1VLAA,
1VMBA, 1VMDA, 1VMHA, 1VPMMA, 1VQ2A, 1VQSA, 1VR3A, 1VR7A, 1VR8A,
1VSRA, 1VYIA, 1VYKA, 1W0NA, 1W1HA, 1W2WB, 1W4SA, 1W7OA, 1W8IA,
1W94A, 1W9AA, 1W9EA, 1WC2A, 1WDJA, 1WEHA, 1WJXA, 1WKAA, 1WKOA,
1WKQA, 1WL8A, 1WLJA, 1WLTA, 1WLUA, 1WN2A, 1WNAA, 1WNYA, 1WOLA,
1WOUA, 1WPBA, 1WPNA, 1WQGA, 1WS8A, 1WVHA, 1WVQA, 1WWIA, 1WWRA,
1WWZA, 1X0GA, 1X0TA, 1X3KA, 1X6OA, 1X6ZA, 1X8QA, 1X91A, 1X9UA,
1XA8A, 1XBIA, 1XERA, 1XG0C, 1XHNA, 1XJCA, 1XJTA, 1XJUA, 1XKPB,
1XKPC, 1XODA, 1XQAA, 1XRKA, 1XS1A, 1XSVA, 1XT5A, 1XTEA, 1XU1A,
1XVWA, 1Y07A, 1Y0BA, 1Y0GA, 1Y0HA, 1Y12A, 1Y43B, 1Y5HA, 1Y60A, 1Y63A,
1Y7RA, 1Y81A, 1Y88A, 1Y93A, 1Y9IA, 1Y9KA, 1Y9QA, 1Y9TA, 1YA9A, 1YB0A,
1YB3A, 1YBTA, 1YE8A, 1YF9A, 1YGTA, 1YHFA, 1YJ7A, 1YLFA, 1YLLA, 1YLMMA,
1YLLA, 1YLMA, 1YLXA, 1YNBA, 1YOAA, 1YOCMA, 1YPYA, 1YQ5A, 1YQHA, 1YRKA, 1YSRA,
1YTLA, 1YWMA, 1YWQA, 1YYVA, 1YZ1A, 1YZFA, 1Z2UA, 1Z2WA, 1Z3AA,
1Z3EA, 1Z3QA, 1Z4EA, 1Z4RA, 1Z67A, 1Z6BA, 1Z6MA, 1Z6NA, 1ZBFA, 1ZC3B,
1ZCEA, 1ZE3H, 1ZHSA, 1ZHVA, 1ZK5A, 1ZK8A, 1ZKIA, 1ZLDA, 1ZMAA, 1ZOQA,
1ZPSA, 1ZSOA, 1ZUOA, 1ZVPA, 1ZX8A, 2A0BA, 2A15A, 2A1IA, 2A1VA, 2A3MA,

2A50B, 2A5DB, 2A67A, 2A9IA, 2A9SA, 2AALA, 2AANA, 2AG4A, 2AJ6A, 2AJ7A, 2AJGA, 2ANXA, 2AO9A, 2APLA, 2AR1A, 2AR5A, 2ARCA, 2ARKA, 2ASFA, 2AUWA, 2AXWA, 2AZWA, 2B06A, 2B0AA, 2B0VA, 2B18A, 2B1KA, 2B1LA, 2B1XB, 2B1YA, 2B4AA, 2B59B, 2B5GA, 2B71A, 2B7KA, 2B8MA, 2BBAA, 2BBEA, 2BBRA, 2BC3A, 2BCMA, 2BDRA, 2BFWA, 2BJNA, 2BK9A, 2BKMA, 2BL2A, 2BM5A, 2BMOB, 2BNLA, 2BNMA, 2BRFA, 2BSFA, 2BSJA, 2BT6A, 2BZ1A, 2BZVA, 2C21A, 2C2IA, 2C2PA, 2C4BA, 2C5LB, 2C8SA, 2C92A, 2C95A, 2C9WA, 2CALA, 2CARA, 2CAYA, 2CBOA, 2CC0A, 2CCMA, 2CE2X, 2CF7A, 2CHCA, 2CIUA, 2CJ4A, 2CJSA, 2CJTA, 2CKKA, 2CLBA, 2CM5A, 2CNTA, 2CO3A, 2CVBA, 2CVDA, 2CVEA, 2CW9A, 2CWZA, 2CX1A, 2CXCA, 2CXXA, 2CXYA, 2CY2A, 2CY5A, 2CYEA, 2CYJA, 2CZ4A, 2D0OB, 2D1PB, 2D27A, 2D48A, 2D4PA, 2D58A, 2D59A, 2D5MA, 2DBBA, 2DC3A, 2DC4A, 2DDZA, 2DG8A, 2DJ6A, 2DJFA, 2DKOA, 2DKOB, 2DP9A, 2DRVA, 2DSTA, 2DT4A, 2DTCA, 2DTJA, 2DVKA, 2DWKA, 2DXAA, 2DXQA, 2DY0A, 2E1NA, 2E2AA, 2E2DC, 2E5YA, 2E7PA, 2E7VA, 2E8EA, 2E9XA, 2E9XB, 2EB1A, 2EBYA, 2ECUA, 2EGJA, 2EH3A, 2EHPA, 2EI5A, 2EISA, 2EJ8A, 2EJNA, 2EK5A, 2EKNA, 2EO4A, 2ERVA, 2ETSA, 2EULA, 2EV0A, 2EW0A, 2EWCA, 2EXUA, 2F01A, 2F06A, 2F08A, 2F0CA, 2F1FA, 2F22A, 2F46A, 2F4IA, 2F4PA, 2F5GA, 2F5JA, 2F62A, 2F6EA, 2F9FA, 2F9HA, 2FA1A, 2FA5A, 2FB6A, 2FBHA, 2FBNA, 2FCJA, 2FCKA, 2FCLA, 2FCOA, 2FCWA, 2FD5A, 2FE1A, 2FEXA, 2FG1A, 2FG9A, 2FGCA, 2FH5A, 2FHPA, 2FHZA, 2FI1A, 2FI9A, 2FJ8A, 2FJRA, 2FK5A, 2FK9A, 2FKBA, 2FL4A, 2FM8A, 2FOMB, 2FP1A, 2FP7B, 2FPRA, 2FR2A, 2FR5A, 2FREA, 2FRGP, 2FS2A, 2FSRA, 2FSXA, 2FTRA, 2FULA, 2FUPA, 2FVVA, 2FW5A, 2FWHA, 2FWTA, 2FY6A, 2FYGA, 2FZPA, 2FZSA, 2G1UA, 2G2SB, 2G3AA, 2G3RA, 2G3WA, 2G5RA, 2G7SA, 2G84A, 2GA1A, 2GC7D, 2GECA, 2GEDA, 2GENA, 2GEYA, 2GHTA, 2GIAB, 2GIYA, 2GJ3A, 2GJAA, 2GJVA, 2GKGA, 2GKPA, 2GL9A, 2GLZA, 2GMYA, 2GRRB, 2GS5A, 2GU3A, 2GUDA, 2GUIA, 2GWFA, 2GWMA, 2GWOA, 2GXGA, 2GYQA, 2GZQA, 2H0BA, 2H0DB, 2H1CA, 2H1TA, 2H2TB, 2H30A, 2H3LA, 2H5CA, 2H5NA, 2H88C, 2H88D, 2H8EA, 2H9UA, 2HA8A, 2HBOA, 2HC8A, 2HCMA, 2HD9A, 2HDIB, 2HDVA, 2HEWF, 2HEYG, 2HH6A, 2HHGA, 2HHZA, 2HIYA, 2HKVA, 2HLJA, 2HMA, 2HNGA, 2HP4A, 2HPLA, 2HPSA, 2HQ4A, 2HQ7A, 2HQ9A, 2HQSC,

