Predicting and Characterising Protein-Protein Complexes

Iain Hervé Moal

June 2011

Biomolecular Modelling Laboratory,
Cancer Research UK London Research Institute
and
Department of Biochemistry and Molecular Biology,
University College London

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
at the University College London.

I, Iain Hervé Moal, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Macromolecular interactions play a key role in all life processes. The construction and annotation of protein interaction networks is pivotal for the understanding of these processes, and how their perturbation leads to disease. However the extent of the human interactome and the limitations of the experimental techniques which can be brought to bear upon it necessitate theoretical approaches. Presented here are computational investigations into the interactions between biological macromolecules, focusing on the structural prediction of interactions, docking, and their kinetic and thermodynamic characterisation via empirical functions. Firstly, the use of normal modes in docking is investigated. Vibrational analysis of proteins are shown to indicate the motions which proteins are intrinsically disposed to undertake, and the use of this information to model flexible deformations upon protein-protein binding is evaluated. Subsequently SwarmDock, a docking algorithm which models flexibility as a linear combination of normal modes, is presented and benchmarked on a wide variety of test cases. This algorithm utilises state of the art energy functions and metaheuristics to navigate the free energy landscape. Information derived from Langevin dynamics simulations of encounter complex formation in the crowded cytosolic environment can be incorporated into SwarmDock and enhances its performance. Finally, a benchmark of binding free energies derived from the literature is presented. For this benchmark, a large number of molecular descriptors are derived. Machine learning methods are then applied to these in order to derive empirical binding free energy, association rate and dissociation rate functions which take account of the conformational changes which occur upon complexation.

Acknowledgements

It would be impossible for me to thank all the friends, family, colleagues and teachers who have inspired and educated me, supported and enriched me. To those I have omitted, I apologise.

I would like to thank my examiners Prof. Michael Sternberg and Dr. Andrew Martin for agreeing to review this thesis and allowing me to defend it *viva voce*.

Funding for this work was provided by Cancer Research UK, and I would like to extend my gratitude to the charity and its supporters. In particular, I would like to thank my thesis committee, Prof. Neil McDonald and Dr. Helen Walden. I would also like to thank the various members of the Biomolecular Modelling Laboratory, without any of whom my time at the London Research Institute would have been duller. The indomitable scientific officer, Raphaël Chaleil, for all the help and conversations, and the wizardly ability to compile even the most multifarious code. Xiaofan Li, a solid wall for bouncing ideas off, for the good company and friendship, and the productive collaboration. Marcin Król, for donating code and helpful discussions during the preliminary design of SwarmDock. Alexander Tournier and Özge Kürkçüoğlu for reading manuscripts and providing constructive feedback. Katie Bentley, Marc Offman, Yanlan Mao, Tammy Cheng, Mieczysław Torchała, Melda Tozluoğlu, Rudi Agius and Torbj̈orn Klatt, for all the spirited discussions, entertainment and food for thought, both at work and outside. I would also like to thank the support staff at the London Research Institute, including but not limited to Sally Leevers, Erin Fortin, Emma Rainbow and Sabina Ebbols. My greatest thanks, however, must go to Paul Bates, for placing his faith in me, taking me under his wing,

allowing me the liberty to follow my thoughts and, above all, for sharing his knowledge. For all of this, I am indebted.

I would like to extend gratefulness to the various collaborators I have had the pleasure of working with. Prof. Xiaodong Zhang at Imperial College London, Prof. Joël Janin at Université Paris-Sud, Prof. Alexandre Bonvin and Panagiotis Kastritis at Universiteit Utrect, and Prof. Zhiping Weng and Dr. Howook Hwang at the University of Massachusetts.

I would also like to thank those without whose imparted wisdom my last four years would never have happened. The Chemistry faculty of the University of Nottingham, who elevated interest to fascination. Particularly, Prof. Jonathan Hirst, under whose tutelage the field of molecular modelling was revealed in diorama, and Dr. Richard Wheatley, for supervising my Masters project. I would also like to thank the staff at The Henley College and Burnham Grammar School. Particularly Dr. Hywel Thomas, Dr. Branfield, Dr. Judge and Ms. Lilly, all of whose clarity illuminated their respective disciplines.

Last, but by no means least, I would like to thank my parents, François and Geraldine, for their love, support and stoicism. Without the curiosity they nurtured and indulged, and their encouragement, I would only be a fraction of myself, and I aspire to some day become a fraction of the parent they have been to me.

This work is dedicated to Ciara - More than six billion people on the surface of a sphere, how lucky I am to have found you.

Contents

Abstract **3**

Acknowledgements **4**

Contents **7**

List of Figures **13**

List of Tables **15**

List of Abbreviations **17**

Peer-reviewed publications **19**

1 Introduction **21**
1.1 Lucretius Vindicated: Of Atoms, Interactions, Life and Disease 21
1.2 Outline of the thesis 22
1.3 A Thesis Justified 23
1.3.1 Systems From the Ground Up 23
1.3.1.1 Cellular Logic 23
1.3.1.2 The Circuitry of Life 27
1.3.1.3 Aspirations, Tribulations and Computations . 30
1.3.2 More Immediate Applications 34
1.3.2.1 Protein-Protein Interactions as Drug Targets . 34
1.3.2.2 Other Applications 36
1.4 The Physical Basis of Reality 36
1.4.1 Quantum and Molecular Mechanics 37
1.4.1.1 The Schrödinger Equation 37
1.4.1.2 Energy Landscapes 38

1.4.1.3 Force Field Construction 43
1.4.2 Dynamics 44
1.4.2.1 Newton's Laws 44
1.4.2.2 The Simple Harmonic Oscillator 46
1.4.2.3 Normal Mode Analysis 48
1.4.3 Interaction Kinetics 50
1.4.4 Interaction Thermodynamics 53
1.4.4.1 The Second Law: A Classical Perspective . . . 54
1.4.4.2 The Second Law: A Statistical Perspective . . 55
1.4.4.3 The Gibbs Free Energy 59
1.4.4.4 Statistical Potentials 61
1.4.4.5 The Free Energy Landscape 63
1.4.4.6 Thermodynamic Cycles 64
1.4.4.7 From Free Energy to Binding Affinity 65
1.4.4.8 Binding Affinity Prediction 66
1.5 Protein-Protein Docking 71
1.5.1 A General Overview 73
1.5.2 Rigid-body Docking 73
1.5.3 The Correlation Method 74
1.5.4 Surface Matching 76
1.5.5 Guided Search 77
1.5.6 Accounting for Flexibility 78
1.5.7 Re-Ranking and Clustering 82
1.5.8 Incorporating Experimental Data 84
1.5.9 High-Throughput Docking 86
1.5.10 The CAPRI Experiment 86

2 Normal Mode Analysis and Conformational Transitions 90
2.1 Introduction 90
2.2 Methods 93
2.2.1 Normal Mode Analysis 93
2.2.1.1 The Elastic Network Model 94
2.2.1.2 Rotation-Translation-of-Blocks 95
2.2.2 Overlap 95
2.2.3 Modes in Linear Combination 96
2.2.4 Data Set 100

2.3 Results . . . 101
2.3.1 Atomistic and Coarse ENM . . . 101
2.3.2 Single Modes . . . 101
2.3.2.1 Rotations-Translation-in-Blocks Method . . . 103
2.3.3 Modes in Combination . . . 105
2.4 Discussion . . . 110

3 SwarmDock **113**
3.1 Introduction . . . 113
3.2 Methods . . . 117
3.2.1 Search Space . . . 117
3.2.2 SwarmDock: An Overview . . . 118
3.2.3 Initialisation . . . 120
3.2.4 Particle Swarm Optimisation . . . 122
3.2.4.1 Neighbourhoods . . . 123
3.2.4.2 Velocity Clamping . . . 124
3.2.4.3 Variations of PSO . . . 124
3.2.5 Local Search . . . 125
3.2.6 Energy Function . . . 127
3.2.6.1 Van der Waals and Electrostatics . . . 127
3.2.6.2 EEF1 Desolvation . . . 129
3.2.6.3 DComplex . . . 130
3.2.7 Clustering . . . 130
3.3 Results . . . 131
3.3.1 Parameterisation . . . 131
3.3.1.1 PSO Variant Selection . . . 131
3.3.1.2 Inertial Weight and Velocity Limits . . . 135
3.3.2 Bound-Bound Benchmark v2.0 . . . 137
3.3.3 The EEF1 Desolvation Term . . . 144
3.3.4 Unbound-Unbound Benchmark v2.0 . . . 146
3.3.5 Docking as a Function of Modes . . . 149
3.3.6 The DComplex Potential . . . 153
3.4 Discussion . . . 158

4 Crowding and Search Space Reduction **162**
4.1 Introduction . . . 162

4.2 Methods . . . 164
4.2.1 Data Set . . . 164
4.2.2 BioSimz . . . 164
4.2.3 Combining BioSimz and SwarmDock . . . 166
4.2.4 Wilcoxon Rank-Sum Test . . . 168
4.2.5 Surface Contact Heatmaps . . . 169
4.3 Results . . . 169
4.3.1 Uncrowded Simulations . . . 169
4.3.2 Crowded Simulations . . . 172
4.3.3 CAPRI Targets . . . 175
4.3.4 Possible Mechanistic Insights . . . 176
4.4 Discussion . . . 178

5 CAPRI **180**
5.1 Introduction . . . 180
5.2 Standard Rounds . . . 181
5.2.1 Round 15 . . . 181
5.2.2 Round 16 . . . 184
5.2.3 Round 17 . . . 185
5.2.4 Round 18 . . . 185
5.2.5 Round 19 . . . 186
5.2.6 Rounds 22 onwards . . . 187
5.3 Interface Design Rounds . . . 188
5.3.1 Round 20 . . . 190
5.3.2 Round 21 . . . 192
5.3.2.1 Molecular Descriptors . . . 193
5.3.2.2 Feature Set Validation . . . 195
5.3.2.3 Empirical Binding Score . . . 197
5.4 Discussion . . . 202

6 The Affinity Benchmark **203**
6.1 Introduction . . . 203
6.2 The Affinity Benchmark . . . 204
6.2.1 The Structures . . . 212
6.2.2 The Affinities . . . 212
6.2.3 The Functions . . . 213

6.2.4 Affinity Prediction 213
6.3 Summary 215

7 Affinity and Kinetics Prediction 216
7.1 Introduction 216
7.2 Molecular Descriptors 217
7.2.1 Statistical Potentials 219
7.2.2 Solvation and Entropy terms 220
7.2.3 Other Potentials 221
7.2.4 Other Descriptors 222
7.2.5 Unbound-Bound Descriptors 222
7.2.6 Ensemble Descriptors 223
7.3 Binding Free Energy Models 224
7.3.1 Methods 224
7.3.1.1 Random Forest 225
7.3.1.2 Multivariate Adaptive Regression Splines . . 226
7.3.1.3 Radial Basis Function Interpolation 227
7.3.1.4 M5′ Decision Tree 227
7.3.1.5 Data Sets 228
7.3.1.6 Model Evaluation 228
7.3.2 Results 231
7.3.2.1 All Features 231
7.3.2.2 Model Details 235
7.3.2.3 Feature Subsets 236
7.3.3 Discussion 238
7.4 Kinetic Rate Functions 239
7.4.1 Methods 240
7.4.1.1 Feature Selection 240
7.4.1.2 Model Selection 240
7.4.1.3 Subset Combinations 242
7.4.2 Results 242
7.4.3 Discussion 245

8 Epilogue 248
8.1 Protein-Protein Docking 249
8.2 Binding Affinity Prediction 250

8.3 Future Work 251
8.4 Concluding Remarks 252

A EEF1 Results **254**

B Rigid-Body Unbound-Unbound Results **257**

C Flexible Unbound-Unbound Results **262**

D Ligand Density Scoring Schemes **267**

Bibliography **271**

List of Figures

1.1 The Functioning of the *lac* Operon 26
1.2 The Phosphorylation Control of the Retinoblastoma Protein . 29
1.3 The Morse and Hooke Potentials 40
1.4 The Dihedral Energy Term 41
1.5 Balls on Springs 46
1.6 Particles in a 2D box 55

2.1 The Complex Between Murine IgG1,λ HC19 Antibody and Influenza Hemagglutinin 92
2.2 The Complex Between Actin and Vitamin D Binding Protein, in the Unbound, Bound and Fitted Conformation 99
2.3 The Mean Overlap Matrix Between Coarse and Fine ENM . . 100
2.4 Mean Maximum Overlap Across the Fold and Interface at 3 Levels of Resolution 102
2.5 Maximum Overlap and Respective Mode Across the Whole Fold and the Interface 104
2.6 Mean Percentage Reduction in RMSD as a Function of the Number of Modes Used. 106
2.7 Percentage Reduction in RMSD upon Regression for 130 High Flexibility Proteins. 107
2.8 Detailed Reduction in RMSD upon Regression for 30 High Flexibility Proteins. 109

3.1 An Overview of the SwarmDock Algorithm 119
3.2 Starting Positions Illustrated 121
3.3 Binding Site Location Probability for PSO Variants 132
3.4 Convergence Behaviour of RPSO 134
3.5 Global SwarmDock as a Function of Modes 151

3.6 Local SwarmDock as a Function of Modes 152
3.7 Energy as a Function of Modes 154

4.1 A Snapshot of a Crowded BioSimz Simulation 165
4.2 BioSimz Ligand Cloud and SwarmDock Starting Positions . . 167
4.3 Contact Heat Map for CAPRI Target 37 169
4.4 SwarmDock Rank Following Filtering 171
4.5 Cluster Size Difference and log *P* Values for Uncrowded Docking 173
4.6 Scored Positions for Target 32 175
4.7 Scored Positions for Target 40 177
4.8 Contact Heatmap for the CDK2-Ckshs1 complex. 178
4.9 Contact Heatmap for CAPRI target 39. 179

5.1 Prediction and Solution for CAPRI Target 37 184
5.2 Prediction and Solution for CAPRI Target 40 186
5.3 Prediction and Solution for CAPRI Target 50 187
5.4 Predicted and Experimental Binding Free Energies. 198
5.5 Distribution of Predicted Binding Free Energy. 200
5.6 Receiver Operating Characteristic Curve for Round 21 201

7.1 Regression Model Validated Set 229
7.2 Consensus Score Performance 230
7.3 Consensus Model Validated Set 231
7.4 Consensus Model All Complexes 232
7.5 MARS Features 234
7.6 Affinity Correlations with Descriptor Subsets 236
7.7 The p38 MAPK/MK2 Interactions 237
7.8 Early Stopping Surface/Curve 241
7.9 Venn Diagram of Models 242
7.10 Empirical Kinetic and Affinity Functions 246
7.11 The Kinetic Model Applied to the Benchmark 247

List of Tables

1.1 The Truth Table for the Material Nonimplication Operator. . . 25
1.2 The Multiplicity of Constrained States. 57
1.3 Empirical Binding Free Energy Functions 68

3.1 Performance of PSO Variants 131
3.2 Parameter Efficieny Using MOLS 1. 136
3.3 Parameter Efficiency Using MOLS 2. 137
3.4 Performance of Bound-Bound Docking 138
3.5 Flexible vs. Rigid-Body Unbound-Unbound Docking 146
3.6 Performance of SwarmDock with DComplex 148
3.7 The mean cluster size for the correct binding site. 150
3.8 DComplex vs. VDW & Elec Unbound-Unbound docking . . . 155
3.9 Performance of SwarmDock with DComplex 157
3.10 Comparison with Other Methods 159

4.1 External Crowder Proteins 166
4.2 P-Values for Uncrowded Simulations 170
4.3 P Values for Crowded Simulations 174

5.1 Summary of CAPRI Predictions 182
5.2 Summary of CAPRI Scoring 183
5.3 Results for CAPRI Round 20, Challenge 1 190
5.4 Results for CAPRI Round 20, Challenge 2 192

6.1 Binding Affinity Benchmark 207
6.2 Summary of Binding Affinity Benchmark 213
6.3 Empirical Energy Functions Performance 214

7.1 Kinetics Prediction Results 243

A.1 Performance With and Without EEF1 Desolvation. 254

B.1 Performance of Rigid-Body Unbound-Unbound Docking . . . 258

C.1 Performance of Flexible Unbound-Unbound Docking 263

D.1 P Values for Binding Site Enhancement 269
D.2 P Values for Binding Site Depreciation 270

List of Abbreviations

ACE	Analytic Continuum Electrostatics
ACE	Atomic Contact Energies
BPTI	Bovine Pancreatic Trypsin Inhibitor
cAMP	Cyclic Adenosine MonoPhosphate
CAP	cAMP Activator Protein
cyc*X*	Cyclin *X*
cdk*X*	Cyclin dependent kinase *X*
DNA	DeoxyriboNucleic Acid
ENM	Elastic Network Model
GPU	Graphics Processing Unit
ITC	Isothermal Titration Calorimetry
MARS	Multivariate Analysis Regression with Splines
MOLS	Mutually Orthogonal Latin Squares
NMR	Nuclear Magnetic Resonance
ODE	Ordinary Differential Equation
PDB	Protein Data Bank
PMF	Potential of Mean Force
Plk1	Polo-like kinase 1
pRb	Retinoblastoma protein
PSO	Particle Swarm Optimisation
RBF	Radial Basis Function Interpolation
RF	Random Forest
RNA	RiboNucleic Acid
RNAP	RNA Polymerase
RTB	Rotation-Translation in Blocks
SPR	Surface Plasmon Resonance
SVM	Support Vector Machine

Peer-reviewed publications

A chronological list of publications published in peer-reviewed journals during the PhD project:

- Fleishman et al. (2011)., Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *Journal of Molecular Biology*, In press.
- Moal IH*, Agius R* and Bates PA (2011)., Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, In press.
- Kastritis PL*, Moal IH*, Hwang H, Weng Z, Bates PA, Bonvin AM and Janin J (2011)., A structure-based benchmark for protein-protein binding affinity. *Protein Science*, **20**(3):482-91.
- Moal IH and Bates PA (2010)., SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *International Journal of Molecular Sciences*, **11**(10):3623-48.
- Li X*, Moal IH* and Bates PA (2010)., Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. *Proteins*, **78**(15):3189-96.

* These authors contributed equally to this work.

"No rest is rendered to the primal bodies Along the unfathomable inane; but rather, Inveterately plied by motions mixed, Some, at their jamming, bound aback and leave Huge gaps between, and some from off the blow Are hurried about with spaces small between. And all which, brought together with slight gaps, In more condensed union bound aback, Linked by their own all intertangled shapes,- These form the irrefragable roots of rocks And the brute bulks of iron, and what else Is of their kind..."

"What seems to us the hardened and condensed Must be of atoms among themselves more hooked, Be held compacted deep within, as 'twere By branch-like atoms- of which sort the chief Are diamond stones, despisers of all blows, And stalwart flint and strength of solid iron, And brazen bars, which, budging hard in locks, Do grate and scream. But what are liquid, formed Of fluid body, they indeed must be Of elements more smooth and round- because Their globules severally will not cohere: To suck the poppy-seeds from palm of hand Is quite as easy as drinking water down, And they, once struck, roll like unto the same. But that thou seest among the things that flow Some bitter, as the brine of ocean is, Is not the least a marvel... For since 'tis fluid, smooth its atoms are And round, with painful rough ones mixed therein; Yet need not these be held together hooked: In fact, though rough, they're globular besides, Able at once to roll, and rasp the sense. "

Lucretius - De Rerum Natura, Book II, 1st century BCE

Chapter 1

Introduction

1.1 Lucretius Vindicated: Of Atoms, Interactions, Life and Disease

The Latin philosopher Lucretius pre-dated the Roman Empire, yet in his 6 volume poem *De Rerum Natura* (On Natural Things), he laid forth a view that the origin of disease lies within nature and that nature is composed of discrete and finite bodies which interact with each other. With his physical theories, then unsupported by what we would now call rigorous science, he speculated that the interactions between atoms could be understood in terms of hooks and notches. Whilst he knew nothing of electrons, protons or thermodynamics, it is in consideration of these that atomic and molecular interactions can be understood and his notions vindicated. Furthermore, without knowledge of cells, viruses, proteins or evolution, he realised that life and the causative agents of disease were constructed from atoms themselves; a realisation now so well understood as to be trite. These two ideas underpin molecular biology and molecular medicine. His works were destroyed by the early Roman Church for their Epicureanism, and Lucretius was slandered and ridiculed for his unorthodox views about religion, philosophy, physics and biology. Only an unfinished version of *De Rerum Natura* survived, but it is undoubtedly his magnum opus. We have almost no biographical knowledge of his life, but we do know that he had a deep and penetrating intellect, without fear of grappling with new ideas. An expanse of time separates us from him, over a millennium and a half of which was

dominated by the protection of received wisdom and the violent oppression of heresy. Despite the ocean of time between us, I think Lucretius would feel more comfortable in the modern world than in his own era. He would immediately recognise the concept of molecular interactions and their role in biological systems, health and disease. He would stand in awe at the progress we have made and at the questions that remain unanswered, not just for the joy of learning and understanding, but also for the betterment of the human condition. It is a pleasure to live in an era where the pursuit of knowledge is not only unfettered, but encouraged and funded. Understanding biological systems will play a key role not only in finding new molecular targets for drugs against cancer and other illnesses, but also in modifying those systems to biosynthesise drugs, biofuels and novel materials. The characterisation and prediction of biomolecular interactions is the topic of my thesis, and a long-standing problem which needs to be solved before the seed of an idea in Lucretius's mind, is brought to full fruition.

1.2 Outline of the thesis

In section 1.3, I outline what has motivated my work by placing it within the wider context of the current paradigm. I will attempt to justify my thesis by showing how the accurate and efficient computational determination of structure, affinity and kinetics could revolutionise biology. I will also show how the elucidation of the structure of biomolecular assemblies can provide crucial information for rational drug design, and how docking can aid in the *de novo* design of protein-protein interactions. In section 1.4, I give a cursory review of the physical basis of molecular interactions in terms of energetics, dynamics, kinetics and thermodynamics, and show some of the conceptual frameworks within which macromolecular binding phenomena can be viewed, as well as a brief review of the current state of empirical binding free energy calculation. In section 1.5 is presented an introduction to protein-protein docking and a survey of modern approaches and algorithms.

In Chapter 2, a comprehensive analysis of the use of normal modes in protein-protein docking is presented. The ability for normal modes, on their own and in linear combination, to capture conformational changes

when proteins bind to one another is assessed at three levels of resolution, both across the whole fold and at the interface. Following this analysis, the development of a novel docking algorithm, which makes use of the particle swarm optimisation metaheuristic, is shown in Chapter 3. The algorithm, which models flexibility as a linear combination of elastic network normal modes, is extensively benchmarked. Chapter 4 shows how SwarmDock can be combined with rigid-body Langevin dynamics simulations of encounter complex formation and its performance enhanced, along with the presentation of investigations into the consequences of macromolecular crowding effects on protein-protein interactions. A summary of how SwarmDock has performed in the CAPRI blind assessment experiment is shown in Chapter 5. The performance of SwarmDock, combined with other molecular descriptors, to the selection of *de novo* designed protein-protein interactions is also shown. In Chapter 6, a benchmark of empirical binding free energies is presented, along with a large set of molecular descriptors which characterise the interactions. Finally, in Chapter 7 is outlined preliminary investigations into how machine learning and feature selection can be used to derive empirical models of binding free energy and kinetic rate constants.

1.3 A Thesis Justified

1.3.1 Systems From the Ground Up

> "That which I cannot build, I cannot understand."
> Richard Feynman

1.3.1.1 Cellular Logic

Biomolecular complexes feature in most biological processes (Gavin *et al.*, 2002; Alberts, 1998). However, knowledge of interacting partners can tell us a lot more than just the relative positions of different molecules which are colocalised. The word 'biology' is a portmanteau which shares etymological roots with the word 'logic'. This etymology refers to the application of reason to the study of living things, however life itself is rational and logic

is abundantly observed in the behaviour of living things. Since the renunciation of vitalism, it has been known that biological systems observe the laws of nature and that these laws govern the behaviour of life processes from the atomic all the way up the hierarchy of biological structures. It is increasingly realised that the true understanding of life, and the holy grail of systems biology, depends on reductionism and reconstruction - finding the fundamental elements in living systems and understanding how higher levels of organisation emerge from them, via the construction of predictive models.

Logic was established as a discipline by Aristotle, as a formal symbolic representation of correct reasoning, and has percolated through almost all fields of human thought. Once translated from natural language, the structures of philosophical arguments are based in logic. The whole theory of computation boils down to the study of that which can and cannot be calculated algorithmically using Boolean logic (Turing, 1939). Logic is widely believed to provide the foundations for the whole edifice of mathematics (Carnap, 1931), and logical arguments underpin physical reasoning; the laws of nature are themselves logical, even in the ostensibly incomprehensible quantum world (Birkhoff, 1936). All logical operations involve inputs, processing of those inputs, and outputs; a NOT gate on a computer chip, for instance, takes an electronic signal as a Boolean input, passes this signal through transistors, and outputs the inverse of the input. Life can be cast in the same light. The inputs to a human can be taken as being all that it sees, smells, feels, hears, breathes, eats and so on. Its processing units are cells, which have been shaped by natural selection, and its output is its behaviour. However, cells are not the fundamental unit of computation, as they do not obey simple rules. They are themselves logical operators - the inputs can be seen as the molecules which permeate the cell wall and the panoply of cell surface receptors which detect the external environment. The processing can be seen as the transduction of signals through biological pathways and their subsequent modulation of gene expression. Their output can be seen as their response to external stimuli, their behaviour.

I will illustrate a simple example of cellular logic using the archetype of

genetic regulation: the *lac* operon of *E. coli*. The metabolism of sugars allows the extraction of chemical potential energy and the synthesis of ATP via the glycolytic pathway. Two such sugars are glucose and lactose, the first of which is more abundant and can undergo glycolysis immediately. The metabolism of lactose however, requires additional lactose metabolising proteins. These include lactose permease (*LacY*), a membrane protein which pumps the lactose into the cell (Cohen and Monod, 1957), and β-galactosidase (*LacZ*), which cleaves the ether bond of lactose, forming glucose and galactose (Cohn and Monod, 1951). These lactose metabolising proteins are only synthesised in the presence of lactose, and their expression is supressed if the more nourishing glucose is present, even if lactose is also present. These facts can be consolidated into a logical operator with inputs p and q respectively corresponding to the presence of lactose and glucose in sufficient concentration, and with the output being the production of lactose metabolising proteins. This logical operator, known as the material nonimplication connective, is given the symbol $\not\supset$, and the truth table which surmises the mapping from its inputs to its output is given in Table 1.1.

Table 1.1: The truth table for the material nonimplication operator.

p	q	$p \not\supset q$
T	T	F
T	F	T
F	T	F
F	F	F

In light of the struggle for the survival of the fittest, the expression of these proteins only when they are beneficial is a prime example of evolutionary expediency, and it is clear that natural selection would favour such a regulatory mechanism. Nevertheless, what is this mechanism? At what level is the computation achieved? What is the equivalent of the transistors in the circuitry of life? The mechanism underpinning this logic can be understood in terms of the interactions between the molecular constituents of the cell, as revealed in the seminal work of Jacob and Monod (1961), summarised in Figure 1.1.

In the absence of lactose, the repressor protein lacI binds tightly to the protomer region of the DNA, upstream of the LacY and LacZ coding

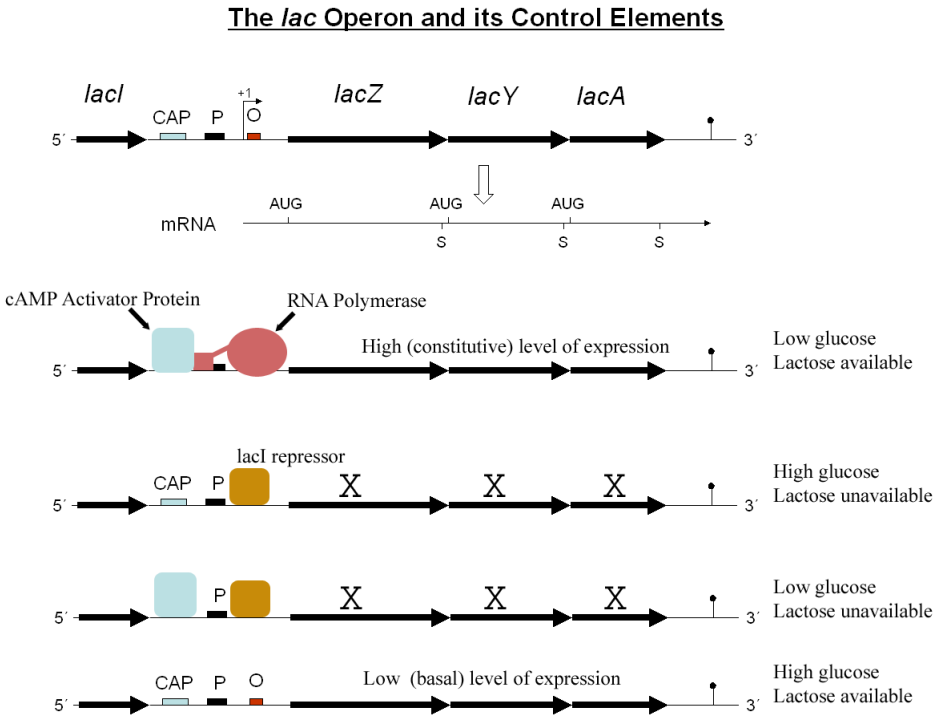

Figure 1.1: Pictoral representation of the *lac* operon and its control elements. The open reading frame, O, contains the *LacY* and *LacZ* genes. The promotor region contains the CAP binding site and an RNAP binding site, P. The CAP protein only binds when also bound to cAMP. RNAP exhibits cooperative binding to the DNA and CAP. A lacI repressor binding site ovelaps the promotor region and the open reading frame. A metabolite of lactose, allolactose, binds to lacI and precludes DNA binding. Image taken from Wikimedia Commons under the Creative Commons Attribution 2.0 Generic license (`http://commons.wikimedia.org/wiki/File:Lac_operon.png`).

region, preventing transcription by occluding RNA polymerase. However, if lactose is present, one of its metabolites, allolactose, can bind lacI and debilitate its DNA binding, allowing the approach of RNAP. Further, CAP can bind to the protomer region, but only when it is bound to cAMP. As cAMP is present at a concentration inversely proportional to the concentration of glucose, DNA bound CAP is a proxy for the absence of glucose. Due to a positive cooperative binding effect between the carboxy-terminal domain of the RNAP α subunit and CAP, the lactose metabolising genes are highly transcribed in the absence of glucose (Ptashne and Gann, 2002). Hence, the regulation of the lactose metabolising proteins boils down to molecular interactions, and the truth table which surmises their production can be deduced from these interactions and their consequences.

1.3.1.2 The Circuitry of Life

It is evident, at least for the *lac* operon system, that it is the interactions between the molecular constituents of the cell which determine cellular logic and information processing. The decades since have shown this to be the case for many more systems, of ever increasing complexity. Indeed, most biological systems are so complex, that the only way to make the information comprehensible is to consolidate it as a schematic much as electrical engineers and chemical engineers consolidate their systems as circuit diagrams and process flow diagrams. Such a graph must not only tell us which entities physically interact with each other, but must also contain the logical structure connecting the inputs to the outputs; all the information required to construct the systems truth table. For instance, in the *lac* operon, just knowing that RNAP and lacI bind the protomer region is not sufficient. Not only must we know that their binding is mutually exclusive, but we must also know that lacI binds tightly, preventing the binding of RNAP. Further, there are many other processes which must be represented, such as covalent modifications, proteolytic cleavage, enzymatic stimulation, movement between cellular compartments and so on. All such processes must be part of the repertoire of symbols employed, and such schemes have been proposed (Kohn, 1999). An example is given in Figure 1.2, using the phosphorylation control of pRb, a protein which is frequently deactivated

in cancer, as an example. The power of this view lies not just in its ability to consolidate information, but also to act as a hypothesis generator, and a method of predicting the effects of perturbing a system. A glance at Figure 1.2 will show that inhibiting the cycD:cdk4 complex will result in no phosphorylation of pRb, and consequently encourage formation of the pRb:E2F1:DP1 trimer, whilst inhibiting the cycE:cdk2 complex will result in phosphorylation of pRb at the D site only, but also encourage trimer formation.

It is also important to relate the logical structure embedded within these schematics to the underlying physical processes. Interactions between molecules are not abstract entities. Interaction events occur at a place and time, for a duration. The addition of glucose to a colony of *E. coli* doesn't result in the immediate dissociation of all lacI proteins from DNA, as a naïve interpretation of logical operators acting as binary switches might imply. Moreover, by the analogy which Monod used, these discrete states correspond to the settings on a 'thermostat' and the system dynamically adjusts itself with the components changing states at various rates. In the language of chemical kinetics, these discrete logical states correspond to a steady-state, in which the rates of conversion to and from each molecular species are equal. While Boolean or discrete state models have been useful for modelling systems as disparate as the mammalian cell cycle (Faure *et al.*, 2006), flower morphogenesis (Mendoza *et al.*, 1999), T-cell differentiation (Mendoza, 2006), cell segmentation and polarity (Sanchez *et al.*, 2008) and many more (Morris *et al.*, 2010), there are cases for which this approximation is not appropriate, such as in biological oscillators (Friesen and Block, 1984). For these situations, a more accurate model can be made using ODEs (Alon, 2006). An example of such a model is of the nucleocytoplasmic shuttling of Smads in the TGF-β pathway (Schmierer *et al.*, 2008). In this model, a system of ODEs was solved which describe the kinetics of relevant processes, including nuclear import and export, Smad homodimer and heterodimer formation, phosphorylation and dephosphorylation, receptor activation and inhibition. Most of the rates were found experimentally, with only 7 parameters needing to be fitted. These parameters were fit to four data sets describing the change in concentration of various constituents

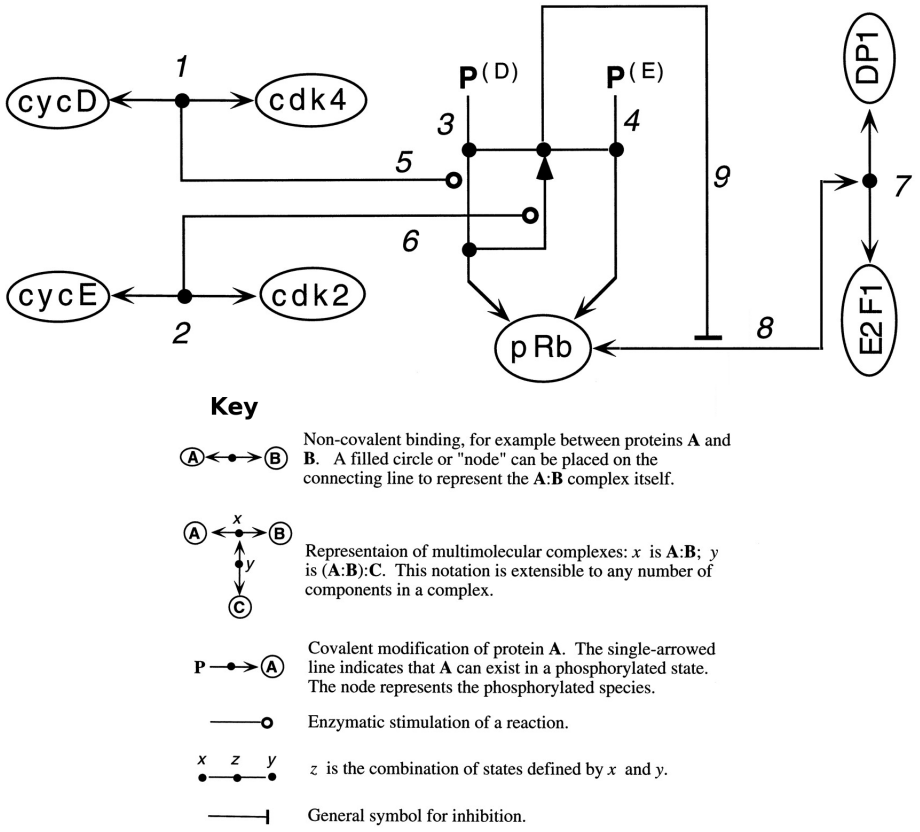

Figure 1.2: Schematic of the phosphrylation control of pRb. (1) A complex is formed between cycD and cdk4. (2) A complex is formed between cycE and cdk2. (3) pRb can be phosphorylated at the D site. (4) pRb can be phosphorylated at the E site. (5) The cycD:cdk4 complex phosphorylates the unphosphorylated pRb at the D site. (6) The cycE:cdk4 complex phosphorylates the D phosphorylated pRb at the E site, forming the hyperphosphorylated pRb. (7) Transcription factors E2F1 and DP1 form a complex. (8) The E2F1:DP1 dimer can form a trimer with pRb. (9) Hyperphosphorylated pRb inhibits the formation of the (E2F1:DP1):pRb complex. Image adapted from Kohn (1999) with permission from ASCB MBC (`http://www.molbiolcell.org/cgi/content/full/10/8/2703/`).

of the system in the nucleus and cytoplasm. The experimental data were well reproduced, and the model could accurately predict both the effect of introduction of a mutant Smad incapable of dimerising, as well as another experimental data set describing the rate of fluorescence recovery of nuclear Smads following photobleaching.

1.3.1.3 Aspirations, Tribulations and Computations

> "If people do not believe that mathematics is simple, it is only because they do not realise how complicated life is"
> John von Neumann

The approach to modelling that I have discussed is immensely powerful. It provides testable predictions, explains perturbations to biological systems, deepens our understanding of how complexity emerges from the laws of nature and indicates where areas of our knowledge are lacking. One notable achievement in this vein is the engineering of completely new metabolic pathways in micro-organisms, by integrating genes from multiple species (Ro *et al.*, 2006; Martin *et al.*, 2003). This has made possible the biosynthesis of anti-malarial drugs on a large scale and at a fraction of the cost of traditional synthetic routes, a cost which is affordable to the third world countries which need them the most.

It should come as no surprise that this approach has received a lot of attention from those involved in biomedical research. What of understanding and manipulating the biological systems of humans? Is systems biology ready to influence clinical practice? In 'The Hallmarks of Cancer', one of the most cited papers in cancer research, Hanahan and Weinberg (2000) state

> "Progress in dissecting signaling pathways has begun to lay out a circuitry that will likely mimic electronic integrated circuits in complexity and finesse, where transistors are replaced by proteins (e.g., kinases and phosphatases) and the electrons by phosphates and lipids, among others".

They predicted that the reductionist approach to cancer research would be supplanted by a systemic view. The notion that a schematic of the processes

involved in cancer could be derived is an appealing one. The prospect of taking a sample of a patient's tumour, finding the pathological rewiring of gene transcription and protein-protein interaction networks, and using the schematic to find molecular targets for personalised cancer therapy, motivates many scientists. However, despite the many successes of the systemic approach to biology, there is a long way to go before this vision is translated from fantasy to the clinic.

Firstly, humans are not *E. coli*. Our cells are compartmentalised and structured and we have sophisticated cell-cell signalling. Our transcriptome has 5 times as many genes and unlike prokaryotes we don't have a 1:1 gene to protein mapping, but many splice variants, so our proteome is bigger still. Our interactome is vast, containing an estimated 650,000 pairwise interactions between proteins (Stumpf *et al.*, 2008) and most of these are unknown (Ramani *et al.*, 2005). Accurate models need to be built which embody cellular logic. However, the construction of these models is far from trivial.

We have already seen how the building of accurate models requires a number of related pieces of information which include but are not limited to the following:

1. Binding partners; what interacts with what?
2. Affinity; how tenuous is the interaction?
3. Kinetics; how quickly do the proteins associate?
4. Physical change; for instance, are there allosteric or hindrance effects?
5. Chemical change; do covalent modifications occur and if so where?

All these pieces of information can be determined, but at a cost. There are many ways of determining which molecules interact with each other (Fu, 2004; Golemis, 2002). High-throughput technology, such as yeast-two-hybrid, phage display and tandem affinity purification have produced vast data sets, but these are plagued with false positives and false negatives and do not provide any data other than the identity of interaction partners. With the possible exception of *S. cerevisiae*, the interactomes

are incomplete. Different experimental techniques have different biases, prone to uncovering different sets of interactions and, with little spatial or temporal annotation, interaction data expressed as an undirected graph does not contain sufficient information to be of use (Kiemer and Cesareni, 2007). The inherent uncertainty and overwhelming volume of this sort of data has failed to impress some of the most revered biologists of our time. "I'm a little bit dubious about systems biology. These sort of static pictures of what interacts with what I don't find very illuminating. You can't even tell if they're true or not actually" exclaimed Tim Hunt at the 2010 Lindau meeting [1], while Sydney Brenner described much of this type of work as "low-input, high-throughput, no-output biology". There is greater cause for optimism than these quotes may suggest, and the data produced from these methods are not without merit. The topology of these networks has some value in predicting the localisation, function, processes, and gene ontology labels of proteins, as well as in the study of the modularity of biological systems (Zhang, 2009). One recent study managed to use protein-protein interaction networks to predict the prognosis of breast cancer patients with higher accuracy than the tools currently available to oncologists (Taylor *et al.*, 2009). However, it is undeniable that these interactions need confirmation and supplementary information before they become anything close to the accuracy of the well studied systems discussed in section 1.3.1.2.

A less bewildering yet still substantial array of techniques can be deployed to study binding affinity and kinetics. These range from sedimentation equilibrium, electrophoretic mobility shift assays and radio-ligand binding, to methods which provide k_{on} and k_{off} kinetic constants, such as stopped-flow fluometry, spectrophotometry, inhibition kinetics and surface plasmon resonance (SPR), to isothermal titration calorimetry (ITC), which provides binding free energy as well as how this is partitioned into entropic (ΔS) and enthalpic (ΔH) parts. The 'gold standards' of SPR and ITC are not amenable to high-throughput processing. SPR requires locating and modifying a residue by which to attach one of the binding partners to a metal surface, whilst for ITC, the calorimeter can only be used to study one interaction at a time. Both methods are limited in the range of

[1] NatureVideoChannel, 14 Sep 2010, `http://www.youtube.com/watch?v=lQzjyKqLJiQ`

affinities they can determine. While there is a significant quantity of data on affinity, kinetics and thermodynamics of biomolecular recognition, most of it languishes in the scientific literature and current databases are poor (Kumar and Gromiha, 2006; Kumar *et al.*, 2009).

Great bearing on these problems can be obtained with structural information as the consequences of the thermodynamic driving forces of biomolecular recognition are there to be seen; salt-bridge formation, electrostatic interactions, burial of hydrophobic surface area and hydrogen bonding amongst others. Functional consequences can also be derived from structure. For instance, the mutual exclusivity of two interactions which share an interaction partner can be ascertained by considering the relative positions and orientations of the binding partners, or conformational changes occurring upon binding. Similarly, the relative positions of a kinase and its substrate can indicate phosphorylation sites. Although some structural information can be derived using low-resolution technology such as small angle X-ray scattering (Svergun and Koch, 2003) or cryo-electron microscopy (Milligan *et al.*, 1984), atomic resolution can only be achieved experimentally using X-ray crystallography or NMR. However, these techniques are expensive and have limitations. Not all complexes will crystallise, especially those involving membrane proteins, transient complexes and complexes containing intrinsic disorder. NMR is limited by the size of the system that can be determined. For these reasons, the structural annotation of known interactions lags far behind the rate at which interactions are discovered by proteomics initiatives; between 2006 and 2008, fewer than 1000 heterocomplexes were deposited in the PDB, compared to over 100,000 new entries in the intAct database (Nussinov and Schreiber, 2009; Aranda *et al.*, 2010). It is clear that this data must be supplemented with computational approaches to confirm and characterise interactions, such as computational docking and using machine learning to relate structure to affinity, thermodynamics and kinetics. Such computational methods have the potential to be high-throughput and accurate, and greatly enhance systems modelling.

1.3.2 More Immediate Applications

Aside from the potential to aid in the *in silico* reconstruction of the molecular mechanisms of life, predicting and characterising protein-protein interactions also has more immediate applications. These include finding interfacial binding pockets for rational drug design, *de novo* protein-protein interface design to probe biological functions, antibody engineering and the rationalisation of mutations.

1.3.2.1 Protein-Protein Interactions as Drug Targets

Most small molecule drugs target cell surface receptors and enzymes, and many others target ion channels, nuclear envelope proteins and transporter proteins (Hopkins and Groom, 2002; Overington *et al.*, 2006; Imming *et al.*, 2006). However, there is increasing interest in inhibiting protein-protein interactions due to their importance in almost all biological processes (Berg, 2003; Arkin and Wells, 2004; Yin and Hamilton, 2005; Hershberger *et al.*, 2007; Wells and McClendon, 2007; Berg, 2008). As the binding energy was believed to be distributed approximately evenly across the whole interface, and the interface was believed to be too flat for a strong interaction with a molecule small enough to be viable as a drug, protein-protein interactions were traditionally seen as very difficult targets. However, since recent developments in the understanding of protein-protein interactions, a number of interaction inhibitors have shown promising therapeutic value (Dechantsreiter *et al.*, 1999; Wang *et al.*, 2006; Tse *et al.*, 2008; Nguyen *et al.*, 2007; Nabors *et al.*, 2007).

For instance, the anti-apoptotic members of the Bcl-2 family of proteins are commonly overexpressed in cancer cells. Their ability to enhance cancer cell viability arises from binding to pro-apoptotic Bcl-2 proteins and preventing them from activating the caspase cascade and initiating cell death (Chao and Korsmeyer, 1998). Targeting the binding site of the anti-apoptotic Bcl-2 proteins is an intriguing strategy in the development of cancer treatments. Indeed, a number of compounds have been found which bind to a hydrophobic cleft at the protein-protein interface with affinity

ranging from micromolar to subnanomolar (Wang *et al.*, 2006; Tang *et al.*, 2007a,b; Wang *et al.*, 2008; Oltersdorf *et al.*, 2005; Nguyen *et al.*, 2007; Tse *et al.*, 2008). Some of these compounds have shown powerful anti-cancer properties in mouse tumour models and have entered clinical trials. Another anti-apoptotic oncogene is the X-linked inhibitor of apoptosis, XIAP. It functions by binding to caspases 3, 7 and 9, ubiquitinating them and tagging them for degredation (Duckett *et al.*, 1998). A natural inhibitor of this interaction is the tumour suppressor Smac/DIABLO which binds XIAP. Smac/DIABLO mimics have shown promising activity against human leukaemia, breast cancer and HeLa cell lines (Rajapakse, 2007; Li *et al.*, 2004; Sun *et al.*, 2007; Nikolovska-Coleska *et al.*, 2008). Anti-cancer properties have also been shown in inhibitors which prevent Plk1 from binding to interaction motifs which serve to anchor it where its enzymatic activity is required to regulate mitosis (Strebhardt and Ullrich, 2006; McInnes *et al.*, 2005; Steegmaier *et al.*, 2007; Gumireddy *et al.*, 2005; Reindl *et al.*, 2008). The interaction between tumour suppressor p53 and oncogene MDM2 has been the target of many drug development efforts (Domling, 2008). Further, inhibition of the oncogene STAT, which transduces signals by binding to cell surface receptors via its SH2 domain, has also yielded compounds which show promising activity in mouse models and human cancer cell lines (Siddiquee *et al.*, 2007; Song *et al.*, 2005).

Structures determined using protein-protein docking can aid rational drug design. Indeed, drugs for which structural information is available are more likely to enter clinical trials (Borshell *et al.*, 2011). Knowing the location of the binding site allows the identification and exploitation of druggable cavities which are likely to inhibit the interaction (Perot *et al.*, 2010; Hajduk *et al.*, 2005). Structure-based virtual ligand screening of these sites can vastly reduce the cost of drug discovery compared to high-throughput screening (Villoutreix *et al.*, 2009). Such an approach, for instance, was used to discover a compound capable of inhibiting the interaction between the nef protein of the HIV-1 virus, and the cell surface receptors of its hosts CD4+ lymphocytes (Betzi *et al.*, 2007), and between p53 and MDM2 (Bowman *et al.*, 2007). Additionally, the structure of a complex can allow the computational identification of hot spots, regions of

the interface which contribute significantly to the overall binding affinity (Grosdidier and Fernandez-Recio, 2008; Li *et al.*, 2006). Fragment screening has shown that hot spots are often generic, and the same contacts are often used in binding molecular fragments, drug-like molecules and proteins (Vajda and Guarnieri, 2006; English *et al.*, 2001; Mattos *et al.*, 2006). Hence, it is no surprise that exploiting these hot spots by mimicking affinity imbuing contacts has become a method of choice in designing protein-protein interaction inhibitors (Trosset *et al.*, 2006; Carr *et al.*, 2005; Arkin and Wells, 2004).

1.3.2.2 Other Applications

Protein-protein docking also has use in other applications. For instance, understanding the functional consequences of mutations and combining docking and double-mutant cycles to build accurate models with confidence (Lacy *et al.*, 2005; Sivasubramanian *et al.*, 2009; Horovitz, 1996). Docking has also been used to investigate ion trafficking and electron transport (Medina *et al.*, 2008; Arnesano *et al.*, 2004). Further, *de novo* design of unnatural protein-protein interactions has been achieved by docking scaffold proteins to biological targets, followed by iterative rounds of structure and sequence optimisation (Liu *et al.*, 2007; Huang *et al.*, 2007; Fleishman *et al.*, 2011). A similar approach has had limited success in antibody engineering (Sircar and Gray, 2010; Pedotti *et al.*, 2011). These case studies highlight the potential applications of protein-protein docking. However, improvements still need to be made before these techniques can form a routine part of the molecular biologists toolkit.

1.4 The Physical Basis of Reality

In this section I aim to show in a loose manner how the principles of macromolecular interactions can be derived from general underlying principles. More comprehensive overviews of most the material presented here can be found in Atkins and De Paula (2006); Atkins and Friedman (2004); Cramer (2004); Leach (2001); Zacharias (2010b); Hinchliffe (2003) and Nussinov and

Schreiber (2009). More rigorous treatments can be found in references therein.

1.4.1 Quantum and Molecular Mechanics

The molecular world can only be truly modelled using quantum mechanics. Due to the difference in size between electrons and nuclei, nuclear and electronic effects can be separated, and the energy of a system can be derived as a function of atomic coordinates, leading to the concept of a potential energy landscape. This energy surface can be fit to simple equations. The form and parameters of these equations constitute a molecular mechanics force field, and a quick and computationally efficient way to approximately calculate the energy of even large biomolecular systems such as protein-protein complexes.

1.4.1.1 The Schrödinger Equation

Biological systems are composed of atoms, which are themselves composed of electrons and nucleons. All observables pertaining to a physical system are contained within its wavefunction, Ψ, and can be yielded by applying operators upon it. The wavefunction is a complex function whose domain is a set of all possible positions of all n electrons and nucleons of the system. When multiplied by its complex conjugate, $\Psi^*\Psi$, the wavefunction tells us the probability of finding the system in the state specified by its coordinates. The energy, E, of the system can be determined by applying the Hamiltonian operator, $\hat{H}$, to the wavefunction, yielding the time independent Schrödinger equation:

$$E\Psi = \hat{H}\Psi \tag{1.1}$$

where the Hamiltonian is composed of two further operators, the kinetic energy operator, proportional to the second derivative of the wavefunction, ∇^2, and a potential energy term, V:

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(x_1, y_1, z_1, x_2, y_2 ... x_n, y_n, z_n) \tag{1.2}$$

where m is the mass of the system and $\hbar$ is Planck's constant. Hence, if

we know the wavefunction, we can calculate the energy of the system in any configuration of electrons and nuclei.

The calculation of a wavefunction for a molecular system is very computationally demanding. However, this can be greatly alleviated by factorising the wavefunction into a nuclear and an electronic part using the Born-Oppenheimer approximation; electrons are much lighter than nuclei and hence the nuclei change position on a much slower time scale than the electrons. For any configuration of nuclei, the system can be assumed to be in the electronic ground state. This means that if we fix the coordinates of the atoms, we can calculate the energy of the electrons. In the Hartree-Fock method, the electronic wavefunction can be modelled using a liner combination of basis functions, where the basis set is a set of atomic spin orbitals centred around the nucli, i.e. $\Psi_{elec} : \mathbb{R}^n \rightarrow \mathbb{Q}$ is given by equation 1.3.

$$\Psi_{elec}(x_1, y_1, z_1, x_2, y_2 ... x_n, y_n, z_n) = \sum_{i=1}^{j} c_i \psi_i \tag{1.3}$$

where the set of j atomic orbitals ψ_i, usually a product of atom centred Gaussian radial functions and spherical harmonic angular functions, are weighted by orbital coefficients c_i. The ground state of the system is its lowest energy state. Knowing this, we can approximate the wavefunction using the variational method; varying the orbital coefficients, applying the Hamiltonian operator, and finally accepting the wavefunction once the energy has been lowered to the desired level of accuracy. There are a number of more sophisticated methods for modelling the interactions between electrons, such as density functional theory, couple-cluster and Møller-Plesset theory, as well as many different basis sets available. These are covered elsewhere (Cramer, 2004).

1.4.1.2 Energy Landscapes

Combining the Born-Oppenheimer approximation and the variational principle results in a very powerful concept, that of a potential energy surface. We can perform a systematic conformational search and solve the electronic Schrödinger equation at each atomic configuration. Knowing how the elec-

tronic energy varies as a function of the positions of nuclei, we can use this as the potential energy term in the nuclear wave equation, yielding a manifold of rovibrational states which can be probed using microwave and infrared spectroscopy. This potential energy function can be approximated as a summation of different terms of different physical origin. If these terms can be accurately modelled, it is possible to construct an energy function which does not require solving the Schrödinger equation. The following are some of these terms.

1. **Bonding terms**: Atoms can be held together by covalent bonds. The energy of a diatomic molecule of internuclear seperation r can be modelled as a Morse function, given in equation 1.4

$$V(r) = D_e(1 + e^{-a(r-r_e)})^2 \tag{1.4}$$

   where D_e is the dissociation energy, r_e is the equilibrium bond length and a is a parameter relating to the width of the energy well. Biological systems operate at a temperature at which the higher energy states are not accessible. Hence, only the bottom of the energy well needs to be modelled. This can be approximated as harmonic, taking the same form as Hooke's law, equation 1.5, governing the energy of two masses connected by springs, where k is the spring constant.

$$V(r) = \frac{1}{2}k(r - r_e)^2 \tag{1.5}$$

   These functions can be seen in Figure 1.3.

2. **Valence angle terms**: Valence shell electron pair repulsion theory, orbital hybridisation theory and their successor, molecular orbital theory, all show that bonds projecting from an atom have equilibrium valence angles separating them. Deviation from equilibrium results in a rise in energy. This rise in energy can be modelled using a modified version of Hooke's law, equation 1.5, in which the deviation from equilibrium internuclear separation between two atoms, $r - r_e$, is replaced by the deviation from equilibrium valence angle between three atoms, $\theta - \theta_e$.

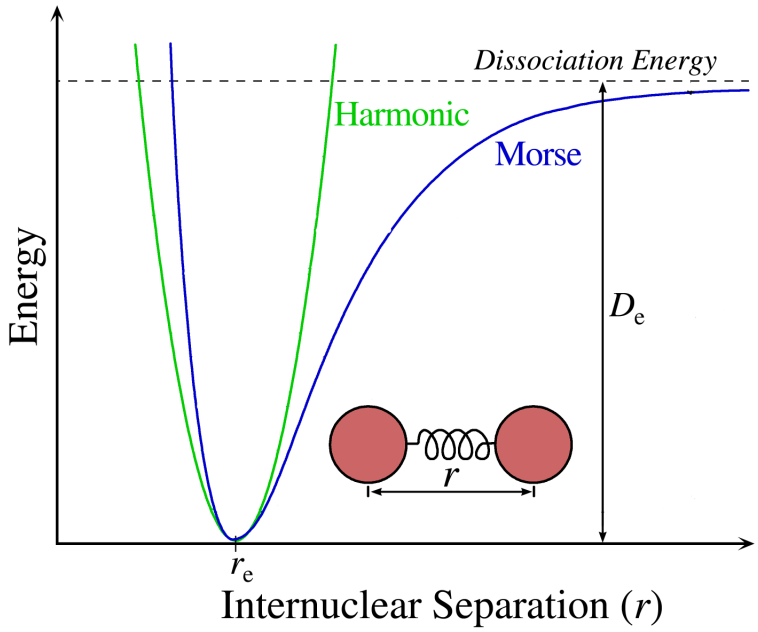

Figure 1.3: The Morse (blue) and Hooke (green) potentials for modelling covalent bonds. Image adapted from Wikimedia Commons under the Creative Commons Attribution-Share Alike 2.5 Generic license (`http://en.wikipedia.org/wiki/File:Morse-potential.png`).

3. **Dihedral angle terms**: Bonds without π character are rotatable. As we rotate about these bonds, the conformation changes from staggered to eclipsed and back, as shown in Figure 1.4. Steric hindrance renders the eclipsed state higher in energy than the staggered state. The variation in energy as a function of the dihedral angle can be modelled by equation 1.6. The height of the energy maximum is k_ϕ and the phase is δ. Of the two bonded atoms, the maximum number of donor orbitals in the hybridised valence orbitals is equal to $n + 1$, such that n represents the number of minima in $V(\phi)$. For instance, a bond involving sp^2 carbon will have $n = 3$, as shown in Figure 1.4. Unless the chemical groups attached to the bonded atoms are the same, the energy minima and maxima are not of the same magnitude. However, this is ignored in most biomolecular force fields due to the additional complexity it would engender.

$$V(\phi) = k_\phi(1 + \cos(n\phi - \delta)) \tag{1.6}$$

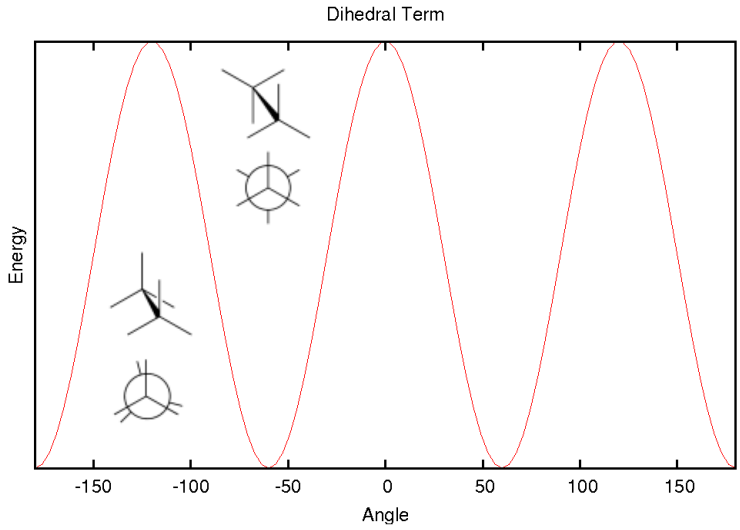

Figure 1.4: The dihedral energy as a function of angle for an sp^2-sp^2 bond. The staggered and eclipsed conformations are shown as see-saw and Neuman projections.

4. **Electrostatic terms**: Nuclei and electrons carry charges and subsequently produce electric fields surrounding them. A charged particle within an electric field has a potential energy, the electrostatic potential, which can be calculated as the product of that charge and the magnitude of the field. For two point particles, such as nuclei, separated by distance r_{ij}, and of charges q_i and q_j, the potential is given by the Coulombic energy, equations 1.7.

$$V(r) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.7}$$

where ϵ_0 is the permittivity of free space. Equation 1.7 cannot deal with electrons, which are not point particles but spatially diffuse entities. However, it can be modified such that if particle i is an electron, the charge q_i is replaced by a triple integral of the charge density over x,y and z, where the density is given by $\Psi^*\Psi$, so that the total energy is the sum of the interaction over all 'chunks' of space, weighted by the chance of finding that electron within those chunks. This presents a serious problem when attempting to model the energy without determining Ψ. The most common approximation used is to treat the system such that all the negative charge is collapsed onto the nuclei:

i.e. treat the nuclei as partial charges. In order to do this, a restrained electrostatic potential fit procedure is done. In this procedure, the partial charges are varied, whilst keeping the total charge an integer, so that the electrostatic potential isosurface at the Van der Waals radii matches as closely as possible that obtained by solving the Schrödinger equation. Thus, the Coulomb equation can be used in its point-particle form (Bayly *et al.*, 1993).

5. **Overlap-Exchange repulsion terms**: The Pauli exclusion principle dictates that no two particles can occupy the same region of space. Thus, electrons cannot collapse upon their nuclei, despite their opposite charge, and must reside in distinct orbitals surrounding the nucleus. This gives rise to the full diversity of chemical elements we see around us. As two particle approach one another, their wavefunctions overlap, resulting in an exchange repulsion which increases exponentially. Biomolecular modellers rarely calculate this as an exponential term, but instead treat it as a repulsion which scales as r_{ij}^{-12} within the Lennard-Jones equation, as explained below.

6. **Electrostatic induction terms**: The electron distribution of a molecule in an electric field does not remain fixed, but adjusts itself according to that field so as to lower the total energy due to the Debye force. Like charges tend towards each other and opposite charges repel each other. As electrons are difficult to model explicitly, this term is usually ignored by the biomolecular modelling community, except when the inducing multipole arises from spontaneous quantum fluctuations, giving rise to the London dispersion energy.

7. **London dispersion terms**: The electrons in a molecule are in constant flux, and at any moment in time there may be higher or lower electron density in one part of the molecule than at the next moment, giving rise to spontaneous dipoles or higher multipoles. These spontaneous multipoles in one molecule cause complimentary electrostatic induction in the other, lowering the energy of the system. This energy change scales as r_{ij}^{-6}.

8. **Lennard-Jones terms**: The repulsive overlap-Exchange and attactive

London dispersion terms are usually combined into the Lennard-Jones potential function, equation 1.8.

$$V(r_{ij}) = \sqrt{\epsilon_i \epsilon_j}\left[\left(\frac{R_{min_{i,j}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{i,j}}}{r_{ij}}\right)^{6}\right] \tag{1.8}$$

where $R_{min_{i,j}}$ corresponds to the minima, and $\sqrt{\epsilon_i \epsilon_j}$ is the depth of the potential well. These terms take this form due to its computational efficiency; $r_{ij}^{-12} = (r_{ij}^{-6})^2$. This is often referred to as the Van der Waals energy, as it reflects the factors implicit in the Van der Waals equation of state, a modification of the ideal gas law which takes account of the volume of molecules and the attractive forces between them.

9. **Cross terms**: All the above terms have a clear physical origin. However, additional terms can also be employed to reduce the discrepancies between the energy surface calculated using quantum mechanics and a fitted force field. One correction is the cross-term, which couples the ϕ and ψ protein backbone dihedral angles, and takes the form of nm terms given by equation 1.9 (Mackerell *et al.*, 2004).

$$V(r_{ij}) = k_{nm}(1 + \cos(n\phi + m\psi - \delta_{nm})) \tag{1.9}$$

where δ_{nm} is the phase.

1.4.1.3 Force Field Construction

For any given system, terms in section 1.4.1.2 can be fitted to a potential energy surface derived from quantum mechanics to give a reasonable fit. While parameters describing, say, a oxygen-hydrogen bond, are difference from those describing an $sp^2 - sp^2$ carbon-carbon bond, the parameters describing different O-H bonds, or O-H bonds in different systems, are often very close to one another. If it is possible to classify atoms in a system, or set of systems, such that the parameters relating to them are the same as they are in many other systems, then it would be possible to approximate the energy surface of a system whose parameters have not been derived from quantum mechanical calculations. This principle, the transferability hypothesis, makes it possible to construct a generic force

field, a set of terms and parameters derived from, for example, dipeptides, which can then be used to model tripeptides or even whole proteins. The art of force field construction is beyond the scope of this introduction and the topic has been dealt with extensively within the scientific literature. In the menagerie of modern force fields, some are general (Wang *et al.*, 2004), others are for specific classes of compound, such as heterocycles (McDonald and Jorgensen, 1998) or organic liquids (Jorgensen *et al.*, 1996). Some force fields are compatible with one another. One of the most popular set are those which are compatible with the force field released with version 22 of the molecular simulations program CHARMM (MacKerell *et al.*, 1998; Brooks *et al.*, 2009; Patel and Brooks, 2004; Patel *et al.*, 2004). The original CHARMM22 force field was developed specifically for simulating proteins. However, since then, a compatible force field was developed for nucleic acids (MacKerell *et al.*, 2000) as well as a general force fields for modelling small molecules (Vanommeslaeghe *et al.*, 2010). These force fields have proven remarkably accurate given their meagre computational requirement. As a back-of-the-envelope estimate, doing a single point energy calculation on a moderate protein of around 1,000 atoms using the cc-pVTZ basis set and the MP4 quantum mechanics method (which scales as N^6 with the number of basis functions), will take around 800 times the age of the universe. Doing a single point energy calculation on the same protein with the CHARMM22 force field takes a fraction of a second.

1.4.2 Dynamics

1.4.2.1 Newton's Laws

The temporal evolution of non-relativistic macroscopic physical systems are governed by Newton's laws of motion:

1. Without external force acting upon it, an object will remain stationary or move with constant velocity.

2. The acceleration of a body is equal to the force applied upon it divided by its mass.

3. The forces applied on each other by two bodies are equal and opposite.

For a single body in one dimension, the second law can be variously written as

$$f = ma = m\frac{dv}{dt} = m\frac{d^2x}{dt^2} = -\frac{d}{dx}V(x) \tag{1.10}$$

where x is position, v is velocity, a is acceleration, m is the mass and $V(x)$ is the potential energy. The equivalence of force and the gradient of the potential arises due to conservation of energy; any increase in kinetic energy has a corresponding decrease in potential energy and thus the rate of change of kinetic energy (the force) is equal and opposite to the rate of change of potential energy. Generalised to any number of particles in any dimensionality, Newton's equation of motion for a classical system is given by equation 1.11.

$$\mathbf{M}\frac{\partial^2}{\partial t^2}\mathbf{x}(t) = F\mathbf{x}(t) = -\nabla V(\mathbf{x}(t)) \tag{1.11}$$

where $\mathbf{M}$ is the mass matrix. As noted in section 1.4.1 these laws do not apply to molecules. Nevertheless, some molecular processes can be studied under the regime of classical mechanics, but only if the process is slow and the thermal energy of the system is considerably larger than the energy of quanta, i.e. if $k_BT >> \hbar\lambda$, where k_B is the Boltzmann constant, T is the temperature, $\hbar$ is Planck's constant and λ is the frequency of the motion in question. At 300K, classical mechanics can meaningfully model processes on a time scale above around 200fs, which is significantly lower than the time scale of most biological functions not involving the movement of electrons.

Thus, it is possible to take a protein structure, such as one given by NMR or crystallography, set up a force field and perform molecular dynamics by solving the differential equation 1.11 using a suitable integrator, such as Verlet-Stoermer numerical integration (Alder and Wainwright, 1959). Numerous technicalities need to be considered, however, assuming that the potential function is exact, then this numerical approach can be seen as an approximate solution to the exact problem. There is no general analytical solution to Newton's equation of motion. However, if we optimise the system until we find the minima, express $V(\mathbf{x}(t))$ as a Taylor series and

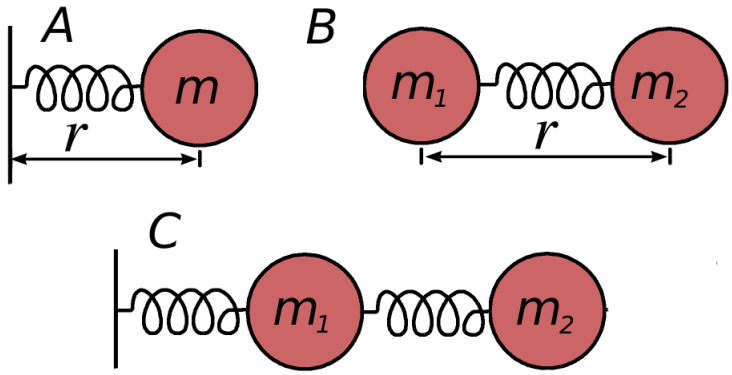

Figure 1.5: Various simple systems of masses connected by springs.

truncate it after the third term, the potential function becomes a harmonic well, with each coordinate of each particle obeying Hooke's law, equation 1.5. For such systems there are an infinite number of analytical solutions to equations 1.11, and one can obtain the exact solution to the approximate problem. Of particular interest are the set of solutions which have the same frequency, the normal modes of vibration, the acquisition of which is best illustrated with simple systems of balls on springs.

1.4.2.2 The Simple Harmonic Oscillator

Take the simple system in Figure 1.5A, which obeys Hooke's law. In this system, the force acting on the mass is linearly proportional to the extension of the spring, thus

$$f(r) = -k(r - r_e) \tag{1.12}$$

Combining 1.10 and 1.12, we get the second-order differential equation 1.13.

$$m\frac{d^2r}{dt^2} = -k(r - r_e) \tag{1.13}$$

This has the general solution

$$r = r_e + A\sin(t\sqrt{k/m}) + B\cos(t\sqrt{k/m}) \tag{1.14}$$

which repeats every $t = 2\pi\sqrt{k/m}$. The constants of integration, A and B, can be found by imposing boundary conditions. Stipulating that $r = r_e$ at time $t = 0$, then $B = 0$, so that the rightmost term becomes zero. If we then stipulate E, the total energy of the system, then $A = \sqrt{2E/k}$, by noting that when at full extension, $\sin(\sqrt{k/m}) = 1$, the kinetic energy is zero and the total energy is given by the potential energy, equation 1.5. The quantity $\sqrt{k/m}$, which has units s^{-1}, is the angular vibration frequency, ω, leaving the final form of this simple harmonic oscillator:

$$r = r_e + A\sin(t\omega) + B\cos(t\omega) \tag{1.15}$$

A slightly more complex example, with two balls and one spring, is given in Figure 1.5B. Here, the length of the spring is given by $r = x_2 - x_1$, and its extension, ξ is

$$\xi = x_2 - x_1 - r_e \tag{1.16}$$

The equation of motion for the first mass is given by

$$m_1\frac{d^2x_1}{dt^2} = k\xi \tag{1.17}$$

and for the second mass, which moves in the opposite direction:

$$m_2\frac{d^2x_2}{dt^2} = -k\xi \tag{1.18}$$

Combining 1.17 and 1.18, we can express the equations of motion in terms of a single internal coordinate.

$$\frac{d^2r}{dt^2} = -\frac{k}{m_2}\xi - \frac{k}{m_1}\xi = -k\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\xi \tag{1.19}$$

This equation is of the same form as 1.13, and thus has solutions similar to 1.14, only with the mass, m, replaced by the reduced mass, $\mu = \left(\frac{1}{m_1} + \frac{1}{m_2}\right)$. In the next section, we shall go one step further in complexity by adding a second spring. Using this as a basis, I will then derive a general form which is useful for studying larger systems as well as systems whose potential

energy function may be much more complex than balls on springs, such as one from a molecular mechanics force field.

1.4.2.3 Normal Mode Analysis

In the system given in Figure 1.5C, the first mass, at position x_1 is subject to forces originating from two springs. For these springs, the deviations from their equilibrium positions are $\xi_1 = x_1 - r_{e1}$ and $\xi_2 = x_1 - r_{e2}$. For the first spring, the force exerted upon the first mass is $-k_1\xi_1$. The force exerted by the second spring on both masses depends not just on its deviation from its equilibrium position, but also on the position of the first mass. Thus it exerts $k_2(\xi_2 - \xi_1)$ force upon the first mass and $-k_2(\xi_2 - \xi_1)$ upon the second. Newton's equations of motion for this system can be expressed in terms of two internal coordinates:

$$m_1 \frac{d^2}{dt^2}\xi_1(t) = k_2(\xi_2 - \xi_1) - k_1\xi_1 = -(k_2 + k_1)\xi_1 + k_2\xi_2 \tag{1.20}$$

$$m_2 \frac{d^2}{dt^2}\xi_2(t) = -k_2(\xi_2 - \xi_1) = k_2\xi_1 - k_2\xi_2 \tag{1.21}$$

There are many solutions to these simultaneous differential equation, including sets for which the angular frequency, ω, is the same for both particles, the normal modes. All possible harmonic motions can be expressed as a linear combination of these normal modes. These solutions are of the form

$$\xi_1(t) = A\sin(\omega t + \delta_1) \tag{1.22}$$

$$\xi_2(t) = A\sin(\omega t + \delta_2) \tag{1.23}$$

Differentiating, and then substituting 1.22 and 1.23 back in:

$$\frac{d^2}{dt^2}\xi_1(t) = -\omega^2 A\sin(\omega t + \delta_1) = -\omega^2\xi_1(t) \tag{1.24}$$

$$\frac{d^2}{dt^2}\xi_2(t) = -\omega^2 B\sin(\omega t + \delta_2) = -\omega^2\xi_2(t) \tag{1.25}$$

Substituting 1.24 and 1.25 into equations 1.20 and 1.21 yields equations 1.26 and 1.27, which are only valid when ω has one of two values.

$$-\frac{(k_1+k_2)}{m_1}\xi_1+\frac{k_2}{m_1}\xi_2=-\omega^2\xi_1 \tag{1.26}$$

$$\frac{k_2}{m_2}\xi_1-\frac{k_2}{m_2}\xi_2=-\omega^2\xi_2 \tag{1.27}$$

This system of linear equations can be written in matrix form. The mass-weighted force constant matrix is know as the Hessian.

$$\begin{pmatrix} -\frac{(k_1+k_2)}{m_1} & \frac{k_2}{m_1} \\ \frac{k_2}{m_2} & -\frac{k_2}{m_2} \end{pmatrix}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}=-\omega^2\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \tag{1.28}$$

As this is of the form $\mathbf{H}\boldsymbol{\xi}=\lambda\boldsymbol{\xi}$, the Hessian matrix can be diagonalised, yielding the change in internal coordinates as eigenvectors, $\boldsymbol{\xi}$, and the negative square of their corresponding frequencies, λ, as eigenvalues. These eigenvectors are the normal coordinates.

It should be noted that the elements in the ith row of the Hessian all have a factor of $1/m_i$, thus the matrix is described as mass weighted. Factoring this out, the jth element of the ith row contains the coefficient of the jth coordinate in the ith equation of motion, equations 1.20-1.21 in this case. As 1.20-1.21 are linear with respect to the coordinates, then this coefficient can be obtained by differentiating the ith equation of motion with respect to the jth coordinate. For instance, if $i=1$ and $j=1$, then $\frac{d}{d\xi_1}-(k_1+k_2)\xi_1+k_2\xi_2=-(k_1+k_2)$ which, mass weighted, appears in the top left of the Hessian in equation 1.28.

From here, it is easy to generalise to larger systems and systems of higher dimensionality, such as proteins in 3D. Take a system with n coordinates, such as a protein with $n/3$ atoms, where n is a multiple of 3. Then, a series of equations of motion can be constructed, one for each coordinate.

$$m_i\frac{d^2\xi_n}{dt^2}=\sum_{j=1}^{j=n}c_{i,j}\xi_j \qquad \text{for} \qquad i=1...n \tag{1.29}$$

As shown above, the coefficients $c_{i,j}$ are the derivative of the force in the ith equation of motion with respect to the jth coordinate. However, as shown in equation 1.10, the force is equal to the derivative of the potential

energy, and so the coefficients can be calculated as

$$c_{i,j} = \frac{\partial^2 V}{\partial \xi_i \partial \xi_j} \tag{1.30}$$

Now, the same line of reasoning can be employed as for the system in Figure 1.5C. The solutions in which the angular frequency are the same are

$$\xi_i(t) = C_i \sin(\omega t + \delta_i) \qquad \text{for} \qquad i = 1...n \tag{1.31}$$

Differentiating, we get

$$\frac{d^2}{dt^2}\xi_i(t) = -\omega^2 C_i \sin(\omega t + \delta_i) \qquad \text{for} \qquad i = 1...n \tag{1.32}$$

Substituting 1.31 into 1.32, and then substituting into 1.29, we end up with a series of linear equations which, as matrix form, are

$$\mathbf{H}\boldsymbol{\xi} = \lambda \boldsymbol{\xi} \tag{1.33}$$

Due to equation 1.30, the Hessian, $\mathbf{H}$, has elements

$$H_{i,j} = \frac{1}{m_j}\frac{\partial^2 V}{\partial \xi_i \partial \xi_j} \tag{1.34}$$

and can be diagonalised to yield eigenvectors $\boldsymbol{\xi}$, the normal coordinates, and a set of eigenvalues $\lambda = -\omega^2$. All conformations of the molecule in question can be expressed as a linear combination of normal coordinates, which form an orthogonal basis.

1.4.3 Interaction Kinetics

In the section 1.4.1 we saw how the energy of a physical system varies as a function of the coordinates of its constituents. In section 1.4.2.1 we saw how Newton's equation of motion can be used to model the temporal evolution of the coordinates of a system as well as its kinetic and potential energy. It is natural then to use these concepts as a basis for understanding molecular interactions. Biomolecular interactions occur in solution, constantly buffeted by water and the other molecular constituents of the cell. Interactions between molecules occur all the time, but those that are biologically

meaningful are the ones which last for a long period of time, or between molecules which, at equilibrium, are more likely to be found spatially closer to one another than expected if they were evenly distributed in space.

Take a non-covalent interaction between molecular species A and B, to form complex C:

$$A + B \leftrightharpoons C \tag{1.35}$$

The concentration of these species can be denoted $[A]$, $[B]$ and $[C]$. There are two processes which can occur; association happens when A and B cohere and form C, and dissociation occurs when C decomposes into its constituents A and B. It is clear that these rates depend on the concentration of the species. If we have twice the concentration of $[C]$, then the rate of dissociation will be double. Thus we can model the rate of dissociation as

$$\text{rate of dissociation} = k_{off}[C] \tag{1.36}$$

where constant k_{off} represents the innate disposition for the complex to dissociation. Similarly the number of association events in a given period of time is linearly proportional to $[A]$ and $[B]$, and we can define

$$\text{rate of association} = k_{on}[A][B] \tag{1.37}$$

Using these, we can express the rate of change of concentration of these species

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = k_{off}[C] - k_{on}[A][B] \tag{1.38}$$

$$\frac{d[C]}{dt} = k_{on}[A][B] - k_{off}[C] \tag{1.39}$$

The system is in equilibrium when the rate of association equals the rate of dissociation, and the concentration of all species remains constant, i.e., when equations 1.38 and 1.39 equal zero. Rearranging, we can define the dissociation constant, K_D, as a measure of binding affinity

$$K_D = \frac{[A][B]}{[C]} = \frac{k_{off}}{k_{on}} \tag{1.40}$$

which relates the association and dissociation rates to the proportion of molecules which are found in complex. The association constant, K_A, is the inverse of the dissociation constant. This allows us to calculate Y, the fractional occupancy of the binding site on A or B:

$$Y_A = \frac{[C]}{[A]_{tot}} = \frac{[C]}{[A] + [C]} = \frac{K_A[B]}{1 + K_A[B]} = \frac{[B]}{[B] + K_D} \tag{1.41}$$

Biologically meaningful interactions have a low K_D, ranging from around 10^{-4}M to below 10^{-15}M. One theoretically possible way of obtaining the kinetics and structure of an interaction would be to simulate the association and dissociation processes by solving Newton's equation of motion numerically. However, molecular dynamics simulations of typical protein-protein complexes can only be realistically achieved on the microsecond timescale, at best. Even if the bound structure was known and the complex dissociated quickly, such as the complex between Ras and RalGDS (Kiel *et al.*, 2004), $k_{off} = 1.49\text{s}^{-1}$, the mean lifetime is on the order of hundreds of milliseconds, well beyond the timescale attainable by molecular dynamics. There has been little work done on calculating dissociation rates, and that which had been done remains to be reproduced (Bai *et al.*, 2011).

There has, however, been a significant amount of work concerning the prediction of association rates. Association events commence with a collision. Interactions for which diffusion is the rate limiting step are described as diffusion controlled. In this case, the proteins can be modelled as spheres and the collision probability can be found analytically considering the diffusion constant, D, of the binding partners and their collisional radius r (Smoluchowski, 1917):

$$k_{on} = 4\pi r(D_A + D_B) \tag{1.42}$$

The diffusion constant can, in turn, be calculated using the Einstein-Stokes equation from the viscosity of the solution, η and the temperature T

$$D = \frac{k_B T}{6\pi\eta r} \tag{1.43}$$

The transition from an encounter complex to the bound state can be simulated. The Smoluchowski equation, 1.42, is used to calculate the probability of the proteins finding themselves within a certain separation of each other. The probability of the complex progressing from here to a final bound state can be determined using either using Brownian dynamics, or transition state theory, where the magnitude of the energy barriers separating the bound state from the unbound state are calculated (Northrup *et al.*, 1984; Rojnuckarin *et al.*, 2000; Lee and Karplus, 1987; Cheng *et al.*, 2007b; Song *et al.*, 2004; Zhou, 1997; Selzer and Schreiber, 1999).

Due to the difficulty of calculating binding affinity via kinetics, especially the difficulty of calculating the dissociation rates, a natural alternative would be to think of the physical binding process in terms of its energetics, which is the subject of the next section.

1.4.4 Interaction Thermodynamics

> "Thermodynamics has the same degree of certainty as its postulates. Reasoning in thermodynamics is often subtle, but it is absolutely solid and conclusive. We shall see how Plank and Einstein built on it with absolute trust and how they considered thermodynamics the absolutely firm foundation in which to build a physical theory. Whenever they were confronted by formidable obstacles they turned to it."
> Emilio Segrè, 1980, 'From X-Rays to Quarks'

Newtons equation of motion, 1.11, shows us that atoms accelerate towards regions of low energy. For rigid bodies in the gas phase, collisions are elastic; the kinetic energy gained from this acceleration is sufficient to escape the potential well and no cohesion occurs. Subsequently, for gas phase reactions of the form $A + B \leftrightharpoons AB$, two-body collisions do not result in association, and a third body is required to take away the excess kinetic energy if the reaction is to proceed. For interactions between biological macromolecules, this excess kinetic energy can be taken away by the solvent, or

be absorbed by either of the binding partners as vibrational motion. Either way, the kinetic energy generated dissipates as disorderly motion. At constant volume and pressure, the change in internal energy upon binding is known as the interaction enthalpy and given the symbol ΔH. If ΔH is negative, then it releases energy as heat, and the process is described as exothermic. The reasoning above would suggest that a good strategy to locate the structure of the complex would be to perform conformational sampling by adjusting the coordinates of the binding partners, calculate the enthalpy of interaction, and repeat until a promising enthalpy minimum is found. Indeed, most binding events are exothermic, but some are endothermic. These actually take heat from the surrounding when they bind to one another, such as the interaction between the ephrin B2 ectodomain and the ephrin B4 receptor, $\Delta H = 3.3$kcal/mol (Chrencik *et al.*, 2006), the interaction between eglin c and chymotrypsin, $\Delta H = 2.0$kcal/mol (Ascenzi *et al.*, 1988), the interaction between Rac GTPase and p67 Phox, $\Delta H = 6.5$kcal/mol (Lapouge *et al.*, 2000) or that between cytochrome C and cytochrome C peroxidase, $\Delta H = 2.3$kcal/mol (Erman *et al.*, 1997). Superficially, this presents a conceptual challenge. In order for the system to spontaneously go from unbound to bound would require an increase in energy, and the atoms would have to move against a repulsive force! Despite this, all these interactions have a dissociation constant, K_D, well below 10^{-5}M. This apparent inconsistency arises from falsely equating the driving force of complex formation to the system's tendency to accelerate towards a low enthalpy configuration. Indeed, the example of gas phase two-body collisions shows that it is not enthalpy that drives spontaneous processes. Unless the excess kinetic energy is taken away, ultimately to increase the disordered motion of the surrounding, the reaction cannot proceed. The driving force of not just molecular interactions, but all processes, is the increase of disorder in the universe, its entropy.

