University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

http://go.warwick.ac.uk/wrap/60673

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

THE UNIVERSITY OF WARWICK

AUTHOR: **Jack Heal** DEGREE: **Ph.D.**

TITLE: **Effects of ligand binding on the rigidity and mobility of proteins: a computational and experimental approach**

DATE OF DEPOSIT: ................................

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: ......................................................

---

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

..................................................................................

..................................................................................

..................................................................................

..................................................................................

..................................................................................

www.warwick.ac.uk

# Effects of ligand binding on the rigidity and mobility of proteins: a computational and experimental approach

by

**Jack Heal**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**MOAC Doctoral Training Center**

September 2013

THE UNIVERSITY OF
WARWICK

# Contents

**List of Tables** v

**List of Figures** vii

**Acknowledgments** x

**Declarations** xi

**Abstract** xii

**Abbreviations** xiii

**Chapter 1 Introduction** 1

- 1.1 Proteins . . . 1
  - 1.1.1 Structure and function . . . 2
  - 1.1.2 Flexibility and mobility . . . 4
  - 1.1.3 Protein folding . . . 5
  - 1.1.4 Folding cores . . . 5
- 1.2 HIV-1 protease . . . 7
  - 1.2.1 HIV-1 and AIDS . . . 7
  - 1.2.2 The HIV-1 virion . . . 8
  - 1.2.3 The HIV-1 life-cycle . . . 9
  - 1.2.4 Antiretroviral therapies . . . 11
- 1.3 Cyclophilin A . . . 11
  - 1.3.1 The peptidyl-prolyl *cis-trans* isomerase family . . . 12
  - 1.3.2 Cyclosporin A . . . 14
  - 1.3.3 HIV-1 capsid protein . . . 14
- 1.4 Rigidity analysis . . . 15
  - 1.4.1 FIRST . . . 17
  - 1.4.2 The pebble game algortihm . . . 18

1.4.3 Hydrogen bonds and $E_{\text{cut}}$ . . . . . . . . . . . . . . . . . . . . 19
1.4.4 Rigidity dilution . . . . . . . . . . . . . . . . . . . . . . . . . 19
1.4.5 Hydrophobic interactions . . . . . . . . . . . . . . . . . . . . 20
1.5 Simulating protein motion . . . . . . . . . . . . . . . . . . . . . . 21
1.5.1 FRODA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
1.5.2 Normal mode analysis . . . . . . . . . . . . . . . . . . . . . . 23
1.6 NMR spectroscopy . . . . . . . . . . . . . . . . . . . . . . . . . . 24
1.6.1 Nuclear spin and magnetic resonance . . . . . . . . . . . . . . 24
1.6.2 2D NMR: HSQC, TOCSY and NOESY . . . . . . . . . . . . 28
1.6.3 3D NMR: TOCSY-HSQC and NOESY-HSQC . . . . . . . . . 30
1.6.4 Hydrogen-deuterium exchange NMR . . . . . . . . . . . . . . 31

**Chapter 2 Inhibition of HIV-1 protease: the rigidity perspective 32**
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
2.1.1 The structure of HIV-1 protease . . . . . . . . . . . . . . . . 33
2.1.2 X-ray crystal structures of HIV-1 protease . . . . . . . . . . . 34
2.1.3 Dynamics of HIV-1 protease . . . . . . . . . . . . . . . . . . . 35
2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.3.1 Processing the crystal structures . . . . . . . . . . . . . . . . 37
2.3.2 Rigidity analysis . . . . . . . . . . . . . . . . . . . . . . . . . 38
2.3.3 Rigidity dilution . . . . . . . . . . . . . . . . . . . . . . . . . 39
2.3.4 $E_{\text{body}}$ and $E_{\text{flap}}$ . . . . . . . . . . . . . . . . . . . . . . . . . . 41
2.3.5 The effect of inhibitors on $E_{\text{body}}$ and $E_{\text{flap}}$ . . . . . . . . . . . 42
2.3.6 The flexibility fraction $\Phi$ . . . . . . . . . . . . . . . . . . . . 43
2.3.7 Complexes with darunavir and their $\Delta\Phi$ values . . . . . . . . 45
2.3.8 The effect of inhibitors on flap flexibility . . . . . . . . . . . . 47
2.3.9 Inhibitor binding and inhibition mechanisms . . . . . . . . . 49
2.3.10 The $\Delta\Phi$ value is not the direct result of backbone mutation . 51
2.3.11 Analysis of structures crystallised with ritonavir . . . . . . . 51
2.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

**Chapter 3 Rigidity analysis of CypA 57**
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
3.1.1 The structure of CypA . . . . . . . . . . . . . . . . . . . . . . 57
3.1.2 X-ray crystal structures of CypA. . . . . . . . . . . . . . . . . 58
3.1.3 Structures of CypA with CsA or CsA analogs . . . . . . . . . 60
3.1.4 The CsA binding site . . . . . . . . . . . . . . . . . . . . . . 61

3.1.5 Measures of global and local rigidity . . . . . . . . . . . . . . 62
3.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
3.2.1 RD plots for structures of unbound CypA . . . . . . . . . . . 63
3.2.2 The effect of ligand binding on RD plots . . . . . . . . . . . . 65
3.2.3 The impact of ligand binding on global and local rigidity . . 65
3.2.4 The rigidity difference map . . . . . . . . . . . . . . . . . . . . 68
3.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

**Chapter 4 Simulations of CypA mobility 74**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.1.1 Folding cores . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.1.2 Pseudodihedral angles . . . . . . . . . . . . . . . . . . . . . . 75
4.1.3 Hydrophobic interactions . . . . . . . . . . . . . . . . . . . . . 75
4.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
4.2.1 Simulating motion with FRODA . . . . . . . . . . . . . . . . . 76
4.2.2 Determining folding cores . . . . . . . . . . . . . . . . . . . . 77
4.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78
4.3.1 Theoretical folding cores and the HDX folding core . . . . . . 78
4.3.2 The mobility of CypA characterised by pseudodihedral angles. 81
4.3.3 The effect of ligand binding on $M$ . . . . . . . . . . . . . . . 82
4.3.4 Burial distances before and after ligand removal . . . . . . . 85
4.3.5 Rigidity analysis with different HB/HP settings . . . . . . . . 87
4.3.6 The effect of HB/HP settings on $M$ . . . . . . . . . . . . . . 89
4.3.7 Further simulation data . . . . . . . . . . . . . . . . . . . . . 90
4.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91
4.4.1 Predicting the HDX folding core . . . . . . . . . . . . . . . . 91
4.4.2 Simulations of protein motion . . . . . . . . . . . . . . . . . . 92

**Chapter 5 Expression, purification and characterisation CypA 94**
5.1 Protein Expression . . . . . . . . . . . . . . . . . . . . . . . . . . 94
5.1.1 Expression of $^{15}$N-labelled CypA . . . . . . . . . . . . . . . 96
5.1.2 Electrophoresis analysis: SDS-PAGE . . . . . . . . . . . . . . 97
5.2 Purification of CypA . . . . . . . . . . . . . . . . . . . . . . . . . 98
5.2.1 Ion exchange chromatography . . . . . . . . . . . . . . . . . . 99
5.2.2 Size exclusion chromatography . . . . . . . . . . . . . . . . . 102
5.2.3 Estimating protein concentration . . . . . . . . . . . . . . . . 103
5.3 Characterisation of CypA . . . . . . . . . . . . . . . . . . . . . . 104
5.3.1 Circular dichroism spectroscopy . . . . . . . . . . . . . . . . . 104

5.3.2 Fluorescence spectroscopy . . . . . . . . . . . . . . . . . . . . . 108
5.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

**Chapter 6 Hydrogen-deuterium exchange NMR experiments 112**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
6.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
6.2.1 Sample preparation . . . . . . . . . . . . . . . . . . . . . . . . 113
6.2.2 Data acquisition, processing and analysis. . . . . . . . . . . . 113
6.2.3 HDX experiments . . . . . . . . . . . . . . . . . . . . . . . . . 115
6.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
6.3.1 The $^1$H-$^{15}$N HSQC spectra for unbound CypA and the CypA-CsA complex . . . . . . . . . . . . . . . . . . . . . . . . . 115
6.3.2 Assignment of the HSQC spectra . . . . . . . . . . . . . . . . . 115
6.3.3 Chemical shift perturbation . . . . . . . . . . . . . . . . . . . . 119
6.3.4 HDX experiments on the CypA-CsA complex . . . . . . . . . 123
6.3.5 HDX experiments on unbound CypA . . . . . . . . . . . . . . 125
6.3.6 Comparison of HDX folding cores . . . . . . . . . . . . . . . . 128
6.3.7 Theoretical folding cores for the CypA-CsA complex . . . . . 135
6.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136

**Chapter 7 Conclusions 137**
7.1 Summary and outcomes . . . . . . . . . . . . . . . . . . . . . . . . . 137
7.2 Context and outlook . . . . . . . . . . . . . . . . . . . . . . . . . . . 140

**Appendix A Amino acid codes 144**

**Appendix B Further rigidity analysis of HIV-1 protease 145**
B.1 Adding hydrogen atoms to PDB files . . . . . . . . . . . . . . . . . . 145
B.2 X-ray crystal B-factors and their influence on $\Delta\Phi$ values . . . . . . . 146
B.3 Complete data list for rigidity of HIV-1 protease . . . . . . . . . . . 147

**Appendix C NMR chemical shift assignment tables 154**

# List of Tables

2.1 The $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ values for structures crystallised with DRV 47

2.2 The values of $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ for structures crystallised with TPV and DRV . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

3.1 X-ray crystal structures of CypA . . . . . . . . . . . . . . . . . . . . . . . 59

3.2 Structural data for 13 CypA PDB entries crystallised with CsA or CsA-derivatives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

3.3 The values of $\Delta X_{\mathrm{LRC}}$, $Z_{\mathrm{LRC}}$, $Z_{\mathrm{LRC}}(\mathrm{U})$ and $\Delta Z_{\mathrm{LRC}}$ for 19 CypA structures, at $E_{\mathrm{cut}} = -1$ kcal/mol . . . . . . . . . . . . . . . . . . . . . 67

4.1 Agreement between theoretical methods and HDX folding core . . . 81

4.2 The number of interactions for each HB and HP setting in 1CWA with no ligand bound . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

4.3 The effect of the HB/HP bond setting on $E_{\mathrm{fc}}$ and folding core size . 88

5.1 The nutrients added to water to make 1 L of MM . . . . . . . . . . . 97

6.1 Chemical shift perturbations and chemical shift differences . . . . . . 123

6.2 The sequence of time intervals for recording HSQC spectra during HDX experiments on CypA-CsA . . . . . . . . . . . . . . . . . . . . 124

6.3 The sequence of time intervals for recording HSQC spectra during HDX experiments on unbound CypA . . . . . . . . . . . . . . . . . . 125

6.4 The folding core residues for the CypA-CsA complex and for unbound CypA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133

6.5 Agreement between theoretical methods and HDX folding core for CypA-CsA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135

A.1 The amino acid codes used throughout this work . . . . . . . . . . . 144

B.1 Running REDUCE before removing water molecules does not affect $\Delta\Phi$.145

B.2 The $\Phi(I)$, $\Phi(U)$ and $\Delta\Phi$ values for all high resolution structures. . . 147

C.1 Assignment of backbone N-H pairs. . . . . . . . . . . . . . . . . . . . . . 154

# List of Figures

1.1 An example of protein structure: HIV-1 protease . . . . . . . . . . . 3
1.2 Timescales of protein motion. . . . . . . . . . . . . . . . . . . . . . . . . 4
1.3 An energy diagram of the one-step folding process . . . . . . . . . . 6
1.4 The HIV-1 virion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
1.5 The HIV lifecycle . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
1.6 The *trans* and *cis* arrangements of a peptide bond . . . . . . . . . . 13
1.7 The CypA-CA complex is essential for productive HIV-1 infection . 16
1.8 Simple examples of overconstrained, isostatic and underconstrained structures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
1.9 Fourier transform of the FID . . . . . . . . . . . . . . . . . . . . . . . 26
1.10 A typical 1D NMR protein spectrum . . . . . . . . . . . . . . . . . . 27
1.11 An example of a $^1$H-$^{15}$N HSQC spectrum: unbound CypA . . . . . . 29
1.12 A diagram showing the distribution of cross-peaks in 3D NMR . . . 30

2.1 The amino acid sequence of HIV-1 protease . . . . . . . . . . . . . . 33
2.2 The structure of HIV-1 protease . . . . . . . . . . . . . . . . . . . . . 34
2.3 An overview of the 212 HIV-1 protease structures . . . . . . . . . . . 36
2.4 The mainchain rigidity of HIV-1 protease . . . . . . . . . . . . . . . 39
2.5 Rigidity dilutions of 2HS1 and 3LZU . . . . . . . . . . . . . . . . . . 40
2.6 The HIV-1 protease mainchain loses rigidity at $E_{\text{cut}} = E_{\text{body}}$ . . . . 42
2.7 Distributions of $E_{\text{body}}$ and $E_{\text{flap}}$ values . . . . . . . . . . . . . . . . 44
2.8 Comparison between $E_{\text{body}}$ and $E_{\text{fc}}$ . . . . . . . . . . . . . . . . . . . 46
2.9 Distributions of $\Phi(\text{I})$, $\Phi(\text{U})$ and $\Delta\Phi$ . . . . . . . . . . . . . . . . . 48
2.10 The number of interactions between enzyme and inhibitor . . . . . . 50
2.11 Tipranavir binds directly to the flap tips . . . . . . . . . . . . . . . . 51
2.12 Rigidity dilutions of 1HXW and 1SH9 . . . . . . . . . . . . . . . . . 53

3.1 The structure of CypA . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
3.2 The CypA binding site . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

3.3 Rigidity dilutions of 3K0N and 1W8V . . . . . . . . . . . . . . . . . 64
3.4 Rigidity dilutions of 1AWQ and 1AWS . . . . . . . . . . . . . . . . . 66
3.5 The $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ values for 13 structures . . . . . . . . . . . 69
3.6 The RDM for the structure 1CWA . . . . . . . . . . . . . . . . . . . 70
3.7 The cumulative RDM for 13 structures of CypA . . . . . . . . . . . 71
3.8 The effect of TPV on the rigidity of HIV-1 protease shown as a cumulative RDM . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

4.1 Construction of the pseudodihedral angle . . . . . . . . . . . . . . . 76
4.2 Burial distances for residues in CypA . . . . . . . . . . . . . . . . . . 79
4.3 Computationally determined folding core residues as predictors of the HDX folding core . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
4.4 Mobility of CypA, determined using FRODA and pseudodihedral angles. 82
4.5 The effect of $E_{\text{cut}}$ on $M$ . . . . . . . . . . . . . . . . . . . . . . . . . 83
4.6 The $\Delta M$ value for each residue at $E_{\text{cut}} = -2$ kcal/mol . . . . . . . . 84
4.7 The $\Delta M$ value for each residue at $E_{\text{cut}} = -0.5, -1.0$ and $-1.5$ kcal/mol 85
4.8 The RDM for 1CWA $\Delta M$ for each residue at $E_{\text{cut}} = -0.5$ and $-1.5$ kcal/mol . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
4.9 Ligand binding affects $R$ for binding site residues . . . . . . . . . . . 87
4.10 Mobility of CypA for the three HB settings. . . . . . . . . . . . . . . 89
4.11 The HB setting can influence protein mobillity. . . . . . . . . . . . . 90

5.1 A diagram of the DNA plasmid used to express CypA . . . . . . . . 95
5.2 An SDS-PAGE gel showing that CypA has been expressed . . . . . . 99
5.3 An ion exchange chromatogram showing elution of CypA . . . . . . 101
5.4 An SDS-PAGE gel showing incomplete purification as a result of IEC 102
5.5 An SDS-PAGE gel showing purified CypA following SEC . . . . . . 103
5.6 The far-UV CD spectrum of CypA . . . . . . . . . . . . . . . . . . . 105
5.7 Thermal denaturation of CypA using CD . . . . . . . . . . . . . . . 106
5.8 The CD spectrum of CypA after heating and then cooling . . . . . . 107
5.9 Thermal denaturation of CypA is consistent between pH = 5.3 and pH = 7.3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.10 Fluorescence emission spectra of CypA with increasing CsA concentration . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
5.11 Fractional fluoresence change during the CypA-CsA titration . . . . 111

6.1 The HSQC spectrum of CypA with CsA bound . . . . . . . . . . . . 116
6.2 The HSQC spectrum of unbound CypA . . . . . . . . . . . . . . . . 117

6.3 Overlaid HSQC spectra of CypA with and without CsA bound . . . 118
6.4 Strip plots of NMR data demonstrating sequential assignment . . . . 120
6.5 A section of the overlaid HSQC spectra showing chemical shift perturbation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121
6.6 Chemical shift perturbations as a result of CypA-CsA complex formation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122
6.7 Chemical shift perturbations shown on 1CWA . . . . . . . . . . . . . 124
6.8 HSQC spectra of the CypA-CsA complex at different time intervals following initiation of HDX . . . . . . . . . . . . . . . . . . . . . . . . 126
6.9 Cross-peaks for residues 55, 97, 122 and 139 in the HSQC spectra of the CypA-CsA complex at different time intervals following initiation of HDX . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 127
6.10 HSQC spectra of unbound CypA at different time intervals following initiation of HDX . . . . . . . . . . . . . . . . . . . . . . . . . . . . 129
6.11 Cross-peaks for residues 55, 97, 122 and 139 in the HSQC spectra of unbound CypA at different time intervals following initiation of HDX 130
6.12 HSQC plot showing the HDX folding core of the CypA-CsA complex and the unbound CypA . . . . . . . . . . . . . . . . . . . . . . . . . 131
6.13 HDX folding cores shown on 1CWA . . . . . . . . . . . . . . . . . . 132
6.14 The effect of CsA binding on the HDX folding core shown on 1CWA 134

B.1 Flap rigidity against relative flap B-factors for 26 structures . . . . . 146

# Acknowledgments

I thank my supervisors, Professor Robert Freedman and Professor Rudolf Römer, for their valuable insights throughout the course of my study. Professor Freedman has continually been informative with respect to the biological aspects of my work, both in terms of laboratory work and in directing our simulations towards biologically relevant questions. Professor Römer has always encouraged me to strive for excellence in my approach to research, with his keen sense of what science should be and what a scientist should do. Dr Stephen Wells provided expert guidance and supervision during our work on HIV-1 protease and wrote the codes we use in Chapter 4 to calculate burial distances and pseudodihedral angles from simulations of protein motion. Dr Claudia Blindauer kindly facilitated the NMR experiments presented in Chapter 6 and offered invaluable help with the assignment procedure. I thank Professor Gunther Fischer, who provided the plasmid for expressing CypA.

I am grateful to the friends I have worked alongside in C10 and PS001. Dr Katrine Wallis, Kelly Walker and John Blood of the Freedman group saved me from drowning (metaphor) when I began working in the lab. Dr Emilio Jimenez-Roldan helped to initiate the study in Chapter 2 and along with Dr Sebastian Pinski and Dr Andrea Fischer made sure that I was happily settled in the Römer group. I am also thankful to the MOAC staff and the nine other students in my cohort.

Outside of the university, I have always been supported by my sisters, Liz and Steph, and my parents, “Mum” and “Dad”. To the newest members of my family, George (4) and Ruby (1), I am grateful, not so much for their insightful questions as for the joy they bring to my visits home.

Thank you Neesha, for being The Best.

# Declarations

The work contained within this thesis is entirely my own original work, except where acknowledged in the text. The work was carried out under the supervision of Professor R. B. Freedman and Professor R. A. Römer. Dr C. A. Blindauer joined my supervisory team at the beginning of my third year of study. The term *we* is used at various points in this thesis to indicate that research strategies and interpretations were the result of discussion between me and my supervisors, but all the experiments and computations leading to the results reported in this thesis represent entirely my own work.

The work in this thesis has not been submitted for a degree at another time or institution. Some of the work has been published in a journal, or is in preparation for future publication.

# Abstract

The interplay between protein structure, flexibility, mobility and function is a central concern in structural biology. Here, we have studied the interactions of two different proteins, HIV-1 protease and cyclophilin A (CypA), with a variety of ligands. HIV-1 protease is a key modern drug target due to its role in the life-cycle of the virus HIV-1. CypA is a multifunctional protein which acts principally as an enzyme during protein folding, but also as the primary binding partner for the immunosuppressant drug cyclosporin A (CsA). Using a computational approach we have studied the manner in which different drugs affect the flexibility of HIV-1 protease. We have used experimental and computational approaches to investigate the binding effect of CsA on the structure, flexibility and mobility of CypA.

The wealth of structural data available, particularly from X-ray crystallography studies, creates an opportunity for computational scientists to provide data on flexibility and mobility in order to augment the structural data and inform future experiments. We use rapid computational tools to model the flexibility (FIRST) and mobility (FRODA) with a variety of structural data as input.

We characterised, with FIRST, the effect of ligand binding on the rigidity of more than 200 X-ray crystal structures of HIV-1 protease. A similar approach has been applied to the structural dataset available for CypA. In addition, we have used FRODA to simulate the mobility of the protein. A new procedure, of tracking surface exposure during large-scale simulations of protein motion and combining the information with rigidity analysis data, was used to try to predict results of hydrogen-deuterium exchange NMR experiments (HDX). Experimentally, we have characterised CypA and its binding with CsA and subsequently performed NMR spectroscopy including HDX on both the unliganded CypA and the CypA-CsA complex. Small changes in chemical shift and the HDX folding core were observed upon ligand binding.

This is the first study of the CypA-CsA complex using HDX. Our method for predicting the results of HDX for CypA is an improvement on the FIRST-only approach, but we find that it is not yet sufficiently sensitive to predict the effects of ligand binding on CypA using these experiments. It is hoped that this work will contribute to the general understanding of ligand interactions, the particular interactions involving HIV-1 protease and CypA, and the applications of computational simulation using rigidity analysis and large-scale coarse-grained motion.

# Abbreviations

| | | Page |
|---|---|---|
| AIDS | Acquired Immunodeficiency Syndrome | 7 |
| APV | Amprenavir (HIV-1 protease inhibitor) | 35 |
| ATZ | Atazanavir (HIV-1 protease inhibitor) | 35 |
| BMRB | Biological Magnetic Resonance Data Bank | 115 |
| CA | HIV-1 capsid protein | 8 |
| CD | Circular Dichroism | 104 |
| CN | Calcineurin | 14 |
| CsA | Cyclosporin A | 11 |
| CypA | Cyclophilin A | 13 |
| DMP | DMP-323 (HIV-1 protease inhibitor) | 35 |
| DRV | Darunavir (HIV-1 protease inhibitor) | 35 |
| DSSP | Define Secondary Structure of Proteins algorithm | 33 |
| $E_{\mathrm{cut}}$ | Energy cutoff | 19 |
| EDTA | Ethylenediaminetetraacetic acid | 97 |
| ENM | Elastic Network Model | 22 |

| | | |
|---|---|---|
| FDA | Food and Drug Administration | 11 |
| FID | Free Induction Decay | 26 |
| FIRST | Floppy Inclusions and Rigid Substructure Topography | 17 |
| FPLC | Fast Protein Liquid Chromatography | 100 |
| FRET | Förster Resonance Energy Transfer | 3 |
| FRODA | Framework Rigidity Optimized Dynamic Algorithm | 22 |
| HB | Hydrogen bond | 19 |
| HDX | Hydrogen-Deuterium Exchange | 24 |
| HDX MS | HDX coupled with mass spectroscopy | 141 |
| HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid | 96 |
| HIV | Human Immunodificiency Virus | 7 |
| HP | Hydrophobic tether | 20 |
| HSQC | Heteronuclear Single Quantum Correlation | 28 |
| IDV | Indinavir (HIV-1 protease inhibitor) | 35 |
| IEC | Ion exchange chromatography | 99 |
| IPTG | isopropyl-$\beta$-D-thiogalactopyranoside | 94 |
| LB | Lysogeny Broth | 95 |
| LPV | Lopinavir (HIV-1 protease inhibitor) | 35 |
| LRC | Largest Rigid Cluster | 20 |
| MD | Molecular Dynamics | 4 |
| MM | Minimal Medium | 96 |

| | | |
|---|---|---|
| NFV | Nelfinavir (HIV-1 protease inhibitor) | 35 |
| NMA | Normal Mode Analysis | 23 |
| NMR | Nuclear Magnetic Resonance | 2 |
| NOESY | Nuclear Overhauser Effect Spectroscopy | 28 |
| $OD_{600}$ | Optical Density at 600 nm | 95 |
| PDB | Protein Data Bank | 2 |
| PEP | The peptide (ACE)TI(NLE)(NLE)QR (HIV-1 protease inhibitor) | 35 |
| PPI | Peptidyl Prolyl *cis-trans*-Isomerase | 12 |
| ppm | Parts per million | 27 |
| RCD | Rigid Cluster Decomposition. | 19 |
| RD | Rigidity Dilution | 19 |
| RDM | Rigidity Difference Map | 57 |
| RF | Radio Frequency | 25 |
| RMSD | Root-Mean-Square Deviation | 92 |
| RTV | Ritonavir (HIV-1 protease inhibitor) | 35 |
| SDS-PAGE | Sodium Diodecyl Sulphate Polyacrylamide Gel Electrophoresis | 97 |
| SEC | Size exclusion chromatography | 99 |
| SQV | Saquinavir (HIV-1 protease inhibitor) | 35 |
| TOCSY | Total Correlation Spectroscopy | 28 |
| TPV | Tirpranavir (HIV-1 protease inhibitor) | 35 |
| TRIS | Tris(hydroxymethyl)aminomethane | 100 |

# Chapter 1

# Introduction

## 1.1 Proteins

Proteins are large biomolecules which are found in every living cell and essential for all forms of life. They vary dramatically in shape and function and are at the centre of most processes within cells. Signalling, transport, catalysis and DNA replication are all driven by interactions between proteins and other molecules [Henzler-Wildman and Kern, 2007; Fischer et al., 2010; Bewley and ul Hussan, 2013]. Throughout this thesis we focus on interactions between proteins and small molecules referred to as ligands. Protein-ligand interactions play a pivotal role in the functional pathways of many cellular processes [Fielding, 2007; Schreyer and Blundell, 2009] with examples including antigen-antibody interactions, proteins which act as receptors for cellular signalling and the protein-DNA interactions involved in DNA replication. The act of binding to a ligand can have an effect on the structure, flexibility, mobility and therefore function of the protein [Harding and Chowdhry, 2001; Levin et al., 2007]. This can have a cascading effect on the processes dependent on the protein. Clearly, the interactions and their consequences vary according to the identity of both the protein and the ligand. Detailed knowledge of protein-ligand interactions is required so that these interactions can be modulated, making the control of biological processes possible. Consequently, understanding and predicting the effects of ligand binding is a problem of central importance in molecular biology [Harding and Chowdhry, 2001; Bucher et al., 2011; Greenidge et al., 2013].

Enzyme-ligand interactions form a group of important complexes which regulate numerous processes. The two proteins on which we focus in this thesis, HIV-1 protease and cyclophilin A, are both enzymes with functions informed by interactions with other molecules. We use computational models of protein flexibility and mobility to determine how these are affected by ligand binding. X-ray crystal structures taken from the protein data bank (PDB) are used as input for our models. We characterise the effects of ligand binding experimentally using circular dichroism, protein fluorescence and nuclear magnetic resonance spectroscopy (NMR), in particular hydrogen deuterium exchange (HDX) NMR.

### 1.1.1 Structure and function

Proteins are long chains of amino acids, or residues, joined together by peptide bonds. There are 20 amino acids principally used by modern organisms in protein synthesis [Cleaves II, 2010], and their respective three-letter and one-letter codes are given in Appendix A (Table A.1). The sequence of residues which make up the links in the polypeptide chain forms the primary structure of the protein. Secondary structure motifs arise when hydrogen bonds form between amino acids in the chain. These are local sub-structures, most commonly arranged in $\alpha$-helices and $\beta$-strands which can be identified by their hydrogen bonding patterns. Interactions between the sub-structures cause them to fold together in an arrangement known as the tertiary structure of the protein. In certain proteins such as HIV-1 protease, discussed in Section 1.2, two or more polypeptide chains join together non-covalently to form the quaternary structure of the protein. The structure of HIV-1 protease is depicted in Figure 1.1. The molecular visualisation software used throughout this work is PYMOL [Schrödinger, LLC, 2013]. The secondary structural motifs are represented as arrows ($\beta$-sheet) and helices ($\alpha$-helices) in Figure 1.1(a), and individual atoms as spheres in Figure 1.1(b). Figure 1.1(a) shows how the secondary structure motifs fold together to form the tertiary structure in each individual monomer. Figure 1.1(b) demonstrates how densely the atoms in a protein are arranged. The relationship between structure and function is also hinted at in this image, since the cavity in the centre of the protein is key to its role as an enzyme [Ishima et al., 1999]. Intrinsic molecular stability is a function of protein structure, and the three dimensional conformation of the protein plays a big role in determining its function [Burton et al., 2002; Scarff et al., 2008]. As a consequence, much effort has been expended computationally and experimentally in an attempt to elucidate the relationship between the amino acid sequence and the final conformation of the protein [James and Tawfik, 2003]. This is known as the protein folding problem [Dill et al., 2008], and is discussed further in Section 1.1.3. The traditional view is that the sequence of amino acids uniquely determines a three dimensional structure which alone represents the folded state of the protein [Baldwin, 1995]. This view is supported by X-ray crystallography experiments, where the static structure of

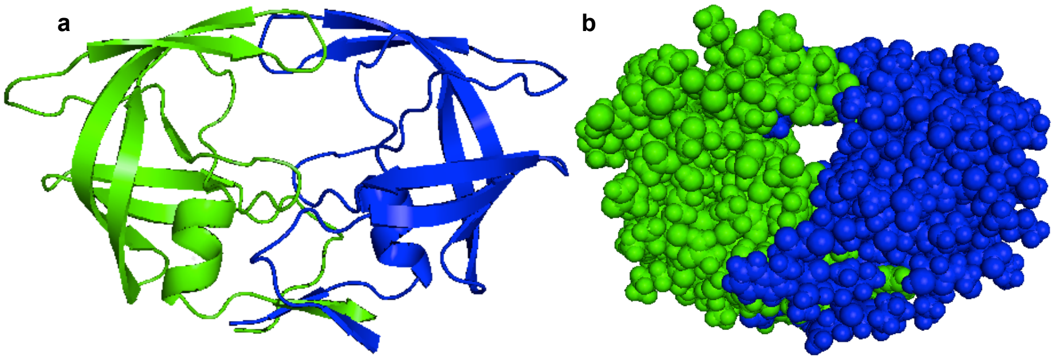


Figure 1.1: An example of protein structure. (a) HIV-1 protease represented as a cartoon. The secondary structure motifs are shown as arrows ($\beta$-sheet) and helices ($\alpha$-helices). Two individual monomers, in green and blue, are joined non-covalently to form the quaternary structure of the protein. (b) The same structure is shown with individual atoms represented as spheres.

conformationally and chemically homogeneous molecules is determined [James and Tawfik, 2003]. Dynamic techniques such as NMR, Förster resonance energy transfer (FRET) and protein folding studies challenged this view, resulting in the new model of protein structure, whereby the molecules exist as part of an ensemble of different conformations in equilibrium [Baldwin, 1995; Dill and Chan, 1997]. Conformational diversity has also been demonstrated using X-ray crystallography [Kohn et al., 2010]. The final conformation of a protein is not an entirely fixed rigid structure [James and Tawfik, 2003]. Many proteins have intrinsically flexible regions which are likely to be more dynamic than the rigid regions [Jacobs et al., 2001]. The dynamism of the folded protein may result in functional promiscuity and is thought to be important for the rapid evolution of proteins [Tokuriki and Tawfik, 2009]. Catalytically promiscuous enzymes [Colman and Smith, 2002] as well as moonlighting enzymes, which perform structural or regulatory roles alongside catalytic ones [Copley, 2003], demonstrate the utility of flexible and dynamic structure. Intrinsically disordered proteins provide extreme examples of flexible, dynamic molecules which subvert the one sequence, one structure, one function model [Tompa, 2012]. The static image of a final protein structure is therefore misleading. The image represents merely a snapshot in time of a dynamic structure which may fluctuate to varying degrees on vast (fs — ms) timescales [Henzler-Wildman and Kern, 2007].