2HSBA, 2HT9A, 2HTDA, 2HU9A, 2HUHA, 2HUJA, 2HVWA, 2HW2A, 2HW4A, 2HWXA, 2HX0A, 2HX5A, 2HY5A, 2HY5B, 2HY5C, 2HYJA, 2HYTA, 2HZQA, 2I02A, 2I1SA, 2I39A, 2I4SA, 2I51A, 2I52A, 2I6JA, 2I74A, 2I7AA, 2I7FA, 2I7HA, 2I8BA, 2I8DA, 2I8TA, 2I9AA, 2I9CA, 2I9WA, 2IA1A, 2IA7A, 2IABA, 2IAFA, 2IAYA, 2IB0A, 2IBLA, 2IC2A, 2ICGA, 2IDCA, 2IDLA, 2IF6A, 2IFXA, 2IG8A, 2IGIA, 2IGPA, 2IIIA, 2IJLA, 2IKBA, 2IKKA, 2IL5A, 2IMJA, 2IMLA, 2IMSA, 2IMZA, 2INWA, 2IONA, 2IQJA, 2IQYA, 2ISBA, 2ISYA, 2IT9A, 2IU5A, 2IYGA, 2IYVA, 2IZ6A, 2J1AA, 2J1VA, 2J2JA, 2J3TC, 2J59B, 2J6AA, 2J6BA, 2J6PA, 2J73A, 2J85A, 2J8HA, 2J9WA, 2JA4A, 2JA9A, 2JCBA, 2JDAA, 2JDCA, 2JDJA, 2JE3A, 2JEKA, 2JI2A, 2JJ7A, 2JJSB, 2JLIA, 2JLJA, 2JLMA, 2LISA, 2NLVA, 2NMLA, 2NN5A, 2NPTA, 2NQ3A, 2NR7A, 2NRKA, 2NRRA, 2NSZA, 2NUJA, 2NV0A, 2NV4A, 2NVNA, 2NW2A, 2NWFA, 2NX2A, 2NX4A, 2NX8A, 2NYBA, 2NYIA, 2O0PA, 2O16A, 2O1AA, 2O1QA, 2O28A, 2O3BB, 2O4DA, 2O5HA, 2O5UA, 2O66A, 2O6FA, 2O6LA, 2O6PA, 2O70A, 2O7AA, 2O7MA, 2O7TA, 2O8QA, 2O90A, 2O9AA, 2OA2A, 2OAFA, 2OB0A, 2OB5A, 2OBBA, 2OD0A, 2OD4A, 2OD6A, 2ODAA, 2OEBA, 2OFCA, 2OFKA, 2OFZA, 2OGFA, 2OGGA, 2OH1A, 2OH3A, 2OHWA, 2OIFA, 2OIKA, 2OIWA, 2OIZD, 2OJ6A, 2OKFA, 2OKMA, 2OKQA, 2OKUA, 2OLMA, 2OMDA, 2OMLA, 2OMOA, 2OMZB, 2ONFA, 2ONUA, 2OOCA, 2OOJA, 2OOKA, 2OPCA, 2OPDA, 2OPKA, 2OPLA, 2OQMA, 2OQOA, 2ORWA, 2OS0A, 2OSOA, 2OTMA, 2OU3A, 2OU5A, 2OU6A, 2OUWA, 2OV0A, 2OV7A, 2OWAA, 2OWPA, 2OX6A, 2OX7A, 2OXGB, 2OYAA, 2OYOA, 2OZHA, 2OZJA, 2OZNA, 2P0BA, 2P0NA, 2P0SA, 2P12A, 2P14A, 2P19A, 2P25A, 2P39A, 2P3WA, 2P57A, 2P58C, 2P65A, 2P8GA, 2P8IA, 2PA7A, 2PAGA, 2PBRA, 2PCNA, 2PD1A, 2PD2A, 2PDRA, 2PEBA, 2PFBA, 2PFIA, 2PH0A, 2PHPA, 2PIEA, 2PIMA, 2PJSA, 2PMAA, 2PN0A, 2PN2A, 2PN6A, 2PNDA, 2PNMA, 2PPNA, 2PQVA, 2PR5A, 2PR7A, 2PRBA, 2PRVA, 2PTMA, 2PTTA, 2PTTB, 2PU9A, 2PV4A, 2PVBA, 2PWWA, 2PYQA, 2PYTA, 2Q03A, 2Q0YA, 2Q22A, 2Q24A, 2Q2HA, 2Q30A, 2Q3PA, 2Q3QA, 2Q3TA, 2Q3WA, 2Q3XA, 2Q48A, 2Q4PA, 2Q5CA, 2Q5XA, 2Q78A, 2Q7AA, 2Q7BA, 2Q82A, 2Q9QC, 2Q9RA, 2QBWA, 2QDLA, 2QE9A, 2QEBA, 2QEUA, 2QF4A, 2QFAA, 2QFEA, 2QG6A, 2QG8A, 2QHKA, 2QHQA, 2QIPA, 2QIYA, 2QJ2A, 2QJWA, 2QJZA, 2QKPA, 2QL8A, 2QLWA, 2QLXA, 2QMLA,

2QMMA, 2QN4A, 2QNGA, 2QNLA, 2QNTA, 2QPVA, 2QPWA, 2QQYA, 2QQZA,
2QS7A, 2QS9A, 2QSAA, 2QSFX, 2QSIA, 2QT5A, 2QUDA, 2QUOA, 2QUPA,
2QVKA, 2QYBA, 2QYCA, 2QZ7A, 2QZQA, 2QZTA, 2R0FA, 2R0XA, 2R16A,
2R1IA, 2R25A, 2R2AA, 2R2CA, 2R2YA, 2R32A, 2R39A, 2R47A, 2R4IA, 2R4QA,
2R5OA, 2R6QA, 2R6VA, 2R78A, 2R7HA, 2R8EA, 2RA9A, 2RBDA, 2RBGA,
2RC3A, 2RD9A, 2RDCA, 2RDGA, 2RE2A, 2RE7A, 2RETB, 2RFFA, 2RFMA,
2RFRA, 2RGQA, 2RH0A, 2RH3A, 2RHKA, 2RHMA, 2RI7A, 2RIHA, 2RIQA,
2RJ2A, 2RJLA, 2RJZA, 2RK9A, 2RKQA, 2RL8A, 2RLDA, 2SAKA, 2TNFA, 2UV4A,