1.4.4.1 The Second Law: A Classical Perspective

When the laws of thermodynamics were discovered, their link to molecular processes was not known. They were discovered via the study of engines and the consideration of processes involving the transfer of work, heat and matter. Comprehension of the second law of thermodynamic was born out

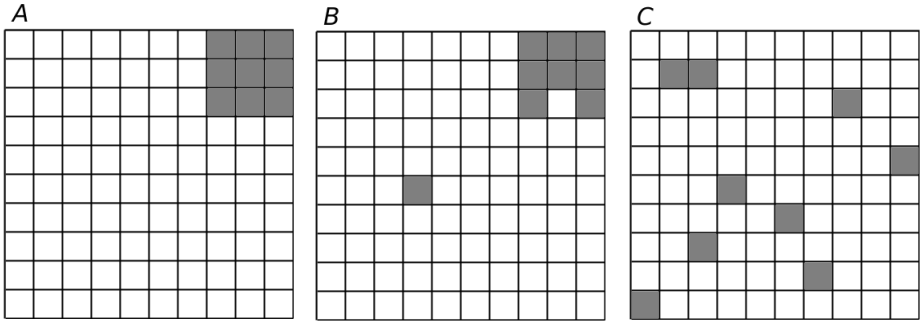

Figure 1.6: Hypothetical distributions of 'gas' particles in a 10 by 10 2D 'box'. (A) Lowest entropy, (B) Low entropy, (C) High entropy.

of efforts to optimise the efficiency of dirty coal-fired machines. Despite these grim origins, it is the second law which drives all change, from the unfurling of a leaf in spring, to the growing of a child in the womb, to the passage of thought through the mind.

One of the earliest definitions of entropy, proposed by Rudolph Clausius, was the quantity of heat supplied to one system from another, divided by the temperature. This is the thermodynamic definition of entropy, S:

$$dS = \frac{dq}{T} \tag{1.44}$$

where q is the quantity of energy transferred as heat. Clausius recognised that for any cyclic process, the process will be spontaneous if and only if the total change in entropy is positive.

1.4.4.2 The Second Law: A Statistical Perspective

A deeper understanding of entropy can be first understood in terms of a tautology; at equilibrium, a system is most likely to be found in a state in which it is statistically more likely to be found. As an illustration, imagine the 9 'gas' particles in the 10×10 2D box in Figure 1.6.

How many ways are there to organise 9 particles into 100 positions? The first particle can occupy any of the 100 positions, the second can occupy any of the remaining 99, and so on. Because the particles are indistinguishable, the ordering is irrelevant, and the total number of possible states is a 9-combination of the set of 100 positions, given by the binomial coefficient

$Z = C(100, 9) \approx 1.90 \times 10^{12}$, known as the partition function. All possible configurations of these particles are one of these states. In Figure 1.6A, the particles all occupy a small region of the box, corresponding to a perfectly ordered macrostate. There is only one configuration, or microstate, of these 9 particles corresponding to this macrostate, so we say the multiplicity, Ω, of the most ordered macrostate, is one. Whilst it is possible to swap any one of the particles for any other, as we are treating them as indistinguishable, we only consider only distinct states.

Now, we can imagine the same number of particles, occupying the same 'volume', but in which the system is slightly less ordered, as in Figure 1.6B. In this state, the coordinates of one of the particles is unknown; it can be in any of the 91 positions other than the top right hand corner, and the 'hole' in the top right hand corner can be in any of the 9 positions. Thus, there are $\Omega = 91 \times 9 = 819$ configurations, or microstates, corresponding to this slightly less ordered 'macrostate'. Clearly, all positions being equally likely, this state dominates the perfectly ordered macrostate.We can then consider an even less ordered system, in which two of the particles are disordered. Here there are $91 \times 98/2$ configurations of the two particles, and $9 \times 8/2$ configuration of the 'hole'. Hence, there are $\Omega = 160,524$ states with 2 disordered particles. Again, this state is statistically dominant compared to the more ordered states. The multiplicity of increasingly disordered 'macrostates', where the number of particles in the 3×3 box is specified, is shown in Table 1.2, as calculated using equation 1.45.

$$\Omega_i = \frac{91!}{(91-i)!i!} \times \frac{9!}{(9-i)!i!} = C(91, 91-i) \times C(9, 9-i) \tag{1.45}$$

It can be seen that over 99.7% of the states have four or fewer particles constrained, as calculated by dividing the sum of the multiplicities of these states by the partition function, Z. In this model system, the most disordered states are the dominant ones, so that as long as there is no *a priori* reason to suspect that certain microstates are more likely than others, disorder is the order of the day. Order could be defined as a different configurations, such as positioned like a chequers board, or as a 9×9 block not in the top right, but anywhere on the grid, or in any shape in which all particles are adjacent. The conclusions, however, are the same; the majority of configurations are non-descript. In fact, if we were to use a finer grid than 10×10, the dominance of

Table 1.2: The multiplicity of macrostates corresponding to i and only i particles in a pre-specified 3×3 box within a 10×10 grid.

i	Ω	Ω cumulative
9	1	1
8	819	820
7	147,420	148240
6	10,204,740	10,352,980
5	336,756,420	347,109,400
4	5,859,561,708	6,206,671,108
3	55,991,367,432	62,198,038,540
2	291,383,646,840	353,581,685,380
1	764,882,072,955	1,118,463,758,335
0	783,768,050,065	1,902,231,808,400 $= Z$

the disordered configuration increases. It we divide space up so finely that, for all intent and purpose, it is continuous, then the most disordered state is so probable that the probability of finding the system in any other macrostate is vanishingly small. Further, if a system were to begin in an ordered state, and were subject to perturbations, it would spontaneously move towards a more disordered state, and the chance of that system spontaneously reverting back to an ordered state is very small indeed. The multiplicity of the macrostate is a precise measure of this disorder. Expressed in a more convenient logarithmic scale, the statistical definition of entropy, S, is given as

$$S = k_B \ln \Omega \tag{1.46}$$

The second law of thermodynamics can then be stated as 'A spontaneous process at fixed volume with a fixed number of particles has a positive change in entropy'.

For real systems, the microstate of a system is not given by its position in configurational space, but its position in phase space; each particle possesses not just a position, but also a velocity and its corresponding kinetic energy. Similar arguments can be used for the distribution of energy as were used for spatial distribution, only the constraint now isn't that the particles are located within the grid, but that the total energy is constant. To illustrate

this, imagine that 10 units of energy are distributed between our 9 particles. In this case, because all the particles have different locations, they can be distinguished from one another. It could be that one single particle has all 10 units of energy, and the remaining 8 particles have none. As the particle with all the energy could be any particle, the multiplicity of this energy distribution is 9. Another possibility is that one particle has 9 units of energy, and another has 1. The multiplicity of this energy distribution is $9 \times 8 = 72$. Similar calculations can be made as above, and similar conclusions reached. The partition function can be calculated, and without *a priori* expectation about the distribution of energy, the probability of finding a system with a certain distribution of energy is calculated as the multiplicity of the distribution divided by the partition function. Just as localised matter dissipates to disorder, localised energy also dissipates to disorder. Just as when the spacing of grid points is reduced to almost continuous, the most disordered spatial distribution dominates all others, when the spacing of energy levels is reduced to a pseudo-continuum, the most disordered distribution of energy dominates. This disordered distribution of energies is the Boltzmann distribution, given by equation 1.47.

$$\frac{n_i}{N} = \frac{e^{-\epsilon_i \beta}}{\sum_i e^{-\epsilon_i \beta}} = \frac{\Omega_i}{Z} \tag{1.47}$$

where n_i/N is the proportion of particles with energy ϵ_i, and $\beta = \frac{1}{k_B T}$. The denominator is the partition function, and the numerator the multiplicity. A derivation of this equation is given in Atkins and De Paula (2006), along with a demonstration of the equivalence of the classical thermodynamic definition of entropy, equation 1.44, and the statistical definition of entropy, equation 1.46. The Boltzmann distribution can be thought of as either a probability distribution when considering a single particle, or an actual energy distribution considering a quasi-infinite number of particles. It is impossible to enumerate all states in most cases, and in all cases for biological macromolecules. Thus the partition function cannot be calculated. However, the probability of finding a particle in one state relative to another can be found by the ratio of their multiplicities.

1.4.4.3 The Gibbs Free Energy

It has been shown that an increase in entropy is the driving force for spontaneous processes, and its molecular basis demonstrated. The inability of the reaction, $A + B \leftrightharpoons AB$ to occur spontaneously in a two-body collision can easily be explained; there are many more ways of distributing A and B in which they are not together, than there are configuration in which they are. The spontaneous nature of this reaction, or indeed a biomolecular interaction, when the excess kinetic energy is removed to the surroundings, can also be explained using equation 1.44. When energy is transferred to a system, in this case the surrounding, more states become available and the entropy of that system increases. Therefore, as long as the increase in entropy of the surroundings is greater than the decrease in entropy associated with the reaction, then the reaction will proceed spontaneously. The chemist Josiah Willard Gibbs derived a quantity, with unit of energy, which for any given process weighs the entropy change of the surroundings with the entropy change of the system in question, and can be used to determine whether a process is spontaneous. This quantity is the Gibbs free energy:

$$\Delta G = -T\Delta S_{tot} = -T\Delta S_{surr} - T\Delta S \tag{1.48}$$

As the entropy in spontaneous processes increases, and the temperature is positive, spontaneous processes are those for which the Gibbs free energy decreases. Using equation 1.44, we see

$$\Delta G = \frac{T\Delta H}{T} - T\Delta S = \Delta H - T\Delta S \tag{1.49}$$

Thus, the spontaneity of the process can be understood in terms of the change in enthalpy and entropy. For a biomolecular interaction, the enthalpy and entropy of interactions can be calculated as the enthalpy and entropy of the binding partners when bound minus their enthalpy and entropy when free in solution, and these two quantities can be used to calculate the free energy of binding.

$$\Delta H_{int} = H_{AB} - H_A - H_B \tag{1.50}$$

$$\Delta S_{int} = S_{AB} - S_A - S_B \tag{1.51}$$

Given the coordinates of a complex, the enthalpy of interaction can be calculated by using a molecular mechanics force field (see section 1.4.1.3). If solvent molecules are not modelled explicitly, an implicit solvation model which treats the solvent as a continuous medium can be employed, such as by solving the Poisson-Boltzmann equation (Fogolari *et al.*, 2002), or its approximation Generalised Born (Chen *et al.*, 2006), or using other methods (Schaefer and Karplus, 1996; Ferrara *et al.*, 2002; Zhang *et al.*, 1997; Lazaridis and Karplus, 1999b; Wesson and Eisenberg, 1992). The entropy can be trickier to calculate. If the potential energy surface is approximated as harmonic, the states of the system and their energy are given by the normal modes and their frequencies (see section 1.4.2.3). Within this approximation, the vibrational entropy of the system is given by equation 1.52 (Mcquarrie, 2000):

$$-TS_{vib} = \sum_{i=7}^{N} \left(\frac{k_B T \ln(1 - \exp(-\hbar\lambda_i / k_B T)) - \hbar\lambda_i}{\exp(\hbar\lambda_i / k_B T) - 1} \right) \tag{1.52}$$

where T is the temperature, k_B is the Boltzmann constant, $\hbar$ is Planck's constant and λ_i is the frequency of the i^{th} normal mode, the square root of the corresponding Hessian eigenvalue. As the six lowest frequency modes correspond to trivial translational and rotational motions, the summation is across the non-trivial modes. Outside of this approximation, calculation of conformational entropy requires extensive conformational sampling, such as Monte Carlo sampling, to determine the accessible states and their energies.

Rotational and translational entropy can be calculated as

$$\frac{S_{rot}}{R} = \ln\left[\frac{\sqrt{\pi I_a I_b I_c}}{\sigma} \left(\frac{8\pi^2 k_B T}{\hbar^2} \right)^{\frac{3}{2}} \right] + \frac{3}{2} \tag{1.53}$$

$$\frac{S_{trans}}{R} = \ln\left[\left(\frac{2\pi m k_B T}{\hbar^2} \right)^{\frac{3}{2}} \frac{k_B T}{P} \right] + \frac{5}{2} \tag{1.54}$$

where m is the mass of the protein or complex, R is the gas constant, I_a, I_b and I_c are the moments of inertia around the x, y and z axes respectively, P is the pressure and σ is the order of rotational symmetry. Other schemes

exist for approximating other forms of entropy, such as the worm-like chain model for disordered loops, or the Gaussian polymer model for disordered chains (Zhou, 2004).

Another important entropic contribution to the binding free energy is due to the hydrophobic effect. This occurs when non-polar atoms are transferred to or from a solvent exposed environment. Water molecules are polar, and form hydrogen bonds with one another. This network of hydrogen bonds constantly dynamically adjusts itself. When a solute is incapable of forming polar interactions with the water, it disrupts this network of hydrogen bonds and the water forms a 'cage' surrounding the hydrophobic region. As this 'cage' is ordered, solvent exposure of hydrophobic atoms causes a decrease in entropy, and burial of hydrophobic surface area causes an increase in entropy. Calculating the magnitude of the hydrophobic effect is difficult. However, thermodynamic measurement of the free energy of transfer of non-polar solutes to the pure liquid phase have shown that the magnitude of the hydrophobic effect is approximately proportional to the exposed surface area. Thus, an estimate of the hydrophobic effect can be made by calculating the solvent exposed hydrophobic surface area (Chothia, 1974).

1.4.4.4 Statistical Potentials

The previous section outlined different terms which can be used to calculate the binding free energy by considering the underlying physical processes. However, another way to calculate free energy changes is by using knowledge based potentials. In this methodology, many observations are used to parameterise a free energy function, by comparing the statistical distribution of the relative positions of atoms or residues to some reference state (Levitt, 1976; Miyazawa and Jernigan, 1985; Sippl, 1990).

To illustrate take, for instance, the distance between the C_α atoms of a tyrosine residue and an isoleucine residue. If we assume that the potential energy of these two atoms can be calculated as a function of their separation then this potential energy function would give rise to a statistical distribution of separations given by the Boltzmann distribution, equation 1.47, where

each separation r, has energy $F(r)$:

$$\frac{n(r)}{N} = \frac{e^{-F(r)\beta}}{Z} \tag{1.55}$$

Z is the partition function. It is possible to rearrange this equation to form the inverse Boltzmann equation, which can be used to derive an energy function from an observed radial distribution:

$$F(r) = -k_B T \ln \frac{n_i}{N} - k_B T \ln Z \tag{1.56}$$

As the rightmost term is constant, it can be removed. Thus, we can take a large number of high resolution crystal structures and derive a potential energy function. However, there are a number of issues associated with this approach. Many effects are averaged out, such as different protonation states and relative orientations of the amino acids. It is possible to modify the form of the potential function, such as to include the orientation of amino acids, or to predict the protonation state of amino acids which have a pKa near physiological pH and treat the same amino acid in different states as different entities. Some effects, however, cannot be modelled using pair potentials, such as changes associated with atomic restriction; rotational, vibrational and translational entropy or entropy changes associated with disorder to order transition. These effect are always averaged out or neglected when deriving the statistical distribution. Further, there may be other factors influencing distribution other than those of an energetic nature. A trivial example is the fact that the volume of space between distance $r - dr/2$ and $r + dr/2$ is proportional to the surface area of a sphere of radius r and hence even if atom or residue pairs were evenly distributed, the radial distribution would be quadratic. Also, some amino acids occur with greater frequency than others, and biases can also be caused by the different shapes of the proteins used in parameterisation. To overcome these problems, the potential energy function is usually expressed relative to a reference state.

$$F(r) = -k_B T \ln \frac{n_i}{n_{Ri}} \tag{1.57}$$

For instance, the reference state may be the distribution expected if the atoms were an ideal gas distributed evenly in a sphere of the same

density as a typical protein, with the partial pressures of the atoms being proportional to the abundance of their corresponding amino acids. This way, the free energy of a system can be estimated from its structure, by doing a pairwise summation of all atom or residue pairs.

There are many different statistical potential functions available, often using different reference states. Most are pair potentials, but multi-body potentials also exist, as do orientation-dependent potentials. Some are course-grained residue-level potentials between C_α atoms, C_β atoms or residue centroids, whilst others are atomistic. Some are trained on intermolecular pairs, others on intramolecular pairs.

1.4.4.5 The Free Energy Landscape

It is now possible to think of the energetics of protein-protein binding not in terms of the potential energy (enthalpy) landscape, but in terms of the free energy landscape. When proteins bind to one another and shift from the unbound state to the bound state, their free energy lowers. It is known that the energy landscape surrounding the bound structure forms a funnel-like structure (Tsai *et al.*, 1999; Ma *et al.*, 1999; Wang and Verkhivker, 2003; Zhang *et al.*, 1999; Hunjan *et al.*, 2008). Indeed, this must be the case, as the combination of side chain angles, backbone conformations and relative positions and orientations is so large that without a funnel-like structure to guide two proteins towards the native state, the timescale on which protein-protein association events occur would be much larger. Quite some work has been done to characterise the ensemble of trajectories which are taken when proteins bind to one another.

One possibility is that binding proceeds without significant conformational change; the lock and key model of binding. In this model, the ensemble of accessible structures adopted by the binding partners is not significantly altered. This model of binding, whilst approximately true for some cases, cannot explain all interactions, as structural data reveals some conformational changes, however small, in most cases. In some cases, the change is dramatic, and disorder to order transitions are seen; folding and binding occur simultaneously (Sugase *et al.*, 2007). The question now

focuses on the degree to which the conformational ensemble of the binding partners in the bound state overlap with the unbound state. At one end of the spectrum lies the 'induced fit' hypothesis, in which the conformations accessible in the bound state are not accessible in the unbound state, and are only stable in the presence of the other binding partner. At the other end of the spectrum in the 'conformational sorting and population shift' model, in which the two ensembles share many conformations and the proportion of proteins in the bound conformation is increased by the stabilising effect of the binding partner. Evidence suggests that both effects occur, however the conformational sorting mechanism is favoured when the energy barriers separating the conformational states are large, when the interaction is weak and when the interaction is dominated by short-range forces (Okazaki and Takada, 2008; Zhou, 2010). For protein-protein interactions, the evidence seems to suggest that the conformational sorting mechanism is preponderant (Boehr *et al.*, 2009; Okazaki and Takada, 2008; Goh *et al.*, 2004; Marsh and Teichmann, 2011; Boehr *et al.*, 2009; Gsponer *et al.*, 2008; Lange *et al.*, 2008; Marsh *et al.*, 2010; Tobi and Bahar, 2005; Stein *et al.*, 2011b).

1.4.4.6 Thermodynamic Cycles

The entropy, enthalpy, Gibbs free energy and potential energy are state functions. This means that their value depends purely on the state of the system, its pressure, temperature, volume and composition. This property is very useful for the calculation of free energies, as it allows the decomposition of binding into separate processes, via the construction of a thermodynamic cycle. For instance, the binding free energy change associated with proteins going from the unbound to the bound state in solution can be calculated as the free energy of solvation of the bound proteins, plus the free energy of binding *in vacuo*, minus the free energy of solvation of the products. Not only can thermodynamic cycles correspond to real processes, but also alchemical processes.

1.4.4.7 From Free Energy to Binding Affinity

The formation of a pure product from pure reactants, under standard conditions, is known as the standard Gibbs reaction energy. This forces a distinction between actual, hypothetical and model systems. The standard Gibbs binding energy, $\Delta G^{\ominus}$, is the change of free energy when every protein A and B, one mole of each, form one mole of the complex AB. This free energy is the Boltzmann weighted average of all bound microstates. When a structure is used to calculate a single point free energy, this corresponds to the energy of one mole of AB at the corresponding set of coordinates, minus the energy of one mole each of A and B all at the energy of the unbound coordinates. In real biological systems, however, not all molecules of A and B are part of a complex, and at equilibrium the relative proportion in the bound state is determined by the dissociation constant, K_D, given by equation 1.40. To establish a link between binding free energy and the binding affinity, we need to consider the relationship between the standard Gibbs reaction energy and the Gibbs energy of mixtures. The free energy change associate with adding or removing a particle from a system is known as the chemical potential, μ. The process of binding is equivalent to a thermodynamic cycle consisting of the removal of one particle each of A and B, and the addition of one particle of AB. The change in free energy for this cycle can be expressed as

$$\Delta G = \mu_{AB} - \mu_A - \mu_B \tag{1.58}$$

At equilibrium, the above equals zero, as there is no driving force to alter the composition. The chemical potential of a substance can be determined by equation 1.59, which is derived from first principles in Atkins and De Paula (2006)

$$\mu_X = \mu_X^{\ominus} + RT \ln a_X \tag{1.59}$$

where a_X is the activity of molecular species X, and $\mu_X^{\ominus}$ is the chemical potential of pure X at standard conditions. The activity can be expressed in terms of $[X]$, the concentration of X

$$a_X = \frac{f_X[X]}{c^{\ominus}} \tag{1.60}$$

where the standard concentration, $c^{\ominus}$, is 1M. For solutions of low concentration the activity coefficient, f_X, is close to 1, and so to a first approximation the activity is equal to $[X]$, and this can be used in equation 1.59. Therefore, combining 1.59 and 1.58 and using the definition of the dissociation constant, at equilibrium

$$\begin{aligned} 0 &= \mu_{AB} - \mu_A - \mu_B = \mu^{\ominus}_{AB} + RT\ln[AB] - \mu^{\ominus}_A - RT\ln[A] - \mu^{\ominus}_B - RT\ln[B] \\ &= \mu^{\ominus}_{AB} - \mu^{\ominus}_A - \mu^{\ominus}_B + RT\ln\frac{[AB]}{[A][B]} \\ &= \Delta G^{\ominus} + RT\ln K_D \end{aligned} \tag{1.61}$$

Rearranging and using the identities in equations 1.40 and 1.49, we get

$$\Delta G^{\ominus} = -RT\ln\frac{k_{off}}{k_{on}} = \Delta H^{\ominus} - T\Delta S^{\ominus} \tag{1.62}$$

This deduction is key, as it provides a solid link between the thermodynamics and the kinetics of the binding process. It is difficult to overstate the importance of this relationship, and it is used implicitly throughout this thesis. Biologically significant interactions, the understanding of which could facilitate systems biology studies, are those that have high affinity and whose bound state engender biological function. This state relates to the microscopic configurations of the atoms; the structural ensemble of the complex. Through the above relationship, the structure of the complex can, in principle, be used to calculate the binding affinity. Similarly, by finding the global free energy minimum, the structures of unbound proteins can, in principle, be used to predict the structure of the complex. Further, if a sufficiently wide conformational search can be undertaken, the presence or absence of a deep minimum can, in principle, confirm or repudiate a putative interaction.

1.4.4.8 Binding Affinity Prediction

In section 1.3.1.3, it was shown how knowing the strength of the interactions between the molecular constituents of biological systems is antecedent to understanding cellular logic. There are a number of methods of calculating affinities from structure, such as thermodynamic integration, free-energy perturbation, MM-PBSA and others (Zacharias, 2010b; Gilson and Zhou,

2007). However, these methods are very expensive, requiring extensive conformational sampling. Even considering the advances in GPU accelerated molecular dynamics, scoring whole interactomes is beyond the remit of these techniques. Faster methods can be broadly split into two categories; knowledge-based potentials, and "master" thermodynamic equations, both of which must be empirically parametrised.

Table 1.3: A summary of empirical binding free energy functions published to date. The number of test cases are shown (Cases), with a subdivision in parentheses (protease-inhibitors/other enzyme-inhibitors/antibody-antigen/small peptides/others). The parameters (Par.), number of variable terms (Var.), reported performance (Per.), the method used (Method) and the reference (Reference) are also given. The method is reported as either a potential of mean force (PMF) or a sum of terms (sum), with the name given where applicable.

Cases	Var.	Perf.	Method	Par.	Reference
3 (0/0/3/0/0)	0	10.5[a]	sum	Electrostatics, hydrophobic burial, side-chain entropy, constant	Novotny *et al.* (1989)
15 (13/0/1/0/1)	3	0.96[b]	sum	Hydrophobic burial, polar burial, constant	Horton and Lewis (1992)
9 (7/0/0/0/0)	0	2.4[ac]	sum	Electrostatics, H-bonding, side-chain entropy, constant	Krystek *et al.* (1993)
9 (7/0/0/0/0)	0	1.3[ac]	sum	Electrostatics, hydrophobic burial, side-chain entropy, constant	Vajda *et al.* (1994)
14 (14/0/0/0/0)	2	0.9[a]	sum	Electrostatics, desolvation	Nauchitel *et al.* (1995)
9 (9/0/0/0/0)	4	0.74[b]	sum	Statistical function	Wallqvist *et al.* (1995)
21 (16/0/2/0/3)	4	0.86[b]	sum	Hydrophobic burial, hydrophilic burial, # hydrophilic pairs, constant	Xu *et al.* (1997)
9 (9/0/0/0/0)	0	0.7[b]	PMF, ACE	Coefficient, offset	Zhang *et al.* (1997)
20 (16/0/3/0/1)	0	0.94[bc]	sum	Electrostatics, hydrophobic burial, side-chain entropy, constant	Weng *et al.* (1997)
2 (1/0/1/0/0)	0	0.54[a]	sum	Electrostatics (water and self), VDW (water and self), cavitation, entropy (translational, rotational, vibrational and configurational)	Noskov and Lim (2001)
28 (16/1/7/0/4)	1	0.75[b]	PMF	Coefficient, offset	Jiang *et al.* (2002)
19 (16/1/1/0/1)	3	0.95[b]	sum	Hydrophobic burial, side-chain entropy, # hydrophilic pairs, constant	Ma *et al.* (2002)
69 (19/2/8/27/13)	2	0.87[b]	PMF, DComplex	Coefficient, offset	Liu *et al.* (2004)
52 (14/1/5/27/5)[d]	0	0.79[be] 0.85[b]	sum, Rosetta	Electrostatics, H-bonding, VDW, EEF1 desolvation, pair-potential, water potential	Jiang *et al.* (2005)
82 (21/3/14/27/18)	2	0.73[b]	PMF, DComplex	Coefficient, offset	Zhang *et al.* (2005b)
24 (18/1/2/0/3)	7	0.98[b] 0.95[f] 0.62[a]	sum, AffinityScore	Interface gap volume, # exposed charges, # salt bridges, # hydrogen bonds, # constricted torsions, # exposed hydrophobic groups, constant	Audie and Scarlata (2007)
20 (7/2/5/0/6)	5	0.83[b]	sum	Hydrophobic burial, polar burial, charge-charge interaction, charge burial, side-chain entropy	Bougouffa and Warwicker (2008)
86 (25/2/13/26/20)	0	0.76[b] 2.24[a]	PMF	None	Su *et al.* (2009)
63 (6/15/4/0/37)	8[g]	0.73[bcf]	sum	Trans/rot entropy, # atom pairs, # non-polar residues, # non-polar atom pairs, interface planarity, interface gap volume, gap volume/interface area ratio, constant	Bai *et al.* (2011)

[a] RMSE, kcal mol^{-1}.
[b] Correlation.
[c] Two complexes omitted from evaluation.
[d] Identities from personal communication.
[e] Without water potential.
[f] Leave-one-out cross-validation.
[g] Feature selection.

In the knowledge-based approach, statistical potentials (described in section 1.4.4.4), are used to predict the binding free energy of complexes with known binding affinity. Appropriately formulated, these do not require training on binding free energies, only structures (Zhang *et al.*, 1997; Su *et al.*, 2009). Having no adjustable parameters, these methods carry no risk of over-fitting. Often, however, a correction to this prediction scheme, in the form of a coefficient and an offset, is derived by linear regression of the predicted energies against empirical binding free energies (Liu *et al.*, 2004; Zhang *et al.*, 2005b), or by just adjusting the gradient (Jiang *et al.*, 2002). Such pair potentials cannot tell us about how the free energy is factorised into entropic and enthapic components, nor can it be factorised according to physical origin: electrostatics, desolvation, Van der Waals and so on.

In the "master equation" scheme, a number of physically relevant terms are calculated explicitly and are taken in linear combination. In some cases, the inclusion and weighting of each term is based on physical law, such as Coulomb's law, or on empirically parametrised functions, such as the function relating the change in hydrophobic surface to the free energy of transfer from aqueous to non-aqueous solvent (Chothia, 1974). This information is derived from data other than the affinity test set and thus, again, no fitting is required (Novotny *et al.*, 1989; Krystek *et al.*, 1993; Vajda *et al.*, 1994; Weng *et al.*, 1997). In other approaches, the weights of some or all terms are determined by linear regression against a training set (Horton and Lewis, 1992; Xu *et al.*, 1997; Ma *et al.*, 2002; Bougouffa and Warwicker, 2008; Bai *et al.*, 2011). Of course, the physical origins are known in these methods, as their contributions are calculated explicitly, and their form can also be related to the enthalpy and entropy of the binding for comparison with isothermal titration calorimetry data (Weng *et al.*, 1997). However, often parameters can be colinear (Bougouffa and Warwicker, 2008; Vajda *et al.*, 1994), and thus it is difficult to be certain that the functional form appropriately reflects the underlying physics. For instance, the affinities of almost identical test sets can be equally predicted by equations of differing functional form (Horton and Lewis, 1992; Xu *et al.*, 1997; Weng *et al.*, 1997). A summary of methods published to date is given in Table 1.3.

In most efforts to predict binding free energy, the rigid-body approxim-

ation was invoked. Either the rigid-body assumption was implicit as the test/training sets were derived from previous publication in which structures which undergo conformational changes were excluded (Jiang *et al.*, 2002, 2005; Su *et al.*, 2009), or flexible cases were explicitly excluded from the set of complexes. Prior to Liu *et al.* (2004), only small sets of proteins were used, and these mostly constituted protease/inhibitor interactions, and other high affinity complexes composed of rigid binding partners, such as barnase/barstar, the insulin dimer, the α and β chains of deoxyhaemoglobin and lysozyme/antibody complexes. In these studies, excellent agreement could be made between experiment and theory, with correlations up to 0.96 reported (Horton and Lewis, 1992; Nauchitel *et al.*, 1995; Weng *et al.*, 1997).

Between 2004 and 2009, four papers were published in which free energy functions were applied to more diverse sets of complexes (Liu *et al.*, 2004; Jiang *et al.*, 2005; Zhang *et al.*, 2005b; Su *et al.*, 2009). Most of this added diversity came from the inclusion of small peptides, typically between 2 and 5 residues in length, and mostly involving interactions with oligopeptide-binding protein (Sleigh *et al.*, 1999, 1997; Tame *et al.*, 1995). However, other interactions such as those including G-binding proteins, signal transduction complexes and hormone/receptor pairs started entering the data sets. Correspondingly, there was a decrease in the correlation between the experimental binding free energies and theoretical results. In the work of Bai *et al.* (2011), feature selection and multiple regression was used to construct models to predict affinity, dissociation and association rates for a diverse set of complexes. However, it is not clear how they impliment their feature selection algorithm, and there are a large number of adjustable parameters. Further, the binding energy function seems incongruous, involving terms such as the number of contacted atom pairs per 100Å^2 interface area, the volume of space between the interface per 100Å^2 interface area, the planarity of the interface and the number of non-polar residues at the interface.

A vastly more diverse set of interactions for which structural and affinity data was available was collected and tested by Kastritis and Bonvin (2010). In this set, no class of interaction is over-represented, as the list of complexes from which it was derived contains no pairs with high sequence identity, except for antibodies (Chen *et al.*, 2003). Post *corregendum*, their test set contained 46 complexes and 12 binding free energy functions were applied.

The greatest correlation with these data is 0.53, highlighting the degree to which previous studies were biased towards rigid proteins and particular classes of interaction.

1.5 Protein-Protein Docking

In sections 1.4.4.3 and 1.4.4.4, I showed two approaches for calculating the binding free energy of a protein-protein interaction from the coordinates of a complex, and how this energy relates to the probability of finding two proteins with that particular set of coordinates relative to a different set of coordinates. This immediately suggests an intractable but sure method of computationally determining the accessible structures a complex can adopt; enumerate all states and their energies by performing an exhaustive search of all conformations of both binding partners at all relative positions and orientations, calculating the binding free energy at each point. However, this approach is impossible in practice, as the combinatorial explosion associated with having more than a handful of atoms renders the number of possible conformations of the two proteins astronomical; hundreds of orders of magnitude greater than the number of particles in the observable universe! Here, I discuss the methods which can be employed to reduce the problem from being equivalent to calculating the partition function, down to something which is computationally feasible given current technology, and discuss the current methods and issues in the protein-protein docking field.

The free energy landscape of most proteins is such that their internal structure is ordered. Indeed, in archaea and bacteria, it is believed that only around 5% of proteins are mostly disordered, although in eukaryotes this figure seems around five times higher (Oldfield *et al.*, 2005). Thus, for the majority of the time, most proteins in isolation adopt one of a relatively narrow ensemble of conformations.

Analyses of the conformational changes associated with binding events show that the topological features of the fold are conserved. Exceptions do exist, such as the complex between thrombin and the heparin cofactor (PDBid 1JMO), where the central strand of an antiparallel β-sheet distal

to the binding site unravels, forming a parallel β-sheet and significant structural rearrangement. However, most interactions for which structural data is available demonstrate that the fold remains intact. Hence, a good starting place for docking is to use an available unbound structure. If this is not available, homology modelling can be employed to predict the structure of the protein if the structure of homologues are known, although docking homology models is inherently less reliable. If no homologues are known, *ab initio* prediction of protein structure can be employed, but these techniques are not currently viewed as capable of producing models of sufficient quality for docking.

The earliest attempt to predict the structure of a complex from its unbound constituents was an attempt to find the stacking of sickle cell haemoglobin molecules into the fibres which exemplify the disease. In this procedure, various relative positions and orientation of the haemoglobin were chosen based on their consistency with electron microscopy data, and side chain angles were optimised to eliminate clashes and optimise electrostatics (Levinthal *et al.*, 1975). A similar approach of optimising electrostatics was used to model the complex between cytochromes b5 and c (Salemme, 1976). The first automated procedure, and the first attempt at modelling which resembles modern docking, was the systematic rigid-body docking of trypsin to BPTI, both small proteins (Wodak and Janin, 1978). In this approach, the relative orientations and positions of the two binding partners were systematically explored and the energy was evaluated in terms of a non-bonded potential and a solvent exclusion term. A coarse-grained model was used, with one interacting center per residue, and the most promising solutions were refined. In broad outline, the modern approach to the docking problem has not changed much since these early studies, and docking is still viewed as an optimisation problem or a systematic rigid-body search followed by refinement. However, the conformational sampling and scoring routines have and are still progressing, and no particular approach has yet been shown to be consistenly superior (Mendez *et al.*, 2005, 2003; Lensink *et al.*, 2007; Lensink and Wodak, 2010b,a).

1.5.1 A General Overview

There is no standard procedure for modern docking algorithms, however many approaches share commonalities. Often a series of steps are performed; an initial search, followed by refinement and then post-processing. The initial step may be a rigid-body search which generates many structures, which can be filtered with an efficient scoring function, or clustered, or both. Then, a reduced list of potential docked poses can undergo flexible refinement, to optimise the positions of sidechains and loops. Finally, the refined structures can be re-ranked using a more expensive scoring function. Not all algorithms undergo all steps and their details vary. Often the algorithms run at the various stages can be substituted for one another. For instance, the initial search can be performed by ZDOCK, PIPER, FTDock or PatchDock (Chen and Weng, 2002; Kozakov *et al.*, 2006; Gabb *et al.*, 1997; Schneidman-Duhovny *et al.*, 2005b), refinement could be achieved with RDOCK, MultiDock or FireDock (Li *et al.*, 2003; Jackson *et al.*, 1998; Andrusier *et al.*, 2007), and the final scoring/clustering could be completed using ClusPro, ZRANK or RPScore (Comeau *et al.*, 2004; Pierce and Weng, 2007; Moont *et al.*, 1999).

1.5.2 Rigid-body Docking

The simplest representation of a protein is as a rigid body. Analysis of complexes with known bound and unbound structure shows that in many cases backbone rearrangements can be significant (Betts and Sternberg, 1999) and almost 40% of interface side-chains swap rotamer (Guharoy *et al.*, 2010; Ruvinsky *et al.*, 2011). However, a rigid-body representation can often still be used to find a reasonable docked pose and the lock and key model suffices as a first approximation (Betts and Sternberg, 1999). Under the rigid-body regime, there are six degrees of freedom, three translational and three rotational, resulting in a manageable search space. This representation of protein structure is frequently used as part of conformational searching and refinement. For a systematic search in rigid-body search space with rotational steps of 12°, in a cube of 120Å, with translational grid spacing of 1Å, around 10^{10} conformations need to be generated (Zacharias, 2010b). This

is a large number of structures, however approximations can be used to significantly accelerate the searching and scoring, and the most common rigid-body approach uses the correlation technique, discussed in section 1.5.3. Alternatively, surface patches can be matched up to generate potential docked poses, as discussed in section 1.5.4. These methods are based on the observation that protein-proteins interfaces are often composed of complementary surfaces, with protrusions matching indentations (Janin, 1995; Chothia *et al.*, 1985; Rebek, 1987; Norel *et al.*, 1999). Rigid-body searching is often a starting point for guided searches, in which the funnel-like energy landscape is exploited, as discussed in section 1.5.5.

1.5.3 The Correlation Method

The correlation method of rigid body docking, developed by Katchalski-Katzir *et al.* (1992), vastly accelerates the translational search by using the fast Fourier transformation algorithm to calculate a correlation between two discrete functions. These three dimensional functions, a and b, partition the binding partners into interior, exterior and surface regions, discretised at a resolution of $0.7 - 0.8$Å:

$$a_{l,m,n} = \begin{cases} 1 & \text{on the surface} \\ \rho & \text{inside} \\ 0 & \text{outside} \end{cases} \tag{1.63}$$

$$b_{l,m,n} = \begin{cases} 1 & \text{on the surface} \\ \delta & \text{inside} \\ 0 & \text{outside} \end{cases} \tag{1.64}$$

The interior parameters, ρ and δ, are set such that their product is large and negative. The correlation function, c is given by

$$c_{a,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} b_{l+\alpha,m+\beta,n+\gamma} \tag{1.65}$$

The indices of the correlation function, α, β and γ, index the relative position of the binding partners. Its value is positive if the surfaces match and clashes are penalised by the large negative product of ρ and δ. The

Fourier transform $X_{o,p,q}$, of a function $x_{l,m,n}$ is given by

$$X_{o,p,q} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} e^{-2\pi i \frac{ol+pm+qn}{N}} x_{l,m,n} \tag{1.66}$$

The Fourier transform of the correlation function can be expressed in terms of $B_{o,p,q}$ and $A^*_{o,p,q}$, the Fourier transform of b and the complex conjugate of the Fourier transform of a respectively

$$C_{o,p,q} = A^*_{o,p,q} B_{o,p,q} \tag{1.67}$$

Hence it possible to use equation 1.66 to obtain the $A_{o,p,q}$ and $B_{o,p,q}$, and use these to to determine $C_{o,p,q}$, the inverse Fourier transformation of which yields $c_{a,\beta,\gamma}$. Hence, a scored list of rigid body positions is obtained. This procedure is repeated at different orientiations, to systematically search both translational and orientational space.

Later, a modified version of the algorithm which included electrostatics was implemented as the program FTDock (Gabb *et al.*, 1997). In this scheme the receptor is represented such that the discrete functions return the electrostatic potential at the exterior of the molecule, and the ligand is represented as its charge, q

$$a_{l,m,n} = \begin{cases} \sum_j \frac{q_j}{\epsilon(r_j) r_j} & \text{outside} \\ 0 & \text{inside} \end{cases} \tag{1.68}$$

$$b_{l,m,n} = q(l,m,n) \tag{1.69}$$

Since then, a number of modifications to this scheme have been proposed, such as the DOT algorithm (Mandell *et al.*, 2001), which includes solvent continuum electrostatics as well as Van der Waals forces, using a more a general form of correlation functions, composite convolution functions. In the ZDOCK algorithm (Chen and Weng, 2002), electrostatics and surface complementarity are modelled along with the atomic contact energy desolvation potential developed by Zhang *et al.* (1997). In another correlation based approach, that of Heifetz *et al.* (2002), the discrete functions are complex functions with the real part containing the surface

complementarity and the electrostatics contained in the imaginary part. In a later modification, a hydrophobic term was added (Berchanski *et al.*, 2004), resulting in improved performance. Alternatively, statistical pair potentials can be encoded, such as is implemented in the docking program PIPER (Kozakov *et al.*, 2006). Here, the reference state for the pair potentials were docking decoys; incorrect structures that were generated by docking.

Another correlation based method of note is that implemented in the program HEX (Ritchie and Kemp, 2000). In this approach, the coordinate system consists of 1 intermolecular distance and 5 angles. The shape and electrostatics are expressed as a linear combination of the product of spherical harmonics and Laguerre polynomials. In doing so, the correlation function can be simultaneously evaluated for the 5 angular degrees of freedom, as opposed to just the three translational degrees of freedom in the earlier Fourier transformation based methods, resulting in considerable speed up (Ritchie *et al.*, 2008). Ported to GPU technology, this algorithm can complete the entire 6D systematic search in as little as 15s, two orders of magnitude faster than the next quickest systematic search methods (Ritchie and Venkatraman, 2010), and has hence been a method of choice for high-throughput docking (Wass *et al.*, 2011).

One disadvantage of the correlation approach is that whilst it works very well for proteins which undergo little conformational change upon binding, it can produce many false positives and often other information, or refinement, is required to determine which generated structures are near-native.

1.5.4 Surface Matching

An alternative to systematic search is to only generate those structures which have surface complementarity. Concave and convex features of the protein surface can be characterised by their size and shape. These 'critical points' on the solvent-accessible surface were first discussed by Connolly, and later adaptations have served as a basis for the geometric hashing technique, in which these surface descriptors are stored in a hash

table (Wang and Levinthal, 1991; Connolly, 1983, 1986), or for sub-matrix matching (Helmer-Citterich and Tramontano, 1994). The first successful attempt at matching 'holes' to 'knobs' using geometric hashing was performed by Norel *et al.* (1994). Following this success, the groups of Nussinov and Wolfson later refined the algorithm and released it as the program PatchDock, one of the most efficient surface matching algorithm available to date (Fischer *et al.*, 1995; Norel *et al.*, 1999; Schneidman-Duhovny *et al.*, 2003, 2005b). The surface matching approach has been extended to other molecular shape representations, such as descriptors based on a linear combination of Zernike polynomials (Venkatraman *et al.*, 2009), or facets generated using the marching cubes algorithm (Bordner and Gorin, 2007; Lesk and Sternberg, 2008). Surface matching can be modified to include flexibility, such as in the 3D-Garden method (Lesk and Sternberg, 2008) and FlexDock, a modified version of the PatchDock algorithm (Schneidman-Duhovny *et al.*, 2005a). A quite different surface matching approach is the SKE-DOCK method (Komatsu *et al.*, 2003; Terashi *et al.*, 2007). Benzene clusters are simulated around the surface of the binding partners by molecular dynamics. Hydrophobic benzene attractor points serve as the equivalent of 'holes' and 'knobs', which are matched up and further refined. Surface matching and guided search docking methods are not mutually exclusive, as some algorithms can fit easily into either category.

1.5.5 Guided Search

Guided docking methods are ones which employ optimisation or simulation techniques to minimise the interaction energy, such as molecular dynamics, Monte Carlo sampling, Newton-Raphson, simplex or steepest descent. This approach takes advantage of the energy funnel surrounding the bound ensemble of structures. Methods of this type include ATTRACT (Zacharias, 2003), which uses a quasi-Newton minimiser, 3D-Dock, which uses steepest descent (Jackson *et al.*, 1998), ICM-disco, which uses a pseudo-Brownian Monte Carlo routine (Fernandez-Recio *et al.*, 2004; Fernández-Recio *et al.*, 2003), HADDOCK, which uses simulated annealing and steepest descent, (Dominguez *et al.*, 2003; de Vries *et al.*, 2007), methods based on the genetic algorithm (Gardiner *et al.*, 2001), MC2, which uses Monte

Carlo sampling (Bastard *et al.*, 2003), SMOOTHDOCK, which uses molecular dynamics, simplex and adopted basis Newton-Raphson (Camacho and Vajda, 2001; Camacho and Gatchell, 2003), RDOCK, which also uses adopted basis Newton-Raphson (Li *et al.*, 2003), RosettaDock, which uses Monte Carlo with quasi-Newton minimisation (Wang *et al.*, 2005; Gray *et al.*, 2003), FireDock, which uses linear programming, Monte Carlo and quasi-Newton minimisation (Andrusier *et al.*, 2007), ReplicOpter, which uses an adapted Hooke-Jeeves optimiser (Demerdash *et al.*, 2010) and FibreDock, which uses Monte Carlo and quasi-Newton minimisation (Mashiach *et al.*, 2010).

1.5.6 Accounting for Flexibility

When proteins bind to one another, a number of possible structural changes can occur. These can be global motion, such as hinge bending, in which domains connected by a flexible region move rigidly relative to one another, or shear motion, in which the interdigitated sidechains of two packed structural elements move parallel along with those elements and repack themselves. Localised rearrangements are more frequent; flexible loops can change their conformation, and sidechains can switch rotamer, or move to a non-rotameric conformation. Accounting for the conformational changes which occur as proteins bind to one another remains a problem in the field of protein-protein docking, and has been subject to a number of recent reviews (May and Zacharias, 2005; Bonvin, 2006; Gray, 2006; Andrusier *et al.*, 2008; Moreira *et al.*, 2010; Zacharias, 2010a; Bastard *et al.*, 2011).

The simplest common method to account for the coformational changes is to use 'soft' potentials to allow clashes and some interpenetration of surfaces (Palma *et al.*, 2000; Fernández-Recio *et al.*, 2003; Jiang and Kim, 1991; Zacharias, 2003; Katchalski-Katzir *et al.*, 1992; Gabb *et al.*, 1997; Mandell *et al.*, 2001; Chen and Weng, 2002; Eisenstein and Katchalski-Katzir, 2004; Gardiner *et al.*, 2001; Gray *et al.*, 2003; Schneidman-Duhovny *et al.*, 2004). Indeed, in the first pioneering early studies, either a reduced model (Wodak and Janin, 1978), or a softened Van der Waals term in which the repulsive $1/r^{12}$ term is replaced by a $1/r^{8}$ term, was used (Levinthal *et al.*, 1975). Soft potentials also reduce the ruggedness, and thus the multimodality, of the energy landscape, which aids optimisation as it reduces

the number of local minima in which the algorithm can become trapped. Soft potentials and coarse-graining are simple methods of soft-docking, however three other approaches stand out as particularly noteworthy. One of these is employed in the GRAMM-X approach (Tovchigrechko and Vakser, 2006). Here, the degree of coarse-graining and smoothing of the potential can be varied, and thus a finer grain can be used when docking high-resolution structures or those deemed to be rigid, whilst a coarser grain can be used for lower resolution structures, homology models or flexible proteins, or for studying low-resolution recognition factors. Another interesting form of coarse-graining is the potential used in the SMOOTHDOCK algorithm (Camacho and Vajda, 2001; Camacho and Gatchell, 2003), where initially electrostatics and desolvation dominates the energy function, and the weight of the Van der Waals interaction is slowly increased as the algorithm proceeds. Thus, the ruggedness of the energy landscape increases as the search is focussed. Finally, the semi-definite programming-based underestimation approach developed by the Vajda lab uses the convex global underestimation method which was originally developed for protein folding (Paschalidis *et al.*, 2007; Shen *et al.*, 2008). In this model, the local energy minima are fitted to a quadratic function, and further sampling is biased towards the minima of this function.

The most basic method to refine structures which are generated with soft potentials is energy minimisation in all degrees of freedom. For instance, following a systematic rigid-body search, Li *et al.* (2003) tooks a series of three minimisations, first with just Van der Waals, then with Van der Waals and uncharged polar groups and finally with Van der Waals and full electrostatics. This method, however, can only deal with clashes and very minor changes.

A more advanced method of including flexibility is ensemble docking. In this approach, ensembles of structures are docked together. The ensembles can be generated in a number of ways, and there are a number of different ensemble-based approaches. In the cross-docking method, molecular dynamics is performed on the ligand and receptor, the trajectories are clustered and the clusters rigidly docked pairwise to one another (Smith

et al., 2005; Grunberg *et al.*, 2004; Krol *et al.*, 2007a). A similar approach is that of Mustard and Ritchie (2005), in which the program CONCOORD (de Groot *et al.*, 1997) was used to generate an ensemble of structures using pseudo-NMR restraints. Principle component analysis was employed and used to generate 'eigenstructures', which were subsequently cross-docked. The other main ensemble based docking method is the mean field approach, in which the whole ensemble is docked simultaneously (Koehl and Delarue, 1994). Take, for instance, two interacting side-chains. Each member of the ensemble of both side chains are weighted equally. Each member of the first side chain ensemble feels the weighted average energy of the second side-chain, and their weights are adjusted such that they follow the Boltzmann distribution. Then the weights of the second side chain are then adjusted so as to follow a Boltzmann distribution in the mean field created by the first side chain. Then, the weights of the first side chain are adjusted again, and the process is repeated iteratively until self-consistency is achieved. This model can be extended without loss of generality. This approach has been used to model side chains in a number of programs, including MultiDock and ATTRACT (Jackson *et al.*, 1998; Zacharias, 2003; Koehl and Delarue, 1994; Mendes *et al.*, 1999), as well as to model loops in RosettaDock, MC2 and ATTRACT amongst others (Bastard *et al.*, 2003; Loriot *et al.*, 2011; Chaudhury and Gray, 2008; Bastard *et al.*, 2006). Of course, the ability to accurately model loops and side chains is predicated upon the native conformation residing within the inital ensemble. This inital ensemble can be derived from loop or rotamer databases (Oliva *et al.*, 1997; Michalsky *et al.*, 2003; Wang and Dunbrack, 2003), from molecular dynamics or Monte Carlo simulations, CONCOORD, NMR ensembles, normal mode analysis or from known homologues (Demerdash *et al.*, 2010). Mean field modelling is commonly applied during a docking procedure, however sometimes loops can be ignored during the docking itself and rebuilt in the post-processing stage (Wang *et al.*, 2007a; Soto *et al.*, 2008).

Others have developed methods of modelling hinge motions, and this is particularly suited to modelling multi-domain proteins. Usually this is done by locating hinge regions either manually or automatically (Emekli *et al.*, 2008), docking both sides of the hinge independently and then

reassembling the complex (Ben-Zeev *et al.*, 2005; Schneidman-Duhovny *et al.*, 2005a, 2007; Sandak *et al.*, 1998b,a; Cheng *et al.*, 2008; Karaca and Bonvin, 2011). The approach detailed in Karaca and Bonvin (2011), which has been implemented in the HADDOCK suite, is of particular interest as it allows the simultaneous modelling of hinge-bending, side-chain and backbone motions. Other interesting approaches to hinge modelling have been outlined by Wang *et al.* (2007a) and Zhao *et al.* (2006), in which flexibility is handled by novel data structures, although it remains to be seen whether these approaches can consistently model hinge motion efficiently and accurately.

Another method to account for flexibility is Monte Carlo sampling, in which backbone or side-chain conformational changes are proposed and either accepted or rejected depending on the energy of the new conformation. For instance, side chain Monte Carlo sampling has been implemented in ICM-DISCO (Abagyan and Totrov, 1994; Fernández-Recio *et al.*, 2003) and RosettaDock (Gray *et al.*, 2003). Side-chain rotamer prediction has also been tackled using molecular dynamics (Camacho and Gatchell, 2003; de Vries *et al.*, 2007), genetic algorithms (Tuffery *et al.*, 1991) and neural networks (Hwang and Liao, 1995). Another approach uses graph-theoretical models in which each side-chain is represented using a node, and those with an interacting rotamer pair are connected by edges. The graph can be decomposed such that the optimal set of rotamers is derived by combining the optimised rotatmer combination corresponding to sub-graphs (Krivov *et al.*, 2009).

Monte Carlo sampling can also be used to model the conformational changes of loops. Fixed-end-moves, the rotation of a number of atoms around two fixed points (Betancourt, 2005), have been observed in crystal structures (Davis *et al.*, 2006) and can capture some known protein motions (Friedland *et al.*, 2009). They have been used to model backbone motions in the RosettaDock program as part of the Monte Carlo move set (Fleishman *et al.*, 2010; Lauck *et al.*, 2010). Modelling backbone flexibility has also been done by varying ϕ and ψ torsion angles, either using Monte Carlo (Wang *et al.*, 2007a) or by simulated annealing molecular dynamics (de Vries *et al.*, 2007).

The final approach to modelling conformational changes that occur upon binding is to use normal modes. As many protein motions can be approximated using a small number of low frequency modes (see section 1.4.2.3), using a linear combination of normal modes is a promising approach for modelling conformational change. Aside from the approach analysed and implemented during the course of the PhD, and presented in later chapters, two other groups have dynamically adjusted movements along normal coordinates as part of a flexible docking strategy. In the ATTRACT program, the 5 lowest frequency non-trivial normal modes are used as degrees of freedom, along with the translation and orientation (May and Zacharias, 2008a). Quasi-Newton minimisation is performed from multiple starting positions, the proteins are represented using a coarse-grained model and the energy is calculated using a soft Van der Waals potential and electrostatics. The other approach is that used in the FibreDock refinement protocol (Mashiach *et al.*, 2010). In this protocol, the side-chains of rigidly soft-docked poses, derived from another method, are optimised using a linear programming routine (Kingsford *et al.*, 2005). Then rigid-body minimisation is performed followed by minimisation in normal mode space. In this step, the overlap between pre-calculated normal modes and the forces acting upon the binding partners is used to select 10 modes, and Monte Carlo sampling is undertaken in this normal mode space. Finally, after another round of rigid-body minimisation, the lowest energy solutions are returned.

1.5.7 Re-Ranking and Clustering

The above methods can be used to generate structures, either *ab initio* or by the refinement of initial structures. Re-ranking and clustering has shown to be an essential component of many docking protocols, either for filtering prior to refinement, or for final scoring and model selection.

The basis of clustering is the fact that long range forces steer the binding partners into regions of low free energy, and it has been shown to significantly improve the ranking of generated structures (Comeau *et al.*,

2004; Kozakov *et al.*, 2005; Lorenzen and Zhang, 2007; Tong and Weng, 2004; Zhang *et al.*, 2005a). A typical clustering protocol is exemplified by the ClusPro methodology (Comeau *et al.*, 2004; Kozakov *et al.*, 2005). In this approach the all-versus-all RMSD matrix is calculated and each structure scored by the number of other structures within a certain cut-off. The first cluster consists of the highest scoring structure and all structures within the cut-off of it. These structures are removed from the matrix and subsequent clusters are similarly determined. Clustering can be performed hierarchically, where the clustering threshold is adjusted, thus revealing low-resolution determinants of binding and sub-clusters within clusters. This information can then be used to guide model selection.

Re-ranking is the process of ranking generated structures using a higher accuracy, and generally more expensive, scoring function than that used in model generation. For instance, in the grid-based docking algorithm BIGGER, amino acid propensities and geometric complementarity is used to rank structures which are rapidly determined using fast heuristics and Boolean logic (Palma *et al.*, 2000; Krippahl *et al.*, 2003). Then, in the post-processing stage, amino acid propensities, surface complementarity, desolvation and electrostatics are used to score all the generated structures. These terms are fed through a neural network which was trained on a set of 25 protein-protein complexed using a back-propagation algorithm, in which the objective function was to maximise the distinction between false structures and near-native complexes. A more recent neural network based scoring function takes atom-pair distance distributions as the input. The neural network was trained on docking decoys for 185 protein-protein complexes, and tested against 65 complexes which weren't used for training (Chae *et al.*, 2010). This method managed to find near-native structures in the top 10 in 22 cases, when ranking unrefined structures generated from a correlation based systematic search. Another interesting approach is that of Bernauer *et al.* (2007), in which the interfaces of native structures and false positive docked solutions are converted to Voronoï diagrams; a set of polygons which encode information about the structure of the interface. Amino acids were categorised according to physiochemical properties: hydrophobic, aromatic, positive, negative, polar and small. The Voronoï

diagrams were used to derive a number of features relating to the composition and interactions at the interface. These features were then used to train a number of machine learning methods, including logistic regression, support vector machines and a genetic algorithm, which maximised the area under receiver operator curves. The latter two methods provided a considerable improvement in the ranking of structures generated using DOCK and HADDOCK.

One of the most popular re-ranking approaches uses pair potentials (Muller and Sticht, 2007; Huang and Zou, 2008; Liang *et al.*, 2007). These can be statistical potentials as described in section 1.4.4.4, or they can be derived to maximise discrimination between true and false docked poses. Alternatively, a combination of terms can be weighted to aid in ranking. For instance, the ZRANK program weights Van der Waals, electrostatics and a desolvation score (Pierce and Weng, 2007). The repulsive and attractive components of the Van der Waals term were separated, and the contribution of electrostatics was decomposed into four terms: short-range (< 5Å) attractive, short-range repulsive, long-range attractive and long-range repulsive. A simplex algorithm was used to determine the weights of these seven parameters to optimise the ranking of near-native structures.

1.5.8 Incorporating Experimental Data

Experimental data can be incorporated into a docking routine in various ways. For instance, many homomultimeric complexes are symmetric. Information about this symmetry can be derived from small angle x-ray scattering (SAXS), cryo-electron microscopy (cryo-EM) and NMR residual dipolar coupling and juxtaposed chemical shift. As the conformations of molecules in symmetric complexes are related by symmetry operations, only one of each molecule needs to be modelled, and the others are reconstructed via these symmetry operations, thus reducing the search space. A number of docking algorithms incorporate C_n symmetry, including M-ZDOCK (Pierce *et al.*, 2005), SymmDock (Schneidman-Duhovny *et al.*, 2005a), HADDOCK (Karaca *et al.*, 2010), RosettaDock (Andre *et al.*, 2007), MultiDock (Jackson *et al.*, 1998), MultiFit (Lasker *et al.*, 2010b), ClusPro (Comeau and Camacho, 2005) and others (Berchanski and Eisenstein, 2003;

Berchanski *et al.*, 2005; Huang *et al.*, 2005).

SAXS curves themselves can be used as part of a docking strategy, and although these offer low-resolution data, they are particularly suitable for characterising large and anisotropic complexes. As such, a number of docking protocols which incorporate SAXS data have been developed, which either filter docking results by comparing the synthetic SAXS profiles of docked solutions to experimental SAXS profiles, or combine the overlap of synthetic and experimental curves into the scoring function (Schneidman-Duhovny *et al.*, 2011; Forster *et al.*, 2008; Petoukhov and Svergun, 2005; Pons *et al.*, 2010).

Another source of experimental data which can be incorporated is cryo-EM, and tools for fitting components into cryo-EM density maps are available, such as Flex-EM (Topf *et al.*, 2008) and the Situs suite (Wriggers, 2010). Incorporating computational docking techniques with the fit between modelled structures and cryo-EM density has been a focus of research in the groups of Haim Wolfson and Andrej Sali (Lasker *et al.*, 2010b, 2009). Indeed, this is currently being incorporated into the Integrative Modelling Platform (IMP), in which restraints based on cryo-EM, SAXS and various other proteomics data can be simultaneously incorporated and weighted for the modelling of multimeric assemblies (Lasker *et al.*, 2010a).

Other sources of experimental information which illuminate biomolecular recognition include site-directed mutagenesis, where the influence of mutations on protein-protein binding can be ascertained (Otzen and Fersht, 1999), and H/D exchange, where the occlusion of interfacial hydrogens can be determined (Garcia *et al.*, 2004). Further, the physical properties of a protein surface and the evolutionary conservation of it amino acids can be used to predict binding sites or binding hot spots (Ezkurdia *et al.*, 2009; Tuncbag *et al.*, 2009; de Vries and Bonvin, 2008). This data, however, can be unreliable and although the identity of interacting residues may be known, specific contacts are usually not. The incorporation of this type of data into docking has been pioneered by the group of Alexandre Bonvin in the HADDOCK package (de Vries *et al.*, 2007). An additional energy term

is added to the scoring function in order to account for these ambiguous restraints. A list of residues which are believed to be involved in binding (active residues) and those which may be involved in binding (passive residues) is provided as an input to the algorithm. For each active residue the additional energy term is evaluated against all passive residues and the other active residues such that contacts between these residues are favoured. The algorithm is run several times, with different random subsets of these restraints, so that residues which may be erroneously identified as being interfacial do not systematically bias the search.

1.5.9 High-Throughput Docking

A number of key issues which, if resolved, could facilitate the ambitions of systems biology, were outlined in section 1.3.1.3. The methods discussed above show some promise at being able to derive structures of complexes. However, the confirmation or rejection of interactions discovered by high-throughput proteomics, the identification of new interactions, and the structural annotation of protein-protein interaction networks have not yet been considered. These questions are beginning to be addressed using high-throughput docking (Stein *et al.*, 2011a; Wass *et al.*, 2011; Matsuzaki *et al.*, 2009; Yoshikawa *et al.*, 2009; Mosca *et al.*, 2009). Attempt to use docking to pick out known interactions from all-versus-all docking or by picking known interactions from decoy docking partners has so far shown only limited success. The precision of these methods can be quite high, but the recall is low (Matsuzaki *et al.*, 2009; Yoshikawa *et al.*, 2009; Wass *et al.*, 2011). A first attempt at the structural annotation of an interactome has been undertaken, however estimates of both recall and sensitivity are both low, with at most a quarter of true structures being ranked as one of the top three docked structures (Mosca *et al.*, 2009).

1.5.10 The CAPRI Experiment

As docking algorithms progressed, the structural bioinformatics community saw a need for docking algorithms to be put to the test. In 1994, following the crystallographic determination of the complex between TEM-1 β-lactamase

and β-lactamase inhibitor 1 by researchers at the University of Alberta, nine laboratories were set the challenge of predicting the structure of the complex given nothing but the coordinates of the unbound proteins (Strynadka *et al.*, 1996). Six of the groups rose to the challenge, submitting between 1 and 15 complexes each. To determine the accuracy of the submitted answers, the RMSD between the submissions and the crystal structure were calculated as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{1.70}$$

where N is the number of atoms, and δ_i is the distance between the ith atom of the bound structure and the ith atom of the predicted structure. The coordinates were optimised to find the minimum RMSD, and this was used to judge the accuracy of the model. All six of the groups produced models with RMSD below 2.5Å, and five below 2.0Å. In hindsight, the target was an easy one, being an enzyme-inhibitor complex, having subnanomolar affinity (Albeck and Schreiber, 1999), and with a change in interface RMSD below 0.5Å, however it was recognised that blind tests would be the acid test for the state of macromolecular docking. Soon after, another docking challenge was issued as part of CASP2, a blind test for the protein/homology modelling community commenced in order to determine the state of the modelling field. The docking community soon followed suit with CAPRI; Critical Assessment of PRediction of Interactions. Generous structural biologists were asked to volunteer their structures prior to release, so that anyone who chooses to register can test their algorithms (Vajda *et al.*, 2002).

One of the earliest questions raised pertained to the evaluation of submissions. The RMSD of the complex is not a perfect score, as if the interface is essentially predicted correctly, a small rotation of a large prolate ligand causes a larger increase in RMSD than for a smaller spherical one, to such an extent that one submission could be deemed better than another with a more accurately modelled interface. Further, physically impossible predictions with many clashes could rank highly. To resolve these issues, and others, a number of different measures are evaluated to judge the goodness of prediction.

1. Interface RMSD, I_RMSD; the minimum RMSD of interface main-chain atoms, where the interface is defined as the residues which lose surface area upon binding.

2. Ligand RMSD, L_RMSD; the RMSD of the smaller of the binding partners, after superimposition of the larger.

3. F_nat; The fraction of native residue-residue contacts correctly predicted, where a ligand residue and a receptor residue are deemed to be in contact if an atom in one is within 5Å of an atom in the other.

4. F_nonnat; The fraction of non-native contacts which appear in the predicted structure.

Submitted models are removed if they contain too many clashes. The remaining models, are categoried as either high accuracy (F_nat $\geq$ 0.5 and either I_RMSD $\leq$ 1.0 or L_RMSD $\leq$ 1.0), medium accuracy (F_nat $\geq$ 0.3 and either 1.0 < I_RMSD $\leq$ 2.0 or 1.0 < L_RMSD $\leq$ 5.0), acceptable (F_nat $\geq$ 0.1 and either 2.0 < I_RMSD $\leq$ 4.0 or 5.0 < L_RMSD $\leq$ 10.0), or incorrect (F_nat < 0.1).

Usually, each round of the CAPRI experiment is split into two parts; prediction and scoring. In the prediction part, the participants are given the unbound structure or, if this is not available, the amino acid sequence for homology modelling. Participants then upload 10 predicted structures, ordered by preference. Participants are also encouraged to upload up to 100 structures for the scoring round. In the scoring round, participants can download all the uploaded structures and are given the challenge of selecting the best models, up to 10 of which are refined and uploaded. Teams are then ranked according to how many targets they successfully predicted, and similarly for how many targets are correctly selected during the scoring round, although this procedure has been criticised and an alternative method of ranking the scorers has been proposed (Feliu and Oliva, 2010).

Approximately every two years, members of the CAPRI community are invited to convene at a conference, and a special issue is published in the journal 'Proteins: Structure, Function and Bioinformatics', in which the

performance of the participants is analysed, and the latest developments in docking algorithms published; Volume 52, Issue 1 (2003), Volume 60, Issue 2 (2005), Volume 69, Issue 4 (2007) and Volume 78, Issue 15 (2010). A number of groups have participated regularly in the CAPRI experiment, and as time goes on this number is growing.

Chapter 2

Normal Mode Analysis and Conformational Transitions

2.1 Introduction

Protein-protein docking can be formulated as an optimisation problem, in which an energy function is optimised in the given degrees of freedom. Rigid-body docking, in which only the 6 translational and rotational degrees of freedom are optimised, cannot capture the conformational changes which occur when proteins bind to one another. For a system of N particles, including full flexibility requires $3N$ degrees of freedom, one for each coordinate. A representation of conformation can take a number of forms. For instance, in the Cartesian representation, each coordinate corresponds to either the x, y or z position of an atom. In the internal coordinate system, each coordinate corresponds to a bond length, valence angle or dihedral angle. The difficulty of an optimisation problem is related to the number of variables which need to be simultaneously optimised, and so increasing the search space by including full flexibility vastly increases the difficulty of optimisation. In section 1.5.6, a number of ways of incorporating flexibility with fewer degrees of freedom were outlined, including restricting flexibility to side-chain and/or flexible loops, or just optimising torsion angles. The potential success of any reduction in complexity is predicated upon the degrees of freedom being eliminated not being relevant to the conformational changes observed. For instance, the observation that bond lengths do not vary significantly when proteins bind

to one another allows these degree of freedom to be eliminated when using an internal coordinate system, leaving only $2N$ variables for optimisation.

Normal coordinates, the derivation of which appears in section 1.4.2.3, are complete, as in any possible conformation can be expressed as a linear combination of some, or all, of the $3N$ modes. They are also orthogonal; no motion along one mode can be expressed as a linear combination of any other modes. Hence, they can be considered a candidate coordinate system. Indeed, representing protein conformation as a linear combination of normal modes has been discussed as a coordinate system for guided protein-protein docking (Andrusier *et al.*, 2008; Bonvin, 2006; May and Zacharias, 2005), has been used used in protein-small molecule docking (May and Zacharias, 2008b; Floquet *et al.*, 2006; Sander *et al.*, 2008; Cavasotto *et al.*, 2005; Kovacs *et al.*, 2005), to refine protein-DNA and protein-small molecule interactions (Zacharias and Sklenar, 1999; Lindahl and Delarue, 2005) and generate structures for protein-small molecule cross-docking (Rueda *et al.*, 2009). It has also been used in two protein-protein docking routines, as discussed in section 1.5.6 (May and Zacharias, 2008a; Mashiach *et al.*, 2010). In these approaches, higher frequency normal modes were eliminated, on the basis that many protein motions resemble a single low frequency normal mode, usually one of the first 5 (Yang *et al.*, 2007a; Krebs *et al.*, 2002; Tama and Sanejouand, 2001; Atilgan *et al.*, 2001). In the context of protein-protein binding, Dobbins *et al.* (2008) demonstrated that the same is true when considering the global conformational changes which occur when highly flexible proteins bind to one another. These results suggests that normal modes can be a very powerful representation of protein conformation. Using an internal coordinate system, even if all bond lengths and valence angles are eliminated, there are still as many degrees of freedom as there are atoms, which can be well over a thousand. The above studies suggest that the conformational changes which occur upon binding could be modelled using only a few low frequency modes, say 5 or 10, conferring a huge advantage to this model.

These studies, however, neglect some salient issues pertaining to the use of normal coordinates when docking. Most of them use a coarse-

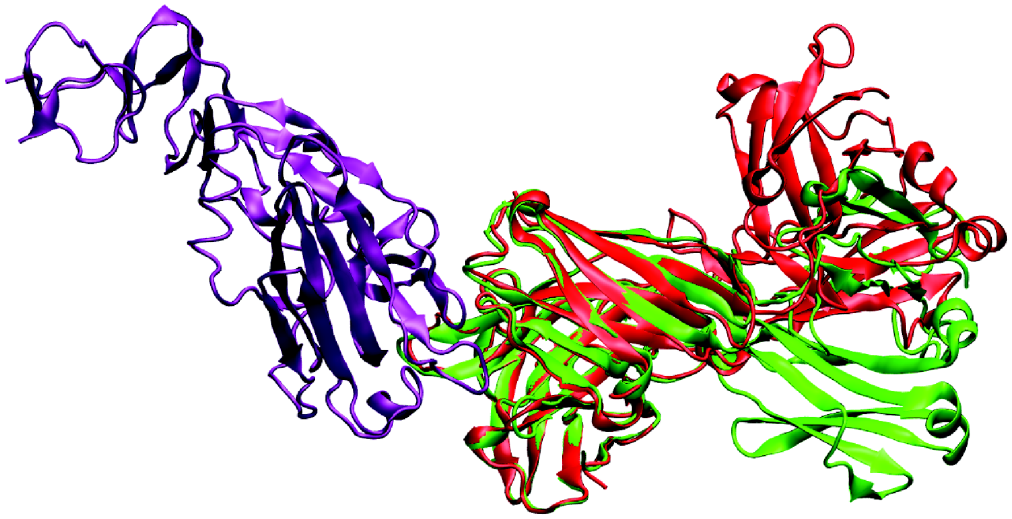

Figure 2.1: The complex between murine IgG1,λ HC19 antibody in the bound (green) and unbound (red) conformation, and influenza hemagglutinin (purple), PDBid 2VIS. The structures are superimposed on the binding domain. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

grained model of protein conformation, whilst most docking algorithms operate at atomic resolution, especially those which aim at producing high-resolution models. All of them focus upon single modes in isolation, whilst for guided docking algorithms the mode which best represents the conformational change is not know, and thus multiple modes must be taken in linear combination. Finally, they all focus on global conformational changes occurring across the whole fold, whilst it is known that *in silico*, as *in vivo*, protein-protein binding is driven by changes which occur at the binding interface, and it is the reconstruction of these changes which is of greatest import in a docking protocol.

To illustrate this last point, consider the interaction between influenza hemagglutinin and the HC19 IgG1,λ antibody, shown in Figure 2.1, for which the lowest frequency mode can capture the observed conformational change (Dobbins *et al.*, 2008). It is clear in this case that the normal mode analysis is identifying the hinge motion, and that the hinge motion is not functionally relevant, as it corresponds to the movement of a domain which is not involved in binding. It is likely that the observed hinge motion is due to crystal packing forces acting upon the intrinsic flexibility of the protein, as is the case for the differing hinge angles observed when D-allose-binding

protein is crystallised in different space groups (Magnusson *et al.*, 2002). Another example of global conformational differences arising due to crystal packing forces is given by two different crystal structures of the T4 lysozyme. Molecular dynamics simulations of these structures show convergence of the interdomain hinge angle after 500ps (Arnold and Ornstein, 1997). Additionally, there are the different hinge angles of the same protein in the asymmetric unit of the rabbit tissue factor crystal structure (Muller *et al.*, 1998b), and different crystallographic conformations of the SIV proteinase (Wilderspin and Sugrue, 1994). Thus, while it is difficult to ascertain which global conformational changes are due to crystal packing, and which are functionally related to binding, interfacial conformational change upon docking is expected to be less susceptible to this effect, as this change is either not associated with global change, or the global movements are functional aspects of the binding, such as when binding to a cleft between two domains.

In this chapter, the above issues are addressed by evaluating the ability for normal coordinates to capture unbound to bound conformational transitions at atomic, main-chain and residue level resolution, across the whole fold, and at the interface, on a large data set of 236 proteins. This reveals the upper limit of what conformational changes can be modelled using low frequency modes in linear combination during a docking run. This work has been published in Moal and Bates (2010).

2.2 Methods

2.2.1 Normal Mode Analysis

Calculating normal modes can be very computationally expensive. The cost arises from having to extensively optimise the system prior to calculating the Hessian. The elastic network model is an alternative formulation which is much less computationally demanding, and is the one used here (Tirion, 1996; Bahar *et al.*, 1997; Atilgan *et al.*, 2001).

2.2.1.1 The Elastic Network Model

In the elastic network model (ENM), the system is treated as balls on springs. For all atom or residue pairs A and B, within a certain cutoff, c, a Hooke potential V_{AB}, equation 1.12, is employed. The same force constant, $k = 1.0$, is used for each spring, and r_e, the equilibrium spring length, is set to be that given in the initial structure, ensuring that this structure is the energy minimum. Thus, the potential energy function becomes

$$V_{tot} = \sum_{r_{AB}<c} V_{AB} \tag{2.1}$$

A cutoff of $c = 10$Å is used for all calculations. This function is used to construct the Hessian and yield the normal modes (see section 1.4.2.3). This model may seem very simplistic, only taking account of the shape of the molecule, and ignoring all electrostatics and hydrogen bonding and so on, as well as only being relevent to small linear motions and ignoring multiple minima and solvent damping effects. Nevertheless, the model, which is robust to the value of the cutoff used (Kurkcuoglu *et al.*, 2006), has been able to reproduce thermal B factors at atomistic resolution (Tirion, 1996), and residue-level resolution (Bahar *et al.*, 1997), as well as atomic fluctuations as determined using NMR (Yang *et al.*, 2007b) and molecular dynamics (Rueda *et al.*, 2007). Additionally, it has been shown that for most known protein motions, the global rearrangements of the fold can be modelled using a single low-frequency mode (Yang *et al.*, 2007a; Krebs *et al.*, 2002; Tama and Sanejouand, 2001; Atilgan *et al.*, 2001; Dobbins *et al.*, 2008; Cui *et al.*, 2004; Petrone and Pande, 2006). In this work, the Hessian matrix is constructed using pdbmat, part of the ElNeMo package (Suhre and Sanejouand, 2004). To diagonalise the Hessian, we used either the Basic Linear Algebra Subprograms (BLAS) library (Dongarra, 2002), or diagrtb, an implimentation of the rotation-translation-of-blocks method (see below) (Suhre and Sanejouand, 2004). Calculations were done on unbound structures, at either a coarse-grain level, or at atomistic detail. For the coarse-grain calculations, only C_α positions were considered. For the atomistic calculations, small molecule ligands, but not crystallographic waters, were used to construct the Hessian. These were ignored for the remainder of the analysis.

2.2.1.2 Rotation-Translation-of-Blocks

The rotation-translation-of-blocks method offers considerable computational saving, as unwanted higher frequency modes are not calculated. Blocks of atoms, usually of one residue or more, are treated as rigid units capable of rotation and translation. The all-atom Hessian is projected into a block translation and rotation subspace by applying a projection matrix. This projected Hessian is used to determine vibrational frequencies and eigenvectors (Tama *et al.*, 2000; Li and Cui, 2002). The all-atom eigenvectors are then constructed by applying the transpose of the projection matrix to the block eigenvectors, to produce approximate low-frequency modes. Subsequently, these approximate all-atom modes are iteratively perturbed toward the exact solution by using higher frequency modes calculated independently for each block (Durand *et al.*, 1994; Durand, 1983). Compared to the standard method of diagonalising the all-atom Hessian, this method can reproduce frequencies and atomic fluctuations with comparable accuracy. In this work, single residue blocks are used.

2.2.2 Overlap

The ability for the j^{th} normal mode to capture the displacement of atoms involved in a protein motion is given by its overlap, O_j, with that motion (Marques and Sanejouand, 1995).

$$O_j = \frac{|\sum_{i=1}^{3n} a_{ij}(r_i^b - r_i^u)|}{\sqrt{\sum_{i=1}^{3n} a_{ij}^2 \sum_{i=1}^{3n}(r_i^b - r_i^u)^2}} \tag{2.2}$$

where r_i^b and r_i^u are the i^{th} coordinates of the bound and unbound structures respectively and a_{ij} is the i^{th} coordinate of the j^{th} normal mode. If the movement is in exactly the same direction as the mode, this function has a value of 1. If the mode and the conformational change are orthogonal, it takes a value of 0. The overlap can be calculated over the whole set of points used in the ENM, or a subset of those points, such as the interface.

Two sets of m modes, **a** and **b**, can be compared using the overlap matrix, for which the i^{th} element of the j^{th} row corresponds to the overlap between the i^{th} mode of **a** and the j^{th} mode of **b**. The degree of similarity between these

two sets of modes can be ascertained by calculating, $\bar{O}$, the root-mean-square inner product between them.

$$\bar{O} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{a}_i\mathbf{b}_j)^2} \tag{2.3}$$

A value of 1 indicates that all the modes in **a** can be expressed as a linear combination of modes in **b**. A value of 0 indicates that all modes in **a** are orthogonal to all modes in **b**.

2.2.3 Modes in Linear Combination

The overlap value only tells us how closely a single mode corresponds to a conformational change, and not how well a number of modes in combination can capture the change. A method of determining the ability of a linear combination of normal modes to capture unbound to bound transitions was used by Lindahl and Delarue (2005) and Mustard and Ritchie (2005). As the modes are orthogonal, a projection was used to calculate the contribution of each mode independently. However, when investigating subsets of atoms such as the interface, orthogonality breaks down and this technique overestimates the degree to which the conformational change can be modelled. As this approach is not applicable in such cases, an alternative approach was formulated; the unbound to bound transitions were decomposed through basis expansion of a limited number of low-frequency modes. When the mode set is orthonormal, this method is identical to that used by Lindahl and Delarue (2005). This basis expansion was achieved using linear least-squares. It is exact and analytical, and was used to determine the extent of deformation along each mode that was necessary to generate the closest possible structure to the bound using a subset of normal coordinates. Before fitting, the atom sets which are being mapped are superimposed. As this approach has not previously been used for this purpose, a full derivation follows.

Take **D** to be a vector of length n containing all the Cartesian coordinates of all the atoms in the unbound structure, or to be the coordinates of a subset of atoms in the unbound structure, such as main chain atoms, or interface

atoms. Similarly, take **E** for the bound conformation. The unbound to bound transition is defined as $\mathbf{T} = \mathbf{E} - \mathbf{D}$. The i^{th} element of **T** is defined as T_i. Take **M** to be an $m \times n$ matrix containing m normal modes, with each column corresponding to a different normal mode and each row corresponding to each coordinate, such that M_{ij} is the displacement of the j^{th} coordinate of the i^{th} mode. We define $\mathbf{M}_j$ as the vector of atomic displacements for the j^{th} mode.