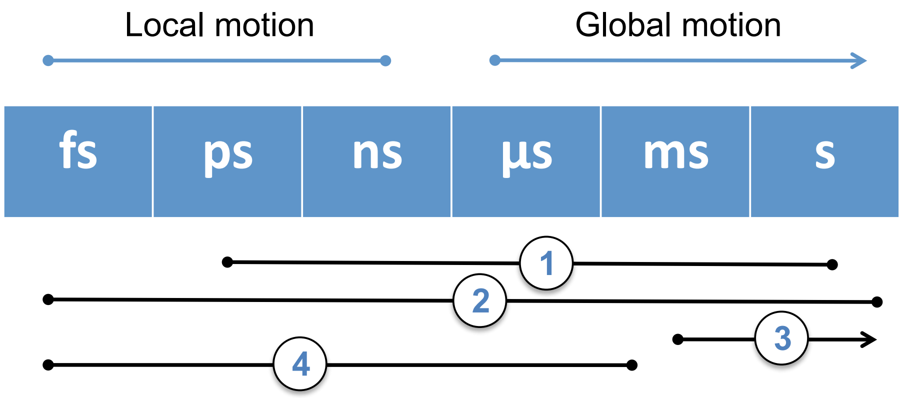


Figure 1.2: Timescales of protein motion along with approaches which can be used to detect fluctuations at the different timescales. (1) NMR, (2) X-ray diffraction, (3) hydrogen-deuterium exchange, (4) molecular dynamics. Figure adapted from [Henzler-Wildman and Kern, 2007].

### 1.1.2 Flexibility and mobility

Proteins are dynamic locally on rather short timescales and globally on much longer timescales. Bond vibrations occur at the femtosecond level, local atomic fluctuations at the level of picoseconds and collective atomic fluctuations on the nanosecond timescale [Henzler-Wildman and Kern, 2007]. Larger scale motion and protein folding occurs typically on a timescale of micro- to milliseconds [Kubelka et al., 2004]. It has been shown that motions on the pico- to nanosecond timescale can facilitate slower, larger-scale motions [Henzler-Wildman et al., 2007]. The timescales of protein motion are shown in Figure 1.2 along with some of the methods discussed in this work which can be used to study them [Henzler-Wildman and Kern, 2007]. NMR spectroscopy is able to probe motion on the picosecond to second timescale, whilst X-ray diffraction can reach a timescale of femtoseconds. HDX experiments, which we introduce in Section 1.6.4 and discuss further in Chapters 4 and 6, deal with longer timescales. Molecular dynamics (MD) simulations are restricted to the $\mu$s timescales in most cases [Klepeis et al., 2009], although simulations on the order of 1 ms have been reported on a highly specialised computational system [Lindorff-Larsen et al., 2011; Piana et al., 2013]. These computationally demanding simulations require huge amounts of computational time [Jimenez-Roldan et al., 2012]. Simplified modelling techniques such as coarse-graining permit the simulation of the large-scale protein motion probed by HDX which are currently inaccessible to MD.

### 1.1.3 Protein folding

The protein folding problem, essentially “how do proteins fold?” has been a prevalent question for scientists during the past 50 years [Dill and MacCallum, 2012]. The interplay between protein structure and function and the impact of misfolding underline its importance. There are a huge number of conformations which are accessible to a protein – so many indeed that if a typical protein achieved its final fold by randomly sampling conformations, the folding time scale would be on the order of billions of years. In reality, proteins fold on the $\mu$s – ms time scale [Kubelka et al., 2004]. This discrepancy of timescales is known as Levinthal’s paradox [Levinthal, 1969]. The paradox can be resolved by assuming an accelerated folding process: conformations are not randomly sampled, rather the process is driven towards the final conformation by local interactions. Identifying the type of interactions which are most influential and the manner in which they drive folding is the essence of the protein folding problem. There are two principal competing theories on how protein folding initiates, with neither being universally accepted. Indeed, it may well be that both are valid depending on which protein is being investigated [Hespenheide et al., 2002]. These two models are diffusion-collision [Karplus and Weaver, 1994] and nucleation-condensation [Itzhaki et al., 1995]. In the diffusion-collision model local interactions guide the independent formation of incipient secondary structure units and hydrophobic clusters. The resulting locally-formed microdomains move diffusively and collide enabling the creation of further constraints as the microdomains coalesce. After a series of coalescence steps the final protein fold is attained [Karplus and Weaver, 1994]. The nucleation-condensation model differs in that the secondary and tertiary structures are formed simultaneously. A nucleus with marginal stability is assembled, which becomes a template around which further structural components are able to condense [Nölting and Agard, 2008].

### 1.1.4 Folding cores

The gap in knowledge of the exact folding process causes problems when defining the folding core of a protein. Intuitively the folding core should refer to the residues which “collapse early during folding” [Woodward, 1993], yet this set of residues is difficult to ascertain. One way of defining the folding core experimentally is through $\phi$-value analysis [Fersht et al., 1992]. The analysis is applied to proteins which fold in a one-step process, whereby proteins in an unfolded state become fully folded via a transition state. The transition state is by definition transient and partially unstructured making it difficult or impossible to capture using standard

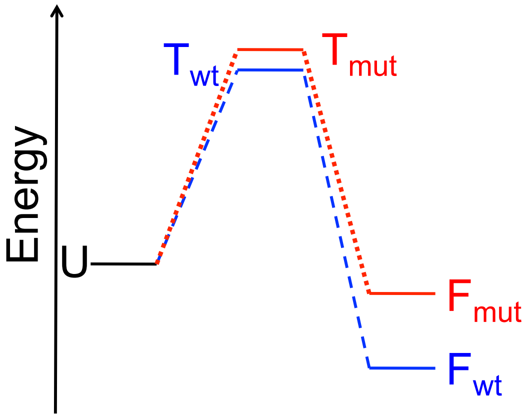


Figure 1.3: Energy diagram of the one-step folding process. A wild type (wt) and mutant (mut) protein in an unfolded state (U) become fully folded (F) via a transition state (T). The ratio of differences in energy levels is used to calculate the $\phi$-value of the mutated residue.

structural techniques such as NMR and X-ray crystallography. An energy diagram for this process is shown in Figure 1.3. The folding kinetics and the conformational stability of the wild type protein and a point mutant are compared in order to gain information about the role of the mutated residue in the transition state. The energy difference between the transition states of the wild type and mutant is compared with the energy difference between the two native states. The $\phi$-value is the ratio between these two energy differences. Values of $\phi$ lie between 0 and 1. When $\phi \simeq 0$, the difference in energy between the transition states is much smaller than the difference in energy between the native states. This means that mutating the residue does not change the structure of the transition state. When $\phi \simeq 1$, the energy difference between the transition states is approximately the same as between the native states, suggesting that the mutated residue is folded during the transition state (its mutation destabilises this state) [Oliveberg and Fersht, 1996]. The folding core can then be identified as those residues for which $\phi \simeq 1$. In Figure 1.3, $\phi \simeq \frac{1}{3}$. The major assumptions of using this method are that the mutation does not change the folding pathway or the structure of either the unfolded or folded protein [Fersht et al., 1992; Itzhaki et al., 1995]. The $\phi$-value analysis focuses on the folding process. An alternative way of considering the folding core of the protein is to study the folded structure and its dynamics. HDX experiments can be used for this purpose, since the results of these experiments can identify which residues are typically buried within a protein. By combining HDX with refolding experiments, the residues which adopt a folded conformation early in the folding process can be determined [Woodward, 1993]. In the cases of barnase and chymotrypsin inhibitor 2, it has been shown that residues identified this way in HDX have high $\phi$-values [Li and Woodward, 1999]. Experimental work focussing on the folding process and the folded structure therefore give different definitions of the folding core which nevertheless overlap in terms of the residues identified. We will introduce HDX experiments in Section 1.6.4 and discuss them, along with the HDX folding core, in Chapters 4 and 6.

## 1.2 HIV-1 protease

HIV-1 protease plays a key role in the life cycle of type 1 Human Immunodeficiency Virus (HIV-1) and as such is an important drug target. Its function is to cleave long peptide chains at specific sites, generating proteins which subsequently assemble to form a new virus particle (a virion). There is a wealth of crystal structures of HIV-1 protease available in the PDB, with a wide range of inhibitors bound. This emphasises the importance of the protein as a drug target and also the difficulty of designing successful inhibitors. A high mutation rate leads to rapid immunity to antiretroviral therapies and so treatments continually need to be modified [Gilks et al., 2006; Levy, 2009; Branson and Stekler, 2012]. The computational tools of rigidity analysis (see Section 1.4) are rapid to implement and offer us the opportunity to provide an analytical overview of many of these structures. The protease enzyme is a symmetrical homodimer with 198 residues in total. Figure 1.1 shows the structure of HIV-1 protease. In Chapter 2 we study the influence of different inhibitors on the rigidity of the protein.

### 1.2.1 HIV-1 and AIDS

HIV-1 is a pandemic virus, infection with which leads to Acquired Immunodeficiency Syndrome (AIDS). There are two types of HIV. The more virulent and infective type is HIV-1 whereas HIV-2 is largely confined to Africa [Sharp and Hahn, 2011]. The virus infects and kills CD4$^+$ T-lymphocytes [Brenchley et al., 2004] resulting in a weakened or non-functioning immune system. The infected individual then becomes susceptible to infection from other maladies. The symptoms of initial acute infection, where huge numbers of infected CD4$^+$ T-cells are killed by activated CD8$^+$ T-cells can arise within days of infection [Kahn and Walker, 1998; Branson and Stekler, 2012]. Upon sufficient reduction in number of CD4$^+$ T-cells, cell-mediated immunity is lost, resulting in AIDS. Between acute infection and the onset of AIDS there is a latency period with few symptoms, or sometimes none at all, which may last a few

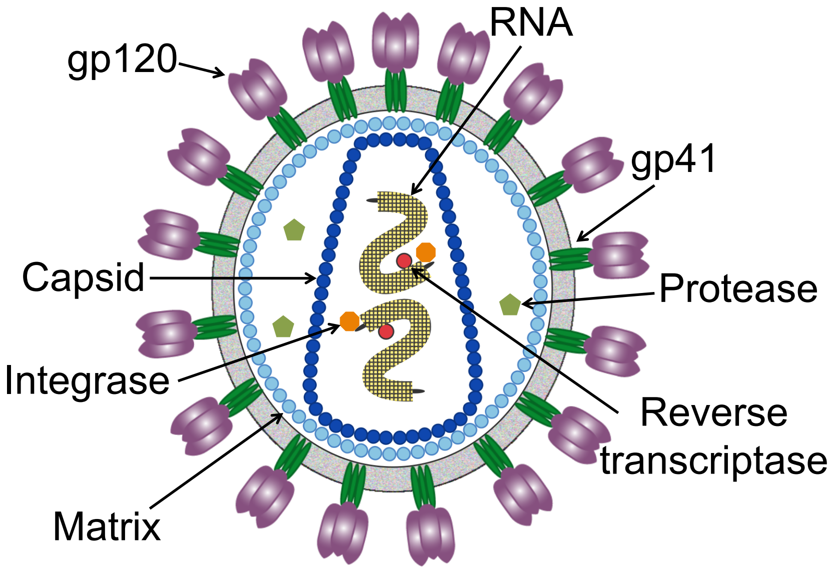


Figure 1.4: A cross-sectional representation of the spherical HIV-1 virion. Glycoprotein spikes formed from gp41 and gp120 protrude from the virion. The matrix lies below the viral envelope, and contains the bullet-shaped viral capsid. Genetic information is contained in the viral RNA. The three HIV-1 enzymes are the reverse transcriptase, integrase and protease enzymes.

weeks but can last for years. After the symptoms of AIDS are observed, then death is highly likely within the following year if no antiretroviral treatment is undertaken [Morgan et al., 2002]. With the help of antiretroviral therapies, life expectancy after the onset of AIDS can be up to five times longer than without therapy [Schneider et al., 2005]. Although treatments can suppress the virus for long time periods in the majority of individuals, viral reservoirs cannot currently be eradicated [Chun and Fauci, 2012].

### 1.2.2 The HIV-1 virion

The HIV-1 virion is depicted in Figure 1.4. The virion is a spiky ball with a diameter of approximately 0.1 $\mu$m. The spikes are the glycoproteins gp120 and gp41, and they protrude from the viral envelope. The matrix protein (MA) forms the matrix which lies below the envelope and contains the viral capsid. MA contains membrane-binding domains and also plays a role in the assembly of the virion [Freed, 1998]. The viral capsid protein (CA) forms the capsid which is the outer layer of the core, containing the three retroviral enzymes as well the two RNA strands which form the genetic material of the virus [Freed, 1998].

HIV-1 is a retrovirus, meaning that its genetic material is RNA which needs to be reverse transcribed into DNA before viral proteins can be produced in the host cell. The genetic material of HIV-1 encodes for some proteins which are common to all retroviruses [Frankel and Young, 1998]. Initially these are synthesised as polyproteins which are subsequently cleaved by viral or host-cell proteases. The Gag polyprotein is cleaved to form the core and structural proteins, including MA and CA [Bosco and Kern, 2004]. The three retroviral enzymes, reverse transcriptase, integrase and protease, are cleaved from the Pol polyprotein [Frankel and Young, 1998]. The Env polyprotein is the precursor for gp120 and gp41 which form the spikes on the outer layer of the virus enabling it to bind to its target cells. HIV-1 protease cleaves both Gag and Pol polyproteins [Darke et al., 1998]; the host-cell protease furin cleaves the Env polyprotein [Hallenberger et al., 1992]. There are further accessory proteins encoded for by the viral genome, which generates fifteen proteins in total [Frankel and Young, 1998].

### 1.2.3 The HIV-1 life-cycle

The HIV lifecycle is shown in Figure 1.5. Here, we briefly discuss the life-cycle in order to demonstrate the importance of HIV-1 protease and the centrality of its role. More comprehensive discussions are available elsewhere [Frankel and Young, 1998; Freed, 1998; Levy, 2009]. Initially, the virion binds to the receptors of the $CD4^+$ T-lymphocytes before fusing with the membrane and injecting the core contents into the host cell. The virus binds to a CD4 receptor *and* one of the CCR5 or CXCR4 co-receptors in order to fuse with the host. The single-stranded RNA is reverse-transcribed HIV-1 reverse transcriptase, resulting in a section of double stranded viral DNA being formed [Himmel et al., 2005]. This takes place in the cytoplasm and the DNA is subsequently transported into the host nucleus before HIV-1 integrase splices the DNA of the host and incorporates the viral DNA. At this stage there may be a latency period during which the viral DNA is dormant within the host cell. In this case, the genetic information may be replicated as the cell divides, but the DNA is not transcribed. Alternatively, the lifecycle may progress resulting in new HIV-1 virions which may in turn infect other cells. For this to happen, cellular transcription factors such as NF-$\kappa$B need to be present at high concentrations – something which occurs when the cells are activated [Hiscott et al., 2001]. When the lifecycle continues viral mRNA is translated forming the polyproteins discussed above. It is the job of HIV-1 protease to cleave the Gag and Pol polyprotein chains. The individual proteins assemble together and bud from the host cell, taking part

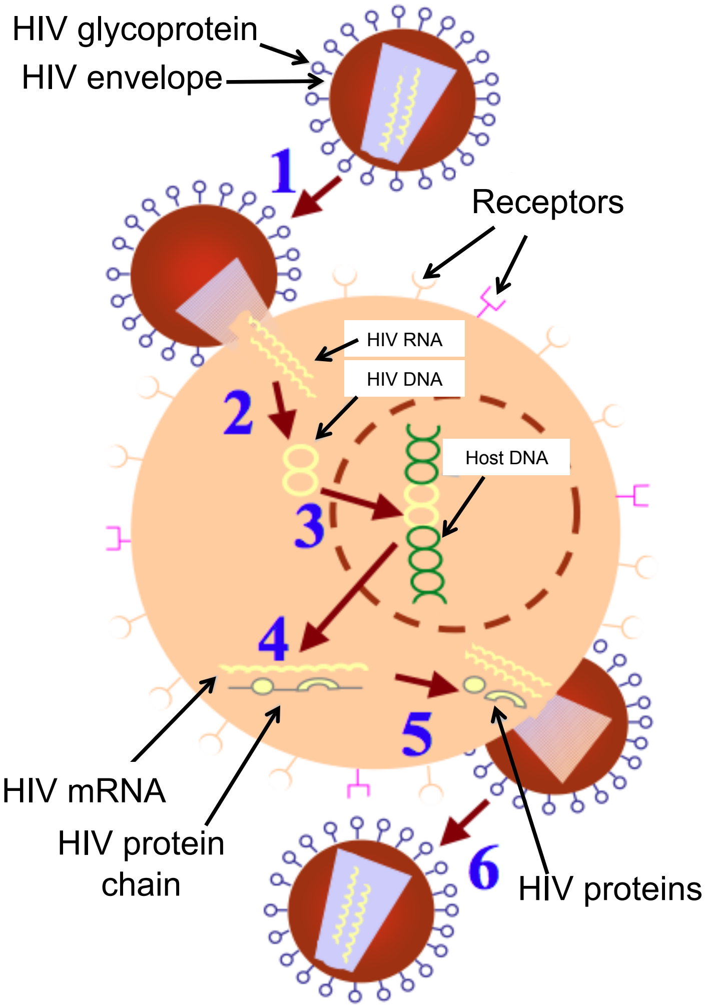

HIV glycoprotein
HIV envelope
1
Receptors
HIV RNA
HIV DNA
2
Host DNA
3
4
5
HIV mRNA
HIV protein chain
6
HIV proteins

Figure 1.5: The lifecycle of the virus HIV. Figure taken from http://www.thebody.com/content/art40989.html. (1) The virus binds to the receptor and releases its viral RNA into the host cell. (2) Viral RNA is reverse-transcribed into double stranded DNA by the enzyme HIV-1 reverse transcriptase. (3) The viral DNA is integrated into the host genome by the HIV-1 integrase enzyme. (4) The viral DNA may lay dormant in the host cell for many years before being transcribed into mRNA by the machinery of the host cell. This mRNA is then translated by the host cell, forming HIV-1 polyprotein chains. (5) HIV-1 protease cleaves some of these HIV-1 protein chains into their constituent parts so that they can function as viral proteins and assemble to form a new virion. (6) The virion buds from the host cell, taking with it part the cellular envelope of the host.

of the cellular envelope with them as they do so in order to form the membrane of the new virion. This then buds from the host cell and the life-cycle continues with infection of another CD4$^+$ T-lymphocyte.

### 1.2.4 Antiretroviral therapies

The exact pathogenesis of the virus is unclear, but there are various mechanisms through which HIV-1 can bring about CD4$^+$ T-cell depletion [Alimonti et al., 2003]. HIV-1 can kill infected cells directly, but can also increase rates of apoptosis in infected cells. Infected CD4$^+$ T-cells can also be killed by CD8 cytotoxic lymphocytes that recognise infected cells [Levy, 2009]. As a result of its ubiquity and lethality, enormous efforts have been made within the scientific community to quell the virus in a range of different ways. One such method at the molecular level is to disrupt the mechanism by which the virus is able to proliferate by preventing one of the three viral enzymes from functioning [Sayasith et al., 2001; Arnold et al., 1996; Hornak and Simmerling, 2007]. The twelve other proteins encoded by the viral enzyme can also serve as drug targets [Tang et al., 2003]. For example, other authors have focussed on the CA [Barrera et al., 2007], HIV-1 viral protein R (Vpr) [Rouzic and Benichou, 2005] and the envelope glycoproteins gp120 and gp41 [Chan et al., 1997; Tan and Rader, 2009; Harvey et al., 2011]. In total, more than twenty such drugs have been licensed for clinical use [Cane, 2009]. HIV-1 has a high mutation rate, resulting in rapid drug resistance [Gilks et al., 2006]. In order to counteract this and maintain effective therapies for longer, antiretroviral therapies are typically administered in the form of a drug cocktail, whereby multiple drugs are given in combination [Cane, 2009]. Such combination therapies reduce the viral load more effectively and can delay the onset of AIDS as a result. In total, ten inhibitors of HIV-1 protease have been approved by the US Food and Drug Administration (FDA) . As well as combining drugs which target different viral proteins, pairs of protease inhibitors have also been used in combination [Hicks et al., 2006; Cane, 2009]. In Chapter 2, we discuss these inhibitors in more detail. Using the FIRST software we focus on their impact on the rigidity of the enzyme.

## 1.3 Cyclophilin A

Cyclophilin A is a multifunctional protein which acts as an enzyme to catalyse a step during protein folding in addition to performing other roles when binding to different molecules such as the immunosuppressant drug cyclosporin A (CsA) and the HIV-1 capsid protein CA. CypA has been studied extensively using X-ray crystallography, with more than 50 structures of CypA (in complex or otherwise) available in the PDB. In Chapter 3, we present a rigidity analysis study of these structures. NMR has been used to solve the structure of CypA [Ottiger et al., 1997] as well as the CypA-CsA complex [Neri et al., 1991; Spitzfaden et al., 1994]. We perform HDX experiments on CypA in the presence and absence of CsA in order to gain insight into the impact of ligand binding on the protein. These results are presented in Chapter 6. We want to use computational methods to try to predict the results of HDX experiments and the effect of ligand binding on these results. When selecting CypA, we selected a protein which satisfied several criteria. There needed to be a wealth of structural data available so that we could adapt the approach of large-scale rigidity analysis which we apply to HIV-1 protease in Chapter 2. Ideally, we required structures which had been crystallised with a variety of ligands bound, as well as some unbound structures. The size of the protein was also important, since smaller proteins (100 – 500 residues) are more amenable to NMR spectroscopy. Although our computational methods are rapid, a smaller protein would also permit faster interpretation of computational results. Other considerations concerned practicality: it would be desirable to have a plasmid available to express the protein, and peptide ligands were preferable to complex carbohydrates for reasons of cost. The published study of HDX experiments on unbound CypA [Shi et al., 2006] also encouraged our selection, since we would be able to compare these data with results from our computational work, which was conducted alongside our own HDX experiments. Finally, and in many ways predominantly, we wished to study a protein which was known to interact with ligands in a biologically interesting manner. Other candidate proteins we considered included included protein kinases A [Pearce et al., 2010], and B [Brazil and Hemmings, 2001; Fayard et al., 2005], adenylate kinase [Pisliakov et al., 2009], c-SRC [Martin, 2001] and maltose binding protein [Bucher et al., 2011]. The satisfaction of the above criteria make CypA an ideal test-bed for our computational studies and experimental work. The importance of the protein, with its many roles and substrates motivate further study into the effects of ligand binding on its flexibility and mobility.

### 1.3.1 The peptidyl-prolyl *cis-trans* isomerase family

Peptidyl-prolyl *cis-trans* isomerases (PPIs) are enzymes which catalyse the *cis-trans* isomerisation of a protein chain at the Xaa-Pro bond, where Xaa is any amino acid. A peptide bond can be arranged with the functional groups of the two amino acids involved in the bond on opposing sides of the chain, *trans*, or same side, *cis*. These two arrangements are illustrated in Figure 1.6 for both the general case and for the

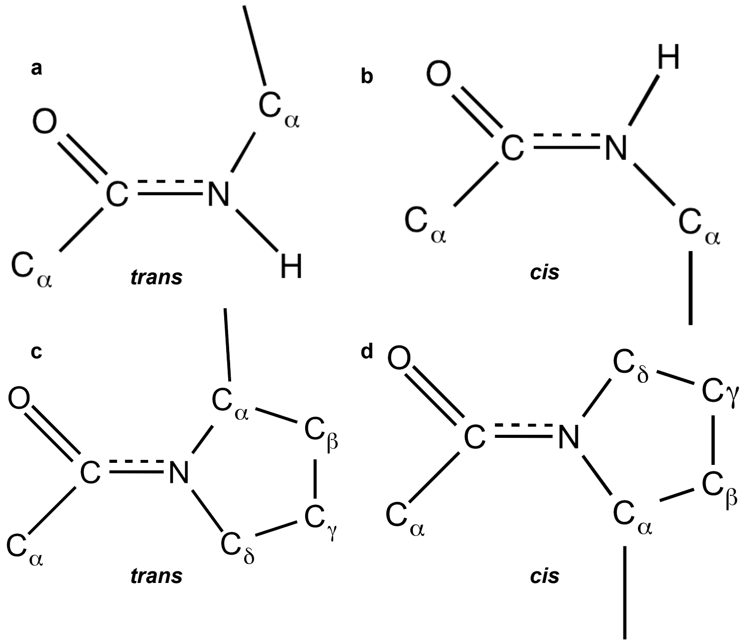


Figure 1.6: The *trans* (a) and *cis* (b) arrangements of a peptide bond. The *trans* (c) and *cis* (d) arrangements for the special case of the Xaa-Pro peptide bond, where Xaa is any amino acid. The partial double bond character of the peptide bond is indicated by a dotted line above the bond. This Figure was drawn using ChemDraw version 13.0.

special case involving proline. The *trans* conformation is generally favoured over the *cis* conformation due to the extra space afforded to the functional groups when they lie on opposite sides of the peptide chain. In the case of an Xaa-Pro bond, the symmetry between the $C_\alpha$ and $C_\delta$ atoms in the functional ring of proline means that the *cis* and *trans* conformations are similarly favourable and the *cis* form is more common than in other peptide bonds. This may lead to an incorrect arrangement at the site of an Xaa-Pro bond. Since the peptide bond has partial double bond character, the energy barrier opposing rotational rearrangement around this bond is large. It is the role of PPIs to lower the activation energy and hence permit faster protein folding [Kern et al., 1995; Lilie et al., 1993].

The principal members of the PPI family are the cyclophilins and the FK506 binding proteins (FKBPs) [Lilie et al., 1993]. Cyclophilin A (CypA) is the prototypical cyclophilin and also the most abundant [Fischer et al., 2010]. CypA is found in almost all tissues in prokaryotes and eukaryotes; human CypA is found in all organs [Satoh et al., 2010b]. Cyclophilins B, C and D are less abundant [Satoh et al., 2010a], with human CypB and murine CypC localised to to the endoplasmic reticulum [Price et al., 1991; Schneider et al., 1994] and CypD to the mitochondria [Bergsma et al., 1991].

### 1.3.2 Cyclosporin A

In addition to its role as a PPI, CypA is also involved in the function of the immunosuppressant drug CsA. Most commonly used to suppress organ rejection following a transplant, CsA has also been administered to treat ulcerative colitis, cardiac disease and a number of autoimmune diseases [Nussenblatt and Palestine, 1986; Lichtiger et al., 1994; Mott et al., 2004]. The compound, a 1.2 kDa cyclic peptide consisting of eleven amino acid residues, was initially isolated from a fungus found in a soil sample. After successfully passing clinical trials it became a front-line drug for combatting organ rejection [Stähelin, 1996]. CypA was identified as the main target for CsA. This interaction gives the cyclophilins their name [Handschumacher et al., 1984; Harding et al., 1986]. The enzyme peptidyl-prolyl *cis-trans* isomerase and CypA were initially presumed to be distinct proteins, but were later discovered to be the same [Fischer et al., 1989]. The cyclophilins all share a CsA-binding domain. There are eight single-domain (including CypA) and ten multi-domain cyclophilins [Fischer et al., 2010]. CypA has a strong binding affinity for CsA, with $K_d = 20$ $\mu$M [Handschumacher et al., 1984], twice that of CypC but ten times lower than CypB [Schneider et al., 1994]. CsA inhibits the T cell activator calcineurin (CN), a phosphatase [Zydowsky et al., 1992; Liu et al., 2006]. In fact, it is the CypA-CsA complex which blocks the activity of CN by forming a CypA-CsA-CN complex. Neither CypA nor CsA bind to CN independently [Luban et al., 1993]. The binding of CypA to CsA has been observed by enhanced fluorescence upon binding due to the change in environment of the Trp121 residue [Husi and Zurini, 1994; Gastmans et al., 1999]. In Chapter 5 we use this result to examine the binding of CsA to purified CypA.

### 1.3.3 HIV-1 capsid protein

CypA has also been shown to play a role in the life-cycle of HIV-1, binding to the CA domain of the HIV-1 Gag polyprotein [Luban et al., 1993; Bosco and Kern, 2004; Wang and Heitman, 2005]. Once Gag has been cleaved, CypA therefore forms a complex with CA. Due to this interaction, CypA has been shown to be essential for efficient HIV-1 replication and therefore its virulence [Bosco and Kern, 2004; Eisenmesser et al., 2005]. Figure 1.7, taken from [Cullen, 2003] shows that the CypA-CA complex prevents the recognition of HIV-1 virions by restriction factors in human cells, leading to productive infection [Towers et al., 2003]. An abortive infection pathway can be accessed by ensuring that the CypA-CA complex is not formed. In owl monkey cells, the CypA-CA complex itself is targeted by restriction factors and productive infection occurs when the complex is not formed [Cullen, 2003]. CypA also binds to the capsid protein of the human papillomavirus [Bienkowska-Haba et al., 2009], and may also regulate the replication of the vesicular stomatitis virus [Bose et al., 2003] and the cytomegalovirus [Kawasaki et al., 2007]. CypA acts catalytically to support RNA replication in the hepatitis C virus [Hanoulle et al., 2009; Kaul et al., 2009; Fischer et al., 2010] and has been identified as having a physiological and pathological role in cardiovasular diseases [Satoh et al., 2010b]. The strong binding affinity of CsA to CypA means that CsA can be used to prevent the activity of CypA in the lifecycle of these viruses. Due to the CypA-CsA complex blocking CN and therefore acting as an immunosuppressant, there is ongoing research with the aim of designing CsA analogs that form a complex with CypA without immunosuppressant properties [Fischer et al., 2010].