2UVOA, 2UVPA, 2UWIA, 2UWNA, 2UX0A, 2V0UA, 2V1MA, 2V1OA, 2V2KA,
2V4XA, 2V57A, 2V5EA, 2V5EB, 2V5TA, 2V79A, 2V7FA, 2V8FA, 2V8QA, 2V9VA,
2VACA, 2VB1A, 2VBUA, 2VEZA, 2VGXA, 2VH3A, 2VI7A, 2VIFA, 2VJ2A, 2VJWA,
2VKJA, 2VKPA, 2VLIA, 2VM5A, 2VN6A, 2VNGA, 2VNLA, 2VO9A, 2VOFA,
2VSWA, 2VSZA, 2VTGA, 2VTYA, 2VV6A, 2VVDA, 2VVWA, 2VXTI, 2VXZA,
2VY8A, 2VZBA, 2VZCA, 2VZPA, 2W0GA, 2W0IA, 2W1RA, 2W1SA, 2W25A,
2W2AA, 2W31A, 2W3GA, 2W3XA, 2W4EA, 2W4LA, 2W5ZA, 2W6KA, 2W7AA,
2W86A, 2W87A, 2WAWA, 2WB3A, 2WBXA, 2WCIA, 2WCRA, 2WCWA, 2WDQC,
2WDQD, 2WDSA, 2WFBA, 2WFIA, 2WFOA, 2WFWA, 2WG7A, 2WGPA, 2WH6A,
2WJ9A, 2WJGA, 2WKKA, 2WL8A, 2WLUA, 2WLVA, 2WM8A, 2WNOA, 2WNXA,
2WNYA, 2WP7A, 2WQ4A, 2WSHA, 2WTGA, 2WV3A, 2WVFA, 2WWEA, 2WY3B,
2WZ8A, 2WZ9A, 2WZOA, 2X29A, 2X2SA, 2X32A, 2X3GA, 2X46A, 2X4HA, 2X4JA,
2X57A, 2X5NA, 2X5PA, 2X5QA, 2X5RA, 2X5YA, 2X7MA, 2XBLA, 2XC8A, 2XDHA,
2XDPA, 2XDYA, 2XEVA, 2XFDA, 2XFGB, 2XFVA, 2XG5B, 2XGWA, 2XHAA,
2XHFA, 2XI7A, 2XIGA, 2XJYA, 2XKIA, 2XODA, 2XOLA, 2XOMA, 2XOVA,
2XPPA, 2XRHA, 2XSEA, 2XT2A, 2XU8A, 2XVMA, 2XVOA, 2XVSA, 2XW6A,
2XW7A, 2XWSA, 2XXNA, 2XY2A, 2XZ8A, 2XZEA, 2Y0OA, 2Y28A, 2Y32A,
2Y3MA, 2Y3VA, 2Y3WA, 2Y69D, 2Y69E, 2Y6XA, 2Y71A, 2Y78A, 2Y8GA, 2YBYA,
2YFDA, 2YFUA, 2YG2A, 2YGNA, 2YGOA, 2YH5A, 2YH6A, 2YILA, 2YKZA,
2YMSA, 2YNQA, 2YNYA, 2YNZA, 2YOPA, 2YP6A, 2YSKA, 2YVEA, 2YVOA,
2YWIA, 2YWLA, 2YXBA, 2YYHA, 2YZIA, 2YZJA, 2YZKA, 2Z08A, 2Z0BA,
2Z0TA, 2Z0XA, 2Z10A, 2Z14A, 2Z15A, 2Z3HA, 2Z3RA, 2Z45A, 2Z51A, 2Z6DA,
2Z6OA, 2Z98A, 2ZCAA, 2ZCMA, 2ZDPA, 2ZEXA, 2ZF9A, 2ZFDDB, 2ZK9X, 2ZMFA,

2ZNRA, 2ZNZA, 2ZOPA, 2ZOUA, 2ZQMA, 2ZQNA, 2ZS0B, 2ZS0D, 2ZSCA, 2ZVYA,
2ZX0A, 2ZXJA, 2ZYCA, 2ZZDA, 2ZZDB, 3A07A, 3A0YA, 3A1BA, 3A1PA, 3A2EA,
3A2ZA, 3A35A, 3A4CA, 3A57A, 3A5PA, 3A6RA, 3A6SA, 3ACAA, 3ADYA, 3AFPA,
3AG3D, 3AG3E, 3AG7A, 3AGNA, 3AGTA, 3AGYA, 3AJ4A, 3AJVA, 3AJVB,
3AK8A, 3AKBA, 3AKOA, 3AKSA, 3AQJA, 3AS5A, 3AY2A, 3AYHA, 3B02A,
3B08A, 3B33A, 3B42A, 3B48A, 3B5KA, 3B64A, 3B79A, 3B7CA, 3B8FA, 3B8LA,
3BA3A, 3BB9A, 3BCIA, 3BCWA, 3BCYA, 3BDEA, 3BDVA, 3BEDA, 3BESL,
3BF4A, 3BFQG, 3BGUA, 3BJ6A, 3BJKA, 3BJNA, 3BJOA, 3BKRA, 3BL4A,
3BLNA, 3BLZA, 3BM7A, 3BMZA, 3BN3B, 3BN7A, 3BODA, 3BOQA, 3BQAA,
3BQEA, 3BQXA, 3BRCA, 3BROA, 3BSWA, 3BT5A, 3BVFA, 3BWLA, 3BWUD,
3BWVA, 3BWZA, 3BX4B, 3BY8A, 3BYQA, 3BZ6A, 3BZHA, 3BZTA, 3C1DA,
3C1QA, 3C5KA, 3C6VA, 3C7MA, 3C7XA, 3C85A, 3C8IA, 3C8LA, 3C9PA, 3CANA,
3CAOA, 3CBNA, 3CCGA, 3CEBA, 3CEWA, 3CEXA, 3CF4G, 3CGGA, 3CHBD,
3CHMA, 3CI3A, 3CI6A, 3CITA, 3CJDA, 3CJEA, 3CJNA, 3CK1A, 3CK2A, 3CLJA,
3CM3A, 3CNEA, 3CNGA, 3CNHA, 3CNIA, 3CNUA, 3CNVA, 3CNXA, 3CO5A,
3COLA, 3COOA, 3CP3A, 3CP5A, 3CQBA, 3CQRA, 3CQVA, 3CRYA, 3CT5A,
3CT6A, 3CT8A, 3CU3A, 3CVBA, 3CWFA, 3CX2A, 3CXKA, 3CYIA, 3CYPB,
3CZ1A, 3CZ6A, 3CZXA, 3CZZA, 3D01A, 3D06A, 3D0FA, 3D0JA, 3D1BA, 3D1PA,
3D22A, 3D32A, 3D3BA, 3D3SA, 3D4EA, 3D4RA, 3D5PA, 3D6MA, 3D6WA, 3D79A,
3D7JA, 3D82A, 3D89A, 3D8PA, 3D9NA, 3D9RA, 3D9XA, 3DA0A, 3DA7A, 3DALA,