Firstly, we wish to derive an expression in which the unbound to bound transition is equal to a linear combination of normal modes, $\sum_{j=1}^{m} \beta_j \mathbf{M}_j$, plus some residual vector **r**

$$\mathbf{T} = \mathbf{E} - \mathbf{D} = \mathbf{r} + \sum_{j=1}^{m} \beta_j \mathbf{M}_j \tag{2.4}$$

where $\boldsymbol{\beta}$ is defined as a vector of the β coefficients. Rearranging 2.4, r_i, the residual of the i^{th} coordinate for any given $\boldsymbol{\beta}$ can be calculated as

$$r_i(\boldsymbol{\beta}) = T_i - \sum_{j=1}^{m} M_{ij}\beta_j \tag{2.5}$$

We wish to find the set of β coefficients which minimise the sum square residuals, S, which can be obtained using

$$S(\mathbf{r}) = \sum_{i=1}^{n} r_i^2 \tag{2.6}$$

The particular set of β coefficients which minimise this function is given the symbol $\hat{\boldsymbol{\beta}}$. In order to find this set, we must find where the gradient of S is equal to zero. Using the chain rule, we have

$$\frac{\partial S}{\partial \beta_j} = \frac{\partial S}{\partial r_i}\frac{\partial r_i}{\partial \beta_j} = 0 \tag{2.7}$$

As the residuals $\mathbf{r}(\boldsymbol{\beta})$ are linear with respect to the β coefficients, so is their sum. Therefore, the sum square residual, $S(\mathbf{r}(\boldsymbol{\beta}))$ is a quadratic function, and its derivative is linear with one unique root. It is evident that $S(\boldsymbol{\beta})$ is also a convex function; if any of the β coefficients are pushed to extremes, then S becomes very large and positive. Hence, this solution corresponds to a minimum. Differentiating 2.6 with respect to the residuals, we find

$$\frac{\partial S}{\partial r_i} = 2\sum_{i=1}^{n} r_i \tag{2.8}$$

Differentiating 2.5 with respect to β_j yields

$$\frac{\partial r_i}{\partial \beta_j} = \sum_{j=1}^{m} -M_{ij} \tag{2.9}$$

Substituting 2.5 into 2.8, and substituting the resulting equation and 2.9 into 2.7, we find the following series of equations

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{n}\left(T_i - \sum_{k=1}^{m} M_{ik}\hat{\beta}_k\right)\left(-M_{ij}\right) = 0 \qquad \text{for} \qquad j = (1,2,3,...,m) \tag{2.10}$$

Expanding the bracket

$$\frac{\partial S}{\partial \beta_j} = -2\sum_{i=1}^{n} M_{ij}T_i + 2\sum_{i=1}^{n}\sum_{k=1}^{m} M_{ij}M_{ik}\hat{\beta}_k = 0 \qquad \text{for} \qquad j = (1,2,3,...,m) \tag{2.11}$$

Rearranging, the 2s cancel and we get the normal equations

$$\sum_{i=1}^{n}\sum_{k=1}^{m} M_{ij}M_{ik}\hat{\beta}_k = \sum_{i=1}^{n} M_{ij}T_i \qquad \text{for} \qquad j = (1,2,3,...,m) \tag{2.12}$$

Expressed in matrix notation

$$(\mathbf{M^T M})\hat{\boldsymbol{\beta}} = \mathbf{M^T T} \tag{2.13}$$

Inverting the normal equations shows how the fitted coefficients, $\hat{\boldsymbol{\beta}}$, are obtained. These coefficients correspond to the magnitudes of deformation along their respective normal coordinates.

$$\hat{\boldsymbol{\beta}} = (\mathbf{M^T M})^{-1}\mathbf{M^T T} \tag{2.14}$$

The inversion was achieved using the Cholesky decomposition routine implemented in the Python NumPy library. For illustration, an example is

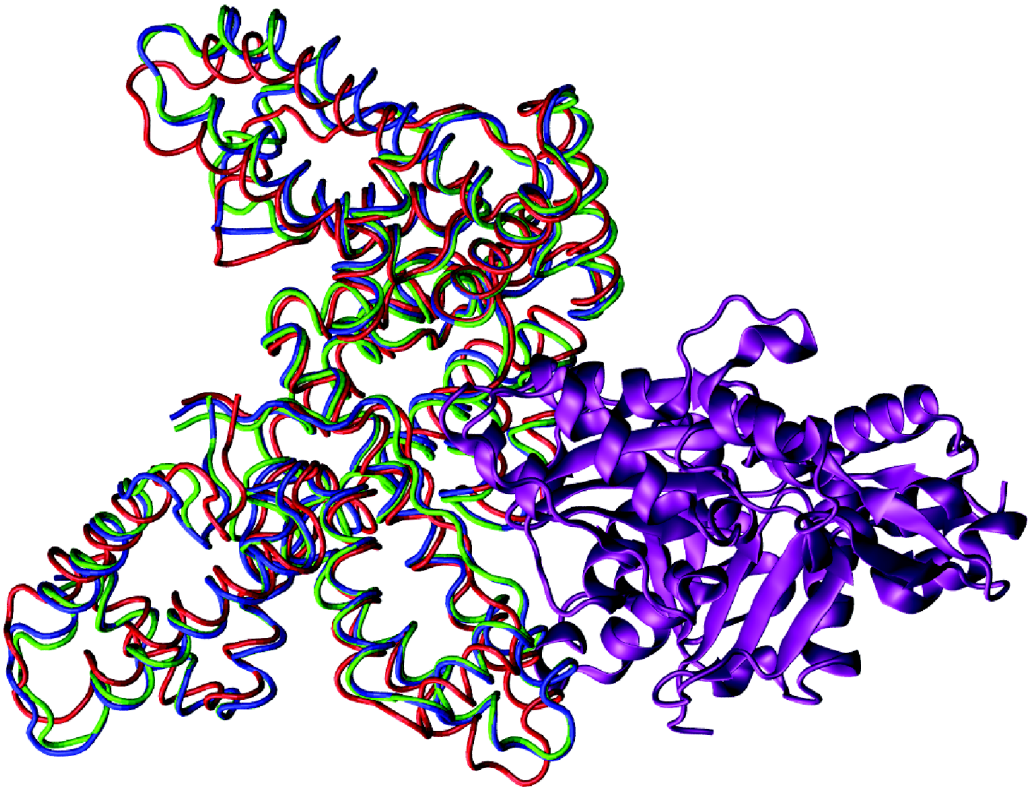

Figure 2.2: The interaction between actin (purple) and vitamin D binding protein (VDBP, PDBid 1KXP). Bound VDBP is shown in green, and unbound in red. The C_α RMSD between these is 2.12Å. In blue is the fitted structure, as determined by linear regression to the unbound to bound transition using 20 normal modes as a basis. This corresponds to the closest possible structure to the bound that can attained with the 20 lowest frequency modes. The RMSD against the bound has reduced to 0.87Å. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

given in Figure 2.2.

In order to achieve the above fitting, a one-to-one correspondence of atoms in the bound and unbound state is necessary. For some complexes in the data set, this was the case. For the other cases, sequences were aligned and non-matching residues were ignored in the mapping, but not in the construction of the Hessian.

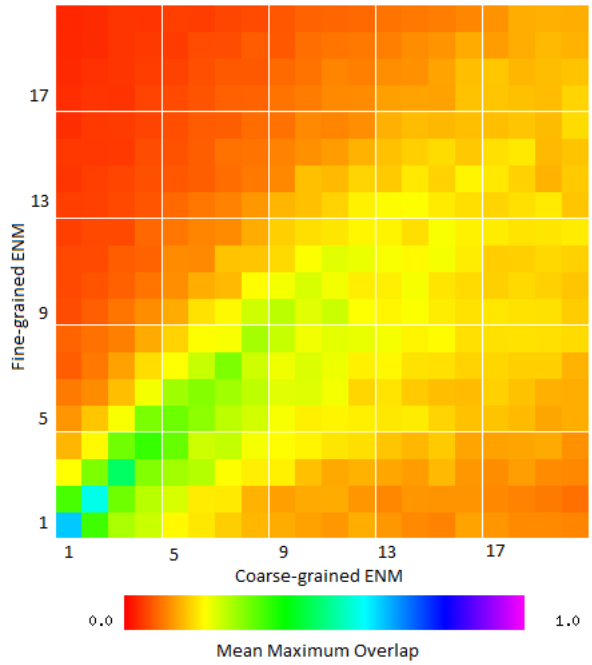

Figure 2.3: The mean overlap matrix between the fine grained (y-axis) and coarse grained (x-axis) ENM for the lowest frequency 20 non-trivial modes. Matrix elements correspond to the mean value of that element across the whole data set. This graph shows that the low frequency modes in the fine-grained model have an approximately 1:1 correspondence with the modes in the coarse-grained model. Only minor deviations in the ordering of modes is observed, which is to be expected when the lowest modes are of similar frequency.

2.2.4 Data Set

For the fitting, 124 complexes in the Protein-Protein docking benchmark v3.0 were used (Hwang *et al.*, 2008). Both binding partners were analysed, giving 236 conformation changes upon binding, after the removal of 12 cases for which the unbound structure is known for only one binding partner. All structures have a resolution below 3.25Å and redundancy is alleviated as described by Chen *et al.* (2003). The interface was determined as amino acids containing a heavy atom within 6Å of a heavy atom on the binding partner.

2.3 Results

2.3.1 Atomistic and Coarse ENM

As the elastic network model has mostly been studied using coarse-graining, and all of the following analysis uses atomistic normal modes, the degree of correspondence between the coarse and atomistic ENM was evaluated to ensure that this model is appropriate. There are $3N - 6$ non-trivial modes for an ENM of N nodes. Moving from the coarse-grained model to the fine-grained ENM, the low-frequency collective modes are preserved whilst the additional modes derived correspond to localised nanosecond scale motions. Between the coarse-grained and fine-grained ENM, using the lowest 20 non-trivial normal modes, there is a mean root-mean-square inner product of 0.97 averaged over the whole data set, indicating great agreement between the low frequency modes derived using both models. Figure 2.3 shows the mean value of elements in the overlap matrices, indicating a prevalence of one to one correspondence for the lowest modes.

2.3.2 Single Modes

The overlap between the conformational change upon binding and low frequency modes was calculated for the whole data set, at the interface and across the whole fold, and at atomic, backbone and C_α resolution. For each protein, the mode amongst the lowest frequency 20 which has the greatest overlap with the conformational change, was found. This maximum overlap, averaged across the whole data set, is shown in Figure 2.4. When the Hessian is diagonalised in the standard way, the mean maximum overlap increases as the size of the subsystem decreases. As backbone atoms move collectively with their adjacent atoms, the difference between the backbone and the C_α mean maximum overlap is small. There is a greater difference in mean maximum overlap when we compare backbone to all atoms, a consequence of the fact that whilst the side chains do move in concert with their adjacent backbone atoms, they can also move independently and change rotameric state. When the mean maximum overlap of the whole protein is compared to that of the interface, a decrease is observed at all levels of

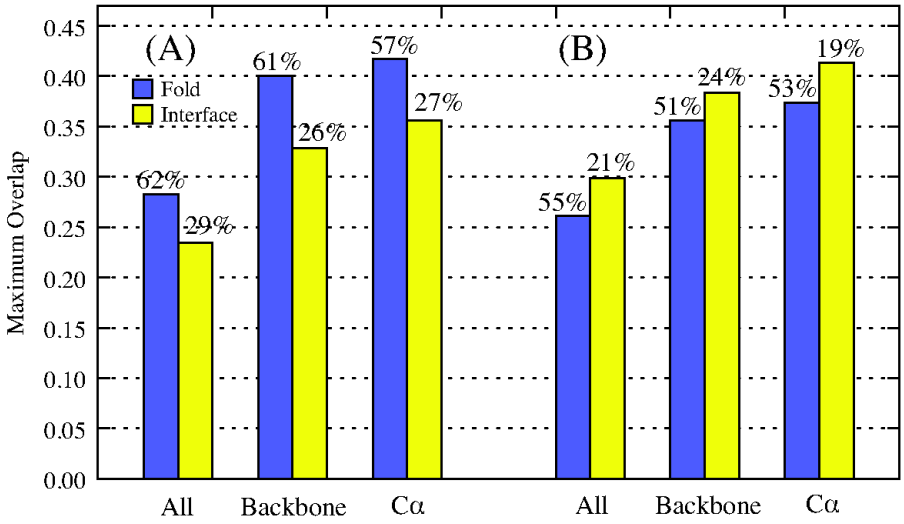

Figure 2.4: Mean maximum overlap with the first 20 modes, across the whole fold and the interface, for all atoms, backbone atoms and C_α atoms, calculated with (A) standard diagonalisation of the Hessian and (B) the RTB approach. The percentage of complexes for which the mode of maximum overlap is one of the first five non-trivial modes is shown above the bars. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

resolution. Additionally, while the global motion is often best represented by one of the 5 lowest frequency modes (57%-62% of the modes of maximum overlap), the same is not true of conformational change at the binding interface (26%-29%).

When this analysis is extended to the modes of maximum overlap covering the lowest 500 normal modes, a difference can be seen between the motion of interface C_α atoms and all interface atoms, as shown in Figure 2.5. Nevertheless, when compared to the distribution of modes of maximum overlap for global motion, much higher frequency modes are clearly involved in the more subtle changes that occur at the binding interface. This demonstrates that the conformational change of the binding interface does not usually resemble one of the five lowest modes, whilst the global change does. Petrone and Pande (2006) suggested that, upon complexation, a proteins binding partner can induce higher frequency modes within it, which is consistent with the activation of higher frequency modes observed here.

2.3.2.1 Rotations-Translation-in-Blocks Method

The normal modes were recalculated using the RTB method, and compared with the results arrived at from the exact Hessian diagonalisation shown above. The mean maximum overlaps are shown in Figure 2.4B. The same trend is observed going from atomic resolution to a coarser representation.

There is, however, a noticeable difference in the methods ability to model conformational change across the whole fold when compared to the interface. The mean overlap across the interface is 0.30 and 0.41 for the atomistic and C_α levels of resolution respectively. This is notably higher than for the more computationally demanding method of exactly diagonalising the Hessian (0.23 and 0.36), and is comparable to the well documented ability for the ENM to capture global motion (0.28 and 0.42). It is not clear the reason why this is the case, but it may be related to the way in which the low frequency modes are perturbed with higher modes calculated for each residue. Nevertheless, this does suggest that not only being a faster and less memory demanding technique, the RTB approach is also a better choice when pre-calculating normal modes for use in docking.

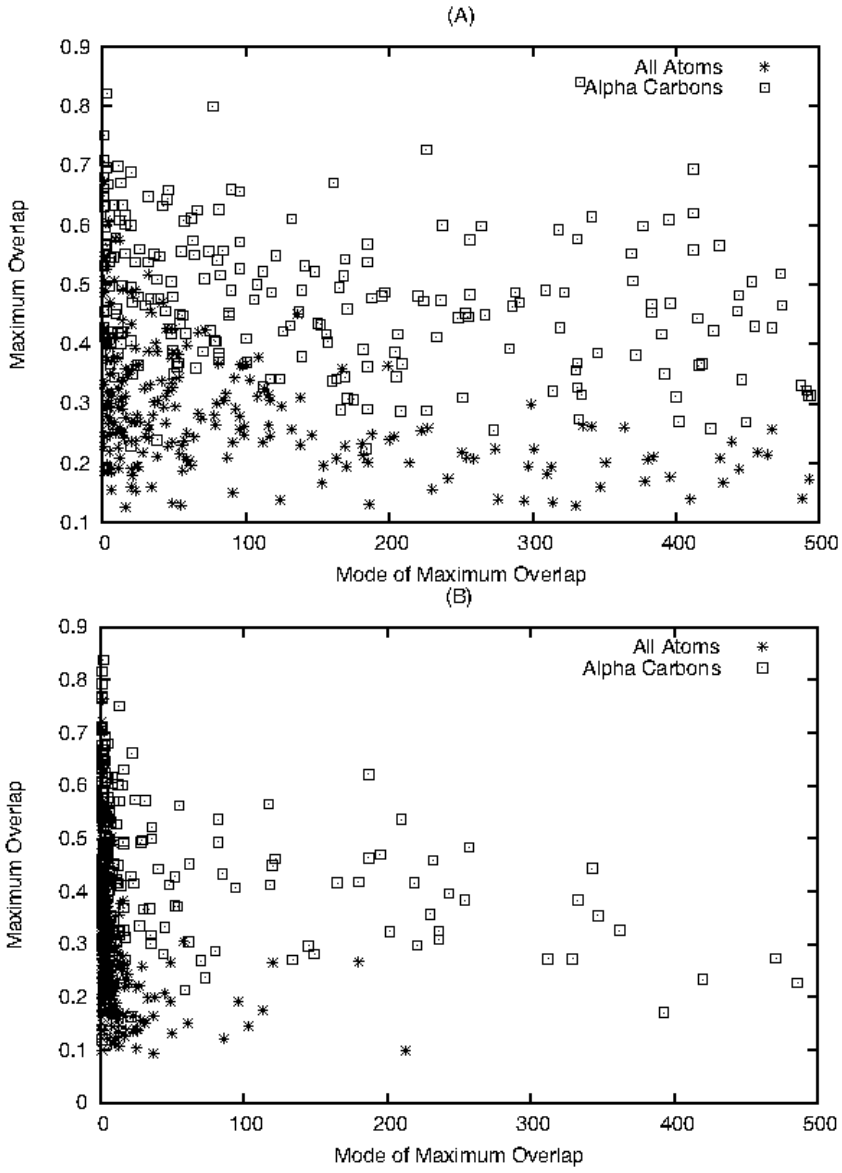

Figure 2.5: Maximum overlap and respective mode for (A) interface residues and (B) the whole fold. This demonstrates that the changes at the interface are most closely associated with higher frequency motions. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

2.3.3 Modes in Combination

Although the overlap gives a good measure of the degree to which the conformational changes resemble any given single normal mode, without the bound structure it is not possible to know which mode has the greatest overlap. Hence, it is preferable to include a number of modes in a docking strategy. Furthermore, if modes are used in linear combination, the ability to model the unbound to bound transition may be significantly enhanced compared to when single modes are used in isolation, whilst still vastly reducing the degrees of freedom which need to be modelled compared to an internal or Cartesian coordinate representation of conformation. To investigate this, the conformational changes that occur upon binding were decomposed into a linear combination of normal modes using the method outlined in section 2.2.3. It should be noted that the purpose of this analysis is not to model dynamics, the original purview of normal mode analysis, but to determine the potential efficacy of using normal coordinates in a docking algorithm. Hence there is no consideration of the phase or energy of each mode.

For each structure in the data set, the coefficients for multiple subsets of low-frequency modes were obtained which, of the infinite structures that can be generated with the linear combination, minimise the RMSD against the bound conformation. The minimisation was performed at three levels of resolution, both across the whole protein and at the interface. The mean reduction in RMSD, as a percentage of the initial bound-unbound RMSD, is shown plotted against the number of low frequency modes used in the fitting in Figure 2.6. At the C_α resolution and using the 5 lowest modes, only 44% of the complexes have a greater reduction in interface RMSD than across the whole fold. However, when 10, 15 and 20 modes are used, this figure increases to 55%, 60% and 64% respectively. The same trend is observed when considering all atoms (41%, 45%, 50% and 52%). Evidently, as the number of modes used increases, the ability to model the unbound to bound transition at the interface improves at a greater rate than across the whole fold, even though the conformational change across the fold has greater overlap with single low frequency modes. Hence, despite the observation that much higher frequency modes require consideration

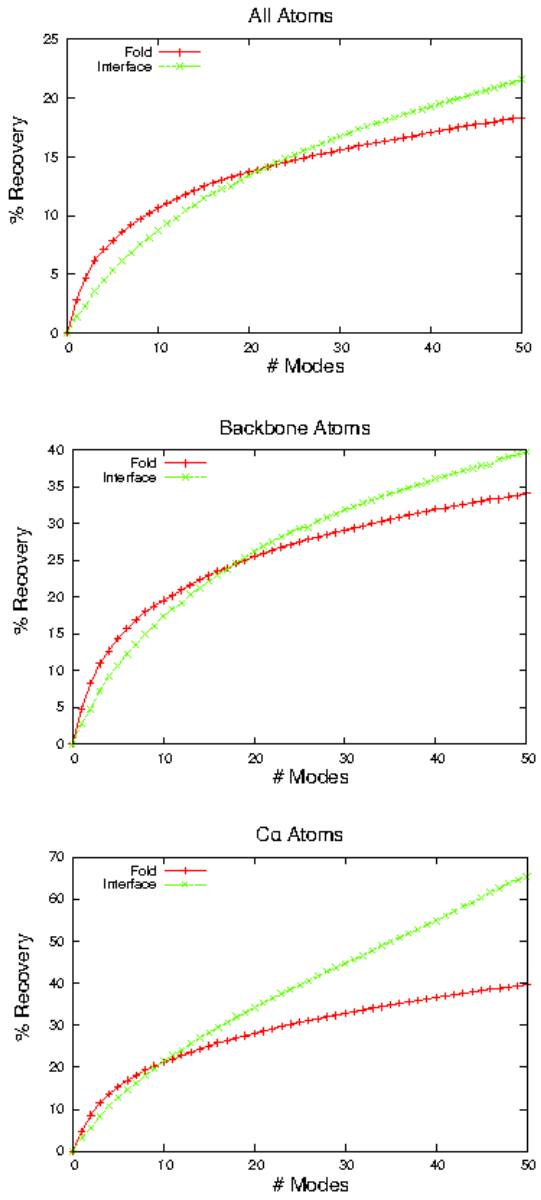

Figure 2.6: Mean percentage reduction in RMSD as a function of the number of modes used. Initially, the rate of improvement with respect to the modes is great, but lowers to a constant as the graph tends towards more modes. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

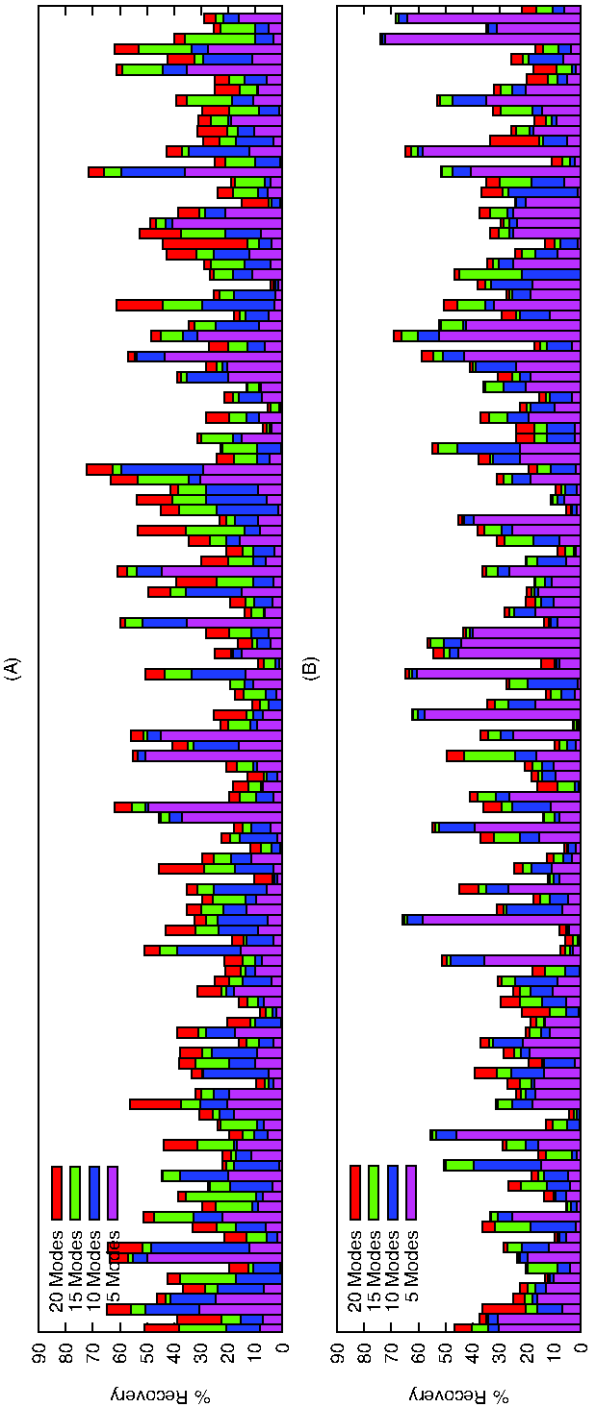

Figure 2.7: Percent decrease in RMSD upon regression of C_α atoms against the conformational change using a basis of 5, 10, 15 and 20 low frequency normal modes. These values are for the 130 structures with the largest initial C_α RMSD for the interface (A) and the whole protein (B). From left to right, structures appear in ascending order of initial C_α RMSD. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

in order to find a single mode which represents conformational change at the interface, not all the modes are required to model this change; a fewer number in linear combination will suffice. Indeed, the RMSD for interface C_α atoms is reduced by almost 70%, on average, when 50 modes are employed.

Usually, only a subset of the modes used in the fitting contribute significantly to the reduction in RMSD. The mode of greatest influence is that with the largest coefficient. When 20 modes are used, the average mode of greatest influence across the fold is 5.0 and 6.8, when using all atoms and C_α atoms respectively. However, when the interface is considered, this increases to 8.1 and 9.3, consistent with Figure 2.5 and the notion that higher frequency motions are involved with the conformational rearrangements at the interface. A broad breakdown of the distribution of significant modes for a number of the higher-flexibility proteins is given in Figure 2.7. The salient features of this graph illustrate a number of conclusions. Firstly, when 20 modes are used, the reduction in RMSD is, on average, greater across the interface than across the whole protein. Secondly, whilst the conformational changes across the fold are mostly modelled using the lowest 5 modes (purple), higher frequency modes are more often involved in the changes which occur at the interface (red and green).

In Figure 2.8, a more detailed breakdown of mode contributions is given for the 30 complexes of highest flexibility, as calculated using 20 modes. A notable feature of this image is structure 5, the structure of influenza hemagglutinin and its paratope (PDBid 2VIS), discussed in section 2.1. Here it is clear that the global hinge-bending motion, which probably arises due to crystal packing forces, is associated with the first mode and is unrelated to the subtler conformational rearrangements occurring at the interface. For most the complexes shown, the distribution of modes at the interface is distinctly different from the modes which contribute to global motion. There are, however, a number of exceptions: I1BR (Ran/Importin β complex), 1E4K (immunoglobulin G Fc fragment/FcγRIII complex) and 1Y64 (actin/BNI1 complex). All these complexes correspond to larger columns. For the Ran/Importin β complex, the ligand spans the

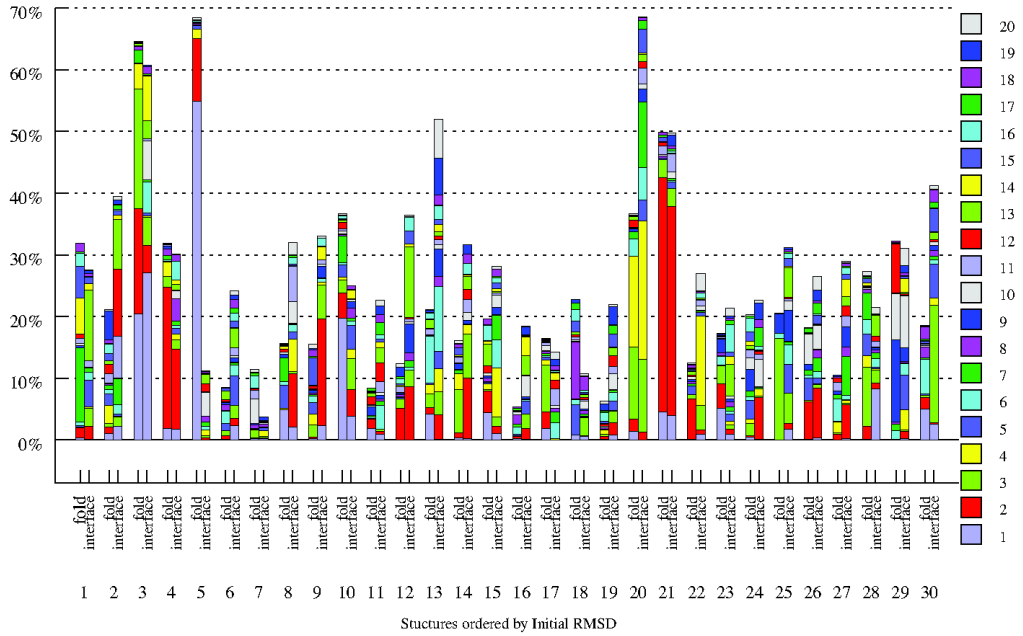

Figure 2.8: Percentage reduction of initial RMSD for the 30 structures of greatest RMSD across the fold. Structures appear from left to right in descending order of initial all-atom RMSD. Effect of inclusion of 1 to 20 modes is shown for both the interface and the whole fold. Structures included are (1) 1IRA_r,(2) 1H1V_l,(3) 1Y64_r,(4) 1FAK_r,(5) 2VIS_r,(6) 1R8S_r,(7) 1IBR_r,(8) 1EER_r,(9) 1FC2_r,(10) 2FD6_l,(11) 2C0L_l,(12) 1FQ1_l,(13) 1JMO_r,(14) 1BKD_l,(15) 1GPW_r,(16) 1I2M_r,(17) 1YVB_l,(18) 2AJF_l,(19) 1NW9_r,(20) 1E4K_r,(21) 1IBR_l,(22) 2CFH_l,(23) 1HIA_l,(24) 2OT3_l,(25) 1KKL_r,(26) 1PXV_r,(27) 1KTZ_r,(28) 1BKD_r,(29) 1EER_l and (30) 1IB1_r. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

entire length of the receptor, which wraps around it. For the actin/BNI1 complex, the ligand is intercalated in a cleft between two domains. For the immunoglobulin G Fc fragment/FcγRIII complex, the receptor opens a region between two domains via hinge bending motion, into which the ligand binds. In all three of these cases, the global motion is functionally related to the binding process.

The observation that in most cases the conformational change at the interface is distinct from global change prompted the question of whether flexibility prediction techniques which have proven successful at predicting global change are also able to predict the extent of local rearrangements. One such method was shown by Dobbins *et al.* (2008). In this method, the conformational change was hypothesised to be driven by thermal motions, and the degree to which available thermal energy can drive a given motion

is related to λ, the frequency of that motion. Frequencies were calculated using normal mode analysis, and the proposed relationship between RMSD and the frequencies is given as

$$\text{RMSD} \propto \sqrt{\frac{1}{N}\sum_{j=1}^{3N-6}\frac{1}{\lambda_j}} \tag{2.15}$$

This method was shown to be a reasonable predictor of the extent of conformational change observed in a large set of known protein-protein interactions. When applied to the extent of conformational change at the interface, a Wilcoxon rank-sum test showed that there is no significant difference between the predicted RMSD and the real RMSD, at significance levels of 0.0197, 0.0197 and 0.0006, when ordered by C_α, backbone and all-atom RMSD respectively, confirming that this method is also capable of estimating the extent of rearrangement at the interface.

2.4 Discussion

The conformational changes which occur upon protein-protein binding have been studied using ENM normal mode analysis, with a focus on its use in the construction of docking algorithms. A large test set was used to assess the ability of the atomistic ENM to capture conformational rearrangements, by quantifying the ability of normal modes to move all atoms, backbone (C, O, N, C_α) and C_α atoms from the unbound to bound conformation. It was confirmed that the differences in the crystal structures of bound and unbound proteins can usually be approximated using a single low-frequency normal mode. However, interfacial changes are the greatest contributor to the free energy of binding; desolvation, electrostatics and Van der Waals effects. Thus, it is the computational reconstruction of these changes which are most significant when designing a docking algorithm, and close attention was paid to the residues which are involved in binding. Global protein motion was shown to be significantly different from localised conformational rearrangement at the interface, except in a minority of cases where global motion is functionally involved in docking. Despite this, the global flexibility predictor developed by Dobbins *et al.* (2008) can still be

used to predict the extent to which the conformation at the interface changes.

Conformational rearrangements of interface residues were found to be associated with higher frequency modes than global changes. Despite hundreds of modes needing consideration to find a single mode which sufficiently captures the unbound-bound transition at the interface, when used in linear combination, much fewer were required. When increasing numbers of modes are used, a greater reduction in RMSD can be made across the interface than across the whole protein, despite higher frequency motions being involved. The applicability of the RTB method was also assessed. Although less accurate at predicting global conformational changes, this technique has a greater ability to model conformational change at the interface. For this reason, and due to its computational efficiency, this method makes a better choice for generating pre-calculated modes for docking.

The ENM is computationally inexpensive, and only needs to be calculated once prior to docking. Further, applying normal motions to proteins is also inexpensive and can easily be incorporated into a guided docking protocol. For almost all cases, this method is capable of modelling some of the conformational changes, and for those where large hinge motion is required to accommodate the ligand, the RMSD to the bound can be reduced by up to 60% with as few as 5 modes. On average, 10 modes allows the RMSD of the backbone to be reduced by 20%, a significant reduction considering the modest increase in computational expense. When 50 modes are included, the RMSD of interfacial α carbons is reduced by 70%, on average.

It is also worth bearing in mind that the complex does not exist as a static structure, but as an ensemble of structures. As protein-protein binding is believed to be predominantly governed by the conformational sorting and population shift mechanism (see section 1.4.4.5), this lends credence to the notion of modelling binding using a pre-calculated representation of conformational diversity. It is not necessary to model exactly the conformation revealed by crystallography or NMR, but moreover find

a structure within the ensemble of bound conformations. Even if this cannot be achieved using a set of low frequency normal modes, the ENM may be capable of alleviating clashes and reveal geometric or electrostatic motifs on the protein surface that steer the complex from the encounter complex towards the final docked pose. Should such a structure have sufficient binding energy to be distinguished from the false positives, then an approximate binding mode will be revealed. Induced fit effects are also thought to be relevant to molecular recognition. As flexible guided docking methods usually allow a protein to adjust its conformation in the presence of the field generated by its binding partner, movement can be induced along higher frequency modes, which are less frequently visited in solution.

Chapter 3

SwarmDock

3.1 Introduction

Elastic network normal mode analysis has shown to reproduce functionally relevant protein dynamics at many levels of resolution, with application ranging from the identification of small molecule binding sites (Tuzmen and Erman, 2011), to the study of molecular leviathans such as viral capsids and the ribosome (Chennubhotla *et al.*, 2005). The analysis in the previous chapter aimed at establishing the potential efficacy of using normal coordinates derived from this method to model conformational changes in a guided protein-protein docking algorithm. Given the encouraging results of this work, such an algorithm was developed and tested, and later entitled 'SwarmDock'. At its inception, no other protein-protein docking protocol based on pre-calculated normal modes had been published, although two methods have since appeared in the literature (May and Zacharias, 2008a; Mashiach *et al.*, 2010).

Aside from the modelling of flexibility, a number of other issues are in need of consideration. One problem with docking methods based on a systematic search is the sheer number of potential solutions which are produced. Although generating tens of thousands of structures increases the chances of finding a near native, it also produces many false positives, and if flexible refinement is needed for the binding energy to stand out from the sea of false structures, then near native structures can easily be removed from the list of poses deemed sufficiently promising for refinement. For instance,

in a comparison of ContextShapes, ZDOCK and PatchDock tested on 84 complexes, a structure with RMSD within 5Å of the bound was discovered for 41, 43 and 38 complexes respectively. However, fewer than 15 of these structures were found in the top 1,000 (Shentu *et al.*, 2008).

For this reason, it was decided that our docking algorithm could benefit from simultaneous refinement and searching, so as not to produce a vast quantity of unrefined false structures capable of swamping out unrefined near-natives. However, flexible refinement is costly, and a compromise between coverage of search space and flexibility was undesirable. As the inclusion of normal modes is a cheap method of perturbing structures into physically plausible regions of conformational space, it made an ideal candidate. However, this approach is still inevitably more computationally expensive than those accelerated using the correlation method or by decomposing the surfaces into convex and concave regions, and an efficient optimisation algorithm is paramount. The best way to achieve this is to exploit the funnel-like energy landscape surrounding the global minimum (see section 1.4.4.5). All the guided docking approaches mentioned in section 1.5.5 work in this fashion. Some of these, such as those based on the steepest descent or quasi-Newton methods, are greedy algorithms; they only lower their energy. These are particularly prone to becoming trapped in local energy minima, a very significant problem in protein-protein docking, as the energy landscapes encountered tend to be highly rugged and multimodal (Ruvinsky and Vakser, 2009, 2008). Others can cross energy barriers. For example, take the popular Monte Carlo simulated annealing method. In this approach, a move is derived from a move set and proposed. If the change in energy after undergoing the proposed move is negative, the move is accepted. Otherwise, the move may be accepted or rejected, with the probability of acceptance being exponentially distributed about the change in energy. Thus, higher energy solutions can be accepted, as long as the increase in energy is not too high. The rate parameter in the exponential distribution can be adjusted - if it is too high, then higher energy regions are too infrequently sampled for efficient barrier crossing. However, if it is too low, then the algorithm spends an inordinate amount of time in high energy regions and optima are not found. A balance must be struck between exploring new regions of search space and exploiting low energy

regions. Without knowing the nature of the energy landscape surround the starting position in advance, the width of minima and the barrier heights between them, it is difficult to tune the algorithm. Almost always with these methods, a multi-start approach is used, such as in the RosettaDock (Wang *et al.*, 2007a), HADDOCK (Dominguez *et al.*, 2003) and ATTRACT (May and Zacharias, 2008b) protocols. This helps alleviate the problems associated with getting trapped in local minima, as long as a sufficient number of starting positions are used. However, computational resources can be wasted, as some runs will inevitably spend much time searching false energy funnels, being unable to lower their energy below that of the minimum, nor be able to escape the barriers that separate it from the next.

The approach which was taken in the SwarmDock protocol uses the Particle Swarm Optimisation algorithm (PSO) (Kennedy and Eberhart, 1995). Based upon the Swarm Intelligence paradigm, it is a self-organising population based metaheuristic which exhibits emergent properties which allow it to efficiently navigate search space and cross energy barriers. It has a number of desirable features which distinguish it from other methods used for protein-protein docking, and is particularly suited to highly multimodal problems, and problems of high dimensionality. Firstly, it is much less susceptible to getting trapped in local minima, as individual members of the swarm, the particles, can get 'dragged' out of basins of low energy if other members of the swarm find more energetically favourable regions elsewhere. The algorithm performs significantly better than many other methods on a wide variety of optimisation problems, the reason for which is partially attributed to this broadcasting ability arising from communication between particles (Engelbrecht, 2006; Eberhart *et al.*, 2001). Secondly, the algorithm automatically samples a large range of search space, dynamically adjusting between exploiting promising low energy regions by intensification of the swarm, and exploring diffuse regions of search space via its diversification. Particles are initially sparsely distributed within the search space, as long as they are not so sparsely distributed that they fail to find, by chance, a point within the rim of the binding energy funnel. During the earliest phases of the algorithm, the search is effectively random. Some particles are high in energy, and others are low. The particles are

programmed to behave such that they tend towards low energy regions they have previously visited. This is their cognitive aspect. Each particle can also communicate with some of the other particles, those which are said to be within its neighbourhood. Particles have a tendency towards the lowest energy region of search space found by the other particles within their neighbourhood. This is their social aspect. As the algorithm proceeds the particles tend towards the lower energy regions of space, forming many small clusters. The energy wells associated with these clusters are explored, some yielding lower energy than others. Particles from the less yielding clusters tend towards those of lower energy, sampling search space as they go, potentially finding new energy basins along the way. As the algorithm proceeds further, the clusters of particles become fewer and more populated, and thus the search is focussed upon the deepest wells. In the particular version of the PSO that is implemented, there is a term specifically aimed at diversification included in the equations governing the propogation of algorithm. If no energy basins consistently stand out as being of low energy, the swarm remains diffuse. Ultimately, the algorithm converges upon the deepest energy funnel discovered during the search, which potentially corresponds to the global minimum.

The PSO algorithm has been successfully applied to hundreds of optimisation problems. Poli (2008) cites over 600 successful applications of the algorithm to a multitude of problems. It has been shown to be significantly more efficient in protein-small molecule docking than simulated annealing Monte Carlo and a genetic algorithm (Namasivayam and Gunther, 2007; Janson *et al.*, 2008; Chen *et al.*, 2007). There are a number of different versions of PSO, and it is frequently combined with a local search as part of a memetic algorithm, as is the case with SwarmDock. Although a comprehensive account of SwarmDock is given here, details have also been published in Moal and Bates (2010) and Li *et al.* (2010b).

3.2 Methods

3.2.1 Search Space

The search space specifies the variables which are optimised in the SwarmDock algorithm. In the search, the position and orientation of the receptor is kept fixed. The algorithm optimises the position and orientation of the ligand and the normal mode space for the receptor and the ligand. The translational dimensions consist of Cartesian coordinates of the ligand centre of mass. The normal mode space for the receptor and the ligand correspond to the coefficients of the modes when they are taken in linear combination (see Chapter 2). The orientation of the ligand is represented using a quaternion. This particular representation was chosen partly due to its ease of implementation, computational efficiency, numerical stability and its ability to avoid the gimbal lock problem, where two axes in Euler angles can become aligned. However, it also is particularly suited to both the PSO algorithm and the local search.

Quaternions are an extension of complex number, and they consist of a 'real' part and three 'imaginary' parts. Take the quaternion z,

$$z = a + bi + cj + dk \tag{3.1}$$

The three 'imaginary' dimensions correspond to an axis around which the rotation takes place, whilst the fourth corresponds to the extent of rotation about that axis. Two inter-convertible ways of doing this are applied. In the first, the b, c and d form a normalised vector representing the axis of rotation in 3D space, and a is the angle of rotation in radians. This is useful for the local search, since if we want to randomly perturb the ligand by approximately 0.1 radians, we can simply alter a by a number of approximately this magnitude. In the second, the whole quaternion is normalised, and each orientation corresponds to a position on the surface of a four dimensional hypersphere. The 'poles' of the hypersphere corresponds to no rotation at all, and orthogonal to the poles lies a unit 3-sphere which corresponds to a rotation of 180° around the axis corresponding to the position on that sphere. This method is particularly suited to the PSO,

as a particle in one region of orientational space can move directly towards a particle in another region of orientational space. As the quaternion is normalised upon each iteration, this approximately corresponds to moving in a straight line across the surface of the hypersphere and represents smooth interpolation between the two positions. If Euler angle space were employed instead, the trajectory of the atoms would depend on the order in which rotations are performed. To eliminate this artefact would be intractable.

To convert from the first quaternion representation to the second simply requires normalisation of z. Conversion from the second to the first is obtained by normalising $< b, c, d >$ and the angle is obtained by $\theta = 2\cos^{-1} a = 2\sin^{-1} \sqrt{b^2 + c^2 + d^2}$. In order to apply the rotation to the ligand, the rotation matrix, $\mathbf{R}$, is obtained from the normalised quaternion as

$$\mathbf{R} = \begin{bmatrix} a^2 + b^2 - c^2 - d^2 & 2bc - 2ad & 2bd + 2ac \\ 2bc + 2ad & a^2 - b^2 + c^2 - d^2 & 2cd - 2ab \\ 2bd - 2ac & 2cd + 2ab & a^2 - b^2 - c^2 + d^2 \end{bmatrix} \tag{3.2}$$

3.2.2 SwarmDock: An Overview

SwarmDock is coded in C++ using current standards (Iso14882, 1998) . Code repetition was kept to a minimum during its development, and the efficient std::vector<T> class was used for all containers. The boost library was used for all random number generation. SwarmDock was developed as a C++ library organised into modules, such as the protein class for constructing proteins and reading pdb files, or the quaternion class which contains methods for manipulating orientation. The swarm class contains the constructor to create an instance of a swarm, as well as the methods related to the propagation of the algorithm. Proteins are organised hierarchically, such that a protein object possesses chains, which in turn possess residues, which possess groups which possess atoms. All the energy functions and algorithms are internal such that with the exception of the normal modes, which must be pre-calculated, SwarmDock is a completely stand-alone program. The clustering algorithm was programmed in Python.

An overview of the SwarmDock algorithm is given in Figure 3.1. Firstly,

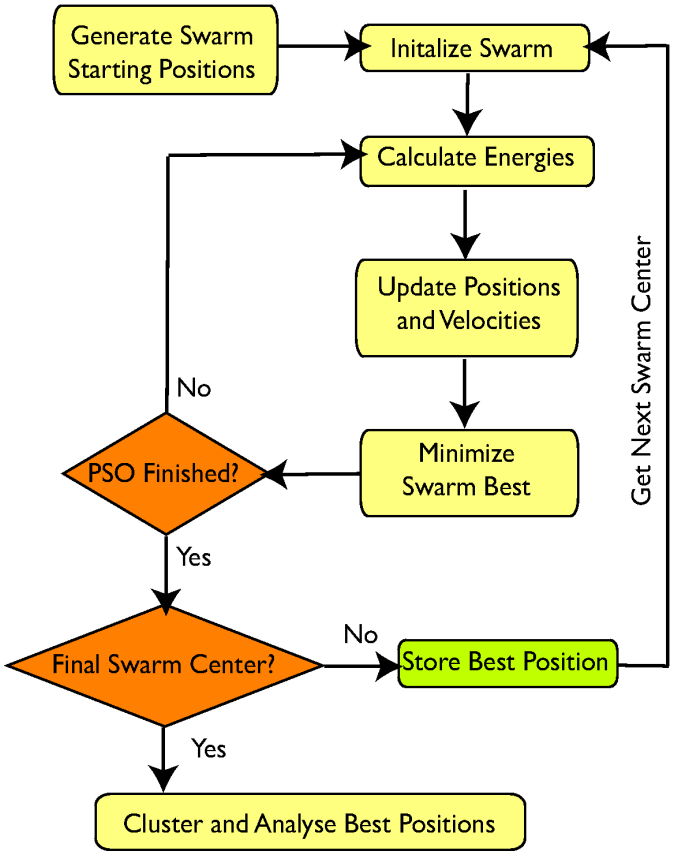

Figure 3.1: An overview of the SwarmDock algorithm. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

starting positions are generated surrounding the receptor. From each of these positions, a swarm of 350 particles is generated as described in section 3.2.3. The orientation of each particle is randomised, and the particles are moved in a random direction from the starting position by an amount drawn from a Gaussian distribution ($\sigma = 10$Å). Normal mode coefficients are also drawn at random from a Gaussian distribution ($\sigma = 3.0$). This range is narrow so that the internal bonded geometry of the complex isn't perturbed beyond that which is physically reasonable. Each particle is also assigned a velocity in search space, all of which are initially set to zero.

Following this, the binding energy of each member of the swarm is calculated using one of the scoring functions outlined in section 3.2.6. The energies are then used to update the velocities of the particles in search space, and the velocities are used to update the particle positions, as described in section 3.2.4. At this point, the lowest energy member of the swarm undergoes a local search, as described in section 3.2.5. The cycle of energy evaluation, velocity and position update, and local search is repeated for 600 iterations. The process is repeated from each of the starting positions. The best energy structure found in each of the runs are amalgamated and clustered as described in section 3.2.7, yielding a list of clusters ordered by energy. A typical run from a single starting position takes around 10 minutes on a single 2.66 GHz CPU.

3.2.3 Initialisation

SwarmDock is run from several positions surrounding the receptor, usually around 120. The strategy for deriving these positions was to approximate the shape of the receptor, and then use this as a basis for generating the final positions.

Specifically, the receptor is first approximated as an ellipsoid. It is initially translated such that its centre of mass coincides with the origin (0,0,0). The moments of the receptor, with atoms i, around any given axis are given by the formula $M = \sum_i m_i r_i$ where m_i is the mass of atom i and r_i is the perpendicular distance from the axis to that atom. If, for every possible line that passes through the centre of mass of a receptor, a point is placed at a distance proportional to $1/M$, then the resulting distribution of points approximates the shape of the receptor. An ellipsoid, $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$, is

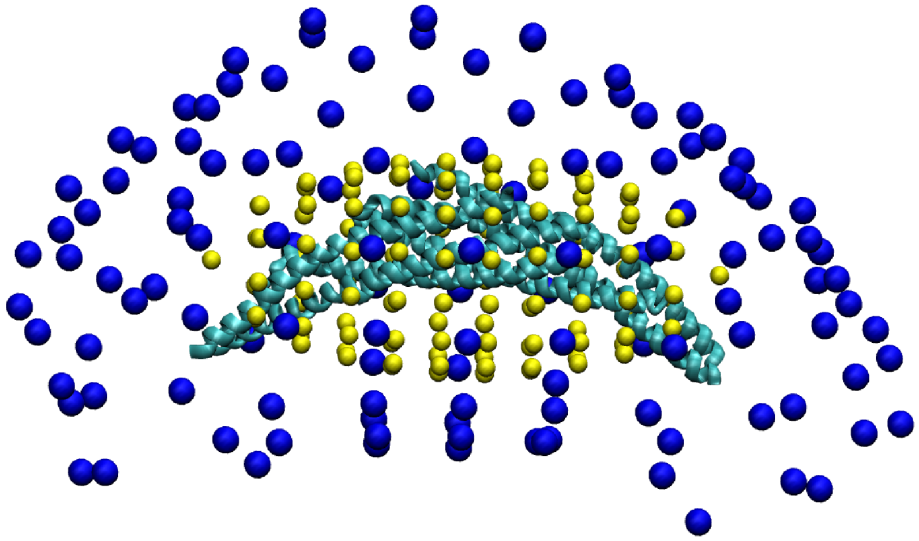

Figure 3.2: This image shows the starting points generated for arfaptin (PDBid 1I49). The points after approximation of the receptor as an ellipsoid are shown in yellow. The final points are shown in blue. This protein is particularly asymmetrical, and thus an even distibution is of particular importance, so as not to bias the algorithm towards any given region of search space.

fit to this shape by setting the equatorial radii, a and b, to $\max(1/M)$ and $\min(1/M)$ respectively and setting the polar radius, c, to $1/M$ around the axis perpendicular to the plane defined by a, b and the centre of mass. The receptor is rotated so that its principal axes, x, y and z are parallel to the semiprincipal axes of the elliposid, (a,0,0), (0,b,0) and (0,0,c). From here, n elliptical cross-sections, $\epsilon_1 \cdots \epsilon_n$ are made in the y-z plane, with x values given by the arithmetic progression $x = a(1-\frac{1}{n}), a(1-\frac{3}{n}), a(1-\frac{5}{n}) \cdots a(1-\frac{2n-1}{n})$. The distance, l, between ellipses along the x axis is given by $l = \frac{2a}{n}$. For all $\epsilon_1 \cdots \epsilon_n$, t points are equally spread along the circumference of the ellipse and $f(\epsilon_j, l) \in \mathbb{Z}$ is a function which returns the value of t for which the arc length between points is closest to l. The value of n is found which gives a value for expression 3.3 which is closest to an arbitrary value of 120.

$$2 + \sum_{j=1}^{n} f(\epsilon_j, l) \tag{3.3}$$

Knowing n, the vectors which describe all the t_j points evenly spaced around all n ellipses, along with the vectors (a,0,0) and ($-a$,0,0), are normalised. Along each of these vectors, a ligand protein is drawn from the

origin, one Ångström at a time, until a negative or zero interaction energy is obtained. It is then drawn a further 15Å away from the receptor, so as to be approximately twice the radius of gyration away from the receptor, for a typical protein. The centres of mass for these ligands then become the starting positions for SwarmDock, and are approximately evenly spaced around the receptor. An example of the distribution of generated starting positions for a particularly prolate protein is shown in Figure 3.2.

3.2.4 Particle Swarm Optimisation

The PSO is a population based algorithm inspired by the behaviour of social animals (Kennedy and Eberhart, 1995). The simplest version of the algorithm is as follows. Individuals (particles) in a population (swarm), are simulated as they navigate search space (the objective function). Each particle, i, has a position vector specifying a position in search space, $\boldsymbol{\chi}_i$, and a velocity vector, $\mathbf{v}_i$

$$\boldsymbol{\chi} = \begin{pmatrix} x \\ y \\ z \\ n_x \\ n_y \\ n_z \\ \theta_n \\ M_{rec,1} \\ M_{rec,2} \\ \vdots \\ M_{rec,r} \\ M_{lig,1} \\ M_{lig,2} \\ \vdots \\ M_{lig,l} \end{pmatrix} \qquad \mathbf{v} = \begin{pmatrix} \Delta_x \\ \Delta_y \\ \Delta_z \\ \Delta_{n_x} \\ \Delta_{n_y} \\ \Delta_{n_z} \\ \Delta_{\theta_n} \\ \Delta_{M_{rec,1}} \\ \Delta_{M_{rec,2}} \\ \vdots \\ \Delta_{M_{rec,r}} \\ \Delta_{M_{lig,1}} \\ \Delta_{M_{lig,2}} \\ \vdots \\ \Delta_{M_{lig,l}} \end{pmatrix} \tag{3.4}$$

The algorithm is iterative, whereby the objective function and two simple transition functions are evaluated for each particle, on each iteration, t. The transition functions, which update the positions and velocities synchron-

ously, are

$$\boldsymbol{\chi}_i(t+1) = \boldsymbol{\chi}_i(t) + \mathbf{v}_i(t+1) \tag{3.5}$$

$$\mathbf{v}_i(t+1) = w\mathbf{v}_i(t) + c_1 r_{1,i}(\mathbf{p}_i(t) - \boldsymbol{\chi}_i(t)) + c_2 r_{2,i}(\mathbf{p}_g(t) - \boldsymbol{\chi}_i(t)) \tag{3.6}$$

where w, the inertial weight, is the degree to which previous velocity contributes to the new velocity (Shi and Eberhart, 1998). The terms $r_{1,i}$ and $r_{2,i}$ are random numbers taken from a uniform distribution between 0 and 1. Parameters c_1 and c_2 are the cognitive and social aspect respectively, both typically set to 2.05 (Clerc, 1999). The cognitive aspect is the propensity for the particle to move toward the best scoring position it has previously experienced. The social aspect is the degree to which the particle moves toward the best position found by any member of the swarm. The position vectors $\mathbf{p}_i$ and $\mathbf{p}_g$ correspond to the best position (i.e. position with lowest energy) found by particle i and the best position found by any particle in the swarm, respectively.

3.2.4.1 Neighbourhoods

A common variant on this algorithm, implemented in SwarmDock, uses the concept of neighbourhood (Eberhart and Kennedy, 1995). When the swarm in initialised, k particles are deemed to be in the neighbourhood of each particle i. These are the particles either side of i in the array of particles. In this scheme, $\mathbf{p}_g$ is replaced with $\mathbf{p}_{g,i}$, the best position found by any particle in the neighbourhood of particle i.

$$\boldsymbol{\chi}_i(t+1) = \boldsymbol{\chi}_i(t) + \mathbf{v}_i(t+1) \tag{3.7}$$

$$\mathbf{v}_i(t+1) = w\mathbf{v}_i(t) + c_1 r_{1,i}(\mathbf{p}_i(t) - \boldsymbol{\chi}_i(t)) + c_2 r_{2,i}(\mathbf{p}_{g,i}(t) - \boldsymbol{\chi}_i(t)) \tag{3.8}$$

In SwarmDock, neighbourhoods are set to wrap around, such that the particle at the end of the particle vector 'see' those at the beginning and

vice versa. Subsequently, the neighbourhood graph forms a ring network topology (Kennedy and Mendes, 2002).

3.2.4.2 Velocity Clamping

To avoid situations in which the swarm explodes, and to control the trade-off between focussing and diversifying the swarm, velocity clamping is often employed to impose a limit on the distance the particles can move in any one iteration. If the maximum velocity is exceeded, it is reduced accordingly

$$v_{ij}(t+1) = \begin{cases} v'_{ij} & \text{if } -V_{j,max} < v'_{ij} < V_{j,max} \\ V_{j,max} & \text{if } v'_{ij} > V_{j,max} \\ -V_{j,max} & \text{if } v'_{ij} < -V_{j,max} \end{cases}$$

where v'_{ij} is the calculated velocity of iteration $t+1$. After parameterisation, V_{max} was set at 5Å for translational parameters, 0.2 rad for the angular term in the quaternion and 0.5Å for the spatial terms (see section 3.3.1). Conformational parameters were found not to reach excessive velocities, and thus no clamping was set on the normal mode coefficients.

3.2.4.3 Variations of PSO

Numerous modifications of the PSO algorithm are available, with various strategies to control the dynamics of the swarm. The following variants were tested in combination with the local search:

BPSO: The basic particle swarm optimisation algorithm, as described above.

CPSO: The constricted PSO. Introduced by Clerc and Kennedy (2002), this PSO variant is designed to prevent explosion of the swarm by using a constriction factor, κ, to dampen motion. This allows faster convergence and velocity clamping is not used. This method uses equation 3.9 in place of equation 3.8.

$$\mathbf{v}_i(t+1) = \kappa(\mathbf{v}_i(t) + c_1 r_{1,i}(\mathbf{p}_i(t) - \boldsymbol{\chi}_i(t)) + c_2 r_{2,i}(\mathbf{p}_{g,i}(t) - \boldsymbol{\chi}_i(t))) \qquad (3.9)$$

where

$$\kappa = \frac{2}{|2 - \psi - \sqrt{\psi^2 - 4}|} \qquad \text{and} \qquad \psi = c_1 + c_2 > 4.0 \tag{3.10}$$

CPSOvc: Constricted PSO with velocity clamping. This method is identical to CPSO but with limits on the velocity, and has shown to have improved performance in some test cases (Eberhart and Shi, 2000).

CPSO2: An alternatice version of the constricted PSO, where the sum of the scalar parts of the last two terms in equation 3.8 is not random. This is reported to give small improvements over CPSO (Clerc and Kennedy, 2002). Velocity clamping was included.

$$\mathbf{v}_i(t+1) = w\mathbf{v}_i(t) + c_1 \frac{r_{1,i}}{r_{1,i} + r_{2,i}}(\mathbf{p}_i(t) - \boldsymbol{\chi}_i(t)) + c_2 \frac{r_{2,i}}{r_{1,i} + r_{2,i}}(\mathbf{p}_{g,i}(t) - \boldsymbol{\chi}_i(t)) \tag{3.11}$$

RPSO: Repulsive PSO. This novel variation of PSO includes a repulsive term. For each particle, a random particle is chosen and a distance dependent repulsive term is included in the velocity update function. This variation was postulated to keep the swarm diffuse.

$$\begin{aligned}\mathbf{v}_i(t+1) = {} & w\mathbf{v}_i(t) + c_1 r_{1,i}(\mathbf{p}_i(t) - \boldsymbol{\chi}_i(t)) + c_2 r_{2,i}(\mathbf{p}_{g,i}(t) - \boldsymbol{\chi}_i(t)) \\ & - c_3 r_{3,i}(\mathbf{p}_{rand}(t) - \boldsymbol{\chi}_i(t))\end{aligned}$$

Ultimately, the RPSO variant was selected for use in SwarmDock, due to its enhanced performance compared to the others (see section 3.3.1). Fuller discussions of these variations are available elsewhere (Van Den Bergh, 2002; Engelbrecht, 2006).

3.2.5 Local Search

SwarmDock is a memetic algorithm which combines PSO with a local search. This search is performed on the lowest energy member of the swarm and is executed after each iteration. It is essentially as that described

by Solis and Wets (1981), and is the same as is used in combination with a genetic algorithm in the AutoDock protein/small-molecule docking package (Morris *et al.*, 1998). It randomly picks a displacement in search space from a probability distribution which is biased towards directions which have previously resulted in a lowering of energy. The magnitudes of the displacements depend upon how successful recent moves have been. Initial test of SwarmDock showed that the inclusion of this local search significantly improved the frequency at which the correct structure was found (data not shown).

For each dimension of search space, j, a deviation is drawn from a Gaussian distribution of standard deviation ρ_j, centred around b_j. These form the vectors $\boldsymbol{\rho}$ and $\mathbf{b}$ respectively. Initially, b is set to zero for all dimensions. For translational and conformational dimensions, the initial ρ is set to 2.5Å and 0.15 respectively. The ρ for the angular part of the quaternion is initially set to 5°, and 0.25Å in the spacial dimensions. The deviates form a vector $\mathbf{d}$, and the energy of the position $\boldsymbol{\chi} + \mathbf{d}$ is evaluated. This is a greedy algorithm, so the move is accepted only if it corresponds to an improvement in energy. Otherwise, the energy is evaluated at the $\boldsymbol{\chi} - \mathbf{d}$ position and this move is accepted if it results in a lowering of the energy. If neither of these positions result in a decrease in energy, a new set of deviations are drawn, after updating the bias vector $\mathbf{b}$. This vector is updated according to equation 3.12, such that the bias is attenuated if both positions fail to lower the energy, and the bias is bolstered in the sampled direction if successful.

$$\mathbf{b}(t+1) = \begin{cases} 0.6\mathbf{b}(t) + 0.4\mathbf{d} & \text{if } E(\boldsymbol{\chi}+\mathbf{d}) < E(\boldsymbol{\chi}) \\ 0.6\mathbf{b}(t) - 0.4\mathbf{d} & \text{if } E(\boldsymbol{\chi}+\mathbf{d}) > E(\boldsymbol{\chi}) \text{ and } E(\boldsymbol{\chi}-\mathbf{d}) < E(\boldsymbol{\chi}) \\ \mathbf{b}(t)/2 & \text{if } E(\boldsymbol{\chi}+\mathbf{d}) > E(\boldsymbol{\chi}) \text{ and } E(\boldsymbol{\chi}-\mathbf{d}) > E(\boldsymbol{\chi}) \end{cases} \tag{3.12}$$

There is also a success counter and a failure counter. Upon each successful move, the success counter is incremented and the failure counter is reset to zero. Upon each unsuccessful move, the success counter is reset to zero and the failure counter is incremented. If the success counter reaches 5, then the magnitude of the jumps is increased by doubling the standard deviations,

ρ. If the failure counter reaches 5, then the jumps are contracted by halving ρ. The algorithm terminates after 5 consecutive contractions.

3.2.6 Energy Function

In total, three forms of the energy function were tested. The first of these includes only the electrostatics and Van der Waals terms (see section 1.4.1.2). The second has an additional desolvation term (Lazaridis and Karplus, 1999b). The third is the DComplex statistical pair potential (Liu *et al.*, 2004).

3.2.6.1 Van der Waals and Electrostatics

The first scoring function implemented was composed of an electrostatic term, modelled using partial charges and the Coulomb equation (eqn. 1.7), and a Van der Waals terms, modelled using the Lennard-Jones equation (eqn. 1.8). It is a pairwise term, summed over atom pairs

$$E_{int} = \sum_{i}^{atoms} \sum_{j}^{atoms} E_{i,j} \tag{3.13}$$

The most time consuming component of the docking algorithm is the evaluation of equation 3.13. To alleviate this bottleneck, methods of reducing computational cost were employed. Firstly, as the energy is dominated by short and medium range interactions, long ranges interactions can be neglected. The simplest method for achieving this is to use a simple cut-off, in which interactions between atom pairs separated by a distance greater than that cut-off are neglected. This method is fine for Van der Waals interactions, as these interactions scales as $1/r^6$ and quickly tend to zero. For the electrostatics, however, this approach gives rise to artefacts. Energy discontinuities give rise to 'jumps' in the energy surface, with the energy artificially low on one side of the discontinuity and high on the other. Thus, in the greedy local search algorithm, moves which should be accepted may be rejected due to an artificial increase in energy. This is most pronounced when the cut-off splits polar groups. Long-range interaction should appear approximately neutral, due to the effects of the opposite charges on the group cancelling out. However, if only one half of the polar

group is considered, energy can be artificially raised or lowered significantly.

To overcome this problem, a switching function was employed with group based cut-offs (Leach, 2001). This switching function uses two cut-off distances, $r_{on} = 7\text{Å}$ and $r_{off} = 9\text{Å}$, and smooths the potential between these points from the full interaction energy below r_{on} to zero above r_{off}. Groups of atoms are defined as they are in the CHARMM19 force field, and are electrically neutral (MacKerell *et al.*, 1998). When calculating the energy between two groups, the degree of smoothing used is determined by $\hat{r}$, the distance between the centres of mass of the groups. Thus, the $E_{i,j}$ term in equation 3.13 becomes

$$E_{i,j} = \begin{cases} \frac{q_i q_j}{\epsilon r_{i,j}} + \sqrt{\epsilon_i \epsilon_j}\left[\left(\frac{R_{m_{i,j}}}{r_{i,j}}\right)^{12} - \left(\frac{R_{m_{i,j}}}{r_{i,j}}\right)^{6}\right] & \text{if } \hat{r}_{i,j} < 7 \\ \left(\frac{(r_{off}-\hat{r}_{i,j})^2(r_{off}+2\hat{r}_{i,j}-3r_{on})}{(r_{off}-r_{on})^3}\right)\left[\frac{q_i q_j}{\epsilon r_{i,j}} + \sqrt{\epsilon_i \epsilon_j}\left[\left(\frac{R_{m_{i,j}}}{r_{i,j}}\right)^{12} - \left(\frac{R_{m_{i,j}}}{r_{i,j}}\right)^{6}\right]\right] & \text{if } 7 < \hat{r}_{i,j} < 9 \\ 0 & \text{if } 9 < \hat{r}_{i,j} \end{cases} \tag{3.14}$$

All parameters are taken from the CHARMM19 force field.

A further method of accelerating the energy evaluation is undertaken by partitioning space into boxes of length r_{off}. Only groups whose centre of mass are in the same box or adjacent boxes have their distance evaluated, as otherwise they will be lie beyond r_{off}. The indices of the box in which a group appears is easily obtained from its centre of mass by dividing its coordinates by r_{off}, and truncating them at the decimal place by converting to integers. Groups are deemed to be in the same or adjacent boxes as long as the difference between any index does not exceed one. Thus effectively, an initial distance filter is applied using very efficient integer operations. Subsequently, the remaining group pair distances are evaluated. The full atomistic pairwise interaction is only calculated between groups whose centre of mass are below r_{off}. The SwarmDock implementation of this scoring function was verified by comparing energies of a number of structures to energies calculated using CHARMM and finding identical values.

3.2.6.2 EEF1 Desolvation

The desolvation effective energy function (EEF1) developed by Lazaridis and Karplus (1999b) was also coded and implemented in SwarmDock. This model is an empirically parametrised method whose functional form is based on statistical mechanics. It is very computationally inexpensive as it uses the idea that the interaction energy between the solvent and the solute can be calculated as a sum of pairwise interactions between the atoms in the solute. The basis for this idea is that each atom has a solvation free energy, and that nearby atoms hide it from the solvent and reduce this energy. Thus, a completely buried atom makes no contribution to the total solvation free energy. This method is fundamentally different from those based on solving or approximating the Poisson-Boltzmann equation. These techniques calculate the effect of placing the solvent around atoms which, by default, are not solvated. In the EEF1 method, each atom is assumed to be fully solvated by default and the effect of solvent exclusion by the other solute atoms is calculated. The functional form of the pairwise desolvation energy between atom i and j is given by

$$\Delta G_{ij} = -\frac{2\Delta G_i}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2}\exp\left(-\left(\frac{r-R_i}{\lambda_i}\right)^2\right)V_j - \frac{2\Delta G_j}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2}\exp\left(-\left(\frac{r-R_i}{\lambda_j}\right)^2\right)V_i \tag{3.15}$$

where λ_i is the thickness of the hydration shell around atom i and r_{ij} is the internuclear separation between i and j. The solvation free energy of atom i is ΔG_i, which was determined empirically by considering the solvation free energy of small molecules in which the same atom type is largely exposed, and extrapolated to approximate the energy of a fully exposed atom. The Van der Waals radius and volume of atom i are R_i and V_i respectively. The rationale behind the form of this equation is given in Lazaridis and Karplus (1999b). All the volume, radii, λ_i and ΔG_i parameters were taken from the CHARMM implementation of this energy function. As the model was empirically parametrised to be compatible with the CHARMM19 force field, it is already correctly weighted with the

electrostatics and Van der Waals terms. The same cutoffs and switching term are used as above.

3.2.6.3 DComplex

The third energy function tested was the DComplex statistical pair potential. This simple interaction potential was originally derived from the crystal structures of individual proteins (Zhou and Zhou, 2002), however it has shown promise when studying protein-protein interactions (Liu *et al.*, 2004). Its derivation is essentially as described in section 1.4.4.4. This potential was re-coded and implemented in SwarmDock using the original DComplex data files. The implementation was validated by obtaining identical energies from SwarmDock and the DComplex binary.

3.2.7 Clustering

After the docking has been completed, the resulting structures are clustered. The clustering algorithm used is more efficient than the algorithm outlined in section 1.5.7, as it does not require the calculation of the whole all-versus-all RMSD matrix. The lowest energy structure is assigned as the first member of the first cluster. The remaining poses are then clustered in ascending order of energy; for each structure, the list of clusters is iterated upon. If the structure is found to be within the cutoff RMSD of the first member of a cluster, then the structure is added to that cluster. If all clusters have been checked and the structure has not been assigned, then it becomes the first member of a new cluster. Clusters are returned in the same order they were constructed. Tests in which the cluster resolution was varied between 2.0 Å and 6.0 Å have shown that the algorithm is robust to the value used at 2.5 Å and above (data not shown). Below 2.5 Å resolution, the cluster corresponding to the correctly docked pose can be split.

3.3 Results

3.3.1 Parameterisation

3.3.1.1 PSO Variant Selection

Table 3.1: The performance of different PSO variants described in section 3.2.4.3, for local bound-bound docking; agglomerated results for 7 test complexes. The number of times a correct solution is found, irrespective of its ranking, is shown. The average neighbourhood size, population size and number of steps required before successfully finding the bound complex is also shown, in order to demonstrate the most successful neighbourhood sizes and population sizes for the variants, as well as how many iterations are required, on average, in order to find the correct solution.

Method	# Hits	av. k/n	av. n	av. steps
BPSO	100	0.36	339	219
CPSO	68	0.23	350	317
CPSOvc	223	0.35	326	246
CPSO2	56	0.19	339	169
RPSO	728	0.48	323	273

The first test of the algorithm was performed to select the PSO variant which performs most favourably. Initial benchmarking with the five variations of the PSO algorithm described in section 3.2.4.3 was done by rigid body docking of seven diverse structures in their bound conformation. These complexes are as follows: antibody/antigen complexes 1MLC and 1E6J, enzyme/inhibitor complexes 2MTA and 1F34, and other complexes 1GCQ, 1I4D and 1H1V. 1MLC, a FAB/Lysozyme complex, has a highly hydrophobic interface whilst 1E6J, FAB/HIV1 capsid protein p24 complex, has significant electrostatic contribution to the binding energy. 2MTA, methylamine dehydrogenase/amicyanin complex is hydrophobic, with no intermolecular hydrogen bonds and one salt bridge and 1F34, pepsin/pepsin inhibitor complex, has a large interface with 13 hydrogen bonds. 1GCQ, GRB2/VAV complex, has a fairly small interface. 1I4D, arfaptin/RAC1-GDP complex is hydrophobic, with only 2 hydrogen bonds. Finally, actin/gelsolin complex 1H1V has a large interface which is very electrostatic in nature.

As this test was performed to establish the behaviour of the swarm, only local docking was undertaken. Ligands were pulled away from the

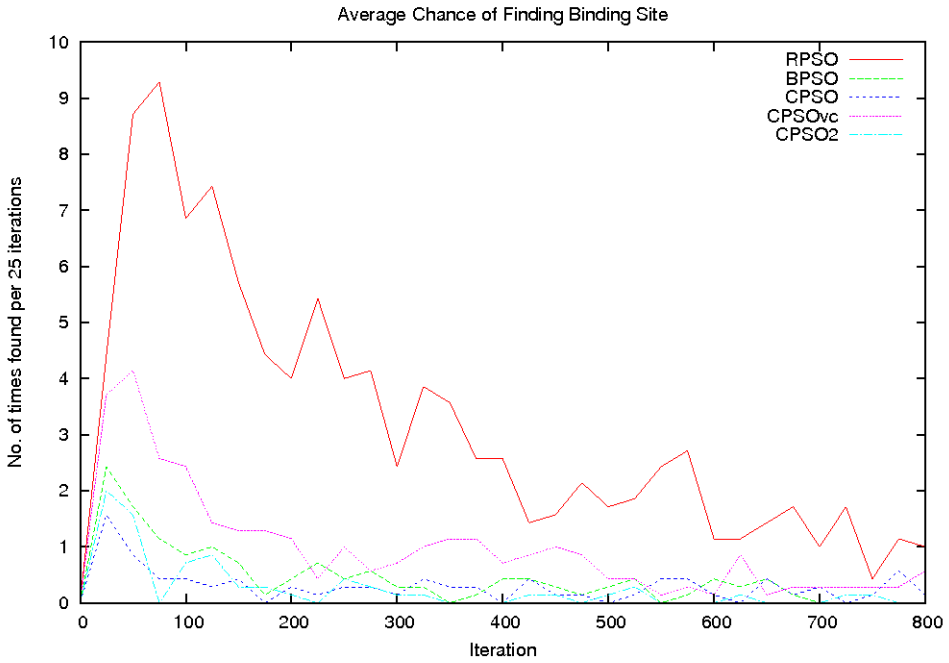

Figure 3.3: Correctly determined poses were placed into bins of 25 iterations for the 5 different PSO variants. The RPSO method is shown to be considerably more likely to find the correct binding site than the other methods. The downward trend after the first hundred or so iterations is due to the swarm focusing on low energy regions away from the binding site and no longer exploring new regions of search space.

receptor by 20Å, moved randomly by a number generated from a Gaussian distribution (σ = 30Å) and randomly oriented. Different swarm sizes, n, and neighbourhoods, k, were tested. For each complex, 25 parameter sets were tested, being all permutations of $n = 100, 200, 300, 400, 500$ and $k = 2, 0.25n, 0.5n, 0.75n, n - 1$. Runs for each parameter set were repeated 10 times. The number of times the binding site was found (RMSD < 1.0Å compared to crystal structure) is shown in Table 3.1, along with the average neighbourhood size, swarm size and number of iterations for successful runs. In Figure 3.3, correct binding site identification events are put into bins of 25 iterations for the various methods.

It is clear from Table 3.1 and Figure 3.3 that the RPSO variant performs significantly better than the others. For this reason, this variant was chosen for further parametrisation and for all subsequent docking runs. Based on the average population size and neighbourhood size of successful runs shown in Table 3.1, the number of particles used per run was set to $n = 350$, with a neighbourhood size of $k = 114$ for all subsequent runs.

It was speculated that the enhanced performance of the RPSO variant compared to the other variants was due to the distance-dependent repulsion term maintaining the diversity of the swarm when no particular region of search space had yet shown to be consistently lower in energy. Mechanistically, contraction of the swarm will reduce the magnitude of the repulsion term and hasten further contraction. The initial collapse will occur when the $\mathbf{p}_i$ and $\mathbf{p}_{g,i}$ for many particles coincide and the particles all head to toward close regions in search space. To test whether this effect prevented premature convergence, the diversity of the swarm was tracked as the algorithm progressed. As a measure of how diffuse the swarm is, the mean Euclidean distance between the particles was calculated every 25 iterations. This is the distance separating centres of mass of the particles averaged over all particle pairs. The results, averaged over all runs, is shown in Figure 3.4A. Upon termination, the centre of mass of particles for successful runs are around 5Å closer to one another than for runs in which the correct binding site was not found, indicating that the repulsion term is behaving as expected.

To look into this effect further, successful runs were separated into 6 groups depending on when the correct binding site was found. The results

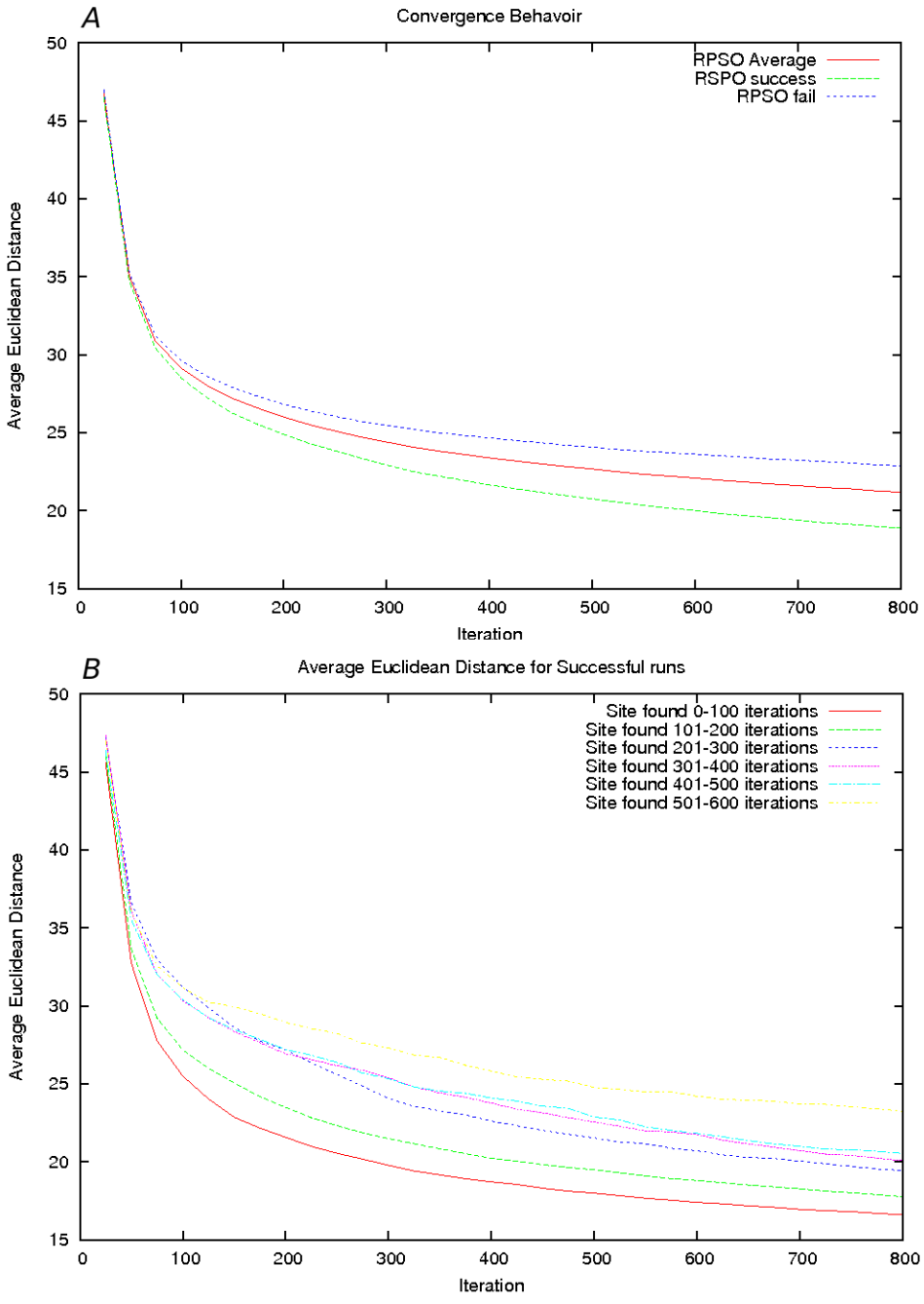

Figure 3.4: The convergence behaviour of the RPSO. (A) The mean Euclidean distance between ligand centres of mass, averaged over all runs, successful runs and runs which failed to correctly locate the binding site. (B) The mean Euclidean distance for successful runs.

of this can be seen in Figure 3.4B. It is clear that the swarm collapses earlier if the binding site is found earlier, an effect which was not observed for any of the other PSO methods tested. For runs in which the correct binding site is found between iteration 101 and 200, the swarm had already started to collapse prior to iteration 100. This indicates that the low energy attractors, $\mathbf{p}_i$ and $\mathbf{p}_{g,i}$, had already flagged the binding funnel as a promising region worthy of focusing the search, in keeping with the above mechanism.

3.3.1.2 Inertial Weight and Velocity Limits

The model contains a number of variable parameters. Finding the best combination of parameters is difficult as we do not know the range to investigate and if we did, an exhaustive scan presents a combinatorial explosion. Rigid-body parameters were tested using mutually orthogonal Latin squares. This method generates parameter sets giving good coverage spanning a large range of values. It vastly reduces the computational expense of parametrisation compared to an exhaustive search, and has been used in a number of applications (Vengadesan and Gautham, 2003; Mandl, 1985; Viji *et al.*, 2009; Arunachalam *et al.*, 2006; Vengadesan *et al.*, 2004; Prasad *et al.*, 2005). As the quaternion representation of orientations consists of three numbers specifying a unit vector and one angle of rotation around that vector, there are two maximum velocities to be optimised; $V_{max,n}$ for the first three parameters and $V_{max,A}$ for the angular parameter. Maximum Euclidean velocity and the inertial weight were also tested. The method was tested on 10 diverse complexes: 1E6J, 1F34, 1GCQ, 1H1V, 1I4D, 1MLC and 2MTA in the bound state and 1I9T, 1NCA and 2QFW in the unbound state.

Latin squares are o order square matrices filled with o symbols such that each symbol appears once in each row and each column. Two Latin squares are orthogonal if each symbol appears in the same element of both squares only once. For example, the following Latin squares are orthogonal, as the symbols only appear in the same position in the leftmost column.

1	3	2
2	1	3
3	2	1

1	2	3
2	3	1
3	1	2

A set of mutually orthogonal Latin squares (MOLS) is a set of Latin squares in which all are orthogonal to one another. A set of 4 mutually orthogonal Latin squares of order 5 were generated using the algorithm outlined by Vengadesan and Gautham (2003). Each Latin square corresponds to a different parameter in the SwarmDock algorithm. Each symbol corresponds to a value of that parameter. When the squares are superimposed, each element has a set of parameters for the SwarmDock algorithm. For all tested complexes, for all elements of the superimposed Latin squares, two docking runs were performed. For all values of all parameters, correctly bound structures were counted, the results of which appear in Table 3.2.

Table 3.2: Parameter efficiency using MOLS. The number of correctly bound structures for each parameter value was summed across all parameter sets containing that parameter value.

Parameter	Value	Count
W	0.4	1020
	0.525	996
	0.65	1022
	0.775	1037
	0.9	967
V_{max}	0.5	890
	1.1	947
	1.7	1019
	2.3	1080
	3.0	1106

Parameter	Value	Count
$V_{max,n}$	0.2	972
	0.325	1014
	0.45	990
	0.575	1024
	0.7	1042
$V_{max,A}$	0.05	956
	0.1	1007
	0.15	1056
	0.2	987
	0.25	1036

As the only clear trend is in V_{max}, another run was completed with a larger range, as is shown in Table 3.3. This shows that this trend stops at around V_{max}=2.5. As the variation between the second set of runs is not significant, the following values were chosen arbitrarily from within the acceptable range for all subsequent calculations: W=0.8, V_{max} = 5.0Å, $V_{max,n}$ = 0.5Å, $V_{max,a}$ = 0.2 rad.

Table 3.3: Parameter efficiency using MOLS. The number of correctly bound structures for each parameter value was summed across all parameter sets containing that parameter value.

Parameter	Value	Count
W	0.4	1159
	0.525	1187
	0.65	1154
	0.775	1147
	0.9	1172
V_{max}	2.5	1159
	3.2	1187
	3.9	1151
	4.6	1136
	5.3	1152

Parameter	Value	Count
$V_{max,n}$	0.2	1155
	0.35	1134
	0.5	1190
	0.65	1187
	0.8	1153
$V_{max,A}$	0.1	1155
	0.2	1134
	0.3	1190
	0.4	1187
	0.5	1153

3.3.2 Bound-Bound Benchmark v2.0

The performance of rigid-body bound-bound docking of all 84 complexes in the Benchmark v2.0 (Mintseris *et al.*, 2005), using the electrostatics and Van der Waals energy function and clustering at 3.5Å resolution, is shown in Table 3.4. These runs are not intended to represent a biologically useful docking scenario, but are merely a preliminary test of the energy function and the algorithm. Starting positions were generated as described in section 3.2.3, and the algorithm was run twice from each point.

Table 3.4: The performance of bound-bound docking.