## 1.4 Rigidity analysis

Protein rigidity analysis is a computational method which rapidly identifies rigid and flexible regions in a protein crystal structure [Rader et al., 1999; Jacobs et al., 2001]. The structure is considered as a molecular framework in which bond lengths and angles are considered fixed while dihedral angles are permitted to vary. Degrees of freedom of the atoms are matched against the constraints due to bonding using an integer algorithm, the pebble game. Covalent bonds, polar interactions (including hydrogen bonds and salt bridges), and hydrophobic tethers can all be included as bonding constraints. The output of the algorithm is a division of the structure into rigid clusters and flexible regions. Rigidity analysis is rapid, informative and complementary to more computationally expensive methods such as MD [Gohlke et al., 2004]. In a rigidity analysis, it is flexibility and therefore the potential for motion which is of central importance. This is akin to "identifying the hinges on a door, without moving the door" [Wells et al., 2005]. Rigidity analysis on a single protein structure may be carried out on a standard desktop computer on a timeframe of seconds. This is on the order of $10^6$ times faster than MD simulations [Jacobs et al., 2001]. The short computational time means that the techniques can be applied to a large number of protein structures. When studying a particular protein, a more comprehensive analysis of the available structural data is therefore possible as an

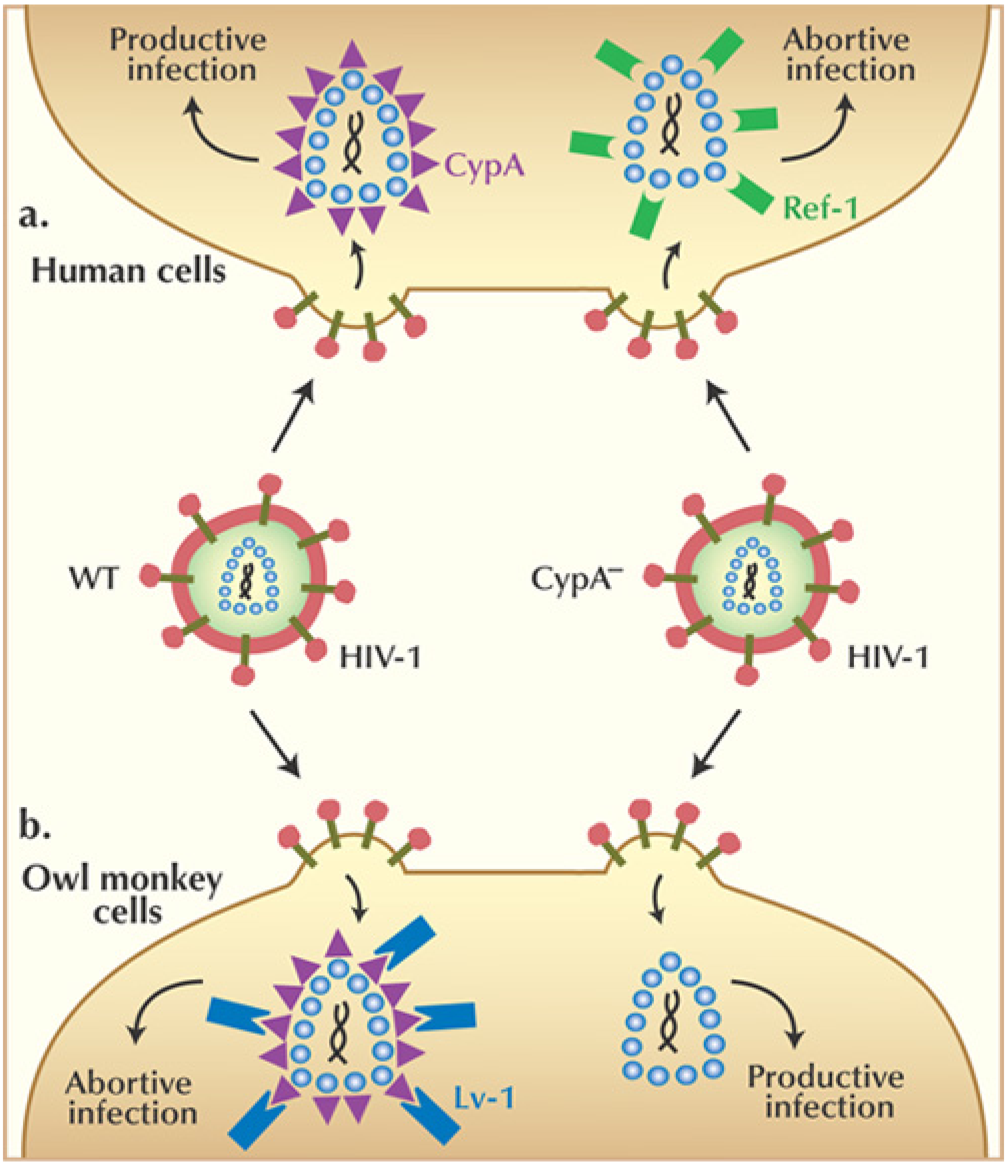

Productive infection
CypA
Abortive infection
Ref-1
a.
Human cells
WT
HIV-1
CypA⁻
HIV-1
b.
Owl monkey cells
Abortive infection
Lv-1
Productive infection

Figure 1.7: The CypA-CA complex leads to productive infection of HIV-1 in human cells. In owl monkey cells, the complex is targeted by the restriction factor Lv-1, resulting in abortive infection. When the complex is not formed, the capsid is targeted by the restriction factor Ref-1 in human cells and infection is abortive. Contrastingly in owl monkey cells when the complex is not formed, Lv-1 is unable to bind and the HIV-1 infection is productive. Figure from [Cullen, 2003].

alternative to investigating a single structure in detail. In Chapter 2 for example, we study 212 structures of HIV-1 protease, which are listed in Appendix B.3, Table B.2.

### 1.4.1 FIRST

We study the rigidity of proteins with the software FIRST (Floppy Inclusions and Rigid Substructure Topography). Using an X-ray crystal structure as an input, FIRST rapidly categorises sections of the protein as being rigid or flexible. The 'pebble game' algorithm employed by the software is an integer counting algorithm which scales linearly with system size [Jacobs and Hendrickson, 1997]. As a result there is effectively no size limit to the proteins or protein systems which can be investigated. FIRST has been used to study the assembly of the viral capsid of the cowpea chlorotic mottle virus and both of the ribosomal subunits [Wang et al., 2004]. A range of proteins including rhodopsin [Rader et al., 2004] ubiquitin [Jacobs and Dallakyan, 2005], RNase H [Livesay and Jacobs, 2006], adenylate kinase [Jolley et al., 2008] and rubredoxin [Rader, 2010] have also been subjected to rigidity analysis using FIRST. Rigidity analysis has informed studies on the similarities between evolutionarily distinct proteins [Thorpe et al., 2000] as well as on the mechanisms of allostery [Rader and Brown, 2011] and thermostability [Radestock and Gohlke, 2008].

FIRST can be used to inform MD [Gohlke et al., 2004; Fuxreiter et al., 2005] or coarse-grained simulations [Wells et al., 2005, 2009]. The information gained from rigidity analysis can be fed into the motion software so that the dynamics can be targeted around the flexible regions of the protein. FIRST has been used in this way to facilitate simulations using Gaussian network models [Rader et al., 2004; Rader and Bahar, 2004; Wang et al., 2004] and normal mode analysis models with FRODA [Gohlke and Thorpe, 2006; **?**]. We apply this approach to CypA in Chapter 4 by running coarse-grained simulations based on normal mode analysis with FRODA. The results from FIRST have also been compared with experimental results such as hydrogen-deuterium exchange NMR [Hespenheide et al., 2002; Rader and Bahar, 2004; Zavodszky et al., 2004]. This type of comparison is discussed in more detail and expanded upon in Chapter 4. Fluorescence spectroscopy, circular dichroism spectroscopy [Rader, 2010] and thermal denaturation experiments [Radestock and Gohlke, 2008; Rader, 2010] have also been utilised alongside rigidity analysis.

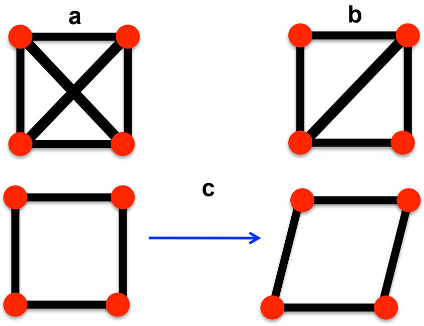


Figure 1.8: A simple example of overconstrained (a), isostatic (b) and underconstrained (c) structures. The structures each consist of four atoms shown as red circles, joined by bonds which are represented as black lines. Transitions from (a) to (b) and from (b) to (c) are made by the removal of a single bond. The flexible structure (c) can be perturbed as indicated by the blue arrow.

### 1.4.2 The pebble game algortihm

The pebble game algorithm of FIRST matches the degrees of freedom of the atoms in a structure with constraints and hence categorises the atoms as being either flexible or rigid. For more detail on the development of the algorithm and how it works, the reader is referred to previous publications [Jacobs and Thorpe, 1995; Jacobs and Hendrickson, 1997; Jacobs, 1998; Jacobs et al., 1999; Thorpe et al., 2000; Jacobs et al., 2001; Hespenheide et al., 2004].

The atoms of the protein are viewed as the nodes on a graph connected by edges which are the bonds between the atoms. These bonds are inflexible rods which are either present or absent — this is a balls and sticks rather than a balls and springs representation of the protein. The set of bonds which determine the edges of the graph in FIRST include covalent bonds, hydrogen bonds and hydrophobic interactions. The location and number of bonds in the network determine the distribution of rigidity within the structure. Based on the bond network, the structure is divided the into overconstrained, isostatic or underconstrained regions. Examples of each of these types of region are given in Figure 1.8. An underconstrained region is flexible, in the sense that dihedral angles can vary and so atoms can move without violating distance constraints [Thorpe et al., 2001]. An isostatic region is rigid but with an exact balance of constraints and degrees of freedom; the removal of any constraint would make the region flexible. Overconstrained regions are rigid with constraints to spare, and so it is possible to remove a constraint without altering the rigidity of the region. Clusters of atoms form mutually rigid substructures, known as rigid clusters. A division of a molecule into rigid clusters separated by flexible regions is called a rigid cluster decomposition (RCD).

### 1.4.3 Hydrogen bonds and $E_{\text{cut}}$

The RCD clearly depends on the bonds in the structure: the rigidity of the protein is affected by the number and location of bonds. In FIRST the strong bonding forces such as covalent bonds and hydrophobic interactions are always included. Long-range electrostatic interactions and van der Waals forces do not contribute to the distance constraints between atoms because they are generally weak at this level and not highly directional [Jacobs and Thorpe, 1996]. These forces are excluded from the bond network. Hydrogen bonds (HBs) and salt bridges are incorporated into the constraint network on a selective basis. In FIRST, the strength of each HB in the structure, measured in kcal/mol, is calculated as a function of the geometry of the donor, hydrogen and acceptor atoms using the Mayo potential [Dahiyat et al., 1997]. The potential is highly distance and angle dependent [Wells et al., 2009]. The principal variable parameter in FIRST is the HB energy cutoff, $E_{\text{cut}}$. A bond is included in the network if its bond energy is more negative than $E_{\text{cut}}$, otherwise it is excluded. The choice of $E_{\text{cut}}$ determines the bond network and therefore affects the results of rigidity analysis. It is difficult to define a single $E_{\text{cut}}$ which can be used to simulate any given protein in native conditions. In previous studies using FIRST, $E_{\text{cut}}$ has been varied considerably. The default option is $-1.0$ kcal/mol and this value has been used extensively [Jolley et al., 2006; Macchiarulo et al., 2007; Rader and Brown, 2011]. Other values which have been used include $-0.6$ kcal/mol [Gohlke et al., 2004] and $-0.3$ kcal/mol [Tan and Rader, 2009]. A more robust approach is to repeat the calculations using different values of $E_{\text{cut}}$ in order to check the dependence of the results on the choice of $E_{\text{cut}}$ [Jacobs et al., 2001; Gohlke et al., 2004]. Alternatively, one can select the appropriate $E_{\text{cut}}$ by examining a set of values and determining which yields the RCD which best reflects the available experimental data [Hespenheide et al., 2004; Zavodszky et al., 2004; Jolley et al., 2008]. This approach yields $E_{\text{cut}}$ values as diverse as $-0.35$ kcal/mol [Hespenheide et al., 2004] and $-2.3$ kcal/mol [Jolley et al., 2008]. The selection of an appropriate $E_{\text{cut}}$ value is considered in more detail in Chapters 2, 3 and 4.

### 1.4.4 Rigidity dilution

A rigidity dilution (RD) involves removing the hydrogen bonds in order of strength, and running rigidity analysis after each change to the bond network. The RCD of the protein changes with the bond network, becoming more flexible as bonds are removed. This procedure is a systematic lowering of $E_{\text{cut}}$, and the pattern of rigidity loss can be used to gain insight into structural and functional properties of the protein [Rader et al., 1999; Heal et al., 2012]. General insights into the rigidity of proteins can be obtained by studying the rigidity dilution of a diverse collection of structures. It has been observed that proteins which are predominantly $\beta$-sheet lose rigidity abruptly (first order rigidity loss) and those which are mostly $\alpha$-helix lose their rigidity in a more gradual manner (second order rigidity loss) [Wells et al., 2009].

The result of rigidity analysis in FIRST can be processed to suit the particular investigation. The measure $X_{\text{LRC}}$, the size of the largest rigid cluster (LRC) relative to the whole protein, has been introduced to study the rigidity of the protein at the molecular level [Tan and Rader, 2009; Rader and Brown, 2011]. We note that $X_{\text{LRC}}$ is the same as the measure $\text{P}_\infty$ (type 2) discussed elsewhere [Pfleger et al., 2013a]. The propensity for each residue to be in the largest rigid cluster can also be evaluated using data from RD simulation [Rader et al., 2004; Rader, 2010; Pfleger et al., 2013a]. The measure $f_k$ [Wells et al., 2009] extends $X_{\text{LRC}}$ to incorporate the $k$ largest rigid clusters (so that $f_1 = X_{\text{LRC}}$). The similarly defined $Z_{\text{LRC}}$ is used to analyse the local rigidity within a structure, for example surrounding the binding site of a protein [Rader and Brown, 2011]. In Chapter 3 we discuss these measures further and apply them to CypA.

### 1.4.5 Hydrophobic interactions

Hydrophobic tethers (HPs) are indirect, entropy-driven interactions [Folch et al., 2008] and as such are different from direct interactions such as HBs. HPs are formed as a result of hydrophobic groups being repulsed by the surrounding polar environment. They are repulsion-driven therefore indirect, and the resulting organisation of hydrophobic groups away from solvent exposure results in a lower entropy. Their action is thought to contribute significantly to protein folding [Dill, 1990]. In FIRST, HPs are modelled as flexible constraints, restricting the separation distance between atoms involved in the interaction, but not the angle between them. In this way, the interacting atoms are permitted to slip relative to each other [Hespenheide et al., 2002; Gohlke et al., 2004]. The HPs are less specific than the HBs, with each HP removing two degrees of freedom instead of three [Rader et al., 1999]. HPs between carbon or sulfur atoms are typically included if the distance between these atoms is less than the sum of their van der Waals radii plus a distance cutoff, $D_{\text{HP}}$. The van der Waals radii of carbon and sulfur are 1.7 Å and 1.8 Å respectively, and $D_{\text{HP}}$

is typically set to 0.25 Å [Hespenheide et al., 2002; Gohlke et al., 2004; Radestock and Gohlke, 2008].

HPs have not always been included in FIRST simulations, and their effect on modelling the flexibility and mobility of proteins is still not fully understood. In early papers using FIRST, HPs were not included in the bond network [Thorpe et al., 2000; Jacobs et al., 2001], whereas more recently it is usual practice to include HPs and maintain them throughout [Zavodszky et al., 2004; Radestock and Gohlke, 2008]. The reasons given for not including HPs are that they are "fairly slippery and nonspecific" [Thorpe et al., 2000] or "generally weak" and "not highly directional" [Jacobs et al., 2001]. Motivation for their inclusion is due to their role in the folding and stability of proteins [Dill, 1990] as well as the hydrophobic effect whereby the strength of HPs appears to increase with temperature [Tanford, 1980; Schellman, 1997], suggesting that they may become more important and influential as HBs are removed during a rigidity dilution. Since there are typically as many as half the number of HPs as there are HBs [Gohlke et al., 2004], the contribution of these interactions is at least worthy of our attention. Aside from these two opposing approaches – including or discounting all of the HPs – there has been little discussion of alternatives. In one paper the authors consider the importance of HPs and reevaluate their simulations with and without HPs, concluding that finding the appropriate balance between HPs and HBs is "crucial for an accurate representation of the flexibility characteristics of proteins by FIRST" [Gohlke et al., 2004]. More recently a system was developed to account for the enhancement of HPs with temperature by linearly increasing $D_{\mathrm{HC}}$ from 0.25 Å at 300 K to 0.40 Å at 420 K [Rathi et al., 2012; Pfleger et al., 2013b]. This method yields better prediction of the thermostability of proteins in (hyper)thermophilic organisms, but also requires defining an explicit relationship between $E_{\mathrm{cut}}$ and temperature. In Chapter 4 we use different settings for HBs and HPs in order to try to predict the results of HDX experiments.

## 1.5 Simulating protein motion

The study of protein motion is of fundamental importance in structural biology due to its intimate link with protein function. The most prevalent method for modelling protein motion computationally is through MD simulations [Gohlke and Thorpe, 2006]. An MD force field describes the interactions between the atoms of the protein, and by numerically solving Newton's equations of motion, structural fluctuations are calculated over time [Adcock and McCammon, 2006]. There are a number of widely used programs to implement the simulations [Karplus and McCammon, 2002]. The detailed atomic trajectories derived make MD simulations easier to probe than experimental data [Karplus and Petsko, 1990]. Since proteins are made up of many thousands of atoms, the large number of equations which need to be solved ensure that MD simulations are computationally intensive. Accurate trajectories require small timesteps on the order of 1 fs [Wells et al., 2005]. Some of the most biologically relevant motions take place on the ms – s timescale currently inaccessible to MD. In order to simulate larger amplitude protein motion, methods of coarse graining can be applied [Clementi, 2008]. In such simulations, atomistic detail is neglected and instead conglomerates form the unit particles. As a result the force fields are simplified, the number of calculations is reduced and longer simulations are possible. Geometric simulations, such as those applied with FRODA, involve the exploration of a conformational space defined by the structure of the protein and the constraints placed on its atoms by the bond network. FIRST can be used to establish the bond network and the rigidity it confers on the protein. The application of normal mode analysis to a coarse-grained elastic network model (ENM) [Tirion, 1996] can be used to generate vectors along which to direct such an exploration. These concepts are introduced in the following sections, and described further elsewhere [Wells et al., 2005; Jimenez-Roldan et al., 2012].

### 1.5.1 FRODA

The constrained geometric simulation software FRODA (framework rigidity optimized dynamic algorithm) rapidly simulates protein motion [Wells et al., 2005]. FRODA is a module within FIRST. The framework approach searches for conformers which satisfy the constraints of the network defined by FIRST. Random motions are used to generate new conformations, and the re-application of constraints within the network ensures that the conformer becomes valid. Trajectories are formed from continuous pathways between sets of acceptable conformers. The progression of motion is not measured using time, as in MD, but in distances calculated using the newly generated conformers and a reference conformer, typically the initial input structure [Wells et al., 2005]. The algorithm scales linearly with system size, and large-amplitude motions in proteins with hundreds or thousands of residues can be rapidly explored using a desktop computer on a timescale of minutes [Jimenez-Roldan et al., 2012]. The speed of the algorithm comes with the price of sacrificed detail. There is no information on the energy of different conformational states, and the lack of time in a FRODA simulation makes direct quantitative comparison with experimental data problematic [Jolley et al., 2006]. Distance-based comparisons between FRODA simulations and NMR experiments have been made [Wells et al., 2005; Gohlke and Thorpe, 2006]. Details of the algorithm can be found elsewhere [Wells et al., 2005]. FRODA has been used to investigate protein-protein docking problems [Jolley et al., 2006] as well as substrate recognition and conformational changes [Macchiarulo et al., 2007].

### 1.5.2 Normal mode analysis

In addition to using random motions, it is possible to direct the motion so that particular conformational changes can be modelled. Using normal mode analysis (NMA) [Diamond, 1990], we direct the motion of the protein along low-frequency normal mode vectors, the superposition of which describes large-scale motion [Hinsen, 1998]. Previous applications of NMA include predictions of conformational changes in proteins, steered MD simulations and high-throughput comparisons of dynamics in protein families [Bahar and Rader, 2005]. In Chapter 6 we will discuss the results of hydrogen-deuterium exchange NMR experiments which, as shown in Figure 1.2, concern long timescale protein motion. We use NMA to model the the slow large-amplitude motions of the protein using normal mode vectors so that we can make predictions concerning such long timescales of protein motion.

Global protein motion can be described as the superposition of a set of independent concerted motions [Suhre and Sanejouand, 2004; Petrone and Pande, 2006]. Each independent motion describes a state of the system where all of the particles oscillate with the same characteristic frequency [Dobbins et al., 2008]. The vectors describing these independent motions are the normal mode vectors; NMA determines the vectors along with their respective frequencies. The calculation of normal mode vectors from an input structure involves approximating the potential energy of the system around a global minimum. In standard NMA the potential energy function is described using bond lengths, bond angles, and dihedral angles between bonded atoms as well as steric repulsions, van der Waals attractions, and electrostatic interactions between non-bonded atoms [Levitt, 1983]. Minimising this potential energy function is computationally intensive and often inaccurate [Tirion, 1996]. The computational demands of standard NMA applied to proteins are compounded by the large system sizes — tens of thousands of atoms — typically involved [Hinsen, 1998].

The potential energy function may simplified in order to remove the minimisation step [Tirion, 1996]. In this simplification, the pairwise interactions between neighbouring atoms are modelled as springs, and the spring constant is the same for each interaction [Tirion, 1996]. This simple balls and springs representation is the ENM. The advantage of the ENM is that the input structure is *by definition* a minimum for the potential and minimisation is no longer required. The lowest frequency normal modes are not significantly affected when calculated in this way [Tirion, 1996]. Large amplitude protein motion can be accurately described by a small number of low-frequency normal modes [Krebs et al., 2002; Alexandrov et al., 2005; Jimenez-Roldan et al., 2012].

Further simplification is achieved through coarse-graining of the protein. In the coarse-grained ENM, each residue of the protein is represented by a point mass located at the $C_\alpha$ position. This simplification remains sufficient to calculate the backbone motion of the protein and therefore accurately characterise the low-frequency normal modes [Hinsen, 1998]. Further coarse-graining methods have also been successfully implemented [Gohlke and Thorpe, 2006]. We implement the NMA of the coarse-grained ENM using the ELNEMO software [Suhre and Sanejouand, 2004]. Normal modes are rapidly generated. The first six normal modes, $m_1, ..., m_6$ are the three rotations and translations of a rigid body in space, with frequency of zero. These are neglected, and the next normal modes are used. In the work in chapter 4 we simulate motion along $m_7 - m_{16}$, the first ten non-trivial modes. These are the normal modes with the lowest non-zero frequencies. In Chapter 4 we use an approach which combines random motion with directed motion along the mode direction [Jimenez-Roldan et al., 2011].

## 1.6 NMR spectroscopy

We study CypA experimentally using NMR spectroscopy. In particular, we study the effect of CsA binding on CypA experimentally using hydrogen-deuterium exchange NMR (HDX). In HDX experiments, the surface exposure at the residue level is detected for a protein in solution over a long timeframe [Henzler-Wildman and Kern, 2007]. We are interested in how the results of such experiments correlate with our coarse-grained simulations, which are able to probe large-scale protein motion. A brief introduction to the NMR experiments conducted is given here. The reader is also referred to the excellent, more detailed explanations available elsewhere [Hore, 1995; James, 1998; Doucleff et al., 2011; Hornak, 2011].

### 1.6.1 Nuclear spin and magnetic resonance

Atomic nuclei and elementary particles have a fundamental property called spin. As the mass of a particle responds to a gravitational field or its charge responds to electrical fields, so spin responds to magnetic fields [Doucleff et al., 2011]. Each atomic nucleus has a spin quantum number $S$, which describes how it will respond to the magnetic field. NMR is most commonly applied to nuclei with $S = \frac{1}{2}$, examples of which are $^1$H, $^{13}$C and $^{15}$N. The angular momentum of a spin-$\frac{1}{2}$ nucleus has two permitted directions, and in the absence of a magnetic field these have the same energy [Hore, 1995]. When a magnetic field is applied to the sample, the magnetic moments of the nuclei align with the magnetic field or oppose it. The lower energy state $E_l$ is to align with the field, opposing the field is the upper energy state $E_u$. The population becomes biased towards $E_l$ so that the ratio of populations is

$$\frac{N(E_l)}{N(E_u)} = e^{\Delta E/kT} \tag{1.1}$$

where $N(E_l)$ and $N(E_u)$ are the number of nuclei in $E_l$ and $E_u$ respectively, $\Delta E$ is the energy gap $E_u - E_l$, $k$ is Boltzmann's constant and $T$ is the temperature in Kelvin [Hore, 1995]. When $\Delta E$ is larger, the bias increases and the intensity of the NMR signal increases. It should be noted however, that the population bias is tiny. On the 700 MHz spectrometer used in some of our NMR experiments, at 25°C, the bias is $\frac{N(E_l)}{N(E_u)} = 1.00011268$. This means that for every $1,000,000$ nuclei in $E_u$ there are $1,000,113$ nuclei in $E_l$. The population bias is referred to as the bulk magnetisation of the sample. Conventionally the magnetic field is said to lie along the $z$-axis. The magnetic moments precess around this axis with a bias along the direction of the magnetic field. The vector of bulk magnetisation lies in the positive $z$-direction with no component in the $x$-$y$ plane [James, 1998]. The larger the magnetic field, the larger the $\Delta E$, and therefore the greater the magnitude of the bulk magnetisation. Energy $E$ is related to frequency $\nu$ via Planck's constant $h$, as $E = h\nu$. When electromagnetic radiation with a frequency corresponding to the energy gap between the two states is passed through the sample, magnetic resonant absorption occurs [Hore, 1995]. An NMR spectrum is a plot of frequency against absorption: a resonant frequency leads to an absorption which is a signal in the spectrum, referred to as a resonance. The energy gap can be made larger by increasing the strength of the magnetic field. NMR spectrometers are typically categorised based on this strength, but are often referred to in terms of the resonant frequency of a proton in this magnetic field. For example, a 16.45 Tesla magnet results in a resonant frequency for $^1$H nuclei of 700 MHz, and so a spectrometer with a 16.45 Tesla magnet would be referred to as a 700 MHz spectrometer. NMR spectrometers are typically in the region of $400 - 900$ MHz and these frequencies fall within the radiofrequency (RF) part of the electromagnetic spectrum. This means that the electromagnetic radiation which is passed through the sample is an

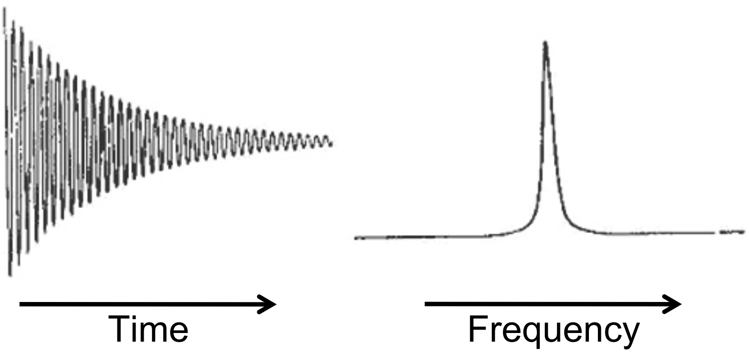


Figure 1.9: Left: an example of an FID, where intensity of the signal is plotted as a function of time. Right: after applying a Fourier transform, the intensity of the signal is plotted as a function of the frequency of the signal. Figure from [Harvard Medical School, 2013].

RF pulse. A short RF pulse is applied, causing the bulk magnetisation to rotate towards the $x$-$y$ plane. The time for which to apply the pulse in order to induce a 90° rotation is determined. This is referred to as the 90° pulse. After the 90° rotation the bulk magnetisation vector has no $z$ component and non-zero $x$ and $y$ components. The precessing magnetic moments in the $x$-$y$ plane induce an oscillating voltage in a detection coil surrounding the sample. As the bulk magnetisation relaxes back towards equilibrium, the magnitude of the vector in the x-y plane decreases. The oscillating voltage consequently decreases. This is known as free induction decay (FID) [James, 1998]. The FID is the NMR signal — absorption of a resonant frequency induces a change in the bulk magnetisation which is detected as an induced oscillating voltage. The FID contains information from all the individual excited spins which make up the bulk magnetisation. By applying a Fourier transform to the FID, the signal is converted from the time domain into the frequency domain. Figure 1.9 illustrates the effect of applying a Fourier transform to a simplified version of the FID.