3DB0A, 3DB7A, 3DBOB, 3DCMX, 3DCXA, 3DD7A, 3DDCB, 3DEOA, 3DF8A,
3DFGA, 3DLOA, 3DLQI, 3DM8A, 3DMCA, 3DMNA, 3DN7A, 3DNSA, 3DNXA,
3DO8A, 3DOUA, 3DP1A, 3DPJA, 3DQGA, 3DQYA, 3DRZA, 3DSBA, 3DT5A,
3DXEA, 3DZAA, 3E0HA, 3E0ZA, 3E3UA, 3E4GA, 3E4VA, 3E6QA, 3E7HA, 3E8MA,
3E8OA, 3E8PA, 3E99A, 3E9VA, 3EBEA, 3EBNA, 3EBTA, 3EBYA, 3EC6A, 3ECFA,
3ECHA, 3ECYA, 3EDOA, 3EEAA, 3EEFA, 3EEHA, 3EERA, 3EF4A, 3EF8A,
3EFAA, 3EFYA, 3EGNA, 3EHCA, 3EHDA, 3EHGA, 3EJGA, 3EJKA, 3EJVA,
3EK3A, 3ELKA, 3ELSA, 3EMFA, 3EMIA, 3EN8A, 3ENUA, 3EO4A, 3EO5A,
3EO6A, 3EOIA, 3ER7A, 3ES1A, 3ES4A, 3ETZA, 3EURA, 3EXNA, 3EXQA, 3EXZA,
3EY8A, 3EYEA, 3EYTA, 3EZIA, 3F08A, 3F0DA, 3F0IA, 3F14A, 3F1ZA, 3F2IA,
3F2ZA, 3F40A, 3F43A, 3F4AA, 3F4MA, 3F5BA, 3F5OA, 3F5RA, 3F62A, 3F7CA,

3F7EA, 3F7LA, 3F7SA, 3F7XA, 3F81A, 3F8KA, 3F8XA, 3F95A, 3F9SA, 3F9UA,
3F9XA, 3FAJA, 3FBUA, 3FCNA, 3FD4A, 3FDXA, 3FETA, 3FEUA, 3FF0A, 3FF2A,
3FF4A, 3FFLA, 3FFVA, 3FFYA, 3FG8A, 3FG9A, 3FGYA, 3FH1A, 3FHKA, 3FHVA,
3FI7A, 3FJ2A, 3FJSA, 3FJVA, 3FK8A, 3FKAA, 3FL2A, 3FLHA, 3FLJA, 3FM2A,
3FN5A, 3FNCA, 3FOJA, 3FOVA, 3FP5A, 3FPNA, 3FPWA, 3FRQA, 3FRRRA,
3FS5A, 3FSAA, 3FSDA, 3FSOA, 3FT1A, 3FTTA, 3FUYA, 3FVBA, 3FVWA,
3FWZA, 3FXHA, 3FY6A, 3FYNA, 3FZ4A, 3FZEA, 3FZGA, 3G0KA, 3G0MA,
3G13A, 3G14A, 3G16A, 3G2EA, 3G2SA, 3G39A, 3G46A, 3G48A, 3G5JA, 3G7PA,
3G7XA, 3G8ZA, 3G98A, 3GA3A, 3GA4A, 3GAXA, 3GBOA, 3GBWA, 3GBYA,
3GDWA, 3GDZA, 3GE3E, 3GFAA, 3GFSA, 3GGYA, 3GHJA, 3GI7A, 3GIOA,
3GK6A, 3GKMA, 3GLVA, 3GM5A, 3GMGA, 3GMXA, 3GN3A, 3GN5A, 3GP4A,
3GP6A, 3GPKA, 3GQQA, 3GQSA, 3GRDA, 3GT6A, 3GUDA, 3GUZA, 3GV1A,
3GWIA, 3GWNA, 3GWRA, 3GX8A, 3GXHA, 3GY9A, 3GYDA, 3GYKA, 3GYZA,
3GZ8A, 3GZBA, 3GZRA, 3GZXB, 3H05A, 3H0NA, 3H16A, 3H35A, 3H3HA, 3H3IA,
3H4OA, 3H50A, 3H51A, 3H5JA, 3H6QA, 3H6RA, 3H6XA, 3H79A, 3H7HA, 3H86A,
3H87A, 3H8TA, 3H8UA, 3H8ZA, 3H95A, 3H9WA, 3HA2A, 3HA9A, 3HCGA, 3HCJA,
3HD4A, 3HDEA, 3HDFA, 3HE0A, 3HF5A, 3HH1A, 3HIEA, 3HJ7A, 3HLYA, 3HM4A,
3HMCa, 3HMZA, 3HNSA, 3HNYM, 3HOIA, 3HP4A, 3HPCX, 3HPKA, 3HQXA,
3HSRA, 3HT1A, 3HTNA, 3HVAA, 3HVWA, 3HWUA, 3HX1A, 3HX8A, 3HX9A,
3HXSa, 3HYNA, 3HZ4A, 3HZPA, 3I0YA, 3I2VA, 3I3FA, 3I7DA, 3I7MA, 3I7TA,
3IBZA, 3ICUA, 3IDFA, 3IDUA, 3IE4A, 3IF4A, 3IGRA, 3IHTA, 3II2A, 3IIJA,
3IISM, 3IJMA, 3IKBA, 3ILCA, 3ILXA, 3IMKA, 3IMOA, 3IO2A, 3IP0A, 3IPRA,
3IQ2A, 3IQTA, 3IR9A, 3IRAA, 3IRBA, 3IT4A, 3IT7A, 3ITFA, 3IU5A, 3IU6A,
3IUOA, 3IV4A, 3IVPA, 3IVVA, 3IwTA, 3IWUA, 3IX3A, 3IXCA, 3IXSA, 3JQWA,
3JRNA, 3JRta, 3JRVA, 3JSJA, 3JSRA, 3JT0A, 3JTWA, 3JU3A, 3JUDA, 3JUMA,
3JUWA, 3JX9A, 3JYBA, 3JYGA, 3JYZA, 3JZ9A, 3K05A, 3K0XA, 3K0ZA, 3K12A,
3K1HA, 3K1YA, 3K21A, 3K2VA, 3K2ZA, 3K3CA, 3K44A, 3K5JA, 3K63A, 3K67A,
3K69A, 3K6QA, 3K7IB, 3K7PA, 3K8UA, 3KASB, 3KBQA, 3KBYA, 3KCPB,
3KCWA, 3KDFa, 3KDFB, 3KDGA, 3KE7A, 3KEPA, 3KEVA, 3KEYA, 3KF6A,
3KFFA, 3KGZA, 3KH0A, 3KH1A, 3KH7A, 3KK4A, 3KKFA, 3KKGA, 3KKWA,
3KLQA, 3KLUA, 3KM5A, 3KMIA, 3KMTA, 3KMVA, 3KNVA, 3KOLA, 3KOPA,

3KORA, 3KPBA, 3KQ0A, 3KQGA, 3KS4A, 3KSNA, 3KT9A, 3KTAA, 3KTAB,