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1A2K	1	16	yes	0.13	0.44	0.98	0.00	0.05(1/16)	0.05(1/16)	1.00(1/16)	0.00(1/16)	1/1/1
1ACB	1	2	no	0.19	0.58	0.97	0.02	0.19(1/2)	0.19(1/2)	0.97(1/2)	0.00(1/2)	1/1/1
1AHW	1	23	yes	0.24	1.21	0.92	0.03	0.17(1/23)	0.17(1/23)	0.92(1/23)	0.00(1/23)	1/1/1
1AK4	1	5	no	0.12	0.34	1.00	0.02	0.12(1/5)	0.12(1/5)	1.00(1/5)	0.02(1/5)	1/1/1
1AKJ	1	6	no	0.04	0.07	0.97	0.02	0.04(1/6)	0.04(1/6)	0.97(1/6)	0.00(1/6)	1/1/1
1ATN	1	2	no	0.21	0.85	0.97	0.00	0.21(1/2)	0.21(1/2)	0.97(1/2)	0.00(1/2)	1/1/1
1AVX	1	19	yes	0.31	1.44	0.91	0.03	0.12(1/19)	0.12(1/19)	0.97(1/19)	0.02(1/19)	1/1/1
1AY7	1	54	yes	0.31	1.35	0.97	0.05	0.11(1/54)	0.11(1/54)	1.00(1/54)	0.00(1/54)	1/1/1
1B6C	1	2	no	0.18	0.46	0.93	0.00	0.18(1/2)	0.18(1/2)	0.93(1/2)	0.00(1/2)	1/1/1
1BGX	17	1	no	6.38	18.82	0.05	0.72	6.38(17/1)	10.50(2/1)	0.05(17/1)	0.72(17/1)	-/-/-
1BJ1	1	3	no	0.23	0.72	0.94	0.03	0.07(1/3)	0.07(1/3)	0.97(1/3)	0.00(1/3)	1/1/1
1BUH	1	55	yes	0.17	0.48	0.96	0.00	0.04(1/55)	0.04(1/55)	1.00(1/55)	0.00(1/55)	1/1/1
1BVK	1	38	yes	0.14	0.27	0.98	0.00	0.06(1/38)	0.06(1/38)	0.98(1/38)	0.00(1/38)	1/1/1
1BVN	1	14	yes	0.23	0.52	0.90	0.00	0.14(1/14)	0.14(1/14)	0.95(1/14)	0.00(1/14)	1/1/1
1CGI	1	3	no	0.20	0.67	0.93	0.01	0.19(1/3)	0.19(1/3)	0.94(1/3)	0.01(1/3)	1/1/1
1D6R	1	27	yes	0.20	0.50	0.98	0.03	0.11(1/27)	0.11(1/27)	1.00(1/27)	0.00(1/27)	1/1/1
1DE4	1	1	no	0.30	1.25	0.91	0.02	0.30(1/1)	0.30(1/1)	0.91(1/1)	0.02(1/1)	1/1/1
1DFJ	1	1	no	0.19	0.66	0.93	0.01	0.19(1/1)	0.19(1/1)	0.93(1/1)	0.01(1/1)	1/1/1
1DQJ	1	29	yes	0.26	0.63	0.93	0.03	0.15(1/29)	0.15(1/29)	0.99(1/29)	0.00(1/29)	1/1/1
1E6E	54	1	no	2.91	8.47	0.30	0.39	2.91(54/1)	2.91(54/1)	0.30(54/1)	0.39(54/1)	-/-/54
1E6J	1	13	no	0.03	0.09	0.98	0.00	0.03(1/13)	0.03(1/13)	1.00(1/13)	0.00(1/13)	1/1/1
1E96	1	29	yes	0.10	0.23	0.93	0.00	0.06(1/29)	0.06(1/29)	0.95(1/29)	0.00(1/29)	1/1/1

Table 3.4 continued on next page

Table 3.4 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1EAW	1	6	no	0.17	0.44	0.96	0.03	0.14(1/6)	0.14(1/6)	0.96(1/6)	0.03(1/6)	1/1/1
1EER	1	1	no	0.10	0.20	0.98	0.01	0.10(1/1)	0.10(1/1)	0.98(1/1)	0.01(1/1)	1/1/1
1EWY	76	12	yes	0.73	1.80	0.72	0.08	0.67(76/12)	0.67(76/12)	0.72(76/12)	0.06(76/12)	76/76/76
1EZU	1	3	yes	0.04	0.09	1.00	0.01	0.04(1/3)	0.04(1/3)	1.00(1/3)	0.01(1/3)	1/1/1
1F34	1	20	yes	0.16	0.43	0.97	0.00	0.04(1/20)	0.04(1/20)	1.00(1/20)	0.00(1/20)	1/1/1
1F51	1	9	no	0.19	0.38	1.00	0.03	0.13(1/9)	0.13(1/9)	1.00(1/9)	0.00(1/9)	1/1/1
1FAK	121	1	no	8.01	17.78	0.00	1.00	8.01(121/1)	16.81(7/2)	0.04(122/1)	0.88(122/1)	-/-/-
1FC2	2	18	yes	0.10	0.52	1.00	0.02	0.08(2/18)	0.08(2/18)	1.00(2/18)	0.00(2/18)	2/2/2
1FQ1	1	23	yes	0.39	2.35	0.83	0.00	0.22(1/23)	0.22(1/23)	0.87(1/23)	0.00(1/23)	1/1/1
1FQJ	1	23	yes	0.25	0.81	0.97	0.02	0.13(1/23)	0.13(1/23)	0.98(1/23)	0.00(1/23)	1/1/1
1FSK	1	15	yes	0.25	1.04	0.94	0.02	0.07(1/15)	0.07(1/15)	0.97(1/15)	0.00(1/15)	1/1/1
1GCQ	1	25	yes	0.09	0.23	0.94	0.02	0.08(1/25)	0.08(1/25)	0.98(1/25)	0.02(1/25)	1/1/1
1GHQ	132	5	no	0.34	1.32	1.00	0.00	0.11(132/5)	0.11(132/5)	1.00(132/5)	0.00(132/5)	132/132/86
1GP2	1	12	yes	0.15	0.41	1.00	0.03	0.10(1/12)	0.10(1/12)	1.00(1/12)	0.01(1/12)	1/1/1
1GRN	1	9	yes	0.10	0.26	1.00	0.01	0.06(1/9)	0.06(1/9)	1.00(1/9)	0.00(1/9)	1/1/1
1H1V	1	12	yes	0.10	0.21	0.93	0.00	0.07(1/12)	0.07(1/12)	0.97(1/12)	0.00(1/12)	1/1/1
1HE1	1	21	yes	0.08	0.20	0.99	0.03	0.06(1/21)	0.06(1/21)	1.00(1/21)	0.00(1/21)	1/1/1
1HE8	1	18	yes	0.53	1.12	0.89	0.06	0.06(1/18)	0.06(1/18)	0.95(1/18)	0.00(1/18)	1/1/1
1HIA	1	24	yes	0.16	0.47	0.94	0.03	0.05(1/24)	0.05(1/24)	0.98(1/24)	0.00(1/24)	1/1/1
1I2M	1	31	yes	0.11	0.28	0.97	0.01	0.08(1/31)	0.08(1/31)	0.98(1/31)	0.00(1/31)	1/1/1
1I4D	1	25	yes	0.18	0.33	0.95	0.00	0.14(1/25)	0.14(1/25)	0.96(1/25)	0.00(1/25)	1/1/1
1I9R	1	30	yes	0.26	1.47	0.91	0.06	0.12(1/30)	0.12(1/30)	0.95(1/30)	0.00(1/30)	1/1/1
1IB1	1	3	no	0.06	0.14	1.00	0.00	0.06(1/3)	0.06(1/3)	1.00(1/3)	0.00(1/3)	1/1/1
1IBR	1	1	no	0.21	0.45	0.92	0.00	0.21(1/1)	0.21(1/1)	0.92(1/1)	0.00(1/1)	1/1/1

Table 3.4 continued on next page

Table 3.4 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1IJK	1	26	yes	0.30	0.99	0.98	0.00	0.12(1/26)	0.12(1/26)	1.00(1/26)	0.00(1/26)	1/1/1
1IQD	1	5	no	0.16	0.64	0.95	0.01	0.16(1/5)	0.16(1/5)	0.95(1/5)	0.00(1/5)	1/1/1
1JPS	1	22	yes	0.16	0.61	0.97	0.01	0.09(1/22)	0.09(1/22)	0.99(1/22)	0.00(1/22)	1/1/1
1K4C	1	40	yes	0.05	0.13	1.00	0.00	0.03(1/40)	0.03(1/40)	1.00(1/40)	0.00(1/40)	1/1/1
1K5D	1	9	no	0.13	0.33	1.00	0.02	0.11(1/9)	0.11(1/9)	1.00(1/9)	0.01(1/9)	1/1/1
1KAC	1	61	yes	0.24	0.99	0.98	0.04	0.05(1/61)	0.05(1/61)	1.00(1/61)	0.00(1/61)	1/1/1
1KKL	13	20	yes	0.14	0.27	0.95	0.00	0.07(13/20)	0.07(13/20)	0.97(13/20)	0.00(13/20)	13/13/13
1KLU	1	23	yes	0.11	0.66	1.00	0.02	0.04(1/23)	0.04(1/23)	1.00(1/23)	0.00(1/23)	1/1/1
1KTZ	1	37	yes	0.09	0.31	1.00	0.06	0.04(1/37)	0.04(1/37)	1.00(1/37)	0.00(1/37)	1/1/1
1KXP	1	3	no	0.10	0.29	0.97	0.02	0.10(1/3)	0.10(1/3)	0.97(1/3)	0.01(1/3)	1/1/1
1KXQ	8	4	no	1.41	4.14	0.58	0.16	1.16(8/4)	1.16(8/4)	0.66(8/4)	0.08(8/4)	-/8/8
1M10	1	29	yes	0.20	0.79	0.93	0.02	0.17(1/29)	0.17(1/29)	0.96(1/29)	0.00(1/29)	1/1/1
1MAH	1	13	yes	0.17	0.43	0.93	0.00	0.16(1/13)	0.16(1/13)	0.94(1/13)	0.00(1/13)	1/1/1
1ML0	1	35	yes	0.04	0.08	1.00	0.03	0.03(1/35)	0.03(1/35)	1.00(1/35)	0.00(1/35)	1/1/1
1MLC	1	25	yes	0.22	0.96	0.95	0.02	0.13(1/25)	0.13(1/25)	0.98(1/25)	0.00(1/25)	1/1/1
1N2C	1	19	yes	0.09	0.23	0.96	0.01	0.07(1/19)	0.07(1/19)	0.98(1/19)	0.00(1/19)	1/1/1
1NCA	1	24	yes	0.18	0.48	0.93	0.03	0.09(1/24)	0.09(1/24)	0.95(1/24)	0.00(1/24)	1/1/1
1NSN	3	13	yes	0.31	0.81	0.82	0.00	0.26(3/13)	0.26(3/13)	0.92(3/13)	0.00(3/13)	3/3/3
1PPE	1	18	yes	0.09	0.25	0.97	0.00	0.04(1/18)	0.04(1/18)	0.99(1/18)	0.00(1/18)	1/1/1
1QA9	4	47	yes	0.49	1.06	0.87	0.08	0.32(4/47)	0.32(4/47)	0.89(4/47)	0.00(4/47)	4/4/4
1QFW	1	17	yes	0.32	1.27	0.91	0.06	0.09(1/17)	0.09(1/17)	0.96(1/17)	0.00(1/17)	1/1/1
1RLB	12	6	no	0.17	0.45	0.93	0.00	0.16(12/6)	0.16(12/6)	0.93(12/6)	0.00(12/6)	12/12/12
1SBB	20	18	yes	0.57	1.77	0.82	0.27	0.12(20/18)	0.12(20/18)	1.00(20/18)	0.03(20/18)	20/20/20
1TMQ	1	2	no	0.07	0.16	0.99	0.00	0.04(1/2)	0.04(1/2)	0.99(1/2)	0.00(1/2)	1/1/1

Table 3.4 continued on next page

Table 3.4 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1UDI	1	29	yes	0.19	0.46	0.99	0.00	0.08(1/29)	0.08(1/29)	1.00(1/29)	0.00(1/29)	1/1/1
1VFB	1	46	yes	0.19	0.37	0.96	0.04	0.09(1/46)	0.09(1/46)	0.98(1/46)	0.00(1/46)	1/1/1
1WEJ	1	51	yes	0.15	0.56	0.98	0.02	0.10(1/51)	0.10(1/51)	1.00(1/51)	0.00(1/51)	1/1/1
1WQ1	1	6	no	0.12	0.33	0.95	0.01	0.12(1/6)	0.12(1/6)	0.96(1/6)	0.01(1/6)	1/1/1
2BTF	1	34	yes	0.10	0.32	0.96	0.01	0.06(1/34)	0.06(1/34)	0.99(1/34)	0.00(1/34)	1/1/1
2HMI	102	1	no	5.50	23.43	0.32	0.65	5.50(102/1)	15.95(1/1)	0.32(102/1)	0.65(102/1)	-/-/-
2JEL	1	14	no	0.06	0.13	0.98	0.04	0.03(1/14)	0.03(1/14)	1.00(1/14)	0.00(1/14)	1/1/1
2MTA	19	27	yes	0.20	0.78	1.00	0.08	0.08(19/27)	0.08(19/27)	1.00(19/27)	0.00(19/27)	19/19/19
2PCC	135	8	no	0.35	0.59	0.97	0.12	0.33(135/8)	0.33(135/8)	1.00(135/8)	0.10(135/8)	135/135/135
2QFW	1	30	yes	0.36	2.10	0.94	0.04	0.26(1/30)	0.26(1/30)	0.96(1/30)	0.00(1/30)	1/1/1
2SIC	1	3	no	0.16	0.38	1.00	0.04	0.16(1/3)	0.16(1/3)	1.00(1/3)	0.04(1/3)	1/1/1
2SNI	1	4	no	0.12	0.26	0.97	0.01	0.12(1/4)	0.12(1/4)	0.97(1/4)	0.01(1/4)	1/1/1
2VIS	1	6	no	0.18	0.63	1.00	0.02	0.08(1/6)	0.08(1/6)	1.00(1/6)	0.00(1/6)	1/1/1
7CEI	1	33	yes	0.13	0.40	0.96	0.00	0.09(1/33)	0.09(1/33)	0.98(1/33)	0.00(1/33)	1/1/1

The performance of bound-bound docking using the electrostatics and Van der Waals energy function, clustering at 3.5Å resolution. All structures in the benchmark v2.0 are docked (Mintseris *et al.*, 2005). A number of performance metrics are evaluated. For the lowest IRMSD first cluster member, the cluster rank (Rank), the cluster size (# found), interface RMSD (IRMSD), ligand RMSD (LRMSD), fraction of native contacts (FNat), fraction of non-native contacts (FNNat), and whether this is the largest cluster (bCluster) are shown. Also, the best IRMSD (bIRMSD), LRMSD (bLRMSD), FNat (bFNat) and FNNat (bFNNat) found, with the their corresponding cluster rank and size in parentheses, are reported, as is the best ranked high, medium and acceptable solution (bRank). All metrics are evaluated per the CAPRI standards (see section 1.5.10 or Mendez *et al.* (2003)).

It is clear that most structures are bound with high accuracy. For all but three of the complexes, an acceptable structure was found, according to the CAPRI criteria (Mendez *et al.*, 2003). All but five were found with a ligand RMSD below 1.0Å. Most complexes, 69/84, were ranked first and 73/84 were in one of the top 10 clusters. Even of the 11 structures which didn't rank in the top 10, five of these were amongst the 52 cases where the largest cluster contained the lowest IRMSD. SwarmDock performed very well by all metrics. However, for a number of cases, the correct structure was found infrequently. Considering the complexes 1DE4, 1DFJ, 1EER and 1IBR, for instance, the most accurate cluster only had a population of one, even though this cluster still ranked first. A number of cases did perform particularly badly. These were investigated and for most of these the reasons for the failure of the algorithm is evident. These are the exceptions which prove the rule; bound-bound docking is a solved problem.

The first is 1BGX, the complex between Taq polymerase and a murine antibody. This is a highly suspicious structure, which is likely fake and should be removed from the protein databank and the docking benchmark. It was deposited by H. M. Krishna Murthy, a known perpetrator of scientific fraud who has forged numerous crystal structures (Borrell, 2009). While some of his suspect papers have been forcibly retracted (Retraction, 2010), others remain. This complex was published just before the time period investigated by the University of Alabama at Birmingham, and is thus not one of the 12 structures explicitly demonstrated to be forged. However, it is likely to be fake, as its WHAT IF check report shows that it has an anomylous Ramachandran plot, severly atypical B-factors, unusual chirality deviations, bond lengths, valence angles, torsion angles, ϕ angles, ψ angles, ω angles and χ angles, untenable unit cell dimensions, unusual proline puckering phases and amplitudes, Van der Waals clashes, abnormal residue packing, water molecules without hydrogen bonds and many unsatisfied buried hydrogen bond donors and acceptors.

The complex between adrenodoxin reductase and adrenodoxin (PDBid 1E6E) should, also, arguably not be present in the docking benchmark. The complex contains a covalent crosslink between LYS27 of adrenodoxin and the ASP39 residue of adrenodoxin reductase. In light of this, it is of little

surprise that only a single low RMSD structure was found, and that it did not rank highly; covalent bonds are shorter that non-covalent interactions, and thus the true binding pose would be deemed to have a serious Van der Waals clash using the current force field.

The Anabaena Ferredoxin I/Ferredoxin-NADP reductase complex (PDBid 1EWY) is also easily explained. This complex contains an Fe_2S_2 iron cluster and a flavin-adenine dinucleotide at the binding interface, which were ignored during the docking protocol. Nevertheless, although it doesn't rank first, the correct structure still has the largest cluster size, being found 12 times. The next biggest clusters, ranked 1 and 17, both have only 6 members, so the true structure still stands out as being found more frequently by the swarm, even if it doesn't stand out as being the most energetically favourable.

The complex between soluble tissue factor and blood coagulation factor VIIA (PDBid 1FAK) is not so easily explained, although the factor VIIA wraps significantly around the tissue factor, which may proclude association. Irrespectively, no reasonable structures were found by SwarmDock.

The complex between complement C3 and the Epstein-Barr virus CR2 receptor (PDBid 1GHQ) is highly suspicious. Recent mutagenesis studies add to a mounting quantity of evidence which suggest that the interface in this crystal structure is not biologically significant. In the words of Isenman *et al.* (2010), "The results with CR2 confirm our earlier mapping studies and cast even further doubt on the physiologic relevance of the complex visualized in the C3d:CR2 cocrystal". Nevertheless, a highly accurate structure was found five times, the lowest energy of which contained all the native contacts and no non-native contacts. This structure, however, did not rank first.

It is unknown why the interaction between HIV-1 reverse transcripase and a murine antibody (PDBid 2HMI) failed to dock. However, it is possible that the interface is not amenable to detection using the current scoring function, due to the domination of desolvation effects driving binding. Interestingly, this is one of the few complexes which performed better when the EEF1 desolvation term is included in the energy function (see section 3.3.3).

Similarly, it is unknown why the yeast cytochrome C and its peroxidase

(PDBid 2PCC) fail to dock correctly. It may be related to the fact that this is only a weak interaction of micromolar affinity, and is driven predominantly by entropic forces (Pielak and Wang, 2001). A very close non-cognate cytochrome C/peroxidase complex, similarly driven by entropy, has been shown to be highly sensitive to ionic strength (Erman *et al.*, 1997). It could well be that the protonation states of interfacial histidines are important for binding, a factor which is not considered during the assignment of atomic charges. Nevertheless, the cluster corresponding to the bound structure still has a very reasonable size of 8, the third largest cluster found.

Despite the three examples for which there is no immediate explaination for failure, the algorithm performs very well. Whilst bound-bound docking is of little scientific value, it acts as a good proof of principle and suggests that the algorithm and scoring function may serve as a suitable starting point for tackling the more difficult unbound-unbound problem.

3.3.3 The EEF1 Desolvation Term

The EEF1 desolvation model was included in the first design of the SwarmDock algorithm. In an early investigation, SwarmDock was run with and without this term in the energy function. As this was an early test, undertaken before flexibility or clustering were included in the protocol, structures were docked as rigid bodies. Their bound conformations were used. As such, the runs are also not intended to represent a biologically useful docking scenario, but merely as a preliminary test of the energy function and the algorithm. A subset of the docking benchmark 2.0 was used, containing 68 complexes (Mintseris *et al.*, 2005). Points were generated as described in section 3.2.3, and the algorithm was run twice from each point. The structures were ranked by energy without clustering. Table A.1, which appears in the appendix, shows the lowest ligand RMSD after superimposing the bound receptors, the best ranked structure with a ligand RMSD below 5Å, and the number of times a structure was found with ligand RMSD below 5Å.

For 40 of the structures, the correct structure ranked first both with and without the EEF1 term. However, the Van der Waals and electrostatics only runs produced better ranking models for 23 cases, as opposed to 4 which

rank better with the EEF1 term. The ligand RMSD was lower for 61 of the cases when the desolvation term was omitted, as opposed to only 6 when it is wasn't. In addition, for most cases, the desolvation term resulted in the bound structure being found less frequently. There are some cases where the EEF1 term did improve the results. For the complex between cyclophilin A and the HIV-1 capsid (PDBid 1AK4), the correct binding mode was found twice as often and the ligand RMSD was significantly reduced. Similarly, for the actin/Dnase I complex (PDBid 1ATN), the correct structure was found 5 times more frequently, and the ligand RMSD dropped by more than half. For the Methylamine dehydrogenase/Amicyanin complex, the complex was found more frequently and its rank improved from 27 to 1. The EEF1 term also greatly increased the number of correctly determined structures for complexes 1E6J, 1E96, 1I4D and 1UDI. Despite these cases, the algorithm does perform significantly better when this term is discarded. The EEF1 terms were similarly deleterious for rigid-body unbound-unbound docking (data not shown).

The reason why the EEF1 term is detrimental to the algorithm is not known. However, there are a number of problems associated with the function, amongst them those discussed by Lazaridis and Karplus (1999b) themselves. The suitability of a pairwise summation in modelling desolvation is by no means assured. However, although the EEF1 model does fail to distinguish near native poses from decoys for some small molecule/protein interactions (Seok *et al.*, 2003), the extent to which this term impairs SwarmDock is difficult to reconcile with the fact that it can help in ranking small molecule complexes (Davis and Baker, 2009), as well as in distinguishing near native conformations from false positive conformations in protein folding (Lazaridis and Karplus, 1999a) and in reproducing protein dynamics (Krol, 2003; Lazaridis and Karplus, 1999b). Interestingly, when used for flexible refinement of docked protein-protein interactions, it does not perform as well as using a simple distance dependent dielectric constant in the Coulombic term, a very crude approximation of solvation effects (Krol *et al.*, 2007b). Similarly, a modified version of the function performed worse at refining docked poses than an alternative desolvation term when tested by Solernou and Fernandez-Recio (2011). This, along

with the results presented here, suggest that whilst the EEF1 model may be useful in modelling the folding and dynamics of single proteins, it is not fit for modelling protein-protein interactions. The EEF1 term was omitted from all subsequent simulations.

3.3.4 Unbound-Unbound Benchmark v2.0

The SwarmDock algorithm was applied to all 84 complexes in the docking benchmark v2.0 (Mintseris *et al.*, 2005), using the unbound structures. This was done both rigidly and including normal mode flexibility. For the flexible docking the five lowest frequency modes, as calculated using the RTB method, were used for both the receptor and the ligand. For both sets of runs, the algorithm was run twice from each starting position. The tables of results, B.1 and C.1 appear in the appendix. As expected, the success rate was considerably lower than for the bound-bound problem. Judging by the CAPRI criteria (see section 1.5.10 or Mendez *et al.* (2003)), rigid-body docking generated an acceptable structure for 35 cases, medium quality structures for 16 cases and high quality structures for 6 cases. Including flexibility improved the results, as acceptable structures were generated for 47 cases, medium quality structures for 21 cases and high quality structures for 5. A summary of all the complexes for which a structure was found with an interface RMSD below 5Å is given in Table 3.5, showing the best cluster size, cluster rank and interface RMSD for each complex. By this criterion, 63 complexes (75% of the data set) bound in either the rigid or the flexible runs, or both. A summary of the results for the flexible method is shown in Table 3.6. As expected, the success rate drops markedly as the extent of conformational changes upon binding increases, with no high accuracy solutions found for the complexes of medium flexibility, and no high or medium accuracy solutions found for the highly flexible complexes.

Table 3.5: Flexible vs. Rigid-Body Unbound-Unbound docking. The best performance for each metric is highlighted in bold.

	Flexible			Rigid-Body		
Complex	RMSD	Rank	Clus	RMSD	Rank	Clus
1NCA	0.391	1	6	**0.356**	1	**28**

Table 3.5 continued on next page

Table 3.5 continued from previous page

	Flexible			Rigid-Body		
Complex	RMSD	Rank	Clus	RMSD	Rank	Clus
1KTZ	**0.660**	**106**	**6**	-	-	-
1QFW	0.821	5	4	**0.793**	5	**20**
1AY7	**0.872**	**19**	6	0.935	21	**16**
1QA9	1.158	**155**	**3**	**0.996**	164	2
2QFW	**1.272**	143	1	1.288	**104**	**3**
7CEI	**1.313**	11	10	1.625	**5**	**22**
1GCQ	**1.399**	**169**	**1**	-	-	-
1VFB	1.403	117	1	**1.295**	**93**	**5**
1FSK	**1.437**	**65**	1	3.728	72	**2**
1HE8	1.448	**46**	1	**1.042**	90	**4**
1I9R	1.483	1	10	**1.445**	1	**34**
1ML0	**1.553**	**5**	1	2.274	79	**2**
1JPS	**1.567**	23	**4**	2.066	23	1
1GRN	**1.578**	**16**	7	1.590	42	**26**
1AVX	1.591	**25**	1	**1.565**	60	**4**
1E6E	**1.745**	**94**	2	1.781	127	2
1WEJ	1.785	**6**	1	**1.349**	9	**4**
1EWY	1.845	**13**	3	**1.838**	77	**8**
1EAW	**1.866**	**1**	**1**	-	-	-
1BUH	1.921	**59**	**6**	**1.896**	124	3
1KAC	2.238	**1**	9	**2.194**	3	**20**
1FQJ	**2.264**	**11**	2	2.341	27	**10**
1PPE	**2.467**	**62**	1	3.687	112	1
1NSN	**2.523**	**20**	2	2.766	45	2
1TMQ	**2.589**	**72**	**3**	3.185	81	1
1HIA	**2.606**	**2**	**1**	-	-	-
1WEJ	**2.750**	116	1	4.851	**88**	**3**
2SNI	**2.801**	178	1	4.487	**47**	1
1WQ1	2.925	**14**	1	**2.787**	30	**6**
1BJ1	**2.990**	**183**	**1**	-	-	-
1KXQ	**3.058**	**97**	**1**	-	-	-
1E96	**3.099**	174	**2**	4.562	**170**	1
1K5D	**3.141**	**30**	**1**	-	-	-
1F34	3.259	**15**	**2**	**2.892**	138	1
1KXP	**3.277**	**122**	1	4.966	185	1
1E6J	3.441	**6**	3	**3.322**	35	**7**
2BTF	**3.624**	**93**	**1**	-	-	-
2PCC	**3.711**	48	2	3.770	**37**	**4**
1DQJ	3.714	47	1	**3.029**	**38**	**3**

Table 3.5 continued on next page

Table 3.5 continued from previous page

	Flexible			Rigid-Body		
Complex	RMSD	Rank	Clus	RMSD	Rank	Clus
1I2M	3.765	**2**	7	**3.648**	4	**17**
1HE1	**3.814**	**205**	**1**	-	-	-
1MLC	**3.834**	**47**	**1**	-	-	-
1GP2	**3.884**	**11**	**1**	-	-	-
2SIC	**3.935**	**41**	**2**	-	-	-
1KKL	**4.009**	**32**	**1**	-	-	-
2JEL	**4.362**	183	1	4.552	**76**	1
1BVN	**4.409**	**115**	**1**	-	-	-
1BVK	4.437	**128**	1	**3.525**	141	1
1B6C	**4.587**	**145**	**1**	-	-	-
1K4C	**4.597**	**168**	**1**	-	-	-
1ACB	**4.625**	**121**	**1**	-	-	-
1I4D	**4.708**	**66**	**1**	-	-	-
2HMI	**4.721**	**217**	**1**	-	-	-
1F51	**4.765**	**179**	**1**	-	-	-
1D6R	4.852	**19**	**4**	**3.460**	27	1
2MTA	4.947	**21**	1	**4.525**	35	**4**
1CGI	-	-	-	**4.527**	**123**	**2**
2VIS	-	-	-	**3.164**	**124**	**1**
1MAH	-	-	-	**1.795**	**148**	**1**
1GHQ	-	-	-	**4.547**	**124**	**1**
1M10	-	-	-	**4.822**	**58**	**2**
1EER	-	-	-	**4.365**	**93**	**1**

Table 3.6: The performance of SwarmDock using the Van der Waals and electrostatics potentials. Complexes are categorised as enzyme/inhibitor (EI), antibody/antigen (AB) or other (OT). Complexes are also categorised according to their difficultly: rigid (Rig., IRMSD < 1.5Å), difficult (Diff., IRMSD > 2.2Å) and medium (Med., the remainder). Models are classified as high (High), medium (Med.) or acceptable (Acc.) as per the CAPRI criteria (see section 1.5.10 or Mendez *et al.* (2003)). Three criteria for success are also used; whetner the model is found, whether it is found and ranked in the top 100 clusters, or whether it is found and ranked in the top 10 clusters.

	Found.			Top 100			Top 10		
	Acc.	Med.	High	Acc.	Med.	High	Acc.	Med.	High
All (84)	56% (47)	26% (22)	6% (5)	42% (35)	20% (17)	5% (4)	14% (12)	8% (7)	4% (3)
Rig. (63)	63% (40)	32% (20)	8% (5)	46% (29)	24% (15)	6% (4)	17% (11)	11% (7)	5% (3)
Med. (13)	46% (6)	15% (2)	0% (0)	46% (6)	15% (2)	0% (0)	8% (1)	0% (0)	0% (0)
Diff. (8)	12% (1)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
EI (22)	54% (14)	23% (6)	4% (1)	50% (13)	23% (6)	4% (1)	12% (3)	4% (1)	0% (0)
AB (26)	73% (16)	36% (8)	14% (3)	50% (11)	27% (6)	14% (3)	27% (6)	18% (4)	14% (3)
OT (36)	47% (17)	22% (8)	3% (1)	31% (11)	14% (5)	0% (0)	8% (3)	6% (2)	0% (0)

Of the complexes in Table 3.5, 20 were not found by the rigid body method, whilst 6 were not found by the flexible method. Hence inclusion

of flexibility does increase the chance of finding the correct binding site. However, of the 37 complexes for which both runs bound, the interface RMSD found using the flexible method was lower for only 18 complexes, indicating that neither method is consistently superior in terms of RMSD. Furthermore, whilst the chances of finding the correct binding site is greater with the flexible method, the site is found less frequently; 34 of the sites found with the flexible method had a cluster size of 1, compared to 14 for those found using the rigid-body method. However, for the flexible method, the best structure was found in the top 10 ranked clusters for 10 cases, 4 of which were not found with the rigid-body method. Rigid-body docking ranked the correct structure in the top 10 for 7 complexes. Of these, 3 were better ranked using the flexible method, 3 were equally ranked with both methods and only 1 was not found with the flexible method. Of the correctly bound structures found using both methods, which were not ranked in the top 10 of either methods, 18 were best ranked using the flexible method and 8 using the rigid-body method. This lead us to conclude that modelling flexibility using a linear combination of normal modes improves both the chance of finding the correct binding site and the rank of the solution once found.

3.3.5 Docking as a Function of Modes

The previous analysis showed that the inclusion of flexibility notably improved the performance of SwarmDock. However, it remained uncertain as to how many modes should be used to model flexibility. In order to investigate this, eight of the complexes were chosen for further docking runs, in which the number of modes included are varied.

Four of the chosen cases were 'easy' targets, for which only small conformational changes occur upon binding. These were docked globally; four times from each starting position. One of these was the complex 1AY7, which originally docked with low RMSD and a good rank. This was chosen to investigate whether the inclusion of extra flexibility increased the number of false positives. Another complex, 1GCQ, was chosen to see if it improved in rank because, as shown in the previous section, low RMSD solutions were found, but these ranked poorly. Conversely, 1E6J was chosen as it ranked

Table 3.7: The mean cluster size for the correct binding site (RMSD < 5Å), averaged over runs including different numbers of modes. The greatest values are marked in bold.

	Modes			
Complex	1-10	11-20	21-30	31-40
1AY7	**15.1**	11.2	12.3	14.3
1GCQ	1.2	1.9	**2.9**	2.6
1E6J	10.0	9.9	11.0	**11.5**
1TMQ	2.1	2.9	**4.7**	2.9
1EER	3.6	3.1	5.3	**5.9**
1K5D	5.9	6.9	**7.4**	6.6
1GRN	26.6	28.7	**29.3**	27.3
1KKL	2.8	2.9	3.9	**5.6**

well, but the RMSD had room for improvement. 1TMQ was chosed for its mediocrity, since its rank and RMSD were both moderate.

The other four complexes were those which contain significant backbone rearrangements at the interface. These complexes, 1EER, 1GRN, 1K5D and 1KKL, have C_α RMSDs at the interface of 2.44Å, 1.22Å, 1.19Å and 2.20Å respectively. These structures were docked locally, with the algorithm being run 8 times from the 10 starting positions nearest the binding site.

For each of these complexes, unbound-unbound docking was performed with between zero and 40 normal modes included in both the receptor and the ligand. For some of the runs in which a large number of modes were used, some non-physical structural deformations were observed, compromising the integrity of the complex. To correct any such perturbations, the structures underwent extensive minimisation in CHARMM (600 steps steepest decent for rapid minimisation, 500 steps conjugate gradient for higher accuracy and 4000 steps adopted basis Newton-Raphson for fine-tuning). For all runs, the structures were clustered at 2.5Å RMSD. The rank and RMSD of the global runs are shown in Figure 3.5, and of the local runs in Figure 3.6.

For the global docking, the gradient of the regression line for both the rank and RMSD is negative in all cases, suggesting that the inclusion of more modes generally improves the ability of the algorithm to model the conformation of the interfaces as well as discriminate between false

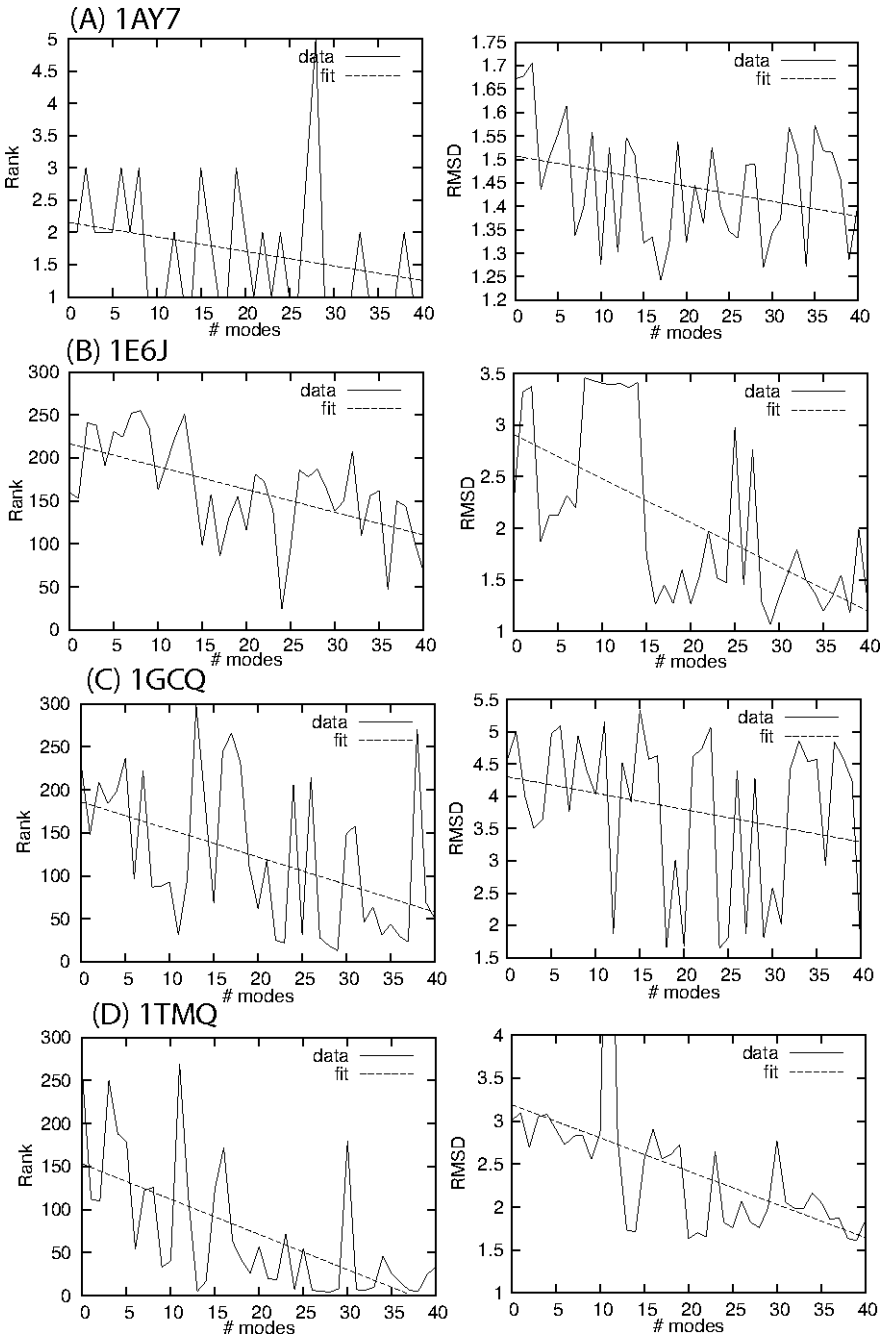

Figure 3.5: Docking results and linear regression for global docking. I_RMSD is the lowest value found during the run. Image adapted from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

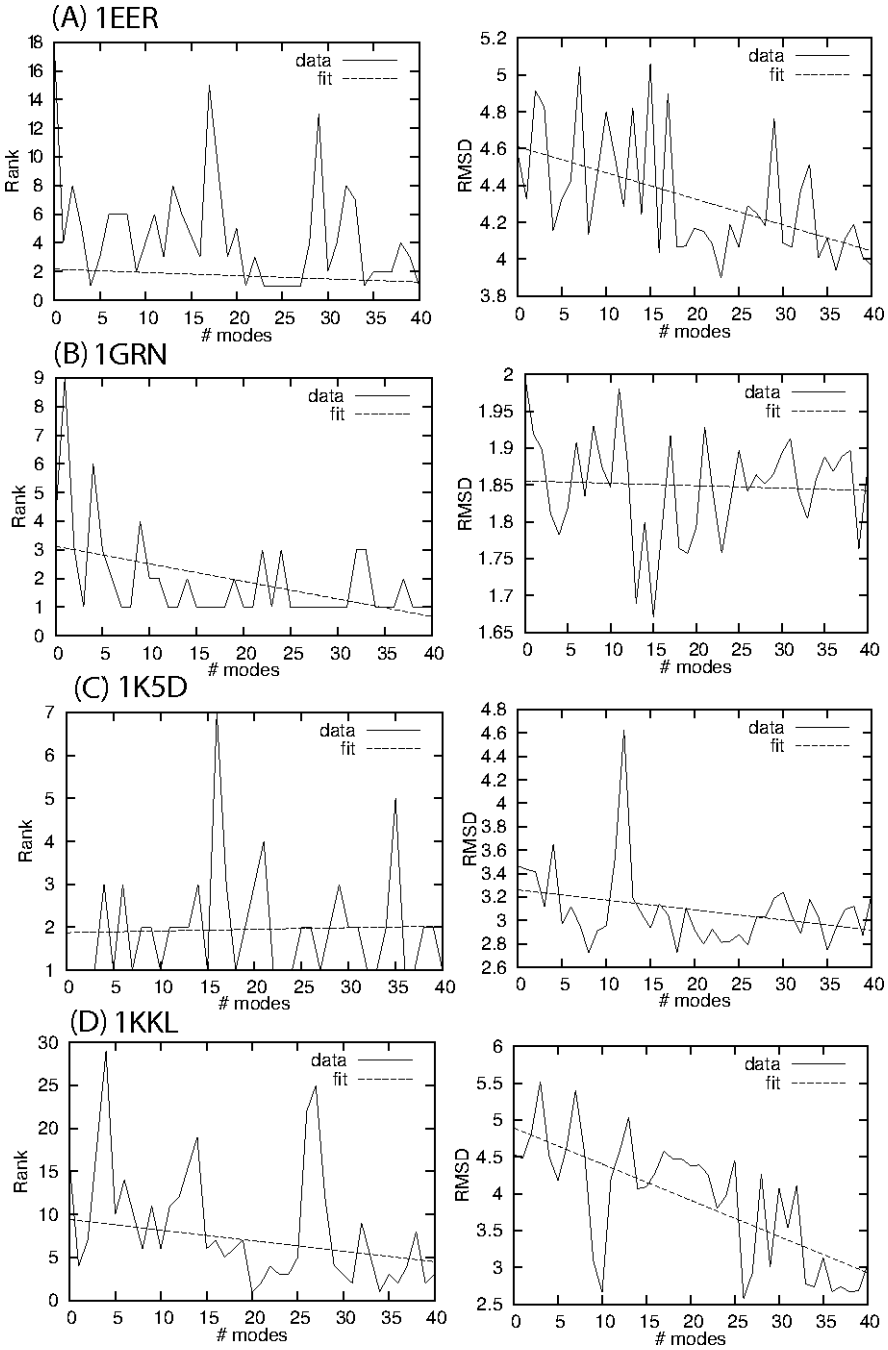

Figure 3.6: Docking results and linear regression for local docking. I_RMSD is the lowest value found during the run. Image adapted from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

positives and the true structure, despite the fact that these complexes undergo only small conformational changes upon binding. This suggested that perhaps the algorithm preferentially broadens or deepens the binding funnel relative to the energy wells surrounding false positive structures. For the local docking runs, near-native structures with low RMSD could be found in all cases, and for 1EER and 1KKL, significant improvements in RMSD were found as the number of included modes was increased.

As a broader binding funnel has a wider rim, the search algorithm is more likely to stumble upon it by chance, and subsequently locate the minimum. To test the idea that the inclusion of extra normal modes broadens the true binding funnel, the size of the cluster corresponding to the binding interface was investigated. The mean cluster sizes, averaged over runs with various numbers of modes, is shown in Table 3.7. For all but one of the eight cases, the binding site was located more frequently in the runs including 21-30 modes, or 31-40 modes, suggesting that the binding funnel was broadened by the inclusion of a higher number of modes.

To investigate the possibility that as more modes are included, the true binding funnel is preferentially deepened compared to false positive funnels, the minimum energy of a near-native structure found by SwarmDock was compared to the lowest energy false positive structures. The results of this analysis are shown in Figure 3.7. With the exception of 1E6J, the energy of the false positive structures did not decrease significantly as the number of modes included increased. The energy of the docked structures closest to the native, however, did seem to decrease significantly in most cases, suggesting that the inclusion of ever higher frequency modes does preferentially deepen the true binding funnel.

3.3.6 The DComplex Potential

After the EEF1 desolvation term was abandoned (see section 3.3.3), we were concerned about the neglection of this important driving force for protein association. Due to the importance of hydrophobic burial, it was decided that the DComplex statistical pair potential should be programmed into

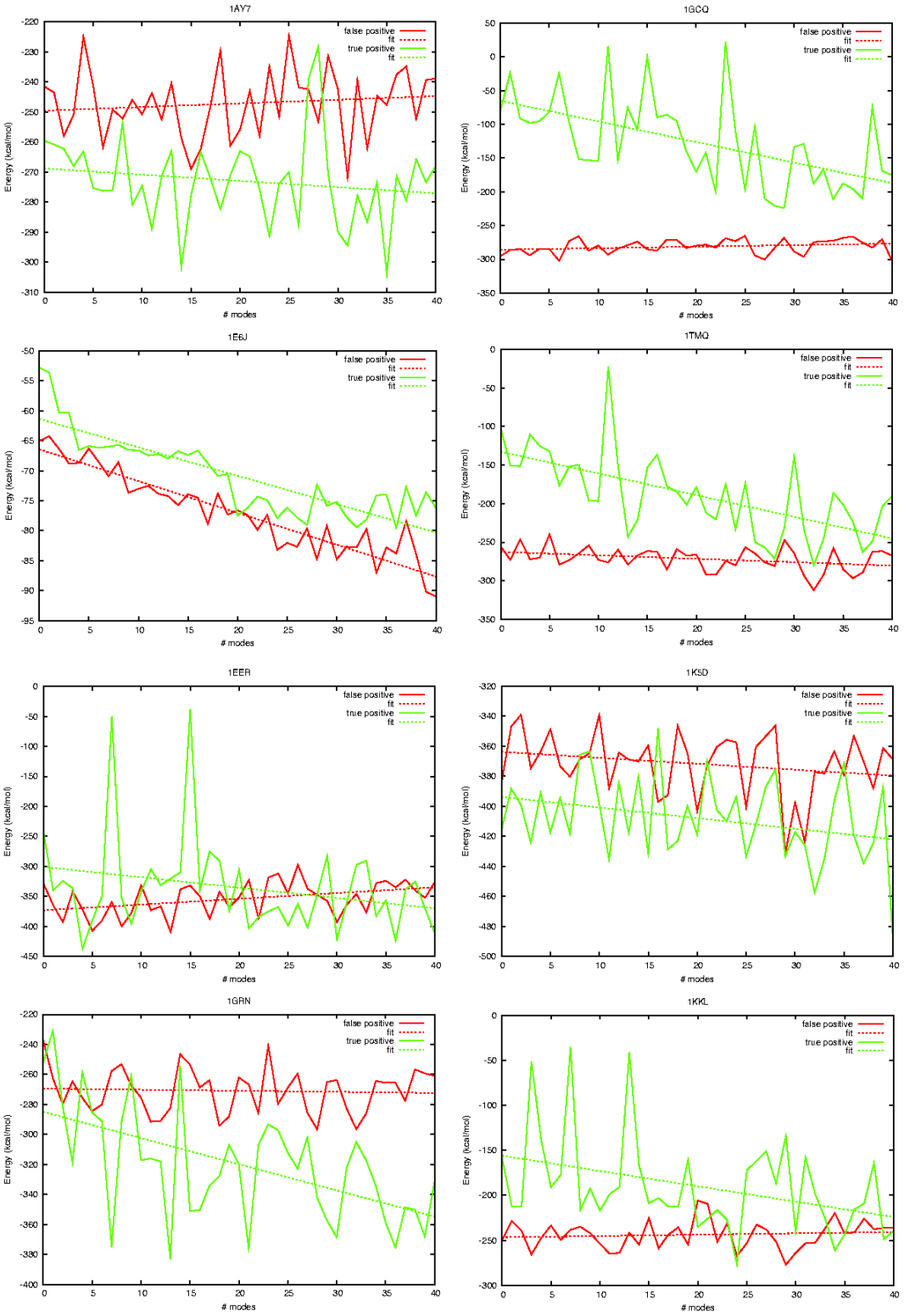

Figure 3.7: Energy of lowest energy true positive structure and mean energy of the lowest energy member of the 5 lowest energy false positive clusters. Image taken from Moal and Bates (2010) under the Creative Commons Attribution 3.0 licence.

SwarmDock, as it implicitly accounts for some desolvation effects. With this new energy function, flexible unbound-unbound docking was performed on 78 cases from the benchmark and compared with the results for the previous scoring function. Aside from the different energy function, the runs were identical. For 67 of the cases tested, a structure with interface RMSD below 5Å was found. A summary of these results is found in Table 3.8, showing the best cluster size, cluster rank and interface RMSD for each complex.

Table 3.8: DComplex vs. VDW & Elec Unbound-Unbound docking. The best performance for each metric is highlighted in bold.

	DComplex			VDW & Elec.		
Complex	RMSD	Rank	Clus	RMSD	Rank	Clus
1NCA	0.467	10	6	**0.391**	**1**	6
1PPE	**0.510**	**1**	**15**	2.467	23	1
1JPS	**0.624**	26	**3**	1.567	**19**	2
1BJ1	**0.668**	**112**	**2**	2.990	134	1
1KXQ	**0.756**	**6**	**8**	3.058	57	1
1MAH	**0.829**	**36**	**2**	-	-	-
1QFW	0.845	84	2	**0.821**	**5**	**3**
1FSK	**0.846**	6	**9**	1.437	**4**	1
1AHW	**0.868**	22	**8**	1.443	**6**	1
1IQD	**0.892**	**83**	**3**	-	-	-
1EAW	**0.913**	5	**5**	1.866	**1**	1
1GCQ	**0.921**	**41**	**6**	1.399	122	1
1TMQ	**0.951**	**1**	3	2.589	72	3
1MLC	**1.002**	**20**	**4**	3.834	48	1
1AVX	**1.017**	30	**2**	1.591	**25**	1
1AY7	1.027	19	5	**0.872**	**2**	**6**
7CEI	**1.141**	13	**11**	1.313	**12**	9
1E6J	**1.173**	19	**7**	3.441	**6**	3
1KXP	**1.184**	**2**	**9**	3.277	107	1
1ML0	**1.192**	**1**	**12**	1.553	5	1
1QA9	1.311	76	**11**	**1.158**	**35**	3
1KTZ	1.340	**98**	1	**0.660**	114	**6**
1BVN	**1.341**	**22**	**3**	4.409	104	1
2PCC	**1.422**	**36**	**4**	3.711	49	2
1I9R	**1.448**	36	2	1.483	**1**	**10**
1WEJ	**1.451**	**22**	1	2.750	36	**2**
1GRN	**1.532**	55	**7**	1.578	**17**	4
1F51	**1.603**	**28**	**4**	4.765	184	1

Table 3.8 continued on next page

Table 3.8 continued from previous page

	DComplex			VDW & Elec.		
Complex	RMSD	Rank	Clus	RMSD	Rank	Clus
1NSN	**1.620**	**3**	**8**	2.523	20	2
1BUH	**1.657**	**11**	**29**	1.921	60	6
1UDI	**1.726**	**4**	**13**	-	-	-
2QFW	1.757	**20**	**6**	**1.272**	83	1
1RLB	**1.798**	**84**	**2**	-	-	-
2SNI	**1.839**	**38**	1	2.801	180	1
1HE1	**1.850**	**27**	3	3.814	171	3
1KAC	**1.925**	30	5	2.238	**1**	**9**
1DFJ	**1.939**	**1**	**3**	-	-	-
1E96	**1.939**	**4**	**9**	3.099	180	2
1EWY	2.015	**6**	**15**	**1.845**	13	3
1B6C	**2.157**	**3**	**15**	4.587	148	1
2SIC	**2.239**	113	1	3.935	**34**	**2**
1F34	**2.386**	**3**	**6**	3.259	16	3
2JEL	**2.466**	**9**	**3**	4.362	184	1
1E6E	2.512	**1**	**13**	**1.745**	94	2
1I2M	**3.013**	3	**9**	3.765	**2**	7
1D6R	**3.044**	**15**	**5**	4.852	85	1
2BTF	**3.238**	**78**	**3**	3.624	95	1
1CGI	**3.312**	**8**	**2**	-	-	-
2MTA	**3.574**	**3**	**5**	4.947	21	1
1HE8	3.644	58	**2**	**1.448**	**46**	1
1FQJ	3.728	71	2	**2.264**	**11**	2
1VFB	3.890	**32**	1	**1.403**	118	1
1GP2	4.032	91	2	**3.884**	**11**	2
1DQJ	4.079	**10**	**2**	**3.714**	42	1
1M10	**4.324**	**51**	**1**	-	-	-
1I4D	**4.350**	**36**	1	4.708	66	1
1HIA	4.365	2	**4**	**2.606**	2	2
1K4C	**4.413**	**6**	**3**	4.597	170	1
1SBB	**4.558**	**150**	**3**	-	-	-
1AKJ	**4.723**	**107**	**1**	-	-	-
1A2K	**4.756**	**151**	**2**	-	-	-
1KKL	4.768	**25**	**2**	**4.009**	33	1
1EER	**4.964**	**154**	**1**	-	-	-
1K5D	-	-	-	**3.141**	**31**	**1**
2HMI	-	-	-	**4.721**	**217**	**1**
1BVK	-	-	-	**4.437**	**133**	**1**
1ACB	-	-	-	**4.625**	**124**	**1**

The DComplex energy function outperformed the Van der Waals and electrostatics energy function by all metrics. It correctly found the bound structure for 11 cases which were not found by the previous energy function, as opposed to only four complexes which were found with the previous scoring function but not DComplex. Further, of the structures found by both methods, DComplex had a lower interface RMSD in 37 cases, as opposed to 15 cases. Further, the cluster size was larger for 37 cases versus 7 for the previous function. These metrics show that this potential is better able to generate the correct docked structure. Additionally, the correct structure was ranked higher in 32 cases, as opposed to 19, showing that the DComplex potential is also better able to discriminate the correct structure from the other generated poses. A summary of the performance of SwarmDock using the DComplex potential is shown in Table 3.9. As shown above, the success rates are greater than for the runs in which the Van der Waals and electrostatics potentials are used (Table 3.6). As with the previous energy function, the success rates drop dramatically as the extent of conformational change increases. Table 3.9 also gives us absolute measures for both the ability to generate the correct docked pose and the ability to select is; whilst an acceptable or better structure could be found for over two thirds of the complexes in the benchmark, only around a quarter of the complexes were ranked in the top 10 clusters of lowest energy. As the DComplex potential was added later in the PhD period, it was not included in the runs for the analyses in previous sections, and was used only in later rounds of CAPRI.

Table 3.9: The performance of SwarmDock using the DComplex potential. Complexes are categorised as enzyme/inhibitor (EI), antibody/antigen (AB) or other (OT). Complexes are also categorised according to their difficultly: rigid (Rig., IRMSD < 1.5Å), difficult (Diff., IRMSD > 2.2Å) and medium (Med., the remainder). Models are classified as high (High), medium (Med.) or acceptable (Acc.) as per the CAPRI criteria (see section 1.5.10 or Mendez *et al.* (2003)). Three criteria for success are also used; whetner the model is found, whether it is found and ranked in the top 100 clusters, or whether it is found and ranked in the top 10 clusters.

	Found.			Top 100			Top 10		
	Acc.	Med.	High	Acc.	Med.	High	Acc.	Med.	High
All (78)	68% (53)	50% (39)	17% (13)	62% (48)	49% (38)	15% (12)	26% (20)	14% (11)	6% (5)
Rig. (61)	80% (49)	62% (38)	21% (13)	75% (46)	61% (37)	20% (12)	31% (19)	18% (11)	8% (5)
Med. (10)	30% (3)	10% (1)	0% (0)	20% (2)	10% (1)	0% (0)	10% (1)	0% (0)	0% (0)
Diff. (7)	14% (1)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
EI (25)	80% (20)	52% (13)	16% (4)	76% (19)	52% (13)	16% (4)	44% (11)	16% (4)	8% (2)
AB (20)	75% (15)	65% (13)	40% (8)	70% (14)	60% (12)	35% (7)	25% (5)	20% (4)	15% (3)
OT (33)	55% (18)	39% (13)	3% (1)	45% (15)	39% (13)	3% (1)	12% (4)	9% (3)	0% (0)

In order to see how well this energy function performs in SwarmDock compared to other methods, a comparison was done with a number of other docking algorithms. Unfortunately, there is no standard method to evaluate the performance of an algorithm which is used consistently throughout the literature. However, Table 3.10 compares the SwarmDock results with results from two recent publications in which the criteria for a successful hit varies markedly in their tolerance of deviations from the crystal structure of the complex (Ravikant and Elber, 2011; Shentu *et al.*, 2008). When the more lenient criterion is used (rank < 250 and IRMSD < 4Å), SwarmDock correctly identifiess two structures more than ZDOCK/ZRANK, but eight structures fewer than DOCK/PIE. Interestingly, SwarmDock performs best on the rigid cases and fails to identify any of the difficult cases. Presumably, this is due to the softness of the potential used in the initial search compared to the relatively hard DComplex potential used by SwarmDock. It should be noted that whilst the ZDOCK/ZRANK and DOCK/PIE methods include a re-ranking stage, no such stage is included for SwarmDock. Further, the PIE and ZRANK re-ranking methods were trained using the same complexes which are used to evaluate their performance, and thus overfitting cannot be ruled out. No such biases exist for the DComplex potential. When the stricter criterion is used (rank < 250 and IRMSD < 2.5Å) and no re-ranking stage is included, SwarmDock performs very favourably compared to Patchdock, Context Shapes and ZDOCK. Although no highly flexible complexes are correctly identified, SwarmDock outperforms in all other categories. However, this performance comes at much higher computational expense relative to these other methods, and thus this may not constitute a fair comparison.

3.4 Discussion

SwarmDock, a protein-protein docking algorithm, has been developed. Flexibility is modelled with elastic network normal modes, calculated using the RTB method (see section 2.3.2.1). Putative docked poses are generated using a memetic algorithm composed of a PSO and a local search. The algorithm has been tested with three different scoring functions, on a large set of protein-protein interactions, and the inclusion of flexibility was shown

Table 3.10: A comparison of SwarmDock with a number of other methods, in terms of the number of hits found. Complexes are categorised as enzyme/inhibitor (EI), antibody/antigen (AB) or other (OT). Complexes are also categorised according to their difficultly: rigid (Rig., IRMSD < 1.5Å), difficult (Diff., IRMSD > 2.2Å) and medium (Med., the remainder). The methods are SwarmDock (SD), DOCK/PIE (DP), ZDOCK/ZRANK (ZDR), Patchdock (PD), ZDOCK (ZD) and Context Shapes (CS).

	All (78)	Rig. (61)	Med. (10)	Diff. (7)	EI (25)	AB (20)	OT (33)
SD†	67% (52)	80% (49)	30% (3)	0% (0)	76% (19)	80% (16)	52% (17)
DP†	74% (58)	79% (48)	50% (5)	71% (5)	96% (24)	70% (14)	61% (20)
ZDR†	64% (50)	74% (45)	30% (3)	29% (2)	76% (19)	80% (16)	45% (15
SD‡	55% (43)	69% (42)	10% (1)	0% (0)	60% (15)	75% (15)	39% (13)
PD‡	8% (6)	10% (6)	0% (0)	0% (0)	12% (3)	10% (2)	3% (1)
ZD‡	13% (10)	16% (10)	0% (0)	0% (0)	24% (6)	15% (3)	3% (1)
CS‡	10% (8)	13% (8)	0% (0)	0% (0)	12% (3)	25% (5)	0% (0)

† The criteria for success is rank < 250 and IRMSD < 2.5Å. Results taken from Ravikant and Elber (2011).
‡ The criteria for success is rank < 250 and IRMSD < 4.0Å. Results taken from Shentu *et al.* (2008).

to produce a marked improvement in performance compared to rigid-body. The best results were obtained using the DComplex statistical pair potential, for which reasonable structures could be generate for around 80% of the unbound-unbound cases.

The algorithm was further tested on 8 cases of varying difficulty, with between zero and 40 normal modes included in the search. The rank and RMSD of the generated structures was shown to improve as more modes were included. These simulations also suggest that the inclusion of higher modes broadens and preferentially deepens the binding funnel.

Whilst SwarmDock has shown promise, there is still much scope for future development. Automated re-ranking, accounting for cluster size and rank simultaneously is underway. Further, even with the DComplex potential, only 20 of the structures were ranked in the top 10 clusters. Of the 63 structures which bound correctly, 9 were only found once, and 36 of them were found fewer than 5 times. This strongly suggests that a higher proportion of the complexes would have been correctly identified had more runs been undertaken. However, running the algorithm more times runs the risk of generating false positives and potentially lowering the rank of the correct structures. The criticality of the energy function for both searching and ranking is evident, and there are many avenues of exploration which can be taken to optimise it for the task of docking. For re-ranking, machine learning and artificial intelligence techniques can be

used to combine terms describing the interaction, such as pair potentials or desolvation energy models (see section 1.5.7). The SwarmDock pipeline would benefit from such a step, and its implementation should not prove a difficult future development. However, these methods can only help choose the most likely structure. To find an energy function to be used within the algorithm, which facilitates the generation of near-native poses is a more difficult. One possibility would be to sample the energy wells surrounding both true and false positive structures. Machine learning tools could be used in an attempt to derive an energy function which broadens and deepens the true binding funnel, whilst attenuating the false energy wells, thus bolstering the swarm's ability to navigate the energy surface towards the native structure.

Another important shortcoming regarding SwarmDock is its inefficient side-chain modelling. While modelling flexibility with normal modes does allow linear movement of side-chain atoms, it does not permit rotamer switching, repacking or specific torsion angle changes. There are a number of possibilities which can be used for this. One of these would be to include a sidechain repacking algorithm for each particle at each iteration, requiring up to 500,000 repackings per run. In tests where all rotamers are packed in a test set of 65 proteins, the scwrl and OPUS-rota methods took on average 2 and 8 seconds per protein respectively (Lu *et al.*, 2008b). In SwarmDock, only the interface atoms would require repacking, and this figure would drop dramatically. However, even this would still represent an inordinate computational expense, and a back of the envelope calculation in which an interface repacking process takes on average 0.1s would still increase the running time a SwarmDock run from around 10 minutes to 10 hours, far beyond what is desirable considering that the algorithm is currently run around 240 times per complex.

An alternative side-chain modelling approach would be to derive a discrete 'side-chain space' for both binding partners prior to running SwarmDock, and explore this simultaneously with the translational, orientational and conformational space. Sub-optimal solutions to the solvent exposed rotamer optimisation problem could be linked together on a graph, with each solution corresponding to a node. Solutions which differ by one

rotamer can be linked by an edge. This graph can be cheaply pre-computed. The PSO can navigate this graph, using a discrete version of the same equations which propagate SwarmDock through the rest of search space (discrete PSOs are discussed, for instance, in Engelbrecht (2006)). The distance between any given pair of positions on the graph can be pre-calculated using Dijkstra's algorithm, as can the identity of the edge which needs to be crossed to move from one position towards another. This approach would be fast, as the potential packed solutions would be pre-calculated, and the only additional computations required during the SwarmDock run itself would be handled by fast indexing with the Boost Graph Library, and the reassignment of the rotamer corresponding to the edge. As long as a pair of rotamer sets (one for the ligand and one for the receptor) correspond to a set of rotamers in the ensemble of bound structure, this approach has the potential to significantly enhance the SwarmDock pipeline.

Chapter 4

Crowding and Search Space Reduction

4.1 Introduction

Protein-protein docking is traditionally undertaken by sampling the position, orientation and conformation of two proteins in isolation. However, proteins have evolved to interact *in vivo*, where they are constantly buffeted by water and the other molecular constituents of the cell, the latter of which can occupy a large proportion of the volume. This crowded environment gives rise to phenomena collectively entitled 'macromolecular crowding effects' (Zimmerman and Trach, 1991; Minton, 1981; Hall, 2003; Ellis, 2001; Zhou *et al.*, 2008; Hall, 2003). Proteins constantly undergo non-specific interactions, forming temporary associations held together by electrostatics and desolvation forces, only to disengage swiftly. For binding partners of biological significance, the formation of this encounter complex is likely antecedent to the formation of the final docked ensemble (Tang *et al.*, 2006; Camacho *et al.*, 1999; Blundell and Fernandez-Recio, 2006). The phenomenon of encounter complex formation is surely an exaptation to the evolution of specific protein-protein interactions. Steering effects brought about by complementary electrostatics will result in preferential contact of certain surface patches. From these favoured regions, proteins can explore each others surfaces. The surface search will not be entirely random, but electrostatics and shorter range forces will likely bias it in certain directions. During the surface search, the two proteins will spend a greater amount of

time in lower energy regions. Should the lifetime of this interaction have consequences for the fitness of the organism, then it can be acted upon by natural selection. The mutation of surface residues could increase the efficiency of electrostatic steering, bias the surface search, or further deepen the energy well. Such a mechanism seems a highly plausible explanation for the origin of the binding funnel. It is a particularly efficient mechanism, as the 2D surface search constitutes a reduction in dimensionality compared to a 3D search in which the binding partners must, by chance, come into contact with their binding surfaces correctly positioned and oriented.

In light of this, an understanding of encounter complex formation has the potential to be informative for docking; should it possible to glean information regarding the encounter complex ensemble, docking searches could be focussed upon regions of frequent or tenuous contact. During the development of SwarmDock a colleague, Xiaofan Li, was developing BioSimz, a Rigid-body Langevin dynamics package for the purpose of studying encounter complex formation and macromolecular crowding effects (Li *et al.*, 2010b,a; Li, 2011). Information regarding encounter complex formation was derived from BioSimz simulations and used to restrict the search space during SwarmDock runs. Simulations were run both with and without external crowder proteins included to replicate the effects of the crowded *in vivo* environment. This work has been published (Li *et al.*, 2010b), and was recently described as "an important new research area that will clearly be very relevant to CAPRI in the future" (Lensink and Wodak, 2010b). In this work, Langevin dynamics simulation and trajectory merging was performed by Xiaofan Li. The scoring of SwarmDock starting positions, statistical tests and all docking runs were performed by myself. The analysis of the influence of crowding effects by molecular type was performed in unison.

4.2 Methods

4.2.1 Data Set

Unbound structures for 26 complexes in the easy and medium categories of the docking benchmark 2.0 were studied (Mintseris *et al.*, 2005). These include enzyme-inhibitor (1AVX, 1AY7, 1PPE, 7CEI, 1TMQ, 1EAW, and 1HIA), enzyme-substrate (1EWY and 1E6E), antibody-antigen (1QFW, 1JPS, 1NCA, 1VFB, 1AHW, 1NSN, 1I9R, and 1FSK), and signal-effector/receptor (1KTZ, 1GCQ, 1GRN, 1FQJ, 1BUH, 1KAC, 1ML0, 1QA9, and 1HE8) complexes. The term 'receptor' refers to the larger of the two binding partners, and 'ligand' to the smaller. The methods presented here were also applied to a number of CAPRI targets.

4.2.2 BioSimz

The BioSimz package is an efficient rigid-body dynamics simulator for the simulation of biological molecular systems at atomic resolution. Full details of the package are given elsewhere (Li *et al.*, 2010b,a; Li, 2011). Briefly, BioSimz models dynamics based on the Langevin dynamics scheme whereby the temporal evolution of a molecule at position $\mathbf{x}$ is modelled using the following form of Newton's equation of motion

$$m\frac{d^2\mathbf{x}}{dt^2} = -\nabla\sum_i^N u(\mathbf{x}_i) - \gamma\frac{d\mathbf{x}}{dt} + \sqrt{2\gamma k_B T m}\xi(t) \tag{4.1}$$

where $\gamma = 10^{-11}$s is the damping constant which models the hydrodynamic drag interaction between the solvent and the solute, and the last term models random Brownian motion in which $\xi(t)$ is a normally distributed random force. The potential energy $u(\mathbf{x}_i)$ of atom i is calculated using the Van der Waals and electrostatics terms in the CHARMM27 force field, along with a desolvation and hydrogen bonding term.

For all runs, molecules are randomly distributed in a $240 \times 240 \times 240$Å^3 box with periodic boundary conditions. The molecules were allowed to equilibrate in a 10ns high-energy run. Subsequently, 200ns of dynamics with a 1ps timestep was performed at 298K. Each simulation was run

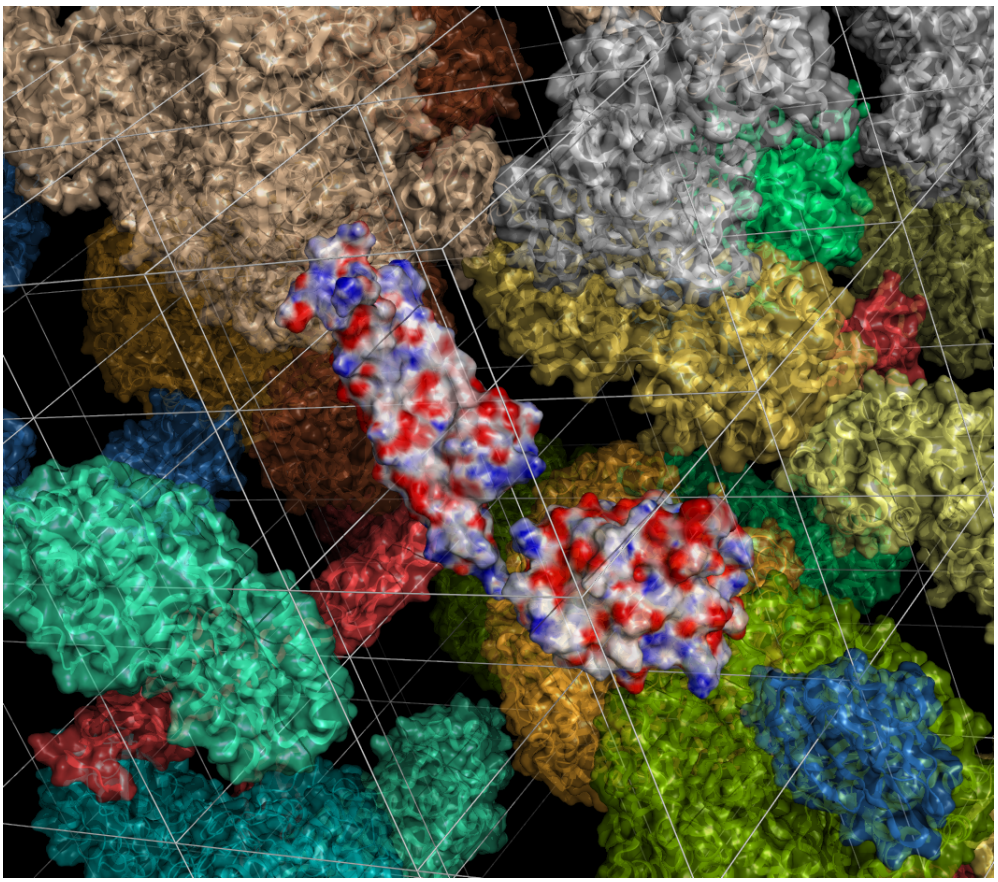

Figure 4.1: A snapshot of a crowded BioSimz simulation of the interaction between TGFβ3 and TβII (PDBid 1KTZ). The binding partners, which are forming an encounter complex, are coloured according to their surface electrostatics as calculated with APBS. Figure reproduced from Li *et al.* (2010b) with permission from Wiley.

Table 4.1: External crowder proteins used in BioSimz simulations

#	Name	Weight (Da)	PDB	Copy #
1	Hexokinase	52,209.9	1IG8	2
2	Phosphoglucose Isomerase	125,225.0	1HOX	2
3	Phosphofructokinase	35,263.2	4PFK	2
4	Fructose 1,6-bisphosphate Adolase	37,046.1	1ZEN	2
5	Triose Phosphate Isomerase	53,326.5	2YPI	1
6	Glyceraldehyde-3-phosphate Dehydrogenase	73,379.5	3GPD	2
7	Phosphoglycerate Kinase	45,315.5	3PGK	1
8	Phosphoglycerate Mutase	56,864.5	1EQJ	2
9	Endolase	93,739.4	2ONE	2
10	Pyruvate Kinase	198,428.0	1E0U	1

10 times and the results amalgamated for the remainder of the analysis. Eight receptors and eight ligands were included in each simulation, corresponding to a concentration of 1.92mM. Due to the variation in the size of the proteins, this corresponds to between 18.5 and 19.2 gL^{-1}. Runs were performed either with or without the presence of external crowder molecules not involved with the interaction in question. The crowder molecules chosen were taken from the glycolytic pathway, as these are present in high abundance, are highly conserved between species and pervasive throughout almost all forms of life (Ishihama *et al.*, 2008). The identities of the crowders are shown in Table 4.1. The crowded simulations correspond to a total protein density of between 168 and 252gL^{-1}, a value comparable to the 300gL^{-1}, the approximate protein concentration of the *in vivo* cytosol (Zimmerman and Trach, 1991). Figure 4.1 shows a snapshot of a crowded simulation of the interaction between TGFβ3 and TβII (PDBid 1KTZ).

4.2.3 Combining BioSimz and SwarmDock

The BioSimz trajectories were used to generate a 'cloud' of ligand density surrounding the receptor. Firstly, the trajectories were converted to trajectory points, which are composed of one point per ligand molecule per picosecond, positioned at the ligand centre of mass. The density of trajectory points within a given region act as a measure of both how frequently that region is visited and the duration of those visits. In order to generate the ligand density 'cloud', the receptors for all timesteps are superimposed along

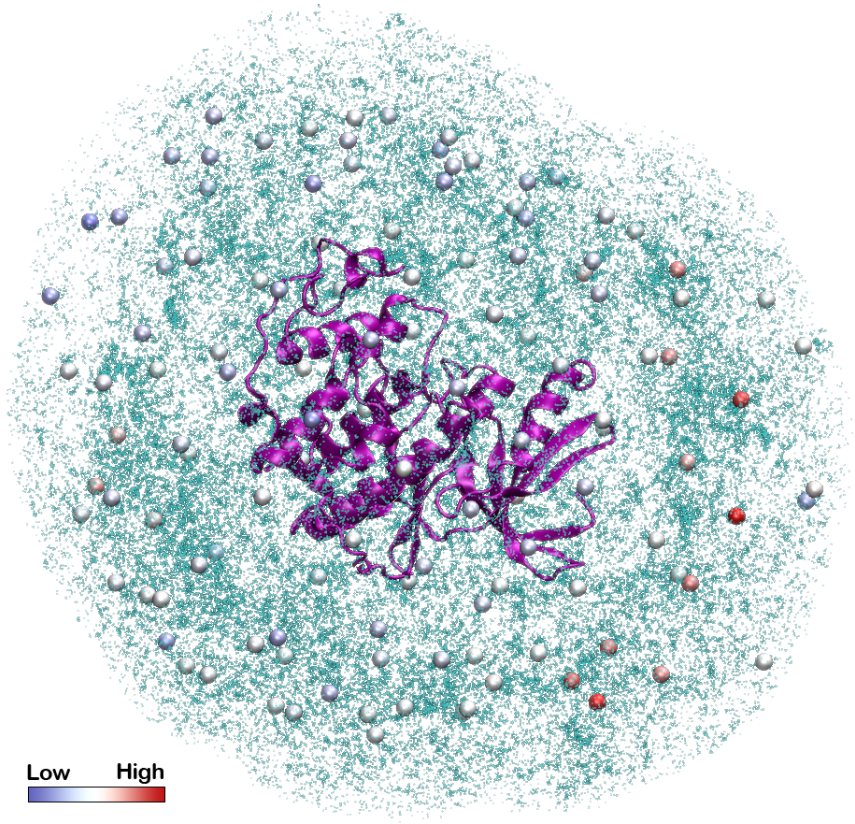

Figure 4.2: A ligand cloud derived from superimposing receptors and merging trajectory points derived from BioSimz simulations. The SwarmDock starting positions are shown, coloured according to their score.

with their associated trajectory points, which are subsequently merged. To ensure that only trajectories associated with interacting binding partners are considered, all trajectory points beyond a cutoff of a receptor atom are removed. This cutoff is determined as the longest distance between the centre of mass of the ligand and a ligand atom, plus an additional 7Å.

This 'cloud' of ligand density was used to restrict the search space used by SwarmDock, by eliminating half the starting positions on the basis that the ligand density surrounding them is low. The starting positions are initially generated using the method outlined in section 3.2.3. They are then scaled so as their mean distance from the receptor centre of mass is equal to that of the trajectory points. In order to score the starting positions based on the ligand density, the following procedure was employed. For each trajectory point, the scores of the nearest 5 starting points are incremented by 1. Due to slight unevenness in the spacing of starting positions, a correction term is applied; the score of each starting position is scaled by a quantity proportional to the inverse cube to the nearest starting point. This correction was chosen on the basis that the volume of space within a given distance of a point is proportional to the inverse cube of that distance. To illustrate, an example of a set of scored starting positions, along with the merged trajectory points, is shown in Figure 4.2.

4.2.4 Wilcoxon Rank-Sum Test

The ranked list of starting positions was used in two ways. Firstly, it was used to remove positions which scored poorly prior to SwarmDock runs. Secondly, it was used to determine whether the region surrounding the binding site was engaged in more frequently and tenuous encounter complexes than regions away from the binding site, with statistical significance. In order to determine this statistical significance, a one-tailed Wilcoxon rank-sum test was employed. For the 10 starting positions nearest the bound ligand centre of mass, the scores were tested against the null hypothesis that they were drawn from the same distribution as the scores of the remaining starting positions. The alternative hypothesis was that the scores of the 10 starting positions nearest the binding region were greater than the scores of the remaining starting positions. The tests were repeated with the alternative that the scores of the starting positions nearest the binding

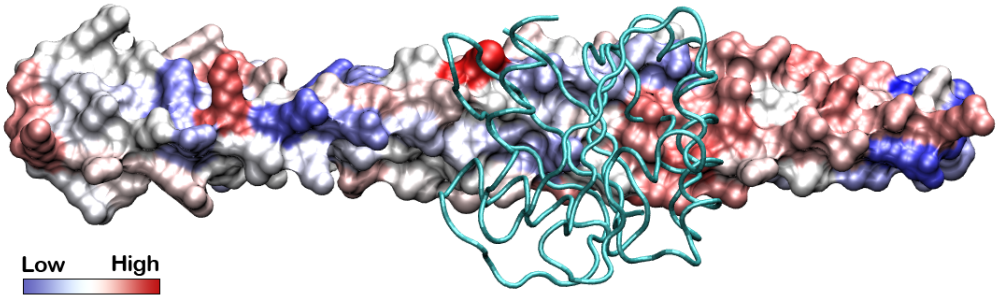

Figure 4.3: A contact heatmap for CAPRI target 37, the complex between human ADP-Ribosylation factor 6 (ARF6) and the second leucine zipper of JNK-interacting protein 4 (JIP4-LZ2), which was built by homology modelling. The most contacted zipper residue is in contingence with ARF6, which also contacts a diffuse region of frequent visitation to the right.

site were lower than those of the remaining points, to give an indication of whether the binding regions could be disfavoured during the simulations.

4.2.5 Surface Contact Heatmaps

BioSimz trajectories were also used to derive surface contact heatmaps showing which residues are most frequently involved in tenuous contacts. Residues are scored by the number of times they contact their binding partner. Only contacts associated with interactions which last for longer than 100ps are counted, so as to negate glancing blows which bias the count towards peripheral residues and are not relevant to encounter complex formation. An example of such a heat map is given in Figure 4.3, for which the colour is shown on a logarithmic scale.

4.3 Results

4.3.1 Uncrowded Simulations

Langevin dynamics simulations as outlined above were applied to the 26 complexes in the data set and used to score the SwarmDock starting positions. Wilcoxon rank-sum tests were then applied to test whether the 10

positions nearest the binding region scored significantly higher than the remaining positions at the 5% significance level. This was found to be the case for 12 of the complexes (1KAC, 1KTZ, 1FQJ, 1E6E, 1EWY, 1AHW, 1GRN, 1BUH, 1VFB, 1FSK, 7CEI and 1AY7). Subsequently, the test was applied to ascertain whether for any of the complexes, the 10 starting positions nearest the binding site scored significantly lower than the remaining positions at the same significance level. This was found to be the case for three of the complexes (1I9R, 1EAW and 1TMQ). It is interesting to note that for two of these complexes, 1EAW and 1TMQ, a protrusion from one of the binding partners intercalates deeply within a grove within the other, and as such it is unlikely that the binding process can be adequately modelled using rigid-body dynamics. The associated p-values of these tests are shown in Table 4.2.

Table 4.2: P-values for uncrowded simulations using one-tailed Wilcoxon rank-sum tests, testing whether the scores of the 10 starting positions nearest the ligand centre of mass are greater in value than the remaining points (Pg) or lesser in value (Pl). Significant results are highlighted in bold.

Complex	Pg	Pl	Complex	Pg	Pl
1GRN	**0.014**	0.986	1BUH	**0.016**	0.984
1HE8	0.335	0.668	1NCA	0.856	0.146
1QFW	0.256	0.747	1JPS	0.094	0.907
1GCQ	0.621	0.384	1AVX	0.372	0.632
1AY7	**0.049**	0.951	1AHW	**0.010**	0.990
1KTZ	**0.000**	1.000	1EWY	**0.007**	0.994
1VFB	**0.026**	0.975	1EAW	0.997	**0.003**
7CEI	**0.036**	0.965	1KAC	**0.000**	1.000
1QA9	0.361	0.643	1PPE	0.051	0.950
1FSK	**0.029**	0.972	1NSN	0.190	0.813
1FQJ	**0.002**	0.998	1TMQ	0.965	**0.036**
1I9R	0.998	**0.002**	1HIA	0.730	0.273
1ML0	0.221	0.781	1E6E	**0.004**	0.997

Following the scoring of the SwarmDock starting positions, two sets of SwarmDock runs were initiated. The first of these was identical to the flexible method used in section 3.3.4. The second set was the same, however with only half the number of starting positions, where the lowest scoring half are removed. For all runs, the complexes were clustered at 3.5Å resolution, ranked, and compared. The results are surmised in Figure 4.4. For 17 of the complexes, the rank was improved and for five, the rank

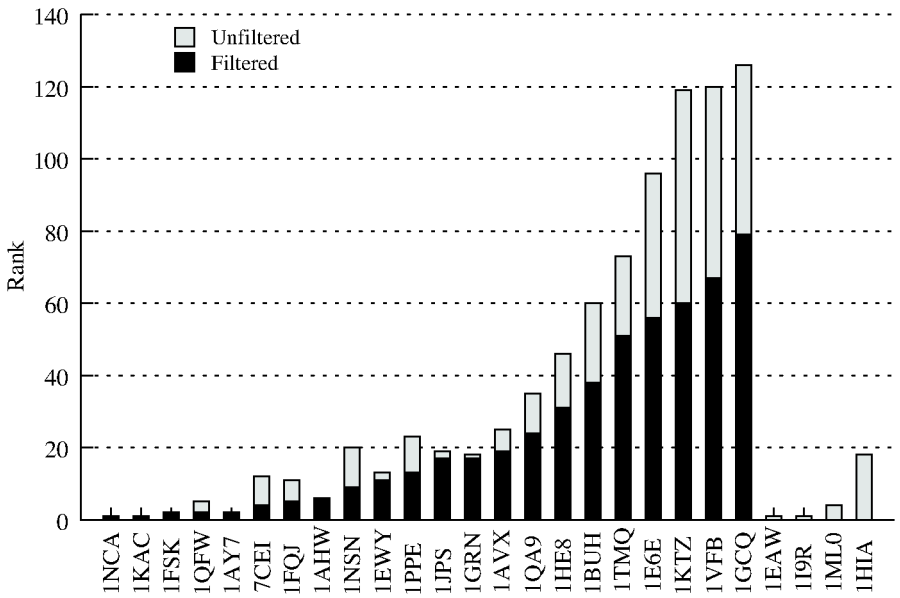

Figure 4.4: SwarmDock ranks following unbound-unbound docking, for the best ranking clusters less than 5Å away from the bound crystal structure. For the rightmost four complex, filtered runs failed to successfully dock. Image reproduced from Li *et al.* (2010b) with permission from Wiley.

remained the same (note that for four of these, the correct structure was already ranked in the top two). The remaining four structures failed to dock correctly using the filtered list of starting positions. For three of these (1EAW, 1I9R and 1HIA), the binding region was disfavoured in the BioSimz simulations, whilst for the other (1ML0), the binding region was enhanced, but SwarmDock failed to correctly identify the final docked pose.