The magnetic field experienced by individual protons in the sample varies according to their chemical environment and hence to their position within the molecule. Surrounding electrons shield the proton from the magnetic field, and so a proton surrounded by many electrons will experience a weaker magnetic field. As a result the energy gap will be reduced and so will its resonance frequency $\nu$ [Hore, 1995]. The reduction in frequency is on the order of Hz — a millionth of the resonance frequency itself. This effect is known as the chemical shift. The shifted resonance frequency is compared with $\nu_{\mathrm{ref}}$, the resonance frequency of the reference

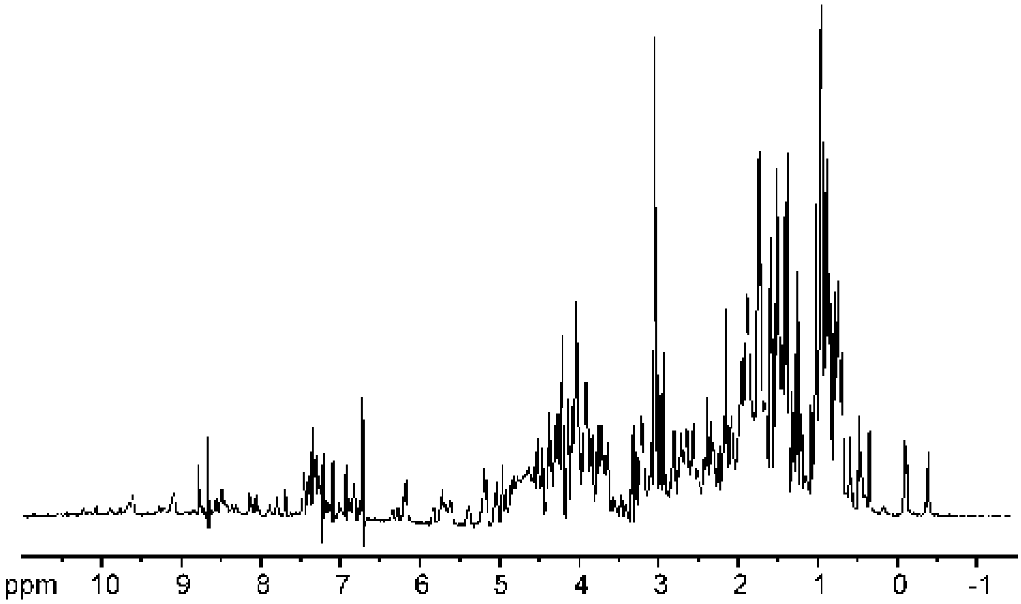


Figure 1.10: A typical 1D NMR protein spectrum. NMR signal intensity is plotted against proton chemical shift in ppm. Figure taken from [Schirra, 2013].

compound to give the chemical shift $\delta$ in parts per million (ppm)

$$\delta = 10^6 \times \frac{\nu - \nu_{\rm ref}}{\nu_{\rm ref}}. \tag{1.2}$$

The protons of the reference compound thus have $\delta = 0$, and since $\nu_{\rm ref}$ changes depending on the strength of the magnet in the spectrometer, normalising by $\nu_{\rm ref}$ in Equation 1.2 means that chemical shifts can be compared directly between spectrometers [Hore, 1995]. Since in proteins there are many protons in different environments resulting in different resonant frequencies, the FID signal is more complex than that shown in Figure 1.9. CypA, for example, contains 1236 protons [ProtParam tool, 2013]. As a result, after Fourier transform, there are many peaks in the frequency domain. The intensity of each peak is proportional to the number of protons resonating at that particular frequency. This type of plot is a 1D NMR spectrum. A typical example of a 1D NMR protein spectrum is shown in Figure 1.10.

### 1.6.2 2D NMR: HSQC, TOCSY and NOESY

Extra dimensions can be employed in NMR spectroscopy in order to spread the peaks and make it easier to identify which peak corresponds to which proton environment. It is possible to add another dimension by transferring magnetisation between two atoms, either through a bond (J-coupling) or through space [James, 1998]. Heteronuclear Single Quantum Correlation (HSQC) involves the through-bond transfer of magnetisation between two different types of nuclei. We use $^{1}$H-$^{15}$N HSQC, which involves transfer between an amide proton and its associated $^{15}$N nucleus. This through-bond transfer is a $^{1}\mathrm{J}_{^{1}\mathrm{H},^{15}\mathrm{N}}$ coupling. Using simultaneous RF pulses on both $^{1}$H and $^{15}$N channels, magnetisation is transferred from $^{1}$H to $^{15}$N and back. By varying the pulse sequences, the amplitude of the signal obtained from the protons oscillates at the frequency of the $^{15}$N nucleus. In this way, the proton broadcasts the chemical shift of its associated $^{15}$N nucleus by amplitude modulation [Doucleff et al., 2011]. The chemical shifts of the nitrogen and its associated proton determined using $^{1}$H-$^{15}$N HSQC are plotted on a HSQC spectrum, an example of which is given in Figure 1.11. Each cross-peak in the HSQC represents an N-H pair, a series of contour lines indicating signal intensity. The location of each cross-peak in the spectrum is determined by the chemical shifts of both of the $^{1}$H and $^{15}$N nuclei. This means that there will be approximately one peak per amino acid. Some amino acids contribute more than one peak due to the presence of $^{15}$N in their side-chains, and proline does not contribute a peak since it does not have an amide proton. The example spectrum in Figure 1.11 is in fact the HSQC spectrum that we recorded for unbound CypA, and is shown again in Chapter 6 as Figure 6.2, where it is discussed in detail.

In a total correlation spectroscopy (TOCSY) experiment, repetitive series of pulses are used to detect through-bond couplings (J-couplings) between protons belonging to the same amino acids. Nuclear Overhauser effect spectroscopy (NOESY) uses NOEs instead of through-bond couplings. NOEs are through-space interactions which involve non-radiative (instantaneous) transfer of magnetisation in a way which is highly distance dependent [James, 1998]. Transfer through NOEs can occur between nuclei which are not separated by more than 5 Å [Hore, 1995; Doucleff et al., 2011]. NOESY cross-peaks are therefore between proximal nuclei rather than between those which are connected through a small number of bonds.

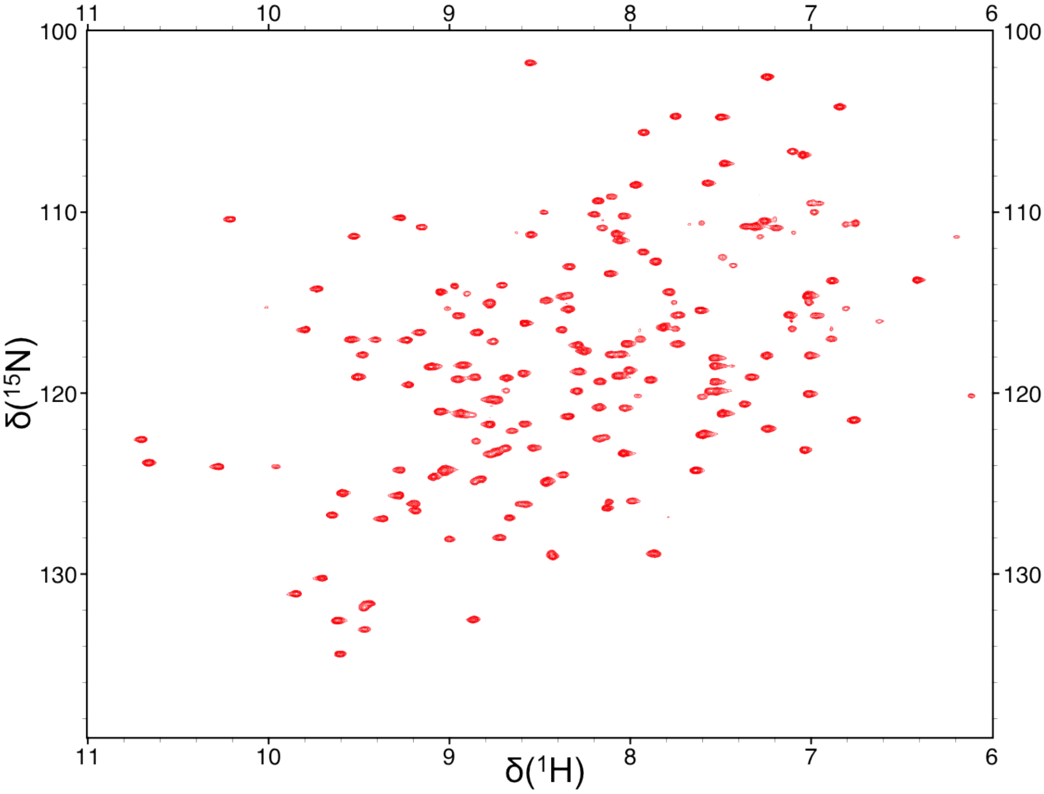


Figure 1.11: An example of a $^1$H-$^{15}$N HSQC spectrum. Contour lines in red show the intensity of the signal for each cross-peak. This is the $^1$H-$^{15}$N HSQC spectrum we generated for CypA, as detailed in Chapter 6. Chemical shifts $\delta(^{15}\mathrm{N})$ and $\delta(^1\mathrm{H})$ are given in ppm.

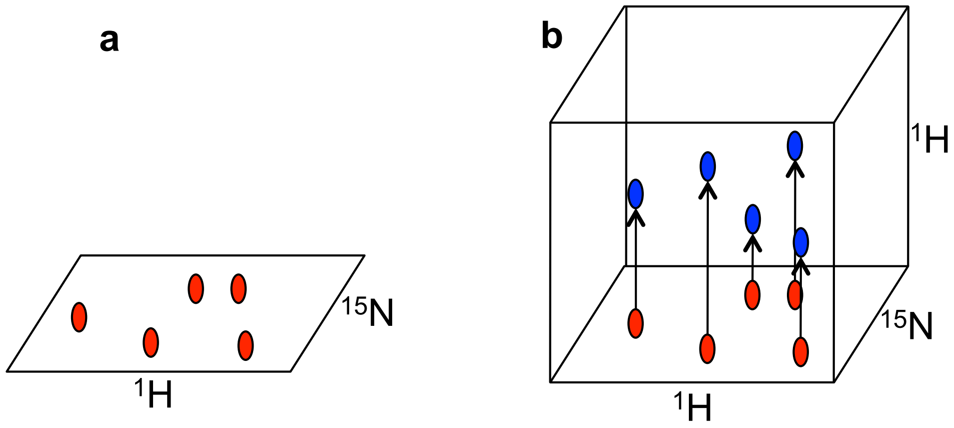


Figure 1.12: A diagram showing the distribution of cross-peaks in 3D NMR. (a) the $^1$H-$^{15}$N HSQC spectrum, illustrated with five cross peaks in red. (b) In the 3D spectrum, the cross-peaks in the HSQC plane have additional cross-peaks associated with them in the third dimension, shown in blue. These are the result of through-bond or through-space couplings, for TOCSY-HSQC and NOESY-HSQC experiments, respectively.

### 1.6.3 3D NMR: TOCSY-HSQC and NOESY-HSQC

For each backbone N-H group in the protein, we need to identify the location of its cross-peak in the HSQC spectrum so that we can analyse the HDX data. We do this using data obtained from 3D NMR experiments. 3D TOCSY-HSQC and 3D NOESY-HSQC are 2D TOCSY/NOESY experiments extended by a HSQC step. In a 3D experiment, the cross-peaks are distributed in a cube instead of a plane. Each axis of the cube corresponds to a type of nucleus, as shown in the diagram in Figure 1.12.

Above each cross-peak in the HSQC plane are TOCSY/NOESY strips containing associated peaks for other protons in the same residue or proximal to the given amide proton, as appropriate. Examples of such strips can be seen in Chapter 6, Figure 6.4. The cross-peaks in the HSQC spectrum can be associated with particular N-H groups in the protein using a combination of the 3D approaches. TOCSY-HSQC is used to identify the type of amino acid which is associated with a particular cross-peak in the HSQC. NOESY-HSQC allows us to determine which amino acid a particular peak is close to in the 3D fold of the protein [James, 1998]. Since the sequence of CypA is known, cross-peaks in the HSQC which correspond to two amino acids of the same type may be distinguished on the basis of their sequential neighbours, protons of which will be visible in the NOESY strip of the

NOESY-HSQC experiment. This procedure of linking amino acids together in sequence using NOESY cross-peaks is called sequential assignment.

### 1.6.4 Hydrogen-deuterium exchange NMR

3D NMR experiments are used to help assign the HSQC spectrum – to identify which peak in the spectrum corresponds to which N-H group in the protein. Once this has been accomplished, results from HDX can be analysed, yielding information on the mobility and surface of the protein [Woodward et al., 1982]. The deuteron is sufficiently different from the proton in its magnetic properties not to be detected in the $^{1}$H-$^{15}$N HSQC experiment. If an amide proton is swapped with a deuteron prior to recording the HSQC spectrum, the peak corresponding to the proton will therefore disappear. It is this effect which is studied using HDX NMR [Woodward et al., 1982]. When a protein is placed into $D_2O$, the protons which are exposed may exchange with the deuterons of the solvent. Depending on the level of surface exposure and the mobility of the protein, the amide protons will exchange at different rates. By recording a series of HSQC experiments at different time intervals after immersion of the protein in $D_2O$, one can track the loss of signal over time. We have conducted HDX experiments on CypA in the presence and absence of CsA in order to measure the effect of ligand binding on the intramolecular mobility of the protein. The results for these experiments are given in Chapter 6.

# Chapter 2

# Inhibition of HIV-1 protease: the rigidity perspective

## 2.1 Introduction

We use the FIRST software, introduced in Section 1.4, to perform rigidity analysis on a set of 212 crystal structures of HIV-1 protease. This work has been published in the journal Bioinformatics [Heal et al., 2012]. Preliminary data has also been published in a Journal of Physics Conference Series article [Heal et al., 2011]. We demonstrate the use of rigidity analysis to rapidly generate information on the effect of ligand binding upon a protein. In this case the protein of interest, HIV-1 protease, is a key drug target and a comparison of available inhibitors is made. The virus HIV-1 and the role of HIV-1 protease in its life-cycle was discussed in Section 1.2. The large number of high resolution crystal structures of the enzyme in complex with a range of inhibitors (see Figure 2.3) make HIV-1 protease an ideal subject for a comprehensive analysis using FIRST. The interest in the enzyme as a drug target and the implications of successful inhibition are also key motivators for pursuing this study. The results obtained lead to the categorisation of the FDA approved inhibitors and prompt a discussion on the best way to continue with drug design. During the study, we discuss rigidity dilution plots, rigid cluster decompositions and energy cut-off values, all of which were introduced in the previous chapter. For each structure crystallised with a ligand bound, we compare the results of rigidity analysis before and after removing the inhibitor.

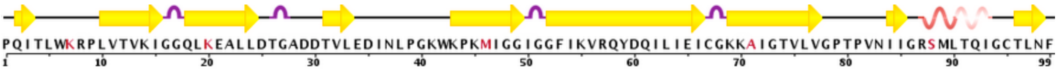

PQITLWKRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQILIEICGKKAIGTVLVGPTPVNIIGRSMLTQIGCTLNF
1
10
20
30
40
50
60
70
80
90
99

Figure 2.1: The primary structure of 3LZU, taken from the PDB. Since the protein is a homodimer, the structure of one monomer is shown. The sequence contains two mutations — Q7K and N88S, which are shown in red. The other residues shown in red show are highlighted in the PDB as mismatches with the sequence in the UniProt Knowledge Base [UniProt knowledge base, 2013], but are not listed as mutations by the authors publishing the structure. The secondary structure elements are shown as yellow arrows ($\beta$-sheet), red waves ($\alpha$-helix) and purple humps (turn).

### 2.1.1 The structure of HIV-1 protease

Wild-type HIV-1 protease is a symmetrical homodimer comprising polypeptide chains each containing 99 amino acids. We refer to the two monomers as chain A and chain B, with residues numbered from 1 – 99 in chain A and from 101 – 199 in chain B. The amino acid sequence for chain A, taken from the crystal structure 3LZU, is given in Figure 2.1. This structure has two mutations from the wild-type protease, highlighted in red: Q7K and N88S. The catalytic triad (Asp-Thr-Gly) is located at residues 25-27 and is a form of active site common to aspartic proteases [Toh et al., 1985]. The glycine-rich flap tips surround residue 50, which itself is an isoleucine residue (Ile50). The secondary structure, as identified by the DSSP (Define Secondary Structure of Proteins) algorithm, is shown above the primary structure. The DSSP was created with the aim of approximating the intuitive notion of secondary structure use of an objective algorithm [Kabsch and Sander, 1983]. The geometry of the atoms in a crystal structure are used to assign hydrogen bonds and subsequently a secondary structure. The secondary structure is formed of the elements turns and bridges, where extended turns form helices and extended bridges $\beta$-strands. The turns are defined by hydrogen bonds between residues $i$ and $i + n$ where $n \in [1, 5]$, and bridges by hydrogen bonds between residues which are more separated along the primary structure. In the structure 3LZU, there are four turns identified along the backbone along (shown in purple in Figure 2.1) with one $\alpha$-helix (red) defined by the characteristic $i, i+4$ hydrogen bonding pattern. There are nine extended $\beta$-strands (yellow arrows).

Figure 2.2 shows how the secondary structures arrange to form the monomers of the tertiary structure, and how these monomers associate to form the dimeric quaternary struture. The structure 3LZU has been chosen from the many available partly due its high resolution (1.76 Å) and partly because the ligand in the structure is the inhibitor darunavir — an inhibitor which will be important in discussion. In

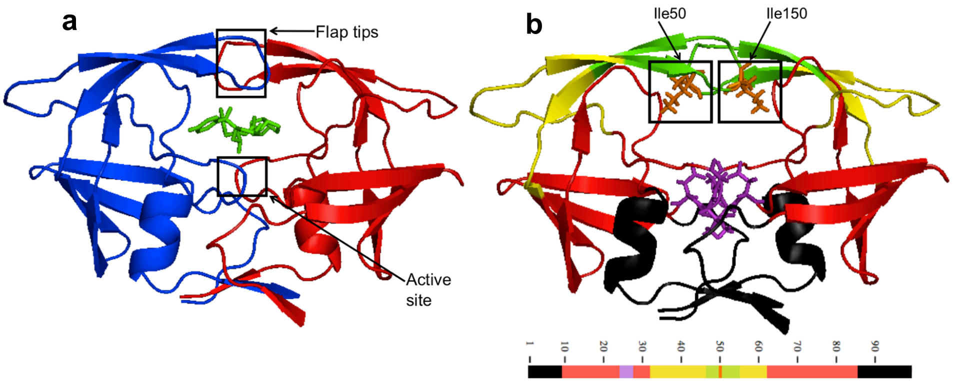


Figure 2.2: The structure of HIV-1 protease. (a) The structure 3LZU presented as a cartoon with chain A in red and chain B in blue. This highlights the secondary structure of the protein, which is largely $\beta$-sheet. The heavy atoms of the inhibitor darunavir are shown in green. The $\beta$-hairpin flaps and the active site are indicated. (b) The same structure is shown after deletion of the inhibitor, with the terminal region in black, the core in red and the flaps in yellow. The active site within the core is coloured purple and the flap tip residues in green. The residues Ile50 and Ile150 at the very tips of the flaps are coloured orange and labelled. These residues and the active site residues are represented as sticks. The primary structure of chain A is shown beneath with the same colour code.

Figure 2.2, secondary structure motifs are clearly depicted as arrows ($\beta$-sheet) and helices ($\alpha$-helix). The terminal region is situated at the base of the enzyme and consists of the N- and C-termini of both monomers, the turn located immediately above this and the two $\alpha$-helices towards the C-terminal of the monomers. The core region is composed primarily of $\beta$-strands. The active site is situated within the core, consisting of residues 25 – 27 and, due to symmetry, residues 125 – 127. The active site lies beneath the flaps, the tips of which are indicated in Figure 2.2. The flap region extends out to the flap elbows which, it has been suggested, act as a cantilever for flap motion [Swaminathan et al., 1991; Harte et al., 1992]. These are located at the widest section of the protein, coloured yellow in Figure 2.2(b). The flap tips are a glycine-rich area of the protein and thus have enhanced conformational flexibility [Hong et al., 1997].

### 2.1.2 X-ray crystal structures of HIV-1 protease

There is a wealth of high resolution structures available in the PDB. This is demonstrated in Figure 2.3, which gives an overview of the 212 structures that we have investigated using rigidity analysis. The structures are organised according to the location of any mutations they have, as well as the identity of the inhibitors with which they have been crystallised. Some subsets of structures with inhibitors from the same family are also given, labelled with asterisks. For example, we have the subset KNI-*** consisting of structures which were crystallised with the inhibitors KNI-272, KNI-577, KNI-764, KNI-10075, KNI-10033, KNI-10265, KNI-10265, KNI-1689. These inhibitors may exhibit different binding properties with the protease [Velazquez-Campoy et al., 2001], but are grouped together here for interest of comparison within an inhibitor family. We have identified only some of these subsets in Figure 2.3, more could be selected from within the category labeled "other inhibitors". The structures inevitably have a wide range of crystallisation conditions, inhibitors and mutations. It is therefore difficult to extract useful information about the effect of a set of *individual* inhibitors on the rigidity of the same protein. Within the full set of 206 structures a subset of main inhibitors has been identified based on some of the commercially available protease inhibitors approved for use in antiretroviral therapies. We refer to the FDA approved inhibitors by their three-letter abbreviations: APV (amprenavir, available as the pro-drug fosampranavir) , ATZ (atazanavir), DRV (darunavir), IDV (indinavir), NFV (nelfinavir), SQV (saquinavir) and TPV (tipranavir) [Pokorná et al., 2009]. We also include DMP-323 and the polypeptidic inhibitor (ACE)TI(NLE)(NLE)QR amongst the main inhibitors due to the relatively large number of high resolution structures crystallised with these inhibitors. We label DMP-323 as DMP and (ACE)TI(NLE)(NLE)QR as PEP. As can be seen in Figure 2.3, there are a different number of structures for each of the main inhibitors, which are as follows: APV (8), ATZ (4), DRV (16), IDV (5), NFV (4), SQV (6), TPV (4), DMP (5), and PEP (5), making a total of 57 structures. The FDA-approved lopinavir (LPV) and ritonavir (RTV) are omitted from our list of main inhibitors, since there are insufficient high resolution crystal structures of HIV-1 protease in complex with these ligands. Ritonavir is the focus of Section 2.3.6.

### 2.1.3 Dynamics of HIV-1 protease

As well as X-ray crystal structures, which are used as input for rigidity analysis in this study, other structural studies have been conducted experimentally. NMR spectroscopy [Nicholson et al., 1995; Ishima et al., 1999; Freedberg et al., 2001; Katoh et al., 2003] and fluorescence spectroscopy [Rodríguez et al., 1993] have been used to determine the structure and to study the flexibility of the protein. The flaps have been found to be highly flexible in the unbound enzyme, with internal flap motion occurring on a sub-ns timescale [Freedberg et al., 2001; Katoh et al., 2003]. On a timescale of 100 $\mu$s, they undergo a different large amplitude motion in slow dynamic

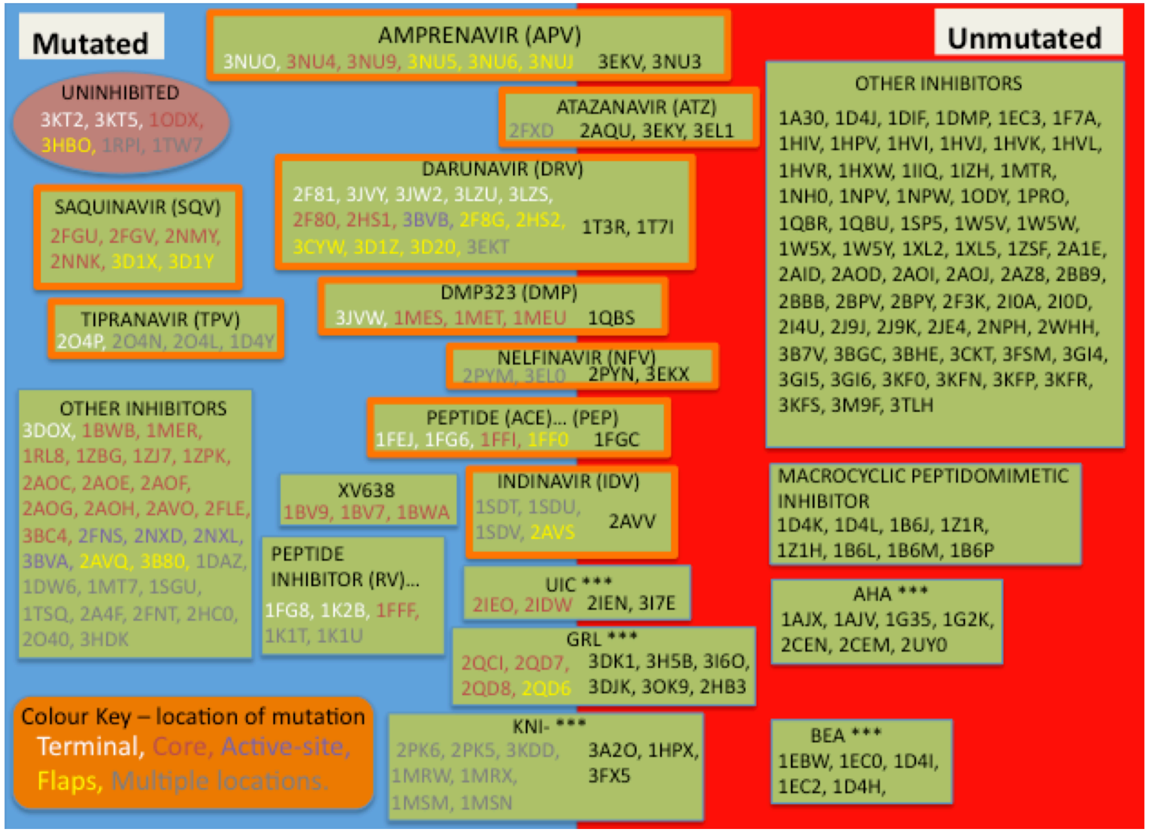

Mutated
Unmutated
AMPRENAVIR (APV)
3NUO, 3NU4, 3NU9, 3NU5, 3NU6, 3NUJ 3EKV, 3NU3
UNINHIBITED
3KT2, 3KT5, 1ODX,
3HBO, 1RPI, 1TW7
ATAZANAVIR (ATZ)
2FXD 2AQU, 3EKY, 3EL1
DARUNAVIR (DRV)
2F81, 3JVY, 3JW2, 3LZU, 3LZS,
2F80, 2HS1, 3BVB, 2F8G, 2HS2, 1T3R, 1T7I
3CYW, 3D1Z, 3D20, 3EKT
SAQUINAVIR (SQV)
2FGU, 2FGV, 2NMY,
2NNK, 3D1X, 3D1Y
DMP323 (DMP)
3JVW, 1MES, 1MET, 1MEU 1QBS
TIPRANAVIR (TPV)
2O4P, 2O4N, 2O4L, 1D4Y
NELFINAVIR (NFV)
2PYM, 3EL0 2PYN, 3EKX
OTHER INHIBITORS
3DOX, 1BWB, 1MER,
1RL8, 1ZBG, 1ZJ7, 1ZPK,
2AOC, 2AOE, 2AOF,
2AOG, 2AOH, 2AVO, 2FLE,
3BC4, 2FNS, 2NXD, 2NXL,
3BVA, 2AVQ, 3B80, 1DAZ,
1DW6, 1MT7, 1SGU,
1TSQ, 2A4F, 2FNT, 2HC0,
2O40, 3HDK
PEPTIDE (ACE)... (PEP)
1FEJ, 1FG6, 1FFI, 1FF0 1FGC
XV638
1BV9, 1BV7, 1BWA
INDINAVIR (IDV)
1SDT, 1SDU,
1SDV, 2AVS 2AVV
PEPTIDE
INHIBITOR (RV)...
1FG8, 1K2B, 1FFF,
1K1T, 1K1U
UIC ***
2IEO, 2IDW 2IEN, 3I7E
GRL ***
2QCI, 2QD7, 3DK1, 3H5B, 3I6O,
2QD8, 2QD6 3DJK, 3OK9, 2HB3
Colour Key – location of mutation
Terminal, Core, Active-site,
Flaps, Multiple locations.
KNI- ***
2PK6, 2PK5, 3KDD, 3A2O, 1HPX,
1MRW, 1MRX, 3FX5
1MSM, 1MSN
OTHER INHIBITORS
1A30, 1D4J, 1DIF, 1DMP, 1EC3, 1F7A,
1HIV, 1HPV, 1HVI, 1HVJ, 1HVK, 1HVL,
1HVR, 1HXW, 1IIQ, 1IZH, 1MTR,
1NH0, 1NPV, 1NPW, 1ODY, 1PRO,
1QBR, 1QBU, 1SP5, 1W5V, 1W5W,
1W5X, 1W5Y, 1XL2, 1XL5, 1ZSF, 2A1E,
2AID, 2AOD, 2AOI, 2AOJ, 2AZ8, 2BB9,
2BBB, 2BPV, 2BPY, 2F3K, 2I0A, 2I0D,
2I4U, 2J9J, 2J9K, 2JE4, 2NPH, 2WHH,
3B7V, 3BGC, 3BHE, 3CKT, 3FSM, 3GI4,
3GI5, 3GI6, 3KF0, 3KFN, 3KFP, 3KFR,
3KFS, 3M9F, 3TLH
MACROCYCLIC PEPTIDOMIMETIC
INHIBITOR
1D4K, 1D4L, 1B6J, 1Z1R,
1Z1H, 1B6L, 1B6M, 1B6P
AHA ***
1AJX, 1AJV, 1G35, 1G2K,
2CEN, 2CEM, 2UY0
BEA ***
1EBW, 1EC0, 1D4I,
1EC2, 1D4H,

Figure 2.3: An overview of the 212 HIV-1 protease structures. The blue half of the diagram lists the structures that have at least one mutation. The colour key indicates how the structures are listed to show which regions contain mutations. In the red half of the diagram are the wild-type structures. Each individual box lists the structures complexed with a particular inhibitor, and the boxes containing our set of main inhibitors are highlighted with an orange border. The two large boxes labelled "other inhibitors" list structures crystallised with various inhibitors for which only a small number of structures are present in the PDB. More details can be found in Table B.2, and in the PDB itself. Inhibitors labelled with asterisks, such as KNI ***, represent a subset of inhibitors from the same family.

equilibrium between semi-open and less ordered open conformations [Ishima et al., 1999]. Computational work has also been used to investigate the flexibility and mobility of HIV-1 protease. Conformational dynamics have been investigated using MD simulations [Swaminathan et al., 1991; York et al., 1993; Harte and Beveridge, 1993; Zhu et al., 2003] and coarse grained models [Chang et al., 2006; Tozzini and McCammon, 2005]. The flaps have been found to be highly flexible with distances between their tips ranging from 7 Å to 25 Å [Collins et al., 1995]. The conversion between closed, semiopen and fully open flaps has been simulated in qualitative agreement with experimental data [Hornak et al., 2006; Hornak and Simmerling, 2007].