3KTOA, 3KTYA, 3KU3B, 3KUUA, 3KUVA, 3KVTa, 3KWUA, 3KY8A, 3KYEA,
3KYJA, 3KYJB, 3KYZA, 3L00A, 3L11A, 3L18A, 3L1EA, 3L1NA, 3L20A, 3L2HA,
3L34A, 3L3UA, 3L4AA, 3L4HA, 3L4NA, 3L4RA, 3L51A, 3L51B, 3L6IA, 3L7XA,
3L8EA, 3L8HA, 3L8UA, 3L9BA, 3LAAA, 3LASA, 3LAXA, 3LB2A, 3LBEA, 3LEQA,
3LFHA, 3LFJA, 3LFKA, 3LFRA, 3LGBA, 3LHEA, 3LHNA, 3LIYA, 3LLOA, 3LLUA,
3LLVA, 3LMBA, 3LMFA, 3LMNA, 3LOIA, 3LPWA, 3LQ9A, 3LQBA, 3LQCA,
3LQCB, 3LQHA, 3LQNA, 3LQYA, 3LR3A, 3LURA, 3LW7A, 3LW9A, 3LWAA,
3LWCA, 3LWGA, 3LWXA, 3LX3A, 3LXRF, 3LYDA, 3LYGA, 3LYHA, 3LYUA,
3LYYA, 3LZAA, 3LZLA, 3M1IB, 3M1XA, 3M45A, 3M5BA, 3M5RA, 3M6JA, 3M7AA,
3M7FB, 3M7OA, 3M86A, 3M97X, 3M9ZA, 3MA2B, 3MAHA, 3MALA, 3MAOA,
3MC3A, 3MCFA, 3MCWA, 3MDPA, 3ME0B, 3ME7A, 3MEAA, 3MEWA, 3MF7A,
3MFNA, 3MGDA, 3MGMA, 3MK9A, 3MKLA, 3MKOA, 3ML1B, 3ML3A, 3MLAA,
3MMHA, 3MN2A, 3MNMA, 3MQHA, 3MQOA, 3MR0A, 3MSWA, 3MTQA, 3MUJA,
3MVCA, 3MVNA, 3MW6A, 3MWMA, 3MWZA, 3MXNA, 3MXNB, 3MXZA, 3MY2A,
3MZIA, 3N08A, 3N1EA, 3N4HA, 3N4JA, 3N6YA, 3N72A, 3N77A, 3N79A, 3N8KA,
3N9BA, 3NA2A, 3NBCA, 3NBMA, 3NCEA, 3NCTA, 3NECA, 3NEUA, 3NHVA,
3NJCA, 3NJNA, 3NKEA, 3NKGa, 3NKLA, 3NL9A, 3NNGA, 3NOHA, 3NOIA,
3NPDA, 3NPHB, 3NQNA, 3NQOA, 3NQZA, 3NR1A, 3NR5A, 3NRFA, 3NRHA,
3NRVA, 3NRWA, 3NSWA, 3NTKA, 3NUFA, 3NV0B, 3NVWA, 3NWCA, 3NY7A,
3NYMA, 3NZNA, 3O0LA, 3O10A, 3O1CA, 3O1KA, 3O22A, 3O2RA, 3O2UA,
3O3XA, 3O48A, 3O5QA, 3O7IA, 3O8ZA, 3OBFA, 3OBLA, 3OBQA, 3OC8A, 3OCMA,
3OCPA, 3OCQA, 3OG6A, 3OGAA, 3OGHA, 3OGNA, 3OHEA, 3OI8A, 3OIOA,
3OJ0A, 3OJ7A, 3OKQA, 3OKXA, 3OLOA, 3OMDA, 3ON2A, 3ON4A, 3ON9A,
3ONHA, 3OONA, 3OOPA, 3OOUA, 3OP6A, 3OP9A, 3OQ4A, 3OSEA, 3OSTA,
3OT2A, 3OUGA, 3OURB, 3OVKA, 3OWCA, 3OWRA, 3OXPA, 3P06A, 3P1GA,
3P2AA, 3P2TA, 3P3VA, 3P4HA, 3P4UA, 3P51A, 3P6IA, 3P8CF, 3P9AA, 3P9NA,
3P9VA, 3P9YA, 3PASA, 3PCVA, 3PDDA, 3PFYA, 3PG0A, 3PG6A, 3PH9A, 3PHXA,
3PILA, 3PIVA, 3PIWA, 3PJPA, 3PJVD, 3PJYA, 3PK1A, 3PKZA, 3PLWA, 3PM2A,
3PMCA, 3PN3A, 3PNAA, 3PNXA, 3POJA, 3PP2A, 3PP9A, 3PPBA, 3PQCA,
3PQHA, 3PR6A, 3PROB, 3PT8A, 3PVEA, 3PVHA, 3PWFA, 3PXVA, 3PYCA,

3Q0HA, 3Q20A, 3Q2EA, 3Q3FA, 3Q46A, 3Q49B, 3Q4OA, 3Q62A, 3Q63A, 3Q64A, 3Q6BA, 3Q7RA, 3QAYA, 3QB8A, 3QBMA, 3QBQA, 3QBTB, 3QC7A, 3QD7X, 3QH6A, 3QHBA, 3QHPA, 3QIOA, 3QL9A, 3QM9A, 3QNAA, 3QOOA, 3QORA, 3QP3A, 3QP4A, 3QPAA, 3QPIA, 3QR5A, 3QR7A, 3QRAA, 3QRLA, 3QS2A, 3QS3A, 3QTHA, 3QU3A, 3QUAA, 3QVEA, 3QW9A, 3QY3A, 3QYYA, 3QZBA, 3QZMA, 3QZRA, 3QZTA, 3QZXA, 3R0NA, 3R1PA, 3R24B, 3R3QA, 3R3RA, 3R5GA, 3R5ZA, 3R62A, 3R6AA, 3R6FA, 3R72A, 3R87A, 3R90A, 3R9FA, 3RDOA, 3RDRA, 3REGA, 3RFEA, 3RFQA, 3RGQA, 3RH0A, 3RHBA, 3RIOA, 3RKCA, 3RLKA, 3RLNA, 3RLSA, 3RMHA, 3RMQA, 3RMUA, 3RNQB, 3RO3A, 3ROBA, 3ROFA, 3RPFA, 3RPJA, 3RQIA, 3RRIA, 3RT2A, 3RTL A, 3RVCA, 3RWNA, 3RX6A, 3RX9A, 3RY4A, 3S06A, 3S0AA, 3S2WA, 3S3TA, 3S4EA, 3S4OA, 3S6EA, 3S6FA, 3S6LA, 3S8GB, 3S8IA, 3S8SA, 3S9DA, 3S9DB, 3S9FA, 3S9XA, 3SB1A, 3SE5A, 3SEIA, 3SFTA, 3SH4A, 3SHOA, 3SHPA, 3SI5A, 3SJ5A, 3SK2A, 