Another pair of SwarmDock runs were set up. Again, the first of these was identical to that in section 3.3.4, with the algorithm run twice from each starting position. The second set of runs was same except the algorithm was run four times from half the starting positions, whereby the lowest scoring half of the starting positions were removed. The difference in cluster size between the unfiltered runs versus the filtered runs is shown in Figure 4.5, plotted against the $\min(\log P)$ value of the two Wilcoxon rank-sum tests. Of the 12 complexes for which the starting positions surrounding the binding site scored significantly greater than those away from the binding site, the correct docked pose was found more frequently during the filtered runs than the unfiltered in all cases. Of the 11 complexes

which did not exhibit a significantly enhanced or diminished binding site ligand density, one was found an equal number of times, five docked more frequently and five docked less frequently. For the three complexes for which the binding site had unfavourable ligand density, SwarmDock found the binding site less frequently. For two of these (1I9R and 1EAW), using the filtered starting positions resulted in no correctly docked solutions. In conclusion, information gleaned from simulations in which external crowding molecules are not included either enhance the performance of SwarmDock or at least do not depreciate it in most cases. Performing such a BioSimz simulation prior to docking takes around 4h of computing time using a typical 8 core CPU.

4.3.2 Crowded Simulations

The dynamics simulations and statistical tests were repeated on the test set in the presence of the external crowder molecules, and results are shown in Table 4.3. For eight of the complexes, higher ligand occupancy was observed around the binding site irrespective of whether crowders are included in the simulation or not (1GRN, 1FSK, 1FQJ, 1BUH, 1EWY, 1KAC, 1E6E, and 1KTZ). For six of the complexes, the inclusion of crowder molecules enriched the ligand density around the binding site relative to the uncrowded simulations (1HE8, 1GCQ, 1QA9, 1JPS, 1ML0, and 1NCA), while for six others, crowders diminished the signal (1AVX, 1VFB, 1AY7, 1AHW, 1PPE and 7CEI). The complexes 1NSN and 1QFW did not have enhanced specific binding, irrespective of their environment. For four of complexes (1I9R, 1TMQ, 1EAW and 1HIA), the binding site is consistently visited less frequently than non-binding regions in both sets of simulations.

As the scoring scheme used to rank the starting positions by the ligand density surrounding them was chosen arbitrarily, a number of different metrics were devised. These schemes, and their corresponding p-values are shown in the appendix (Table D.1 and Table D.2). For all metrics, a greater or equal number of cases had a significantly enhanced binding region for the crowded simulations compared to the uncrowded simulations ($p \leq 0.05$). Similarly, for all metrics, an equal number of cases or fewer

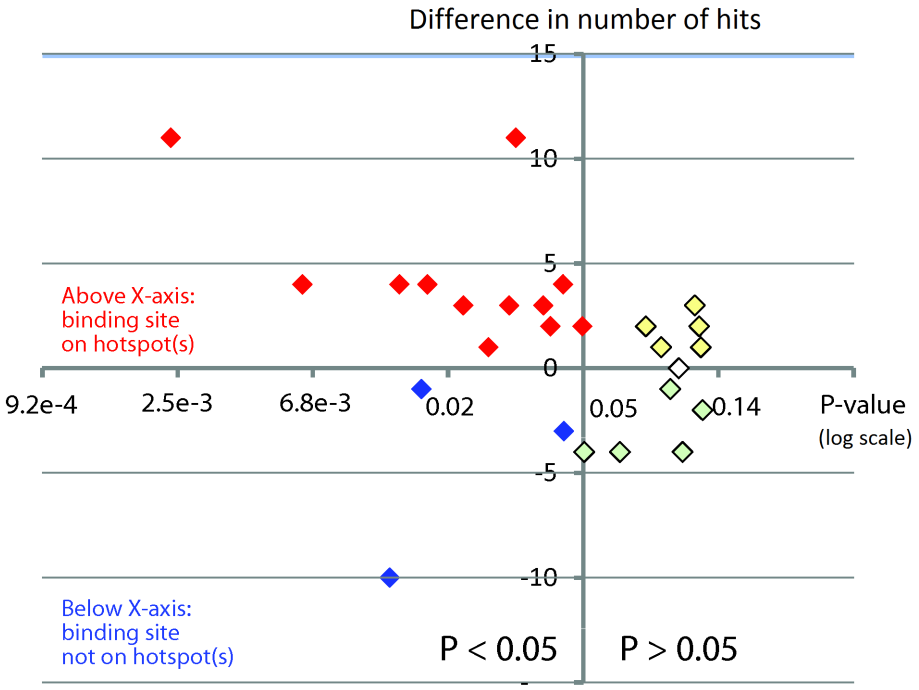

Figure 4.5: Difference in hits between filtered and unfiltered runs (y-axis) versus minimum Wilcoxon rank-sum test p-value, shown on a logarithmic scale (x-axis). SwarmDock performance is enhanced or diminished as expected depending on whether the binding site corresponds to ligand density hotspots; all the complexes which exhibit a significant signal around the binding site (red) are found more frequently after filtering while those which have a diminished ligand density around the binding site (blue) are found less frequently. However, there is not sufficient data to show that the separation of data is significant using a chi-square test (with Yates' correction for a 2×2 matrix).

Table 4.3: P Values for crowded simulations using one-tailed Wilcoxon rank-sum test, testing whether the scores of the 10 starting positions nearest the ligand centre of mass are greater in value than the remaining points (Pg) or lesser in value (Pl). Significant results are highlighted in bold.

Complex	Pg	Pl	Complex	Pg	Pl
1GRN	**0.039**	0.961	1BUH	**0.037**	0.964
1HE8	**0.016**	0.984	1NCA	0.138	0.864
1QFW	0.169	0.834	1JPS	**0.050**	0.951
1GCQ	**0.005**	0.995	1AVX	0.842	0.161
1AY7	0.366	0.638	1AHW	0.150	0.852
1KTZ	**0.050**	0.951	1EWY	**0.005**	0.995
1VFB	0.491	0.513	1EAW	0.990	**0.011**
7CEI	0.386	0.618	1KAC	**0.000**	1.000
1QA9	**0.006**	0.995	1PPE	0.557	0.447
1FSK	**0.047**	0.954	1NSN	0.232	0.771
1FQJ	**0.003**	0.997	1TMQ	0.859	0.143
1I9R	0.995	**0.006**	1HIA	0.810	0.193
1ML0	**0.041**	0.960	1E6E	**0.041**	0.960

had a significantly diminished binding region for the crowded simulation compared to the uncrowded simulations. These results suggest that the inclusion of external crowder molecules offers a small improvement over the simulations in which they are omitted. Of greater interest, however, are the different biases for biological types observed in the crowded and uncrowded simulations. No particular pattern is observed for the antibody-antigen complexes; one performed well in both the crowded and uncrowded simulations, two performed better in the crowded simulations, two performed better in the uncrowded simulations, one performed consistently badly and two showed no particular enrichment of the binding site in either simulation. Both enzyme-substrate complexes, however, performed consistently well in both the crowded and uncrowded simulations, unlike the seven enzyme-inhibitor complexes, none of which performed consistently well. Four of these performed well in uncrowded but not the crowded simulations, and three of them performed consistently badly. The best results, however, were observed for the signal/effector-receptor complexes. None of these performed badly in both sets of simulations and none of them performed better in the uncrowded simulations compared to the crowded simulations. Four performed better in the crowed simulations and five performed well in both simulations.

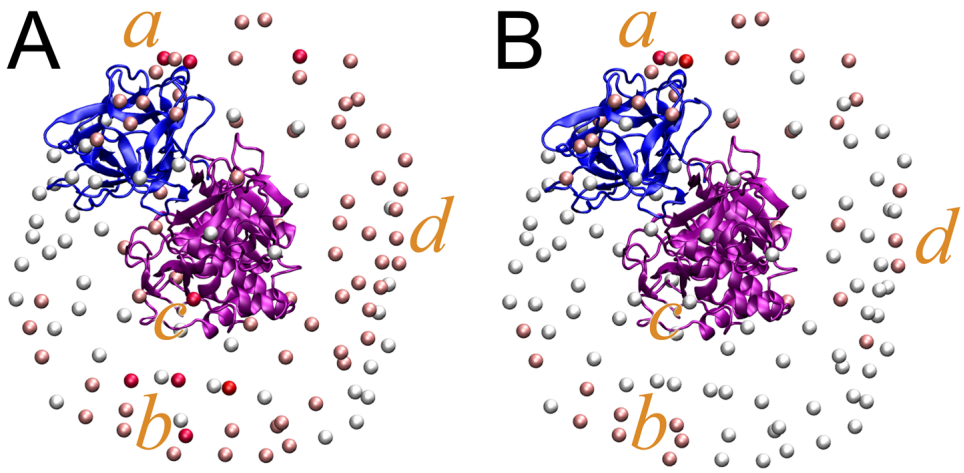

Figure 4.6: The SwarmDock starting positions for CAPRI target 32 (PDBid 3BX1), coloured by their score, for (A) uncrowded and (B) crowded simulations. Image adapted from Li *et al.* (2010b) with permission from Wiley.

This result is of particular interest, as signal transduction complexes are seen as difficult docking targets. They are typically of weaker affinity and often have multiple binding sites. Enzyme-inhibitor complexes, in contrast, are easier targets, being of high affinity, fast binding kinetics, high specificity and often irreversible in nature. The reasons why this particular pattern is observed is not clear, and the mechanistic rationale behind macromolecular crowding effects have not yet been fully elucidated. However, both positive and negative effects on binding kinetics have been observed in crowded simulations (Zhou *et al.*, 2008). One possibility which is consistent with the presented results is that crowders act as ambient momentum providers, allowing the weaker-binding complexes to overcome energy barriers and arrive at the bound state. The more tightly bound complexes, however, are more likely to be found in the bound state, and the additional momentum provided could act to destabilise the complex.

4.3.3 CAPRI Targets

Simulation and statistical tests were performed on CAPRI targets 32 and 38-40, both with and without environmental crowder proteins. Target 32 is

the complex between the enzyme subtilisin savinase and its barley inhibitor (PDBid 3BX1). Unlike the seven enzyme-inhibitor complexes above, this complex performed better in the crowded simulations ($P = 0.025$) compared to the uncrowded simulations ($P = 0.354$). The SwarmDock starting positions, coloured by score, are shown in Figure 4.6. In both simulations, the correct binding site was found. However, without the crowders, three false positive binding sites, *b*, *c* and *d*, were also prominent, one of which exhibited a greater signal than the true binding site, *a*. When the crowding proteins were included, however, one of the false positive sites, *c*, disappears altogether, whilst the signals for *b* and *d* were significantly attenuated, leaving a single prominent signal near the correct binding region.

Similar results were found when the method was applied to target 38 (PDBid 3FM8), a signal-effector complex between centaurin-α1 and the FHA domain of kinesin family member 13B (KIF13B). As the unbound structure of KIF13B was not available, it was built by homology modelling using the POPULUS server (Offman *et al.*, 2008). For this complex, crowded simulations ($P = 0.057$) performed significantly better than in the uncrowded simulations ($P = 0.388$). Target 39 was the same target as target 38, only with bound conformation of the KIF13B FHA domain, and the results were essentially identical.

Target 40 was a bivalent complex between the double-headed arrowhead protease inhibitor and two trypsins (PDBid 3E8L). For one site, p-values of 0.016 and 0.003 were obtained using the uncrowded and crowded simulations respectively. P-values for the second site were 0.159 and 0.169. The SwarmDock starting positions, coloured by score, are shown in Figure 4.7.

4.3.4 Possible Mechanistic Insights

During the above analysis it was observed that for many of the complexes, the high-scoring starting positions were not directly above the binding site but, to a greater or lesser degree, proximal to it (see Figures 4.6 and 4.7). This phenomenon was also observed in the contact heat maps. Intriguingly, for three cases (1BUH, 1QA9 and CAPRI target 39), patches are observed on the ligand and the receptor such that the two patches could be matched

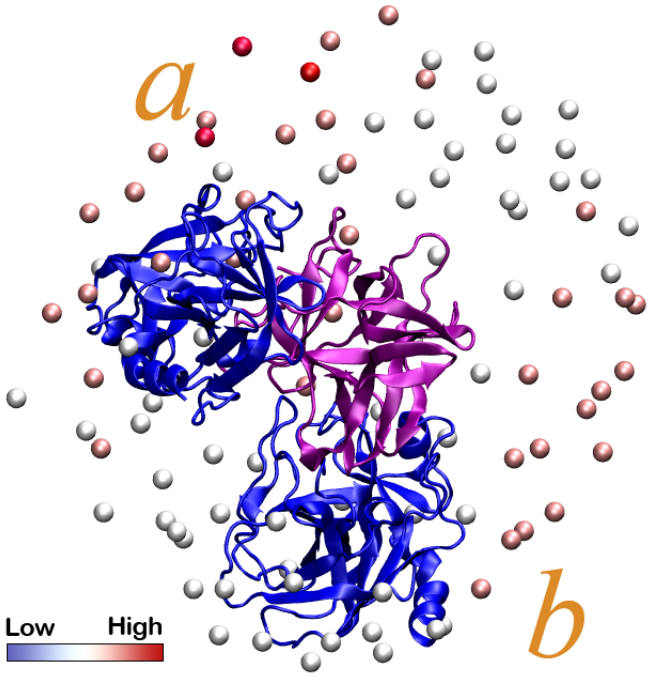

Figure 4.7: The SwarmDock starting positions for CAPRI target 40 (PDBid 3BX1), coloured by their score, for the crowded simulations. The ligands for both binding sites are shown. Image adapted from Li *et al.* (2010b) with permission from Wiley.

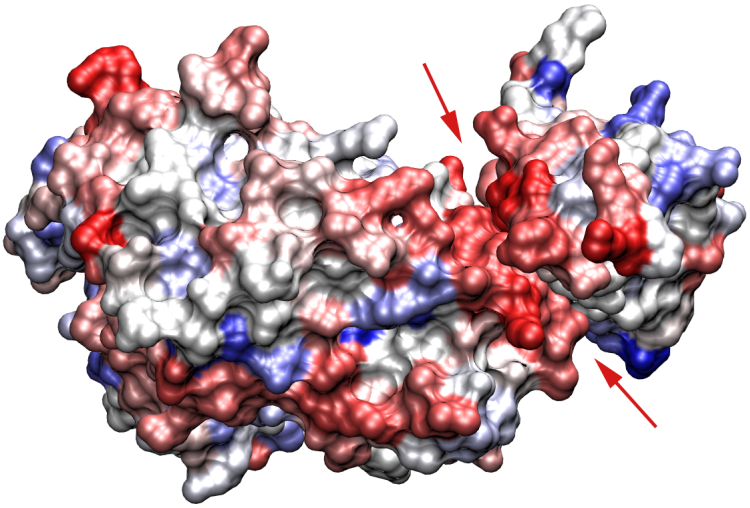

Figure 4.8: The contact heatmap for the CDK2-Ckshs1 complex (PDBid 1BUH), constructed from the crowded simulation. Although hotspots are not located across the whole crystallographic binding site, numerous residues of frequent contacts form patches adjacent to and overlapping with the crystallographic site.

up by a concerted rotation from the crystallographic binding site, indicating that the site of preferential encounter complex formation can proceed to the final docked pose via the rolling of the binding partners across each others surfaces. For the complex between CDK2 and Ckshs1 shown in Figure 4.8 (PDBid 1BUH), the proximal patches are only slightly shifted and overlap the crystallographic site. For target 39, however, the patches do not form part of the crystallographic binding site, but are located directly opposite from one another, as shown in Figure 4.9. These results are consistent with the encounter complex formation/2D surface search mechanism laid fourth in the introduction to this chapter, and described by Blundell and Fernandez-Recio (2006).

4.4 Discussion

BioSimz and SwarmDock were combined into a simulation and docking protocol. Of the complexes tested, around half exhibited preferential encounter complex formation near the binding site region, and for all of these, taking account of this information to restrict the SwarmDock search space resulted in improved docking performance. For those which did not exhibit

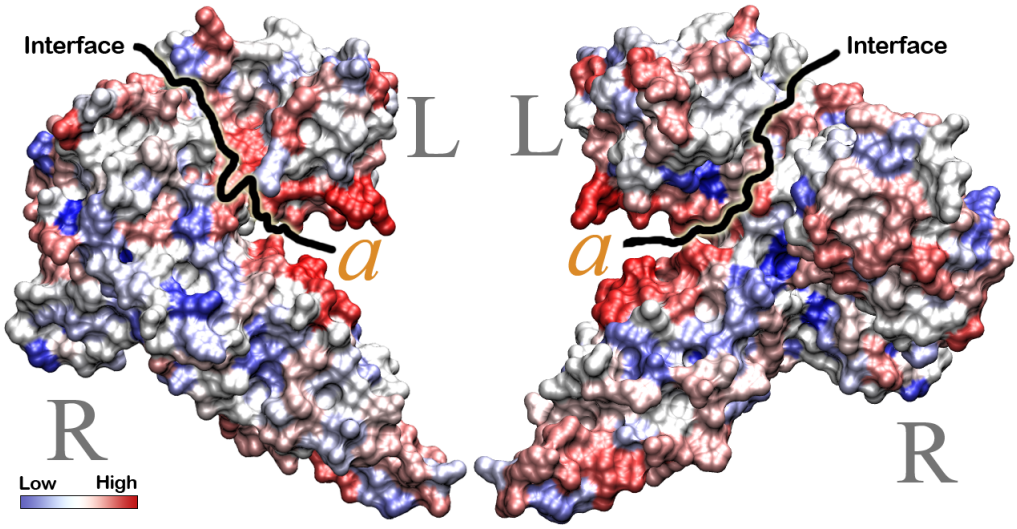

Figure 4.9: The contact heatmap for CAPRI target 39, constructed from the crowded simulation. Neither the crystallographic contacts nor sites distal to the binding site are noteworthy. However, a very strong signal can be observed on opposing patches on the ligand and receptor. Image adapted from Li *et al.* (2010b) with permission from Wiley.

preferential visitation near the binding site, the site was not disfavoured and the restriction of search space did not adversely affect docking performance, except in a minority of cases. The inclusion of external crowding molecules in the simulations aids in the location of the binding site in some cases, and reduces the signal in others. However, averaged over the whole data set, the inclusion of these molecules does seem to facillitate binding site recognition. Intriguingly, the crowders were found to attenuate the signal for enzyme-inhibitor complexes, but enhance the signal for signal/effector-receptor complexes, although the reasons for this are unknown. These simulations also appear to give mechanistic insights into the transition from unbound binding partners to the final docked complex, and complement kinetic and NMR data which suggest that biomolecular recognition proceeds via biased encounter complex formation and surface search.

Chapter 5

CAPRI

5.1 Introduction

SwarmDock was used to model complexes in the CAPRI experiment from round 15 (see section 1.5.10) and has performed respectably, coming joint 11th out of the 54 groups which participated during the last assessment period (Lensink and Wodak, 2010b). For rounds 16-17, the method previously employed in the laboratory, developed by Dr. Marcin Król, was used in tandem with SwarmDock (Krol *et al.*, 2007a). Briefly, molecular dynamics was performed on both the receptor and the ligand. These structures were clustered at a resolution which generated approximately 10 clusters each. These were then crossdocked and the results amalgamated (see section 1.5.6). Following a short minimisation in CHARMM, the structures were ranked by electrostatics and the DComplex pair potential. The top 500 structures according to electrostatics, and the top 300 structures according to DComplex, were then clustered. The lowest energy members of the largest clusters are then selected.

In the scoring part, the approach employed vaguely follows the concept of 'Crowd Intelligence' (Surowiecki, 2005), whereby the results from all competitors are amalgamated, filtered and clustered in order to find a consensus structure. Initially, all the submitted structures are minimised and ranked using DComplex. Subsequently, the top 300 structures are clustered hierarchically. The lowest energy structures from the largest clusters are then selected for submission. This approach has proven highly

fruitful, as during the last assessment period no other participating group managed to correctly identify the structure of more targets (see Lensink and Wodak (2010b), supplementary table S13).

Two very different challenges were also recently posed to the CAPRI community, in which the ability to discriminate between pairs of proteins which bind and those which do not, was tested. The following sections describe each round individually. To summarise, Table 5.1 shows the performance of SwarmDock in the prediction rounds of CAPRI compared to the other participating groups for targets 37 to 42. An acceptable or better structure was found for five of these targets, which is equal to or better than all but two other groups, those of Zho and Zacharias. Despite frequently locating the correct structure, only one of these was classified as good accuracy, and two as medium accuracy. The scoring rounds are summarised in Table 5.2.

5.2 Standard Rounds

5.2.1 Round 15

The first round of CAPRI in which SwarmDock was entered was round 15. Just seven months into the postgraduate course, the algorithm was in a primitive state; no flexibility had been included, the deleterious EEF1 desolvation term was included, and no proper benchmarking had yet been undertaken. Considering that this early set-up is incapable of consistently docking proteins already in their bound conformation (see section 3.2.6.2), it is no surprise that we failed to generate any acceptable structures. Even for the relatively easy target 32 (the Subtilisin Savinase enzyme/BASI inhibitor complex, PDBid 3BX1), for which acceptable solutions or better were submitted by 11 groups (5 of which were high accuracy), retroactive analysis showed that SwarmDock failed to even generate an acceptable structure when docking the Savinase (unbound, PDBid 1SVN) to BASI (unbound, PDBid 1AVA).

Targets 33 and 34 were for the same complex, the interaction between a homology model of Rlma(II) methyltransferase and a portion of bacterial

Table 5.1: A summary of the performance of the participating groups in CAPRI targets 37 to 42. Predictions are categorised as incorrect (0), not participated (-), acceptable (*), medium (**) and good (***) as per the CAPRI criteria (see section 1.5.10 or Mendez *et al.* (2003)). Servers are highlighted in capital letters. A summary is given for each participant, showing the number of complexes found, with the number of medium and high accuracy solutions shown in parentheses. Table adapted from Lensink and Wodak (2010b).

Predictor	T37.1	T37.2	T38	T39	T40A	T40B	T41	T42A	T42B	Summary
Zacharias	**	*	0	0	***	***	***	***	0	6(4***/1**)
Zou	**	*	0	0	***	***	***	0	***	6(4***/1**)
Vajda	0	0	0	***	***	***	**	***	0	5(4***/1**)
Wolfson	*	0	0	0	***	***	**	***	0	5(3***/1**)
Weng	0	***	0	0	**	***	**	0	**	5(2***/3**)
Bonvin	0	**	0	0	***	***	**	*	0	5(2***/2**)
CLUSPRO	0	0	0	**	**	*	**	***	0	5(1***/3**)
Bates	*	*	**0**	**0**	**	***	**	**0**	**0**	**5(1***/2**)**
Eisenstein	0	0	0	0	***	***	***	0	***	4(4***/0**)
Nakamura	-	-	0	0	***	***	**	***	0	4(3***/1**)
Wang	0	0	0	**	***	***	*	0	0	4(2***/1**)
Zhou	0	*	0	0	***	0	***	0	0	3(2***/0**)
Gray	0	0	0	0	***	***	**	0	0	3(2***/1**)
Camacho	0	*	0	0	**	***	-	-	-	3(1***/1**)
HADDOCK	0	0	0	0	0	***	**	0	*	3(1***/1**)
Ritchie	0	0	0	0	**	**	**	0	0	3(0***/3**)
GRAMM-X	0	0	0	0	***	***	***	0	0	3(3***/0**)
Takeda-Shitaka	0	0	0	0	***	**	***	0	0	3(2***/1**)
Xiao	0	0	-	-	*	***	***	0	0	3(2***/0**)
Vakser	0	0	0	0	***	***	**	0	0	3(2***/1**)
SKE-DOCK	*	*	0	0	0	***	0	0	0	3(1***/0**)
Fernandez-Recio	0	0	0	0	0	0	*	**	0	2(0***/1**)
Ten_Eyck	0	0	0	0	***	***	-	-	-	2(2***/0**)
Tovchigrechko	0	0	0	-	***	***	0	0	0	2(2***/0**)
F_Jiang	0	0	0	0	0	*	*	0	0	2(0***/0**)
Comeau	-	-	-	-	***	***	0	0	0	2(2***/0**)
Elofsson	**	**	0	0	0	0	0	0	0	2(0***/2**)
Baker	0	0	0	0	***	0	0	0	0	1(1***/0**)
Mitchell	0	0	0	0	0	0	0	**	0	1(0***/1**)
PATCHDOCK	0	0	0	0	0	0	0	***	0	1(1***/0**)
FIBERDOCK	-	-	-	-	-	-	***	0	0	1(1***/0**)
FIREDOCK	0	0	0	0	***	0	-	-	-	1(1***/0**)
Alexov	0	0	0	0	0	0	**	0	0	1(0***/1**)
Bajaj	-	-	-	-	-	-	**	0	0	1(0***/1**)
TOPDOWN	0	0	0	-	0	**	0	0	0	1(0***/1**)
Elber	-	-	0	0	0	*	0	0	0	1(0***/0**)
Günther	0	0	0	-	-	-	*	0	0	1(0***/0**)
Kihara	0	0	-	-	0	0	*	0	0	1(0***/0**)
Kinoshita	*	0	-	-	-	-	-	-	-	1(0***/0**)

ribsomal RNA, the first protein-RNA complex to be modelled in CAPRI. As SwarmDock was designed to read in CHARMM parameter files, only small changes needed to be made in order to use the CHARMM-27 force field, which includes nucleotide parameters. For target 33, Rlma(II) methyltransferase was to be docked to a homology model the rRNA, and in target 34 the bound structure of the rRNA was provided. The Rlma(II) homology model was built using POPULUS (Offman *et al.*, 2008), and the rRNA was built from PDB structure 1MT4 by nucleotide replacement and 'stitching' helical

Table 5.2: A summary of the performance of the participating groups in CAPRI targets 37 to 42. Scores are categorised as incorrect (0), not participated (-), acceptable (*), medium (**) and good (***) as per the CAPRI criteria (see section 1.5.10 or Mendez *et al.* (2003)). A summary is given for each participant, showing the number of complexes found, with the number of medium and high accuracy solutions shown in parentheses. Table adapted from Lensink and Wodak (2010b).

Scorer	T37A	T37B	T38	T39	T40A	T40B	T41	Summary
Bates	***	**	**0**	**0**	**	**	*	**5(1***/3**)**
Wang	**	**	0	0	**	***	*	5(1***/3**)
Bonvin	**	*	0	0	**	***	*	5(1***/2**)
Zou	***	0	0	0	***	***	***	4(4***/0**)
Haliloglu	**	**	0	0	**	***	0	4(1***/3**)
Weng	***	0	0	0	***	0	*	3(2***/0**)
Wolfson	*	0	0	0	***	***	0	3(2***/0**)
Elber	-	-	0	0	***	***	*	3(2***/0**)
Camacho	0	0	-	-	***	***	-	2(2***/0**)
Takeda-Shitaka	0	0	0	0	***	0	**	2(1***/1**)
Liu	-	-	-	-	**	**	-	2(0***/2**)
Aze	***	0	0	0	0	0	0	1(1***/0**)
Fernandez-Recio	0	0	0	0	0	0	**	1(0***/1**)
Kihara	-	-	-	-	0	0	**	1(0***/1**)
SAMSON+HEX	-	-	-	-	-	-	**	1(0***/1**)
Xiao	-	-	-	-	-	-	**	1(0***/1**)
Mitchell	0	*	0	0	0	0	0	1(0***/0**)

RNA generated using CHARMM.

For target 33, molecular dynamics was run on the built RNA structure, and the resultant trajectory was clustered in order to generate a diverse conformational ensemble against which to dock the Rlma(II). Having six backbone degrees of freedom per nucleotide, RNA is intrinsically more flexible than protein and none of the generated rRNA structures resembled the bound conformation and no accurate structures were generated. None of the other CAPRI participants submitted an acceptable structure or higher quality solution.

For target 34, with the bound conformation of rRNA, we also failed to generate an acceptable structure. However, 13 other groups did manage to find acceptable solutions. For the scoring of this target, only the CHARMM energy was taken into account for selecting structures for clustering, as DComplex cannot score nucleotides. One large, low energy cluster was found. This cluster was split into sub-clusters by clustering at lower resolution. Of the structures submitted for scoring, nine were of acceptable quality.

Targets 35 and 36 were not a protein-protein complex, but two domains of the same protein, xylanase Xyn10B, connected by a disordered linker (PDBid 2W5F). Both domains had to be modelled by homology for target

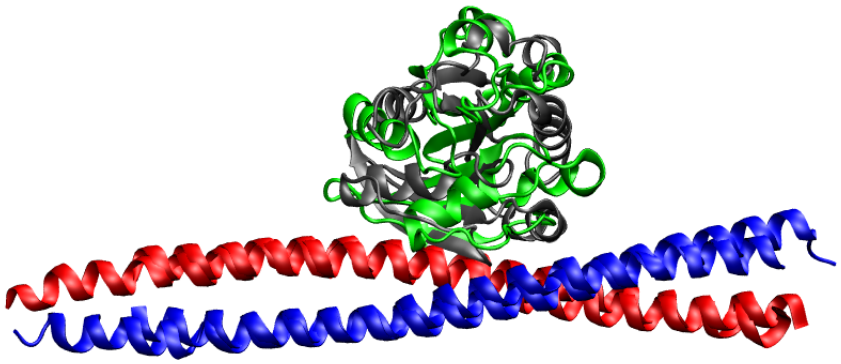

Figure 5.1: The crystal structure of CAPRI target 37 (PDBid 2W5F). The best submitted SwarmDock model is also shown, superimposed on the leucine zipper, with the ARF6 in green, IRMSD = 2.93Å, fnat = 0.31.

35, and the bound structure for the CBM22 domain was given for target 36. As this is an intramolecular interaction rather than intermolecular, in order to adopt the crystallographic structure the interaction energy between the two domains need not be sufficiently low as to overcome the entropic penalty associated with the loss of three translational degrees of freedom. Hence, this was considered a difficult target, compounded by the fact that homology models had to be used. Only one acceptable solution was submitted for target 35 by any of the participants, and only one for target 36.

5.2.2 Round 16

By the time round 16 came around, normal mode flexibility had been added to SwarmDock and the EEF1 term had been removed from the energy function. Only one target, 37 (PDBid 2W5F), was included in this round, a C_2 symmetrical interaction between two ARF6 proteins (unbound, PDBid 2A5D) and a symmetrical homodimer, the second leucine zipper of JIP4 (built by homology). Four acceptable solutions were submitted, derived as follows, the best of which is shown in Figure 5.1. Nine other participating groups also managed to identify correctly the structure of the complex.

Molecular dynamics was performed on both binding partners, and the trajectories clustered. The derived structures were used for crossdocking

both with SwarmDock and with FTDock. For each of the pairwise SwarmDock runs, clusters of low energy and/or large size were amalgamated. Of these structures, a number of similar poses clustered together. The same cluster was found after collecting and clustering the low energy FTDock crossdocked structures. We submitted five SwarmDock and four FTDock structures, including two of acceptable accuracy from each method. The final structure we submitted, which was incorrect, was constructed by superimposition of ARF6 and the leucine zipper onto their respective homologues, RhoA and ROCKI in the PDB structure 1S1C. In the scoring round, six correct structures were submitted, including one of high accuracy.

5.2.3 Round 17

Round 17 was comprised of two targets, both the complex between centaurin-α1 and the KIF13B FHA domain (PDBid 3FM8). In target 38, the FHA domain was constructed by homology modelling, and the bound conformation was given for target 39. Pairwise and crossdocking was performed with SwarmDock and FTDock. No correct solutions were submitted for target 38 by any of the participating groups, and only two groups submitted correct solutions for target 39. Retroactive analysis of target 39 showed that SwarmDock had found a medium accuracy structure (IRMSD = 1.67Å, LRMSD = 3.55Å). However, it only ranked 102nd, highlighting the need for a refinement and/or re-ranking stage to be implemented in SwarmDock. No groups managed to identify correctly the complex in the scoring rounds for either target.

5.2.4 Round 18

Round 18 also consisted of a single bivalent complex, target 40 (PDBid 3E8L). This target contained the double-headed arrowhead protease inhibitor A in the bound conformation, bound to two trypsin molecules (unbound, PDBid 1BTY). During this round, Dr. Zhiping Weng discovered the identity of two residues involved in binding, one from each site. This information was passed on to the other participants, although for the primary site this information was redundant due to an obvious protease binding motif. High-ranking and clustering solutions were filtered according to the involvement

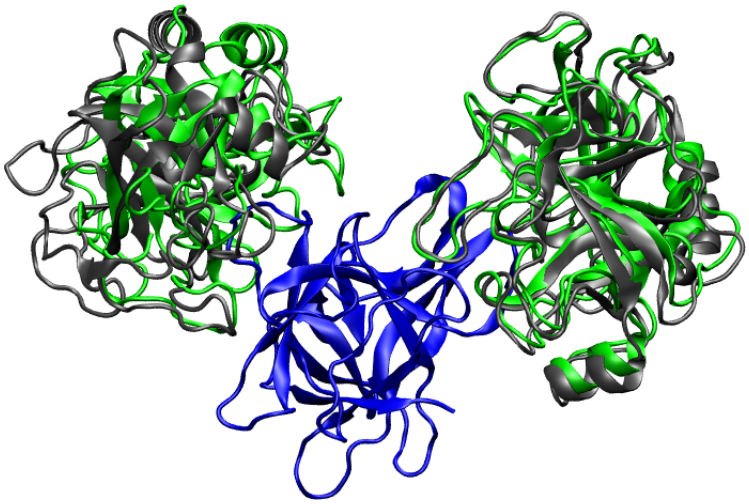

Figure 5.2: The crystal structure of CAPRI target 40 (PDBid 3E8L). The best submitted SwarmDock models for both sites are shown in green, after superimposition of the protease inhibitor. For the primary site (left), a medium solution is shown (IRMSD = 1.49Å, fnat = 0.81). For the secondary site (right), the high quality solution is shown (IRMSD = 0.94Å, fnat = 0.88).

of this binding motif, and molecular dynamics was performed on the best fitting structure. The trajectory was clustered, and the clusters ranked by binding energy. The nine lowest energy structures were submitted, four of which were of acceptable quality and five of medium quality. The tenth structure submitted was for the secondary site, being the only solution of reasonable energy found by SwarmDock for which the residue identified by Dr. Weng participated significantly in the interaction. This structure was of high quality. The best submitted structures are shown in Figure 5.2. Of the participating groups, 23 correctly modelled the binding site of at least one site. In the scoring round for this target, all ten submitted structures were of medium quality and covered both sites.

5.2.5 Round 19

Round 19 contained two targets. The first of these (PDBid 2WPT), is the complex between non-cognate mutants; Colicin E9 DNase (unbound, PDBid 1FSJ) and IM2 immunity protein (unbound, PDBid 2NO8). Wild-type cognate Colicin/IM interactions are of very high affinity, and a number of structures are available in the databank. SwarmDock managed to generate the correct structure a number of times, and comparisons with homologues

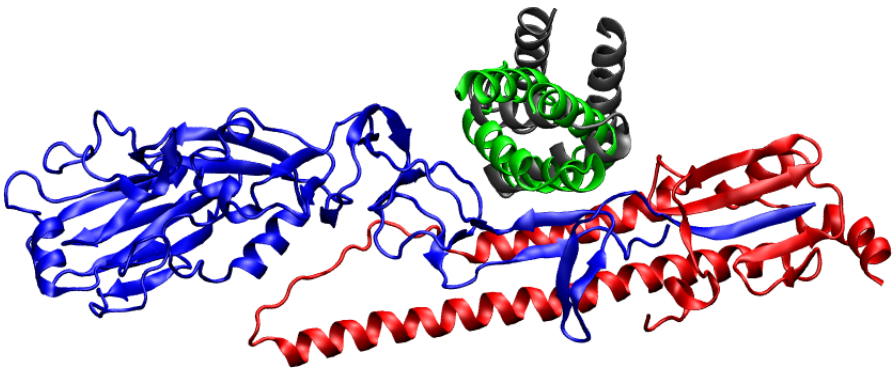

Figure 5.3: The crystal structure of CAPRI target 50 (PDBid 3R2X). The top ranked SwarmDock model is shown in green.

allowed identification and submission of five acceptable and one medium quality structure, along with 21 other groups with correct predictions. Four acceptable structures were submitted in the scoring round.

The second structure in round 19, target 42, was a synthetic homodimer of a tetratricopeptide repeat (PDBid 2WQH), where the monomer required homology modelling. In the crystal structure, two different homodimeric assemblies appear in the asymmetric unit, one with C_2 symmetry and the other with screw axis symmetry. Although 10 groups had successfully found one or the other, no group correctly found both structures. None of the submitted SwarmDock structures were of acceptable quality, although one submission was close (IRMSD = 5.3Å, Fnat = 0.12), and a refinement stage in the SwarmDock pipeline may have been able to nudge the structure towards the native.

5.2.6 Rounds 22 onwards

Rounds 20 and 21 were scoring only rounds, and are dealt with in section 5.3. Round 22 contained a single target, 46, the interaction between MTQ2 and TRM112 (PDBid 3Q87), which required the construction of homology models of both domains. Only the Bonvin group submitted correct solutions. Whilst none of the submitted SwarmDock structures were of acceptable quality, one was close (IRMSD = 4.68Å, Fnat = 0.15). In the scoring round, two acceptable structures were submitted. Again, a refinement protocol may have been able to refine in from this position.

Rounds 23 and 24 contained two and three targets respectively, but have not yet been assessed. Round 25 had one target, but was cancelled as its structure was found by the Bonvin group to be available on the internet. The structure of target 50, part of round 24, a *de novo* designed protein which binds to influenza hemagglutinin, has since been deposited in the protein databank (PDBid 3R2X). The top ranked SwarmDock model that was submitted is shown in Figure 5.3. Although the crystallographic complex is rotated relative to the SwarmDock prediction, the recognition helix was essentially replicated.

5.3 Interface Design Rounds

Rounds 20 and 21 of CAPRI were not typical rounds in which the crystal structure of a complex was to be predicted. Moreover, they were challenges commissioned by Sarel Fleishman and David Baker at the University of Washington, aimed at assessing the ability of the CAPRI community to recognise the features which identify experimentally verified interactions amidst ostensibly indistinguishable interactions for which experiments have failed to show binding.

The *de novo* design of protein-protein interactions is an important problem in structural bioinformatics. Attempts to design such interactions not only provide a test of our fundamental understanding of protein-protein interactions, but could also have considerable biomedical significance. Whilst proteins can be targeted specifically by antibodies, no technology yet exists to target specific surface patches. For instance, antibodies which target variable regions of pathogenic proteins may be capable of blocking the pathogen. However, the variability renders such antibodies prone to resistance, a phenomenon which is less problematic when targeting functionally conserved regions. Such an endeavour to design protein-protein interactions from scratch has been undertaken by Fleishman *et al.* (2011). In their approach, individual residues were docked in order to find hot spots. Following this, scaffold proteins are docked using PatchDock and refined using RosettaDock. Structures are ranked according to the degree to which their

residues match the positions and orientations of the hot spot residues. These hotspots are then grafted onto the scaffold and further refined. Those with the best energetics are selected for experimental testing. Their corresponding DNA constructs are then cloned, expressed and the interaction verified with SPR, yeast display or lysate ELISA.

In principle, with efficient sampling and an accurate energy function, this protocol should be able to identify the sequence of proteins which bind to specific surface patches of any target of interest. However, when tested, very few of the designs showed evidence of binding, despite the RosettaDock scoring function being unable to consistently distinguish between the designs and known protein-protein complexes on the basis of energetics. This indicates a disparity between reality and the free energy function; some physical process or phenomenon is not accounted for or is badly modelled, and this disparity is of sufficient magnitude to thwart the design process in all but a small minority of cases.

From this disparity were commissioned three challenges, two of which constituted CAPRI round 20, and the third of which formed round 21. The first challenge in round 20 was to distinguish the crystal structure of a true interaction from 20 designed interactions which failed to display binding function. The crystal structure was modified such as to mask any telling signs of its crystallographic origin, so that the answer was not known to the participants until the results were announced. In the second challenge of round 20, another set of 21 structures was given, all of which were designed. The participants were asked to predict which, if any, had shown experimental evidence of binding when tested. Again, the participants had no *a priori* knowledge about any of the structures. In these rounds, there was always the possibility that binding did occur, but that the affinity was below the detection threshold of around 7 kcal mol^{-1}. Further, absence of detectable binding may be due to poor monomer solubility, not unfavourable energetics.

The third challenge, round 21, was essentially a binary classification problem. Structures for 120 known protein-protein interactions, and for 87 designs, were given. The aim of the round was to distinguish between the structures by deriving a single metric with which to rank them. As a

condition of the challenge, no information pertaining to the 87 designed structures was to be used in the derivation of this metric. A manuscript outlining this work, including our contribution, is in preparation at the Baker laboratory.

5.3.1 Round 20

Table 5.3: The performance measures applied in CAPRI round 20, challenge one. Structures are ranked in order of preference. Metrics are the SwarmDock rank (SD rank), ligand RMSD (LRMSD), SwarmDock cluster size (Csize), whether the cluster corresponding to the given structure was the largest (Bclus), BioSimz association scores for the receptor (Ar rank) and ligand (Al rank), and the BioSimz dissociation score (D rank).

Rank	Model	SD rank	LRMSD	Csize	Bclus	Ar rank	Al rank	D rank
1)	10	1	1.046	58	Yes	3	1	17.0888
2)	4	1	1.083	19	No	0	0	18.0174
3)	6	1	0.354	74	Yes	3	1	15.632
4)	8	3	1.061	93	Yes	0	0	17.2262
5)	9	2	1.267	45	Yes	0	0	17.6508
6)	5	20	0.854	37	Yes	3	1	17.9287
7)	1	5	0.410	15	Yes	0	1	16.9812
8)	3	8	0.504	17	Yes	0	1	16.0287
9)	11	5	1.185	31	Yes	0	2	17.4597
10)	2	5	0.560	16	No	0	0	16.9510
11)	19	79	0.308	17	No	3	3	17.1189
12)	14	27	0.708	14	No	0	0	17.7264
13)	15	112	0.787	12	No	3	0	18.9913
14)	7	21	0.614	20	No	0	2	15.4912
15)	21	84	1.894	20	No	3	3	17.1113
16)	16	N/A	N/A	N/A	N/A	3	0	19.5187
17)	17	204	1.130	9	No	3	0	18.2994
18)	12	12	6.938	16	No	0	0	18.2452
19)	18	107	2.695	12	No	0	0	19.7383
20)	13	171	2.532	3	No	0	1	17.7149
21)	20	N/A	N/A	N/A	N/A	0	1	16.5006

Seeing as the first challenge in this round was to identify which complex corresponds to a crystal structure in the bound conformation, the approach which was taken was to derive metrics describing the interaction and compare these to a benchmark of known complexes. As bound-bound docking had been benchmarked already (see section 3.3.2), SwarmDock was used to re-dock the 21 structures and check whether the results deviated from those typical of bound-bound docking. After redocking globally, the generated structures were minimised, clustered and ranked. The results are shown in Table 5.3. All but five of the structures (4, 6, 8, 9 and 10) were discarded on the basis that they either did not dock or had an anomalous

ligand RMSD, rank or cluster size. At this stage, models 4, 6 and 10 were the most promising because, like the majority of bound-bound docking cases, the redocked structures ranked first. Models 6, 8, 9 and 10 had very large cluster sizes, larger than all the other clusters for those models, another good indicator of a bound complex. To distinguish between these models, we studied their association and dissociation dynamics using BioSimz. Firstly, the encounter complex formation process was studied using the method outlined in section 4.2.3. The degree to which the binding site received preferential visitation was determined by visual inspection of the scored SwarmDock starting positions, and rated on a scale of 1 to 3 for both the visitation of the receptor to the ligand site, and *vice versa*. This information was used to whittle the options down to either model 6 or model 10, both of which have favourable association dynamics. In order to discriminate between these two models, the dissociation dynamics was studied by performing rigid-body Langevin dynamics simulations starting from the bound conformation. Multiple runs were performed and the tendency for the complex to dissociate was measured (for a more detailed description, see Li (2011)). These simulations were also performed on known complexes from the docking benchmark, as a control. Of models 6 and 10, the former dissociated far quicker than the latter, and the bound-bound benchmark complexes. Hence, by a process of elimination, we chose model 10 as the most likely to be the crystal structure, on the basis that it is the only structure which measures as a typical true interaction by all metrics used. When the results for this challenge were released, not only was model 10 the correct answer, but only one other participating group managed to correctly identify it. This challenge further highlights the utility of combining BioSimz and SwarmDock.

For the second challenge in round 20, the same metrics were used on the 21 structures. The results for these appear in Table 5.4. When this challenge was commissioned, most of the proteins had not yet been assessed for binding affinity. Unfortunately, the bacterial system which was used failed to express a number of these complexes (1, 6, 7, 9-11, 13), and thus it is not known whether binding occurs between the two complex which performed most favourably according the metrics presented here. However, a weak

Table 5.4: The performance measures applied in CAPRI round 20, challenge two. Structures are ranked in order of preference. Metrics are the SwarmDock rank (SD rank), ligand RMSD (LRMSD), SwarmDock cluster size (Csize), whether the cluster corresponding to the given structure was the largest (Bclus), BioSimz association scores for the receptor (Ar rank) and ligand (Al rank), and the BioSimz dissociation score (D rank).

Rank	Model	SD rank	LRMSD	Csize	Bclus	Ar rank	Al rank	D rank
1)	1	1	1.167	69	Yes	0	0	16.3744
2)	6	32	1.104	36	Yes	3	2	19.7654
3)	8	15	0.394	15	Yes	0	0	19.1811
4)	4	25	0.797	22	No	0	0	16.9905
5)	5	23	1.230	22	Yes	0	0	18.5996
6)	7	86	1.335	14	No	0	0	16.7699
7)	3	150	0.655	20	No	0	2	17.1649
8)	18	121	0.475	12	No	2	1	17.6775
9)	2	115	4.856	12	No	3	0	18.5311
10)	13	55	0.974	6	No	1	0	17.2667
11)	15	107	1.663	7	No	1	3	18.3724
12)	20	165	0.762	9	Yes	1	0	17.382
13)	11	152	0.666	2	No	3	0	18.4147
14)	21	146	1.936	7	No	0	0	18.2067
15)	16	239	1.277	5	No	0	0	16.7891
16)	19	184	1.063	7	No	0	2	19.0746
17)	9	220	9.044	2	No	2	2	18.829
18)	14	107	5.653	4	No	0	0	19.3307
19)	10	90	7.630	1	No	0	0	18.2486
20)	12	N/A	N/A	N/A	N/A	1	0	19.2216
21)	17	N/A	N/A	N/A	N/A	0	0	17.5926

binding signal was detected for model 2. We did not rank this model favourably due to its poor SwarmDock rank and RMSD. The performance of the other participating groups has not yet been announced.

5.3.2 Round 21

The aim of round 21 was to derive a score capable of distinguishing 120 known protein-protein interactions from 87 failed designed interaction without incorporating information about the designed proteins in the derivation of the score. The approach used was as follows. Firstly, the structures were evaluated using a number of different molecular descriptors, described in section 5.3.2.1. In order to establish whether these descriptors contained sufficient information for distinguishing between the failed designs and the true interactions, both the failed and the true interactions were used to train support vector machines (see section 5.3.2.2). Cross-validation was used to minimise over-fitting the data, and a population based forward greedy feature selection algorithm was employed to select descriptor subsets capable of

achieving the discrimination. Once the ability to capture the differences was established, a number of empirical scoring functions were trained against experimental protein-protein binding affinities (section 5.3.2.3). These functions were applied to the whole set of failed designed and true interactions, giving good discrimination between the two sets. The binding affinity score was used to bin the complexes into one of five categories: 'binds', 'likely to bind', 'uncertain', 'likely not to bind', and 'does not bind'. This approach yielded one of the best performances in round 22, and we were offered the opportunity to validate our results. The Baker lab sent us nine new structures and asked us to select which, if any, exhibited binding function when tested experimentally. After applying our empirical scoring function, eight of the structures were found to be in the 'does not bind' or 'unlikely to bind' categories, and one structure was allocated to the 'binds' category. After providing our results to the Baker lab, Dr. Fleishman confirmed that the design we identified as a true interaction does indeed exhibit binding function experimentally, and that the eight designs we predicted as not interacting all failed to show evidence of binding. This work is of particular interest, as the experimental evaluation of the designs is a bottleneck in the *de novo* design process, requiring weeks of work to synthesise the DNA constructs, clone the genes, express them and test for binding function. The empirical scoring function presented here can be evaluated in a matter of hours, and the ability to discriminate promising designs from ones which will fail can vastly reduce the number of complexes requiring this experimental evaluation.

5.3.2.1 Molecular Descriptors

A large number of descriptors were derived. The complexes were re-docked with SwarmDock, and the interface RMSD, rank and cluster size were used as features. In addition, two binary descriptors were derived, first whether the top ranked structure had an interface RMSD below 5Å, and second whether the top cluster corresponded to the biggest cluster. These took the value of 1 if true, and 0 if false.

Also included were interaction energy features derived using the PyRosetta package (Chaudhury *et al.*, 2010), calculated as the energy of

the complex minus the energy of the individual binding partners. These included the repulsive and attractive Van der Waals components, the Rosetta desolvation energy, the hydrogen-bonding potential and all-atom statistical pair potentials. The coarse-grain Rosetta terms were also calculated; Van der Waals, residue pair and C_β potentials and the environment score. The total Rosetta energy was also decomposed into residue contributions, and each complex was also assigned a hot-spot count, defined as the number of residues with interaction energy below 1.5kcal mol^{-1}.

The change in solvent accessible surface area upon binding was evaluated with the NACCESS package (Hubbard and Thornton, 1993). Interface residues were categorised based on their degree of burial of solvent exposed surface area upon binding. Those with total burial are 'core', partial 'periphery' and the remaining 'unburied'. Total buried surface are was taken as a feature, as was the number of residues in each category, and the total Rosetta energy of residues in each category. The PyRosetta package was also used for re-docking using the default two stage RosettaDock protocol, in which a coarse-grained search is followed by an atomistic refinement (Chaudhury *et al.*, 2010). The number of correct redocking hits, as well as the lowest interface RMSD found and the rank, were included as descriptors.

Two descriptors used by London and Schueler-Furman (2008) were also included, both of which have shown use in discriminating between true docked solutions and false positives. The first of these is the difference in atomistic energy after RosettaDock coarse-grain docking, and the energy after refinement at atomic resolution. This was speculated to relate to the energy difference between the near-native encounter complex and the final docked solution, in which the interface is well packed. The second of these descriptors is the number of unsatisfied buried hydrogen bond donors or acceptors. The rationale behind this descriptor is that the energetic contribution of interfacial hydrogen bonds is negligible, as these bonds can be equally satisfied by solvent molecules when unbound. However, if a hydrogen bond is not satisfied upon burial, this corresponds to an energetic penalty.

In addition, following the observation by Smith *et al.* (2005) that the core and peripheral interface residues undergo different dynamics, three features were derived to characterise the flexibility of these residues. Elastic network normal mode analysis was applied to all the binding partners, and the ElNeMo tools suite was used to predict thermal B factors (Suhre and Sanejouand, 2004). The mean interface B factor, the mean B factor for core residues, and the mean B factor for peripheral residues were evaluated and taken as features.

As a crude estimate of the degree of interface packing, the number of atom contacts (atom pairs, one on each binding partner, within 5Å of each other) was calculated and taken as a descriptor, as was the ratio between this descriptor and the change in accessible surface area. The interface packing and surface complementarity scores developed by Mitra and Pal (2010) were also determined, as was the DComplex interaction potential (Liu *et al.*, 2004). BioSimz association and dissociation simulations were also undertaken. Association and dissociation scores were used as features, as were the p-values calculated as per section 4.2.3. Finally, a number of terms derived from CHARMM were included. These included the SASA implicit solvation energy (Ferrara *et al.*, 2002), the GBSW generalised Born non-polar and solvation energy, and the various components of the solvation energy using the analytic continuum electrostatics (ACE) model developed by Schaefer and Karplus (1996).

5.3.2.2 Feature Set Validation

A condition of the challenge was that no information regarding the failed designs be used in the training of the model. The approach to this problem was to derive a binding free energy score using the above descriptors, trained on experimental binding free energies for the true interactions. However, whilst a list of binding free energies for these complexes had been published shortly before round 21 was initiated (Kastritis and Bonvin, 2010), this list had not yet come to our attention. For this reason, we believed that the binding free energies would have to be manually amalgamated from the scientific literature, an endeavour which would require a significant amount

of work. As the success of our approach was inevitably predicated upon the descriptor set containing sufficient information for the discrimination of false designs from the true binding partners, an initial test of this was undertaken prior to assembling the list. This would not only demonstrate that the information required to discriminate between the two classes of structures resides within phenomena already studied, but most importantly, validate the above feature set and justify the effort of manually compiling a list of empirical binding free energies. To do this, a binary classification model was derived using a feature selection algorithm and an SVM with cross-validation. This model was quickly set up using the RapidMiner machine learning environment (Mierswa *et al.*, 2006).

In the machine learning routine used here, a feature selection algorithm is used to derive a feature subset. It is a population based algorithm, with a population of three; upon each iteration, three feature subsets are carried on to the next iteration. It is also a forward selection algorithm, in which the feature set grows by one feature each iteration. Further, it is a greedy algorithm, so that the three feature sets which are carried onto the next iteration are the three which give the highest classification accuracy when evaluated. For each evaluation, ten-fold cross-validation is employed. The data set is split into ten sets using stratified sampling and 90% of the data is used to train an SVM which is evaluated on the remaining 10%. This process is repeated ten times until every data point has been classified by an SVM for which it hadn't been used for trained. The results of these ten evaluations are then collected and the overall classification accuracy calculated. An analysis of variance (anova) kernel is used in the SVM.

Upon the first iteration of the feature selection algorithm, each feature is tested on its own. The three best features then form the three feature sets carried onto the next iteration. In the second iteration and all subsequent iterations, for each feature set in the population, each feature not in that set is evaluated with the features in that set. The three new sets which classify most accurately with the cross-validation are then carried forward to the next iteration. The algorithm continues up to three speculative rounds; should no feature set perform with higher accuracy than the best performing set previously evaluated, for three consecutive iterations, then

the algorithm terminates and the best performing set is returned.

When the above algorithm was applied to the bionomial classification challenge of round 21, a feature set was found which could classify the structure with an accuracy of 97.6% (two complexes were incorrectly classified as non-binding, three incorrectly classified true binders, and the remainder were correctly classified). When leave-one-out cross-validated, the accuracy dropped slightly, to 95.7%. The features selected were: SwarmDock cluster size, hot-spot count, attractive and repulsive VDW, the Rosetta C_β and atomistic pair potentials, the number of Rosetta redocking hits, the Rosetta redocking energy change, the number of unsatisfied buried hydrogen bond donors and acceptors, BioSimz p-values and association and dissociation scores, the interface packing score, electrostatics and the ACE interaction, solvation and Coulomb terms. Due to the kernel used, it is difficult to assess the relative importance of each feature. A number of other machine learning routines available in RapidMiner were also performed, including logistic regression, naïve Bayes, discriminant analysis, decision trees, neural networks and rule induction. These were combined with backward selection and evolutionary feature selection algorithms as well as the above forward selection algorithm, all with cross-validation. These methods also performed well, typically with over 80% classification accuracy.

This analysis confirmed that the information needed to discriminate between the failed designs and true interactions was within the derived feature set, and that this information was robust to the method used to extract it. Given this, we proceeded to manually collect empirical binding energies from the scientific literature.

5.3.2.3 Empirical Binding Score

Experimental binding affinities for 95 of the 120 protein-protein interactions were obtained by scouring the scientific literature. Every effort was made to ensure that the binding affinity corresponded as closely as possible to the structure. As the affinities were reported variously as binding free energies, dissociation constants, independent enthalpic and entropic contributions, or by their k_{on} and k_{off} rate constants, they were all converted to free energies using the equalities in equation 1.62. These

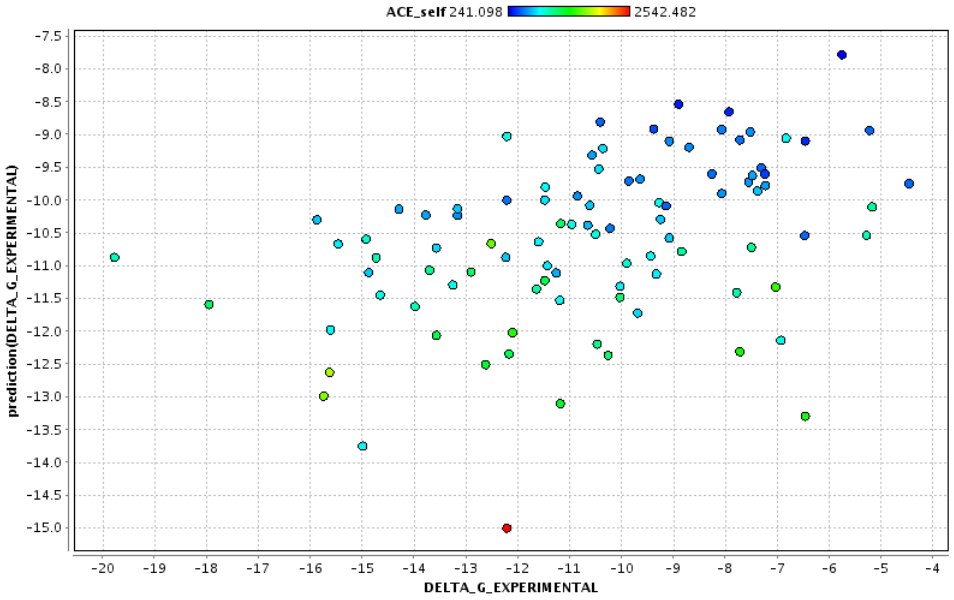

Figure 5.4: Predicted and Experimental binding free energies for the round 21 test cases. Data points are coloured by their ACE self electrostatics energies.

binding affinities were later expanded and checked in triplicate by us and our collaborators, and formed part of a binding affinity benchmark (see next chapter). Multiple regression was performed using feature subsets. Initially, the feature selection algorithms outlined above were used to explore descriptor subset space. However, it quickly became apparent that even with the cross-validation scheme employed, 95 data points were not sufficient to avoid overfitting. The feature space was evidently sufficiently large for ensembles of feature subsets to posses the property of being able to consistently reproduce test data when regressed against the noise in the training data, irrespective of how the data is split into test and training sets. Further, these subsets were being selected by the feature selection algorithms. Indeed, regressions with accuracy ostensibly greater than the experimental error associated with binding affinity measurements were found, even when the predictions were derived using leave-one-out cross-validation. For this reason, a change of tack was needed.

Instead, subsets were chosen on the basis of their physical plausibility and their ability to aid in the discrimination in round 20. A number of these subsets were tested, and when an appropriate binding energy threshold is

chosen, these were capable of distinguishing the failed designs from the true interaction with accuracies between 70% and 90%. The terms of one such subset is as follows, with absolute regression coefficients shown in parentheses. As the training features are normalised to z-scores prior to regression, these values reflect the relative contributions of each term to the final score:

1. Is SwarmDock top ranked structure under 5Å RMSD to bound? (0.076)
2. Does the biggest SwarmDock cluster correspond to the bound? (0.009)
3. Rosetta atomistic statistical pair potential (0.299)
4. Rosetta coarse-grained pair potential (0.122)
5. Van der Waals (0.192)
6. BioSimz association score (0.027)
7. interface packing (0.181)
8. surface complementarity (0.176)
9. ACE self energy (0.780)
10. GBSW solvation energy (0.071).

The first two terms are Boolean, and had been successfully applied to the first challenge of round 20, as had the BioSimz association score. The coarse and fine statistical pair potentials were used to account for low-resolution recognition factors, and to supplement the other atomistic terms respectively. The Van der Waals term was included due to its fundamental physical origin. The interface packing and surface complementarity scores were included to favour tightly interdigitated interfaces. The ACE self energy was included as it is the most accurate electrostatics descriptor; it treats atoms as a distribution of charge rather than collapsing the charge density onto the nuclei. The GBSW solvation term was included as this was regarded as the best, if most computationally demanding, solvation model in the feature

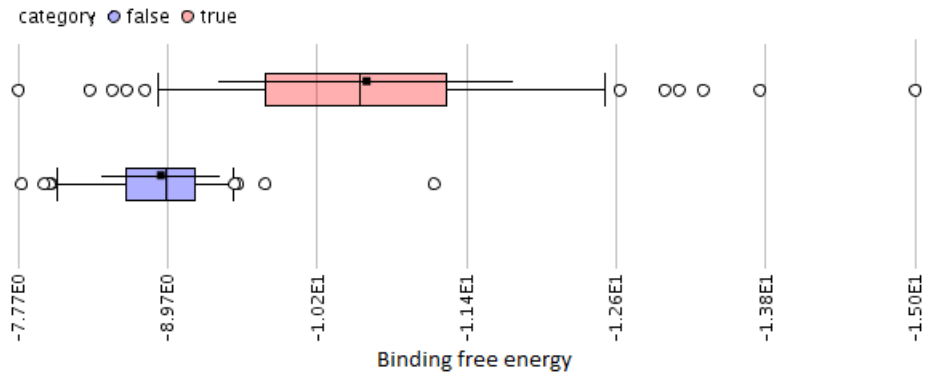

Figure 5.5: Box and whisker plots of the distribution of predicted binding free energy (kcal mol^{-1}) for the known interactors (true) and failed designs (false). The box shows the median and upper and lower quartiles, with whiskers showing the 5% and 95% percentiles. Mean and standard deviation are also shown.

set. Clearly the most significant contribution is given by the ACE self electrostatics term. Regression with this feature set yielded a binding energy function which reproduced the experimental data with an RMS error of 2.76 kcal mol^{-1} and a correlation of 0.414. The predicted binding free energies plotted against the experimental values are shown in Figure 5.4, with data points coloured by their ACE self energy. Interestingly, the greatest outlier, with anomalously favourable electrostatics, is 1BGX, a forged structure submitted by H. M. Krishna Murthy which should be removed from the protein databank. The only reasons that this structure was included in the analysis was because its dubious provenance was not know to us at the time.

Applying this score to all the structures yields the score distribution shown in Figure 5.5. The sensitivity of this score when applied to the round 21 binary classification problem is shown by the ROC curve in Figure 5.6, the area under which constitutes 91.2% of the plot. Using a threshold of -9.55 kcal mol^{-1}, this score can classify the complexes with 88.4% accuracy. Two of the designs are misclassified as true binders, and 22 of the true interactions misclassified as non-interacting, yielding 98.0% precision, 81.7% recall and specificity of 97.7%. As the Baker lab wished to have the results in a comparable format, the participating groups were requested not just to submit raw scores, but also to place all of the structures into one of the following five categories: 'binds', 'likely

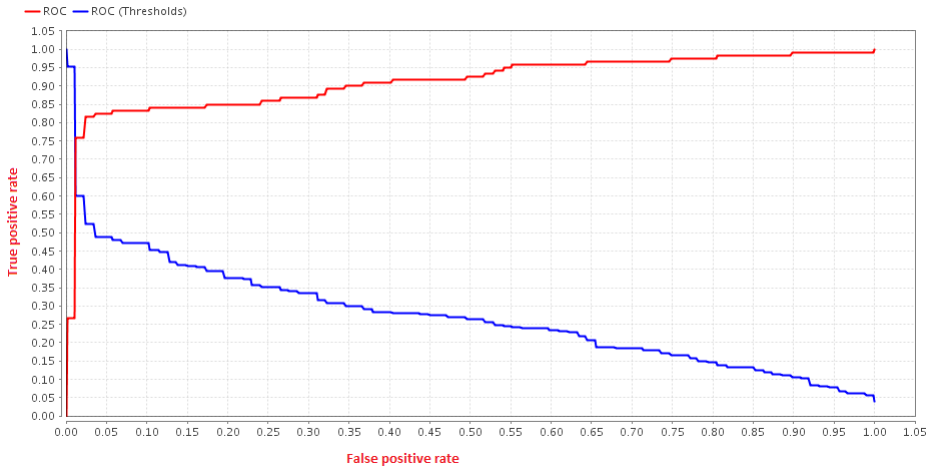

Figure 5.6: Receiver operating characteristic curve for the bionomical classification of true interactions and failed designs. The threshold, as a fraction of the range of the binding affinities, is also shown.

to bind', 'uncertain', 'likely not to bind' and 'does not bind'. To assign the structures, the threshold separating 'binds' from 'likely to bind' was set to 9.8 kcal mol^{-1}, and the thresholds between the higher categories at intervals of 0.2 kcal mol^{-1}. By this allocation, one of the designs was found to score highly enough to place it in the 'binds' category. However, no evidence for binding was found for this protein. Further, one of the designs was later found to show experimental evidence for binding, and this was not predicted by the above allocation. None of the other participating groups managed to identify this model as possessing binding function.

As the results presented above constituted one of the best sets submitted, Dr. Fleishman wished to verify the efficacy of the score in a blind test. After round 21 was closed, we were sent nine structures and asked to score them with our energy function. Seven of the structures were of poor score, landing firmly in the 'does not bind' category. One scored slightly better, and was allocated to the 'likely not to bind' category, and the final structure scored well, with -9.95 kcal mol^{-1}, placing it in the 'binds' category. After relaying these results, Dr. Fleishman informed us that upon experimental evaluation, the complex which was predicted to bind did show binding function, and the remaining eight complexes showed no evidence of binding.

5.4 Discussion

The CAPRI experiment has presented the structural bioinformatics community with important and challenging tasks over the last three years. This level playing field has allowed the testing of algorithms and a comparison of results. SwarmDock has been used in a number of rounds of CAPRI, and has performed reasonably well at predicting structures. A scoring scheme, loosely based on the 'Wisdom of Crowds' paradigm, employing hierarchical clustering, has performed very well at selecting the complex from the aggregated submissions of the participating groups. Two unusual scoring rounds were initiated with the aim of assessing the ability of the community to select *de novo* designed proteins that are likely to interact with their target proteins. For the first of these, a combination of SwarmDock and BioSimz characterisation allowed the correct identification of a true interaction amidst 20 structures which, upon testing, failed to show evidence of binding, a feat only achieved by one other participating group. In the second of these rounds, an empirically parametrised binding energy function was determined to classify 207 structures as true or false interactions with 88.4% accuracy. This method was validated by a blind test of nine structures, all of which were correctly classified.

Chapter 6

The Affinity Benchmark

6.1 Introduction

The current state of binding affinity prediction using empirical functions was outlined in section 1.4.4.8. It was noted how most studies to date have focussed on interactions involved in a limited set of biological functions, and between rigid proteins; when tested on more diverse data sets, these methods performed significantly worse. Many factors pertaining to biomolecular recognition have yet to be incorporated into empirical binding free energy functions. For instance, the effects of conformational changes upon binding, and linkage phenomena between affinity and allostery, pH, ionic environment and temperature. The framework in which the linkage of thermodynamic and environmental properties has already been formulated (Wyman and Gill, 1990; Woodbury, 2008), although much work remains to be done before this can be fully related to structure.

Before the subtle effects of thermodynamic linkage can be addressed and used in the annotation of protein-protein interaction networks, it is necessary to derive robust and accurate empirical binding free energy models for pairwise interactions. To do so requires extensive benchmarking on large and diverse protein sets. The construction of such an affinity benchmark, which was published in Kastritis *et al.* (2011), and a broad description of the complexes within it is outlined in section 6.2. A number of empirical free energy functions were tested on this benchmark and subsets of it, and the results of this are outlined in section 6.2.4.

6.2 The Affinity Benchmark

The first thing that must be done in order to benchmark and/or parametrise a binding affinity metric is to obtain a reliable data set of sufficient size. Whilst collecting binding affinity data for use in CAPRI round 21 (see section 5.3.2.3), the binding affinity benchmark of Kastritis and Bonvin (2010), which contained affinities for 81 interactions, came to our attention. As both this benchmark and the true interaction set of CAPRI round 21 were based on the docking benchmark 3.0 (Hwang *et al.*, 2008), there was considerable overlap between the set of 95 complexes gathered for round 21 and the data in the benchmark of Kastritis and Bonvin (2010). For these overlapping complexes, of which there were 70, the two data sets were plotted against each other. This revealed that whilst for most complexes the same or similar affinity had been recorded, for many of the data points there were discrepancies between our data sets, which for some complexes exceeded several orders of magnitude in binding affinity. Upon further investigation, it was found that mistakes had been made for around 40% of the data in Kastritis and Bonvin (2010). Upon contacting the Bonvin group and relaying concerns about the quality of the data, it was revealed that Prof. Joël Janin at Université Paris-Sud had also raised misgivings about the values presented, and the group subsequently prepared and published a *corrigendum*.

This highlighted the difficulty in procuring data from the literature. Many different methods of determining affinity are used, not all of which are applicable in all cases (Albeck and Schreiber, 1999). Results are presented in different units and scales. Some proteins have different names in different papers and some names correspond to multiple splice variants or are used for all the homologues in different species. For some complexes, a mutant form appears in the protein databank or was used for the affinity determination. Affinities are often dependent on the kinetic model used; for example, the complement C3/CR2 receptor complex fits well to a single-site model according to Guthridge *et al.* (2001), whilst Sarrias *et al.* (2001) claim that the data will only fit a bivalent binding model. Some complexes are also very sensitive to pH or ionic strength, such as the interaction between cytochrome c and its peroxidase (Kresheck *et al.*, 1995). Sometimes different

authors give widely different affinities, such as the interaction between ubiquitin and UCH-L3, for which Hirayama *et al.* (2007) report an affinity of tens of nanomolar, whilst Reyes-Turcu and Wilkinson (2009) claim the affinity is in the range of hundreds of micromolar. Sometimes a number of different constructs are used for the same protein (Eathiraj *et al.*, 2005), and absence or presence of cofactors can have a dramatic effect of binding affinities. For instance, the presence of GDP or GTP can reduce the affinity of Ran GTPase for RCC1 by six orders of magnitude (Klebe *et al.*, 1995). The degree to which comparisons between affinities determined at different temperatures is also questionable in some cases, such as for the interaction between the subunits of tryptophan synthase in the hyperthermophile *Pyrococcus furiosus*, which becomes more exothermic as the temperature increases (Ogasahara *et al.*, 2003). Even the buffer can have an effect, such as for the interaction between α-amylase and tendamistat, which has shown affinity differences when determined in the ionisable TRIS-HCl buffer when compared to non-ionisable HEPES (Piervincenzi and Chilkoti, 2004).

Whilst some lists of binding affinities had been previously published, these were either small (Brooijmans *et al.*, 2002; Krystek *et al.*, 1993; Xu *et al.*, 1997; Horton and Lewis, 1992; Gray *et al.*, 2003), had no primary references (Zhang *et al.*, 2005b), or had no structural cross-referencing (Stites, 1997). Further, these lists had only a single value for each complex, when often differing or corroborating values were also available. None of the lists made reference to structures of the binding partners in their unbound state, and no details of experimental conditions or the method used were included. At the time at which we contacted the Bonvin lab, a new docking benchmark, the benchmark 4.0, had been released by the group of Zhiping Weng (Hwang *et al.*, 2010). The Bonvin, Janin and Weng groups had decided to expand upon the affinity benchmark in Kastritis and Bonvin (2010), and welcomed us into their collaboration. Together, we collated the largest database of binding affinities to date. For each entry, all differing and corroborating values were added, as were the identities of cofactors, the pH, temperature, buffer conditions, experimental methods and further notes regarding enthalpies, entropies and rate constants, as well as cross-referencing with the protein databank entries for the complex and for the individual binding partners

in their unbound state. Each entry was vigilantly checked in triplicate, with care to ensure the maximal correspondence between the affinity and structure. After months of trawling the literature and reading hundreds of papers, the benchmark was finalised and appears in Table 6.1. Due to space limitations, the temperature, pH, buffer conditions, cofactors present in the structure and affinity determination, corroborating affinities and references, kinetic and thermodynamic data, the method of affinity determination and further notes are omitted from this table. This information can be found at the website for the affinity benchmark, `http://bmm.cancerresearchuk.org/~bmmadmin/Affinity/`.

Table 6.1: The binding affinity benchmark. †Personal communication.