## 2.2 Methods

We selected 212 X-ray crystal structures of HIV-1 protease from the PDB on the basis of their resolution. Bond strength in FIRST is determined directly from the geometry of the structure using the Mayo potential [Dahiyat et al., 1997], and so high resolution improves the accuracy of the inferred bond network. We used a resolution cutoff of 2.0 Å when selecting the structures of HIV-1 protease, demanding better resolution than some previous studies [Wang et al., 2004; Rader et al., 2004; Wells et al., 2009; Radestock and Gohlke, 2008; Rader and Brown, 2011]. We were able to do this thanks to the wealth of structures available. Crystal water molecules were removed, before the REDUCE software [Word et al., 1999] was used to add the hydrogen atoms to the X-ray crystal structure and to flip side chains of Asn, Gln and His residues where necessary. Oxygen and nitrogen atoms cannot be distinguished by their X-ray diffraction patterns and for these particular residues, this causes ambiguity in side-chain orientation. Through REDUCE, the orientation is chosen so as to better fit the non-covalent bond network. For each structure of HIV-1 protease crystallised with an inhibitor bound, an unhibited copy of the structure was generated by manually deleting the inhibitor in PYMOL. Rigidity analysis was conducted using FIRST version 6.1 [Flexweb, 2013].

## 2.3 Results

### 2.3.1 Processing the crystal structures

To process the structures in this study, we first removed the crystal water molecules and subsequently used REDUCE to flip sidechains where necessary. An alternative method would be to run the REDUCE software initially, and to remove the crystal water molecules subsequently. To our knowledge there is no discussion in the literature regarding the order of these processes. In some studies the crystal waters are removed before adding the hydrogen atoms to the protein [Radestock and Gohlke, 2008; Wells et al., 2009]; in some studies the order is not specified [Zavodszky et al., 2004; Hespenheide et al., 2002]. In order to address the issue of the order of water removal and hydrogenation of the protein the results were processed twice for a set of eight crystal structures, changing the order for the second run. The two sets of results were compared and no significant difference was found. These data can be found in Appendix B.3.

It has previously been suggested that the presence of buried water is important as it coordinates the protease flap residues [Wlodawer and Vondrasek, 1998]. We note that this should not affect the rigidity analysis presented by FIRST, as MD simulations [Mamonova et al., 2005] have shown that the hydrogen bonds formed between the protein and the solvent have a short duty cycle of typically less than 1 ps, comparable to those in liquid water [Matsumoto et al., 2002], and should be neglected when considering rigidity.

### 2.3.2 Rigidity analysis

The mechanisms of FIRST and its outputs were discussed in Section 1.4. Of particular relevance here are the concepts of the RCD and $E_{\text{cut}}$. We represent the RCD of the protein in terms of its backbone. The RCD of the $C_\alpha$ atoms is referred to as the *mainchain rigidity* of the protein. Examples of the mainchain rigidity of HIV-1 protease are given in Figure 2.4. The structure 3LZU is shown in both 3D and 1D representations. Rigid areas are illustrated as coloured blocks and flexible areas as thin black lines. Rigid clusters are coloured individually. In Figure 2.4(a), $E_{\text{cut}}$ is chosen so that the bond network is relatively large and the structure is mostly rigid. A single rigid cluster shown in red dominates the structure. The 1D representation clarifies some key features of the RCD. One rigid cluster includes residues from several non-contiguous sections of the primary sequence. Indeed, this large rigid cluster includes residues from both chains of the dimer. The $E_{\text{cut}}$ is lowered in Figure 2.4(b) from $E_{\text{cut}} = -1.970$ kcal/mol to $E_{\text{cut}} = -1.985$ kcal/mol, reducing the number of bonds in the network and hence making the structure more flexible. A selection of small rigid clusters shown in different colours are distributed throughout the protein. The 1D representation shows clearly the contrast between the largely rigid structure and the mostly flexible structure. This change occurred when $E_{\text{cut}}$ was lowered by just 0.015 kcal/mol.

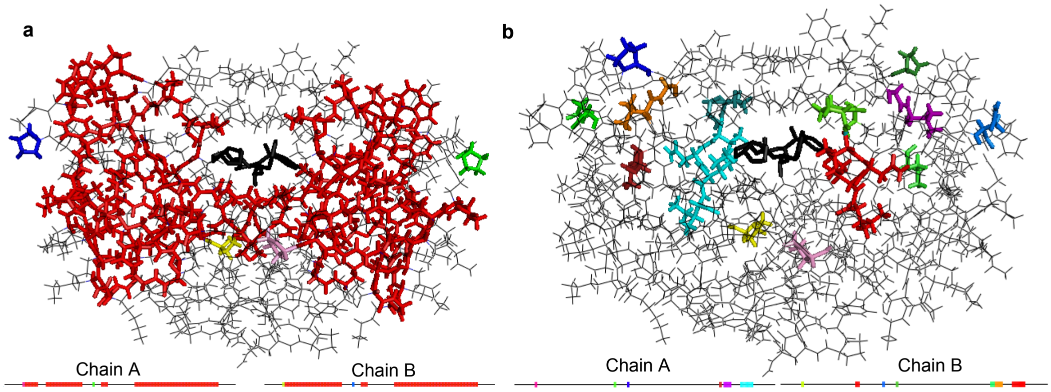


Figure 2.4: The mainchain rigidity of the HIV-1 protease structure 3LZU. Rigid clusters are represented as thick coloured regions and flexible parts of the protein as thin black lines. The heavy atoms of the inhibitor are shown in black at the centre of the enzyme. The RCD of 3LZU is shown when the structure is (a) mostly rigid at $E_{\text{cut}} = -1.970$ kcal/mol and (b) largely flexible at $E_{\text{cut}} = -1.985$ kcal/mol. The corresponding protein backbone is given below each plot, with chains A and B separated for clarity.

### 2.3.3 Rigidity dilution

As explained in Section 1.4.4, a rigidity dilution involves the systematic removal of hydrogen bonds in order of strength and running rigidity analysis after each change to the bond network. With the removal of bonds the protein undergoes a transition from being rigid to being flexible, and the manner in which this occurs may be informative [Rader et al., 1999]. We use rigidity dilution plots to show this loss of rigidity. In such a plot, we display a new 1D representation of the RCD whenever main chain rigidity changes. This highlights where in the structure most of the rigidity resides, and where rigidity is most readily lost. We note that the manner in which these plots are constructed means that the vertical axis is non-linear with $E_{\text{cut}}$. Figure 2.5 shows four such rigidity dilution plots, describing the rigidity of two HIV-1 protease crystal structures – 3LZU and 2HS1 – before and after the deletion of an inhibitor. The structures 2HS1 and 3LZU were both selected because they were crystallised in complex with darunavir, an antiviral drug that was approved by the FDA in June 2006 and was designed to form strong interactions with many different mutated structures of the protease [Ghosh et al., 2007]. The dilution plots of the structures as they were crystallised are given in Figures 2.5(a) and 2.5(c). In both cases the protein loses the majority of its rigidity abruptly, as the 1D profiles change from being mostly rigid to being mostly flexible with a single step in $E_{\text{cut}}$. This is first order rigidity loss, as expected for a protein with a largely $\beta$-sheet

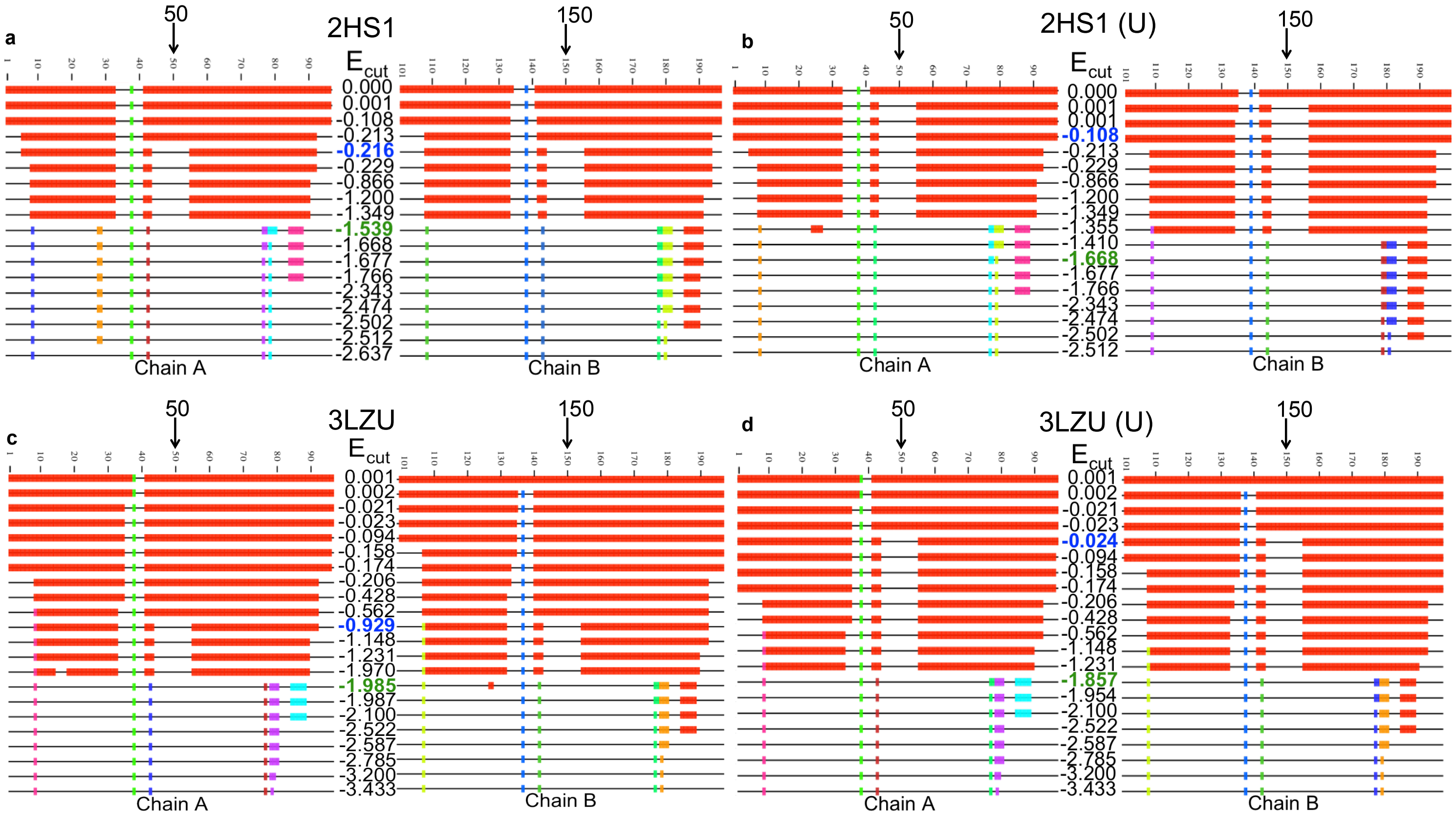


Figure 2.5: Rigidity dilution plots for the crystal structures 2HS1 and 3LZU. Rigidity dilution was carried out on the structures as crystallised (a), (c) and after deletion of the inhibitor (b), (d), indicated by the suffix (U). The 50th and 150th residue of each structure — the flap tips residues — are indicated by labelled arrows. The $E_{\text{cut}}$ used to determine the RCD for each row is shown between chains A and B. $E_{\text{body}}$ and $E_{\text{flap}}$, defined in Section 2.3.4, are shown in green and blue respectively.

secondary structure [Wells et al., 2009]. The flap regions around residues 50 and 150 become flexible at high $E_{\mathrm{cut}}$ relative to the main body of the protein. The helical regions near the terminal residues 86 – 94 remain rigid after the main body of the protein has become flexible. Plots for the same structures with the inhibitors deleted are given in Figures 2.5(b) and 2.5(d). Deletion of the inhibitor does not appear to alter the basic pattern of rigidity loss during the dilution. The flap regions become flexible at even higher $E_{\mathrm{cut}}$, however. The active site of the enzyme is not distinctive in the rigidity analysis of either the inhibited or uninhibited structures, forming part of the largest rigid cluster until the protein becomes mostly flexible. Therefore the principal effect of the inhibitor upon the rigidity profile of the protein is to rigidify the flap tip region rather than the active site.

### 2.3.4 $E_{\mathrm{body}}$ and $E_{\mathrm{flap}}$

For quantitative analysis of the rigidity dilution plots, it is typical to extract from them significant values of $E_{\mathrm{cut}}$. Previous studies such as those on rigidity in the context of thermostability [Radestock and Gohlke, 2008; Rader, 2010; Radestock and Gohlke, 2011] have used measures such as the folding core energy. This is defined to be the $E_{\mathrm{cut}}$ value of lowest line in the dilution plot where at least three consecutive residues are mutually rigid with at least three other consecutive residues of another secondary structural element [Hespenheide et al., 2002; Tastan et al., 2007]. In this instance we are interested in the function rather than the folding or melting point of the protein and so, in order to quantify the influence of inhibitor binding on the overall rigidity of the protein and on the flexibility of the flap regions, we define two significant values of $E_{\mathrm{cut}}$ for our analysis. We define $E_{\mathrm{body}}$ to be the highest $E_{\mathrm{cut}}$ at which at least 70 % of the protein is flexible. Due to the first order nature of the rigidity loss in HIV-1 protease, this value would not change significantly if we made our definition of $E_{\mathrm{body}}$ based on any flexibility in the 50 – 90 % range. Indeed, the value $E_{70}$ typically represents the entirely flexible protein, as seen in Figure 2.6. We also note that for HIV-1 protease the folding core energy and $E_{\mathrm{body}}$ are similar, typically separated in the dilution plots by just one line. The impact of choosing $E_{\mathrm{body}}$ over the folding core energy in the definition of our next measure is discussed in more detail in Section 2.3.11, and in particular in Figure 2.8. Residues Ile50 and Ile150 are residues at the tip of the flaps and have been tracked previously in MD to assess the motion of the flaps as the protein samples an ensemble of conformations in its native state [Zhu et al., 2003]. Due to the symmetry of the crystal structures used here these residues become flexible at the same $E_{\mathrm{cut}}$ in almost all of the structures. We select $E_{\mathrm{flap}}$ as our second important value of $E_{\mathrm{cut}}$ to be the $E_{\mathrm{cut}}$ at which Ile50

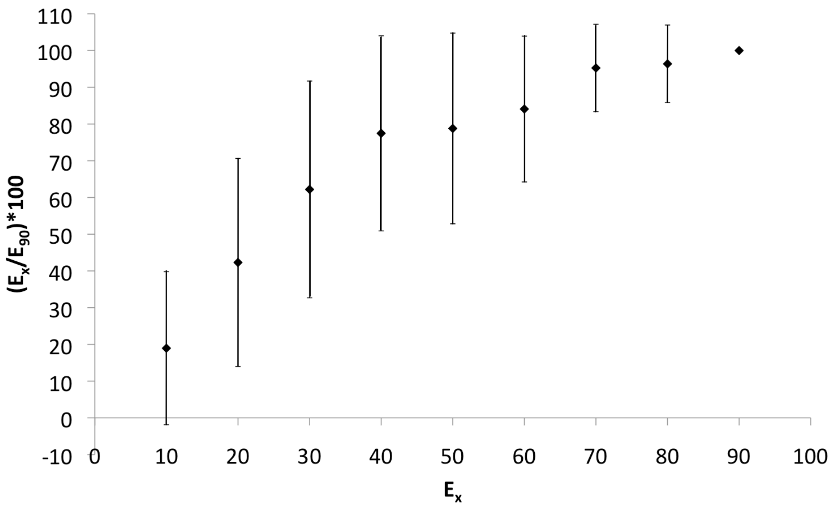


Figure 2.6: For $x \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$, $E_x$ is calculated: the highest $E_{\text{cut}}$ at which at least $x$ % of the protein residues are flexible. The value of each $E_x$ as a percentage of $E_{90}$ is then calculated. At $x = 70$, as chosen for the definition of $E_{\text{body}}$, this percentage is above 95 on average. The graph shows average values for the entire population of HIV-1 protease structures, with error bars equal to one standard deviation.

first becomes flexible in the rigidity dilution process. In each plot of Figure 2.5, the values of $E_{\text{body}}$ and $E_{\text{flap}}$ are coloured green and blue respectively. In Figure 2.5(a) for example, $E_{\text{body}} = -1.539$ kcal/mol and $E_{\text{flap}} = -0.216$ kcal/mol. For numerical reasons we sometimes observe that $E_{\text{flap}}$ has a small positive value, as observed in Figure 2.5(b). In this case, we take $E_{\text{flap}}$ to be the highest negative value of $E_{\text{cut}}$.

Previous studies [Wells et al., 2009; Tan and Rader, 2009] have included structures with resolution worse than 2 Å, and so we also examined a small selection of such structures and verified that rigidity analysis was qualitatively consistent: each showed first order rigidity loss with the flaps becoming flexible before the main body of the protein.

### 2.3.5 The effect of inhibitors on $E_{\text{body}}$ and $E_{\text{flap}}$

There is an observed variation in $E_{\text{body}}$ values across the set of crystal structures. For example, $E_{\text{body}} = -1.539$ kcal/mol for 2HS1 in Figure 2.5(a) and $E_{\text{body}} = -1.985$ kcal/mol for 3LZU in Figure 2.5(c). Deletion of the inhibitor can shift $E_{\text{body}}$ to higher values: $-1.355$ kcal/mol in Figure 2.5(b) and $-1.857$ kcal/mol in Figure 2.5(d). We now consider whether the presence of the inhibitor significantly affects the *overall* rigidity of the protease.

In Figure 2.7(a) we show the distribution of $E_{\text{body}}$ values for structures with the inhibitor present, which we call $E_{\text{body}}(\text{I})$. The distribution is approximately normal with a peak around $-1.5$ kcal/mol; almost all values lie between $-0.5$ kcal/mol and $-2.5$ kcal/mol [Wells et al., 2009]. Figure 2.7(b) is the distribution of $E_{\text{body}}$ values for uninhibited structures, that is, with the inhibitor deleted. This distribution of $E_{\text{body}}(\text{U})$ is similar to that of $E_{\text{body}}(\text{I})$; deletion of the inhibitor does not appear to have a strong effect on overall rigidity. Figures 2.7(c) and 2.7(d) are the corresponding plots for the distribution of $E_{\text{flap}}$ values. In contrast to the situation with $E_{\text{body}}$, the $E_{\text{flap}}$ values in general shift towards zero after the inhibitors are removed. We note that these distributions do not tell us about the effect of inhibitor deletion on each structure individually. The distribution of individual differences $E_{\text{body}}(\text{I}) - E_{\text{body}}(\text{U})$ and $E_{\text{flap}}(\text{I}) - E_{\text{flap}}(\text{U})$ are shown in Figures 2.7(e) and 2.7(f), respectively. In Figure 2.7(e) the distribution strongly peaks around a modal value close to zero, with few structures having a difference of less than $-0.5$ kcal/mol. This confirms that deletion of the inhibitor does not typically have a strong effect on overall rigidity. Contrastingly, Figure 2.7(f) is a bimodal distribution, with many structures showing a large change in $E_{\text{flap}}$ following the removal of an inhibitor. In all parts of Figure 2.7 the distribution of the main inhibitors is shown. The distributions for this set closely resemble the distributions for all structures, illustrating that in terms of $E_{\text{body}}$ and $E_{\text{flap}}$ the selection of main inhibitors is representative of the entire population.

### 2.3.6 The flexibility fraction $\Phi$

We have seen that the flap region becomes flexible at a higher $E_{\text{cut}}$ than the main body of the dimer. We can quantify this by comparing the $E_{\text{cut}}$ value at which the flap tips become flexible, $E_{\text{flap}}$, to the value at which the main body becomes flexible, $E_{\text{body}}$. We define the *flexibility fraction* $\Phi = \frac{E_{\text{flap}}}{E_{\text{body}}}$. This measure runs from zero, when the flaps are flexible from the beginning of the dilution (in which case $E_{\text{flap}} = 0$), to 1, if the flaps become flexible only when the main body becomes flexible (i.e. $E_{\text{flap}} = E_{\text{body}}$). This normalisation takes account of the variation observed in the values of $E_{\text{body}}$ (see Figure 2.7(a)) and hence is useful for making comparisons between different structures. In general we denote the flexibility fraction of an inhibited structure as $\Phi(\text{I})$ and the flexibility fraction of a structure with the inhibitor deleted as $\Phi(\text{U})$. The influence of the inhibitor on the flap rigidity is measured as the difference between them and is denoted $\Delta\Phi$, so that $\Delta\Phi = \Phi(\text{I}) - \Phi(\text{U})$. Since $\Phi(\text{I}), \Phi(\text{U}) \in [0, 1]$, we have that $\Delta\Phi \in [-1, 1]$ but note that $\Delta\Phi < 0$ corresponds to

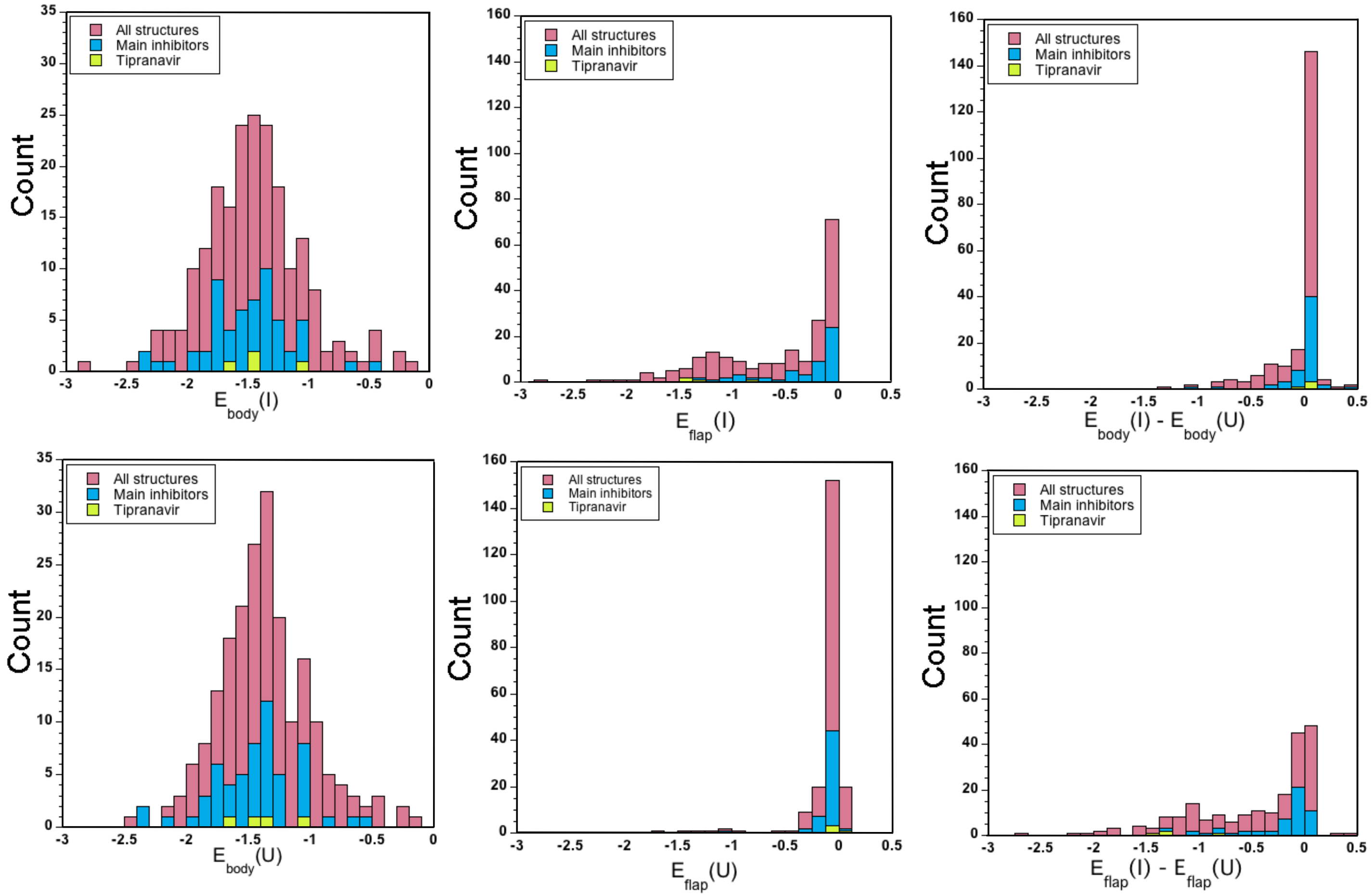


Figure 2.7: Distributions of $E_{\text{body}}$ and $E_{\text{flap}}$ values for 206 structures, coloured according to the inhibitor. Those complexed with TPV are shown in green, the other main inhibitors in blue, and the rest of the population in pink. Distributions of $E_{\text{body}}$ and $E_{\text{flap}}$ are given for structures with inhibitors present (a) and (c), and after the inhibitors have been removed (b) and (d). The distributions of differences for $E_{\text{body}}(\mathrm{I}) - E_{\text{body}}(\mathrm{U})$ are given in (e) and for $E_{\text{flap}}(\mathrm{I}) - E_{\text{flap}}(\mathrm{U})$ in (f).

the physically improbable situation where the flaps show enhanced rigidity after the deletion of the inhibitor. The other possible explanation for $\Delta\Phi < 0$ with small absolute value is that there is a greater variation between $E_{\text{body}}(I)$ and $E_{\text{body}}(U)$ than between $E_{\text{flap}}(I)$ and $E_{\text{flap}}(U)$, which can occur when $E_{\text{flap}}(I)$ is small. In the majority of cases, we observe that $\Delta\Phi \in [0, 1]$. Low positive values of $\Delta\Phi$ indicate that the presence of the inhibitor makes little difference to the flap rigidity, and high positive values indicate that the flaps are rigidified by the presence of the inhibitor. It has been previously noted that stabilising mutations can affect the rigidity analysis computed using FIRST [Tan and Rader, 2009]. The value of $\Phi$ is therefore influenced not only by the presence and identity of the inhibitor but also the presence of any stabilising mutations in the enzyme. However $\Phi$(I) and $\Phi$(U) are calculated each time from *the same structure* and so the effects of mutations are accounted for by considering the difference $\Delta\Phi$. It is worth noting that the choice of $E_{\text{body}}$ over the similarly defined folding core energy $E_{\text{fc}}$ does not substantially change the value of $\Delta\Phi$ or the conclusions drawn from these values. The $\Delta\Phi$ values for a selection of 20 structures, calculated using both our measure of $E_{\text{body}}$ and the analogous definition of $\Phi = \frac{E_{\text{flap}}}{E_{\text{fc}}}$ are given in Figure 2.8. This selection contains all of the structures that were crystallised with DRV or TPV, and were chosen since the two inhibitors have contrasting $\Delta\Phi$ values — something that becomes a focus of this analysis. For each structure presented the pair of $\Delta\Phi$ values are highly similar. The correlation between $\Delta\Phi$ ($E_{\text{fc}}$) and $\Delta\Phi$ ($E_{\text{body}}$) is shown, where linear regression yields an $R^2$ value of 0.998, indicating strong positive correlation. This demonstrates that our choice, $E_{\text{body}}$, does not affect our conclusions.

### 2.3.7 Complexes with darunavir and their $\Delta\Phi$ values

The measure of $\Delta\Phi$ has been presented and its use has been justified over other possible measures. Initially, we examine structures crystallised with the inhibitor darunavir, as was the case for 2HS1 and 3LZU, the rigidity dilutions of which were given in Figure 2.5. Darunavir is chosen as a starting point for presenting the results as it is a recently FDA approved protease inhibitor [Ghosh et al., 2007], and there are a relatively high number — sixteen — of high resolution structures of HIV-1 protease crystallised with darunavir in the PDB. The $\Phi$ values for all sixteen of these structures are presented in Table 2.1. The $\Phi$(I) values range from 0.031 to 0.468. Two of the values — for 3D1Z and 3LZU — are noticeably higher than the others. Discounting these, there is little variation and the remaining values have a mean of 0.074. There is similar variation in the $\Phi$(U) values, although as they are smaller than the $\Phi$(I), the variation itself is of a smaller magnitude. In almost every

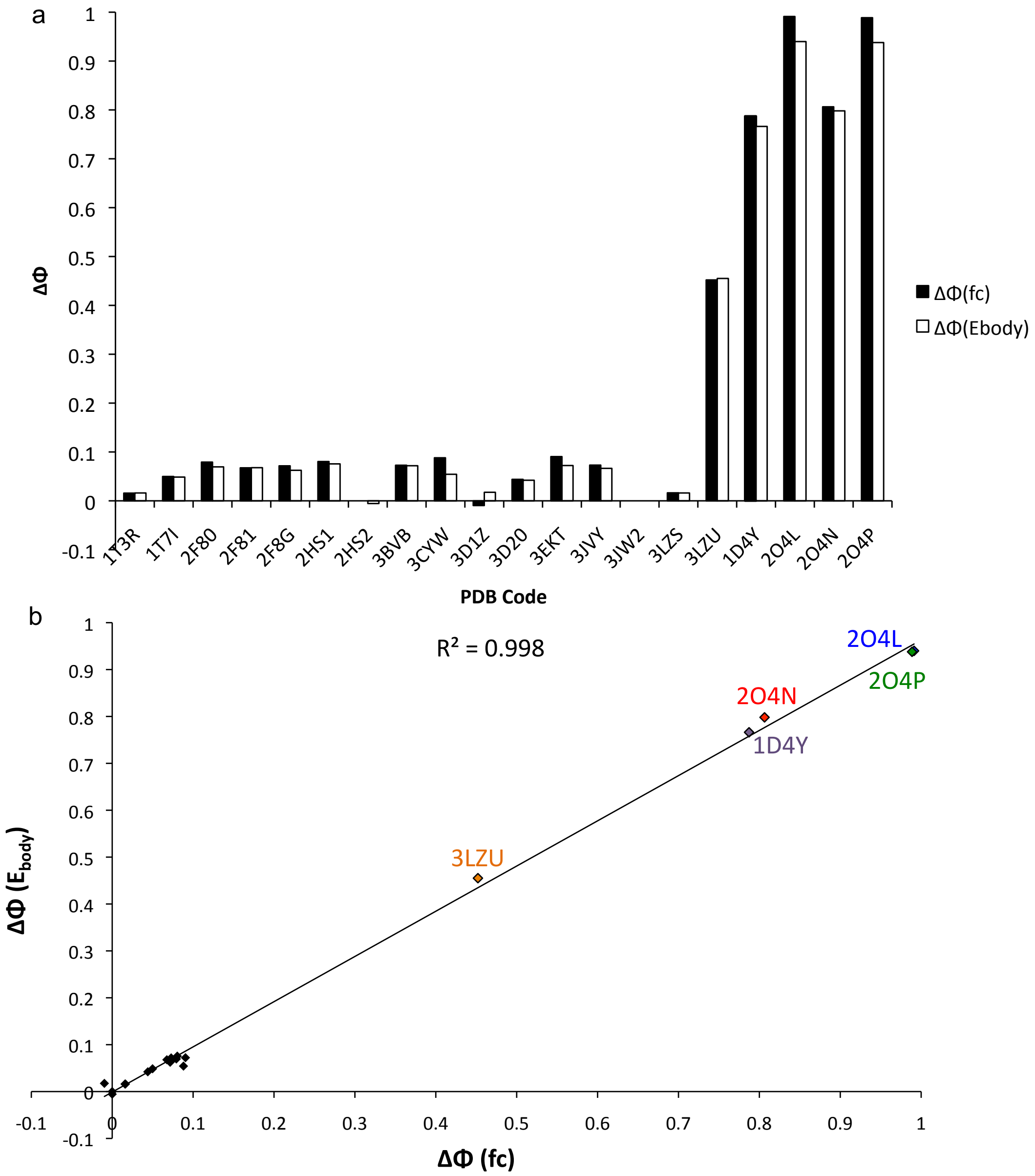


Figure 2.8: (a) The ΔΦ values for twenty structures crystallised with darunavir or tipranavir, calculated using ΔΦ ($E_{\text{body}}$) and ΔΦ(fc). The tipranavir structures have PDB codes 1D4Y, 2O4L, 2O4N and 2O4P. (b) The correlation of ΔΦ values calculated using the two methods. Linear regression gives an $R^2$ value of 0.998. The structures with the largest ΔΦ values — 3LZU, 1D4Y, 2O4L, 2O4N and 2O4P — are labelled.