3SK7A, 3SL2A, 3SL7A, 3SL9A, 3SLZA, 3SNKA, 3SO6A, 3SOEA, 3SOJA, 3SONA, 3SOYA, 3SP7A, 3STBB, 3STBC, 3SU0A, 3SUBA, 3SUKA, 3SUMA, 3SW0X, 3SXMA, 3SXUA, 3SXUB, 3SZ7A, 3T12B, 3T1OA, 3T3LA, 3T43A, 3T64A, 3T6OA, 3T7ZA, 3T8KA, 3T8YA, 3T92A, 3T9GA, 3T9OA, 3T9YA, 3TC5A, 3TD3A, 3TE8A, 3TEQA, 3TGNA, 3TGVA, 3TIPA, 3TIWA, 3TJ7A, 3TJ8A, 3TL4X, 3TLPA, 3TOWA, 3TQ5A, 3TS9A, 3TU3A, 3TUFB, 3TUOA, 3TVQA, 3TYJA, 3U12A, 3U1DA, 3U1UA, 3U2AA, 3U2KA, 3U35A, 3U3ZA, 3U4VA, 3U5SA, 3U6GA, 3U99A, 3U9JA, 3UAFA, 3UB6A, 3UC2A, 3UE2A, 3UFEA, 3UI4A, 3UIDA, 3ULLA, 3ULTA, 3UNNA, 3UNOA, 3UO3A, 3UPSA, 3URGA, 3URRA, 3USHA, 3UT8A, 3UULA, 3UV2A, 3UV9A, 3UWSA, 3UWTA, 3UX2A, 3V0RA, 3V2BA, 3V34A, 3V43A, 3V46A, 3V4GA, 3V7BA, 3V9OA, 3VA4A, 3VA9A, 3VAAA, 3VBAA, 3VBCA, 3VCXA, 3VFIA, 3VH5A, 3VHJA, 3VOQA, 3VORA, 3VPYA, 3VQLA, 3VRDA, 3VTXA, 3VU0A, 3VUBA, 3VUQA, 3VV1A, 3VVVA, 3VVYA, 3VWCA, 3VZ7A, 3VZ9B, 3W0TA, 3W1OA, 3W2ZA, 3W42A, 3W57A, 3W6BA, 3W8JA, 3W9KA, 3WA5B, 3WDCA, 3WDGB, 3WDNA, 3WE5A, 3WG3A, 3WGXA, 3WH2A, 3WHJA, 3WHRB, 3WISA, 3WJDA, 3WJTA, 3WMVA, 3WOEA, 3WOEB, 3WP0A, 3WTDA, 3WU4A, 3WURA, 3WUZA, 3WVAA, 3WVTA, 3WW0A, 3WW8A, 3WZ3A, 3WZSA, 3X0TA, 3X2MA, 3X37A, 3X37B, 3ZBDA, 3ZCBA, 3ZE3A, 3ZFPA, 3ZG1A, 3ZH5A, 3ZHGA, 3ZHNA,

3ZHOA, 3ZJAA, 3ZJBA, 3ZLMA, 3ZMDA, 3ZN4A, 3ZPNA, 3ZQKA, 3ZQSA,
3ZRDA, 3ZRIA, 3ZRXA, 3ZSJA, 3ZSUA, 3ZT9A, 3ZUCA, 3ZUIA, 3ZVSA, 3ZW5A,
3ZXNA, 3ZXOA, 3ZY7A, 3ZYWA, 3ZZYA, 4A02A, 4A0ZA, 4A1IA, 4A2VA, 4A3ZA,
4A42A, 4A4YA, 4A5NA, 4A5UA, 4A5ZA, 4A6HA, 4A6QA, 4A8UA, 4ABLA, 4AC7B,
4ACIA, 4ACJA, 4ACVA, 4AE4A, 4AFFA, 4AFMA, 4AG7A, 4AGKA, 4AIFA,
4AIWA, 4AJYB, 4AJYV, 4AKZA, 4AL0A, 4ALZA, 4AP4A, 4APOA,
4ATGA, 4AX2A, 4AXIA, 4AXOA, 4AYCA, 4AZ3B, 4AZZA, 4B0MA, 4B1MA,
4B4SA, 4B53A, 4B5CA, 4B5OA, 4B62A, 4B6IA, 4B8EA, 4B8XA, 4B94A, 4B9GA,
4B9IA, 4B9KA, 4BBKA, 4BEGA, 4BFOA, 4BGCA, 4BH5A, 4BHQA, 4BHUA,
4BI3A, 4BJ0A, 4BJJA, 4BK8A, 4BMWA, 4BOEA, 4BOPA, 4BOQA, 4BOUA,
4BPYA, 4BSXA, 4BU0A, 4BV1A, 4BVXB, 4BWCA, 4BXIA, 4BXMA, 4BYZA,
4BZPA, 4C0NA, 4C1BA, 4C1WA, 4C3EA, 4C6AA, 4C6RA, 4C6SA, 4C9RB, 4C9XA,
4C9YA, 4CA1A, 4CAHB, 4CBUG, 4CCVA, 4CDJA, 4CDOA, 4CE8A, 4CGSA,
4CHDA, 4CHEA, 4CHGA, 4CHHA, 4CICA, 4CJ2A, 4CK4A, 4CKMA, 4CNIC,
4CNLA, 4CPCA, 4CQ1A, 4CQIA, 4CRIA, 4CRRA, 4CRWB, 4CRYB, 4CS9A,
4CSBA, 4CT3A, 4CU9A, 4CUAA, 4CV4A, 4CV7A, 4CVNA, 4CVNB, 4CVRA,
4CXFA, 4CYBA, 4CZDA, 4CZDB, 4CZXB, 4D0QA, 4D0UA, 4D0YA, 4D53A,
4D5RA, 4D63A, 4D70A, 4D73A, 4DAMA, 4DB5A, 4DBBA, 4DCKA, 4DCKC,
4DGFA, 4DGHA, 4DHKA, 4DIDB, 4DIPA, 4DKAB, 4DKCA, 4DKKA, 4DLHA,
4DMIA, 4DNEA, 4DNYA, 4DOVA, 4DPZX, 4DQ9A, 4DQJA, 4DRIA, 4DT4A,
4DT5A, 4DUIA, 4DVCA, 4DVEA, 4DVKA, 4DXTA, 4DYQA, 4DZOA, 4E0AA,
4E2GA, 4E2UA, 4E6FA, 4E9EA, 4EAEA, 4EETB, 4EFVA, 4EGUA, 4EHSa, 4EIBA,
4EJOA, 4EJQA, 4EKXA, 4EMTA, 4ENFA, 4EO0A, 4EP4A, 4EP8B, 4EPZA,
4EQ8A, 4EQAA, 4EQAB, 4EQPA, 4EQQA, 4ERCA, 4ESBA, 4ESQA, 4ESYA,
4ETYA, 4EUKA, 4EUNA, 4EVDA, 4EVMA, 4EVYA, 4EW5A, 4EW7A, 4EXOA,
4EXRA, 4EYCA, 4EZGA, 4F07A, 4F0HB, 4F0WA, 4F2FA, 4F54A, 4F55A, 4F7HA,