Complex	Type	Protein 1	PDB1	Protein 2	PDB2	DASA (Å^2)	I-RMSD (Å)	Reference	K_D (M)	ΔG (kcal mol^{-1})
1A2K_C:AB	OG	Ran GTPase-GDP	1QG4_A	Nuclear transport factor 2	1OUN_AB	1603	1.11	Chaillan-Huntington *et al.* (2000)	1.50E-07	9.31
1ACB_E:I	EI	Chymotrypsin	4CHA_ABC	Eglin C	1EGL_A	1544	1.08	Ascenzi *et al.* (1988)	2.00E-10	13.05
1AHW_AB:C	A	Fab 5g9	1FGN_LH	Tissue factor	1TFH_A	1899	0.69	Huang *et al.* (1998)	3.40E-09	11.55
1AK4_A:D	OX	Cyclophilin	2CPL_A	HIV capsid	1E6J_P	1029	1.33	Yoo *et al.* (1997)	1.60E-05	6.43
1AKJ_AB:DE	OX	MHC class 1 HLA-A2	2CLR_DE	T-cell CD8 coreceptor	1CD8_AB	1995	1.14	Wyer *et al.* (1999)	1.20E-04	5.32
1ATN_A:D	OX	Actin	1JJJ_B	Dnase I	3DNI_A	1774	3.28	Mannherz *et al.* (1980)	2.00E-09	12.07
1AVX_A:B	EI	Porcine trypsin	1QQU_A	Soybean trypsin inhibitor	1BA7_B	1585	0.47	Lebowitz and Laskowski (1962)	4.80E-10	12.5
1AVZ_B:C	OX	HIV-1-NEF protein	1AVV_A	Fyn kinase SH3 domain	1FYN_A	1259	0.73	Arold *et al.* (1998)	1.58E-05	6.55
1AY7_A:B	EI	Rnase SA	1RGH_B	Barstar	1A19_B	1237	0.54	Hartley *et al.* (1996)	2.00E-10	13.23
1B6C_A:B	OX	FKBP binding protein	1D6O_A	TGFβ receptor	1IAS_A	1752	1.96	Huse *et al.* (2001)	2.80E-07	8.94
1BJ1_HL:VW	AB	Fab - vEGF	1BJ1_HL	vEGF	2VPF_GH	1731	0.5	Muller *et al.* (1998a)	3.40E-09	11.55
1BRS_A:D	EI	Barnase	1A2P_A	Barstar	1A19_B	1555	0.42	Hartley (1993)	2.00E-13	17.32
1BUH_A:B	EI	CDK2 kinase	1HCL_A	Ckshs1	1DKS_A	1324	0.75	Bourne *et al.* (1996)	7.70E-08	9.7
1BVK_DE:F	A	Fv Hulys11	1BVL_BA	HEW lysozyme	3LZT_A	1321	1.24	Foote and Winter (1992)	1.40E-08	10.53
1BVN_P:T	EI	α-amylase	1PIG_A	Tendamistat	1HOE_A	2222	0.87	Piervincenzi and Chilkoti (2004)	9.20E-12	15.06
1CBW_ABC:D	EI	Chymotrypsin	4CHA_ABC	BPTI	9PTI_A	1457	0.74	Castro and Anderson (1996)	1.10E-08	10.75
1DE4_AB:CF	OX	Hemochromatosis protein HFE	1A6Z_AB	Transferrin receptor ectodom.	1CX8_AB	2066	2.59	West *et al.* (2001)	6.80E-08	9.78
1DFJ_E:I	EI	Ribonuclease A	9RSA_B	Rnase inhibitor	2BNH_A	2582	1.02	Vicentini *et al.* (1990)	5.90E-14	18.05
1DQJ_AB:C	A	Fab Hyhel63	1DQQ_CD	HEW lysozyme	3LZT_A	1765	0.75	Li *et al.* (2000)	2.80E-09	11.67
1E4K_AB:C	OR	FC fragment of human IgG 1	2DTQ_AB	Human FCGR III	1FNL_A	1634	2.59	Maenaka *et al.* (2001)	1.70E-06	7.87
1E6E_A:B	ES	Adrenoxin reductase	1E1N_A	Adrenoxin	1CJE_D	2315	1.33	Schiffler *et al.* (2004)	8.56E-07	8.28
1E6J_HL:P	A	Fab 13B5	1E6O_HL	HIV-1 capsid protein p24	1A43_A	1245	1.05	Monaco-Malbet *et al.* (2000)	2.90E-08	10.28
1E96_A:B	OG	Rac GTPase	1MH1_A	p67 Phox	1HH8_A	1179	0.71	Lapouge *et al.* (2000)	2.70E-06	7.42
1EAW_A:B	EI	Matriptase	1EAX_A	BPTI	9PTI_A	1866	0.54	Farady *et al.* (2007)	4.97E-11	14.06
1EER_A:BC	OR	Erythropoietin	1BUY_A	EPO receptor	1ERN_AB	3347	2.44	Darling *et al.* (2002)	3.70E-12	15.59
1EFN_B:A	OX	HIV-1-NEF protein	1AVV_A	Fyn kinase SH3 domain	1FYN_A	1253	0.9	Lee *et al.* (1995)	3.80E-08	10.12
1EMV_A:B	EI	Colicin E9 nuclease	1FSJ_B	Im9 immunity protein	1IMQ_A	1535	1.28	Wallis *et al.* (1995)	2.40E-14	18.58
1EWY_A:C	ES	Ferredoxin reductase	1GJR_A	Ferredoxin	1CZP_A	1502	0.8	Sancho and Gomez-Moreno (1991)	3.57E-06	7.43
1EZU_C:AB	EI	D102N Trypsin	1TRM_A	Y69F D70P Ecotin	1ECZ_AB	2751	1.21	Yang *et al.* (1998)	8.00E-11	13.77
1F34_A:B	EI	Porcine pepsin	4PEP_A	Ascaris inhibitor 3	1F32_A	3038	0.93	Abu-Erreish and Peanasky (1974)	1.00E-10	14.19
1F6M_A:C	ES	Thioredoxin reductase	1CL0_A	Thioredoxin 1	2TIR_A	1830	4.9	Obiero *et al.* (2010)	2.70E-06	7.6
1FC2_C:D	OX	Staphylococcus Protein A	1BDD_A	Human Fc fragment	1FC1_AB	1307	1.69	Walker *et al.* (1995)	2.25E-08	10.43
1FFW_A:B	OX	Chemotaxis protein CheY	3CHY_A	Chemotaxis protein CheA	1FWP_A	1170	1.43	Schuster *et al.* (1993)	1.35E-06	8.09
1FLE_E:I	EI	Elastase	9EST_A	Elafin	2REL_A(4)	1771	1.02	Wiedow *et al.* (1990)	1.00E-09	12.28
1FQJ_A:B	OG	Gt-α	1TND_C	RGS9	1FQI_A	1806	0.91	Skiba *et al.* (1999)	6.70E-08	9.79
1FSK_BC:A	AB	Fab - Birch pollen antigen Bet V1	1FSK_BC	Birch pollen antigen Bet V1	1BV1_A	1623	0.45	†Jørgen Nedergaard Larsen	2.40E-10	13.12

Affinity benchmark continued on next page

Affinity benchmark continued

Complex	Type	Protein 1	PDB1	Protein 2	PDB2	DASA	I-RMSD	Reference	K_D	ΔG
1GCQ_B:C	OX	GRB2 C-ter SH3 domain	1GRI_B	Vav N-ter SH3 domain	1GCP_B	1208	0.92	Nishida *et al.* (2001)	1.70E-05	6.51
1GL1_A:I	EI	Chymotrypsin	4CHA_ABC	PMP-C (LCMI II)	1PMC_A(6)	1595	1.2	Kellenberger *et al.* (1995)	2.00E-10	13.23
1GLA_G:F	ER	Glycerol Kinase	1BU6_0	Glucose specific IIIGlc	1F3Z_A	1304	0.98	Pettigrew *et al.* (1998)	1.10E-05	6.76
1GPW_A:B	OX	HISF protein	1THF_D	Amidotransferase HISH	1K9V_F	2097	0.65	†Reinhard Sterner	5.00E-09	11.32
1GRN_A:B	OG	CDC42 GTPase	1A4R_A	CDC42 GAP	1RGP_A	2332	1.22	Hoffman *et al.* (1998)	2.39E-07	9.03
1GXD_A:C	EI	ProMMP2 type IV collagenase	1CK7_A	Metalloproteinase inhibitor 2	1BR9_A	2445	1.39	Olson *et al.* (1997)	5.20E-09	11.3
1H1V_A:G	OX	Actin	1JJ_B	Gelsolin precursor C-term	1P8X_A	2071	1.05	Bryan (1988)	2.50E-08	10.2
1H9D_A:B	OX	Runx1 domain of CBFα1	1EAN_A	Dimerisation domain of CBF-β	1ILF_A(1)	2125	1.32	Tang *et al.* (2000)	4.50E-08	9.18
1HCF_AB:X	OR	Neurotrophin-4	1B98_AM	TrkB-d5 growth factor receptor	1WWB_X	2146	0.88	Naylor *et al.* (2002)	2.60E-10	13.08
1HE8_B:A	OG	Ras GTPase	821P_A	PIP3 kinase	1E8Z_A	1305	0.92	Pacold *et al.* (2000)	3.20E-06	7.37
1HIA_AB:I	EI	Kallikrein	2PKA_XY	Hirustatin	1BX8_A	1737	1.4	Sollner *et al.* (1994)	1.30E-08	10.76
1I2M_A:B	OG	Ran GTPase-GDP	1QG4_A	RCC1	1A12_A	2779	2.12	Klebe *et al.* (1995)	2.50E-12	15.83
1I4D_D:AB	OG	Rac GTPase	1MH1_A	Arfaptin	1I49_AB	1657	1.41	Tarricone *et al.* (2001)	3.00E-06	7.46
1IB1_AB:E	OX	14-3-3 protein	1QJB_AB	Serotonin N-acteylase	1KUY_A	2808	2.09	Obsil *et al.* (2001)	2.00E-08	9.76
1IBR_A:B	OG	Ran GTPase-GDP	1QG4_A	Importin β	1F59_A	3370	2.54	Villa Braslavsky *et al.* (2000)	1.00E-09	12.07
1IJK_A:BC	ER	Von Willebrand Factor dom. A1	1AUQ_A	Botrocetin	1FVU_AB	1648	0.68	Miura *et al.* (2000)	2.30E-08	10.42
1IQD_AB:C	AB	Fab - Factor VIII domain C2	1IQD_AB	Factor VIII domain C2	1D7P_M	1976	0.48	Jacquemin *et al.* (1998)	<1.4e-11	15
1J2J_A:B	OG	Arf1 GTPase.GNP-RanBD1	1O3Y_A	GAT domain of GGA1	1OXZ_A	1209	0.63	Shiba *et al.* (2003)	1.10E-06	8.13
1JIW_P:I	EI	Alkaline metallo-proteinase	1AKL_A	Proteinase inhibitor	2RN4_A(1)	2018	2.07	Feltzer *et al.* (2000)	4.00E-12	15.55
1JMO_A:HL	ER	Heparin cofactor	1JMJ_A	Thrombin	2CN0_HL	3461	3.21	Ciaccia *et al.* (1997)	1.15E-07	9.47
1JPS_HL:T	A	Fab D3H44	1JPT_HL	Tissue factor	1TFH_B	1852	0.51	Presta *et al.* (2001)	1.00E-10	13.64
1JTG_B:A	EI	β-lactamase inhibitor protein	3GMU_B	β-lactamase TEM-1	1ZG4_A	2600	0.49	Albeck and Schreiber (1999)	4.00E-10	12.82
1JWH_CD:A	ER	Casein kinase II β chain	3EED_AB	Casein kinase II α chain	3C13_A	1451	1.27	Raaf *et al.* (2008)	1.26E-08	11.14
1K5D_AB:C	OG	Ran GTPase	1RRP_AB	Ran GAP	1YRG_B	2527	1.19	Seewald *et al.* (2003)	3.00E-10	12.77
1KAC_A:B	OR	Adenovirus fiber knob protein	1NOB_F	Adenovirus receptor	1F5W_B	1456	0.95	Kirby *et al.* (2000)	1.48E-08	10.68
1KKL_ABC:H	ES	HPr kinase C-ter domain	1JB1_ABC	HPr	2HPR_A	1641	2.2	Lavergne *et al.* (2002)	4.50E-08	10.02
1KLU_AB:D	OX	MHC class 2 HLA-DR1	1H15_AB	Staphylococcus enterotoxin C3	1STE_A	1254	0.43	Andersen *et al.* (1999)	4.60E-06	7.28
1KTZ_A:B	OR	TGF-β	1TGK_A	TGF-β receptor	1M9Z_A	989	0.39	De Crescenzo *et al.* (2006)	2.90E-07	8.92
1KXP_A:D	OX	Actin	1JJ_B	Vitamin D binding protein	1KW2_B	3341	1.12	Mc Leod *et al.* (1989)	9.00E-10	12.34
1KXQ_H:A	AB	Camel VHH - Pancreatic α-amylase	1KXQ_H	Pancreatic α-amylase	1PPI_A	2172	0.72	Lauwereys *et al.* (1998)	3.50E-09	11.54
1LFD_B:A	OG	Ras.GNP	5P21_A	RalGDS Ras-interacting domain	1LXD_A	1167	1.79	Kiel *et al.* (2004)	1.94E-06	7.79
1M10_A:B	ER	Von Willebrand Factor dom. A1	1AUQ_A	Glycoprotein IB-α	1M0Z_B	2097	2.1	Huizinga *et al.* (2002)	5.80E-09	11.24
1MAH_A:F	EI	Acetylcholinesterase	1J06_B	Fasciculin	1FSC_A	2145	0.61	Marchot *et al.* (1993)	2.50E-11	14.51
1MLC_AB:E	A	Fab44.1	1MLB_AB	HEW lysozyme	3LZT_A	1392	0.6	Goldbaum *et al.* (1999)	9.10E-08	9.61
1MQ8_A:B	OX	ICAM-1 domain 1-2	1IAM_A	Integrin α-L I domain	1MQ9_A	1241	1.76	Shimaoka *et al.* (2003)	3.00E-06	7.53
1NB5_AP:I	EI	Cathepsin H	8PCH_A	Stefin A	1DVC_A	1802	1.58	Lenarcic *et al.* (1996)	6.90E-11	13.86
1NCA_HL:N	AB	Fab - Flu virus neuraminidase N9	1NCA_HL	Flu virus neuraminidase N9	7NN9_A	1953	0.24	Pruett and Air (1998)	8.30E-09	11.02
1NSN_HL:S	AB	Fab N10 - Staphylococcal nuclease	1NSN_HL	Staphylococcal nuclease	1KDC_A	1776	0.35	Smith *et al.* (1991)	<1E-10	14
1NVU_Q:S	OG	Ras GTPase.GTP	1LF0_A	Son of sevenless	2II0_B	3127	1.98	Sondermann *et al.* (2004)	3.60E-06	7.43

Affinity benchmark continued on next page

Affinity benchmark continued

Complex	Type	Protein 1	PDB1	Protein 2	PDB2	DASA	I-RMSD	Reference	K_D	ΔG
1NVU_R:S	OG	Ras GTPase.GTP	1LF0_A	Son of sevenless	2II0_B	3452	3.09	Sondermann *et al.* (2004)	1.90E-06	7.8
1NW9_B:A	ER	Capase-9	1JXQ_A	BIR3 domain of XIAP (249-354)	2OPY_A	2112	1.97	Vucic *et al.* (2005)	1.30E-08	11.19
1OC0_A:B	ER	Plasminogen activator inhibitor-1	1B3K_A	Vitronectin Somatomedin Bd	2JQ8_A(4)	1313	1	Zhou *et al.* (2003)	1.00E-09	12.28
1OPH_A:B	EI	α-1-antitrypsin	1QLP_A	Trypsinogen	2PTN_A	1360	1.2	Stratikos and Gettins (1997)	5.00E-09	11.32
1P2C_AB:C	A	FabF10.6.6	2Q76_AB	HEW lysozyme	3LZT_A	1456	0.46	Cauerhff *et al.* (2004)	1.02E-10	13.63
1PPE_E:I	EI	Trypsinogen	2PTN_A	CMTI-1 squash inhibitor	1LU0_A	1688	0.34	Bolewska *et al.* (1995)	3.00E-12	15.56
1PVH_A:B	OR	IL6 receptor β chain D2-D3 domains	1BQU_A	Leukemia inhibitory factor	1EMR_A	1403	0.34	Boulanger *et al.* (2003)	8.00E-08	9.52
1PXV_A:C	EI	Staphylococcus aureus cystein protease	1X9Y_A	Cystein protease inhibitor	1NYC_A	2336	2.63	Dubin *et al.* (2007)	3.10E-10	12.97
1QA9_A:B	OX	CD2	1HNF_A	CD58	1CCZ_A	1353	0.73	van der Merwe *et al.* (1994)	9.00E-06	7.16
1R0R_E:I	EI	Subtilisin carlsberg	1SCN_E	OMTKY	2GKR_I	1409	0.45	Empie and Laskowski (1982)	2.94E-11	14.17
1R6Q_A:C	ER	Clp protease subunit ClpA	1R6C_X	Clp protease adaptor protein ClpS	2W9R_A	2450	1.67	Zeth *et al.* (2002)	3.30E-07	8.84
1RLB_ABCD:E	OX	Transthyretin	2PAB_ABCD	Retinol binding protein	1HBP_A	1439	0.66	Malpeli *et al.* (1996)	8.00E-07	8.18
1RV6_VW:X	OR	PIGF receptor binding domain	1FZV_AB	Flt1 protein domain 2	1QSZ_A	1639	1.09	Sawano *et al.* (1996)	1.70E-10	13.86
1S1Q_A:B	OX	UEV domain	2F0R_A	Ubiquitin	1YJ1_A	1288	0.98	Pornillos *et al.* (2002)	6.36E-04	4.29
1T6B_X:Y	OR	Anthrax protective antigen	1ACC_A	Anthrax toxin receptor	1SHU_X	1948	0.62	Wigelsworth *et al.* (2004)	1.70E-10	13.1
1US7_A:B	ER	Heat shock protein 82 N-ter domain	2FXS_A	HSP90 co-chaperone CDC37 C-ter domain	2W0G_A	1106	1.06	Roe *et al.* (2004)	1.46E-06	8.09
1UUG_A:B	EI	Uracyl-DNA glycosylase	3EUG_A	Glycosylase inhibitor	2UGI_B	2121	0.77	Bennett *et al.* (1993)	<1.00E-13	18
1VFB_AB:C	A	Fv D1.3	1VFA_AB	HEW lysozyme	8LYZ_A	1383	1.02	Bhat *et al.* (1994)	3.70E-09	11.46
1WDW_BD:A	ER	Tryptophan synthase β chain 1	1V8Z_AB	Tryptophan synthase α chain	1GEQ_A	3159	1.29	Ogasahara *et al.* (2003)	2.50E-09	12.72
1WEJ_HL:F	A	Fab E8	1QBL_HL	Cytochrome C	1HRC_A	1177	0.31	Carbone and Paterson (1985)	7.14E-10	12.48
1WQ1_R:G	OG	Ras GTPase.GDP	6Q21_D	Ras GAP	1WER_A	2913	1.16	Eccleston *et al.* (1993)	1.70E-05	6.62
1XD3_A:B	OX	UCH-L3	1UCH_A	Ubiquitin	1YJ1_A	2281	1.24	Dang *et al.* (1998)	3.00E-07	8.9
1XQS_A:C	OX	HspBP1	1XQR_A	Hsp70 ATPase domain	1S3X_A	2350	1.77	Shomura *et al.* (2005)	6.50E-06	7.08
1XU1_ABD:T	OR	TNF domain of APRIL	1U5Y_ABD	TACI CRD2 domain	1XUT_A(11)	1700	1.3	Wu *et al.* (2000)	6.40E-09	11.18
1YVB_A:I	EI	Falcipain 2	2GHU_A	Cystatin	1CEW_I	1743	0.51	Wang *et al.* (2007b)	6.50E-09	11.17
1Z0K_A:B	OG	Rab4A GTPase.GNP	2BME_A	RAB4 binding domain of Rabenosyn	1YZM_A	1787	0.53	Eathiraj *et al.* (2005)	7.70E-06	6.98
1ZHI_A:B	OX	BAH domain of Orc1	1M4Z_A	Sir Orc-interaction domain	1Z1A_A	1322	0.68	Hou *et al.* (2009)	2.00E-07	9.08
1ZLI_A:B	EI	Carboxypeptidase B	1KWM_A	Tick carboxypeptidase inhibitor	2JTO_A(6)	2087	2.53	Arolas *et al.* (2005a)	1.30E-09	12.04
1ZM4_A:B	ES	Elongation factor 2	1N0V_C	Diphtheria toxin A catalytic domain	1XK9_A	1554	2.94	Armstrong *et al.* (2002)	1.30E-06	8.03
2A9K_A:B	ES	Ral-A.GDP	1U8Z_A	Mono-ADP-ribosyltransferase C3	2C8B_X	1751	0.85	Pautsch *et al.* (2005)	6.00E-08	10.25
2ABZ_B:E	EI	Carboxypeptidase A1	3I1U_A	Leech carboxypeptidase inhibitor	1ZFI_A(1)	1447	0.9	Arolas *et al.* (2005b)	2.80E-09	11.67
2AJF_A:E	OR	Angiotensin-converting enzyme 2	1R42_A	SARS spike protein receptor binding domain	2GHV_E	1704	0.65	Li *et al.* (2005)	1.62E-08	10.63
2AQ3_A:B	OX	TCR Vβ8.2	1BEC_A	SEC3	1CK1_A	1105	1.82	Cho *et al.* (2010)	1.20E-05	6.71

Affinity benchmark continued on next page

Affinity benchmark continued

Complex	Type	Protein 1	PDB1	Protein 2	PDB2	DASA	I-RMSD	Reference	K_D	ΔG
2B42_A:B	EI	Xylanase	2DCY_A	Xylanase inhibitor	1T6E_X	2520	0.72	Fierens *et al.* (2005)	1.07E-09	12.11
2B4J_AB:C	OX	Integrase (HIV-1)	1BIZ_AB	PC4 and SFRS1 interacting protein	1Z9E_A(1)	1259	0.99	Tsiang *et al.* (2009)	1.09E-08	10.86
2BTF_A:P	OX	Actin	1JJ_B	Profilin	1PNE_A	2063	0.75	Schluter *et al.* (1998)	2.30E-06	7.69
2C0L_A:B	OX	TRP region of PEX5	1FCH_A	Sterol carrier protein 2	1C44_A	2013	2.62	Stanley *et al.* (2006)	1.09E-07	9.82
2FJU_B:A	OG	Phospholipase β 2	2ZKM_X	Rac GTPase	1MH1_A	1245	1.04	Snyder *et al.* (2003)	5.30E-06	7.2
2GOX_A:B	OX	Complement C3d fragment	1C3D_A	Staphylococcus aureus Efb-C	2GOM_A	1631	0.6	Haspel *et al.* (2008)	1.40E-09	12.08
2HLE_A:B	OR	Ephrin B4 receptor	2BBA_A	Ephrin B2 ectodomain	1IKO_P	2116	1.4	Chrencik *et al.* (2006)	4.00E-08	10.09
2HQS_A:H	OX	TolB	1CRZ_A	Pal	1OAP_A	2333	1.14	Bonsor *et al.* (2007)	2.70E-08	10.15
2HRK_A:B	OX	Glutamyl-t-RNA synthetase	2HRA_A	GU-4 nucleic binding protein	2HQT_A	1595	2.03	Karanasios *et al.* (2007)	9.00E-09	10.98
2I25_N:L	A	Shark single domain antigen receptor	2I24_N	HEW lysozyme	3LZT_A	1425	1.21	Dooley *et al.* (2006)	1.00E-09	12.28
2I9B_E:A	OR	uPAR surface receptor	1YWH_A	Urokinase-type plasminogen activator	2I9A_A	2382	3.79	Gardsvoll *et al.* (2004)	3.30E-10	12.93
2J0T_A:D	EI	MMP1 Intersitial collagenase	966C_A	Metalloproteinase inhibitor 1	1D2B_A(20)	1471	1.23	Wei *et al.* (2003)	4.00E-10	13.34
2JEL_HL:P	AB	Fab Jel42 - HPr	2JEL_HL	HPr	1POH_A	1501	0.17	Smallshaw *et al.* (1998)	2.80E-09	11.59
2MTA_HL:A	ES	Methylamine dehydrogenase	2BBK_JM	Amicyanin	2RAC_A	1461	0.41	Davidson *et al.* (1993)	4.50E-06	7.42
2NYZ_AB:D	OR	Viral chemokine binding p. M3	1MKF_AB	Chemokine XCL1	1J9O_A	2160	2.09	Alexander-Brett and Fremont (2007)	5.00E-10	12.69
2O3B_A:B	EI	NucA nuclease	1ZM8_A	NuiA nuclease inhibitor	1J57_A	1684	3.13	Ghosh *et al.* (2007)	3.20E-12	15.68
2OOB_A:B	ES	E3 ubiquitin-protein ligase CBL-B UBA domain	2OOA_A	Ubiquitin	1YJ1_A	808	0.85	Kozlov *et al.* (2007)	6.00E-05	5.66
2OOR_AB:C	ER	NAD(P) transhydrogenase subunit α part 1	1L7E_AB	NAD(P) transhydrogenase subunit β	1E3T_A	2065	1.42	Diggle *et al.* (1996)	1.55E-08	10.65
2OUL_A:B	EI	Falcipain 2	3BPF_A	Chagasin	2NNR_A	1933	0.53	Wang *et al.* (2007b)	1.70E-09	11.96
2OZA_B:A	OX	MAP kinase 14	3HEC_A	MAP kinase-activated protein kinase 2	3FYK_X	6254	1.89	Lukas *et al.* (2004)	2.50E-09	11.73
2PCB_A:B	ES	Cyt C peroxidase	1CCP_A	Cytochrome C	1HRC_A	1029	0.45	Erman *et al.* (1997)	1.00E-05	6.82
2PCC_A:B	ES	Cyt C peroxidase	1CCP_A	Cytochrome C, yeast	1YCC_A	1141	0.39	Pielak and Wang (2001)	1.60E-06	7.91
2PTC_E:I	EI	Trypsinogen	2PTN_A	BPTI	9PTI_A	1429	0.28	Vincent and Lazdunski (1972)	6.00E-14	18.04
2SIC_E:I	EI	Subtilisin	1SUP_A	Streptomyces subtilisin inhibitor	3SSI_A	1617	0.36	Uehara *et al.* (1978)	7.12E-11	13.84
2SNI_E:I	EI	Subtilisin	1UBN_A	Chymotrypsin inhibitor 2	2CI2_I	1628	0.35	Otzen and Fersht (1999)	2.00E-12	15.96
2TGP_Z:I	EI	Trypsinogen	1TGB_A	BPTI	9PTI_A	1432	0.57	Bode (1979)	2.40E-06	7.54
2UUY_A:B	EI	Trypsinogen	2PTN_A	Tryptase inhibitor from tick	2UUX_A	1280	0.44	Paesen *et al.* (2007)	5.60E-09	11.26
2VDB_A:B	OX	Serum albumin	3CX9_A	Peptostreptococcalalbumin-binding protein	2J5Y_A	1807	0.47	de Chateau *et al.* (1996)	1.50E-10	13.4
2VIR_AB:C	A	Fab	1GIG_LH	Flu virus hemagglutinin	2HMG_AB	1263	0.8	Fleury *et al.* (1998)	1.00E-09	12.28
2VIS_AB:C	A	Fab	1GIG_LH	Flu virus hemagglutinin	2VIU_ACE	1296	0.8	Fleury *et al.* (1998)	4.00E-06	7.36
2WPT_A:B	EI	Colicin E9 nuclease	1FSJ_B	Im2 immunity protein	2NO8_A	1581	1.61	Li *et al.* (1998)	1.50E-08	10.67
3BP8_AB:C	OX	Mlc transcription regulator	1Z6R_AB	PTS glucose-specific enzyme EIICB	3BP3_A	1398	0.45	Nam *et al.* (2008)	4.14E-09	11.44

Affinity benchmark continued on next page

Affinity benchmark continued

Complex	Type	Protein 1	PDB1	Protein 2	PDB2	DASA	I-RMSD	Reference	K_D	ΔG
3BZD_A:B	OX	TCR Vβ8.2	1BEC_A	SEC3-1A4	3BVZ_A	1312	1.08	Cho *et al.* (2010)	9.60E-08	9.57
3CPH_G:A	OG	Ras-related protein Sec4	1G16_A	Rab GDP-dissociation inhibitor	3CPI_G	1685	2.12	Ignatev *et al.* (2008)	3.30E-07	8.84
3SGB_E:I	EI	Streptogrisin B	2QA9_E	Ovomucoid inhibitor third domain	2OVO_A	1268	0.36	Wieczorek *et al.* (1987)	1.79E-11	14.51
4CPA_A:I	EI	Carboxypeptidase A	8CPA_A	Potato carboxypeptidase inhibitor	1H20_A(9)	1177	1.52	Ryan *et al.* (1974)	5.00E-09	11.32

6.2.1 The Structures

The benchmark contains values for 144 complexes. For all of the interactions, high resolution (< 3.25Å) crystal structures are available for the complex and the individual binding partners in isolation, except for a few antibody/antigen complexes. As the affinity benchmark was based on docking benchmark 4.0 (Hwang *et al.*, 2010), no pairs of complexes with high sequence identity are included, in order to alleviate redundancy. Exceptions were made for nine pairs of complexes, in which one is a high affinity cognate complex, and the other is a closely related non-cognate complex with much lower binding affinity. Many of the complexes undergo extensive conformational changes, some involving disorder to order transitions, and others with interface C_α RMSD up to 5Å between the bound and the unbound conformations.

6.2.2 The Affinities

The affinity data comes from a wide variety of sources, mostly isothermal titration calorimetry, surface plasmon resonance, stopped-flow fluorimetry and other spectroscopic techniques, and they span nine orders of magnitude. In many cases, affinities were derived from inhibition constants, used as part of a kinetic model. Corroborating K_D values could be found for around half the complexes, and typically differed from each other by about a factor of 2, corresponding to around 0.4 kcal mol^{-1} error, significantly larger than the 20-50% standard error typically reported. Greater differences in affinity were observed when the conditions were varied; temperature, pH and ionic strength. All but three of the affinities were obtained in the 18-35°C range. The cases in which affinity is determined at various temperatures suggest that this can affect the affinity by up to a factor of two (Kresheck *et al.*, 1995; Schwarz *et al.*, 1995; Ogasahara *et al.*, 2003; Smallshaw *et al.*, 1998). Studies in which the effect of varying ionic strength is ascertained show that salt concentrations within the typical range of 0.1-0.5M can results in differences in binding affinity up to an order of magnitude (Davidson *et al.*, 1993; Erman *et al.*, 1997; Kresheck *et al.*, 1995; Mc Leod *et al.*, 1989; Eccleston *et al.*, 1993; Alexander-Brett and Fremont, 2007; Armstrong *et al.*, 2002). However, the

greatest variations in binding affinities are observed when the pH is varied; this can result in differences up to a factor of 50, corresponding to around 2.3 kcal mol^{-1} in the typical pH range of 5.5-8.5 (Faller and Bieth, 1991; Ascenzi *et al.*, 1988; Mc Leod *et al.*, 1989). For these reasons, any predictive binding affinity function capable of determining free energies within around 2 kcal mol^{-1} should be deemed to be accurate to within the limitations of the data unless these environmental factors are accounted for.

Table 6.2: A summary of the binding affinity benchmark. High affinity is deemed to be $K_D < 10^{-10}$M, medium 10^{-10} to 10^{-6} and low $> 10^{-6}$. Complexes deemed to have large conformational changes (Flex.) are those with interface C_α RMSD > 1.5Å. Means and standard deviations are calculated with non-cognate complexes excluded. Table adapted from Kastritis *et al.* (2011).

	Number		Affinity class			ΔG/kcal mol^{-1}		
Class	All	Noncognate	High	Med.	Low	Mean	σ	Flex.
Antibody/Antigen	19	2	2	16	1	12.2	1.3	0
Enzyme/Inhibitor	40	4	17	22	1	13.8	2.3	7
Enzyme/Other	21	1	0	12	9	9.2	1.9	7
G-protein	17	-	1	6	10	8.9	2.5	6
Receptor	13	-	1	11	1	11.5	2.1	4
Miscellaneous	34	2	0	22	12	9.3	2.2	11
All	144	9	20	90	34	11.0	2.9	35

6.2.3 The Functions

The complexes are involved in wide ranging functions, which were used to classify them. These included antibody-antigen complexes (A, or AB where no unbound antibody was available), enzyme-inhibitor complexes (EI), enzyme-regulatory/accessory chain complexes (ER), receptor containing complexes (OR), complexes with G-binding proteins (OG) and miscellaneous complexes (OX). The breakdown of all complexes according to their number, affinity and associated conformational changes, is shown in Table 6.2.

6.2.4 Affinity Prediction

A number of potentials were applied to the benchmark and subsets of it, in order to investigate the capabilities of methods such as those described

Table 6.3: Empirical binding free energy functions tested on larger data sets. The number of test cases are shown (Cases), with the corresponding set in parentheses, correlation (Corr.), the method used (Method) and the reference (Reference), are shown. Potentials of mean force are denoted 'PMF', and master thermodynamic equations as 'sum'.

Cases	Corr.	Method	Reference
46 (C)	0.35	sum, HADDOCK	de Vries *et al.* (2007)
46 (C)	0.26	PMF, DComplex	Liu *et al.* (2004)
46 (C)	0.35	sum, ATTRACT	Bastard *et al.* (2006)
46 (C)	0.50	sum, AffinityScore	Audie and Scarlata (2007)
46 (C)	0.53	sum, PyDock	Cheng *et al.* (2007a)
46 (C)	0.40	sum, FireDock	Andrusier *et al.* (2007)
46 (C)	0.52	sum, Rosetta	Jiang *et al.* (2005)
46 (C)	0.36	sum, FastContact	Champ and Camacho (2007)
46 (C)	0.35	sum, ZRANK	Pierce and Weng (2007)
137 (B)	0.59	other	Section 7.3
137 (B)	0.33	PMF, DComplex	Liu *et al.* (2004)
80 (N)	0.3	other	Section 7.3
80 (N)	0.21	PMF, DComplex	Liu *et al.* (2004)
67 (R)	0.72	other	Section 7.3
70 (F)	0.32	other	Section 7.3
57 (V)	0.76	other	Section 7.3
57 (V)	0.44	PMF, DComplex	Liu *et al.* (2004)
57 (V)	0.51	PMF	Su *et al.* (2009)
28 ($V \cap R$)	0.91	other	Section 7.3
28 ($V \cap R$)	0.67	PMF, DComplex	Liu *et al.* (2004)
28 ($V \cap R$)	0.85	PMF	Su *et al.* (2009)
29 ($V \cap F$)	0.54	other	Section 7.3
29 ($V \cap F$)	0.41	PMF, DComplex	Liu *et al.* (2004)
29 ($V \cap F$)	0.35	PMF	Su *et al.* (2009)
14 ($V \cap C$)	0.72	other	Section 7.3
14 ($V \cap C$)	0.61	sum, AffinityScore	Audie and Scarlata (2007)
14 ($V \cap C$)	0.66	sum, PyDock	Cheng *et al.* (2007a)
14 ($V \cap C$)	0.69	PMF, DComplex	Liu *et al.* (2004)
14 ($V \cap C$)	0.64	PMF	Su *et al.* (2009)
37 ($B \cap C$)	0.60	other	Section 7.3
37 ($B \cap C$)	0.40	sum, AffinityScore	Audie and Scarlata (2007)
37 ($B \cap C$)	0.44	sum, PyDock	Cheng *et al.* (2007a)
37 ($B \cap C$)	0.57	PMF, DComplex	Liu *et al.* (2004)
31 ($\subset V$)	0.59	sum	Section 7.4

in section 1.4.4.8, as well as the quality of the data. The benchmark, which we shall denote B, was split into subsets. The subset containing flexible complexes, F, is the sets for which the interface C_α RMSD is greater than 1.0. Its complement, $R = F^c$, contains the rigid complexes. In addition, we compiled a subset of the benchmark containing only entries for which we had confidence in the binding affinity. For this set, which we call the validated set, V, complexes were included only if their binding affinity

has been measured experimentally by more than one group or more than one experimental technique, and that the measurements were within 1 kcal mol^{-1} of one another. The validated set contains 57 complexes, 29 of which are in the intersection with F, and spans an affinity range of 13 kcal mol^{-1}. It comprises of 3 antibody/antigen complexes, 16 enzyme-inhibitor complexes, 5 enzyme-substrate complexes, 5 other complexes with enzymes, 8 complexes containing G-proteins, 7 receptor-ligand complexes and 13 miscellaneous complexes. The complement of the validated set is the non-validated set, $N = V^c$.

The results of this analysis appear in table 6.3. For comparison, this table also includes the results from Kastritis and Bonvin (2010), in which a number of methods were evaluated on a set of 46 complexes, which are denoted C, as well as some of the results from the empirical free energy models explained in the next chapter. These results show that the quality of data is very important, as the performance of these method on V are consistently superior to B, which in turn are superior to N. Further, they highlight the difficulties arising due to conformational changes; the performance of these methods on F are significantly worse than on R.

6.3 Summary

A large structurally cross-referenced database of protein-protein binding affinities has been assembled, with care to ensure environmental conditions and corroborating data is reported. These span a broad variety of functions and affinities, and undergo a wide range of conformational changes upon binding, including disorder to order transitions. A number of empirical binding free energy functions were evaluated on the structures, highlighting the importance of high quality data, and the challenges presented when conformational changes are not negligible.

Chapter 7

Affinity and Kinetics Prediction

7.1 Introduction

The binding affinity benchmark, discussed in the previous chapter, was used to devise binding free energy functions and functions for the prediction of kinetic rate constants. To do so, we first compiled a large descriptor set covering many physical and geometrical aspects of the complexes, which is discussed in section 7.2. In section 7.3, it is shown how these descriptors are used to construct models using random forest regression, M5' regression, multivariate adaptive regression with splines and radial basis function interpolation. These outperform all other methods which were tested, and their consensus prediction is better still. Finally, in section 7.4, I show how cross-validated multiple regression, feature selection and a form of early stopping regularisation can be used to build simple dissociation and association rate functions for the prediction of binding kinetics and affinity. The model is carefully parameterised and benchmarked, and shown to perform well. This performance is speculated to be due to the divide-and-conquor approach, in which the binding affinity prediction is split into two problems which are solved independently: dissociation rate prediction and association rate prediction. Both the affinity and kinetics models only require input data which can be quickly and easily computed, and are thus suitable for high-throughput studies.

7.2 Molecular Descriptors

In order to construct binding free energy models, we derived a large number of molecular descriptors for the binding affinity benchmark. The previous chapter, and section 1.4.4.8, highlight a number of issues regarding the prediction of binding affinities, and we had these in mind when selecting descriptors. The following considerations guided our choices.

1. A first concern was computational efficiency. Most of the chosen descriptors can be calculated in a fraction of a second, and all but two can be calculated within seconds. The first of these, the change in vibrational entropy, is traditionally calculated using normal mode analysis with an all atom force field, and requires hours of minimisation prior to the analysis. We used an approximate method based on the elastic network model, which typically takes less than 5 minutes (Carrington and Mancera, 2004). The second computationally demanding feature included was the BioSimz association score. However, this feature was not important for any of the derived affinity models.

2. As proteins do not exist as static structures, but as structural ensembles (see section 1.4.4.5), the measured binding affinities correspond to the Boltzmann weighted average of the energy of the bound ensemble, minus the energy of the unbound ensembles, and we wished to account for this. However, we did not want to perform computationally expensive molecular dynamics or Monte Carlo simulations. To achieve this, we generated ensembles of each complex and its unbound constituents using CONCOORD (de Groot *et al.*, 1997), a method which uses pseudo-NMR restraints. This typically takes minutes. As the CONCOORD method generates physically plausible structures, all the models were treated equally, and the mean was calculated over the ensemble.

3. We wished to keep the features as physically plausible as possible. With the exception of a few interface composition and geometrical features, all the descriptors can be traced back to putative physical phenomena. However, we did not wish to restrict ourselves to terms relating to single phenomena, such as in the master thermodynamic

equations discussed in section 1.4.4.8. Moreover, we wished to allow for the possibility of, say, mixing statistical pair potentials which account for many phenomena simultaneously, with other terms relating to phenomena which aren't captured by pair potentials, such as entropy changes.

4. We wished to have a number of features relating to entropy, as this effect is harder to account for and is often neglected. Thus we made sure to include terms relating to vibrational entropy, the entropy of disordered loops and termini, side-chain entropy, rotational and vibrational entropy, as well as the hydrophobic effect.

5. Solvation effects are important in driving biomolecular recognition, and so a number of solvation models were included, from simple hydrophobic burial terms to sophisticated continuum electrostatics models.

6. As the pH is the most significant environmental factor influencing binding affinity, and alternative protonation states are included in the CHARMM22 force field, a number of the descriptors were added which account for this. Firstly, PROPKA was used to determine the pKa of the titratable amino acids (Bas *et al.*, 2008). The pKa of the amino acids and the pH of the solution determine the probability that a given residue exists in a given protonation state. The most probable assignment of protonation states, at the experimental pH, was determined using PDB2PQR (Dolinsky *et al.*, 2007). For the 16 complexes for which the experimental pH was not reported, the standard pH of 7.5 was used. All the descriptors prefixed with ACE22 were calculated with these pH adjusted models.

7. Conformational changes are very common in the benchmark, and thus we decided to include many features relating to the change in energetics as the proteins go from the unbound state to the bound.

By no means can the final descriptors be seen as a complete set which fit the above criteria, but we believe they give good coverage of the various physical phenomena at play.

In total, 200 parameters were calculated to describe the interface, interaction and conformational changes upon binding. These are divided into the following sections, although some fit equally into multiple categories. Some parameters describe the interface directly, such as surface complimentarity, while others were calculated as $E_{int} = E_{complex} - (E_{R,b} + E_{L,b})$. Of the values, some were calculated *de novo*, others were calculated by servers, packages or stand-alone programs including the ProtorP server (Reynolds *et al.*, 2009), CHARMM (Brooks *et al.*, 2009), PyRosetta (Chaudhury *et al.*, 2010), FireDock (Andrusier *et al.*, 2007), and the Potentials'R'Us server (Feng *et al.*, 2010).

Before calculating the molecular descriptors, post-translational modifications were treated as follows: S-oxymethionine (2TGP) and selenomethionine (1FQJ, 1KKL, 3BP8, 3CPH, 2HQS, 2OOB, 1S1Q and 1XQS) were changed to methionine, phosphoserine (1F34) and phosphothreonine (1IB1) residues were converted to serine and threonine respectively, O-sulfo-L-tyrosine (1JMO) was changed to tyrosine, D-glutamine (1S1Q and 1XD3) was modelled as L-glutamine, dioxyselenocysteine (2SNI) was changed to cysteine and diphthamide (1ZM4) was changed to histidine. With the exception of the phosphothreonine in 1IB1, these modifications are away from the binding interface. 137 proteins were used, with the following omissions: 1DE4 was removed from the test set due to its size. 1UUG, 1IQD and 1NSN were removed as only upper limits to their dissociation constant are available. 1M10, 1NCA and 1NB5 were removed owing to difficulties in deriving a complete descriptor set.

7.2.1 Statistical Potentials

A total of 35 statistical potentials are used, both coarse grain and atomistic. Most were dervied from intramolecular interactions. However, such potentials have shown considerable utility in charaterising interfaces and disinguishing true interfaces from decoys (Liu *et al.*, 2004).

The potentials considered include the atomistic (ROS_FA_PP) and coarse grained (ROS_CG_PP) Rosetta pair potentials (Simons *et al.*, 1999),

the DFIRE (Zhou and Zhou, 2002) (DFIRE) and directional DFIRE (Yang and Zhou, 2008) (DDFIRE) potentials, the OPUS-Ca (Wu *et al.*, 2007) (OPUS_CA) and OPUS-PSP (Lu *et al.*, 2008a) (OPUS_PSP) potentials, the residue level potential of Rykunov and Fiser (2010) (RF_PP), the geometric potential function of Li and Liang (2011) (GEOMETRIC) and the EMPIRE potential (Liang *et al.*, 2007) (EMPIRE). Further are the 26 two-body and four-body coarse grain potentials provided by the Potentials'R'Us server, which are described by Pokarowski *et al.* (2005) (FOUR_BODY, GEN_4_BODY, SHORT_RANGE, QA_PP, QM_PP, QP_PP, HLPL_PP, SKOB_PP, SKOA_PP, SKJG_PP, MJPL_PP, MJ3H_PP, MJ2H_PP, TS_PP, BT_PP, BFKV_PP, TD_PP, TEL_PP, TES_PP, RO_PP, MS_PP, MJ1_PP, MJ3_PP, GKS_PP, VD_PP and MSBM_PP).

7.2.2 Solvation and Entropy terms

Accounting for solvation are the atomic contact energies (DELISI_SOLV) developed by Zhang *et al.* (1997), the Lazaridis-Karplus effective energy function (Lazaridis and Karplus, 1999b) (LK_SOLV), the SASA model (Ferrara *et al.*, 2002) (SASA), and the Rosetta C_β (ROS_CG_BETA) and environment (ROS_CG_ENV) potentials (Simons *et al.*, 1999). Also included are the electrostatic self energy (ACE22_SELF) and screening (ACE22_SCRE) contribution to the solvation free energy, and their sum (ACE22_SOLV), as calculated using the analytical continuum electrostatics method of Schaefer and Karplus (1996) using the CHARMM22 force field, along with their hydrophobic burial term (ACE22_HYDR). These terms were also calculated using the CHARMM19 force field (ACE19_SELF, ACE19_SCRE, ACE19_SOLV and ACE19_HYDR). The hydrophobic burial term from the STC package was also calculated (STC_S_SOL) (Lavigne *et al.*, 2000).

Other entropic terms include the translational (S_T) and rotational (S_R) entropy changes (see equations 1.53 and 1.54), and their sum (S_TR), as calculated using CHARMM, and the total entropy change calculated by STC (Lavigne *et al.*, 2000). Change in vibrational entropy upon binding (S_VIB) was calculated using normal modes, via equation 1.52. The $N-6$ non-trivial elastic network normal modes were calculated using ElNeMo

(Suhre and Sanejouand, 2004), and the entropy was calculated using the M1 scheme described by Carrington and Mancera (2004).

Entropy changes arising from restriction of side chain conformation upon binding (STC_S_SC) were calculated using STC (Lavigne *et al.*, 2000). Two models of entropy changes associated with disorder to order transitions were included, as reviewed by Zhou (2004), with a Gaussian polymer model for free chains at termini, and a wormlike chain model for loops linking ordered regions. Disordered residues are taken as those which are either not resolved in the crystal structure or that have a C_α occupancy below 0.5. Residues which are disordered in both the bound and the unbound structure are neglected, as are loops of fewer than 3 residues. In the Gaussian polymer model (S_GP_ALL2), configurational entropy is proportional to $\ln(\det(\mathbf{W}))$ where $\mathbf{W}$ is the Wang-Uhlenbeck matrix (Wang and Uhlenbeck, 1945). For the non-interacting loops in the data set, this reduces to the logarithm of the number of residues. For the worm-like chain model (S_WLC_ALL2), the method of Zhou (2001) was employed (with C=0). Two further parameters (S_GP_INT2 and S_WLC_INT2) are also included, for which only the entropy changes of loops involving interfacial residues (those with a non-hydrogen atom within 10Å of a non-hydrogen atom on the binding partner) are considered.

7.2.3 Other Potentials

Other potentials include the directional H-bonding potential implimented in Rosetta (Kortemme *et al.*, 2003) (ROS_HBOND) and the 12-10 potential implimented in FireDock (Andrusier *et al.*, 2007) (H_BOND), as well as orientation independent π-π (PI_PI), cation-π (CATION_PI) and aliphatic-alphatic (ALIPHATIC) potentials (Misura *et al.*, 2004). Further are the total Rosetta energy (ROS_TOTAL), total CHARMM22 energy (ACE22_ALL), total STC enthalpy (STC_H) and free energy (STC_G), attractive (ROS_FA_ATR), repulsive (ROS_FA_REP) and coarse grained (ROS_CG_VDW) Van der Waals terms, CHARMM 22 Coulombic (ACE22_COUL), electrostatic interaction (ACE22_INTE, the sum of ACE22_COUL and ACE22_SCRE) and total electrostatic (ACE22_ELEC, the sum of ACE22_INTE and ACE22_SELF), as well

as the CHARMM19 electrostatic interaction (ACE19_INTE) and Coulombic terms (ACE19_COUL).

7.2.4 Other Descriptors

The proportion of interface residues which are in alpha helices (INT_ALPHA) and beta sheets (INT_BETA), were determined with DSSP (Kabsch and Sander, 1983) as well as the number of interfacial H-bonds (NUM_HB), salt bridges (NUM_SB) and water bridges (NUM_WB) as calculated by HBPLUS (McDonald and Thornton, 1994; Barlow and Thornton, 1983). Further are parameters obtained with NACCESS (Hubbard and Thornton, 1993), including DASA, the change in surface area upon binding, RES_P, RES_NP and RES_C, the percentage of interface residues that are polar, non-polar and charged respectively and ATOM_P, ATOM_NP and ATOM_N, the percentage of interface atoms which are polar, non-polar and neutral. Geometrical properties of the interface are characterised by the planarity (PLANARITY), numerical eccentricity (ECCENTRIC), the volume of empty space at the interface (GAP_VOL) and this volume divided by the interface area (GAP_INDEX) as calculated with the SURFNET package (Laskowski, 1995). NSC, a surface complimentarity score and NIP, an interface packing score (Mitra and Pal, 2010), were also calculated. BIOSIMZ_KON is the predicted log(kon), calculated using BioSimz (Li, 2011). STC_CP is the change in specific heat upon binding, as calculated with STC.

7.2.5 Unbound-Bound Descriptors

As well as the loop entropy changes described above, a number of other parameters were used to quantify the effects of conformational changes upon binding. Only residues which appear in both the bound and unbound structures are included, in order to make the two conformational states comparible. One of these terms is the change in internal energy (INTERNAL_UB), as calculated with the CHARMM19 force field. The remaining terms, which are used as described above, are followed by the suffix _UB and are calculated as

$E_{UB \to B} = (E_{R,b} - E_{R,u}) + (E_{L,b} - E_{L,u})$: ROS_FA_ATR_UB, ROS_FA_REP_UB, LK_SOLV_UB, ROS_FA_PP_UB, ROS_CG_VDW_UB, ROS_CG_PP_UB, ROS_CG_ENV_UB, ROS_CG_BETA_UB, ROS_HBOND_UB, ROS_TOTAL_UB, DFIRE_UB, DDFIRE_UB, OPUS_CA_UB, OPUS_PSP_UB, RF_PP_UB, GEOMETRIC_UB, FOUR_BODY_UB, GEN_4_BODY_UB, SHORT_RANGE_UB, QA_PP_UB, QM_PP_UB, QP_PP_UB, HLPL_PP_UB, SKOB_PP_UB, SKOA_PP_UB, SKJG_PP_UB, MJPL_PP_UB, MJ3H_PP_UB, MJ2H_PP_UB, TS_PP_UB, BT_PP_UB, BFKV_PP_UB, TD_PP_UB, TEL_PP_UB, TES_PP_UB, RO_PP_UB, MS_PP_UB, MJ1_PP_UB, MJ3_PP_UB, GKS_PP_UB, VD_PP_UB, MSBM_PP_UB, ACE19_HYDR_UB, ACE19_SELF_UB, ACE19_SCRE_UB, ACE19_COUL_UB, ACE19_SOLV_UB, ACE19_INTE_UB and SASA_UB.

7.2.6 Ensemble Descriptors

Ensembles of structures were generated using CONCOORD 2.1 (de Groot *et al.*, 1997) with dynamic tolerence setting. For each case, 100 structures were generated for the ligand, the receptor and the bound complex. Many of the above parameters were calculated on the ensembles, and their mean value reported with the suffix _ENS for parameters relating to the interaction and _EBU for parameters relating to unbound to bound conformational change: STC_CP_ENS, STC_H_ENS, STC_S_ENS, STC_S_SOL_ENS, STC_S_SC_ENS, STC_G_ENS, ROS_FA_ATR_ENS, ROS_FA_REP_ENS, LK_SOLV_ENS, ROS_FA_PP_ENS, ROS_CG_VDW_ENS, ROS_CG_PP_ENS, ROS_CG_ENV_ENS, ROS_CG_BETA_ENS, ROS_HBOND_ENS, ROS_TOTAL_ENS, NIP_ENS, NSC_ENS, DELISI_SOLV_ENS, H_BOND_ENS, PI_PI_ENS, CATION_PI_ENS, ALIPHATIC_ENS, DFIRE_ENS, DDFIRE_ENS, OPUS_CA_ENS, OPUS_PSP_ENS, RF_PP_ENS, GEOMETRIC_ENS, ROS_FA_ATR_EBU, ROS_FA_REP_EBU, LK_SOLV_EBU, ROS_FA_PP_EBU, ROS_CG_VDW_EBU, ROS_CG_PP_EBU, ROS_CG_ENV_EBU, ROS_CG_BETA_EBU, ROS_HBOND_EBU, ROS_TOTAL_EBU, ACE19_HYDR_EBU, ACE19_SELF_EBU, ACE19_SCRE_EBU, ACE19_COUL_EBU, ACE19_SOLV_EBU, ACE19_INTE_EBU, SASA_EBU, INTERNAL_EBU, GEOMETRIC_EBU, DFIRE_EBU, DDFIRE_EBU, OPUS_CA_EBU,

OPUS_PSP_EBU and RF_PP_EBU.

7.3 Binding Free Energy Models

In collaboration with colleague Rudi Agius, the descriptors described in the previous section were used to train a random forest regression machine learner (RF), a multivariate adaptive regression with splines function (MARS), a radial basis function interpolation model (RBF), and an M5' decision tree regression model (M5'). These were all implimented in MatLab. The efficacy of these learners was tested with leave-one-out cross-validation as an outer wrapper, and were shown to perform well compared to the other binding prediction methods described in section 1.4.4.8. The predictions were then combined into a consensus model, in which the final value is taken as the mean of the four models.

7.3.1 Methods

A wide variety of machine learning tools are available, and their performance is highly dependent on the task at hand. It was our intention not to pick methods at random, and use them as black boxes, but to select methods which are particularly suited to the problem. The four methods were chosen on the basis of the following considerations.

1. All the chosen methods allow their internal mechanisms to be scrutinised. This was deemed to be important so that the physical plausibility of the model could be evaluated, and to ascertain the relative contribution of the various factors at play.

2. Methods were chosen that had different philosophical rhetoric. The RF, for instance, is derived from the consensus of many low-accuracy models. The MARS method allows the exploitation of parameters which have predictive value only over certain ranges. The M5' method works by splitting hairs, and finding local differences between quantitatively similar examples, whilst the RBF method exploits global features of the training set, and the predicted affinity of any given case is largely determined by examples further away in feature space. These different approaches were considered likely to capture different aspects of

the descriptor set, and hence more likely to work synergistically when combined together into a consensus model.

3. None of the chosen methods requires fine tuning, and they are robust to the model parameters. This property is valuable, as it eliminates biases which could originate from tweaking parameters until the desired result is obtained. Default parameters were used in all cases, and were not optimised.

4. As we have more features than examples, overfitting is a serious concern. The methods chosen account for this problem either explicitly or implicitly. Random Forests do not overfit as more trees are added, rather the test error converges to a limiting value. The RF is able to achieve low bias predictions through trees built from different subsets of the data and descriptors, and low variance through averaging the output of all trees for the output, and the cancellation of errors. The M5' and MARS methods both have backward elimination routines in them to eliminate redundant features. In the RBF method, each feature is equally weighted and it is the significance of the examples in the training set which are determined. This is particularly suited to situations in which there are more features than examples, as is the case here.

5. As many of the features are correlated with each other, we chose methods which did not allow the exploitation of colinearity. In linear regression, for instance, two highly correlated features can be assigned large positive and negative coefficients, accentuating their differences and allowing overfitting. None of the methods used are susceptible to this effect.

7.3.1.1 Random Forest

The RF algorithm, as described by Breiman (2001), was used as its MatLab implementation (Liaw and Wiener, 2002). Briefly, take N to be the number of complexes in the training set. Bootstrap sampling is used to generate 750 lists of binding affinities, each of length N; *i.e.* each member of each list is chosen at random such that for each list, some affinities are duplicated and

others are omitted. These 750 lists are then used as a basis for constructing 750 decision trees. For each node of each tree, 20 molecular descriptors are chosen at random and of these, the descriptor which is best able to split the data by affinity is chosen. The tree is grown until each affinity on its list corresponds to a leaf. Once the forest is constructed, it can be used for prediction; the features are fed into each decision tree, and the predicted binding affinity is the mean output of all trees.

7.3.1.2 Multivariate Adaptive Regression Splines

A version of the MARS regression algorithm was used for model building (Friedman, 1991). In this method, the binding free energy function consists of a linear combination of basis functions. These basis functions can take one of three forms: (1) A constant; this is only included once, and is the mean of the training affinities. (2) A hinge function; one linear term, consisting of a feature multiplied by a coefficient, is used up until a hinge point. Beyond this point a second linear term is used. The two terms are equal at the hinge point. (3) An interaction function; the product of two hinge terms.

Initially, during model construction, the basis set consists of only the constant. Additional basis functions are then added one by one by greedy forward selection. The pair of features for each additional basis function consists of either the constant or a feature which has already been used, and a new feature. A brute force search selects the feature pair, hinge point and coefficients which maximally reduce the square residuals of the regression. The algorithm terminates either when 21 basis functions have been selected, or when the addition of a new basis function reduces the square residuals by less than 0.0001 $\text{kcal}^2\ \text{mol}^{-2}$. As this initial construction often produces a model which overfits the data, a feature elimination routine is used. This backward greedy algorithm removes descriptors based on cross-validation. The elimination routine can remove whole basis functions, but it can also remove just one of the linear terms, thus leaving features which are used only over certain ranges. In the version used here, hinge points are replaced with hinge regions, within which there is a smooth interpolation between the two terms.

7.3.1.3 Radial Basis Function Interpolation

The RBF method was also used to construct an empirical free energy function (Hardy, 1971). Before the function is parametrised, the values for each descriptor are normalised in the range [0,1]. In this method, the energy function consists of the mean affinity of the training set, μ, plus a linear combination of multiquadratic basis functions, $\phi(d) = \sqrt{d^2 + 1}$, summed across the n cases in the training set.

$$F(\mathbf{x}) = \mu + \sum_{i=1}^{n} a_i \phi(\|\mathbf{x} - \mathbf{x}_i\|) \tag{7.1}$$

where $\mathbf{x}$ is the descriptor vector. The a regression coefficients are determined by multiple regression. In this intriguing model, all features are equally weighted, and it is the weights of the cases in the training set which are optimised.

7.3.1.4 M5′ Decision Tree

The M5′ decision tree regression, as explained in Wang and Witten (1997), was also applied. Briefly, a decision tree is constructed in which the criteria at each node is selected such that the standard deviations within the branches are minimised. The tree is fully grown until each leaf corresponds to a single affinity. For each internal node, a linear regression model is built using all molecular descriptors. The RMS error of each model is weighted by the number of features used in the regression. Subsequently, a greedy backward selection algorithm is applied to each regression model until the weighted error is minimised. The tree is then pruned back from the leaves for as long as the weighted error of each nodes regression model decreases. When used for prediction, the tree is descended according to the splitting criteria at each node. The tree is then ascended back to the root, and the value predicted from the regression model of the leaf is combined with the value predicted from each internal node as it is crossed, using the method of Quinlan (1992). In the implementation used here, 16 M5′ decision trees are constructed, in four sets of four. Each tree recieves a random subset of features to work with, in such a way that each set of four trees contains every feature and each feature is used by only one tree. When predicting,

the mean output of the 16 trees is returned.

7.3.1.5 Data Sets

All 137 complexes for which the molecular descriptors were calculated were used in this investigation. We also investigate performance on the subsets of examples outlined in section 6.2.4; the validated set, the non-validated set, the flexible set and the rigid set. Also considered was the set of complexes that overlap with the data set tested by Kastritis and Bonvin (2010), the predictions for which were generously supplied to us by the authors.

7.3.1.6 Model Evaluation

The ability for the models to predict binding affinity was determined by calculating the Pearson's product-moment correlation coefficient between the leave-one-out cross-validated predictions and the experimental data. When comparing models, a Fisher r to z transformation of the correlation coefficients is performed to yield the test statistic.

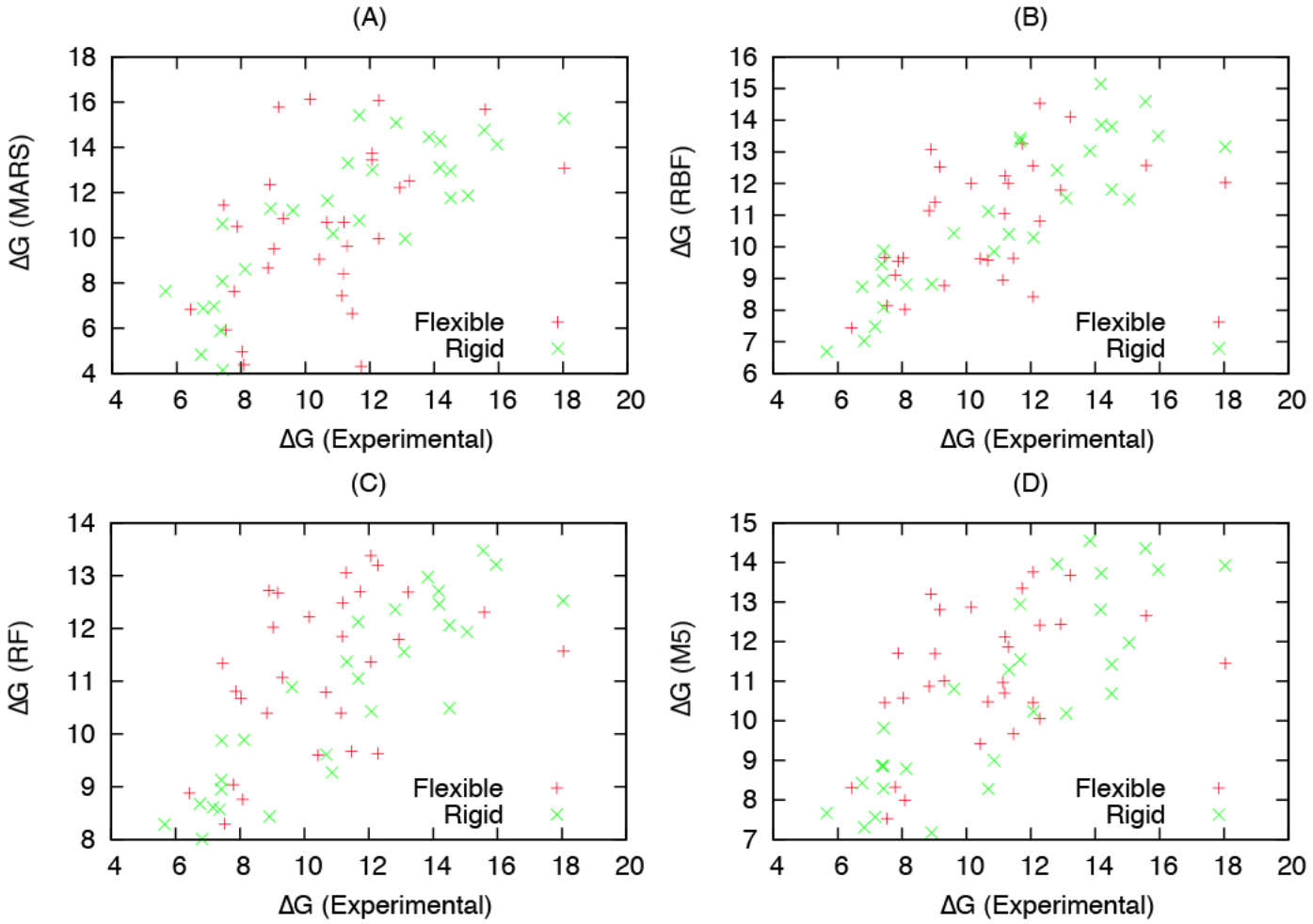

Figure 7.1: Scatter plots of leave-one-out cross-validated results for the four regression models. (A) MARS, r_{all}=0.65, r_{rig}=0.82, r_{flex}=0.45. (B) RBF, r_{all}=0.75, r_{rig}=0.87, r_{flex}=0.54. (C) RF, r_{all}=0.68, r_{rig}=0.89, r_{flex}=0.46. (D) M5', r_{all}=0.69, r_{rig}=0.84, r_{flex}=0.48.

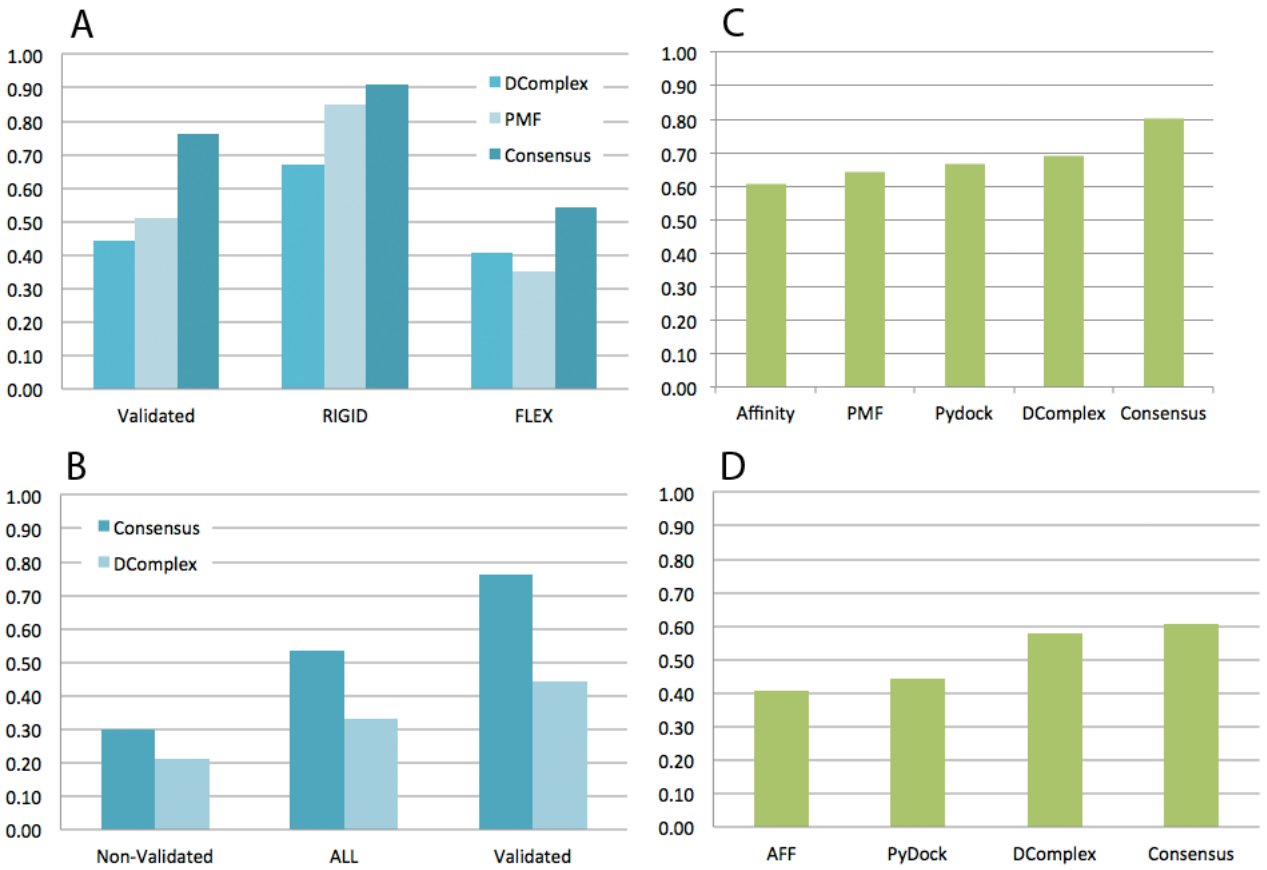

Figure 7.2: The performance of the empirical energy functions and the consensus model, tested on various data sets, as measured by correlation coefficient. Leave-one-out cross-validation is used where applicable to ensure no over-fitting. (A) The validated test set. (B) The affinity benchmark. (C) The overlap between the data set of Kastritis and Bonvin (2010) and the validated set. (D) The overlap between the data set of Kastritis and Bonvin (2010) and the affinity benchmark.

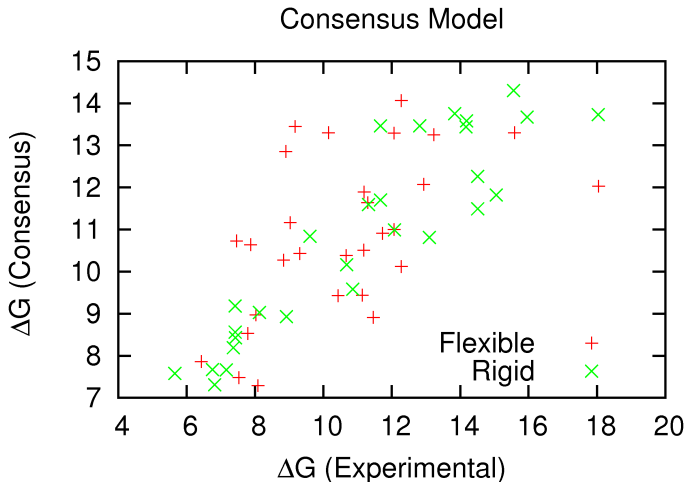

Figure 7.3: Scatter plots of leave-one-out cross-validated results for the consensus model, r_{all}=0.76, r_{rig}=0.91, r_{flex}=0.54.

7.3.2 Results

7.3.2.1 All Features

The four regression models were trained on the validated data set. Scatter plots for leave-one-out cross-validation are shown in Figure 7.1. Although the RBF method performed the best, all methods perform well, with correlations of between 0.65 and 0.75. For comparison, binding affinities were calculated using DComplex (Liu *et al.*, 2004) and the PMF determined by Su *et al.* (2009). For DComplex, the results were r_{all}=0.44, r_{rig}=0.67, r_{flex}=0.41. For the PMF, correlations were found to be r_{all}=0.51, r_{rig}=0.85, r_{flex}=0.35.

The correlations between the predictions were evaluated, to see if the different models were yielding different values. Whilst the predictions of the RF and M5′ methods were correlated with the RBF method (r=0.87 and r=0.86 respectively), and highly correlated with each other (r=0.95), the MARS model was less correlated with the RBF, RF and M5′ methods (r=0.65 , r=0.69 and r=0.68 respectively). This suggested that the models may be detecting different aspects of the feature set, and a consensus model was built in which the predicted affinity was the mean output of the four models. The performance of the consensus model was comparable to that of RBF of its own, although it does perform slightly better when only the rigid complexes are taken into consideration (r_{rig}=0.91). The scatter plot

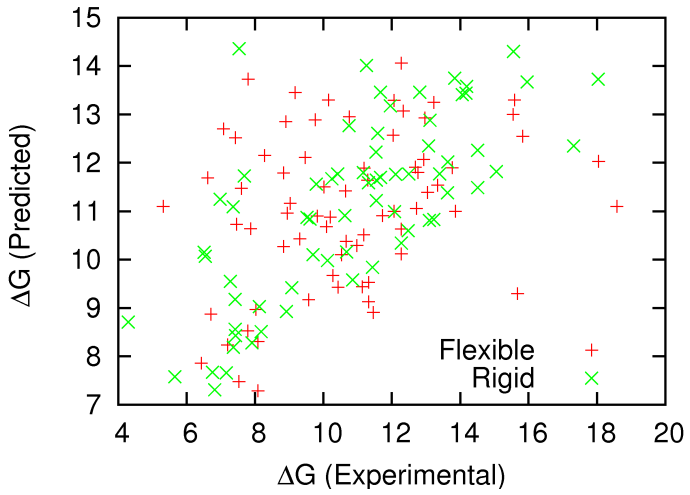

Figure 7.4: Scatter plots for the consensus model for all complexes in the affinity benchmark (r_{all}=0.53, r_{rig}=0.72, r_{flex}=0.32). This is the unison of the predictions for the non-validated set, where the model is trained on the whole validated set, and of the leave-one-out cross-validated predictions for the validated set.

for this model is shown in Figure 7.3. The consensus model performs significantly better than both DComplex (z=2.27, p=0.003) and the other PMF (z=2.25, p=0.012), as shown in Figure 7.2A. The predicted binding affinity for the rigid cases were found to have a much higher correlation with the experimental data than the flexible (r_{rig}=0.91, r_{flex}=0.54, z=3.3, p=0.0005).

When applied to the non-validated set, the predictions correlated with the experimental data with r=0.30, compared to r=0.21 for the DComplex potential (see Figure 7.2B). This drop in correlation highlights the importance of using high quality experimental data. When applied to the whole data set, as shown in Figure 7.4, the correlation of 0.53 was significantly higher than for DComplex (r=0.33, z=2.05, p=0.022). For this larger test set, the consensus model performed noticeably better than the MARS, RBF, RF and M5' models on their own (r=0.45, r=0.42, r=0.47 and r=0.46 respectively). The fact that the bottom right corner of Figure 7.4 is considerably less populated than the top left demonstrates that the model is overestimating the strength of the interaction more frequently than underestimating it. Coupled with the fact that this effect is most prevalent for the flexible cases, this suggests that the deficiency lies in the

modelling of configurational entropy changes and/or enthalpy changes associated with conformational rearrangement. As a further test of the consensus model, the performance of the leave-one-out cross-validated predictions was compared to the performance of AffinityScore (Audie and Scarlata, 2007) and the PyDock scoring function (Cheng *et al.*, 2007a) tested on the 14 complexes which overlap the validated set and the complexes tested by Kastritis and Bonvin (2010) (Figure 7.2C), and the 37 complexes which overlap with the whole data set (Figure 7.2D). The consensus score performed better than all other methods tested on both sets.

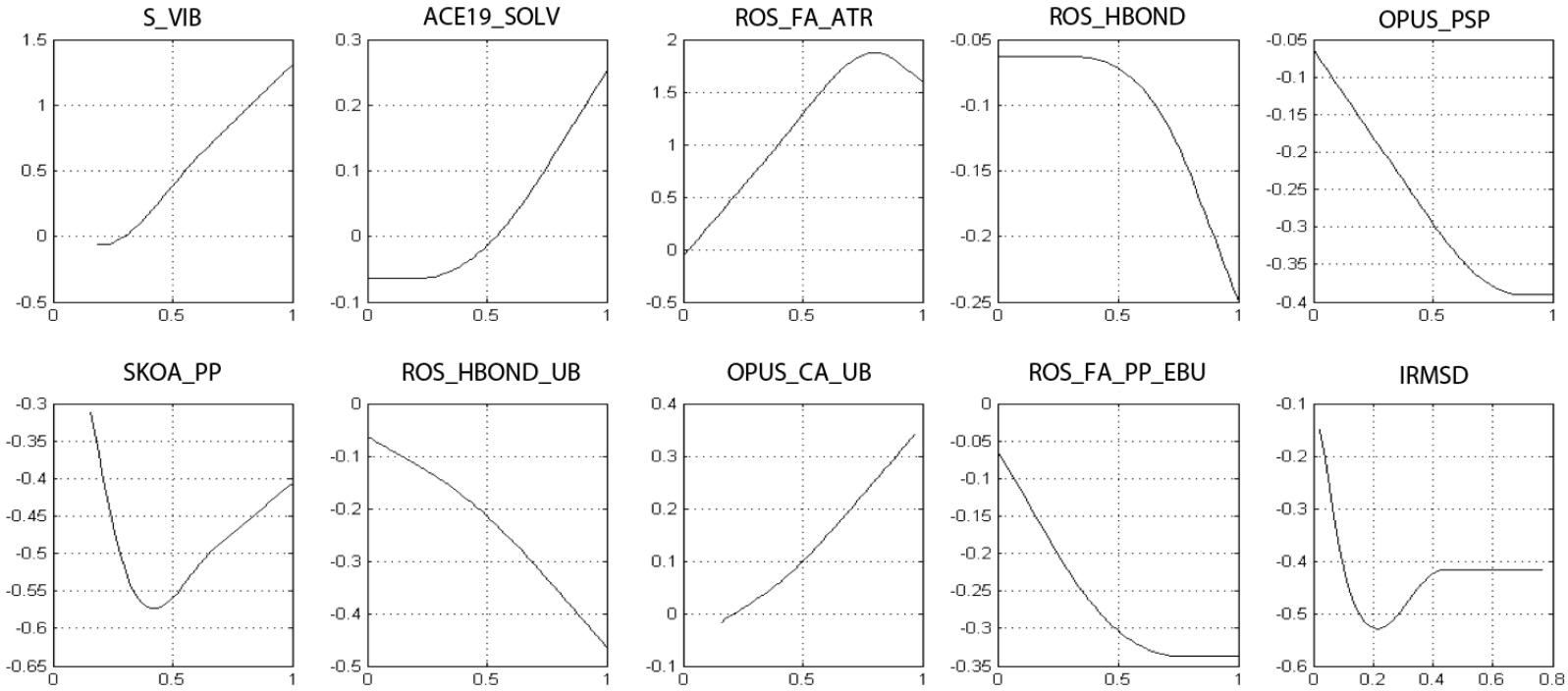

Figure 7.5: MARS feature contribution across the affinity range for the 10 most important features. The contribution, shown on the y-axis, is normalised such that a negative value signifies disruption and *vice versa*, relative to the mean. The parameter range, on the x-axis, is normalised between 0.0 and 1.0.

7.3.2.2 Model Details

All four of the models allow their internal workings to be scrutinised. For instance, for each complex in the RF model, the various features which are used as the decision trees are navigated can be ascertained. Upon doing this, it was shown that a descriptor relating to the unbound-bound transition was one of the 5 most frequently used features for 16 of the 29 flexible cases (55.2%), as opposed to just 3 of the 28 rigid complexes (10.7%), demonstrating that the method was selecting the correct features for the cases in hand.

The MARS method allows different features to make different contributions to the binding affinity depending on their value, with the importance of each feature varying across the range in which it is used. The degree of contribution across the parameter range for the 10 most important features in shown in figure 7.5.

The significance profile of S_VIB, the vibrational entropy change, is interesting. The method used is approximate and noisy (Carrington and Mancera, 2004). The MARS algorithm chooses to ignore this feature when it is low, in the lowest 10% of the range of the training set, but its contribution to the binding free energy increases approximately linearly as its magnitude increases beyond this point. Presumably, this signifies that the algorithm uses this feature when the benefit of including it outweighs the detrimental effect of introducing noise into the calculation.

The change in electrostatic solvation free energy (ACE19_SOLV), is closely related to the change in electrostatic self energy, the feature which was of most significance in CAPRI round 21 (see section 5.3.2.3). The influence of this term is approximately linear over most of its range. Interestingly, the contribution at low values is constant; most of this region contains only a single outlier, 1ZM4. Similar plots, linear over most the range, and sometimes with a constant, or neglection, at extreme values, are seen for most the other descriptors. The plot for the SKOA pair potential is difficult to rationalise, as is the exact form the interface RMSD descriptor, which is involved in two hinge functions. However, for the latter, it does makes sense that the degree of conformational change makes little contribution for the rigid cases, and that increasing flexibility corresponds to destabilisation of

Figure 7.6: Correlation coefficient when the consensus model is trained with feature subsets. The UB features are those derived using the unbound structure. The ENS features are those calculated using the CONCOORD ensembles. The EBU features are those which are calculated on both the ensembles and the unbound structures. The BASIC feature are all the features that are neither UB, ENS, nor EBU.

the final complex; most of the data set is concentrated in the leftmost 20% of the plot.

7.3.2.3 Feature Subsets

The consensus model was retrained four times with different feature subsets (see Figure 7.6). The consensus model still outperforms the PMF and DComplex methods when the information relating to the unbound-bound conformational change is omitted, and thus the superiority of the current approach cannot be explain solely by the inclusion of this data. As expected, the removal of all details pertaining to the unbound-bound transition had no effect on the correlation of the rigid cases; it remained around 0.9 for all the sets. Surprisingly, however, the features relating to the unbound-bound transition only aided the flexible cases noticeably when they were calculated on the structural ensembles; adding these features resulted in a reduction of RMS error of 2.1 kcal mol^{-1} for the flexible cases, compared to only 0.2 kcal mol^{-1} for the rigid.

One of the most interesting results is that for the interaction between MK2 and p36 MAPK (PDBid 2OZA). Upon binding MK2 undergoes a significant disorder-order transition at both the C-terminus and a large loop (Figure

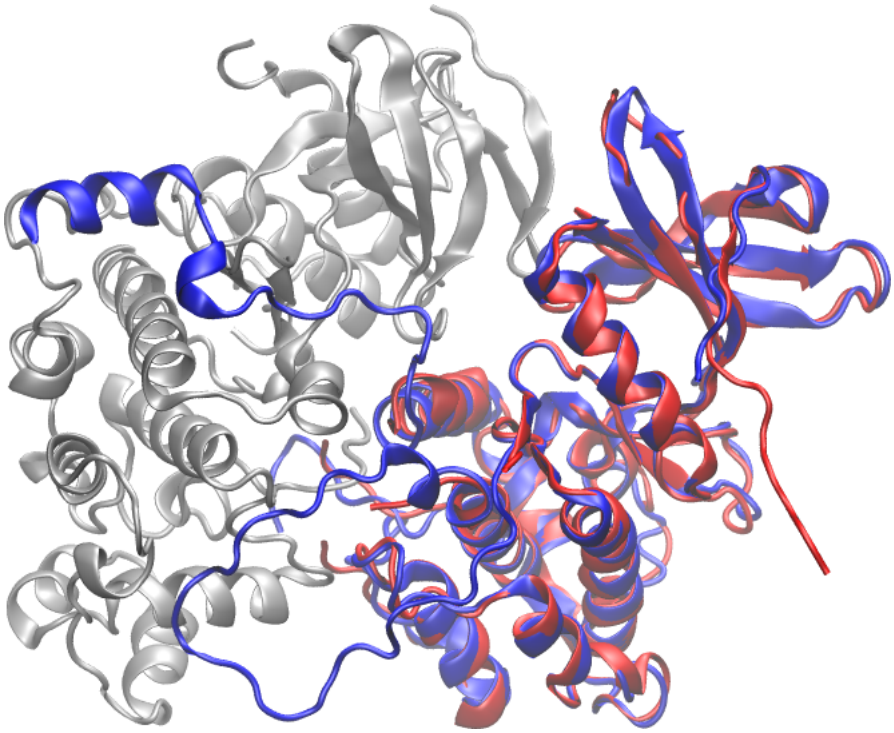

Figure 7.7: The interaction between p38 MAPK (silver) and MK2 (blue). The unbound crystal strucutre of MK2 is shown superimposed (red), revealing the two large regions which undergo a disorder to order transition upon binding.

7.7), by far the largest disorder-order transition in the data set, and has a very large surface area. For these reasons, it is usually omitted from binding free energy calculations (Pons and Fernandez-Recio, 2011; Bai *et al.*, 2011). Indeed, the inability to account for the loss of configurational entropy when the descriptors pertaining to the unbound-bound transition were omitted meant that the affinity was greatly overestimated (17.4 kcal mol^{-1} compared to the experimental value of 11.7 kcal mol^{-}1). However, when these data were included, the descriptors for the entropy change due to disorder to order transitions were picked up, and could be correctly extrapolated to an affinity much closer to the experimental (10.9 kcal mol^{-1}).

7.3.3 Discussion

Accurate energy functions are important for a wide variety of purposes, from protein engineering and interaction design to designing peptide inhibitors and docking. Their construction is predicated upon ascertaining the factors that influence binding and their relative importance. Machine learning techniques were used to combine a large set of molecular descriptors covering a wide variety of physical features describing the affinity benchmark, for the prediction of binding affinities. Leave-one-out cross-validation shows that this method performs better than other empirical free energy functions. In particular, the correlation of the prediction of rigid cases was around 0.9, signifying that their affinity can be predicted within experimental limitations and the inherent approximations used in descriptor calculation. Whilst the inclusion of a number of features relating to the unbound-bound transition aided the prediction of affinity for interactions involving flexible proteins when descriptors were averaged over structural ensembles, this group of complexes still poses significant challenges. Evidently, either energetic factors associated with conformational change are not included in the descriptor set, or these were not being detected in the models.

This work also demonstrates the utility of approximate methods of calculating entropy changes, such as using an elastic network model to calculate vibrational entropy changes (Carrington and Mancera, 2004) and the worm-like-chain and Gaussian polymer models of the conformational entropy of disordered loops (Zhou, 2004). These approximation techniques will play an increasingly important role now that it has become clear that proteins with

high flexibility are significantly under-represented amongst those for which structural details are known at atomic resolution (Marsh and Teichmann, 2011). A manuscript outlining this work is in preparation.

7.4 Kinetic Rate Functions

To date, only one group have attempted to predict dissociation rate constants from structure (Bai *et al.*, 2011). However, an incongruous feature set, and the fact that feature selection, training and validation were performed on the same set of complexes, suggest that the reported correlations highly overstimate the predictive value of their model. Additionally, spurious parameter values render their K_D and k_{off} models unreproducible. Using their $\log_{10} k_{on}$ function, for which they report a correlation of 0.73, results in a correlation of 0.10 when applied to the 44 complexes in the affinity benchmark for which kinetic data is available.

The approach taken by us was to use the descriptor set outlined in section 7.2 for the 44 complexes which had kinetic data, or a subset of these complexes. For these, feature selection and linear regression was used to derive a series of descriptor subsets of increasing cardinality for both k_{on} and k_{off}. From these sets of feature sets, a pair was chosen on the basis of their combined ability to predict binding affinity on a separate set of complexes, by using equation 1.62 which equates ΔG with a function of k_{on} and k_{off}. Finally, the selected k_{on} and k_{off} models were used to predict the binding affinities for a third set of complexes, none of which were used in the previous stages. As feature selection, model selection and model testing are applied to separate subsets, the final affinity estimate is believed to be an accurate representation of the predictive power of the k_{on} and k_{off} models when combined for affinity prediction. This work was implemented in Python, using the rpy package.

This approach can be seen not only as the derivation of two kinetic functions, but as a divide-and-conquer approach to affinity prediction; the binding free energy problem is decomposed into two kinetics prediction problems, which are solved independently and then combined. The procedure

outlined here was also repeated by splitting the affinity prediction problem into enthalpy and entropy prediction problems, using experimental data derived from isothermal titration calorimetry (results not shown). However, experimental entropy and enthalpy data could only be gathered for 28 complexes, which was insufficient for the generation and selection of a good model.

7.4.1 Methods

7.4.1.1 Feature Selection

The feature selection and regression routine used was identical to that outlined in section 5.3.2.2, only with a population size of 20, up to 10 speculative rounds and 5-fold cross-validation as the internal estimator of predictive value. The regression model was not used to predict k_{on} and k_{off} directly, but their base-10 logarithm.

7.4.1.2 Model Selection

Model selection was performed using a type of early stopping. Early stopping is a technique used with feature selection in order to determine when to stop adding features so as to minimise overfitting (see Figure 7.8A). The data is split into a training and a test set. Feature selection is performed with the training set, and the performance on the test set is evaluated each time a new feature is added. Initially, the performance of both the training and the test set decrease as features are added. However, as more features are added, the error of the training data continues to decrease whilst the error of the test set starts to increase, due to the model being overfit to the training data. The feature set which performs best on the test data is selected.

Due to only a limited number of complexes available with kinetic data, splitting it into test and training data was undesirable. Hence, a variation of early stopping was used for model selection, in which the feature sets for the x-axis and y-axis correspond to the models for k_{on} and k_{off}. For every pair of models, the results are combined using equation 1.62 to predict the binding affinity. Thus, the model selection set can consist of complexes for which kinetic data is not available. The performance measure used for

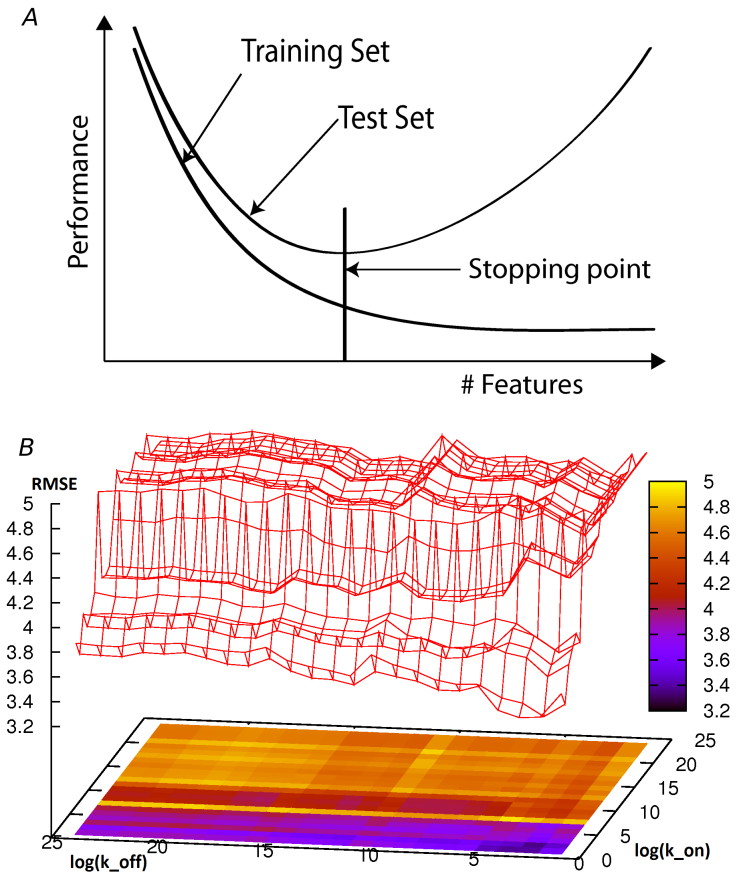

Figure 7.8: (A) A fictitious early stopping curve for model selection. Features are added by a greedy algorithm and trained on a training set. The model which corresponds to the best performance on the test set is selected. (B) An early stopping surface. A pair of feature sets are chosen, one for k_{on} and one for k_{off}. This pair is the one which produces the best performance when combined to predict the affinity of the test set, in this case two features are selected for k_{on} and two for k_{off}. The curve shown corresponds to scheme 2 (see Figure 7.9).