Table 2.1: The values of Φ(I), Φ(U) and ΔΦ for 16 high resolution structures of HIV-1 protease crystallised under various conditions with the antiviral drug darunavir. There is no apparent correlation between ΔΦ and the location of mutations or the resolution of the structures.

| PDB code | Mutation | Resolution (Å) | Φ(I) | Φ(U) | ΔΦ |
|---|---|---|---|---|---|
| 1T3R | No mutation | 1.20 | 0.053 | 0.037 | 0.016 |
| 1T7I | No mutation | 1.35 | 0.102 | 0.053 | 0.049 |
| 2F80 | Core | 1.45 | 0.083 | 0.014 | 0.069 |
| 2F81 | Terminal | 1.25 | 0.076 | 0.008 | 0.068 |
| 2F8G | Flaps | 1.22 | 0.071 | 0.009 | 0.062 |
| 2HS1 | Core | 0.84 | 0.140 | 0.065 | 0.075 |
| 2HS2 | Flaps | 1.22 | 0.064 | 0.070 | -0.006 |
| 3BVB | Active-site | 1.30 | 0.083 | 0.011 | 0.072 |
| 3CYW | Flaps | 1.40 | 0.068 | 0.013 | 0.055 |
| 3D1Z | Flaps | 1.30 | 0.250 | 0.232 | 0.018 |
| 3D20 | Flaps | 1.05 | 0.053 | 0.011 | 0.042 |
| 3EKT | Multiple sites | 1.97 | 0.093 | 0.020 | 0.073 |
| 3JVY | Terminal | 1.60 | 0.079 | 0.013 | 0.066 |
| 3JW2 | Terminal | 1.80 | 0.034 | 0.034 | 0.000 |
| 3LZS | Terminal | 1.95 | 0.031 | 0.015 | 0.016 |
| 3LZU | Terminal | 1.76 | 0.468 | 0.013 | 0.455 |

case we have that $0 \leq \Delta\Phi \leq 0.075$, suggesting that darunavir has in general only a small rigidifying effect on the flap tips. 2HS2 is different in that $\Delta\Phi < 0$, but the magnitude of the negative value is so small that the effect of the inhibitor on this structure is minimal, similar to the structure 3JW2. Contrastingly in 3LZU, $\Delta\Phi = 0.455$ and the inhibitor appears to have a sizeable rigidifying effect, as can be seen by comparing Figures 2.5(c) and 2.5(d). The differences in values of ΔΦ underline the fact that different crystallisation conditions can affect the rigidity of the enzyme-inhibitor complex.

### 2.3.8 The effect of inhibitors on flap flexibility

The principal results of this chapter are given in Figure 2.9. We review the distributions of Φ(I), Φ(U) and ΔΦ for the full dataset of 206 structures and for our subset of main inhibitors. Figures 2.9(a), 2.9(c) and 2.9(e) show the distributions of Φ(I), Φ(U) and ΔΦ respectively. We find that the distributions for the set of main inhibitors are similar to the distributions for all structures. Figure 2.9(a) shows that, unlike the distribution of $E_{\text{body}}$ values, the distribution of Φ(I) values is bimodal; there are peaks near values of zero, indicating that the flaps are flexible from near the beginning of the rigidity dilution, and also near 1, indicating that the flaps form

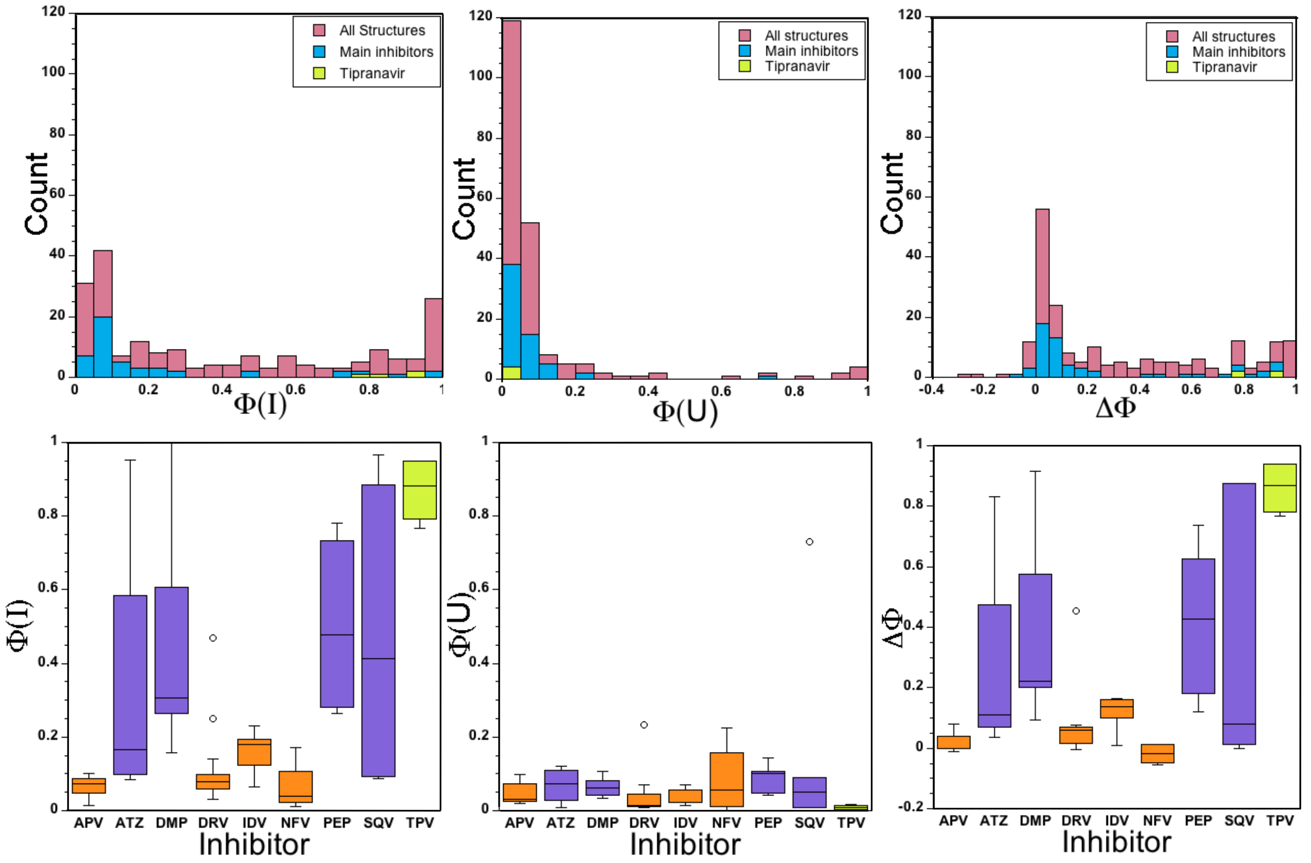


Figure 2.9: Population distribution of flexibility fractions. Histograms for the entire population of structures are shown for $\Phi(I)$, $\Phi(U)$ and $\Delta\Phi$ in (a), (c) and (e) respectively. Colours are as in Figure 2.7. Box plots of the distributions for the main inhibitors are shown in (b) $\Phi(I)$, (d) $\Phi(U)$ and (f) $\Delta\Phi$. Inhibitors with a small effect on flap rigidity are shown in orange and those with generally larger effect in purple. Structures inhibited by tipranavir are shown in green. Note that for each box, the horizontal line in the box corresponds to the median value and the height of the box denotes the interquartile range. Error bars correspond to maximum and minimum values, with circles indicating extreme outliers.

an integral part of the main rigid body of the dimer. In contrast, the distribution of $\Phi(\mathrm{U})$ has a single peak at near-zero values, seen in Figure 2.9(c). After deletion of the inhibitor from the structure it is much more common for the flaps to be flexible relative to the rest of the protein at high $E_{\mathrm{cut}}$ values. The distribution of $\Delta\Phi$ values in 2.9(e) confirms this impression. We see a broad and bimodal distribution of values with peaks around zero (minimal effect on flap rigidity) and around values near 1 – indicating a strong effect on flap rigidity. The bimodality of the distribution suggests that inhibitors might fall into different classes- those which typically affect flap rigidity and those which do not.

We now consider our set of main inhibitors more carefully. The distributions of $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ for these inhibitors are shown using box plots in Figures 2.9(b), 2.9(d) and 2.9(f), respectively. The box plots demonstrate the range of values observed for each inhibitor by indicating the total range of the distribution as well as the median and the inter-quartile range. In Figure 2.9(b), we see that the inhibitors can be divided into three groups based upon the spread of the $\Phi(\mathrm{I})$ values. APV, DRV, IDV and NFV all have consistently low $\Phi(\mathrm{I})$ values, indicating that the flaps are typically flexible in structures crystallised with these inhibitors. ATZ, DMP, PEP and SQV have a wide spread of values, with some of the individual structures crystallised with these inhibitors having much more rigid flaps. Tipranavir (TPV) is distinctive in that the flaps are consistently highly rigid in structures crystallised with this inhibitor. The groups have been colour coded in Figure 2.9(b) according to their different rigidifying effects. The colour coding is maintained in Figures 2.9(d) and 2.9(f). Figure 2.9(d) shows that $\Phi(\mathrm{U})$ values are low across all of the main inhibitors including tipranavir. That is, with the inhibitors deleted, all of the structures now have flexible flaps. This is evidence that the variations in flap rigidity observed in the $\Phi(\mathrm{I})$ values are attributable to the inhibitors, rather than being coincidental effects of the conditions of crystallisation. The distribution of $\Delta\Phi$ values in Figure 2.9(f) confirms this impression.

### 2.3.9 Inhibitor binding and inhibition mechanisms

Tipranavir is distinctive in that $\Delta\Phi$ is consistently high for structures inhibited by tipranavir. The $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ values for these structures are given in Table 2.2. Reviewing the distributions of $E_{\mathrm{body}}$ values in Figure 2.7, where the tipranavir structures are highlighted, demonstrates that tipranavir is not distinctive in its effect on the overall rigidity of the protease. Thus tipranavir appears to act specifically by rigidifying the flaps above the active site. This effect is not due to tipranavir forming an unusually large number of hydrogen bonds, or indeed

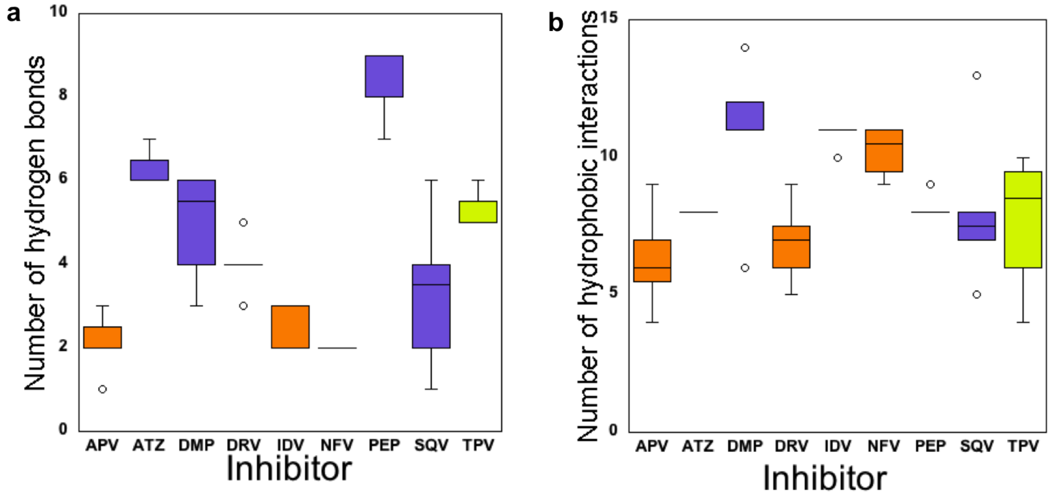


Figure 2.10: The distributions of the number of (a) hydrogen bonds and (b) hydrophobic interactions between the different inhibitors and the protease are shown. The groups are coloured as in Figure 2.9.

hydrophobic interactions, with the protease. In Figure 2.10, we show distributions of the number of bonds formed by inhibitors. Inhibitors with a large effect on flap rigidity have a slight tendency to form more hydrogen bonds than inhibitors with a small effect; the number of hydrophobic interactions shows little difference between the two classes. Tipranavir is not distinctive in either case, suggesting that the number of interactions is not the driving force for its large rigidifying effect on the flaps. The value of rigidity analysis is demonstrated by its sensitivity not only to the number but also the location of constraint formation. Examining the interactions identified by FIRST between tipranavir and the protease using PYMOL shows that tipranavir forms hydrogen bonds with the isoleucine residues 50 and 150, and not the residues of the active site. In contrast, other inhibitors such as darunavir typically bond with the active site and not with the flap tip residues. The difference between these two types of interaction can be seen in Figure 2.11.

Rigidity analysis indicates two different modes of action for HIV-1 protease inhibitors: direct interaction with the active site, or interaction with the flaps to rigidify them in a closed conformation and thus indirectly prevent access to the active site.

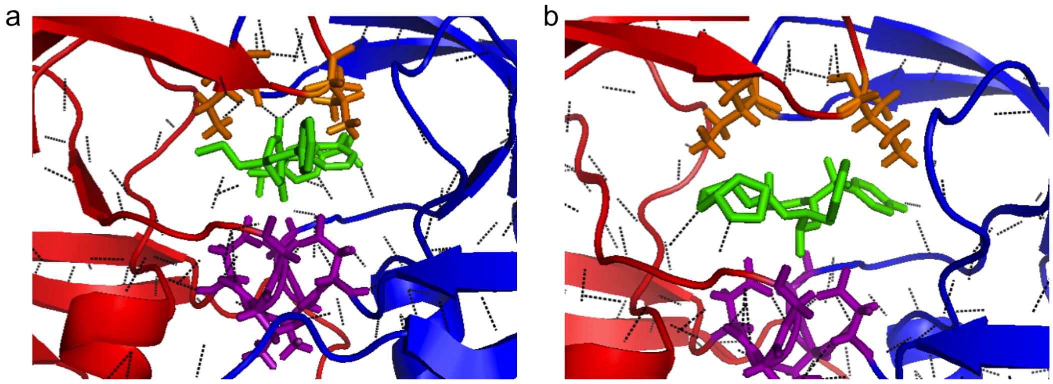


Figure 2.11: Tipranavir binds directly to the flap tips. The hydrogen bond interactions in structures complexed with (a) tipranavir and (b) darunavir. We illustrate the secondary structure motifs of the protein monomers in red and blue, the active site in purple and the flap tip residues in orange. The heavy atoms of the inhibitor are shown in green. The active site, flap tips and the inhibitor are shown as sticks. Each hydrogen bond is shown as a dashed black line.

### 2.3.10 The $\Delta\Phi$ value is not the direct result of backbone mutation

The mutations in the protease sequence are accounted for by the way we calculate $\Delta\Phi$. To verify this further, we consider examples where structures have the same mutation but different inhibitors. The pair 2O4P and 3LZS are crystallised with TPV and DRV respectively, and have the same single mutation, Q7K. Their respective $\Delta\Phi$ values are 0.9376 and 0.0160. Another concern is that the common stabilising flap-tip mutation I50V, which is present in the TPV structure 2O4L ($\Delta\Phi$ = 0.9399), may cause enhanced flap rigidity. This mutation also occurs in other structures with different inhibitors, however, such as 2AVS (IDV), 2F8G (DRV) and 3NU5 (APV). The $\Delta\Phi$ values for these structures are 0.1369, 0.0625 and 0.0450, respectively. These values are listed in Table B.2.

### 2.3.11 Analysis of structures crystallised with ritonavir

The HIV-1 protease inhibitor ritonavir (RTV) became FDA-approved in 1996 and is often administered in combination with other protease inhibitors during antiretroviral therapy. Such combination therapies form important parts of antiretroviral therapies [Hicks et al., 2006; Cane, 2009]. Consequently there is a need for determining sets of drugs which will work well in combination. A criterion for effective inhibitor combination may be to employ one inhibitor from each of the classes, so as to target the enzyme in two different ways. The complementary approach of using one inhibitor which targets the flaps of the protease and one which does not may account for the efficacy of the tipranavir-ritonavir combination [Hicks et al., 2006]. That the effects of TPV and RTV on the rigidity of the protease are indeed complementary is demonstrated in the results of Table 2.2. There is a small selection of HIV-1 protease structures in the PDB which have been crystallised in complex with RTV. Although some of these structures have a lower resolution than 2 Å, we have performed a rigidity analysis and determined that they have consistently low $\Delta\Phi$ values, that is, ritonavir does not rigidify the flaps. Figure 2.12 shows rigidity dilution plots for the crystal structures with PDB codes 1HXW and 1SH9 along with their uninhibited copies 1HXW (U) and 1SH9 (U) made by deletion of an inhibitor. The figure is analogous to Figure 2.4. 1HXV was crystallised with the RTV at a resolution of 1.8 Å; 1SH9 was crystallised with the same inhibitor at 2.5 Å — a lower resolution than our cutoff of 2 Å. The rigidity dilution plots are qualitatively consistent with each other and with those presented in Figure 2.4: they show the key features of first order rigidity loss and of the flap regions becoming flexible before the main body of the protein. The dilution plot for 1SH9 is typical of the poor resolution (i.e. worse than 2 Å) structures. This demonstrates that the restriction to high resolution is purely for the sake of precision — accurate geometry leads to an accurate bond network in FIRST — and not to shield any unwanted artefacts which arise in the results of rigidity analysis on the lower resolution structures. Table 2.2 gives a comparison of the rigidity analysis of crystal structures of the protease complexed with RTV and TPV, by displaying the $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ for each available structure of HIV-1 protease crystallised with these ligands. From the $\Delta\Phi$ values in the table it is clear that TPV has a consistently large effect upon flap rigidity, where RTV does not. The $\Phi(\mathrm{U})$ values for the poor resolution RTV structures are also consistent with the $\Phi(\mathrm{U})$ of other high resolution structures, the results for all of which are given in Appendix B.3.

## 2.4 Discussion

The principal finding is that inhibitors of HIV-1 protease fall into two classes which differ in their effects on the rigidity of the $\beta$-hairpin flaps: those which restrict the flexibility and those which do not. Nine inhibitors, including seven FDA-approved retroviral drugs were studied in detail and of those which do restrict the flaps, tipranavir stands out as being consistently effective for this purpose. Low resolution structures of the enzyme in complex with ritonavir were also studied. Rigidity analysis presents a way of studying the interactions an inhibitor creates with the enzyme, and hence can offer a new perspective during the process of drug design.

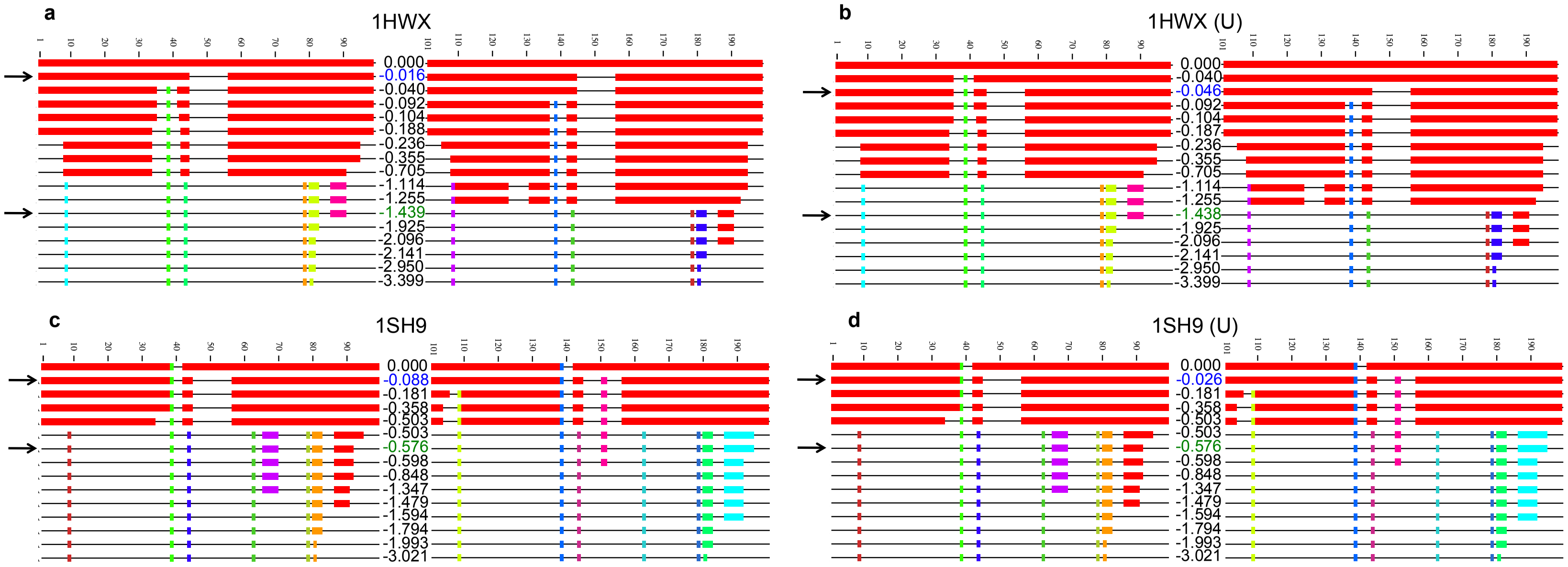


Figure 2.12: Rigidity dilution plots are shown for the structures 1HXW and 1SH9 along with their uninhibited copies 1HXW (U) and 1SH9 (U). The $E_{\text{flap}}$ and $E_{\text{body}}$ values are labelled blue and green respectively and indicated in each figure by arrows.

Table 2.2: The values of $\Phi(\mathrm{I})$, $\Phi(\mathrm{U})$ and $\Delta\Phi$ for 4 structures of HIV-1 protease crystallised under various conditions with TPV and RTV.

| PDB code | Inhibitor | Mutation | Resolution (Å) | $\Phi(\mathrm{I})$ | $\Phi(\mathrm{U})$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 1D4Y | TPV | Multiple sites | 1.97 | 0.766 | 0.000 | 0.766 |
| 2O4L | TPV | Multiple sites | 1.33 | 0.949 | 0.009 | 0.940 |
| 2O4N | TPV | Multiple sites | 2.00 | 0.816 | 0.018 | 0.798 |
| 2O4P | TPV | Terminal | 1.80 | 0.949 | 0.011 | 0.938 |
| 1HXW | RTV | No mutation | 1.80 | 0.032 | 0.011 | 0.021 |
| 1N49 | RTV | Multiple sites | 2.20 | 0.120 | 0.038 | 0.082 |
| 1RL8 | RTV | Core | 2.00 | 0.005 | 0.024 | -0.019 |
| 1SH9 | RTV | Multiple sites | 2.50 | 0.153 | 0.045 | 0.108 |
| 2B60 | RTV | Multiple sites | 2.20 | 0.066 | 0.066 | 0.000 |

An extensive comparative rigidity analysis on a set of 206 high-resolution X-ray crystal structures of HIV-1 protease in complex with a variety of inhibitors has been presented. For each structure, rigidity dilution using FIRST was performed with the inhibitor present and again with the inhibitor deleted. It has been possible to describe the effect of inhibitor binding on the rigidity of the protein using measures of overall protein rigidity and of the rigidity of the $\beta$-hairpin flaps above the active site. HIV-1 protease in solution can have the flaps either open or closed. It is a prediction of this work that inhibitors which rigidify the flaps will bias the structure towards having the flaps closed, and that this bias will be weaker for inhibitors that do not rigidify the flaps. The existence of open-flap and closed-flap conformations could be probed experimentally using spectroscopic methods such as fluorescence spectroscopy. We confirm the finding [Wells et al., 2009] that HIV-1 protease shows an abrupt or first-order loss of rigidity. The first-order loss is preserved regardless of the presence or identity of the inhibitor. The presence or absence of an inhibitor does not typically have a significant effect on the energy cutoff at which rigidity is lost during the dilution. The active site of the protease is not distinctive in the rigidity analysis, typically forming part of the main rigid cluster until overall rigidity is lost. In contrast, the $\beta$-hairpin flaps are notable as they typically become flexible at higher $E_{\mathrm{cut}}$ than the main body of the protein. This effect is observed across the population of X-ray crystal structures. The flexible nature of the flaps is in qualitative agreement with MD simulations as well as experimental work using NMR [Nicholson et al., 1995; Zhu et al., 2003; Hornak et al., 2006]. The flaps are usually considered to have high conformational flexibility – due in part to their high glycine content [Freedberg et al., 2001]. In MD simulations, the presence of glycine is conducive to flexibility as its minimal R group allows wider variation in backbone dihedral angles than for other amino acids. Here we find that glycine is also conducive to flexibility in *rigidity analysis*, but for a different reason: it is not capable of forming side-chain interactions, either hydrogen bonds or hydrophobic tethers, and so tends to introduce fewer constraints into the system.

The flexibility fraction is a normalised measure of the rigidity of the flaps compared to the main body of the protein. We find that the effect of inhibitor binding on flap flexibility varies greatly, from zero/minimal to large rigidifying effect. Detailed examination of a set of inhibitors including FDA-approved antiretroviral medications shows that inhibitors can be divided into two classes: those which typically do not rigidify the flaps, and those which do. The rigidifying effect of the inhibitors is not a direct consequence of the number of hydrogen bonds or hydrophobic interactions that are formed with the protease, but rather the location of the interactions. Tipranavir is distinctive for the strength and consistency of its effect in rigidifying the flaps. The flexibility of the flaps could serve to permit the flexibility-assisted catalysis mechanism [Piana et al., 2002] and thus tipranavir may be acting to combat this mechanism. Tipranavir has a unique structure among the nine main inhibitors. It is the first true non-peptidic HIV-1 protease inhibitor and has shown considerable activity against many drug-resistant strains of HIV [Pokorná et al., 2009]. Its flexibility is not hampered by the presence of a rigid peptide bond. Flexibility is a property which increases the chance of a drug being able to resist some mutations in the protease [Velazquez-Campoy et al., 2001]. Darunavir was also designed to be adaptable to drug resistant strains [Ghosh et al., 2007] but rigidity analysis indicates that tipranavir and darunavir have quite different modes of action. Tipranavir eschews interaction with the active site while forming direct hydrogen-bond interactions with the residues at the tips of the flaps. This suggests that its mode of action is quite distinct from that of inhibitors such as darunavir, which interact with the active site rather than the flaps. The combination of one protease inhibitor from each class appears to be effective in current multi-drug therapies [Hicks et al., 2006] and should be borne in mind in the selection of future multi-drug therapies.

Throughout this work on HIV-1 protease, we did not include the flap water molecule that is conserved in many crystal structures of the protease [Wlodawer and Vondrasek, 1998; Piana et al., 2002]. The possible impact of this water molecule on the rigidity of the flaps and the flexibility fraction has not been considered. A further caveat is that we have not studied how the protonation state of the active site aspartate residues may affect the impact of inhibitor binding on protein rigidity.

It is possible that these considerations may alter the values of $\Delta\Phi$ and thus our categorisation of the protease inhibitors. As such, this is certainly an avenue for further work.

Subsequent to publication, the results from this chapter were discussed in a study designed to predict binding affinities for a series of protein-ligand interactions [Greenidge et al., 2013]. A medium-throughput hybrid computational approach was used to calculate protein-ligand binding energies from over 1000 X-ray crystal structures of protein-ligand complexes. The resulting energies are compared favourably with experiment. A number of the structures examined were HIV-1 protease complexes which we have examined in this chapter. Notably, the structures with low $\Delta\Phi$ values yield poor estimates of binding affinity and those with high $\Delta\Phi$ values estimate the binding affinities well. Inhibitors which rigidify the flaps bias the structure to the closed conformation; the bias is smaller where the inhibitors do not rigidify the flaps. Thus the inhibitors with low $\Delta\Phi$ values do not rigidify the flaps and thus pay a larger enthalpic cost to close the flaps. That this was not accounted for in the calculations resulted in the poor estimation of binding energies for the inhibitors with low $\Delta\Phi$. The results from this chapter have thus already been used to inform another study on protein-ligand interactions.

# Chapter 3

# Rigidity analysis of CypA

## 3.1 Introduction

Following on from the rigidity analysis study on HIV-1 protease in Chapter 2 we apply a similar approach to CypA, which was introduced in Section 1.3. We use computational and experimental approaches to study the effects of ligand binding on the rigidity and mobility of CypA. In this chapter, we focus on rigidity analysis, and study 59 structures of CypA crystallised under a variety of conditions. We apply standard measures to extract information from the rigidity simulations about the effects of ligand binding. In addition we create a rigidity difference map (RDM) which is introduced to show the effects of ligand binding at a range of $E_{\text{cut}}$ values. We use a cumulative RDM to display this information for a number of structures crystallised under similar conditions.