4F8KA, 4F8LA, 4F8YA, 4FBSA, 4FCJA, 4FDFA, 4FFUA, 4FGQA, 4FHRA, 4FLBA,
4FMWA, 4FO5A, 4FQGA, 4FR9A, 4FTFA, 4FTXA, 4FVCA, 4FVDA, 4FVGA,
4FYUA, 4FYVB, 4G08A, 4G0XA, 4G29A, 4G3VA, 4G6QA, 4G6TA, 4G6XA, 4G78A,
4G79A, 4G7VS, 4G7XB, 4G91C, 4G94B, 4G9MA, 4G9SA, 4G9SB, 4G9YA, 4GA2A,
4GB5A, 4GBSA, 4GCOA, 4GDAA, 4GDZA, 4GEIA, 4GF3A, 4GFKA, 4GGFC,

4GGJA, 4GH9A, 4GIWA, 4GJ4A, 4GK6A, 4GLKA, 4GLQA, 4GN0A, 4GN4A, 4GOQA, 4GOSA, 4GQMA, 4GR6A, 4GRDA, 4GT8A, 4GUCA, 4GWBA, 4GYMA, 4GYTA, 4GZCA, 4H08A, 4H0SA, 4H2DA, 4H3UA, 4H5BA, 4H6CA, 4H7LA, 4H7WA, 4H7YA, 4H87A, 4H89A, 4HAMA, 4HATB, 4HBZA, 4HC5A, 4HCEA, 4HCJA, 4HDEA, 4HECA, 4HEMA, 4HF0A, 4HFVA, 4HGNA, 4HH5A, 4HHUA, 4HHVA, 4HHXA, 4HI4A, 4HI8A, 4HIAA, 4HKEA, 4HKHA, 4HMZA, 4HPMA, 4HQZA, 4HRGA, 4HRZA, 4HSPA, 4HU2A, 4HUTA, 4HVOA, 4HVYA, 4HW4A, 4HWMA, 4HWXA, 4HYLA, 4HYZA, 4HZ9A, 4HZ9B, 4I1KA, 4I4KA, 4I4OA, 4I6XA, 4IABA, 4IAMA, 4IAUA, 4IBGA, 4ICIA, 4ICWA, 4ID2A, 4IENA, 4IEUA, 4IGVA, 4IKBA, 4IKDA, 4IN0A, 4INWA, 4IOSA, 4IPCA, 4IPUA, 4IQHA, 4IRFA, 4IRHA, 4ITKA, 4IUMA, 4IUPA, 4IVVA, 4IWBA, 4IY0A, 4IZXA, 4J1VA, 4J2CA, 4J32A, 4J32B, 4J4ZA, 4J5RA, 4JB8A, 4JDEA, 4JDNA, 4JEAA, 4JEMA, 4JF8A, 4JFNA, 4JG4A, 4JG9A, 4JGLA, 4JGXA, 4JHYA, 4JIFA, 4JIUA, 4JMIA, 4JNHA, 4JO5A, 4JOIA, 4JOIB, 4JONA, 4JP6A, 4JPHA, 4JQFA, 4JT4A, 4JV8B, 4JVUA, 4JXEA, 4JXRA, 4JYSA, 4JZSA, 4K00A, 4K02A, 4K08A, 4K2JA, 4K7BA, 4K8WA, 4K9ZA, 4KBMB, 4KDLA, 4KDWA, 4KEFA, 4KG0A, 4KG3A, 4KHBA, 4KHBB, 4KIAA, 4KJMA, 4KL5A, 4KLXA, 4KMNA, 4KOPA, 4KQDA, 4KT3A, 4KT3B, 4KT6B, 4KTBA, 4KTWA, 4KUKA, 4KVXA, 4KWYA, 4KYUA, 4L1DA, 4L3RA, 4L3UA, 4L57A, 4L6EA, 4L8AA, 4L8IA, 4L8PA, 4L9EA, 4L9HA, 4L9NA, 4LA2A, 4LBAA, 4LD1A, 4LD6A, 4LDAA, 4LDFA, 4LDMA, 4LE1A, 4LFLB, 4LG3A, 4LGOA, 4LIZA, 4LJ1A, 4LJXA, 4LK9A, 4LKBA, 4LLDB, 4LLOA, 4LLOB, 4LLQB, 4LMIA, 4LMYA, 4LO6A, 4LOSA, 4LQBA, 4LQDA, 4LQKA, 4LQZA, 4LRJA, 4LRVA, 4LTNA, 4LUAA, 4LUKA, 4LVPA, 4LWRA, 4LWSA, 4LX3A, 4LXOA, 4LY7A, 4LZKA, 4LZLA, 4M1AA, 4M20A, 4M2OA, 4M62L, 4M6BA, 4M7RA, 4M85A, 4M91A, 4MAXA, 4MCJA, 4MD5A, 4ME9A, 4MEJA, 4MESA, 4MGEA, 4MGQA, 4MHLA, 4MI4A, 4MI7A, 4MI8A, 4MIWA, 4MJDA, 4MK6A, 4MLMA, 4MN7A, 4MN9A, 4MNNA, 4MNUA, 4MO1A, 4MPHA, 4MQ3A, 4MQDA, 4MS8C, 4MSPA, 4MT8A, 4MTMA, 4MTSA, 4MTUA, 4MUVA, 4MVEA, 4MVNA, 4MXTA, 4MYMA, 4MZ3A, 4MZCA, 4MZJA, 4N04A, 4N0HF, 4N0KA, 4N0PA, 4N1DA, 4N21A, 4N2PA, 4N31A, 4N3TA, 4N5BA, 4N5UA, 4N5XA, 4N6CA, 4N6OB, 4N6QA, 4N7CA, 4N7IA, 4N80B, 4N82A, 4N8MA, 4N8NA, 4N9OA, 4NB5A, 4NBPA, 4NBXA, 4NC2A, 4NCZA,

4NEXA, 4NF1A, 4NI2A, 4NI6A, 4NK1A, 4NKPA, 4NMUA, 4NN2A, 4NN5A, 4NNRA, 4NOAA, 4NOBA, 4NOCA, 4NOHA, 4NPFX, 4NRHB, 4NTKA, 4NTQB, 4NV4A, 4NWPB, 4NWYA, 4NXYA, 4NYHA, 4NYQA, 4O06A, 4O1RA, 4O2TA, 4O36B, 4O3VA, 4O42A, 4O64A, 4O65A, 4O6KA, 4O6UA, 4O7JA, 4O7KA, 4O8YA, 4O9GA, 4OA3A, 4OCIA, 4ODKA, 4ODPA, 4OE8B, 4OE9A, 4OEBA, 4OFFA, 4OFKA, 4OGQB, 4OGQD, 4OJ7A, 4OJUA, 4OK9A, 4OKEA, 4ONMA, 4ONRA, 4OOPA, 4OSNA, 4OTNA, 4OUHA, 4OUSA, 4OWKA, 