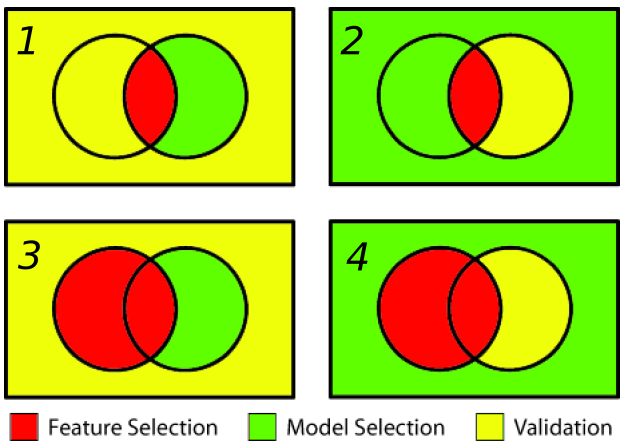

Figure 7.9: Venn Diagrams showing the four different combinations of feature selection, model selection and validation sets. The rectangle corresponds to all 137 complexes in the affinity benchmark. The circle on the left corresponds to the set of 44 complexes for which kinetic data is available. The circle on the right corresponds to the 57 complexes in the validated set (see section 7.3.1.5). The intersection of these two sets consists of 26 interactions.

model selection is the RMS error. An early stopping surface is shown in Figure 7.8B.

7.4.1.3 Subset Combinations

Although the feature selection set must be trained on the kinetic data, it wasn't clear whether to use all the data for which kinetic details were known, or just those in the validated set. Further, it wasn't clear how to separate the data for training and validation. To test the trade off between the number of complexes used for feature selection and the quality of that data, and how to split the data which isn't used for feature selection, four different schemes were devised as shown in Figure 7.9.

7.4.2 Results

Feature selection, model selection and validation were performed. The results are summarised in Table 7.1. Scheme 1 shows that empirical kinetic functions trained on the kinetic data in the validated set, with two descriptors for the k_{on} model and five for the k_{off}, can be combined and predict the affinities of the remainder of the validated set with a correlation of 0.69, considerably higher than both PMFs in Figure 7.2A. However, it

Table 7.1: Results for feature selection, model selection and validation. The number of features for the k_{on} and k_{off} models is shown, alongside their leave-one-out cross-validation correlations and RMSE. The RMSE and correlation of the ΔG values used for selecting these models is also shown, as are the those when applied to the validation set. RMSEs are accordingly reported in either kcal mol^{-1} or as dimensionless numbers.

	$\log_{10} k_{on}$			$\log_{10} k_{off}$			ΔG_{sel}		ΔG_{val}	
Scheme	#	Corr.	RMSE	#	Corr.	RMSE	RMSE	Corr.	RMSE	Corr.
1	2	0.70	0.89	5	0.79	1.17	2.45	0.69	**3.59**	**0.09**
2	2	0.70	0.89	2	0.56	1.58	3.36	0.10	**2.61**	**0.59**
3	8	0.77	0.86	2	0.45	1.47	2.50	0.60	**3.67**	**0.19**
4	2	0.52	1.14	2	0.45	1.47	3.26	0.17	**2.80**	**0.51**

performs poorly when tested on the non-validated set. Scheme 2, in which the non-validated complexes are used for model selection, performs badly on this poor quality set. Despite this, the selected models, both of which have only two terms, can be combined to predict the binding affinities of the validation set with a correlation of 0.59, again considerably higher than the two methods in Figure 7.2A. Scatter plots for these k_{on} and k_{off} models are shown in Figure 7.10, along with their combined ΔG prediction. The fact that such a simple model, with few adjustable parameters, can be selected using poor quality data and perform well on the validated test set, demonstrates the power of the approach taken here. Results for schemes 3 and 4, in which all the kinetic data was used for feature selection, are similar to those for schemes 1 and 2.

In scheme 2, the two descriptors selected for the k_{on} model are NUM_HB and DFIRE_EBU, and the descriptors for the k_{off} model are ROS_CG_BETA and OPUS_CA_ENS. These descriptors are telling, although they must be considered in the context of potential confounding factors. For the k_{off} model, the first is a pair potential calculated on the crystal structure of the complex. It strongly correlates with other pair potentials in the descriptor set (0.8-0.9). The second is also a pair potential, only averaged over the structural ensembles. This also correlates with the other pair potentials calculated over ensembles (0.7-0.8). It is interesting to note that, in theory, these pair potentials should correlate with binding affinity. In principle, the binding affinity can be exactly decomposed into a the sum of $-\log k_{on}$ and $\log k_{off}$, and these pair potentials appear to be capturing the latter; it is reasonable to assume that these are properties associated with the interface, accounting for

effects such as tight packing, which prevent the complex from dissociating once formed. The k_{on} descriptors, however, are more telling. The first is the number of interfacial hydrogen bonds. H-bonds are largely electrostatic in nature, and indeed this feature correlated with the continuum electrostatic models in the descriptor set (0.5-0.6). Electrostatics has long been known to be important for association due to its long range of action (Zacharias, 2010b). The second descriptor is the most interesting. It is the average energy difference between the structural ensembles of the complex, excluding the intermolecular term, and the unbound proteins, as calculated with DFIRE. It correlates well with the other unbound-bound pair potential energies calculated over the ensembles (up to 0.86). This provides strong support for the conformational sorting and population shift mechanism (see section 1.4.4.5). The rate of association is proportional to the concentration of the binding partners, and can be factorised into concentration and association rate constant (equation 1.37). The DFIRE_EBU term is the average difference in energy between the proteins in their bound conformational state and their unbound state, both in the absence of their binding partners. When under thermodynamic control, the population of a state is governed by its energy and the temperature, as quantified by the Boltzmann distribution (equations 1.47). Thus the ratio between the number of unbound proteins in the bound conformation and the unbound conformation is the ratio of their Boltzmann factors, exp $\frac{-\Delta E}{k_B T}$, in which the difference in energy appears in the exponent.

Finally, the k_{on} and k_{off} models of scheme 2 were applied to the whole affinity benchmark, to see if any patterns emerged regarding the functional classes of the interactions. The results for this are shown in Figure 7.11. As expected from the experimental data, the enzyme-inhibitor interactions are the strongest. However, it is observed that receptor/accessory chain interactions share a feature with the enzyme-inhibitor interactions; phenomena related to dissociation contributes more to the binding affinity than phenomena relating to association, relative to the other functional types. This is not seen, however, for the enzyme-substrate interactions, presumably due to the fact that were the substrate to dissociate slowly, it would act as an inhibitor. The classes which appear to dissociate the quickest, given their affinity, are the complexes containing G-proteins and, surprisingly, the miscillaneous complexes. It was expected that the receptor complexes would have fast as-

sociation kinetics, due to their general role in signal transduction, however, this was not observed.

7.4.3 Discussion

The prediction of kinetic constants is imperative for understanding the dynamics of biological systems. Here we have presented a method of deriving a'simple empirical function for the prediction of these constants using a divide-and-conquer approach. Combined, the k_{on} and k_{off} models can successfully predict binding affinities with greater accuracy than the methods outlined in section 1.4.4.8. Additionally, the features which were selected offer insights into the mechanism of protein association, giving strong evidence for the conformational shift scheme. Just like the binding affinity prediction methods previously presented, the quality of the data is important for parametrisation and validation. Should a larger amount of high quality kinetic data become available, it would become possible to add more features without overfitting, and derive further insights into the binding process. Further, this method can be applied to the derivation of empirical enthalpy and entropy functions using ITC data, and as more interactions become structurally and thermodynamically characterised, this approach will become viable. A manuscript outlining this work is in preparation.

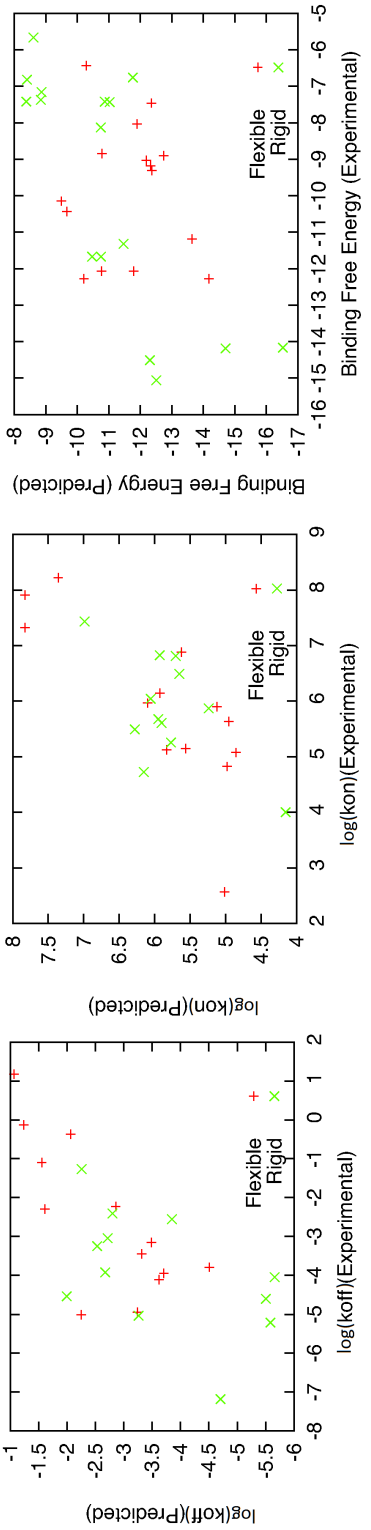

Figure 7.10: The k_{on} and k_{off} models derived from scheme 2, applied to the intersection of the validated set and the set of complexes with kinetic data. The predictions for the combined binding affinity score, applied to the rest of the validated set, is also shown.

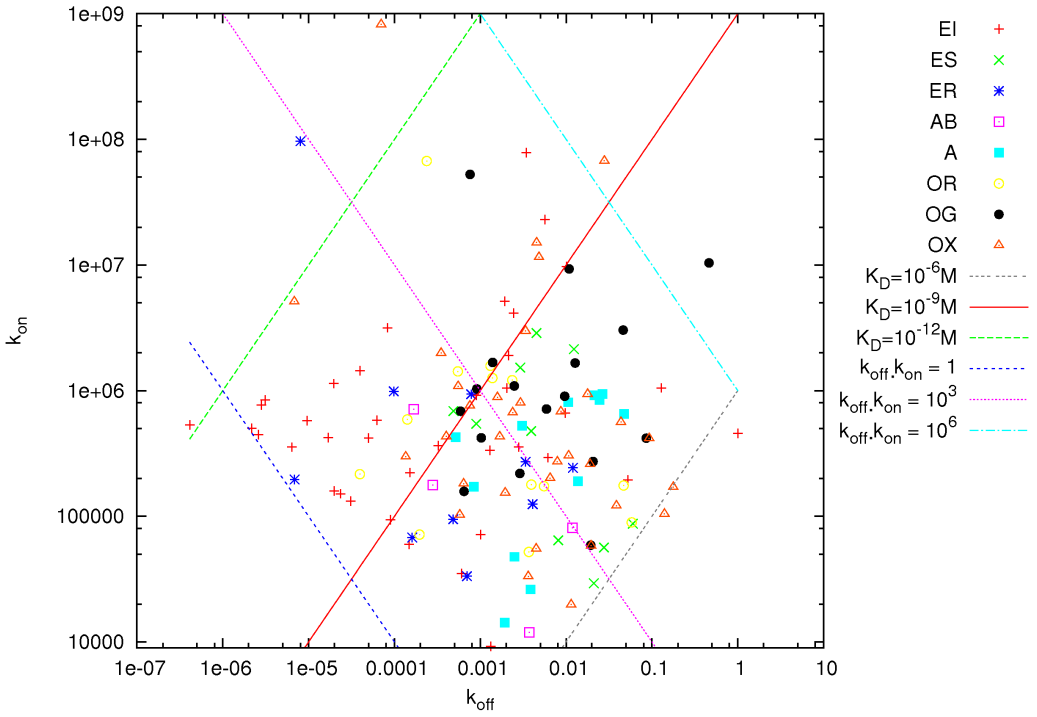

Figure 7.11: The k_{on} and k_{off} models derived from scheme 2 applied to the entire affinity benchmark. Categories are enzyme/inhibitor (EI), enzyme/substrate (ES), enzyme/accessory chain (ER), antibody/antigen with (A) and without (AB) the unbound structure of the antibody, interactions with receptors (OR), G-protein (OG) and miscellaneous interactions. Affinity isolines are shown, running from bottom left to top right. Thus, points in the top left correspond to high affinity complexes whilst those in the bottom right correspond to low affinity complexes. Isolines for the $k_{on}.k_{off}$ product are also shown running from top left to bottom right, which signify the relative contribution of association related and dissociation related phenomena to the overall binding affinity. Thus, complexes in the top right corner are dominate by association forces, whilst those near the dotted blue line in the bottom left corner have binding free energies which can be equally attributed to association and dissociation forces.

Chapter 8

Epilogue

In this thesis, I have approached two main problems; flexible docking and binding affinity prediction. As theoretical problems, they are interesting and test our understanding of the fundamental physics governing biomolecular recognition. However, it is the practical implications of their resolution which prompts interest from both the physical and biological sciences. In an ideal world, a molecular biologist would be able to select a protein and quickly find a structure, either experimental or modelled, and use it to locate and characterise binding partners. In turn, this may immediately reveal the biological processes in which that protein is involved. Atomic details regarding the structure of the protein and its binding partners can shed light on its function within that process; a tight interaction with a reactive site on an enzyme would imply inhibitory function. A more transient interaction with a kinase, with an appropriately positioned threonine or serine, would suggest a role in signal transduction. The possibilities for quickly obtaining such details, without ever having to leave a desktop computer, would hugely enrich the vast wealth of information being generated by cancer genomics and other initiatives which flag up disease-related genes. Such a vision for the future of biomedicine is no pipe dream; the computational power available to each generation dwarfs that of the previous, algorithms are becoming more sophisticated and, as more data become available, theoretical considerations ever more nuanced. The confluence of these forces cannot fail to provide incremental improvements in our ability to model molecular interactions.

8.1 Protein-Protein Docking

Protein-protein docking is a long-standing problem in molecular biology. Some complexes can be modelled with reasonable accuracy by treating the protein as rigid bodiess. However, for many interactions, such a treatment is untenable. Attempts to model flexibility explicitly requires additional degrees of freedom to be incorporated into the docking protocol. As low frequency elastic network normal modes have been shown to correspond to a wide variety of protein motions, their ability to capture the conformational changes across the whole fold and at the interface was evaluated (Chapter 2.1). Following this analysis, a protein-protein docking algorithm called SwarmDock was developed which incorporates flexibility using normal modes in linear combination (Chapter 3). The efficient particle swarm optimisation metaheuristic, in combination with a local search, was used to navigate search space. When tested on bound-bound docking, its performance is near perfect, although unbound-unbound docking still presents a challenge. The algorithm was tested with three different scoring functions, one of which, DComplex, could locate a complex within 5Å of the native for around 80% of cases, although less than a third of the complexes which did bind were ranked in the top 10 structures. The algorithm was shown to be enhanced when information derived from rigid-body Langevin dynamics simulations of encounter complex formation was used (Chapter 4), and it also performed competitively in the CAPRI experiment (Chapter 5). However, despite these modest successes, the algorithm requires much work. Future developments which are expected to further improve performance include the testing of other energy functions, such as statistical pair potentials trained on interactions as opposed to protein structures (Su *et al.*, 2009), the inclusion of a refinement and re-ranking stage, and methods of dealing with side-chain and/or loop flexibility. The incorporation of side-chain flexibility can be made possible using mean-fielding or other methods (Lindahl and Delarue, 2005). Further, conformational sampling methods for flexible loops, which keep the bond lengths and angles within physically plausible ranges, have been developed and could be incorporated as part of a Monte Carlo move set or, possibly, as variables in the PSO search space (Dinner, 2000; Betancourt, 2005; Go and Scheraga, 1970; Dodd, 1993; Vitalis and Pappu, 2009).

8.2 Binding Affinity Prediction

A large database which links the structures of unbound proteins and complexes to experimental data on binding affinities has been constructed (Chapter 6). From these structures, a large number of molecular descriptors were calculated to describe the physical and geometrical properties of the interactions, including some features relating to pH, conformational changes, disorder to order transitions and descriptors calculated on structural ensembles. These were then combined using a number of machine learning techniques in order to derive empirical binding free energy functions which, after testing with leave-one-out cross-validation, have shown to outperform previous empirical energy functions, even when the playing field is leveled by the removal of all descriptors pertaining to the unbound-bound conformational change (Chapter 7). The internal workings of these models was also shown to shed light on the relative importance of the various descriptors.

In addition, 44 of the complexes had kinetic data available for them in the scientific literature. A divide-and-conquer approach to binding affinity prediction was taken, in which the association and dissociation rate constants were first determined and then combined to derive the dissociation constant and the binding free energy. This revealed that the energetics of the unbound to bound conformational change was important for association, but not dissociation, lending support to the conformational sorting and population shift mechanism of protein binding.

Both the binding affinity prediction and the kinetic constant prediction methods highlight the use of approximate approaches, such as using pseudo-NMR restraints for deriving structural ensembles, and the calculation of disorder to order transitions entropy changes and vibrational entropy changes. Although the binding affinity of rigid proteins can be predicted with accuracy, work still needs to be done in order to account for the energetics of conformational changes. As more data become available, the divide-and-conquer approach shown here should be able to shed light on the problem by partitioning the binding free energy into entropy and enthalpy, or association energy and dissociation energy, and bringing into clearer focus the phenomena needing to be modelled.

8.3 Future Work

A number of projects have been undertaken as part of this thesis, and there is much scope for further work. In particular, the SwarmDock program should benefit significantly from being embedded within a pipeline, as typifies the most successful docking approaches today. One possibility is to employ a re-ranking method similar to that taken for the prediction of binding affinity; a descriptor set can be evaluated for all the generated structures and fed into machine learning models which either perform binary classification or generate scores which allows re-ranking of the structures. Between docking and re-ranking, a refinement step could be employed in order to nudge structures closer to the native. As only 200 or so clusters are generated by SwarmDock, as opposed to thousands of structures generated using Fourier transform methods, significant computational time could be allocated to each structure. Monte Carlo simulations or torsion angle dynamics could be employed for this purpose. Finally, the inability for normal coordinate flexibility to allow side-chains to swap rotamer, or rotate around bonds, suggests that the algorithm could benefit from only applying normal modes to backbone motions, and handling side chain motions with a method such as the mean field approach.

Much further work can also be made to the prediction of binding affinities, and in the application of this approach further afield. The descriptor set described in section 7.2 should be automated, so that they can be easily determined and used for other purposes. These include re-ranking of docked poses, the identification of biological interfaces from crystal interfaces in x-ray structures, and computational mutagenesis. Ultimately, these applications and the selection of *de novo* designed complexes are related to binding free energy calculation, and the learners employed will likely contain mutually useful information which can be transferred from one problem to the other. Recent developments in the field machine learning, such as hierarchical Bayesian models, are beginning to allow such transfer of information, a development which could greatly benefit these problems. Further, highly efficient versions of the binding affinity models could allow an unprecedented study of interaction evolution by extensively mapping out how binding affinity changes with sequence. Not only could ancestral

protein sequence reconstruction methods be used to study the evolution of interactions by determining the affinity of intermediates connecting modern proteins that bind to ancestral proteins that do not, but it would also be possible to map out the functionally viable paths connecting orthologs and locate the pathways though which protein interactions can change over time.

8.4 Concluding Remarks

The normal processes of life and the misregulation of these processes in cancer boil down to molecules influencing one another; the machinery of morphogenesis, cell signalling and adhesion, apoptosis, DNA repair and the cell cycle are but a few examples for which a lot is known, yet a lot remains to be discovered. DNA and protein sequencing, along with NMR, crystallography and other physical techniques, have acquainted us with the molecules of life. Genetics and cellular biology have told us something of their roles in living systems. With the information gleaned from these disciplines and others it is possible, in principle, to put the pieces of the puzzle together and build a network which would act as a circuit diagram connecting the genes and molecules for any living cell, be it healthy or cancerous, and to use that knowledge to improve the human condition and relieve suffering. It is physics which can determine the exact nature of the molecular interaction, and physical models will play a key role in turning a list of parts into a model of life processes, and in the structural, thermodynamic and kinetic annotation of interaction and transcription networks.

We live imprisoned within our heads; all that we touch, smell, see and hear is a reconstruction of reality based on the meagre information for which the blind search of evolution has endowed us the ability to detect; the light that enters our eyes, the vibrations passing our ears and the molecules whose stochastic meanderings have lead them to our nostrils. Beyond our perceptions lies a seemingly unfaltering external reality. This reality is the stage within which all of our lives are played out. It is the stage in which atoms in flux make the temporary associations which choreograph all life

processes from the transit of sperm to the egg, to the final electrical impulse which marks our last heartbeat. Physics is the art of making sense of this stage, uncovering the rules which govern it, and showing how reality extends far beyond what is immediately discernible by our senses. It is my intent to use the knowledge of this stage, and the relationship between free energy and molecular structure, to shed light on life and disease. I hope that at least some of the ideas which underpinned the design of SwarmDock will play a part in determining the role of proteins in their cellular context and in reconstructing the circuitry of life. I hope that one day the approach taken to binding affinity prediction, or the affinity benchmark, will play a role in the characterisation of interactomes.

Appendix A

EEF1 Results

Performance of bound-bound docking with and without EEF1 desolvation are shown. The lowest ligand RMSD after superimposing the bound receptors (RMSD), the best ranked structure with a ligand RMSD below 5Å (Rank), and the number of times a structure was found with ligand RMSD below 5Å (# found), are shown. For each complex and for each performance metric, the better performance is highlighted in bold. A discussion of these results appears in section 3.3.3.

Table A.1: Performance of bound-bound docking with and without EEF1 desolvation

	VDW + Elec			VDW + Elec + EEF1		
Complex	Rank	# found	RMSD	Rank	# found	RMSD
1a2k	1	**17**	**0.11**	1	14	0.19
1acb	**1**	**2**	**0.59**	N/A	0	7.68
1ahw	**1**	**23**	**0.39**	N/A	0	17.61
1ak4	1	7	0.35	1	**14**	**0.20**
1akj	**1**	**6**	**0.07**	N/A	0	18.47
1atn	1	2	0.85	1	**10**	**0.41**
1ay7	1	**54**	**0.26**	1	19	0.36
1b6c	**1**	**2**	**0.47**	N/A	0	16.80
1bj1	1	**3**	**0.30**	1	2	1.62
1buh	1	**55**	**0.11**	1	37	0.24
1bvk	**1**	**38**	**0.15**	12	5	0.31

Table A.1 continued on next page

Table A.1 continued

1bvn	**1**	**14**	**0.31**	N/A	0	14.81
1cgi	1	3	**0.47**	1	**4**	0.62
1dfj	**1**	**1**	**0.66**	N/A	0	25.75
1dqj	**1**	**29**	**0.32**	N/A	0	15.51
1e6e	N/A	0	8.62	N/A	0	**6.72**
1e6j	1	15	**0.09**	1	**37**	0.11
1e96	1	29	**0.16**	1	**54**	0.24
1eer	**1**	**1**	**0.20**	N/A	0	23.99
1ewy	**116**	**12**	**1.58**	N/A	0	15.05
1f34	1	**21**	**0.08**	1	2	1.19
1f51	1	**9**	**0.29**	1	2	0.86
1fc2	5	**18**	**0.22**	**2**	17	0.31
1fq1	1	**24**	**0.75**	1	6	0.79
1fqj	1	**25**	**0.31**	1	1	0.48
1fsk	1	**15**	**0.17**	1	3	0.54
1gcq	1	**26**	**0.19**	1	18	0.29
1ghq	169	5	0.62	**66**	**8**	**0.22**
1gp2	1	**13**	**0.24**	1	4	0.32
1grn	1	**9**	**0.14**	1	1	0.27
1h1v	1	**12**	**0.17**	1	10	0.43
1he1	1	**21**	0.12	1	10	**0.06**
1he8	1	**18**	**0.13**	1	12	0.32
1i2m	1	**31**	**0.21**	1	4	0.69
1i4d	1	25	**0.33**	1	**40**	0.52
1i9r	1	**30**	**0.26**	1	14	0.38
1ib1	1	**3**	**0.15**	1	1	0.47
1ibr	**1**	**1**	**0.45**	N/A	0	24.52
1ijk	1	**26**	**0.25**	1	16	0.59
1jps	1	**25**	**0.23**	1	4	1.03
1k4c	1	**40**	**0.07**	1	22	0.22
1k5d	**1**	**9**	**0.24**	N/A	0	20.86
1kac	1	**61**	**0.13**	1	21	0.22
1kkl	**32**	**20**	**0.18**	121	4	0.84

Table A.1 continued on next page

Table A.1 continued

1klu	1	**23**	**0.11**	1	22	0.31
1ktz	1	**37**	**0.07**	1	23	0.15
1kxp	1	**3**	**0.29**	1	1	0.33
1kxq	**12**	**6**	**2.93**	94	3	3.57
1m10	1	**30**	**0.40**	1	5	0.49
1mah	1	**13**	**0.32**	1	8	0.41
1ml0	1	**35**	**0.05**	1	20	0.11
1mlc	**1**	25	**0.33**	2	**27**	0.45
1qa9	**6**	**48**	**0.69**	N/A	0	18.72
1qfw	**1**	**18**	**0.21**	31	9	0.87
1rlb	**15**	**6**	**0.34**	165	2	1.22
1tmq	**1**	**2**	**0.11**	N/A	0	19.22
1udi	1	29	**0.22**	1	**48**	0.25
1vfb	1	**46**	**0.21**	1	27	0.28
1wej	1	**51**	**0.21**	1	19	0.35
1wq1	**1**	**6**	**0.33**	N/A	0	13.53
2btf	1	**34**	**0.12**	1	3	0.59
2hmi	N/A	0	23.12	**53**	**3**	**1.09**
2mta	27	27	**0.22**	**1**	**43**	0.34
2pcc	**218**	**8**	**0.56**	N/A	0	16.53
2sic	**1**	**3**	**0.38**	N/A	0	8.11
2sni	1	**4**	**0.27**	1	1	0.70
2vis	1	6	**0.19**	1	**8**	0.28
7cei	**1**	**34**	**0.19**	5	8	0.44

Appendix B

Rigid-Body Unbound-Unbound Results

The performance of rigid-body unbound-unbound docking using the electrostatics and Van der Waals energy function, clustering at 3.5Å resolution, is shown. All structures in the benchmark v2.0 are docked (Mintseris *et al.*, 2005). A number of performance metrics are evaluated. For the lowest IRMSD first cluster member, the cluster rank (Rank), the cluster size (# found), interface RMSD (IRMSD), ligand RMSD (LRMSD), fraction of native contacts (FNat), fraction of non-native contacts (FNNat), and whether this is the largest cluster (bCluster) are shown. Also, the best IRMSD (bIRMSD), LRMSD (bLRMSD), FNat (bFNat) and FNNat (bFNNat) found, with the their corresponding cluster rank and size in parentheses, are reported, as is the best ranked high, medium and acceptable solution (bRank). All metrics are evaluated per the CAPRI standards (see section 1.5.10 or Mendez *et al.* (2003)).

Table B.1: The performance of rigid-body unbound-unbound docking.

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1A2K	184	1	no	6.29	16.74	0.06	0.88	6.29(184/1)	9.16(34/2)	0.09(40/1)	0.83(189/1)	-/-/-
1ACB	69	2	no	6.11	11.01	0.00	1.00	6.11(69/1)	8.05(16/1)	0.06(72/1)	0.83(72/1)	-/-/-
1AHW	96	1	no	1.35	3.86	0.47	0.11	1.30(33/1)	1.72(9/3)	0.52(9/3)	0.11(96/1)	-/9/9
1AK4	160	1	no	7.88	24.79	0.02	0.97	7.88(160/1)	11.08(15/2)	0.14(52/1)	0.79(52/1)	-/-/-
1AKJ	171	1	no	6.82	14.62	0.18	0.56	6.82(171/1)	11.50(50/1)	0.18(171/1)	0.56(171/1)	-/-/-
1ATN	145	1	no	6.17	9.40	0.04	0.90	6.17(145/1)	7.40(114/1)	0.09(7/1)	0.85(28/1)	-/-/-
1AVX	62	4	no	1.58	3.50	0.49	0.11	1.56(62/4)	1.56(62/4)	0.49(62/4)	0.11(62/4)	-/62/62
1AY7	23	16	yes	0.96	1.87	0.60	0.14	0.94(23/16)	0.94(23/16)	0.65(23/16)	0.14(23/16)	23/23/23
1B6C	99	1	no	6.13	10.12	0.05	0.91	6.13(99/1)	6.13(99/1)	0.07(105/1)	0.90(105/1)	-/-/-
1BGX	103	1	no	7.06	23.54	0.06	0.69	7.06(103/1)	10.40(42/1)	0.06(103/1)	0.69(103/1)	-/-/-
1BJ1	67	3	no	7.82	22.50	0.00	1.00	7.65(67/3)	18.03(1/3)	0.03(64/1)	0.95(64/3)	-/-/-
1BUH	131	3	no	1.95	5.77	0.34	0.48	1.90(131/3)	1.90(131/3)	0.34(131/3)	0.48(131/3)	-/131/131
1BVK	144	1	no	3.52	5.17	0.19	0.68	3.52(144/1)	3.52(144/1)	0.19(2/3)	0.68(144/1)	-/-/144
1BVN	69	2	no	6.86	13.95	0.00	1.00	6.86(69/2)	7.24(33/1)	0.01(19/2)	0.96(20/2)	-/-/-
1CGI	130	2	no	4.53	10.88	0.09	0.64	4.53(130/2)	6.67(20/1)	0.09(20/1)	0.64(130/2)	-/-/-
1D6R	27	1	no	3.46	9.38	0.33	0.50	3.46(27/1)	4.94(3/2)	0.33(27/1)	0.50(27/1)	-/-/27
1DE4	130	1	no	6.13	14.71	0.04	0.89	6.13(130/1)	6.29(1/3)	0.09(1/3)	0.87(1/1)	-/-/-
1DFJ	123	1	no	5.48	19.64	0.12	0.73	5.48(123/1)	5.48(123/1)	0.12(123/1)	0.73(123/1)	-/-/-
1DQJ	100	1	no	3.03	8.13	0.30	0.36	3.03(100/1)	4.71(61/4)	0.48(38/2)	0.21(38/1)	-/-/38
1E6E	130	2	no	1.78	3.45	0.40	0.22	1.78(130/2)	1.78(130/2)	0.40(130/2)	0.22(130/2)	-/130/130
1E6J	36	7	no	3.35	7.81	0.22	0.62	3.32(36/7)	3.32(36/7)	0.24(36/7)	0.58(36/7)	-/-/36

Table B.1 continued on next page

Table B.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1E96	185	1	no	4.56	14.42	0.28	0.48	4.56(185/1)	5.51(19/3)	0.28(185/1)	0.48(185/1)	-/-/-
1EAW	96	1	no	6.60	19.75	0.05	0.79	6.60(96/1)	7.45(29/1)	0.09(58/1)	0.76(58/1)	-/-/-
1EER	97	1	no	4.37	8.73	0.13	0.59	4.37(97/1)	5.76(33/2)	0.13(97/1)	0.59(97/1)	-/-/97
1EWY	80	7	no	1.85	4.85	0.51	0.42	1.84(80/1)	3.98(70/1)	0.53(80/7)	0.41(150/1)	-/80/54
1EZU	149	1	no	7.06	17.82	0.01	0.96	7.06(149/1)	16.18(29/1)	0.02(40/1)	0.94(40/1)	-/-/-
1F34	143	1	no	2.89	9.22	0.25	0.47	2.89(143/1)	7.06(128/1)	0.25(143/1)	0.47(143/1)	-/-/143
1F51	41	2	no	5.58	11.55	0.05	0.93	5.58(41/1)	9.47(4/2)	0.08(22/2)	0.83(84/1)	-/-/-
1FAK	49	1	no	9.90	20.15	0.04	0.88	9.90(49/1)	15.68(5/1)	0.04(15/1)	0.88(49/1)	-/-/-
1FC2	163	1	no	7.45	18.67	0.00	1.00	7.45(163/1)	11.16(58/1)	0.12(98/1)	0.75(98/1)	-/-/-
1FQ1	57	1	no	7.71	33.78	0.02	0.97	7.71(57/1)	16.27(1/4)	0.04(92/2)	0.93(142/1)	-/-/-
1FQJ	27	10	no	2.63	7.32	0.40	0.28	2.34(27/10)	2.34(27/10)	0.40(27/10)	0.28(27/10)	-/-/27
1FSK	72	1	no	3.73	10.39	0.38	0.31	3.73(72/1)	4.72(3/4)	0.38(72/1)	0.26(143/1)	-/-/72
1GCQ	44	2	no	6.29	19.38	0.00	1.00	6.29(44/1)	12.46(1/3)	0.02(108/1)	0.88(108/1)	-/-/-
1GHQ	130	1	no	4.55	14.01	0.00	1.00	4.55(130/1)	4.55(130/1)	0.20(160/1)	0.81(160/1)	-/-/-
1GP2	3	3	no	6.07	22.43	0.02	0.97	5.84(3/1)	5.84(3/3)	0.07(127/1)	0.86(127/1)	-/-/-
1GRN	69	4	no	1.84	4.39	0.25	0.45	1.59(44/22)	1.59(44/22)	0.38(44/22)	0.24(44/22)	-/44/44
1H1V	231	1	no	10.29	21.45	0.00	1.00	10.29(231/1)	12.54(30/2)	0.02(139/1)	0.97(139/1)	-/-/-
1HE1	153	5	no	10.22	21.84	0.00	1.00	10.19(153/5)	16.00(1/1)	0.03(33/1)	0.78(140/5)	-/-/-
1HE8	93	4	no	1.43	2.26	0.76	0.27	1.04(93/4)	1.04(93/4)	0.76(93/4)	0.21(93/4)	93/93/93
1HIA	85	1	no	5.85	14.70	0.16	0.57	5.85(85/1)	11.21(1/1)	0.16(85/1)	0.57(85/1)	-/-/-
1I2M	4	17	yes	3.81	8.08	0.16	0.69	3.65(4/17)	3.65(4/17)	0.16(4/17)	0.65(4/17)	-/-/4
1I4D	14	3	no	6.50	12.39	0.04	0.94	6.50(14/1)	7.33(1/4)	0.09(15/1)	0.81(172/1)	-/-/-
1I9R	1	33	yes	1.54	1.95	0.80	0.16	1.45(1/33)	1.45(1/33)	0.83(1/33)	0.02(1/33)	1/1/1
1IB1	146	1	no	5.58	10.58	0.05	0.86	5.58(146/1)	5.58(146/1)	0.09(141/2)	0.72(141/1)	-/-/-

Table B.1 continued on next page

Table B.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1IBR	108	1	no	8.28	22.32	0.03	0.88	8.28(108/1)	21.42(1/2)	0.03(108/1)	0.88(108/1)	-/-/-
1IJK	87	1	no	5.39	9.25	0.06	0.92	5.39(87/1)	7.98(10/2)	0.06(87/1)	0.92(87/1)	-/-/-
1IQD	57	2	no	5.46	19.06	0.12	0.75	5.46(57/2)	7.18(12/1)	0.12(57/2)	0.75(57/2)	-/-/-
1JPS	23	1	no	2.07	8.44	0.41	0.26	2.07(23/1)	2.07(23/1)	0.41(23/1)	0.26(23/1)	-/-/23
1K4C	171	1	no	5.78	15.96	0.08	0.81	5.78(171/1)	11.61(12/1)	0.08(171/1)	0.60(91/1)	-/-/-
1K5D	9	2	no	5.20	19.90	0.03	0.93	5.20(9/1)	9.15(1/1)	0.09(134/1)	0.74(134/1)	-/-/-
1KAC	3	19	yes	2.22	5.38	0.34	0.55	2.19(3/19)	2.19(3/19)	0.38(3/19)	0.51(3/19)	-/3/3
1KKL	121	1	no	5.19	9.02	0.17	0.70	5.19(121/1)	5.19(121/1)	0.17(121/1)	0.70(121/1)	-/-/121
1KLU	76	1	no	9.32	29.98	0.00	1.00	9.32(76/1)	12.47(20/1)	0.14(111/1)	0.86(111/1)	-/-/-
1KTZ	141	1	no	5.88	23.60	0.07	0.91	5.88(141/1)	15.04(1/2)	0.20(15/6)	0.82(15/1)	-/-/-
1KXP	191	1	no	4.97	22.25	0.05	0.83	4.97(191/1)	4.97(191/1)	0.05(191/1)	0.83(191/1)	-/-/-
1KXQ	150	1	no	7.12	12.65	0.06	0.85	7.12(150/1)	7.12(150/1)	0.06(150/1)	0.85(150/1)	-/-/-
1M10	62	2	no	4.86	12.29	0.08	0.84	4.82(62/2)	5.79(35/1)	0.08(62/2)	0.83(62/2)	-/-/-
1MAH	156	1	no	1.79	4.38	0.28	0.31	1.79(156/1)	1.79(156/1)	0.28(156/1)	0.31(156/1)	-/-/156
1ML0	184	1	no	2.27	4.85	0.27	0.33	2.27(184/1)	2.27(184/1)	0.29(80/1)	0.33(184/1)	-/-/80
1MLC	52	1	no	7.01	18.94	0.00	1.00	7.01(52/2)	11.18(18/1)	0.12(55/2)	0.74(55/2)	-/-/-
1N2C	27	4	no	7.31	18.12	0.00	1.00	7.22(27/1)	7.22(27/4)	0.01(52/1)	0.96(52/1)	-/-/-
1NCA	1	26	yes	0.40	0.84	0.74	0.07	0.36(1/26)	0.36(1/26)	0.79(1/26)	0.02(1/26)	1/1/1
1NSN	45	2	no	2.77	7.13	0.31	0.62	2.77(45/2)	2.77(45/2)	0.31(45/2)	0.62(45/2)	-/-/45
1PPE	122	1	no	3.69	9.97	0.13	0.73	3.69(122/1)	7.20(1/6)	0.13(122/1)	0.73(122/1)	-/-/122
1QA9	169	2	no	1.19	2.59	0.62	0.21	1.00(169/2)	1.00(169/2)	0.62(169/2)	0.12(169/2)	169/169/85
1QFW	5	14	yes	0.91	2.92	0.70	0.00	0.79(5/14)	0.79(5/14)	0.72(5/14)	0.00(5/14)	5/5/5
1RLB	52	1	no	9.54	30.07	0.00	1.00	9.54(52/1)	12.61(138/1)	0.02(185/1)	0.94(185/1)	-/-/-
1SBB	135	1	no	6.10	33.22	0.03	0.95	6.10(135/1)	11.79(10/4)	0.12(51/1)	0.87(169/1)	-/-/-

Table B.1 continued on next page

Table B.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1TMQ	83	1	no	3.19	8.27	0.17	0.64	3.19(83/1)	3.19(83/1)	0.17(83/1)	0.64(83/1)	-/-/83
1UDI	179	1	no	7.27	17.55	0.00	1.00	7.27(179/1)	8.59(35/1)	0.03(35/1)	0.94(35/1)	-/-/-
1VFB	150	1	no	1.33	2.57	0.55	0.04	1.29(98/4)	1.29(98/4)	0.57(98/4)	0.04(98/4)	-/98/98
1WEJ	115	2	no	4.85	10.57	0.05	0.90	4.85(115/1)	6.57(9/4)	0.21(132/1)	0.73(132/1)	-/-/-
1WQ1	31	6	yes	2.97	6.46	0.20	0.38	2.79(31/6)	2.79(31/6)	0.20(31/6)	0.37(31/6)	-/-/31
2BTF	131	1	no	6.72	18.06	0.02	0.96	6.72(131/1)	6.93(4/3)	0.08(172/1)	0.79(172/1)	-/-/-
2HMI	146	1	no	8.64	36.37	0.00	1.00	8.64(146/1)	9.79(4/1)	0.03(4/1)	0.98(4/1)	-/-/-
2JEL	79	1	no	4.55	17.09	0.09	0.79	4.55(79/1)	8.38(6/2)	0.11(11/1)	0.79(79/1)	-/-/-
2MTA	38	3	no	4.65	14.11	0.17	0.77	4.53(38/3)	11.07(5/4)	0.17(38/3)	0.77(38/3)	-/-/-
2PCC	37	4	no	3.79	8.16	0.21	0.76	3.77(37/4)	3.77(37/4)	0.21(37/4)	0.75(37/4)	-/-/37
2QFW	109	3	no	1.33	3.48	0.63	0.17	1.29(109/3)	1.29(109/3)	0.63(109/3)	0.11(109/3)	-/109/109
2SIC	60	1	no	6.14	22.60	0.01	0.96	6.14(60/1)	6.47(30/1)	0.03(8/2)	0.93(134/1)	-/-/-
2SNI	49	1	no	4.49	15.05	0.21	0.33	4.49(49/1)	8.70(1/5)	0.21(49/1)	0.33(49/1)	-/-/-
2VIS	129	1	no	3.16	16.11	0.18	0.76	3.16(129/1)	3.16(129/1)	0.18(129/1)	0.76(129/1)	-/-/129
7CEI	21	1	no	1.38	3.49	0.50	0.13	1.38(21/1)	1.62(5/22)	0.50(5/22)	0.07(7/1)	-/5/5

Appendix C

Flexible Unbound-Unbound Results

The performance of flexible unbound-unbound docking using the electrostatics and Van der Waals energy function, clustering at 3.5Å resolution, is shown. The five lowest frequency non-trivial normal modes were used in both the receptor and ligand. All structures in the benchmark v2.0 are docked (Mintseris *et al.*, 2005). A number of performance metrics are evaluated. For the lowest IRMSD first cluster member, the cluster rank (Rank), the cluster size (# found), interface RMSD (IRMSD), ligand RMSD (LRMSD), fraction of native contacts (FNat), fraction of non-native contacts (FNNat), and whether this is the largest cluster (bCluster) are shown. Also, the best IRMSD (bIRMSD), LRMSD (bLRMSD), FNat (bFNat) and FNNat (bFNNat) found, with the their corresponding cluster rank and size in parentheses, are reported, as is the best ranked high, medium and acceptable solution (bRank). All metrics are evaluated per the CAPRI standards (see section 1.5.10 or Mendez *et al.* (2003)).

Table C.1: The performance of flexible unbound-unbound docking.

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1A2K	21	2	no	5.44	15.45	0.00	1.00	5.40(21/1)	5.40(21/2)	0.15(219/1)	0.69(219/1)	-/-/-
1ACB	124	1	no	4.62	10.25	0.09	0.81	4.62(124/1)	5.72(11/2)	0.09(124/1)	0.81(124/1)	-/-/-
1AHW	56	1	no	1.44	4.55	0.48	0.12	1.44(56/1)	1.78(6/1)	0.48(56/1)	0.12(56/1)	-/6/6
1AK4	181	1	no	6.78	27.15	0.02	0.96	6.78(181/1)	8.94(6/1)	0.07(74/1)	0.90(74/1)	-/-/-
1AKJ	31	1	no	7.27	11.80	0.00	1.00	7.27(31/1)	7.53(16/1)	0.05(74/1)	0.86(74/1)	-/-/-
1ATN	182	1	no	5.47	14.80	0.07	0.84	5.47(182/1)	6.19(130/1)	0.10(51/1)	0.71(238/1)	-/-/-
1AVX	25	1	no	1.59	3.37	0.48	0.11	1.59(25/1)	1.59(25/1)	0.48(25/1)	0.11(25/1)	-/25/25
1AY7	20	6	yes	1.13	2.44	0.55	0.15	0.87(20/6)	0.87(20/6)	0.68(20/6)	0.11(20/6)	20/20/2
1B6C	148	1	no	4.59	11.42	0.07	0.88	4.59(148/1)	4.59(148/1)	0.07(48/2)	0.84(48/1)	-/-/-
1BGX	37	2	no	6.96	18.71	0.01	0.93	6.96(37/1)	15.33(23/1)	0.04(105/1)	0.74(105/1)	-/-/-
1BJ1	185	1	no	2.99	13.08	0.13	0.65	2.99(185/1)	11.85(18/1)	0.14(154/1)	0.65(185/1)	-/-/154
1BUH	60	6	yes	1.99	6.73	0.36	0.54	1.92(60/6)	1.92(60/6)	0.45(60/6)	0.45(60/6)	-/60/60
1BVK	133	1	no	4.44	8.28	0.06	0.87	4.44(133/1)	7.42(88/2)	0.17(2/2)	0.78(2/1)	-/-/-
1BVN	122	1	no	4.41	9.65	0.04	0.89	4.41(122/1)	4.41(122/1)	0.10(104/1)	0.79(104/1)	-/-/-
1CGI	127	1	no	5.09	12.04	0.07	0.75	5.09(127/1)	9.03(14/2)	0.09(15/2)	0.75(127/1)	-/-/-
1D6R	85	1	no	4.85	11.75	0.09	0.83	4.85(85/1)	5.05(19/3)	0.12(106/1)	0.73(106/1)	-/-/-
1DE4	87	1	no	7.58	31.07	0.00	1.00	7.58(87/1)	23.59(1/1)	0.04(37/1)	0.91(37/1)	-/-/-
1DFJ	34	1	no	5.86	21.31	0.15	0.75	5.86(34/1)	17.52(1/1)	0.15(34/1)	0.75(34/1)	-/-/-
1DQJ	47	1	no	3.71	11.74	0.45	0.30	3.71(47/1)	4.90(42/1)	0.45(47/1)	0.30(47/1)	-/-/44
1E6E	94	2	no	1.77	3.48	0.44	0.23	1.75(94/2)	1.75(94/2)	0.44(94/2)	0.19(94/2)	-/94/94
1E6J	6	3	no	3.54	7.92	0.20	0.67	3.44(6/1)	3.44(6/3)	0.30(145/1)	0.52(145/1)	-/-/6

Table C.1 continued on next page

Table C.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1E96	180	2	no	3.20	6.95	0.23	0.70	3.10(180/2)	3.10(180/2)	0.25(180/2)	0.60(180/2)	-/-/180
1EAW	1	1	no	1.87	3.48	0.44	0.39	1.87(1/1)	1.87(1/1)	0.44(1/1)	0.39(1/1)	-/1/1
1EER	74	1	no	8.51	24.85	0.04	0.86	8.51(74/1)	12.74(6/1)	0.04(74/1)	0.86(74/1)	-/-/-
1EWY	13	3	no	1.90	5.11	0.53	0.51	1.84(13/3)	1.84(13/3)	0.56(13/3)	0.50(13/3)	-/13/13
1EZU	180	1	no	8.61	19.21	0.00	1.00	8.61(180/1)	10.43(119/1)	0.02(112/1)	0.93(112/1)	-/-/-
1F34	16	2	no	3.29	6.41	0.14	0.60	3.26(16/3)	3.26(16/2)	0.19(64/3)	0.40(64/3)	-/-/16
1F51	184	1	no	4.76	8.73	0.23	0.48	4.76(184/1)	5.25(24/1)	0.23(184/1)	0.48(184/1)	-/-/184
1FAK	46	1	no	14.41	43.09	0.00	1.00	14.41(46/1)	14.86(3/1)	0.02(130/1)	0.90(136/1)	-/-/-
1FC2	176	1	no	7.80	39.96	0.00	1.00	7.80(176/1)	9.64(42/1)	0.07(45/1)	0.88(82/1)	-/-/-
1FQ1	136	2	no	7.77	20.11	0.18	0.68	7.73(136/2)	9.51(19/2)	0.20(136/2)	0.67(136/2)	-/-/-
1FQJ	11	2	no	2.39	6.76	0.36	0.37	2.26(11/2)	2.26(11/2)	0.40(11/2)	0.36(11/2)	-/-/11
1FSK	65	1	no	1.44	3.03	0.53	0.08	1.44(65/1)	1.44(65/1)	0.53(65/1)	0.08(65/1)	-/65/4
1GCQ	172	1	no	1.40	3.45	0.40	0.05	1.40(172/1)	1.40(172/1)	0.40(172/1)	0.05(172/1)	-/172/172
1GHQ	87	1	no	5.70	14.18	0.04	0.96	5.70(87/1)	7.03(5/1)	0.20(129/1)	0.81(129/1)	-/-/-
1GP2	11	1	no	3.88	10.99	0.05	0.91	3.88(11/1)	9.78(1/1)	0.10(106/1)	0.81(106/1)	-/-/-
1GRN	24	3	no	1.65	3.26	0.40	0.31	1.58(24/3)	1.75(17/4)	0.40(17/4)	0.29(17/3)	-/17/17
1H1V	82	1	no	9.77	17.38	0.00	1.00	9.77(82/1)	15.29(1/1)	0.03(189/1)	0.92(189/1)	-/-/-
1HE1	209	1	no	3.81	14.34	0.15	0.61	3.81(209/1)	5.95(118/2)	0.15(209/1)	0.61(209/1)	-/-/209
1HE8	46	1	no	1.45	2.76	0.79	0.32	1.45(46/1)	1.45(46/1)	0.79(46/1)	0.32(46/1)	-/46/46
1HIA	2	1	no	2.61	5.75	0.40	0.65	2.61(2/1)	2.61(2/1)	0.40(2/1)	0.50(141/1)	-/-/2
1I2M	2	7	yes	3.84	7.87	0.13	0.73	3.77(2/1)	3.77(2/7)	0.16(2/7)	0.60(46/1)	-/-/2
1I4D	66	1	no	4.71	19.96	0.13	0.75	4.71(66/1)	8.85(2/2)	0.13(66/1)	0.75(66/1)	-/-/-
1I9R	1	10	yes	2.33	2.56	0.67	0.35	1.48(1/10)	1.48(1/10)	0.84(1/10)	0.11(1/10)	1/1/1
1IB1	23	1	no	5.84	10.87	0.07	0.78	5.84(23/1)	5.84(23/1)	0.07(23/1)	0.78(23/1)	-/-/-

Table C.1 continued on next page

Table C.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1IBR	172	1	no	9.01	21.68	0.00	1.00	9.01(172/1)	10.74(21/1)	0.03(29/1)	0.88(29/1)	-/-/-
1IJK	8	2	no	8.04	23.16	0.00	1.00	7.81(8/2)	7.81(8/2)	0.00(-1/0)	1.00(-1/2)	-/-/-
1IQD	58	1	no	5.60	18.79	0.12	0.76	5.60(58/1)	7.15(12/2)	0.12(58/1)	0.76(58/1)	-/-/-
1JPS	99	2	no	1.57	4.89	0.52	0.05	1.57(99/2)	3.08(19/1)	0.52(99/2)	0.05(99/2)	-/23/19
1K4C	170	1	no	4.60	11.94	0.06	0.79	4.60(170/1)	11.93(6/1)	0.06(170/1)	0.79(170/1)	-/-/-
1K5D	31	1	no	3.14	7.39	0.11	0.70	3.14(31/1)	3.14(31/1)	0.11(31/1)	0.70(31/1)	-/-/31
1KAC	1	9	yes	2.33	5.22	0.36	0.57	2.24(1/9)	2.24(1/9)	0.38(1/9)	0.51(1/9)	-/1/1
1KKL	33	1	no	4.01	4.79	0.13	0.78	4.01(33/1)	4.01(33/1)	0.19(60/1)	0.70(60/1)	-/-/33
1KLU	168	1	no	5.24	12.38	0.03	0.95	5.24(168/1)	9.32(33/1)	0.17(161/1)	0.78(161/1)	-/-/-
1KTZ	114	6	no	0.84	1.63	0.83	0.11	0.66(114/6)	0.66(114/6)	0.83(114/6)	0.04(114/6)	114/114/114
1KXP	126	1	no	3.28	12.65	0.12	0.41	3.28(126/1)	4.73(107/1)	0.12(126/1)	0.41(126/1)	-/-/126
1KXQ	99	1	no	3.06	7.06	0.27	0.41	3.06(99/1)	3.12(57/1)	0.27(99/1)	0.41(99/1)	-/-/57
1M10	70	1	no	5.71	11.58	0.03	0.93	5.71(70/1)	5.71(70/1)	0.03(39/2)	0.92(139/1)	-/-/-
1MAH	61	1	no	5.65	17.75	0.14	0.70	5.65(61/1)	5.74(11/4)	0.14(61/1)	0.70(61/1)	-/-/-
1ML0	5	1	no	1.55	2.52	0.58	0.18	1.55(5/1)	1.55(5/1)	0.58(5/1)	0.18(5/1)	-/5/5
1MLC	48	1	no	3.83	9.82	0.18	0.78	3.83(48/1)	9.22(4/3)	0.18(48/1)	0.70(145/1)	-/-/48
1N2C	118	1	no	4.90	10.45	0.12	0.60	4.90(118/1)	5.18(108/1)	0.12(118/1)	0.60(118/1)	-/-/-
1NCA	1	6	yes	0.55	1.12	0.72	0.07	0.39(1/6)	0.39(1/6)	0.79(1/6)	0.03(1/6)	1/1/1
1NSN	20	2	no	2.52	6.26	0.41	0.55	2.52(20/2)	2.52(20/2)	0.41(20/2)	0.55(20/2)	-/-/20
1PPE	62	1	no	2.47	6.25	0.30	0.25	2.47(62/1)	2.47(62/1)	0.30(62/1)	0.25(62/1)	-/-/62
1QA9	159	3	no	1.17	2.25	0.62	0.23	1.16(159/3)	1.16(159/3)	0.62(159/3)	0.23(159/3)	-/159/35
1QFW	38	1	no	0.82	2.63	0.70	0.03	0.82(38/1)	0.83(5/3)	0.79(5/3)	0.00(5/1)	5/5/5
1RLB	109	1	no	7.96	22.62	0.00	1.00	7.90(106/2)	11.84(10/4)	0.07(18/2)	0.90(64/2)	-/-/-
1SBB	151	1	no	7.43	35.79	0.03	0.97	7.43(151/1)	8.74(62/1)	0.03(90/1)	0.96(90/1)	-/-/-

Table C.1 continued on next page

Table C.1 continued from previous page

Complex	Rank	# found	bCluster	IRMSD	LRMSD	FNat	FNNat	bIRMSD	bLRMSD	bFNat	bFNNat	bRank
1TMQ	72	3	no	2.59	6.29	0.21	0.57	2.59(72/3)	2.59(72/3)	0.21(72/3)	0.57(72/3)	-/-/72
1UDI	85	2	no	5.77	13.38	0.13	0.69	5.47(85/2)	9.09(50/1)	0.16(85/2)	0.65(85/2)	-/-/-
1VFB	118	1	no	1.40	4.23	0.57	0.04	1.40(118/1)	1.40(118/1)	0.57(118/1)	0.04(118/1)	-/118/118
1WEJ	120	1	no	2.75	5.44	0.40	0.32	2.75(120/1)	2.75(120/1)	0.40(120/1)	0.32(120/1)	-/-/120
1WQ1	14	1	no	2.92	4.98	0.21	0.41	2.92(14/1)	2.92(14/1)	0.21(14/1)	0.41(14/1)	-/-/14
2BTF	95	1	no	3.62	5.76	0.15	0.70	3.62(95/1)	3.62(95/1)	0.15(95/1)	0.70(95/1)	-/-/95
2HMI	217	1	no	4.72	7.84	0.18	0.70	4.72(217/1)	4.72(217/1)	0.18(217/1)	0.70(217/1)	-/-/217
2JEL	184	1	no	4.36	15.40	0.18	0.58	4.36(184/1)	5.76(41/3)	0.18(184/1)	0.58(184/1)	-/-/-
2MTA	21	1	no	4.95	15.55	0.11	0.86	4.95(21/1)	5.68(132/1)	0.11(21/1)	0.86(21/1)	-/-/-
2PCC	49	2	no	4.01	8.89	0.21	0.76	3.71(49/2)	3.71(49/2)	0.21(49/2)	0.76(49/2)	-/-/49
2QFW	152	1	no	1.27	2.27	0.62	0.14	1.27(152/1)	1.27(152/1)	0.62(152/1)	0.14(152/1)	-/152/152
2SIC	41	2	no	3.94	20.52	0.13	0.74	3.94(41/2)	12.26(1/1)	0.13(41/2)	0.74(41/2)	-/-/41
2SNI	180	1	no	2.80	6.59	0.24	0.41	2.80(180/1)	5.17(23/3)	0.24(180/1)	0.41(180/1)	-/-/180
2VIS	62	1	no	5.34	10.93	0.04	0.95	5.34(62/1)	5.34(62/1)	0.04(62/1)	0.95(62/1)	-/-/-
7CEI	12	9	yes	1.68	4.07	0.54	0.22	1.31(12/9)	1.31(12/9)	0.54(12/9)	0.07(12/9)	-/12/12

Appendix D

Ligand Density Scoring Schemes

A number of different scoring schemes were used rank the SwarmDock starting positions based on the ligand density surrounding them (see section 4.3.2).

1. For each trajectory point, the score of the nearest starting point is incremented by 1.

2. For each trajectory point, the score of the nearest 5 starting points are incremented by 1.

3. For each trajectory point, the score of the nearest starting point is incremented by 2. The score of the next 4 nearest starting points in incremented by 1.

4. For each trajectory point, the score of the nearest 7 starting points are incremented by 1.

5. For each trajectory point, the score of the nearest starting point is incremented by 2. The score of the next 6 nearest starting points in incremented by 1.

6. For each trajectory point, the score of the nearest starting point is incremented by 3. The score of the next 2 nearest starting points in incremented by 2. The score of the next 4 nearest starting points is incremented by 1.

7. For each trajectory point, the score of every starting point is incremented by the reciprocal of the distance between the trajectory point and the starting point.

8. For each trajectory point, the score of every starting point is incremented by the reciprocal of the square of the distance between the trajectory point and the starting point.

9. For each trajectory point, the score of every starting point is incremented by the reciprocal of the root of the distance between the trajectory point and the starting point.

The following tables show the Wilcoxon rank-sum one-tailed P-values for whether the scores of the 10 SwarmDock starting positions nearest the binding site are significantly higher or lower than the scores of the other starting positions, using the 9 different scoring schemes, for simulations with (c) and without (u) external crowding proteins. Values below the 0.05 significance level are highlighted in bold.

Table D.1: Binding region greater scoring than non-binding region

Complex	c1	u1	c2	u2	c3	u3	c4	u4	c5	u5	c6	u6	c7	u7	c8	u8	c9	u9
1GRN	0.081	**0.027**	**0.039**	**0.014**	**0.042**	**0.013**	**0.026**	**0.009**	**0.032**	**0.010**	**0.040**	**0.014**	**0.004**	**0.002**	**0.003**	**0.000**	**0.004**	**0.003**
1HE8	**0.015**	0.447	**0.016**	0.335	**0.015**	0.332	**0.018**	0.281	**0.017**	0.292	**0.014**	0.314	0.461	0.617	0.304	0.838	0.522	0.563
1QFW	0.155	0.238	0.169	0.256	0.173	0.247	0.185	0.280	0.178	0.271	0.178	0.262	0.060	0.064	0.066	0.102	0.062	0.064
1GCQ	**0.013**	0.388	**0.005**	0.621	**0.005**	0.535	**0.005**	0.648	**0.006**	0.621	**0.006**	0.605	0.098	0.771	**0.019**	0.659	0.384	0.768
1AY7	0.433	0.236	0.366	**0.050**	0.390	0.072	0.359	0.051	0.382	0.057	0.397	0.060	0.329	0.386	0.296	0.162	0.355	0.506
1KTZ	0.082	**0.001**	**0.050**	**0.000**	0.053	**0.000**	0.066	**0.001**	0.066	**0.001**	0.062	**0.001**	0.149	0.086	0.147	**0.021**	0.151	0.112
1VFB	0.564	**0.035**	0.491	**0.026**	0.491	**0.022**	0.451	**0.018**	0.476	**0.019**	0.494	**0.023**	0.185	0.070	0.171	**0.005**	0.195	0.119
7CEI	0.510	0.151	0.386	**0.036**	0.403	0.052	0.431	**0.047**	0.452	0.058	0.444	0.056	0.741	0.515	0.734	0.284	0.730	0.614
1QA9	**0.007**	0.287	**0.006**	0.361	**0.005**	0.342	**0.005**	0.373	**0.006**	0.369	**0.007**	0.361	**0.021**	0.128	**0.010**	0.160	**0.027**	0.113
1FSK	**0.020**	0.091	**0.047**	**0.029**	**0.041**	**0.027**	**0.035**	**0.020**	**0.034**	**0.019**	**0.031**	**0.019**	**0.020**	**0.018**	**0.029**	**0.042**	**0.016**	**0.015**
1FQJ	**0.002**	**0.007**	**0.003**	**0.002**	**0.003**	**0.002**	**0.002**	**0.001**	**0.002**	**0.002**	**0.002**	**0.002**	**0.001**	**0.000**	**0.000**	**0.000**	**0.001**	**0.001**
1I9R	0.998	0.989	0.995	0.998	0.995	0.998	0.995	0.999	0.995	0.998	0.996	0.998	0.998	0.999	0.998	0.998	0.998	0.999
1ML0	0.065	0.343	**0.041**	0.221	**0.045**	0.216	**0.050**	0.161	**0.050**	0.196	0.051	0.219	**0.028**	**0.038**	**0.019**	0.059	**0.033**	**0.036**
1BUH	0.078	0.102	**0.037**	**0.016**	**0.042**	**0.026**	**0.025**	**0.018**	**0.029**	**0.022**	**0.040**	**0.021**	**0.004**	**0.006**	**0.005**	**0.003**	**0.004**	**0.004**
1NCA	0.119	0.793	0.138	0.856	0.136	0.864	0.094	0.862	0.098	0.848	0.106	0.858	0.188	0.398	0.248	0.886	0.166	0.219
1JPS	**0.044**	0.100	**0.050**	0.094	**0.047**	0.092	**0.037**	0.083	**0.038**	0.084	**0.042**	0.081	**0.015**	**0.022**	**0.015**	**0.026**	**0.013**	**0.015**
1AVX	0.814	0.213	0.842	0.372	0.842	0.298	0.831	0.356	0.849	0.344	0.852	0.348	0.842	0.787	0.868	0.591	0.847	0.834
1AHW	0.146	**0.025**	0.150	**0.010**	0.141	**0.011**	0.134	**0.009**	0.141	**0.008**	0.141	**0.010**	0.069	**0.012**	0.079	**0.006**	0.061	**0.013**
1EWY	**0.012**	**0.031**	**0.005**	**0.007**	**0.005**	**0.008**	**0.003**	**0.004**	**0.003**	**0.005**	**0.004**	**0.006**	**0.008**	**0.014**	**0.003**	**0.007**	**0.015**	**0.018**
1EAW	0.964	0.994	0.990	0.997	0.988	0.997	0.989	0.998	0.987	0.997	0.988	0.997	0.979	0.990	0.976	0.991	0.982	0.988
1KAC	**0.002**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
1PPE	0.683	0.052	0.557	0.051	0.612	**0.045**	0.532	**0.035**	0.549	**0.039**	0.570	**0.044**	0.515	0.325	0.562	0.186	0.528	0.447
1NSN	0.184	0.130	0.232	0.190	0.226	0.166	0.251	0.204	0.232	0.190	0.232	0.173	0.576	0.498	0.605	0.663	0.576	0.510
1TMQ	0.829	0.956	0.859	0.965	0.855	0.962	0.857	0.972	0.853	0.973	0.855	0.967	0.868	0.909	0.819	0.956	0.886	0.900
1HIA	0.792	0.510	0.810	0.730	0.804	0.695	0.829	0.748	0.821	0.744	0.812	0.741	0.921	0.897	0.893	0.780	0.942	0.937
1E6E	**0.048**	**0.012**	**0.041**	**0.004**	**0.045**	**0.005**	**0.037**	**0.004**	**0.039**	**0.004**	**0.041**	**0.004**	**0.041**	**0.030**	**0.034**	**0.008**	0.062	0.054

Table D.2: Binding region lower scoring that non-binding region

Complex	c1	u1	c2	u2	c3	u3	c4	u4	c5	u5	c6	u6	c7	u7	c8	u8	c9	u9
1GRN	0.920	0.973	0.961	0.986	0.959	0.987	0.975	0.991	0.969	0.990	0.961	0.987	0.997	0.998	0.997	1.000	0.996	0.997
1HE8	0.985	0.556	0.984	0.668	0.985	0.671	0.983	0.722	0.983	0.710	0.986	0.690	0.543	0.387	0.699	0.165	0.481	0.440
1QFW	0.847	0.764	0.834	0.747	0.829	0.756	0.817	0.723	0.824	0.732	0.824	0.741	0.941	0.937	0.936	0.900	0.939	0.937
1GCQ	0.988	0.617	0.995	0.384	0.995	0.469	0.995	0.356	0.995	0.384	0.994	0.400	0.904	0.232	0.982	0.344	0.621	0.235
1AY7	0.571	0.767	0.638	0.951	0.614	0.930	0.645	0.950	0.622	0.944	0.607	0.941	0.675	0.618	0.707	0.840	0.649	0.498
1KTZ	0.919	0.999	0.951	1.000	0.948	1.000	0.936	0.999	0.936	0.999	0.939	0.999	0.853	0.915	0.855	0.979	0.851	0.890
1VFB	0.440	0.966	0.513	0.975	0.513	0.978	0.553	0.982	0.527	0.981	0.509	0.977	0.817	0.931	0.831	0.996	0.807	0.883
7CEI	0.494	0.852	0.618	0.965	0.602	0.949	0.573	0.954	0.552	0.943	0.560	0.945	0.263	0.489	0.270	0.720	0.273	0.390
1QA9	0.994	0.716	0.995	0.643	0.995	0.662	0.995	0.631	0.995	0.635	0.993	0.643	0.980	0.875	0.990	0.842	0.973	0.889
1FSK	0.981	0.911	0.954	0.972	0.960	0.974	0.966	0.981	0.967	0.982	0.970	0.982	0.980	0.983	0.972	0.959	0.985	0.986
1FQJ	0.998	0.993	0.997	0.998	0.998	0.998	0.998	0.999	0.998	0.999	0.998	0.998	1.000	1.000	1.000	1.000	0.999	0.999
1I9R	**0.002**	**0.011**	**0.006**	**0.002**	**0.005**	**0.002**	**0.005**	**0.001**	**0.005**	**0.002**	**0.004**	**0.003**	**0.003**	**0.001**	**0.003**	**0.002**	**0.003**	**0.002**
1ML0	0.936	0.660	0.960	0.781	0.956	0.787	0.951	0.841	0.951	0.807	0.950	0.784	0.973	0.963	0.981	0.943	0.968	0.965
1BUH	0.923	0.900	0.964	0.984	0.959	0.974	0.976	0.982	0.972	0.979	0.961	0.979	0.996	0.994	0.995	0.997	0.996	0.996
1NCA	0.883	0.209	0.864	0.146	0.865	0.138	0.907	0.140	0.903	0.154	0.896	0.144	0.814	0.605	0.755	0.116	0.836	0.783
1JPS	0.957	0.902	0.951	0.907	0.954	0.909	0.964	0.919	0.963	0.917	0.959	0.921	0.985	0.978	0.986	0.975	0.988	0.986
1AVX	0.189	0.790	0.161	0.632	0.161	0.706	0.172	0.648	0.153	0.660	0.151	0.656	0.161	0.216	0.134	0.413	0.156	0.169
1AHW	0.857	0.976	0.852	0.990	0.861	0.989	0.868	0.991	0.861	0.992	0.861	0.991	0.932	0.988	0.922	0.994	0.940	0.988
1EWY	0.988	0.970	0.995	0.994	0.995	0.992	0.997	0.996	0.997	0.995	0.996	0.994	0.992	0.986	0.997	0.993	0.985	0.982
1EAW	**0.037**	**0.007**	**0.011**	**0.003**	**0.013**	**0.003**	**0.012**	**0.002**	**0.013**	**0.003**	**0.013**	**0.003**	**0.022**	**0.011**	**0.025**	**0.009**	**0.019**	**0.013**
1KAC	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1PPE	0.321	0.949	0.447	0.950	0.393	0.956	0.472	0.966	0.455	0.962	0.434	0.957	0.489	0.679	0.443	0.817	0.476	0.557
1NSN	0.819	0.873	0.771	0.813	0.778	0.837	0.752	0.799	0.771	0.813	0.771	0.829	0.428	0.506	0.400	0.341	0.428	0.494
1TMQ	0.174	**0.045**	0.143	**0.036**	0.147	**0.039**	0.145	**0.029**	0.150	**0.027**	0.147	**0.033**	0.134	0.093	0.184	**0.045**	0.116	0.101
1HIA	0.211	0.494	0.193	0.273	0.199	0.309	0.174	0.256	0.182	0.259	0.190	0.263	0.081	0.105	0.109	0.223	0.060	0.065
1E6E	0.953	0.988	0.960	0.997	0.956	0.996	0.964	0.996	0.962	0.996	0.959	0.996	0.959	0.971	0.967	0.992	0.940	0.947

Bibliography

Abagyan, R. and Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, **235**:983–1002.

Abu-Erreish, G. M. and Peanasky, R. J. (1974). Pepsin inhibitors from Ascaris lumbricoides. Pepsin-inhibitor complex: stoichiometry of formation, dissociation, and stability of the complex. *J. Biol. Chem.*, **249**:1566–1571.

Albeck, S. and Schreiber, G. (1999). Biophysical characterization of the interaction of the beta-lactamase TEM-1 with its protein inhibitor BLIP. *Biochemistry*, **38**:11–21.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**:291–294.

Alder, B. J. and Wainwright, T. E. (1959). Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, **31**:459–466.

Alexander-Brett, J. M. and Fremont, D. H. (2007). Dual GPCR and GAG mimicry by the M3 chemokine decoy receptor. *J. Exp. Med.*, **204**:3157–3172.

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, first edition.

Andersen, P. S., Lavoie, P. M., Sekaly, R. P., Churchill, H., Kranz, D. M., Schlievert, P. M., Karjalainen, K., and Mariuzza, R. A. (1999). Role of the T cell receptor alpha chain in stabilizing TCR-superantigen-MHC class II complexes. *Immunity*, **10**:473–483.

Andre, I., Bradley, P., Wang, C., and Baker, D. (2007). Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. U.S.A.*, **104**:17656–17661.

Andrusier, N., Mashiach, E., Nussinov, R., and Wolfson, H. J. (2008). Principles of flexible protein-protein docking. *Proteins*, **73**:271–289.

Andrusier, N., Nussinov, R., and Wolfson, H. J. (2007). FireDock: fast interaction refinement in molecular docking. *Proteins*, **69**:139–159.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., *et al.* (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**:D525–531.

Arkin, M. R. and Wells, J. A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov*, **3**:301–317.

Armstrong, S., Yates, S. P., and Merrill, A. R. (2002). Insight into the catalytic mechanism of Pseudomonas aeruginosa exotoxin A. Studies of toxin interaction with eukaryotic elongation factor-2. *J. Biol. Chem.*, **277**:46669–46675.

Arnesano, F., Banci, L., Bertini, I., and Bonvin, A. M. (2004). A docking approach to the study of copper trafficking proteins; interaction between metallochaperones and soluble domains of copper ATPases. *Structure*, **12**:669–676.

Arnold, G. E. and Ornstein, R. L. (1997). Protein hinge bending as seen in molecular dynamics simulations of native and M61 mutant T4 lysozymes. *Biopolymers*, **41**:533–544.

Arolas, J. L., Lorenzo, J., Rovira, A., Castella, J., Aviles, F. X., and Sommerhoff, C. P. (2005a). A carboxypeptidase inhibitor from the tick Rhipicephalus bursa: isolation, cDNA cloning, recombinant expression, and characterization. *J. Biol. Chem.*, **280**:3441–3448.

Arolas, J. L., Popowicz, G. M., Bronsoms, S., Aviles, F. X., Huber, R., Holak, T. A., and Ventura, S. (2005b). Study of a major intermediate in the oxidative folding of leech carboxypeptidase inhibitor: contribution of the fourth disulfide bond. *J. Mol. Biol.*, **352**:961–975.

Arold, S., O'Brien, R., Franken, P., Strub, M. P., Hoh, F., Dumas, C., and Ladbury, J. E. (1998). RT loop flexibility enhances the specificity of Src family SH3 domains for HIV-1 Nef. *Biochemistry*, **37**:14683–14691.

Arunachalam, J., Kanagasabai, V., and Gautham, N. (2006). Protein structure prediction using mutually orthogonal Latin squares and a genetic algorithm. *Biochem. Biophys. Res. Commun.*, **342**:424–433.

Ascenzi, P., Amiconi, G., Menegatti, E., Guarneri, M., Bolognesi, M., and Schnebli, H. P. (1988). Binding of the recombinant proteinase inhibitor eglin c from leech Hirudo medicinalis to human leukocyte elastase, bovine alpha-chymotrypsin and subtilisin Carlsberg: thermodynamic study. *J. Enzym. Inhib.*, **2**:167–172.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**:505–515.

Atkins, P. and De Paula, J. (2006). *Atkins' Physical Chemistry*. Oxford University Press, 8th edition.

Atkins, P. and Friedman, R. (2004). *Molecular Quantum Mechanics*. Oxford University Press, 4th edition.

Audie, J. and Scarlata, S. (2007). A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophys. Chem.*, **129**:198–211.

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, **2**:173–181.

Bai, H., Yang, K., Yu, D., Zhang, C., Chen, F., and Lai, L. (2011). Predicting kinetic constants of protein-protein interactions based on structural properties. *Proteins*, **79**:720–734.

Barlow, D. J. and Thornton, J. M. (1983). Ion-pairs in proteins. *J. Mol. Biol.*, **168**:867–885.

Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008). Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*, **73**:765–783.

Bastard, K., Boelens, R., and Prevost, C. (2011). Accounting for Large Amplitude Protein Deformation during in Silico Macromolecular Docking. *Int. J. Mol. Sci.*, **12**:1316–1333.

Bastard, K., Prévost, C., and Zacharias, M. (2006). Accounting for loop flexibility during protein-protein docking. *Proteins*, **62**:956–969.

Bastard, K., Thureau, A., Lavery, R., and Prevost, C. (2003). Docking macromolecules with flexible segments. *J Comput Chem*, **24**:1910–1920.

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, **97**:10269–10280.

Ben-Zeev, E., Kowalsman, N., Ben-Shimon, A., Segal, D., Atarot, T., Noivirt, O., Shay, T., and Eisenstein, M. (2005). Docking to single-domain and multiple-domain proteins: old and new challenges. *Proteins*, **60**:195–201.

Bennett, S. E., Schimerlik, M. I., and Mosbaugh, D. W. (1993). Kinetics of the uracil-DNA glycosylase/inhibitor protein association. Ung interaction with Ugi, nucleic acids, and uracil compounds. *J. Biol. Chem.*, **268**:26879–26885.

Berchanski, A. and Eisenstein, M. (2003). Construction of molecular assemblies via docking: modeling of tetramers with D2 symmetry. *Proteins*, **53**:817–829.

Berchanski, A., Segal, D., and Eisenstein, M. (2005). Modeling oligomers with Cn or Dn symmetry: application to CAPRI target 10. *Proteins*, **60**:202–206.

Berchanski, A., Shapira, B., and Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. *Proteins*, **56**:130–142.

Berg, T. (2003). Modulation of protein-protein interactions with small organic molecules. *Angew. Chem. Int. Ed. Engl.*, **42**:2462–2481.

Berg, T. (2008). Small-molecule inhibitors of protein-protein interactions. *Curr Opin Drug Discov Devel*, **11**:666–674.

Bernauer, J., Aze, J., Janin, J., and Poupon, A. (2007). A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, **23**:555–562.

Betancourt, M. R. (2005). Efficient Monte Carlo trial moves for polypeptide simulations. *J Chem Phys*, **123**:174905.

Betts, M. J. and Sternberg, M. J. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng.*, **12**:271–283.

Betzi, S., Restouin, A., Opi, S., Arold, S. T., Parrot, I., Guerlesquin, F., Morelli, X., and Collette, Y. (2007). Protein protein interaction inhibition (2P2I) combining high throughput and virtual screening: Application to the HIV-1 Nef protein. *Proc. Natl. Acad. Sci. U.S.A.*, **104**:19256–19261.

Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F. P., Mariuzza, R. A., and Poljak, R. J. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. U.S.A.*, **91**:1089–1093.

Birkhoff, G. (1936). The logic of Quantum Mechanics. *Ann. Math.*, **37**:823–843.

Blundell, T. L. and Fernandez-Recio, J. (2006). Cell biology: brief encounters bolster contacts. *Nature*, **444**:279–280.

Bode, W. (1979). The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. II. The binding of the pancreatic trypsin inhibitor and of isoleucine-valine and of sequentially related peptides to trypsinogen and to p-guanidinobenzoate-trypsinogen. *J. Mol. Biol.*, **127**:357–374.

Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, **5**:789–796.

Bolewska, K., Krowarsch, D., Otlewski, J., Jaroszewski, L., and Bierzynski, A. (1995). Leu CMTI I–a representative of the squash inhibitors of serine proteinases. *FEBS Lett.*, **377**:172–174.

Bonsor, D. A., Grishkovskaya, I., Dodson, E. J., and Kleanthous, C. (2007). Molecular mimicry enables competitive recruitment by a natively disordered protein. *J. Am. Chem. Soc.*, **129**:4800–4807.

Bonvin, A. M. (2006). Flexible protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**:194–200.

Bordner, A. J. and Gorin, A. A. (2007). Protein docking using surface matching and supervised machine learning. *Proteins*, **68**:488–502.

Borrell, B. (2009). Fraud rocks protein community. *Nature*, **462**:970.

Borshell, N., Papp, T., and Congreve, M. (2011). Deal watch: Valuation benefits of structure-enabled drug discovery. *Nat Rev Drug Discov*, **10**:166.

Bougouffa, S. and Warwicker, J. (2008). Volume-based solvation models outperform area-based models in combined studies of wild-type and mutated protein-protein interfaces. *BMC Bioinformatics*, **9**:448.

Boulanger, M. J., Bankovich, A. J., Kortemme, T., Baker, D., and Garcia, K. C. (2003). Convergent mechanisms for recognition of divergent cytokines by the shared signaling receptor gp130. *Mol. Cell*, **12**:577–589.

Bourne, Y., Watson, M. H., Hickey, M. J., Holmes, W., Rocque, W., Reed, S. I., and Tainer, J. A. (1996). Crystal structure and mutational analysis of the human CDK2 kinase complex with cell cycle-regulatory protein CksHs1. *Cell*, **84**:863–874.

Bowman, A. L., Nikolovska-Coleska, Z., Zhong, H., Wang, S., and Carlson, H. A. (2007). Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *J. Am. Chem. Soc.*, **129**:12809–12814.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**:5–32.

Brooijmans, N., Sharp, K. A., and Kuntz, I. D. (2002). Stability of macromolecular complexes. *Proteins*, **48**:645–653.

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., *et al.* (2009). CHARMM: the biomolecular simulation program. *J Comput Chem*, **30**:1545–1614.

Bryan, J. (1988). Gelsolin has three actin-binding sites. *J. Cell Biol.*, **106**:1553–1562.

Camacho, C. J. and Gatchell, D. W. (2003). Successful discrimination of protein interactions. *Proteins*, **52**:92–97.

Camacho, C. J. and Vajda, S. (2001). Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:10636–10641.

Camacho, C. J., Weng, Z., Vajda, S., and Delisi, C. (1999). Free Energy Landscapes of Encounter Complexes in Protein-Protein Association. *Biophys. J.*, **76**:1166–1178.

Carbone, F. R. and Paterson, Y. (1985). Monoclonal antibodies to horse cytochrome c expressing four distinct idiotypes distribute among two sites on the native protein. *J. Immunol.*, **135**:2609–2616.

Carnap, R. (1931). Die logizistische Grundlegung der Mathematik. *Erkenntnis*, **2**:91–121.

Carr, R. A., Congreve, M., Murray, C. W., and Rees, D. C. (2005). Fragment-based lead discovery: leads by design. *Drug Discov. Today*, **10**:987–992.

Carrington, B. J. and Mancera, R. L. (2004). Comparative estimation of vibrational entropy changes in proteins through normal modes analysis. *J. Mol. Graph. Model.*, **23**:167–174.

Castro, M. J. and Anderson, S. (1996). Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to beta-trypsin and alpha-chymotrypsin. *Biochemistry*, **35**:11435–11446.

Cauerhff, A., Goldbaum, F. A., and Braden, B. C. (2004). Structural mechanism for affinity maturation of an anti-lysozyme antibody. *Proc. Natl. Acad. Sci. U.S.A.*, **101**:3539–3544.

Cavasotto, C. N., Kovacs, J. A., and Abagyan, R. A. (2005). Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.*, **127**:9632–9640.

Chae, M. H., Krull, F., Lorenzen, S., and Knapp, E. W. (2010). Predicting protein complex geometries with a neural network. *Proteins*, **78**:1026–1039.

Chaillan-Huntington, C., Braslavsky, C. V., Kuhlmann, J., and Stewart, M. (2000). Dissecting the interactions between NTF2, RanGDP, and the nucleoporin XFXFG repeats. *J. Biol. Chem.*, **275**:5874–5879.

Champ, P. C. and Camacho, C. J. (2007). FastContact: a free energy scoring tool for protein-protein complex structures. *Nucleic Acids Res.*, **35**:W556–560.

Chao, D. T. and Korsmeyer, S. J. (1998). BCL-2 family: regulators of cell death. *Annu. Rev. Immunol.*, **16**:395–419.

de Chateau, M., Holst, E., and Bjorck, L. (1996). Protein PAB, an albumin-binding bacterial surface protein promoting growth and virulence. *J. Biol. Chem.*, **271**:26609–26615.

Chaudhury, S. and Gray, J. J. (2008). Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.*, **381**:1068–1087.

Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, **26**:689–691.

Chen, H. M., Liu, B. F., Huang, H. L., Hwang, S. F., and Ho, S. Y. (2007). SODOCK: swarm optimization for highly flexible protein-ligand docking. *J Comput Chem*, **28**:612–623.

Chen, J., Im, W., and Brooks, C. L. (2006). Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *J. Am. Chem. Soc.*, **128**:3728–3736.

Chen, R., Mintseris, J., Janin, J., and Weng, Z. (2003). A protein-protein docking benchmark. *Proteins*, **52**:88–91.

Chen, R. and Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, **47**:281–294.

Cheng, T. M., Blundell, T. L., and Fernandez-Recio, J. (2007a). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, **68**:503–515.

Cheng, T. M., Blundell, T. L., and Fernandez-Recio, J. (2008). Structural assembly of two-domain proteins by rigid-body docking. *BMC Bioinformatics*, **9**:441.

Cheng, Y., Suen, J. K., Zhang, D., Bond, S. D., Zhang, Y., Song, Y., Baker, N. A., Bajaj, C. L., Holst, M. J., and McCammon, J. A. (2007b). Finite element analysis of the time-dependent Smoluchowski equation for acetylcholinesterase reaction rate calculations. *Biophys. J.*, **92**:3397–3406.

Chennubhotla, C., Rader, A. J., Yang, L. W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Phys Biol*, **2**:S173–180.