### 3.1.1 The structure of CypA

Human CypA is a 165 residue, 18 kDa protein consisting of antiparallel $\beta$-strands arranged in a $\beta$-barrel with $\alpha$-helices covering the top and bottom of the barrel [Ke and Huai, 2004]. CypA is approximately 13 % helix and 32 % $\beta$-sheet. The PDB structure 3K0N of CypA crystallised without a ligand is shown in Figure 3.1. The secondary structure at the top of the figure was taken from the PDB and was determined using the DSSP algorithm. There are five bridges identified along the backbone, in addition to the eight extended $\beta$-strands. There is also a $3_{10}$ helix formed by repeated $i, i+3$ bonding. The binding site is the cleft in the top left of Figure 3.1, although 3K0N was crystallised with no ligand bound.

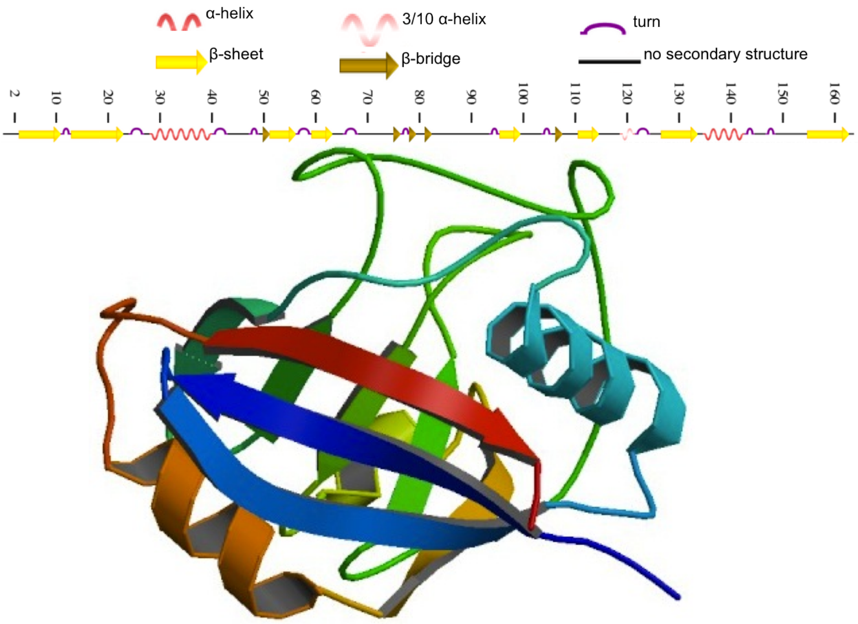


Figure 3.1: A cartoon representation of the structure 3K0N, with the annotated secondary structure as determined by the DSSP given above. The structure consists of a $\beta$-sheet barrel with $\alpha$-helices at either end.

### 3.1.2 X-ray crystal structures of CypA.

We study 59 X-ray crystal structures, listed in Table 3.1. As with HIV-1 protease, the structures were crystallised under various conditions with different ligands bound. The majority of the ligands are CsA, part of the HIV-1 capsid protein CA, or analogs of these ligands. 18 of the structures were crystallised with CsA or analogs thereof and one with CsA and CN. There are 15 structures which were crystallised either with small peptide sequences from CA or with CA itself. Ten were crystallised without ligands bound, 18 with metallic ions or other ligands bound to the protein. Many contain mutations. The resolution of the structures ranges from 1.25 Å to 3.10 Å; the quality of the resolution affects the accuracy of results from FIRST, as discussed in Section 2.2. As before, all these differences were accounted for by studying each structure as it was crystallised as well as after manual removal of the ligands if appropriate. In this way it was possible to analyse the effect of inhibitor binding on the structure in question. Table 3.1 lists all CypA structures in the database, grouped according to the ligand.

Table 3.1: PDB structures of cyclophilin A with the name or a brief description of the ligand and the resolution.

| PDB ID | Ligand | Resolution (Å) |
|---|---|---|
| 1W8V | No ligand | 1.70 |
| 2CPL | No ligand | 1.63 |
| 3K0N | No ligand | 1.39 |
| 3K0O | No ligand, S99T mutation | 1.55 |
| 3K0M | No ligand, cryogenic structure | 1.25 |
| 3K0Q | No ligand, cryogenic structure, S99T mutation | 2.32 |
| 3K0P | No ligand, cryogenic structure, S99T mutation | 1.65 |
| 3K0R | No ligand, cryogenic structure, R55K mutation | 2.42 |
| 2X25 | No ligand, acetylated CypA | 1.20 |
| 2X2A | No ligand, acetylated CypA | 1.40 |
| 1CWA | CsA | 2.10 |
| 2RMA | CsA | 2.10 |
| 2RMB | MEBM2T1-CsA | 2.10 |
| 1CWB | MEBM2T1-CsA | 2.20 |
| 1BCK | 2-Thr CsA | 1.80 |
| 1CWH | 3-D-Ser CsA | 1.86 |
| 1CWF | 2-Val CsA | 1.86 |
| 1CWJ | 2-Val 3-S-methyl-sarcosine CsA | 1.80 |
| 1CWI | 2-Val 3-(N-methyl)-D-Ala CsA | 1.90 |
| 1CWM | 4 meile CsA | 2.00 |
| 1CWL | 4-hydroxy-meleu CsA | 1.80 |
| 1MIK | (5-hydroxynorvaline)-2-CsA | 1.76 |
| 3ODL | Voclosporin Z-ISA247 (CsA analogue) | 2.31 |
| 3ODI | Voclosporin E-ISA247 (CsA analogue) | 2.20 |
| 1CWK | 1-(6,7-dihydro)MEBMT 2-Val 3-D-(2-S-methyl)sarcosine CsA | 1.80 |
| 1CWC | [4,N-dimethylnorleucine]4-cyclosporin | 1.86 |
| 1CWO | Thr2, Leu5, D-HIV8, Leu10 CsA | 1.86 |
| 2X2C | CsA, acetylated CypA | 2.41 |
| 1MF8 | CsA and CN | 3.10 |
| 1AK4 | CA | 2.36 |
| 1M9D | CA, 7 mutations | 1.90 |
| 1M9C | CA N-terminal domain | 2.00 |

| PDB ID | Ligand | Resolution (Å) |
|---|---|---|
| 1M9X | CA N-terminal domain | 1.70 |
| 1M9Y | CA N-terminal domain, H87A, G89A mutations | 1.90 |
| 1M9F | CA N-terminal domain, H87A, A88M mutations | 1.73 |
| 1M9E | CA N-terminal domain, H87A mutation | 1.72 |
| 2X2D | CA N-terminal, acetylated CypA | 1.95 |
| 1FGL | 25 residue sequence from CA | 1.80 |
| 1AWQ | 87HAGPIA92 sequence from CA | 1.58 |
| 1AWR | 87HAGPIA92 sequence from CA | 1.58 |
| 1AWS | 87HAGPIA92 sequence from CA | 2.55 |
| 1AWT | 87HAGPIA92 sequence from CA | 2.55 |
| 1AWU | 87HVGPIA92 sequence from CA | 2.34 |
| 1AWV | 87HVGPIA92 sequence from CA | 2.34 |
| 2CK1 | Acetate ion | 1.80 |
| 2ALF | Magnesium ion | 1.90 |
| 2HQJ | Phosphate ion | 2.00 |
| 3CYH | Dipeptide Ser-Pro | 1.90 |
| 4CYH | Dipeptide His-Pro | 2.10 |
| 5CYH | Dipeptide Gly-Pro | 2.10 |
| 2CYH | Dipeptide Ala-Pro | 1.64 |
| 1VDN | (ACE)AAPA(MCM) | 1.60 |
| 1VBT | Sulfur-substituted AAPF | 2.30 |
| 1LOP | Suc-APA-pNA | 1.80 |
| 1RMH | Suc-AAPF-pNA | 2.40 |
| 1ZKF | Suc-AGPF-pNA | 2.55 |
| 1YND | Sanglifehrin A | 1.60 |
| 1W8M | Ethylpiperidine glyoxylate | 1.65 |
| 1W8L | 1-Acetyl-3-methylpiperidine | 1.80 |

### 3.1.3 Structures of CypA with CsA or CsA analogs

The 13 structures listed in Table 3.2 were crystallised under similar conditions and with CsA or CsA derivatives. We summarise the crystallographic data for the structures, taken from the PDB and the corresponding publications [Papageorgiou et al., 1994; Mikol et al., 1995; Kallen et al., 1998; Mikol et al., 1998]. One of these publications reports on eleven of these structures, stating that all were crystallised at the same pH of 8.0 and notes the lack of conformational difference between the unbound and the liganded forms of CypA [Kallen et al., 1998]. For each structure

Table 3.2: Structural data for 13 CypA PDB entries crystallised with CsA or CsA-derivatives, adapted from [Kallen et al., 1998]. In each case, the resolution of the structure is given along with the R-factor and average B-factors for structural components. The bond length in Å (L) and angle in degrees (A) RMSD from ideality are also listed. There is no R-factor indicated in the PDB for the structure 1CWO.

| PDB code | Resolution (Å) | R-factor | Average B-factor ($Å^2$) | | | Bond RMSD | |
|---|---|---|---|---|---|---|---|
| | | | CypA | Ligand | Solvent | L | A |
| 1BCK | 1.80 | 0.186 | 20.1 | 24.7 | 55.8 | 0.060 | 1.88 |
| 1CWA | 2.10 | 0.167 | 13.0 | 15.9 | 37.2 | 0.013 | 2.64 |
| 1CWB | 2.20 | 0.163 | 13.4 | 20.3 | 37.5 | 0.031 | 2.71 |
| 1CWC | 1.86 | 0.177 | 14.4 | 28.5 | 32.9 | 0.010 | 1.89 |
| 1CWF | 1.86 | 0.177 | 15.1 | 16.9 | 32.9 | 0.012 | 1.59 |
| 1CWH | 2.00 | 0.194 | 28.7 | 24.0 | 45.7 | 0.017 | 2.94 |
| 1CWI | 1.90 | 0.186 | 13.3 | 15.5 | 35.2 | 0.015 | 2.76 |
| 1CWJ | 1.80 | 0.185 | 15.3 | 16.5 | 36.6 | 0.014 | 2.78 |
| 1CWK | 1.80 | 0.165 | 16.4 | 18.5 | 40.7 | 0.014 | 2.64 |
| 1CWL | 1.80 | 0.163 | 14.2 | 15.8 | 36.0 | 0.015 | 2.50 |
| 1CWM | 2.00 | 0.169 | 15.4 | 16.0 | 34.6 | 0.015 | 2.77 |
| 1CWO | 1.90 | N/A | 12.2 | 33.1 | 47.6 | 0.012 | 1.59 |
| 1MIK | 1.76 | 0.175 | 14.8 | 16.9 | 36.4 | 0.011 | 1.57 |

we list the resolution, R-factor, average B-factor, RMSD from ideal bond length in Å ($L$) and RMSD from ideal bond angle in degrees ($A$). The R-factor is a measure of agreement between the crystal model and the diffraction data. The magnitude of the B-factors indicate the relative vibrational motion of different parts of the structure. The bond RMSD from ideality is a measure of how much the bonds in the crystal structure deviate from their expected values, as measured in high resolution models of small molecules. Desirable values for protein crystal structures refined to 2.5 Å are R-value $< 0.02$, $L < 0.02$ Å and $A < 4°$ [Rhodes, 2013]. Each of the 13 structures satisfy these conditions. The similarity of these high quality structures means that variation in the rigidity analysis of the structures is primarily derived from the differences in the ligands. Differences between the liganded and unliganded structures which are consistent across the set suggest universal CsA-derived ligand-binding effects on the protein.

### 3.1.4 The CsA binding site

The binding site of CypA is shown in Figure 3.2. The binding site residues are shown in red, with the rest of the protein coloured green. The ligand CsA is shown in blue.

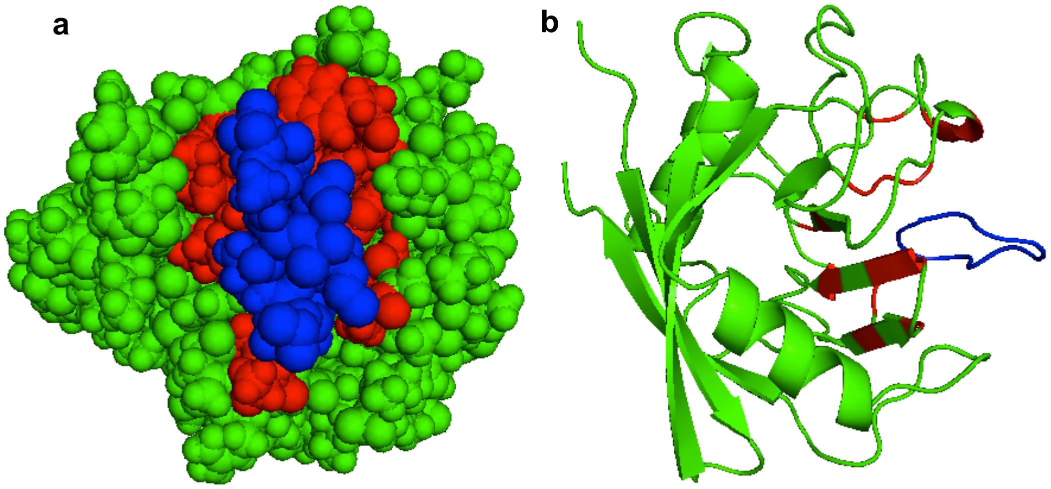


Figure 3.2: The CypA binding site. The structure 1CWA of the CypA-CsA complex is shown, with CsA in blue and CypA in green. The residues of CypA which have a heavy (non-hydrogen) atom within 4 Å of the atoms in CsA are coloured red. The atoms in (a) are shown as spheres and in (b) the protein backbone illustrated using the cartoon representation in PYMOL.

The binding site has been defined as in [Rader and Brown, 2011] for the structure 1CWA. The residues of the site are those which have a heavy (non-hydrogen) atom within 4 Å of the atoms in CsA. The 15 residues of the binding site defined this way are R55, I57, F60, M61, Q63, G72, T73, A101, N102, A103, Q111, F113, W121, L122 and H126.

### 3.1.5 Measures of global and local rigidity

The measures $X_{\mathrm{LRC}}$ and $Z_{\mathrm{LRC}}$ [Rader and Brown, 2011] have been used to determine, respectively, the global and local rigidity in a way which would be comparable across different structures or indeed different proteins. $X_{\mathrm{LRC}}$ is the fraction of atoms in the LRC,

$$X_{\mathrm{LRC}} = \frac{N_{\mathrm{LRC}}}{N}, \tag{3.1}$$

where $N_{\mathrm{LRC}}$ is the number of atoms in the LRC and $N$ is the total number of atoms in the structure. Folding [Tan and Rader, 2009] and stability [Rader et al., 2004] across protein structures have previously been studied using $X_{\mathrm{LRC}}$. We examine local rigidity around the binding site, $Z_{\mathrm{LRC}}$,

$$Z_{\mathrm{LRC}} = \frac{M_{\mathrm{LRC}}}{M}, \tag{3.2}$$

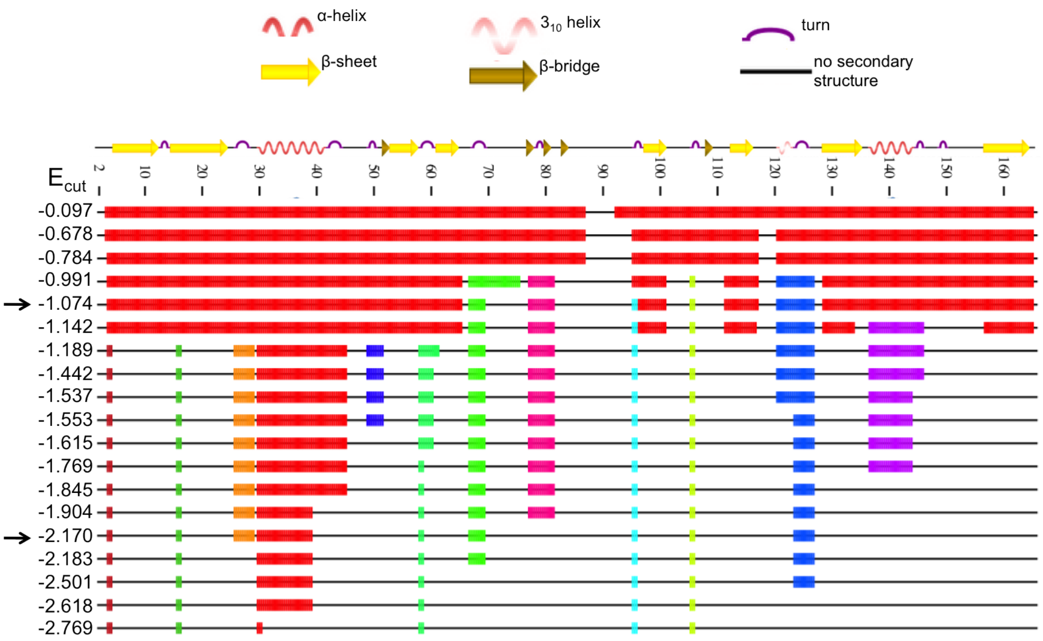
where $M_{\mathrm{LRC}}$ is the number of binding site residues belonging to the LRC and $M$ is the total number of residues in the binding site, so that $M = 15$ for CypA. In a study on allostery in proteins, $Z_{\mathrm{LRC}}$ was used to examine the rigidity of allosteric sites [Rader and Brown, 2011].

## 3.2 Results

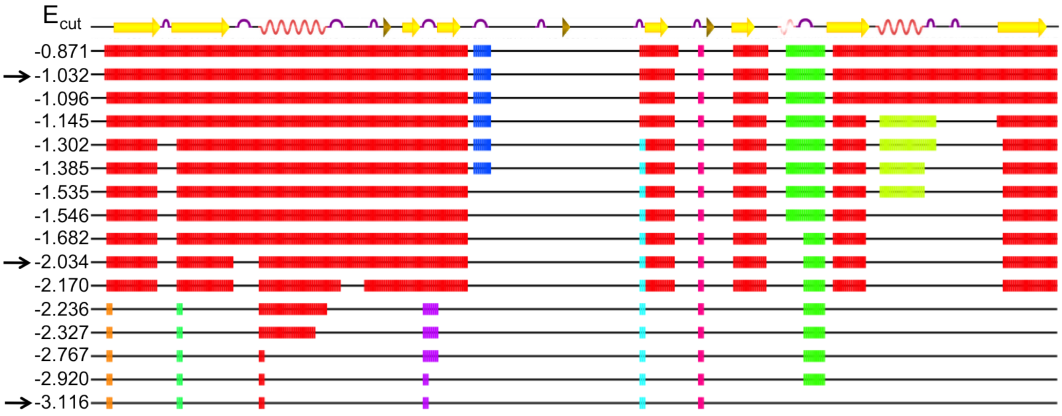
### 3.2.1 RD plots for structures of unbound CypA

Figure 3.3 shows RD plots for two structures crystallised with no ligand bound, 3K0N and 1W8V, of resolution 1.39 Å and 1.70 Å respectively. The secondary structures for each are shown above the RD plots. The mostly $\beta$-sheet protein exhibits first-order rigidity loss during dilution [Wells et al., 2009]. These two plots give us a baseline to use when comparing liganded-unliganded structure pairs, as differences in the plots are not due to ligands but rather the variation shown when comparing X-ray structures which should hypothetically exhibit the same pattern of rigidity loss. Indeed, the key features of the RD plots are in agreement, with the same regions becoming flexible relative to others at different stages of the dilution. In both plots we observe a first-order rigidity transition, at $-1.189$ kcal/mol in (a) and $-2.236$ kcal/mol in (b). The larger $\alpha$-helix maintains rigidity at lower $E_{\mathrm{cut}}$ values than the rest of the protein, with the $\beta$-sheets remaining rigid more than the unstructured regions. The secondary structures determined by the DSSP algorithm are different for residues 52-53 and 76-79. In 3K0N residues 52-53 form part of a $\beta$-sheet, but are identified as having no secondary structure in 1W8V. Residues 76-79 form a bridge–turn–bridge structure in 3K0N, whereas in 1W8V, only the bridge is identified. A small change in geometry resulted in different hydrogen bonds being determined by the DSSP and FIRST. There are more cross-linking hydrogen bonds in this region in 3K0N than in 1W8V, and these establish the enhanced rigidity.

In Figure 3.3 arrows indicate lines in the dilution plots where $|E_{\mathrm{cut}}|$ first exceeds 1, 2 and 3 kcal/mol, emphasising the non-linearity of the $E_{\mathrm{cut}}$ axis. In Figure 3.3(a), the arrows mark the 5th and 15th lines; in 3.3(b) they mark the 2nd, 10th and 16th lines. There is a difference in rigidity between the structures in the region between residues 70 and 100, a largely unstructured section of the protein. The non-linearity of the $E_{\mathrm{cut}}$ axes makes it difficult to determine the significance of the disparity. The rigidity of the structures in this region is largely the same when $E_{\mathrm{cut}} < -1.074$ kcal/mol. Striking differences between the plots do exist. For example, the LRC spans two non contiguous parts of the protein at $E_{\mathrm{cut}} = -2.170$ kcal/mol for 1W8V, but for 3K0N this is not the case for any



(a) 3K0N



(b) 1W8V

Figure 3.3: RD plots for (a) 3K0N and (b) 1W8V, both crystallised with no ligand bound. Black arrows indicate the lines in the dilution where $|E_{\text{cut}}|$ first exceeds 1, 2, and 3 kcal/mol. The secondary structures taken from the PDB entries are given above the RD plots.

$E_{\text{cut}} \leq -1.189$ kcal/mol. Two crystal structures of the same protein may exhibit different geometries and therefore engender different bond networks, which affect the more subtle aspects of the rigidity dilution plot. This underlines the inherent problem of forming conclusions based on the rigidity analysis of a single crystal structure.

### 3.2.2 The effect of ligand binding on RD plots

We generated RD plots for all of the structures in Table 3.1 before and after the ligands were removed. The RD plots for two of these structures, 1AWQ and 1AWS, are presented in Figure 3.4. From Table 3.1, 1AWQ and 1AWS have resolution 1.58 Å and 2.55 Å respectively and were both crystallised in complex with the peptide HAGPIA which forms residues 87-92 of CA. When we characterise the effect of ligand binding on the rigidity of a protein, we perform rigidity analysis on the same structure before and after the removal of the ligand. The alterations made to the bond network are therefore small and the impact on the mainchain rigidity may be subtle. This is what is seen for both 1AWQ and 1AWS. Figures 3.4(a) and 3.4(c) are largely the same, as are the two ligand-free plots, 3.4(b) and 3.4(d). There is not a large difference in the pattern of rigidity loss between the two initial structures, the two ligand-free structures, or the between either structure and its unliganded counterpart. The largest discrepancies occur when the structure is changed rather than when the ligand is removed. These structures and their corresponding RD plots are representative of the structures in Table 3.1. The differences in the RD plots for the ligand-unliganded pairs are outweighed by the impact of the structural variation in the dataset.

### 3.2.3 The impact of ligand binding on global and local rigidity

In our study on HIV-1 protease, the RD plots were sufficient to show consistent effects of ligand binding against the background noise of the different structures. This is mostly not the case with CypA. We therefore examined other measures to determine the effects of ligand binding on rigidity. We applied $X_{\text{LRC}}$ and $Z_{\text{LRC}}$ to the structures in Table 3.2 as well as to 1FGL, 1LOP, 1W8L, 4CYH and 5CYH, which were crystallised with different ligands. These 19 structures were chosen as they all had a single CypA-ligand complex in the crystallographic unit. In a previous study, the difference measure $\Delta X_{\text{LRC}} = X^{T}_{\text{LRC}} - X^{R}_{\text{LRC}}$ has previously been defined in order to be able to compare the active relaxed ($R$) and inactive tense ($T$) states of allosteric proteins [Rader and Brown, 2011]. We adapt this measure so as to focus

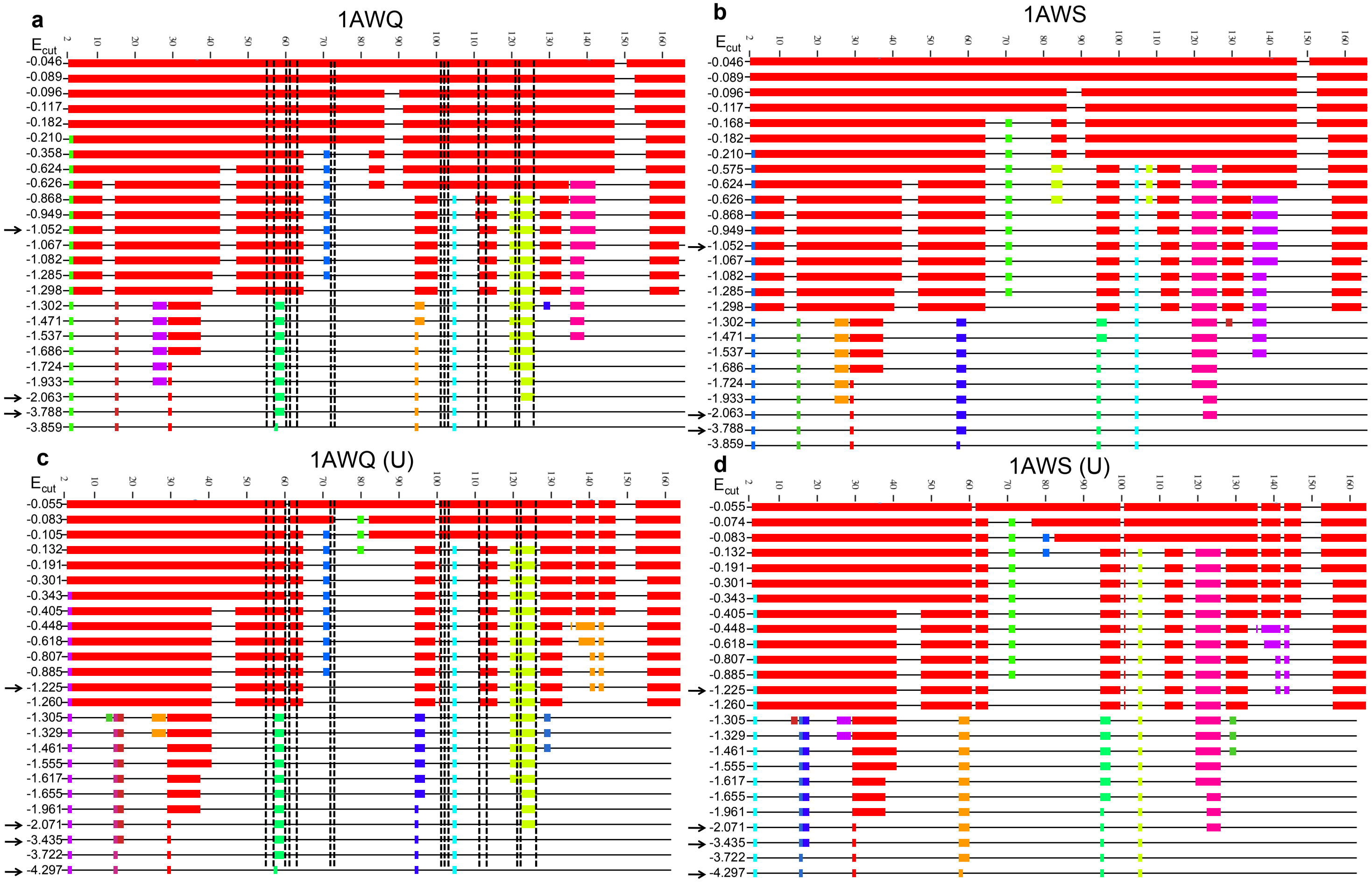


Figure 3.4: RD plots for the structures 1AWQ and 1AWS as crystallised and after the inhibitor removal. The unliganded form is indicated by the suffix (U). The units of $E_{\mathrm{cut}}$ are kcal/mol. Arrows indicate the lines in the dilution where $|E_{\mathrm{cut}}|$ first exceeds 1, 2, 3 and 4 kcal/mol. The dashed lines in (b) and (c) indicate the residues which form part of the ligand binding site.

Table 3.3: The values of $\Delta X_{\text{LRC}}$, $Z_{\text{LRC}}$, $Z_{\text{LRC}}(\text{U})$ and $\Delta Z_{\text{LRC}}$ for 19 structures of CypA, calculated at $E_{\text{cut}} = -1$ kcal/mol.

| PDB Code | $\Delta X_{\text{LRC}}$ | $Z_{\text{LRC}}$ | $Z_{\text{LRC}}(\text{U})$ | $\Delta Z_{\text{LRC}}$ |
|---|---|---|---|---|
| 1BCK | 0.013 | 0.533 | 0.533 | 0.000 |
| 1CWA | 0.020 | 0.533 | 0.467 | 0.067 |
| 1CWB | 0.000 | 0.467 | 0.467 | 0.000 |
| 1CWC | 0.093 | 0.667 | 0.467 | 0.200 |
| 1CWF | 0.267 | 1.000 | 0.467 | 0.533 |
| 1CWH | 0.014 | 0.467 | 0.467 | 0.000 |
| 1CWI | 0.090 | 0.867 | 0.467 | 0.400 |
| 1CWJ | 0.026 | 0.533 | 0.467 | 0.067 |
| 1CWK | 0.160 | 0.867 | 0.467 | 0.400 |
| 1CWL | 0.267 | 1.000 | 0.600 | 0.400 |
| 1CWM | 0.020 | 0.533 | 0.467 | 0.067 |
| 1CWO | 0.008 | 0.733 | 0.733 | 0.000 |
| 1FGL | 0.014 | 1.000 | 1.000 | 0.000 |
| 1LOP | 0.018 | 0.467 | 0.467 | 0.000 |
| 1MF8 | 0.031 | 0.000 | 0.000 | 0.000 |
| 1MIK | 0.094 | 0.667 | 0.467 | 0.200 |
| 1W8L | 0.131 | 1.000 | 1.000 | 0.000 |
| 4CYH | 0.007 | 0.867 | 0.867 | 0.000 |
| 5CYH | 0.011 | 1.000 | 1.000 | 0.000 |

on the impact of ligand binding in Equation (3.3). As in Chapter 2, (U) is used to designate the unliganded form of the structure.

$$\Delta X_{\text{LRC}} = X_{\text{LRC}} - X_{\text{LRC}}(U) \tag{3.3}$$

A positive $\Delta X_{\text{LRC}}$ value indicates an increase in flexibility upon the removal of the ligand. We define $\Delta Z_{\text{LRC}}$ analogously to $\Delta X_{\text{LRC}}$. The sign and magnitude of $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ show the effect of the ligand upon the global and local rigidity respectively. In Table 3.3, $\Delta X_{\text{LRC}}$, $Z_{\text{LRC}}$, $Z_{\text{LRC}}(\text{U})$ and $\Delta Z_{\text{LRC}}$ at are given for the 19 structures at $E_{\text{cut}} = -1.0$ kcal/mol.