4OX6A, 4OXWA, 4OXXA, 4OY6A, 4OY7A, 4OYDB, 4OYUA, 4OZJA, 4P09A, 4P0TA, 4P0ZA, 4P27A, 4P2IA, 4P3HA, 4P3WA, 4P55A, 4P5EA, 4P6IA, 4P82A, 4P9GA, 4P9IA, 4PBDA, 4PDCB, 4PDNA, 4PH2A, 4PH8A, 4PHJA, 4PIBA, 4PJ2A, 4PJ2B, 4PKGG, 4PKLA, 4PL9A, 4PONA, 4PP8B, 4PQ1A, 4PQDA, 4PQQA, 4PS6A, 4PSFA, 4PSYA, 4PT1A, 4PTZA, 4PUXA, 4PV2A, 4PV2B, 4PWQA, 4PWTa, 4PYAA, 4PZ3A, 4Q0YA, 4Q14A, 4Q29A, 4Q2SA, 4Q35B, 4Q53A, 4Q5EA, 4Q5WA, 4Q7OA, 4Q9BA, 4QAFA, 4QAKA, 4QB3A, 4QB5A, 4QBBA, 4QBLA, 4QBSA, 4QC6A, 4QCJA, 4QDNA, 4QE0A, 4QFTA, 4QGPA, 4QHJA, 4QHUC, 4QKDA, 4QKIA, 4QKWA, 4QLPA, 4QLPB, 4QMDA, 4QN8A, 4QP5A, 4QPLA, 4QPRA, 4QPWA, 4QQDA, 4QRLA, 4QT6A, 4QTPA, 4QUSA, 4QWOA, 4QX5A, 4QXAB, 4QXLA, 4QXZA, 4QYTA, 4R03A, 4R1HA, 4R1JA, 4R30A, 4R38A, 4R3HA, 4R3LA, 4R4KA, 4R7KA, 4R7RA, 4R82A, 4R8TB, 4R9KA, 4RA9B, 4RAYA, 4RBRA, 4RCIA, 4RFUA, 4RGUA, 4RHAA, 4RHSA, 4RHZB, 4RI1A, 4RLCA, 4RLEA, 4RLJA, 4RLJB, 4RLRA, 4RMBA, 4RN7A, 4RO3A, 4ROJA, 4ROWA, 4RPTA, 4RQAA, 4RRHA, 4RSUB, 4RT1A, 4RT5A, 4RTHA, 4RVQA, 4RW0A, 4RWHA, 4RXIA, 4RYOA, 4S1PA, 4S24A, 4S2VA, 4S3OB, 4TJVA, 4TKBA, 4TKCA, 4TMDA, 4TMPA, 4TNNA, 4TNSA, 4TPSA, 4TPVA, 4TQ2A, 4TR3A, 4TSDA, 4TSEA, 4TSHA, 4TTWA, 4TUMA, 4TV4A, 4TX5A, 4TXVB, 4TYZA, 4TZHA, 4U13A, 4U3SB, 4U5RA, 4U5WA, 4U5WB, 4U5YD, 4U65B, 4U77A, 4U9BA, 4U9NA, 4U9VB, 4UA3A, 4UAPA, 4UBRA, 4UC1A, 4UE0A, 4UE8A, 4UF1A, 4UHQA, 4UI1B, 4UITA, 4UM7A, 4UMGA, 4UNUA, 4UOSA, 4UQWA, 4USIA, 4USOA, 4UU3A, 4UURA, 4UUUA, 4UXAA, 4UYDA, 4UYIA, 4UYPA, 4UYTA, 4UZNA, 4V29A, 4V4MA, 4W4LB, 4W5XA, 4W64A, 4W6YA, 4W78A, 4W78B, 4W8HA, 4W97A, 4W9ZA, 4WBEA, 4WBJA, 4WDCA, 4WEEA, 4WF0A, 4WFCB, 4WFDB, 4WFQA, 4WH5A, 4WH9A, 4WHIA,

4WHNA, 4WHSa, 4WI0B, 4WILA, 4WJTA, 4WLIA, 4WN5A, 4WP6A, 4WP8A, 4WQKA, 4WQNA, 4WRIA, 4WRPA, 4WSFA, 4WW7B, 4WWFA, 4WXWA, 4X2HA, 4X2HB, 4X5PA, 4X8KA, 4X9UA, 4XAXA, 4XAXB, 4XB4A, 4XB6A, 4XB6B, 4XBAA, 4XBDA, 4XCHA, 4XCPA, 4XDIA, 4XDNB, 4XH7A, 4XHMA, 4XINA, 4XLTA, 4XOSA, 4XPLA, 4XPXA, 4XQAA, 4XTBA, 4XUOA, 4XURA, 4XWXA, 4XXBA, 4XY5A, 4XZGA, 4Y0CA, 4Y1SA, 4Y5OB, 4Y88A, 4Y9BA, 4Y9IA, 4YBMA, 4YCSA, 4YEPA, 4YF1A, 4YF4A, 4YFEA, 4YFJA, 4YG0A, 4YGTA, 4YGUA, 4YHVA, 4YIFA, 4YJWA, 4YN0A, 4YN9A, 4YP6A, 4YSXC, 4YSXD, 4YTKA, 4YTWB, 4YVOA, 4YWAA, 4YWQA, 4YX7C, 4YYNA, 4YZ6A, 4Z04A, 4Z2OA, 4Z4PA, 4Z79A, 4ZA6A, 4ZBGA, 4ZBHA, 4ZBYA, 4ZC4A, 4ZDSA, 4ZF7A, 4ZGFA, 4ZGMA, 4ZHBA, 4ZI0A, 4ZIEA, 4ZOTA, 4ZR7A, 4ZU4A, 4ZV0A, 4ZVCA, 4ZVFA, 4ZXAA, 5A0DA, 5A0RA, 5A1MA, 5A35A, 5A41A, 5A9DA, 5A9HA, 5AEOA, 5AHWA, 5AIGA, 5AILA, 5AIZA, 5AJ8A, 5AJJA, 5AMEA, 5AMHA, 5AUNA, 5BONA, 5BPBA, 5BPXA, 5BPZA, 5BQ5A, 5BQ8A, 5BQPA, 5BS1A, 5BTWA, 5BTYA, 5BVBA, 5BXDA, 5BXGA, 5BXYA, 5BY4A, 5BYUA, 5C2UB, 5C33A, 5C3TA, 5C4YA, 5C50A, 5C50B, 5C5RA, 5C5ZA, 5C6SA, 5C79A, 5C8GA, 5CE7A, 5CEGB, 5CJ3A, 5CPGA, 5CQEA, 5CTAA, 5CV0A, 5CVWA, 5CYVA, 5CYWB, 5D1IA, 5D1MB, 5D66A, 5D70A, 5D71A, 5D9RA, 5DGGA, 5DINA, 5DLEA, 5DMPA, 5DOFA, 5DPOA, 5DUYA, 5DVWA, 5DZOA, 5E2CA, 5E3EB, 5EE2A, 5EGKA, 5ENUA, 5FIWA, 5FJDA, 6FD1A