Cho, S., Swaminathan, C. P., Bonsor, D. A., Kerzic, M. C., Guan, R., Yang, J., Kieke, M. C., Andersen, P. S., Kranz, D. M., Mariuzza, R. A., *et al.* (2010). Assessing energetic contributions to binding from a disordered region in a protein-protein interaction . *Biochemistry*, **49**:9256–9268.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**:338–339.

Chothia, C., Novotny, J., Bruccoleri, R., and Karplus, M. (1985). Domain association in immunoglobulin molecules. The packing of variable domains. *J. Mol. Biol.*, **186**:651–663.

Chrencik, J. E., Brooun, A., Recht, M. I., Kraus, M. L., Koolpe, M., Kolatkar, A. R., Bruce, R. H., Martiny-Baron, G., Widmer, H., Pasquale, E. B., *et al.* (2006). Structure and thermodynamic characterization of the EphB4/Ephrin-B2 antagonist peptide complex reveals the determinants for receptor specificity. *Structure*, **14**:321–330.

Ciaccia, A. V., Monroe, D. M., and Church, F. C. (1997). Arginine 200 of heparin cofactor II promotes intramolecular interactions of the acidic domain. Implication for thrombin inhibition. *J. Biol. Chem.*, **272**:14074–14079.

Clerc, M. (1999). The swarm and the queen: towards a deterministic and adaptive particle swarm optimization. In *Congress on Evolutionary Computation*, volume 3, pages 1951–1957.

Clerc, M. and Kennedy, J. (2002). The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, **6**:58–73.

Cohen, G. N. and Monod, J. (1957). Bacterial permeases. *Bacteriol. Rev.*, **21**:169–194.

Cohn, M. and Monod, J. (1951). Purification and properties of the beta-galactosidase (lactase) of Escherichia coli. *Biochim. Biophys. Acta*, **7**:153–174.

Comeau, S. R. and Camacho, C. J. (2005). Predicting oligomeric assemblies: N-mers a primer. *J. Struct. Biol.*, **150**:233–244.

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**:45–50.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**:709–713.

Connolly, M. L. (1986). Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers*, **25**:1229–1247.

Cramer, C. J. (2004). *Essentials of Computational Chemistry: Theories and Models*. Wiley.

Cui, Q., Li, G., Ma, J., and Karplus, M. (2004). A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase. *J. Mol. Biol.*, **340**:345–372.

Dang, L. C., Melandri, F. D., and Stein, R. L. (1998). Kinetic and mechanistic studies on the hydrolysis of ubiquitin C-terminal 7-amido-4-methylcoumarin by deubiquitinating enzymes. *Biochemistry*, **37**:1868–1879.

Darling, R. J., Kuchibhotla, U., Glaesner, W., Micanovic, R., Witcher, D. R., and Beals, J. M. (2002). Glycosylation of erythropoietin affects receptor binding kinetics: role of electrostatic interactions. *Biochemistry*, **41**:14524–14531.

Davidson, V. L., Graichen, M. E., and Jones, L. H. (1993). Binding constants for a physiologic electron-transfer protein complex between methylamine dehydrogenase and amicyanin. Effects of ionic strength and bound copper on binding. *Biochim. Biophys. Acta*, **1144**:39–45.

Davis, I. W., Arendall, W. B., Richardson, D. C., and Richardson, J. S. (2006). The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*, **14**:265–274.

Davis, I. W. and Baker, D. (2009). RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.*, **385**:381–392.

De Crescenzo, G., Hinck, C. S., Shu, Z., Zuniga, J., Yang, J., Tang, Y., Baardsnes, J., Mendoza, V., Sun, L., Lopez-Casillas, F., *et al.* (2006). Three key residues underlie the differential affinity of the TGFbeta isoforms for the TGFbeta type II receptor. *J. Mol. Biol.*, **355**:47–62.

Dechantsreiter, M. A., Planker, E., Matha, B., Lohof, E., Holzemann, G., Jonczyk, A., Goodman, S. L., and Kessler, H. (1999). N-Methylated cyclic RGD peptides as highly active and selective alpha(V)beta(3) integrin antagonists. *J. Med. Chem.*, **42**:3033–3040.

Demerdash, O. N., Buyan, A., and Mitchell, J. C. (2010). ReplicOpter: a replicate optimizer for flexible docking. *Proteins*, **78**:3156–3165.

Diggle, C., Bizouarn, T., Cotton, N. P., and Jackson, J. B. (1996). Properties of the purified, recombinant, NADP(H)-binding domain III of the proton-translocating nicotinamide nucleotide transhydrogenase from Rhodospirillum rubrum. *Eur. J. Biochem.*, **241**:162–170.

Dinner, A. R. (2000). Local Deformations of Polymers with Nonplanar Rigid Main-Chain Internal Coordinates. *J. Comp. Chem*, **21**:1132–1144.

Dobbins, S. E., Lesk, V. I., and Sternberg, M. J. (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:10390–10395.

Dodd, L. R. (1993). A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.*, **78**:961–996.

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., and Baker, N. A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**:W522–525.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**:1731–1737.

Domling, A. (2008). Small molecular weight protein-protein interaction antagonists: an insurmountable challenge? *Curr Opin Chem Biol*, **12**:281–291.

Dongarra, J. (2002). Basic Linear Algebra Subprograms Technical Forum Standard. *International Journal of High Performance Applications and Supercomputing*, **16**:115–199.

Dooley, H., Stanfield, R. L., Brady, R. A., and Flajnik, M. F. (2006). First molecular and biochemical analysis of in vivo affinity maturation in an ectothermic vertebrate. *Proc. Natl. Acad. Sci. U.S.A.*, **103**:1846–1851.

Dubin, G., Wladyka, B., Stec-Niemczyk, J., Chmiel, D., Zdzalik, M., Dubin, A., and Potempa, J. (2007). The staphostatin family of cysteine protease inhibitors in the genus Staphylococcus as an example of parallel evolution of protease and inhibitor specificity. *Biol. Chem.*, **388**:227–235.

Duckett, C. S., Li, F., Wang, Y., Tomaselli, K. J., Thompson, C. B., and Armstrong, R. C. (1998). Human IAP-like protein regulates programmed cell death downstream of Bcl-xL and cytochrome c. *Mol. Cell. Biol.*, **18**:608–615.

Durand, P. (1983). Direct determination of effective Hamiltonians by wave-operator methods. I. General formalism. *Phys. Rev. A*, **28**:3184–3192.

Durand, P., Trinquier, G., and Sanejouand, Y. H. (1994). A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34**:759–771.

Eathiraj, S., Pan, X., Ritacco, C., and Lambright, D. G. (2005). Structural basis of family-wide Rab GTPase recognition by rabenosyn-5. *Nature*, **436**:415–419.

Eberhart, R. and Kennedy, J. (1995). A new optimizer using particle swarm theory. pages 39–43.

Eberhart, R. C. and Shi, Y. (2000). Comparing inertia weights and constriction factors in particle swarm optimization. volume 1, pages 84–88 vol.1.

Eberhart, R. C., Shi, Y., and Kennedy, J. (2001). *Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation)*. Morgan Kaufmann, 1st edition.

Eccleston, J. F., Moore, K. J., Morgan, L., Skinner, R. H., and Lowe, P. N. (1993). Kinetics of interaction between normal and proline 12 Ras and the GTPase-activating proteins, p120-GAP and neurofibromin. The significance of the intrinsic GTPase rate in determining the transforming ability of ras. *J. Biol. Chem.*, **268**:27012–27019.

Eisenstein, M. and Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *C. R. Biol.*, **327**:409–420.

Ellis, J. R. (2001). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current Opinion in Structural Biology*, **11**:114–119.

Emekli, U., Schneidman-Duhovny, D., Wolfson, H. J., Nussinov, R., and Haliloglu, T. (2008). HingeProt: automated prediction of hinges in protein structures. *Proteins*, **70**:1219–1227.

Empie, M. W. and Laskowski, M. (1982). Thermodynamics and kinetics of single residue replacements in avian ovomucoid third domains: effect on inhibitor interactions with serine proteinases. *Biochemistry*, **21**:2274–2284.

Engelbrecht, A. P. (2006). *Fundamentals of Computational Swarm Intelligence*. Wiley.

English, A. C., Groom, C. R., and Hubbard, R. E. (2001). Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng.*, **14**:47–59.

Erman, J. E., Kresheck, G. C., Vitello, L. B., and Miller, M. A. (1997). Cytochrome c/cytochrome c peroxidase complex: effect of binding-site mutations on the thermodynamics of complex formation. *Biochemistry*, **36**:4054–4060.

Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M. L. (2009). Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinformatics*, **10**:233–246.

Faller, B. and Bieth, J. G. (1991). Kinetics of the interaction of chymotrypsin with eglin c. *Biochem. J.*, **280 (Pt 1)**:27–32.

Farady, C. J., Sun, J., Darragh, M. R., Miller, S. M., and Craik, C. S. (2007). The mechanism of inhibition of antibody-based inhibitors of membrane-type serine protease 1 (MT-SP1). *J. Mol. Biol.*, **369**:1041–1051.

Faure, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, **22**:e124–131.

Feliu, E. and Oliva, B. (2010). How different from random are docking predictions when ranked by scoring functions? *Proteins*, **78**:3376–3385.

Feltzer, R. E., Gray, R. D., Dean, W. L., and Pierce, W. M. (2000). Alkaline proteinase inhibitor of Pseudomonas aeruginosa. Interaction of native and N-terminally truncated inhibitor proteins with Pseudomonas metalloproteinases. *J. Biol. Chem.*, **275**:21002–21009.

Feng, Y., Kloczkowski, A., and Jernigan, R. L. (2010). Potentials 'R' Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics*, **11**:92.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, **335**:843–865.

Fernández-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, **52**:113–117.

Ferrara, P., Apostolakis, J., and Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins*, **46**:24–33.

Fierens, K., Gils, A., Sansen, S., Brijs, K., Courtin, C. M., Declerck, P. J., De Ranter, C. J., Gebruers, K., Rabijns, A., Robben, J., *et al.* (2005). His374 of wheat endoxylanase inhibitor TAXI-I stabilizes complex formation with glycoside hydrolase family 11 endoxylanases. *FEBS J.*, **272**:5872–5882.

Fischer, D., Lin, S. L., Wolfson, H. L., and Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.*, **248**:459–477.

Fleishman, S. J., Corn, J. E., Strauch, E. M., Whitehead, T. A., Andre, I., Thompson, J., Havranek, J. J., Das, R., Bradley, P., and Baker, D. (2010). Rosetta in CAPRI rounds 13-19. *Proteins*, **78**:3212–3218.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**:816–821.

Fleury, D., Wharton, S. A., Skehel, J. J., Knossow, M., and Bizebard, T. (1998). Antigen distortion allows influenza virus to escape neutralization. *Nat. Struct. Biol.*, **5**:119–123.

Floquet, N., Marechal, J. D., Badet-Denisot, M. A., Robert, C. H., Dauchez, M., and Perahia, D. (2006). Normal mode analysis as a prerequisite for drug design: application to matrix metalloproteinases inhibitors. *FEBS Lett.*, **580**:5130–5136.

Fogolari, F., Brigo, A., and Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.*, **15**:377–392.

Foote, J. and Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.*, **224**:487–499.

Forster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A., and Sali, A. (2008). Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. *J. Mol. Biol.*, **382**:1089–1106.

Friedland, G. D., Lakomek, N. A., Griesinger, C., Meiler, J., and Kortemme, T. (2009). A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput. Biol.*, **5**:e1000393.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics*, **19**:1–67.

Friesen, W. O. and Block, G. D. (1984). What is a biological oscillator? *Am. J. Physiol.*, **246**:R847–853.

Fu, H. (2004). *Protein-Protein Interactions: Methods and Applications*. Humana Press.

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**:106–120.

Garcia, R. A., Pantazatos, D., and Villarreal, F. J. (2004). Hydrogen/deuterium exchange mass spectrometry for investigating protein-ligand interactions. *Assay Drug Dev Technol*, **2**:81–91.

Gardiner, E. J., Willett, P., and Artymiuk, P. J. (2001). Protein docking using a genetic algorithm. *Proteins*, **44**:44–56.

Gardsvoll, H., Werner, F., S?ndergaard, L., Dan?, K., and Ploug, M. (2004). Characterization of low-glycosylated forms of soluble human urokinase receptor expressed in Drosophila Schneider 2 cells after deletion of glycosylation-sites. *Protein Expr. Purif.*, **34**:284–295.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**:141–147.

Ghosh, M., Meiss, G., Pingoud, A. M., London, R. E., and Pedersen, L. C. (2007). The nuclease a-inhibitor complex is characterized by a novel metal ion bridge. *J. Biol. Chem.*, **282**:5682–5690.

Gilson, M. K. and Zhou, H. X. (2007). Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct*, **36**:21–42.

Go, N. and Scheraga, H. A. (1970). Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules*, **3**:178–187.

Goh, C. S., Milburn, D., and Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.*, **14**:104–109.

Goldbaum, F. A., Cauerhff, A., Velikovsky, C. A., Llera, A. S., Riottot, M. M., and Poljak, R. J. (1999). Lack of significant differences in association rates and affinities of antibodies from short-term and long-term responses to hen egg lysozyme. *J. Immunol.*, **162**:6040–6045.

Golemis, E. (2002). *Protein-protein interactions: A molecular cloning manual*. Cold Spring Harbor Laboratory Press.

Gray, J. J. (2006). High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**:183–193.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**:281–299.

de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins*, **29**:240–251.

Grosdidier, S. and Fernandez-Recio, J. (2008). Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, **9**:447.

Grunberg, R., Leckner, J., and Nilges, M. (2004). Complementarity of structure ensembles in protein-protein binding. *Structure*, **12**:2125–2136.

Gsponer, J., Christodoulou, J., Cavalli, A., Bui, J. M., Richter, B., Dobson, C. M., and Vendruscolo, M. (2008). A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction. *Structure*, **16**:736–746.

Guharoy, M., Janin, J., and Robert, C. H. (2010). Side-chain rotamer transitions at protein-protein interfaces. *Proteins*, **78**:3219–3225.

Gumireddy, K., Reddy, M. V., Cosenza, S. C., Boominathan, R., Boomi Nathan, R., Baker, S. J., Papathi, N., Jiang, J., Holland, J., and Reddy, E. P. (2005). ON01910, a non-ATP-competitive small molecule inhibitor of Plk1, is a potent anticancer agent. *Cancer Cell*, **7**:275–286.

Guthridge, J. M., Rakstang, J. K., Young, K. A., Hinshelwood, J., Aslam, M., Robertson, A., Gipson, M. G., Sarrias, M. R., Moore, W. T., Meagher, M., *et al.* (2001). Structural studies in solution of the recombinant N-terminal pair of short consensus/complement repeat domains of complement receptor type 2 (CR2/CD21) and interactions with its ligand C3dg. *Biochemistry*, **40**:5931–5941.

Hajduk, P. J., Huth, J. R., and Tse, C. (2005). Predicting protein druggability. *Drug Discov. Today*, **10**:1675–1682.

Hall, D. (2003). Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, **1649**:127–139.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**:57–70.

Hardy, R. L. (1971). Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research*, **76**:1905–1915.

Hartley, R. W. (1993). Directed mutagenesis and barnase-barstar recognition. *Biochemistry*, **32**:5978–5984.

Hartley, W., Both, V., Hebert, E. J., Homerover, D., Jucovic, M., Nazarov, V., Rybajlak, I., and Sevcik, J. (1996). Barstar Inhibits Extracellular Ribonucleases of Streptomyces and Allows their Production from Recombinant Genes. *Protein. Pept. Lett.*, **3**:225–231.

Haspel, N., Ricklin, D., Geisbrecht, B. V., Kavraki, L. E., and Lambris, J. D. (2008). Electrostatic contributions drive the interaction between Staphylococcus aureus protein Efb-C and its complement target C3d. *Protein Sci.*, **17**:1894–1906.

Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M. (2002). Electrostatics in protein-protein docking. *Protein Sci.*, **11**:571–587.

Helmer-Citterich, M. and Tramontano, A. (1994). PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.*, **235**:1021–1031.

Hershberger, S. J., Lee, S. G., and Chmielewski, J. (2007). Scaffolds for blocking protein-protein interactions. *Curr Top Med Chem*, **7**:928–942.

Hinchliffe, A. (2003). *Molecular modelling for beginners*. Wiley.

Hirayama, K., Aoki, S., Nishikawa, K., Matsumoto, T., and Wada, K. (2007). Identification of novel chemical inhibitors for ubiquitin C-terminal hydrolase-L3 by virtual screening. *Bioorg. Med. Chem.*, **15**:6810–6818.

Hoffman, G. R., Nassar, N., Oswald, R. E., and Cerione, R. A. (1998). Fluoride activation of the Rho family GTP-binding protein Cdc42Hs. *J. Biol. Chem.*, **273**:4392–4399.

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nat Rev Drug Discov*, **1**:727–730.

Horovitz, A. (1996). Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold Des*, **1**:R121–126.

Horton, N. and Lewis, M. (1992). Calculation of the free energy of association for protein complexes. *Protein Sci.*, **1**:169–181.

Hou, Z., Danzer, J. R., Mendoza, L., Bose, M. E., Muller, U., Williams, B., and Fox, C. A. (2009). Phylogenetic conservation and homology modeling help reveal a novel domain within the budding yeast heterochromatin protein Sir1. *Mol. Cell. Biol.*, **29**:687–702.

Huang, M., Syed, R., Stura, E. A., Stone, M. J., Stefanko, R. S., Ruf, W., Edgington, T. S., and Wilson, I. A. (1998). The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex. *J. Mol. Biol.*, **275**:873–894.

Huang, P. S., Love, J. J., and Mayo, S. L. (2005). Adaptation of a fast Fourier transform-based docking algorithm for protein design. *J Comput Chem*, **26**:1222–1232.

Huang, P. S., Love, J. J., and Mayo, S. L. (2007). A de novo designed protein protein interface. *Protein Sci.*, **16**:2770–2774.

Huang, S. Y. and Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, **72**:557–579.

Hubbard, S. and Thornton, J. (1993). 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.

Huizinga, E. G., Tsuji, S., Romijn, R. A., Schiphorst, M. E., de Groot, P. G., Sixma, J. J., and Gros, P. (2002). Structures of glycoprotein Ibalpha and its complex with von Willebrand factor A1 domain. *Science*, **297**:1176–1179.

Hunjan, J., Tovchigrechko, A., Gao, Y., and Vakser, I. A. (2008). The size of the intermolecular energy funnel in protein-protein interactions. *Proteins*, **72**:344–352.

Huse, M., Muir, T. W., Xu, L., Chen, Y. G., Kuriyan, J., and Massague, J. (2001). The TGF beta receptor activation process: an inhibitor- to substrate-binding switch. *Mol. Cell*, **8**:671–682.

Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins*, **73**:705–709.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-protein docking benchmark version 4.0. *Proteins*, **78**:3111–3114.

Hwang, J. K. and Liao, W. F. (1995). Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.*, **8**:363–370.

Ignatev, A., Kravchenko, S., Rak, A., Goody, R. S., and Pylypenko, O. (2008). A structural model of the GDP dissociation inhibitor rab membrane extraction mechanism. *J. Biol. Chem.*, **283**:18377–18384.

Imming, P., Sinning, C., and Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov*, **5**:821–834.

Isenman, D. E., Leung, E., Mackay, J. D., Bagby, S., and van den Elsen, J. M. (2010). Mutational analyses reveal that the staphylococcal immune evasion molecule Sbi and complement receptor 2 (CR2) share overlapping contact residues on C3d: implications for the controversy regarding the CR2/C3d cocrystal structure. *J. Immunol.*, **184**:1946–1955.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics*, **9**:102.

Iso14882 (1998). ISO/IEC 14882:1998: Programming languages - C++. Technical report, International Organization for Standardization.

Jackson, R. M., Gabb, H. A., and Sternberg, M. J. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **276**:265–285.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**:318–356.

Jacquemin, M. G., Desqueper, B. G., Benhida, A., Vander Elst, L., Hoylaerts, M. F., Bakkus, M., Thielemans, K., Arnout, J., Peerlinck, K., Gilles, J. G., *et al.* (1998). Mechanism and kinetics of factor VIII inactivation: study with an IgG4 monoclonal antibody derived from a hemophilia A patient with inhibitor. *Blood*, **92**:496–506.

Janin, J. (1995). Principles of protein-protein recognition from structure to thermodynamics. *Biochimie*, **77**:497–505.

Janson, S., Merkle, S., and Middendorf, M. (2008). Molecular docking with multi-objective Particle Swarm Optimization. *Applied Soft Computing*, **8**:666–675.

Jiang, F. and Kim, S. H. (1991). "Soft docking": matching of molecular surface cubes. *J. Mol. Biol.*, **219**:79–102.

Jiang, L., Gao, Y., Mao, F., Liu, Z., and Lai, L. (2002). Potential of mean force for protein-protein interaction studies. *Proteins*, **46**:190–196.

Jiang, L., Kuhlman, B., Kortemme, T., and Baker, D. (2005). A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins*, **58**:893–904.

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, **118**:11225–11236.

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

Karaca, E. and Bonvin, A. M. (2011). A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure*, **19**:555–565.

Karaca, E., Melquiond, A. S., de Vries, S. J., Kastritis, P. L., and Bonvin, A. M. (2010). Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server. *Mol. Cell Proteomics*, **9**:1784–1794.

Karanasios, E., Simader, H., Panayotou, G., Suck, D., and Simos, G. (2007). Molecular determinants of the yeast Arc1p-aminoacyl-tRNA synthetase complex assembly. *J. Mol. Biol.*, **374**:1077–1090.

Kastritis, P. L. and Bonvin, A. M. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.*, **9**:2216–2225.

Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. M. J., and Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Prot. Sci.*

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.*, **89**:2195–2199.

Kellenberger, C., Boudier, C., Bermudez, I., Bieth, J. G., Luu, B., and Hietter, H. (1995). Serine protease inhibition by insect peptides containing a cysteine knot and a triple-stranded beta-sheet. *J. Biol. Chem.*, **270**:25514–25519.

Kennedy, J. and Eberhart, R. C. (1995). Particle Swarm Optimization. *IEEE Int'l. Conf. on Neural Networks*, **4**:1942–1948.

Kennedy, J. and Mendes, R. (2002). Population structure and particle swarm performance. *Computational Intelligence, Proceedings of the World on Congress on*, **2**:1671–1676.

Kiel, C., Selzer, T., Shaul, Y., Schreiber, G., and Herrmann, C. (2004). Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proc. Natl. Acad. Sci. U.S.A.*, **101**:9223–9228.

Kiemer, L. and Cesareni, G. (2007). Comparative interactomics: comparing apples and pears? *Trends Biotechnol.*, **25**:448–454.

Kingsford, C. L., Chazelle, B., and Singh, M. (2005). Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, **21**:1028–1036.

Kirby, I., Davison, E., Beavil, A. J., Soh, C. P., Wickham, T. J., Roelvink, P. W., Kovesdi, I., Sutton, B. J., and Santis, G. (2000). Identification of contact residues and definition of the CAR-binding site of adenovirus type 5 fiber protein. *J. Virol.*, **74**:2804–2813.

Klebe, C., Prinz, H., Wittinghofer, A., and Goody, R. S. (1995). The kinetic mechanism of Ran–nucleotide exchange catalyzed by RCC1. *Biochemistry*, **34**:12543–12552.

Koehl, P. and Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, **239**:249–275.

Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, **10**:2703–2734.

Komatsu, K., Kurihara, Y., Iwadate, M., Takeda-Shitaka, M., and Umeyama, H. (2003). Evaluation of the third solvent clusters fitting procedure for the prediction of protein-protein interactions based on the results at the CAPRI blind docking study. *Proteins*, **52**:15–18.

Kortemme, T., Morozov, A. V., and Baker, D. (2003). 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.

Kovacs, J. A., Cavasotto, C. N., and Abagyan, R. (2005). Conformational Sampling of Protein Flexibility in Generalized Coordinates: Application to Ligand Docking. *J. Comput. Theor. Nanosci.*, **2**:354–361.

Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **65**:392–406.

Kozakov, D., Clodfelter, K. H., Vajda, S., and Camacho, C. J. (2005). Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.*, **89**:867–875.

Kozlov, G., Nguyen, L., Lin, T., De Crescenzo, G., Park, M., and Gehring, K. (2007). Structural basis of ubiquitin recognition by the ubiquitin-associated (UBA) domain of the ubiquitin ligase EDD. *J. Biol. Chem.*, **282**:35787–35795.

Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H., and Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**:682–695.

Kresheck, G. C., Vitello, L. B., and Erman, J. E. (1995). Calorimetric studies on the interaction of horse ferricytochrome c and yeast cytochrome c peroxidase. *Biochemistry*, **34**:8398–8405.

Krippahl, L., Moura, J. J., and Palma, P. N. (2003). Modeling protein complexes with BiGGER. *Proteins*, **52**:19–23.

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**:778–795.

Krol, M. (2003). Comparison of various implicit solvent models in molecular dynamics simulations of immunoglobulin G light chain dimer. *J Comput Chem*, **24**:531–546.

Krol, M., Chaleil, R. A., Tournier, A. L., and Bates, P. A. (2007a). Implicit flexibility in protein docking: cross-docking and local refinement. *Proteins*, **69**:750–757.

Krol, M., Tournier, A. L., and Bates, P. A. (2007b). Flexible relaxation of rigid-body docking solutions. *Proteins*, **68**:159–169.

Krystek, S., Stouch, T., and Novotny, J. (1993). Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J. Mol. Biol.*, **234**:661–679.

Kumar, M. D. and Gromiha, M. M. (2006). PINT: Protein-protein Interactions Thermodynamic Database. *Nucleic Acids Res.*, **34**:D195–198.

Kumar, P., Han, B. C., Shi, Z., Jia, J., Wang, Y. P., Zhang, Y. T., Liang, L., Liu, Q. F., Ji, Z. L., and Chen, Y. Z. (2009). Update of KDBI: Kinetic Data of Bio-molecular Interaction database. *Nucleic Acids Res.*, **37**:D636–641.

Kurkcuoglu, O., Jernigan, R. L., and Doruker, P. (2006). Loop motions of triosephosphate isomerase observed with elastic networks. *Biochemistry*, **45**:1173–1182.

Lacy, D. B., Lin, H. C., Melnyk, R. A., Schueler-Furman, O., Reither, L., Cunningham, K., Baker, D., and Collier, R. J. (2005). A model of anthrax

toxin lethal factor bound to protective antigen. *Proc. Natl. Acad. Sci. U.S.A.*, **102**:16409–16414.

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**:1471–1475.

Lapouge, K., Smith, S. J., Walker, P. A., Gamblin, S. J., Smerdon, S. J., and Rittinger, K. (2000). Structure of the TPR domain of p67phox in complex with Rac.GTP. *Mol. Cell*, **6**:899–907.

Lasker, K., Phillips, J. L., Russel, D., Velazquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A., and Sali, A. (2010a). Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol. Cell Proteomics*, **9**:1689–1702.

Lasker, K., Sali, A., and Wolfson, H. J. (2010b). Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins*, **78**:3205–3211.

Lasker, K., Topf, M., Sali, A., and Wolfson, H. J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J. Mol. Biol.*, **388**:180–194.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, **13**:323–330.

Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., and Kortemme, T. (2010). RosettaBackrub–a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res.*, **38**:W569–575.

Lauwereys, M., Arbabi Ghahroudi, M., Desmyter, A., Kinne, J., Holzer, W., De Genst, E., Wyns, L., and Muyldermans, S. (1998). Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J.*, **17**:3512–3520.

Lavergne, J. P., Jault, J. M., and Galinier, A. (2002). Insights into the functioning of Bacillus subtilis HPr kinase/phosphatase: affinity for its protein substrates and role of cations and phosphate. *Biochemistry*, **41**:6218–6225.

Lavigne, P., Bagu, J. R., Boyko, R., Willard, L., Holmes, C. F., and Sykes, B. D. (2000). Structure-based thermodynamic analysis of the dissociation of protein phosphatase-1 catalytic subunit and microcystin-LR docked complexes. *Protein Sci.*, **9**:252–264.

Lazaridis, T. and Karplus, M. (1999a). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, **288**:477–487.

Lazaridis, T. and Karplus, M. (1999b). Effective energy function for proteins in solution. *Proteins*, **35**:133–152.

Leach, A. (2001). *Molecular Modelling: Principles and Applications*. Prentice Hall, 2 edition.

Lebowitz, J. and Laskowski, M. (1962). Potentiometric Measurement of Protein-Protein Association Constants. Soybean Trypsin Inhibitor-Trypsin Association. *Biochemistry*, **1**:1044–1055.

Lee, C. H., Leung, B., Lemmon, M. A., Zheng, J., Cowburn, D., Kuriyan, J., and Saksela, K. (1995). A single amino acid in the SH3 domain of Hck determines its high affinity and specificity in binding to HIV-1 Nef protein. *EMBO J.*, **14**:5006–5015.

Lee, S. and Karplus, M. (1987). Kinetics of diffusion-influenced bimolecular reactions in solution. I. General formalism and relaxation kinetics of fast reversible reactions. *J. Chem. Phys*, **86**:1883–1903.

Lenarcic, B., Krizaj, I., Zunec, P., and Turk, V. (1996). Differences in specificity for the interactions of stefins A, B and D with cysteine proteinases. *FEBS Lett.*, **395**:113–118.

Lensink, M. F., Mendez, R., and Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*, **69**:704–718.

Lensink, M. F. and Wodak, S. J. (2010a). Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins*, **78**:3085–3095.

Lensink, M. F. and Wodak, S. J. (2010b). Docking and scoring protein interactions: CAPRI 2009. *Proteins*, **78**:3073–3084.

Lesk, V. I. and Sternberg, M. J. (2008). 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics*, **24**:1137–1144.

Levinthal, C., Wodak, S. J., Kahn, P., and Dadivanian, A. K. (1975). Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. *Proc. Natl. Acad. Sci. U.S.A.*, **72**:1330–1334.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, **104**:59–107.

Li, G. and Cui, Q. (2002). A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophys. J.*, **83**:2457–2474.

Li, L., Chen, R., and Weng, Z. (2003). RDOCK: refinement of rigid-body protein docking predictions. *Proteins*, **53**:693–707.

Li, L., Thomas, R. M., Suzuki, H., De Brabander, J. K., Wang, X., and Harran, P. G. (2004). A small molecule Smac mimic potentiates TRAIL- and TNFalpha-mediated cell death. *Science*, **305**:1471–1474.

Li, L., Zhao, B., Cui, Z., Gan, J., Sakharkar, M. K., and Kangueane, P. (2006). Identification of hot spot residues at protein-protein interface. *Bioinformation*, **1**:121–126.

Li, W., Hamill, S. J., Hemmings, A. M., Moore, G. R., James, R., and Kleanthous, C. (1998). Dual recognition and the role of specificity-determining residues in colicin E9 DNase-immunity protein interactions. *Biochemistry*, **37**:11771–11779.

Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., Wong, S. K., Huang, I. C., Xu, K., Vasilieva, N., *et al.* (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.*, **24**:1634–1643.

Li, X. (2011). *Investigation of protein-protein interactions: multibody docking, association/dissociation kinetics and macromolecular crowding*. PhD, University College London.

Li, X. and Liang, J. (2011). Geometric packing potential function for model selection in protein structure and protein-protein binding predictions. *Submitted*.

Li, X., Moal, I., and Bates, P. (2010a). Bridging the gaps: atomic simulation of macromolecular environment brings together protein docking, interaction kinetics and the crowding effects. *BMC Bioinformatics*, **11**:O2.

Li, X., Moal, I. H., and Bates, P. A. (2010b). Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. *Proteins*, **78**:3189–3196.

Li, Y., Li, H., Smith-Gill, S. J., and Mariuzza, R. A. (2000). Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry*, **39**:6296–6309.

Liang, S., Liu, S., Zhang, C., and Zhou, Y. (2007). A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. *Proteins*, **69**:244–253.

Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, **2**:18–22.

Lindahl, E. and Delarue, M. (2005). Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. *Nucleic Acids Res.*, **33**:4496–4506.

Liu, S., Liu, S., Zhu, X., Liang, H., Cao, A., Chang, Z., and Lai, L. (2007). Nonnatural protein-protein interaction-pair design by key residues grafting. *Proc. Natl. Acad. Sci. U.S.A.*, **104**:5330–5335.

Liu, S., Zhang, C., Zhou, H., and Zhou, Y. (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**:93–101.

London, N. and Schueler-Furman, O. (2008). Funnel hunting in a rough terrain: learning and discriminating native energy funnels. *Structure*, **16**:269–279.

Lorenzen, S. and Zhang, Y. (2007). Identification of near-native structures by clustering protein docking conformations. *Proteins*, **68**:187–194.

Loriot, S., Sachdeva, S., Bastard, K., Prevost, C., and Cazals, F. (2011). On the characterization and selection of diverse conformational ensembles with applications to flexible docking. *IEEE/ACM Trans Comput Biol Bioinform*, **8**:487–498.

Lu, M., Dousis, A. D., and Ma, J. (2008a). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.*, **376**:288–301.

Lu, M., Dousis, A. D., and Ma, J. (2008b). OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Sci.*, **17**:1576–1585.

Lukas, S. M., Kroe, R. R., Wildeson, J., Peet, G. W., Frego, L., Davidson, W., Ingraham, R. H., Pargellis, C. A., Labadia, M. E., and Werneburg, B. G. (2004). Catalysis and function of the p38 alpha.MK2a signaling complex. *Biochemistry*, **43**:9950–9960.

Ma, B., Kumar, S., Tsai, C. J., and Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng.*, **12**:713–720.

Ma, X. H., Wang, C. X., Li, C. H., and Chen, W. Z. (2002). A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng.*, **15**:677–681.

MacKerell, A. D., Banavali, N., and Foloppe, N. (2000). Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, **56**:257–265.

MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., *et al.* (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, **102**:3586–3616.

Mackerell, A. D., Feig, M., and Brooks, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*, **25**:1400–1415.

Maenaka, K., van der Merwe, P. A., Stuart, D. I., Jones, E. Y., and Sondermann, P. (2001). The human low affinity Fcgamma receptors IIa, IIb, and III bind IgG with fast kinetics and distinct thermodynamic properties. *J. Biol. Chem.*, **276**:44898–44904.

Magnusson, U., Chaudhuri, B. N., Ko, J., Park, C., Jones, T. A., and Mowbray, S. L. (2002). Hinge-bending motion of D-allose-binding protein from Escherichia coli: three open conformations. *J. Biol. Chem.*, **277**:14077–14084.

Malpeli, G., Folli, C., and Berni, R. (1996). Retinoid binding to retinol-binding protein and the interference with the interaction with transthyretin. *Biochim. Biophys. Acta*, **1294**:48–54.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, **14**:105–113.

Mandl, R. (1985). Orthogonal Latin squares: an application of experiment design to compiler testing. *Commun. ACM*, **28**:1054–1058.

Mannherz, H. G., Goody, R. S., Konrad, M., and Nowak, E. (1980). The interaction of bovine pancreatic deoxyribonuclease I and skeletal muscle actin. *Eur. J. Biochem.*, **104**:367–379.

Marchot, P., Khelif, A., Ji, Y. H., Mansuelle, P., and Bougis, P. E. (1993). Binding of 125I-fasciculin to rat brain acetylcholinesterase. The complex still binds diisopropyl fluorophosphate. *J. Biol. Chem.*, **268**:12458–12467.

Marques, O. and Sanejouand, Y. H. (1995). Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, **23**:557–560.

Marsh, J. A., Dancheck, B., Ragusa, M. J., Allaire, M., Forman-Kay, J. D., and Peti, W. (2010). Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*, **18**:1094–1103.

Marsh, J. A. and Teichmann, S. A. (2011). Relative Solvent Accessible Surface Area Predicts Protein Conformational Changes upon Binding. *Structure*, **19**:859–867.

Martin, V. J., Pitera, D. J., Withers, S. T., Newman, J. D., and Keasling, J. D. (2003). Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.*, **21**:796–802.

Mashiach, E., Nussinov, R., and Wolfson, H. J. (2010). FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins*, **78**:1503–1519.

Matsuzaki, Y., Matsuzaki, Y., Sato, T., and Akiyama, Y. (2009). In silico screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis. *J Bioinform Comput Biol*, **7**:991–1012.

Mattos, C., Bellamacina, C. R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G. A., and Ringe, D. (2006). Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *J. Mol. Biol.*, **357**:1471–1482.

May, A. and Zacharias, M. (2005). Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim. Biophys. Acta*, **1754**:225–231.

May, A. and Zacharias, M. (2008a). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, **70**:794–809.

May, A. and Zacharias, M. (2008b). Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J. Med. Chem.*, **51**:3499–3506.

Mc Leod, J. F., Kowalski, M. A., and Haddad, J. G. (1989). Interactions among serum vitamin D binding protein, monomeric actin, profilin, and profilactin. *J. Biol. Chem.*, **264**:1260–1267.

McDonald, I. K. and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**:777–793.

McDonald, N. A. and Jorgensen, W. L. (1998). Development of an All-Atom Force Field for Heterocycles. Properties of Liquid Pyrrole, Furan, Diazoles, and Oxazoles. *The Journal of Physical Chemistry B*, **102**:8049–8059.

McInnes, C., Mezna, M., and Fischer, P. M. (2005). Progress in the discovery of polo-like kinase inhibitors. *Curr Top Med Chem*, **5**:181–197.

Mcquarrie, D. A. (2000). *Statistical Mechanics*. University Science Books.

Medina, M., Abagyan, R., Gomez-Moreno, C., and Fernandez-Recio, J. (2008). Docking analysis of transient complexes: interaction of ferredoxin-NADP+ reductase with ferredoxin and flavodoxin. *Proteins*, **72**:848–862.

Mendes, J., Soares, C. M., and Carrondo, M. A. (1999). Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers*, **50**:111–131.

Mendez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, **52**:51–67.

Mendez, R., Leplae, R., Lensink, M. F., and Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, **60**:150–169.

Mendoza, L. (2006). A network model for the control of the differentiation process in Th cells. *BioSystems*, **84**:101–114.

Mendoza, L., Thieffry, D., and Alvarez-Buylla, E. R. (1999). Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. *Bioinformatics*, **15**:593–606.

van der Merwe, P. A., Barclay, A. N., Mason, D. W., Davies, E. A., Morgan, B. P., Tone, M., Krishnam, A. K., Ianelli, C., and Davis, S. J. (1994). Human cell-adhesion molecule CD2 binds CD58 (LFA-3) with a very low affinity and an extremely fast dissociation rate but does not bind CD48 or CD59. *Biochemistry*, **33**:10149–10160.

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

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940.

Milligan, R. A., Brisson, A., and Unwin, P. N. (1984). Molecular structure determination of crystalline specimens in frozen aqueous solutions. *Ultramicroscopy*, **13**:1–9.

Minton, A. P. (1981). Excluded volume as a determinant of macromolecular structure and reactivity. *Biopolymers*, **20**:2093–2120.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005). Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, **60**:214–216.

Misura, K. M., Morozov, A. V., and Baker, D. (2004). Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J. Mol. Biol.*, **342**:651–664.

Mitra, P. and Pal, D. (2010). New measures for estimating surface complementarity and packing at protein-protein interfaces. *FEBS Lett.*, **584**:1163–1168.

Miura, S., Li, C. Q., Cao, Z., Wang, H., Wardell, M. R., and Sadler, J. E. (2000). Interaction of von Willebrand factor domain A1 with platelet glycoprotein Ibalpha-(1-289). Slow intrinsic binding kinetics mediate rapid platelet adhesion. *J. Biol. Chem.*, **275**:7539–7546.

Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**:534–552.

Moal, I. H. and Bates, P. A. (2010). SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int J Mol Sci*, **11**:3623–3648.

Monaco-Malbet, S., Berthet-Colominas, C., Novelli, A., Battai, N., Piga, N., Cheynet, V., Mallet, F., and Cusack, S. (2000). Mutual conformational adaptations in antigen and antibody upon complex formation between an Fab and HIV-1 capsid protein p24. *Structure*, **8**:1069–1077.

Moont, G., Gabb, H. A., and Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**:364–373.

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2010). Protein-protein docking dealing with the unknown. *J Comput Chem*, **31**:317–342.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*, **19**:1639–1662.

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., and Lauffenburger, D. A. (2010). Logic-Based Models for the Analysis of Cell Signaling Networks. *Biochemistry*, **49**:3216–3224.

Mosca, R., Pons, C., Fernandez-Recio, J., and Aloy, P. (2009). Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput. Biol.*, **5**:e1000490.

Muller, W. and Sticht, H. (2007). A protein-specifically adapted scoring function for the reranking of docking solutions. *Proteins*, **67**:98–111.

Muller, Y. A., Chen, Y., Christinger, H. W., Li, B., Cunningham, B. C., Lowman, H. B., and de Vos, A. M. (1998a). VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 A resolution and mutational analysis of the interface. *Structure*, **6**:1153–1167.

Muller, Y. A., Kelley, R. F., and de Vos, A. M. (1998b). Hinge bending within the cytokine receptor superfamily revealed by the 2.4 A crystal structure of the extracellular domain of rabbit tissue factor. *Protein Sci.*, **7**:1106–1115.

Mustard, D. and Ritchie, D. W. (2005). Docking essential dynamics eigenstructures. *Proteins*, **60**:269–274.

Nabors, L. B., Mikkelsen, T., Rosenfeld, S. S., Hochberg, F., Akella, N. S., Fisher, J. D., Cloud, G. A., Zhang, Y., Carson, K., Wittemer, S. M., *et al.* (2007). Phase I and correlative biology study of cilengitide in patients with recurrent malignant glioma. *J. Clin. Oncol.*, **25**:1651–1657.

Nam, T. W., Jung, H. I., An, Y. J., Park, Y. H., Lee, S. H., Seok, Y. J., and Cha, S. S. (2008). Analyses of Mlc-IIBGlc interaction and a plausible molecular mechanism of Mlc inactivation by membrane sequestration. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:3751–3756.

Namasivayam, V. and Gunther, R. (2007). pso@autodock: a fast flexible molecular docking program based on Swarm intelligence. *Chem Biol Drug Des*, **70**:475–484.

Nauchitel, V., Villaverde, M. C., and Sussman, F. (1995). Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease. *Protein Sci.*, **4**:1356–1364.

Naylor, R. L., Robertson, A. G., Allen, S. J., Sessions, R. B., Clarke, A. R., Mason, G. G., Burston, J. J., Tyler, S. J., Wilcock, G. K., and Dawbarn, D. (2002). A discrete domain of the human TrkB receptor defines the binding sites for BDNF and NT-4. *Biochem. Biophys. Res. Commun.*, **291**:501–507.

Nguyen, M., Marcellus, R. C., Roulston, A., Watson, M., Serfass, L., Murthy Madiraju, S. R., Goulet, D., Viallet, J., Belec, L., Billot, X., *et al.* (2007). Small molecule obatoclax (GX15-070) antagonizes MCL-1 and overcomes MCL-1-mediated resistance to apoptosis. *Proc. Natl. Acad. Sci. U.S.A.*, **104**:19512–19517.

Nikolovska-Coleska, Z., Meagher, J. L., Jiang, S., Yang, C. Y., Qiu, S., Roller, P. P., Stuckey, J. A., and Wang, S. (2008). Interaction of a cyclic, bivalent smac mimetic with the x-linked inhibitor of apoptosis protein. *Biochemistry*, **47**:9811–9824.

Nishida, M., Nagata, K., Hachimori, Y., Horiuchi, M., Ogura, K., Mandiyan, V., Schlessinger, J., and Inagaki, F. (2001). Novel recognition mode between Vav and Grb2 SH3 domains. *EMBO J.*, **20**:2995–3007.

Norel, R., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1994). Shape complementarity at protein-protein interfaces. *Biopolymers*, **34**:933–940.

Norel, R., Petrey, D., Wolfson, H. J., and Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins*, **36**:307–317.

Northrup, S. H., Allison, S. A., and McCammon, J. A. (1984). Brownian dynamics simulation of diffusion-influenced biomolecular reactions. *J. Chem. Phys.*, **80**:1517–1524.

Noskov, S. Y. and Lim, C. (2001). Free energy decomposition of protein-protein interactions. *Biophys. J.*, **81**:737–750.

Novotny, J., Bruccoleri, R. E., and Saul, F. A. (1989). On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry*, **28**:4735–4749.

Nussinov, R. and Schreiber, G. (2009). *Computational Protein-Protein Interactions*. CRC Press.

Obiero, J., Pittet, V., Bonderoff, S. A., and Sanders, D. A. (2010). Thioredoxin system from Deinococcus radiodurans. *J. Bacteriol.*, **192**:494–501.

Obsil, T., Ghirlando, R., Klein, D. C., Ganguly, S., and Dyda, F. (2001). Crystal structure of the 14-3-3zeta:serotonin N-acetyltransferase complex. a role for scaffolding in enzyme regulation. *Cell*, **105**:257–267.

Offman, M. N., Tournier, A. L., and Bates, P. A. (2008). Alternating evolutionary pressure in a genetic algorithm facilitates protein model selection. *BMC Struct. Biol.*, **8**:34.

Ogasahara, K., Ishida, M., and Yutani, K. (2003). Stimulated interaction between and subunits of tryptophan synthase from hyperthermophile enhances its thermal stability. *J. Biol. Chem.*, **278**:8922–8928.

Okazaki, K. and Takada, S. (2008). Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:11182–11187.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**:1989–2000.

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997). An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**:814–830.

Olson, M. W., Gervasi, D. C., Mobashery, S., and Fridman, R. (1997). Kinetic analysis of the binding of human matrix metalloproteinase-2 and -9 to tissue inhibitor of metalloproteinase (TIMP)-1 and TIMP-2. *J. Biol. Chem.*, **272**:29975–29983.

Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., Bruncko, M., Deckwerth, T. L., Dinges, J., Hajduk, P. J., *et al.* (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, **435**:677–681.

Otzen, D. E. and Fersht, A. R. (1999). Analysis of protein-protein interactions by mutagenesis: direct versus indirect effects. *Protein Eng.*, **12**:41–45.

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nat Rev Drug Discov*, **5**:993–996.

Pacold, M. E., Suire, S., Perisic, O., Lara-Gonzalez, S., Davis, C. T., Walker, E. H., Hawkins, P. T., Stephens, L., Eccleston, J. F., and Williams, R. L. (2000). Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase gamma. *Cell*, **103**:931–943.

Paesen, G. C., Siebold, C., Harlos, K., Peacey, M. F., Nuttall, P. A., and Stuart, D. I. (2007). A tick protein with a modified Kunitz fold inhibits human tryptase. *J. Mol. Biol.*, **368**:1172–1186.

Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. (2000). BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, **39**:372–384.

Paschalidis, I. C., Shen, Y., Vakili, P., and Vajda, S. (2007). SDU: A Semidefinite Programming-Based Underestimation Method for Stochastic Global Optimization in Protein Docking. *IEEE Trans Automat Contr*, **52**:664–676.

Patel, S. and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J Comput Chem*, **25**:1–15.

Patel, S., Mackerell, A. D., and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J Comput Chem*, **25**:1504–1514.

Pautsch, A., Vogelsgesang, M., Trankle, J., Herrmann, C., and Aktories, K. (2005). Crystal structure of the C3bot-RalA complex reveals a novel type of action of a bacterial exoenzyme. *EMBO J.*, **24**:3670–3680.

Pedotti, M., Simonelli, L., Livoti, E., and Varani, L. (2011). Computational docking of antibody-antigen complexes, opportunities and pitfalls illustrated by influenza hemagglutinin. *Int J Mol Sci*, **12**:226–251.

Perot, S., Sperandio, O., Miteva, M. A., Camproux, A. C., and Villoutreix, B. O. (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today*, **15**:656–667.

Petoukhov, M. V. and Svergun, D. I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.*, **89**:1237–1250.

Petrone, P. and Pande, V. S. (2006). Can conformational change be described by only a few normal modes? *Biophys. J.*, **90**:1583–1593.

Pettigrew, D. W., Meadow, N. D., Roseman, S., and Remington, S. J. (1998). Cation-promoted association of Escherichia coli phosphocarrier protein IIAGlc with regulatory target protein glycerol kinase: substitutions of a Zinc(II) ligand and implications for inducer exclusion. *Biochemistry*, **37**:4875–4883.

Pielak, G. J. and Wang, X. (2001). Interactions between yeast iso-1-cytochrome c and its peroxidase. *Biochemistry*, **40**:422–428.

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*, **21**:1472–1478.

Pierce, B. and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, **67**:1078–1086.

Piervincenzi, R. T. and Chilkoti, A. (2004). Effect of genetic circular permutation near the active site on the activity and stability of an enzyme inhibitor. *Biomol. Eng.*, **21**:33–42.

Pokarowski, P., Kloczkowski, A., Jernigan, R. L., Kothari, N. S., Pokarowska, M., and Kolinski, A. (2005). Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins*, **59**:49–57.

Poli, R. (2008). Analysis of the publications on the applications of particle swarm optimisation. *J. Artif. Evol. App.*, **2008**:3:1–3:10.

Pons, C., D'Abramo, M., Svergun, D. I., Orozco, M., Bernado, P., and Fernandez-Recio, J. (2010). Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. *J. Mol. Biol.*, **403**:217–230.

Pons, C. and Fernandez-Recio, J. (2011). Predicting binding affinities in protein-protein interactions by computational docking. *Submitted*.

Pornillos, O., Alam, S. L., Rich, R. L., Myszka, D. G., Davis, D. R., and Sundquist, W. I. (2002). Structure and functional interactions of the Tsg101 UEV domain. *EMBO J.*, **21**:2397–2406.

Prasad, P. A., Vengadesan, K., and Gautham, N. (2005). MOLS–a program to explore the potential energy surface of a peptide and locate its low energy conformations. *In Silico Biol. (Gedrukt)*, **5**:401–405.

Presta, L., Sims, P., Meng, Y. G., Moran, P., Bullens, S., Bunting, S., Schoenfeld, J., Lowe, D., Lai, J., Rancatore, P., *et al.* (2001). Generation of a humanized, high affinity anti-tissue factor antibody for use as a novel antithrombotic therapeutic. *Thromb. Haemost.*, **85**:379–389.

Pruett, P. S. and Air, G. M. (1998). Critical interactions in binding antibody NC41 to influenza N9 neuraminidase: amino acid contacts on the antibody heavy chain. *Biochemistry*, **37**:10660–10670.

Ptashne, M. and Gann, A. (2002). *Genes and Signals*. Cold Spring Harbor Laboratory Press.

Quinlan, J. R. (1992). Learning With Continuous Classes. In *Proceeding 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348. World Scientific.

Raaf, J., Brunstein, E., Issinger, O. G., and Niefind, K. (2008). The interaction of CK2alpha and CK2beta, the subunits of protein kinase CK2, requires CK2beta in a preformed conformation and is enthalpically driven. *Protein Sci.*, **17**:2180–2186.

Rajapakse, H. A. (2007). Small molecule inhibitors of the XIAP protein-protein interaction. *Curr Top Med Chem*, **7**:966–971.

Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, **6**:R40.

Ravikant, D. V. and Elber, R. (2011). Energy design for protein-protein interactions. *J Chem Phys*, **135**:065102.

Rebek, J. (1987). Model studies in molecular recognition. *Science*, **235**:1478–1484.

Reindl, W., Yuan, J., Kramer, A., Strebhardt, K., and Berg, T. (2008). Inhibition of polo-like kinase 1 by blocking polo-box domain-dependent protein-protein interactions. *Chem. Biol.*, **15**:459–466.

Retraction (2010). Retraction of articles by H. M. Krishna Murthy et al. *Acta Crystallogr. D Biol. Crystallogr.*, **66**:222.

Reyes-Turcu, F. E. and Wilkinson, K. D. (2009). Polyubiquitin binding and disassembly by deubiquitinating enzymes. *Chem. Rev.*, **109**:1495–1508.

Reynolds, C., Damerell, D., and Jones, S. (2009). ProtorP: a protein-protein interaction analysis server. *Bioinformatics*, **25**:413–414.

Ritchie, D. W. and Kemp, G. J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins*, **39**:178–194.

Ritchie, D. W., Kozakov, D., and Vajda, S. (2008). Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, **24**:1865–1873.

Ritchie, D. W. and Venkatraman, V. (2010). Ultra-fast FFT protein docking on graphics processors. *Bioinformatics*, **26**:2398–2405.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., *et al.* (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**:940–943.

Roe, S. M., Ali, M. M., Meyer, P., Vaughan, C. K., Panaretou, B., Piper, P. W., Prodromou, C., and Pearl, L. H. (2004). The Mechanism of Hsp90 regulation by the protein kinase-specific cochaperone p50(cdc37). *Cell*, **116**:87–98.

Rojnuckarin, A., Livesay, D. R., and Subramaniam, S. (2000). Bimolecular reaction simulation using Weighted Ensemble Brownian dynamics and the University of Houston Brownian Dynamics program. *Biophys. J.*, **79**:686–693.

Rueda, M., Bottegoni, G., and Abagyan, R. (2009). Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J Chem Inf Model*, **49**:716–725.

Rueda, M., Chacon, P., and Orozco, M. (2007). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**:565–575.

Ruvinsky, A. M., Kirys, T., Tuzikov, A. V., and Vakser, I. A. (2011). Side-Chain Conformational Changes upon Protein-Protein Association. *J. Mol. Biol.*, **408**:356–365.

Ruvinsky, A. M. and Vakser, I. A. (2008). Chasing funnels on protein-protein energy landscapes at different resolutions. *Biophys. J.*, **95**:2150–2159.

Ruvinsky, A. M. and Vakser, I. A. (2009). The ruggedness of protein-protein energy landscape and the cutoff for 1/r(n) potentials. *Bioinformatics*, **25**:1132–1136.

Ryan, C. A., Hass, G. M., and Kuhn, R. W. (1974). Purification and properties of a carboxypeptidase inhibitor from potatoes. *J. Biol. Chem.*, **249**:5495–5499.

Rykunov, D. and Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, **11**:128.

Salemme, F. R. (1976). An hypothetical structure for an intermolecular electron transfer complex of cytochromes c and b5. *J. Mol. Biol.*, **102**:563–568.

Sanchez, L., Chaouiya, C., and Thieffry, D. (2008). Segmenting the fly embryo: logical analysis of the role of the segment polarity cross-regulatory module. *Int. J. Dev. Biol.*, **52**:1059–1075.

Sancho, J. and Gomez-Moreno, C. (1991). Interaction of ferredoxin-NADP+ reductase from Anabaena with its substrates. *Arch. Biochem. Biophys.*, **288**:231–238.

Sandak, B., Nussinov, R., and Wolfson, H. J. (1998a). A method for biomolecular structural recognition and docking allowing conformational flexibility. *J. Comput. Biol.*, **5**:631–654.

Sandak, B., Wolfson, H. J., and Nussinov, R. (1998b). Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins*, **32**:159–174.

Sander, T., Liljefors, T., and Balle, T. (2008). Prediction of the receptor conformation for iGluR2 agonist binding: QM/MM docking to an extensive conformational ensemble generated using normal mode analysis. *J. Mol. Graph. Model.*, **26**:1259–1268.

Sarrias, M. R., Franchini, S., Canziani, G., Argyropoulos, E., Moore, W. T., Sahu, A., and Lambris, J. D. (2001). Kinetic analysis of the interactions of complement receptor 2 (CR2, CD21) with its ligands C3d, iC3b, and the EBV glycoprotein gp350/220. *J. Immunol.*, **167**:1490–1499.

Sawano, A., Takahashi, T., Yamaguchi, S., Aonuma, M., and Shibuya, M. (1996). Flt-1 but not KDR/Flk-1 tyrosine kinase is a receptor for placenta

growth factor, which is related to vascular endothelial growth factor. *Cell Growth Differ.*, **7**:213–221.

Schaefer, M. and Karplus, M. (1996). A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.*, **100**:1578–1599.

Schiffler, B., Zollner, A., and Bernhardt, R. (2004). Stripping down the mitochondrial cholesterol hydroxylase system, a kinetics study. *J. Biol. Chem.*, **279**:34269–34276.

Schluter, K., Schleicher, M., and Jockusch, B. M. (1998). Effects of single amino acid substitutions in the actin-binding site on the biological activity of bovine profilin I. *J. Cell. Sci.*, **111 (Pt 22)**:3261–3273.

Schmierer, B., Tournier, A. L., Bates, P. A., and Hill, C. S. (2008). Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:6608–6613.

Schneidman-Duhovny, D., Hammel, M., and Sali, A. (2011). Macromolecular docking restrained by a small angle X-ray scattering profile. *J. Struct. Biol.*, **173**:461–471.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005a). Geometry-based flexible and symmetric protein docking. *Proteins*, **60**:224–231.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005b). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, **33**:W363–367.

Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., Nussinov, R., *et al.* (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*, **52**:107–112.

Schneidman-Duhovny, D., Nussinov, R., and Wolfson, H. J. (2004). Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking. *Curr. Med. Chem.*, **11**:91–107.

Schneidman-Duhovny, D., Nussinov, R., and Wolfson, H. J. (2007). Automatic prediction of protein interactions with large scale motion. *Proteins*, **69**:764–773.

Schuster, S. C., Swanson, R. V., Alex, L. A., Bourret, R. B., and Simon, M. I. (1993). Assembly and function of a quaternary signal transduction complex monitored by surface plasmon resonance. *Nature*, **365**:343–347.

Schwarz, F. P., Tello, D., Goldbaum, F. A., Mariuzza, R. A., and Poljak, R. J. (1995). Thermodynamics of antigen-antibody binding using specific anti-lysozyme antibodies. *Eur. J. Biochem.*, **228**:388–394.

Seewald, M. J., Kraemer, A., Farkasovsky, M., Korner, C., Wittinghofer, A., and Vetter, I. R. (2003). Biochemical characterization of the Ran-RanBP1-RanGAP system: are RanBP proteins and the acidic tail of RanGAP required for the Ran-RanGAP GTPase reaction? *Mol. Cell. Biol.*, **23**:8124–8136.

Selzer, T. and Schreiber, G. (1999). Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction. *J. Mol. Biol.*, **287**:409–419.

Seok, C., Rosen, J. B., Chodera, J. D., and Dill, K. A. (2003). MOPED: method for optimizing physical energy parameters using decoys. *J Comput Chem*, **24**:89–97.

Shen, Y., Paschalidis, I. C. h., Vakili, P., and Vajda, S. (2008). Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput. Biol.*, **4**:e1000191.

Shentu, Z., Al Hasan, M., Bystroff, C., and Zaki, M. J. (2008). Context shapes: Efficient complementary shape matching for protein-protein docking. *Proteins*, **70**:1056–1073.

Shi, Y. and Eberhart, R. (1998). A modified particle swarm optimizer. In *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, pages 69–73.

Shiba, T., Kawasaki, M., Takatsu, H., Nogi, T., Matsugaki, N., Igarashi, N., Suzuki, M., Kato, R., Nakayama, K., and Wakatsuki, S. (2003). Molecular mechanism of membrane recruitment of GGA by ARF in lysosomal protein transport. *Nat. Struct. Biol.*, **10**:386–393.

Shimaoka, M., Xiao, T., Liu, J. H., Yang, Y., Dong, Y., Jun, C. D., McCormack, A., Zhang, R., Joachimiak, A., Takagi, J., *et al.* (2003). Structures of the alpha L I domain and its complex with ICAM-1 reveal a shape-shifting pathway for integrin regulation. *Cell*, **112**:99–111.

Shomura, Y., Dragovic, Z., Chang, H. C., Tzvetkov, N., Young, J. C., Brodsky, J. L., Guerriero, V., Hartl, F. U., and Bracher, A. (2005). Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Mol. Cell*, **17**:367–379.

Siddiquee, K. A., Gunning, P. T., Glenn, M., Katt, W. P., Zhang, S., Schrock, C., Schroeck, C., Sebti, S. M., Jove, R., Hamilton, A. D., *et al.* (2007). An oxazole-based small-molecule Stat3 inhibitor modulates Stat3 stability and processing and induces antitumor cell effects. *ACS Chem. Biol.*, **2**:787–798.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**:82–95.

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

Sircar, A. and Gray, J. J. (2010). SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.*, **6**:e1000644.

Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J. J. (2009). Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*, **74**:497–514.

Skiba, N. P., Yang, C. S., Huang, T., Bae, H., and Hamm, H. E. (1999). The alpha-helical domain of Galphat determines specific interaction with regulator of G protein signaling 9. *J. Biol. Chem.*, **274**:8770–8778.

Sleigh, S. H., Seavers, P. R., Wilkinson, A. J., Ladbury, J. E., and Tame, J. R. (1999). Crystallographic and calorimetric analysis of peptide binding to OppA protein. *J. Mol. Biol.*, **291**:393–415.

Sleigh, S. H., Tame, J. R., Dodson, E. J., and Wilkinson, A. J. (1997). Peptide binding in OppA, the crystal structures of the periplasmic oligopeptide binding protein in the unliganded form and in complex with lysyllysine. *Biochemistry*, **36**:9747–9758.

Smallshaw, J. E., Brokx, S., Lee, J. S., and Waygood, E. B. (1998). Determination of the binding constants for three HPr-specific monoclonal antibodies and their Fab fragments. *J. Mol. Biol.*, **280**:765–774.

Smith, A. M., Woodward, M. P., Hershey, C. W., Hershey, E. D., and Benjamin, D. C. (1991). The antigenic surface of staphylococcal nuclease. I. Mapping epitopes by site-directed mutagenesis. *J. Immunol.*, **146**:1254–1258.

Smith, G. R., Sternberg, M. J., and Bates, P. A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.*, **347**:1077–1101.

Smoluchowski, M. (1917). Mathematical theory of the kinetics of the coagulation of colloidal solutions. *Z. Phys. Chem.*, **92**:129–168.

Snyder, J. T., Singer, A. U., Wing, M. R., Harden, T. K., and Sondek, J. (2003). The pleckstrin homology domain of phospholipase C-beta2 as an effector site for Rac. *J. Biol. Chem.*, **278**:21099–21104.

Solernou, A. and Fernandez-Recio, J. (2011). pyDockCG: New Coarse-Grained Potential for Protein-Protein Docking. *J Phys Chem B*, **115**:6032–6039.

Solis, F. J. and Wets, R. J.-B. (1981). Minimization by Random Search Techniques. *Math. Oper. Res.*, **6**:19–30.

Sollner, C., Mentele, R., Eckerskorn, C., Fritz, H., and Sommerhoff, C. P. (1994). Isolation and characterization of hirustasin, an antistasin-type serine-proteinase inhibitor from the medical leech Hirudo medicinalis. *Eur. J. Biochem.*, **219**:937–943.

Sondermann, H., Soisson, S. M., Boykevisch, S., Yang, S. S., Bar-Sagi, D., and Kuriyan, J. (2004). Structural analysis of autoinhibition in the Ras activator Son of sevenless. *Cell*, **119**:393–405.

Song, H., Wang, R., Wang, S., and Lin, J. (2005). A low-molecular-weight compound discovered through virtual database screening inhibits Stat3 function in breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **102**:4700–4705.

Song, Y., Zhang, Y., Bajaj, C. L., and Baker, N. A. (2004). Continuum diffusion reaction rate calculations of wild-type and mutant mouse acetylcholinesterase: adaptive finite element analysis. *Biophys. J.*, **87**:1558–1566.

Soto, C. S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins*, **70**:834–843.

Stanley, W. A., Filipp, F. V., Kursula, P., Schuller, N., Erdmann, R., Schliebs, W., Sattler, M., and Wilmanns, M. (2006). Recognition of a functional peroxisome type 1 target by the dynamic import receptor pex5p. *Mol. Cell*, **24**:653–663.

Steegmaier, M., Hoffmann, M., Baum, A., Lenart, P., Petronczki, M., Krssak, M., Gurtler, U., Garin-Chesa, P., Lieb, S., Quant, J., *et al.* (2007). BI 2536, a potent and selective inhibitor of polo-like kinase 1, inhibits tumor growth in vivo. *Curr. Biol.*, **17**:316–322.

Stein, A., Mosca, R., and Aloy, P. (2011a). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr. Opin. Struct. Biol.*, **21**:200–208.

Stein, A., Rueda, M., Panjkovich, A., Orozco, M., and Aloy, P. (2011b). A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure*, **19**:881–889.

Stites, W. E. (1997). Protein-Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. *Chem. Rev.*, **97**:1233–1250.

Stratikos, E. and Gettins, P. G. (1997). Major proteinase movement upon stable serpin-proteinase complex formation. *Proc. Natl. Acad. Sci. U.S.A.*, **94**:453–458.

Strebhardt, K. and Ullrich, A. (2006). Targeting polo-like kinase 1 for cancer therapy. *Nat. Rev. Cancer*, **6**:321–330.

Strynadka, N. C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., *et al.* (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat. Struct. Biol.*, **3**:233–239.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.*, **105**:6959–6964.

Su, Y., Zhou, A., Xia, X., Li, W., and Sun, Z. (2009). Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Sci.*, **18**:2550–2558.

Sugase, K., Dyson, H. J., and Wright, P. E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, **447**:1021–1025.

Suhre, K. and Sanejouand, Y. H. (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **32**:W610–614.

Sun, H., Nikolovska-Coleska, Z., Lu, J., Meagher, J. L., Yang, C. Y., Qiu, S., Tomita, Y., Ueda, Y., Jiang, S., Krajewski, K., *et al.* (2007). Design, synthesis, and characterization of a potent, nonpeptide, cell-permeable, bivalent Smac mimetic that concurrently targets both the BIR2 and BIR3 domains in XIAP. *J. Am. Chem. Soc.*, **129**:15279–15294.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Svergun, D. I. and Koch, M. H. J. (2003). Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics*, **66**:1735.

Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, **41**:1–7.

Tama, F. and Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**:1–6.

Tame, J. R., Dodson, E. J., Murshudov, G., Higgins, C. F., and Wilkinson, A. J. (1995). The crystal structures of the oligopeptide-binding protein OppA complexed with tripeptide and tetrapeptide ligands. *Structure*, **3**:1395–1406.

Tang, C., Iwahara, J., and Clore, G. M. (2006). Visualization of transient encounter complexes in protein-protein association. *Nature*, **444**:383–386.

Tang, G., Ding, K., Nikolovska-Coleska, Z., Yang, C. Y., Qiu, S., Shangary, S., Wang, R., Guo, J., Gao, W., Meagher, J., *et al.* (2007a). Structure-based design of flavonoid compounds as a new class of small-molecule inhibitors of the anti-apoptotic Bcl-2 proteins. *J. Med. Chem.*, **50**:3163–3166.

Tang, G., Yang, C. Y., Nikolovska-Coleska, Z., Guo, J., Qiu, S., Wang, R., Gao, W., Wang, G., Stuckey, J., Krajewski, K., *et al.* (2007b). Pyrogallol-based molecules as potent inhibitors of the antiapoptotic Bcl-2 proteins. *J. Med. Chem.*, **50**:1723–1726.

Tang, Y. Y., Shi, J., Zhang, L., Davis, A., Bravo, J., Warren, A. J., Speck, N. A., and Bushweller, J. H. (2000). Energetic and functional contribution of residues in the core binding factor beta (CBFbeta) subunit to heterodimerization with CBFalpha. *J. Biol. Chem.*, **275**:39579–39588.

Tarricone, C., Xiao, B., Justin, N., Walker, P. A., Rittinger, K., Gamblin, S. J., and Smerdon, S. J. (2001). The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways. *Nature*, **411**:215–219.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**:199–204.

Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D., and Umeyama, H. (2007). The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins*, **69**:866–872.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, **77**:1905–1908.

Tobi, D. and Bahar, I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A.*, **102**:18908–18913.

Tong, W. and Weng, Z. (2004). Clustering protein-protein docking predictions. *Conf Proc IEEE Eng Med Biol Soc*, **4**:2999–3002.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure*, **16**:295–307.

Tovchigrechko, A. and Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.*, **34**:W310–314.

Trosset, J. Y., Dalvit, C., Knapp, S., Fasolini, M., Veronesi, M., Mantegani, S., Gianellini, L. M., Catana, C., Sundstrom, M., Stouten, P. F., *et al.* (2006). Inhibition of protein-protein interactions: the discovery of druglike beta-catenin inhibitors by combining virtual and biophysical screening. *Proteins*, **64**:60–67.

Tsai, C. J., Kumar, S., Ma, B., and Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8**:1181–1190.

Tse, C., Shoemaker, A. R., Adickes, J., Anderson, M. G., Chen, J., Jin, S., Johnson, E. F., Marsh, K. C., Mitten, M. J., Nimmer, P., *et al.* (2008). ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. *Cancer Res.*, **68**:3421–3428.

Tsiang, M., Jones, G. S., Hung, M., Mukund, S., Han, B., Liu, X., Babaoglu, K., Lansdon, E., Chen, X., Todd, J., *et al.* (2009). Affinities between the binding partners of the HIV-1 integrase dimer-lens epithelium-derived growth factor (IN dimer-LEDGF) complex. *J. Biol. Chem.*, **284**:33580–33599.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.*, **8**:1267–1289.

Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief. Bioinformatics*, **10**:217–232.

Turing, A. (1939). Systems of logic based on ordinals. *P. Lond. Math. Soc.*, **2**:161–228.

Tuzmen, C. and Erman, B. (2011). Identification of ligand binding sites of proteins using the Gaussian Network Model. *PLoS ONE*, **6**:e16474.

Uehara, Y., Tonomura, B., and Hiromi, K. (1978). Direct fluorometric determination of a dissociation constant as low as 10(-10) M for the subtilisin BPN'–protein proteinase inhibitor (Streptomyces subtilisin inhibitor) complex by a single photon counting technique. *J. Biochem.*, **84**:1195–1202.

Vajda, S. and Guarnieri, F. (2006). Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr Opin Drug Discov Devel*, **9**:354–362.

Vajda, S., Vakser, I. A., Sternberg, M. J., and Janin, J. (2002). Modeling of protein interactions in genomes. *Proteins*, **47**:444–446.

Vajda, S., Weng, Z., Rosenfeld, R., and DeLisi, C. (1994). Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry*, **33**:13977–13988.

Van Den Bergh, F. (2002). *An analysis of particle swarm optimizers*. Ph.D. thesis, Pretoria, South Africa, South Africa. AAI0804353.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., *et al.* (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem*, **31**:671–690.

Vengadesan, K., Anbupalam, T., and Gautham, N. (2004). An application of experimental design using mutually orthogonal Latin squares in conformational studies of peptides. *Biochem. Biophys. Res. Commun.*, **316**:731–737.

Vengadesan, K. and Gautham, N. (2003). Enhanced sampling of the molecular potential energy surface using mutually orthogonal latin squares: application to peptide structures. *Biophys. J.*, **84**:2897–2906.

Venkatraman, V., Yang, Y. D., Sael, L., and Kihara, D. (2009). Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*, **10**:407.

Vicentini, A. M., Kieffer, B., Matthies, R., Meyhack, B., Hemmings, B. A., Stone, S. R., and Hofsteenge, J. (1990). Protein chemical and kinetic characterization of recombinant porcine ribonuclease inhibitor expressed in Saccharomyces cerevisiae. *Biochemistry*, **29**:8827–8834.

Viji, S. N., Prasad, P. A., and Gautham, N. (2009). Protein-ligand docking using mutually orthogonal Latin squares (MOLSDOCK). *J Chem Inf Model*, **49**:2687–2694.

Villa Braslavsky, C. I., Nowak, C., Gorlich, D., Wittinghofer, A., and Kuhlmann, J. (2000). Different structural and kinetic requirements for the interaction of Ran with the Ran-binding domains from RanBP2 and importin-beta. *Biochemistry*, **39**:11629–11639.

Villoutreix, B. O., Eudes, R., and Miteva, M. A. (2009). Structure-based virtual ligand screening: recent success stories. *Comb. Chem. High Throughput Screen.*, **12**:1000–1016.

Vincent, J. P. and Lazdunski, M. (1972). Trypsin-pancreatic trypsin inhibitor association. Dynamics of the interaction and role of disulfide bridges. *Biochemistry*, **11**:2967–2977.

Vitalis, A. and Pappu, R. V. (2009). Methods for Monte Carlo simulations of biomacromolecules. *Annu Rep Comput Chem*, **5**:49–76.

de Vries, S. J. and Bonvin, A. M. (2008). How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr. Protein Pept. Sci.*, **9**:394–406.

de Vries, S. J., van Dijk, A. D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A. M. (2007). HADDOCK versus HAD-

DOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, **69**:726–733.

Vucic, D., Franklin, M. C., Wallweber, H. J., Das, K., Eckelman, B. P., Shin, H., Elliott, L. O., Kadkhodayan, S., Deshayes, K., Salvesen, G. S., *et al.* (2005). Engineering ML-IAP to produce an extraordinarily potent caspase 9 inhibitor: implications for Smac-dependent anti-apoptotic activity of ML-IAP. *Biochem. J.*, **385**:11–20.

Walker, K. N., Bottomley, S. P., Popplewell, A. G., Sutton, B. J., and Gore, M. G. (1995). Equilibrium and pre-equilibrium fluorescence spectroscopic studies of the binding of a single-immunoglobulin-binding domain derived from protein G to the Fc fragment from human IgG1. *Biochem. J.*, **310 (Pt 1)**:177–184.

Wallis, R., Leung, K. Y., Pommer, A. J., Videler, H., Moore, G. R., James, R., and Kleanthous, C. (1995). Protein-protein interactions in colicin E9 DNase-immunity protein complexes. 2. Cognate and noncognate interactions that span the millimolar to femtomolar affinity range. *Biochemistry*, **34**:13751–13759.

Wallqvist, A., Jernigan, R. L., and Covell, D. G. (1995). A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci.*, **4**:1881–1903.

Wang, C., Bradley, P., and Baker, D. (2007a). Protein-protein docking with backbone flexibility. *J. Mol. Biol.*, **373**:503–519.

Wang, C., Schueler-Furman, O., and Baker, D. (2005). Improved side-chain modeling for protein-protein docking. *Protein Sci.*, **14**:1328–1339.

Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**:1589–1591.

Wang, G., Nikolovska-Coleska, Z., Yang, C. Y., Wang, R., Tang, G., Guo, J., Shangary, S., Qiu, S., Gao, W., Yang, D., *et al.* (2006). Structure-based design of potent small-molecule inhibitors of anti-apoptotic Bcl-2 proteins. *J. Med. Chem.*, **49**:6139–6142.

Wang, H. and Levinthal, C. (1991). A vectorized algorithm for calculating the accessible surface area of macromolecules. *Journal of Computational Chemistry*, **12**:868–871.

Wang, J. and Verkhivker, G. M. (2003). Energy Landscape Theory, Funnels, Specificity, and Optimal Criterion of Biomolecular Binding. *Phys. Rev. Lett.*, **90**:188101.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J Comput Chem*, **25**:1157–1174.

Wang, L., Sloper, D. T., Addo, S. N., Tian, D., Slaton, J. W., and Xing, C. (2008). WL-276, an antagonist against Bcl-2 proteins, overcomes drug resistance and suppresses prostate tumor growth. *Cancer Res.*, **68**:4377–4383.

Wang, M. C. and Uhlenbeck, G. E. (1945). On the Theory of the Brownian Motion II. *Rev. Mod. Phys.*, **17**:323–342.

Wang, S. X., Pandey, K. C., Scharfstein, J., Whisstock, J., Huang, R. K., Jacobelli, J., Fletterick, R. J., Rosenthal, P. J., Abrahamson, M., Brinen, L. S., *et al.* (2007b). The structure of chagasin in complex with a cysteine protease clarifies the binding mode and evolution of an inhibitor family. *Structure*, **15**:535–543.

Wang, Y. and Witten, I. (1997). Induction of model trees for predicting continuous classes. In *Proc European Conference on Machine Learning Poster Papers*, pages 128–137. Prague, Czech Republic.

Wass, M. N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.*, **7**:469.

Wei, S., Chen, Y., Chung, L., Nagase, H., and Brew, K. (2003). Protein engineering of the tissue inhibitor of metalloproteinase 1 (TIMP-1) inhibitory domain. In search of selective matrix metalloproteinase inhibitors. *J. Biol. Chem.*, **278**:9831–9834.

Wells, J. A. and McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **450**:1001–1009.

Weng, Z., Delisi, C., and Vajda, S. (1997). Empirical free energy calculation: comparison to calorimetric data. *Protein Sci.*, **6**:1976–1984.

Wesson, L. and Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.*, **1**:227–235.

West, A. P., Giannetti, A. M., Herr, A. B., Bennett, M. J., Nangiana, J. S., Pierce, J. R., Weiner, L. P., Snow, P. M., and Bjorkman, P. J. (2001). Mutational analysis of the transferrin receptor reveals overlapping HFE and transferrin binding sites. *J. Mol. Biol.*, **313**:385–397.

Wieczorek, M., Park, S. J., and Laskowski, M. (1987). Covalent hybrids of ovomucoid third domains made from one synthetic and one natural peptide chain. *Biochem. Biophys. Res. Commun.*, **144**:499–504.

Wiedow, O., Schroder, J. M., Gregory, H., Young, J. A., and Christophers, E. (1990). Elafin: an elastase-specific inhibitor of human skin. Purification, characterization, and complete amino acid sequence. *J. Biol. Chem.*, **265**:14791–14795.

Wigelsworth, D. J., Krantz, B. A., Christensen, K. A., Lacy, D. B., Juris, S. J., and Collier, R. J. (2004). Binding stoichiometry and kinetics of the interaction of a human anthrax toxin receptor, CMG2, with protective antigen. *J. Biol. Chem.*, **279**:23349–23356.

Wilderspin, A. F. and Sugrue, R. J. (1994). Alternative native flap conformation revealed by 2.3 A resolution structure of SIV proteinase. *J. Mol. Biol.*, **239**:97–103.

Wodak, S. J. and Janin, J. (1978). Computer analysis of protein-protein interaction. *J. Mol. Biol.*, **124**:323–342.

Woodbury, C. (2008). *Introduction to macromolecular binding equilibria*. CRC Press.

Wriggers, W. (2010). Using Situs for the integration of multi-resolution structures. *Biophys Rev*, **2**:21–27.

Wu, Y., Bressette, D., Carrell, J. A., Kaufman, T., Feng, P., Taylor, K., Gan, Y., Cho, Y. H., Garcia, A. D., Gollatz, E., *et al.* (2000). Tumor necrosis factor

(TNF) receptor superfamily member TACI is a high affinity receptor for TNF family members APRIL and BLyS. *J. Biol. Chem.*, **275**:35478–35485.

Wu, Y., Lu, M., Chen, M., Li, J., and Ma, J. (2007). OPUS-Ca: a knowledge-based potential function requiring only Calpha positions. *Protein Sci.*, **16**:1449–1463.

Wyer, J. R., Willcox, B. E., Gao, G. F., Gerth, U. C., Davis, S. J., Bell, J. I., van der Merwe, P. A., and Jakobsen, B. K. (1999). T cell receptor and coreceptor CD8 alphaalpha bind peptide-MHC independently and with distinct kinetics. *Immunity*, **10**:219–225.

Wyman, J. and Gill, S. (1990). *Binding and linkage: functional chemistry of biological macromolecules*. University Science Books.

Xu, D., Lin, S. L., and Nussinov, R. (1997). Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.*, **265**:68–84.

Yang, L., Song, G., and Jernigan, R. L. (2007a). How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.*, **93**:920–929.

Yang, L. W., Eyal, E., Chennubhotla, C., Jee, J., Gronenborn, A. M., and Bahar, I. (2007b). Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure*, **15**:741–749.

Yang, S. Q., Wang, C. I., Gillmor, S. A., Fletterick, R. J., and Craik, C. S. (1998). Ecotin: a serine protease inhibitor with two distinct and interacting binding sites. *J. Mol. Biol.*, **279**:945–957.

Yang, Y. and Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.*, **17**:1212–1219.

Yin, H. and Hamilton, A. D. (2005). Strategies for targeting protein-protein interactions with synthetic agents. *Angew. Chem. Int. Ed. Engl.*, **44**:4130–4163.

Yoo, S., Myszka, D. G., Yeh, C., McMurray, M., Hill, C. P., and Sundquist, W. I. (1997). Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *J. Mol. Biol.*, **269**:780–795.

Yoshikawa, T., Tsukamoto, K., Hourai, Y., and Fukui, K. (2009). Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins. *J Chem Inf Model*, **49**:693–703.

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.*, **12**:1271–1282.

Zacharias, M. (2010a). Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.*, **20**:180–186.

Zacharias, M. (2010b). *Protein-Protein Complexes: Analysis, Modeling and Drug Design*. Imperial College Press.

Zacharias, M. and Sklenar, H. (1999). Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: Application to DNA minor groove ligand complex. *J. Comp. Chem.*, **20**:287–300.

Zeth, K., Ravelli, R. B., Paal, K., Cusack, S., Bukau, B., and Dougan, D. A. (2002). Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA. *Nat. Struct. Biol.*, **9**:906–911.

Zhang, A. (2009). *Protein Interaction Networks*. Cambridge University Press.

Zhang, C., Chen, J., and DeLisi, C. (1999). Protein-protein recognition: exploring the energy funnels near the binding sites. *Proteins*, **34**:255–267.

Zhang, C., Liu, S., and Zhou, Y. (2005a). Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins*, **60**:314–318.

Zhang, C., Liu, S., Zhu, Q., and Zhou, Y. (2005b). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**:2325–2335.

Zhang, C., Vasmatzis, G., Cornette, J. L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, **267**:707–726.

Zhao, Y., Stoffler, D., and Sanner, M. (2006). Hierarchical and multi-resolution representation of protein flexibility. *Bioinformatics*, **22**:2768–2774.

Zhou, A., Huntington, J. A., Pannu, N. S., Carrell, R. W., and Read, R. J. (2003). How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration. *Nat. Struct. Biol.*, **10**:541–544.

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

Zhou, H. X. (1997). Enhancement of protein-protein association rate by interaction potential: accuracy of prediction based on local Boltzmann factor. *Biophys. J.*, **73**:2441–2445.

Zhou, H.-X. (2001). Loops in Proteins Can Be Modeled as Worm-Like Chains. *The Journal of Physical Chemistry B*, **105**:6763–6766.

Zhou, H. X. (2004). Polymer models of protein stability, folding, and interactions. *Biochemistry*, **43**:2141–2154.

Zhou, H. X. (2010). From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophys. J.*, **98**:L15–17.

Zhou, H. X., Rivas, G., and Minton, A. P. (2008). Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences. *Annual Review of Biophysics*, **37**:375–397.

Zimmerman, S. B. and Trach, S. O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *Journal of Molecular Biology*, **222**:599–620.