In all cases, $\Delta X_{\text{LRC}} \geq 0$ and $\Delta Z_{\text{LRC}} \geq 0$. The range of values displayed in both of these columns is wide, with $0 \leq \Delta X_{\text{LRC}} \leq 0.27$ and $0 \leq \Delta Z_{\text{LRC}} \leq 0.54$. In the study on allosteric proteins, $E_{\text{cut}} = -1.0$ kcal/mol was used for calculations involving global and local rigidity [Rader and Brown, 2011]. However, since the LRC will decrease in size as $E_{\text{cut}}$ is lowered, $X_{\text{LRC}}$ and $Z_{\text{LRC}}$ (and therefore $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$) are functions of $E_{\text{cut}}$. The choice of $E_{\text{cut}} = -1.0$ appears arbitrary. We calculated $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ for the 13 structures in Table 3.2 for different $E_{\text{cut}}$ values. These are plotted in Figure 3.5. The $E_{\text{cut}}$ values used are $0.0, -0.5, -1.0, -1.5, -2.0, -2.5$ and $-3.0$ kcal/mol. For all of the structures, $X_{\text{LRC}} = 0$ and $Z_{\text{LRC}} = 0$ when $E_{\text{cut}} \leq -2$ kcal/mol. The choice of $E_{\text{cut}}$ has an effect on $\Delta X_{\text{LRC}}$ for all of the structures; for ten of the structures $\Delta Z_{\text{LRC}}$ is affected. In some of the structures, such as 1CWA and 1CWI, this effect is more dramatic than for others, such as 1CWJ and 1CWH. Comparisons between the structures change depending on $E_{\text{cut}}$. When $E_{\text{cut}} = -1$ kcal/mol, structures 1CWF and 1CWL have the largest $\Delta X_{\text{LRC}}$; when $E_{\text{cut}} = -1.5$ kcal/mol, $\Delta X_{\text{LRC}}$ is largest for 1CWA and 1CWI. 1CWA and 1CWL have the largest $\Delta X_{\text{LRC}}$ values at $E_{\text{cut}} = -0.5$ kcal/mol. The same issue arises for the local rigidity $\Delta Z_{\text{LRC}}$. At $E_{\text{cut}} = -1$ kcal/mol, 1CWF has large $\Delta Z_{\text{LRC}}$ value ($> 0.5$). At $E_{\text{cut}} = -0.5$ kcal/mol and $E_{\text{cut}} = -1.5$ kcal/mol, 1CWF has $\Delta Z_{\text{LRC}} = 0$. These observations make it clear that choosing a single $E_{\text{cut}}$ is unreliable. From Figure 3.5 we can state that the effect of ligand binding on the global and local rigidity appears greater for 1CWA, 1CWF, 1CWI, 1CWK, 1CWL and 1CWM than for 1BCK, 1CWB, 1CWC, 1CWH, 1CWJ, 1CWO and 1MIK. This is likely to be caused by the bonding patterns of the different CsA analogs, although the exact reason for this categorisation has not yet been determined.

### 3.2.4 The rigidity difference map

The preceding analysis demonstrates some of the problems with current methods for dealing with data from FIRST. In response, we construct the RDM, which captures the effect of ligand binding upon the rigidity of a structure, or number of structures, at the residue level. Each residue is given a score based on the RCD determined by FIRST. Following the measure $X_{\text{LRC}}$ we focus on the LRC [Tan and Rader, 2009; Radestock and Gohlke, 2008; Rader and Brown, 2011]. If the residue is in the LRC, it is given a score of 1, otherwise it is given a score of 0. This score is dependent on $E_{\text{cut}}$. As $E_{\text{cut}}$ is lowered, the LRC becomes smaller and more residues are assigned a score of 0. For any particular $E_{\text{cut}}$, the absolute difference between the scores of corresponding residues in the liganded and unliganded structures indicates whether or not the rigidity of the residue is affected by the ligand. This difference is 0 unless, at a particular $E_{\text{cut}}$, a residue is a member of the LRC in one structure and not the other in which case it is 1. These absolute differences were calculated for $E_{\text{cut}}$ between 0 kcal/mol and $-2$ kcal/mol, at intervals of 0.1 kcal/mol. From Figure 3.5, $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ are both 0 for $E_{\text{cut}} \leq -2$ kcal/mol, hence the limit of $-2$ kcal/mol here. Thus for 21 $E_{\text{cut}}$ values, each residue was given a absolute difference score of 0 or 1. It is these scores which are plotted in the RDM. Figure 3.6 is the

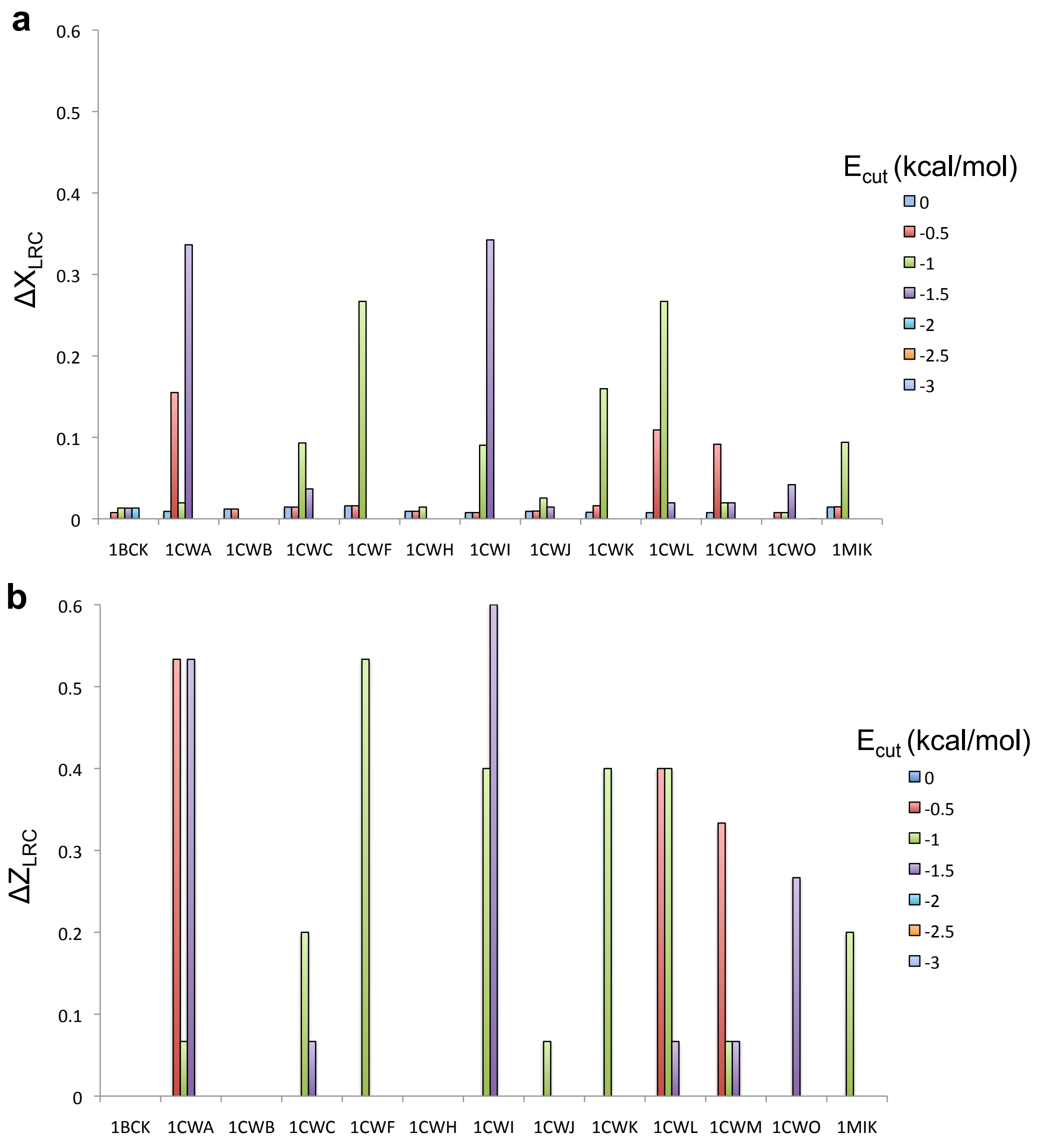


Figure 3.5: (a) $\Delta X_{\mathrm{LRC}}$ and (b) $\Delta Z_{\mathrm{LRC}}$ for 13 structures of CypA crystallised with CsA or CsA derivatives, evaluated at $E_{\mathrm{cut}} = 0, -0.5, -1, -1.5, -2, -2.5$ and $-3$ kcal/mol.

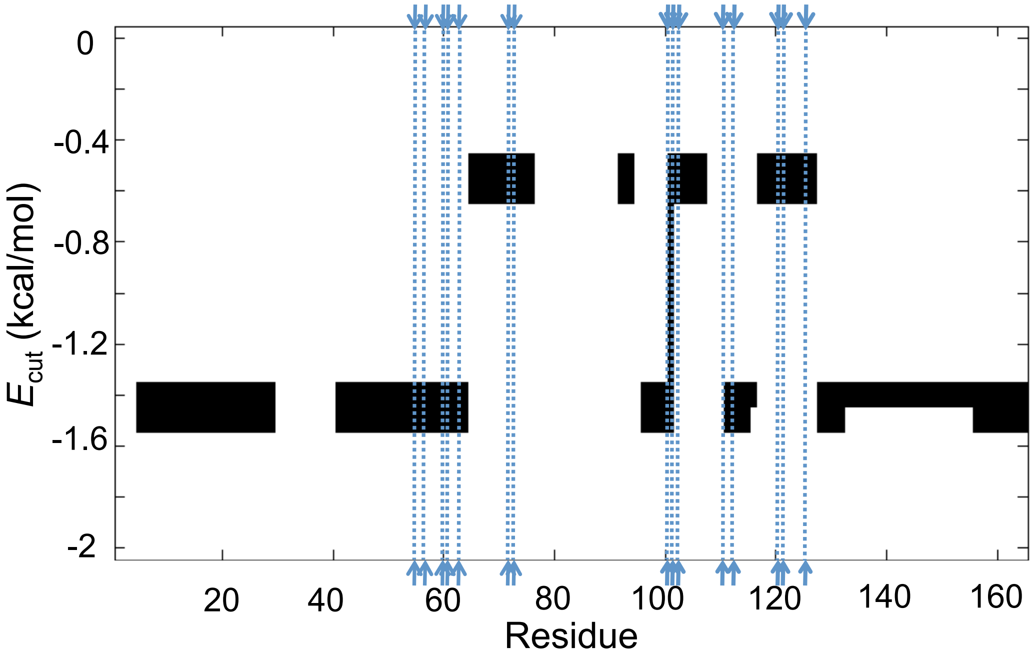


Figure 3.6: The RDM for 1CWA. Black regions show residues at which ligand binding has affected the protein rigidity, and at which $E_{\text{cut}}$ value. White regions indicate the absence of an effect. The binding site residues are indicated with arrows at the top and bottom of the plot, with dotted lines joining them.

RDM for 1CWA, where a score of 1 is shown in black and a score of 0 in white. We see clearly that the effect of ligand binding on the rigidity of the protein is dependent on $E_{\text{cut}}$. For most choices of $E_{\text{cut}}$, there is no difference in the LRC for any residue. When $E_{\text{cut}} = -0.5$ kcal/mol, the set of residues affected by ligand binding is different to the set at $E_{\text{cut}} = -1.5$ kca/mol. Only A101 is affected when $E_{\text{cut}} = -1$ kcal/mol.

In order to present the data for many structures in a single chart, we sum the absolute difference scores across the structures, and use these summed values form the entries of a cumulative RDM. The larger the entry in the cumulative RDM, the more consistent the effect of the ligand. Figure 3.7 is the cumulative RDM for the 13 structures of CypA from Table 3.2 and Figure 3.5. The RDM in Figure 3.7 highlights areas along the backbone of the protein where the rigidity is affected by ligand binding. The area between residues 65 and 130 has large entries at small values of $|E_{\text{cut}}|$. Lower $E_{\text{cut}}$ results in differences in the rest of the chain for some of the structures. A101 is consistently affected by ligand binding, across a broad range of $E_{\text{cut}}$ values. This residue is one of the 15 binding residues and as such it

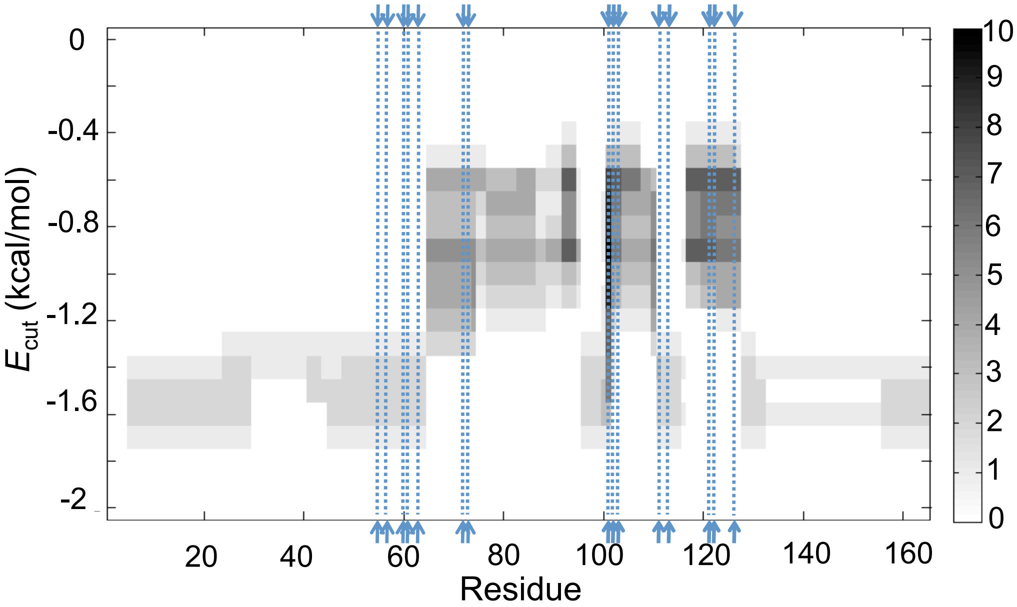


Figure 3.7: The cumulative RDM for the 13 structures of CypA which were also analysed together in Figure 3.5. Blue arrows and dotted lines indicate binding site residues.

is not surprising that it is consistently affected by the removal of the ligand. More intriguing is that other binding residues are less consistently affected — residues 55, 57, 60, 61 and 63 for example. There are also non binding residues which are strongly affected, such as residues 90 – 95.

The RDM identifies the residues affected by ligand binding *and* how this effect is dependent upon $E_{\mathrm{cut}}$. It is quick to generate, avoids confusion arising from a non-linear $E_{\mathrm{cut}}$ axis and gives residue-specific information on the effect of ligand binding on rigidity.

## 3.3 Discussion

Rigidity analysis was carried out on 59 structures of CypA taken from the PDB. Our initial approach was to examine the RD plots for any clear and consistent effects of ligand binding. In Chapter 2, the flap-rigidifying effects of some ligands on HIV-1 protease were clear from many of the RD plots. There were no such clear effects apparent in the RD plots of CypA, as demonstrated for the examples of 1AWQ and 1AWS in Figure 3.4. This is in agreement with experimental data which rule out large conformational change on binding [Kallen et al., 1998; Fanghänel and

Fischer, 2003]. We find that the structural differences between PDB entries caused more variation in the RD plots than ligand binding. Figure 3.3 shows the RD plots for 3K0N and 1W8V, two different structures of unbound CypA. Although overall rigidity loss is similar, there are many smaller discrepancies the kind of which are not apparent between the ligand-unliganded pairs of RD plots in Figure 3.4. The effect of ligand binding on the main chain rigidity of CypA is consistently subtle, and therefore difficult to assess using RD plots alone. Measures such as $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ are also to be treated with care. Their dependence on $E_{\text{cut}}$ was shown in Figure 3.5. It is not therefore appropriate to use a single $E_{\text{cut}}$ to define $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$. Using these measures at a systematic set of $E_{\text{cut}}$ values is a more suitable approach. In this way, potentially important effects are less likely to be overlooked. Figure 3.5 demonstrates that the impact of ligand binding on the rigidity of proteins can be highly sensitive to $E_{\text{cut}}$. Further, in Figure 3.6 we show that it is not merely the *extent* of this impact which is affected by the choice of $E_{\text{cut}}$ but also the *location* of the affected residues. The RDM combines the strengths of the RD plots, $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ and avoids the problems discussed above. In the RDM, residue-specific data is maintained, and a linear $E_{\text{cut}}$ axis is established.

Additionally, the RDM can be adapted so that rigidity analysis data for many structures can be summarised in a single plot (Figure 3.7). The RDM can be generated for any given protein and is an efficient way of showing the effects of ligand binding on a number of structures. We give an example of an informative cumulative RDM for a protein other than CypA. Figure 3.8 is the cumulative RDM for the structures of HIV-1 protease crystallised with the inhibitor tipranavir. Since there are four tipranavir structures, the maximum value for an entry in the cumulative RDM is 4. The figure shows that the most of the LRC is unaffected by the binding of the ligand, since the majority of the entries are 0. However the regions around residues 50 and 150 are consistently affected by the presence of the ligand. This result was discussed in more detail in Chapter 2.

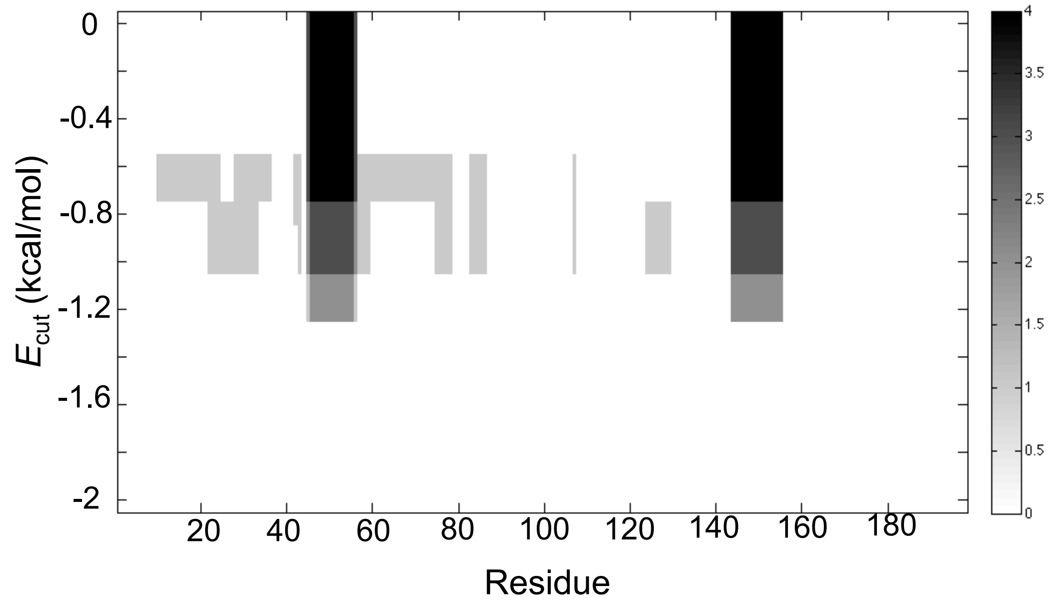


Figure 3.8: The cumulative RDM for four HIV-1 protease structures crystallised with the inhibitor tipranavir. The two regions on the backbone where there is a consistent difference in rigidity between the liganded and unliganded samples correspond to the two $\beta$-hairpin flaps which shield the active site of enzyme.

# Chapter 4

# Simulations of CypA mobility

## 4.1 Introduction

The work in this chapter covers simulations using both FIRST and FRODA, with a particular focus on the mobility simulations with FRODA. We begin by using theoretical methods to try to predict the experimentally determined HDX folding core. We consider different ways of processing the data from FRODA, such as through the use of pseudodihedral angles and by tracking buried residues. Further, we discuss the impact of hydrophobic tethers and hydrogen bonds on the rigidity and mobility of CypA. Extending the mobility simulations, we study the effect of ligand binding on the mobility of the protein by running simulations on the ligand-bound as well as the ligand-free structure. Unless stated otherwise, theoretical results were determined based on the structure 1CWA with the ligand (CsA) removed.

### 4.1.1 Folding cores

In Section 1.1.3 we introduced the protein folding problem and the experimental folding cores which can be identified through $\phi$-value analysis and HDX. Here, we refer to the slowly exchanging residues in HDX as the HDX folding core for a protein. A previously published study on unbound CypA used HDX experiments to categorise each residue based on its exchange rate [Shi et al., 2006]. The data reported by the authors allow us to define the HDX folding core for unbound CypA.

We use FIRST and FRODA to generate theoretical folding cores for comparison with the HDX folding core. We use an established definition for the FIRST folding core [Hespenheide et al., 2002; Rader and Bahar, 2004]. The folding core energy $E_{\mathrm{fc}}$ was discussed in Section 2.3.3 and is defined to be the lowest $E_{\mathrm{cut}}$ in the rigidity dilution plot where at least three consecutive residues are mutually rigid with at least three other consecutive residues of another secondary structural element [Hespenheide et al., 2002; Tastan et al., 2007]. The mutually rigid residues at $E_{\text{fc}}$ make up the folding core of the protein. The FIRST folding core and the HDX folding core were compared for a number of proteins in previous studies [Hespenheide et al., 2002; Rader and Bahar, 2004]. Good agreement was found between the two folding cores for eight of ten different proteins [Hespenheide et al., 2002]. This was expanded upon in a study of 29 different protein structures, where theoretical folding cores were generated with FIRST as well as the Gaussian Network Model [Rader and Bahar, 2004]. The quality of agreement between the different folding cores varied across the set of proteins, with the authors concluding that the best approach for predicting folding cores is to use the techniques in combination [Rader and Bahar, 2004]. The HDX data for unbound CypA was published after these studies, and so we compare theoretical folding cores with this dataset for the first time. We define theoretical folding cores using output from FRODA in two different ways, described below, as well as for a FIRST + FRODA combination approach.

### 4.1.2 Pseudodihedral angles

Pseudodihedral angles are a convenient way of simplifying the representation of protein backbone motion. A dihedral angle is the angle between two planes defined by four points. For example, the familiar backbone dihedral angles of proteins, $\phi$ and $\psi$, are defined using the four adjacent backbone atoms C-N-$C_\alpha$-C and $C_\alpha$-C-N-$C_\alpha$ respectively. Pseudodihedral angles are formed between the two planes defined by four consecutive $C_\alpha$ atoms [DeWitte and Shakhnovich, 1994]. Since the adjacent $C_\alpha$ atoms are not bonded covalently to each other, pseudobonds are constructed by drawing vectors between the atoms, and these pseudobonds are used to construct the pseudodihedral angle. Figure 4.1 shows the construction of a pseudodihedral atom between four consecutive $C_\alpha$ atoms.

### 4.1.3 Hydrophobic interactions

We wish to determine the most appropriate bond network to use when studying the flexibility and mobility of CypA in solution. This is because we will be comparing our simulations with data obtained from experiments, namely HDX using solution NMR. We keep the definition of HP strength constant throughout rigidity dilution and do not change the number of interactions based on temperature. As input to FRODA, we use a single bond network and calculate the mobility of the structure based on that network. HDX is dominated by surface exposure and mobility, and

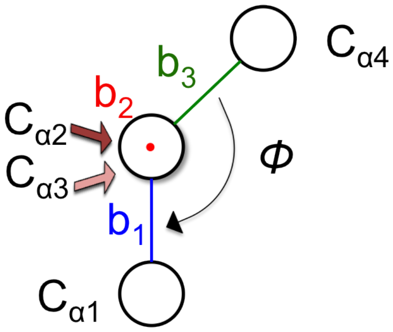


Figure 4.1: Construction of the pseudodihedral angle. Four consecutive $C_\alpha$ atoms, $C_{\alpha 1}$, $C_{\alpha 2}$, $C_{\alpha 3}$ and $C_{\alpha 4}$ are connected by three pseudobonds, labelled $b_1$ (blue), $b_2$ (red) and $b_3$ (green). In the figure, we are looking down $b_2$ directly, so this bond is represented as a dot. In the centre there are in fact two atoms on top of each other, with $C_{\alpha 3}$ above $C_{\alpha 2}$. The pseudodihedral angle is labelled $\phi$.

we change the criteria for HP inclusion based on the surface exposure. In previous work, the exposure level of the interacting atoms has not been considered as the prime criterion for including HPs. There are three different filters we can place on the bond network: (1) both bonding atoms must be buried (HP0), (2) at least one of the bonding atoms must be buried (HP1), (3) both atoms may be buried or exposed (HP2). We also place these conditions on the HBs (HB0, HB1 and HB2), creating nine pairs of HB/HP bond settings. The default setting is HB2HP2, where interacting atoms may be exposed or buried for both HBs and HPs.

## 4.2 Methods

### 4.2.1 Simulating motion with FRODA

For the FRODA simulations we used the PDB structure 1CWA, of the CypA-CsA complex. Unless otherwise stated the results presented are for the structure after the ligand has been removed. The structure was processed as before. Bond networks were generated using FIRST with five different $E_{\text{cut}}$ values ($-0.5, -1.0, -1.5, -2.0$ and $-3.0$ kcal/mol) and the nine different HB/HP settings. The results presented are for setting HB2HP2 and $E_{\text{cut}} = -2.0$ kcal/mol unless we state otherwise. The normal modes of motion and their respective frequencies were determined using the ELNEMO software [Suhre and Sanejouand, 2004]. The FRODA module in FIRST was then used to bias the motion of the protein in the direction of the normal mode vectors. The positive and negative directions of the first ten non-trivial modes, $m_7 - m_{16}$, were used to simulate protein motion. During each run the random step size and the directed step size were both 0.01 Å. A total of 2000 conformers were generated in each run and every 200th conformer was recorded.

We calculated pseudodihedral angles in each recorded conformer. For residue $i$, the pseudodihedral angle is calculated using the $C_\alpha$ atoms of residue $i-1$, $i$, $i+1$ and $i+2$. Since the pseudodihedral angles cannot be generated in this way for the terminal residues (residues 1, 164 and 165), these are assigned zero value. The absolute difference between the cosine of each pseudodihedral angle and of the corresponding pseudodihedral angle in the original structure is calculated. For each residue, we define out mobility measure $M$ to be the mean absolute difference. Typically $M$ is in $[0, 0.5]$, and it is always restricted to the interval $[0, 2]$.

An alternative way of tracking the motion in FRODA is to measure burial distances. For each atom in the protein, a sphere of radius $r_{atom} + r_{solvent}$ is drawn with the atom in the centre. We set $r_{solvent}$ to 1.4 Å, representing a water sphere. The value of $r_{atom}$ varies according to atom type; for nitrogen, $r_{atom} = 1.5$ Å. The points which form the sphere are then individually checked for contact with the neighbouring atoms. If any point on the sphere is not in contact with a neighbour then that is a potential solvent position and the atom is labeled as being exposed. If an atom is not exposed, it is buried and its burial distance, $R$, is the shortest distance to an exposed atom. An exposed atom has $R = 0$ Å. For a given conformer, we calculated $R$ for each nitrogen atom in the amide backbone and used this as the value of $R$ for the residue to which it belongs.

### 4.2.2 Determining folding cores

The published HDX experiments on unbound CypA resulted in a classification of CypA residues in terms of their exchange rates, $k_{\text{ex}}$ [Shi et al., 2006]. The classifications were as follows: immediate (too quick to measure), fast ($k_{\text{ex}} > 10^{-2}$ min$^{-1}$), medium ($10^{-4}$ min$^{-1} < k_{\text{ex}} < 10^{-2}$ min$^{-1}$) and slow ($k_{\text{ex}} < 10^{-4}$ min$^{-1}$). Twelve residues, including the proline residues, were not categorised since they were not identified in the HSQC spectrum. We refer to the slow and medium exchanging residues, with $k_{\text{ex}} < 10^{-2}$ min$^{-1}$ as the HDX folding core. The HDX folding core for unbound CypA has 73 residues.

The FIRST folding core was determined directly from the rigidity dilution plots. Secondary structure classifications were determined using PYMOL.

In a previous study comparing theoretical results with HDX, the measure $s(AB)$ is used to assess the quality of the agreement between the folding cores determined by two methods $A$ and $B$ [Rader and Bahar, 2004]. $s(AB)$ is an 'enhancement factor' – the ratio between the number of residues correctly identified by the theoretical method and the number of residues expected to be identified by a random selection. $s(AB)$ is defined as

$$s(AB) = c(A, B)N/(N_A N_B), \tag{4.1}$$

where $N_A$ and $N_B$ are the number of residues in the folding cores determined by methods $A$ and $B$ respectively, $c(A, B)$ is the overlap between the two folding cores and $N$ is the total number of residues in the protein. The number of residues expected to be identified by a random selection is $N_A N_B/N$. $s(AB) = 1$ corresponds to a random match and the quality of the agreement increases with $s(AB)$. We use this measure here with the HDX experiment as our reference method $A$, so that $N_A = 73$. For CypA, $N = 165$. In addition we introduce two simple percentage-based measures of how well the theoretical method predicts the HDX folding core. These are the *coverage*, $\%_C$, and the *accuracy*, $\%_A$, defined below as

$$\%_C = 100 \times \frac{c(A, B)}{N_A}, \tag{4.2}$$

$$\%_A = 100 \times \frac{c(A, B)}{N_B}. \tag{4.3}$$

## 4.3 Results

### 4.3.1 Theoretical folding cores and the HDX folding core

We shall use four methods for predicting the folding core. In each case, we compare the predicted folding core with the experimentally determined HDX folding core for unbound CypA [Shi et al., 2006]. The folding core is determined using (i) FIRST, (ii) the secondary structure, (iii) FRODA, and (iv) by combining data from FIRST with data from FRODA. Figure 4.2 shows burial distances $R$ for each residue of CypA. For all of the residues, $0 \leq R \leq 4.85$. There are 30 residues for which $R = 0$ and 89 residues for which $R < 1.5$. The remaining 76 residues all have $R > 2.0$; none have $1.5 \leq R \leq 2.0$. We categorise the residues as being buried ($R < 2.0$) or exposed ($R \geq 2.0$).

The FRODA simulations were run as described in Section 4.2.1 and buried residues were evaluated for the final conformer for each mode and direction. There were therefore 20 conformers evaluated. The residues which were classed as being buried in more than half of the final conformations, form the FRODA folding

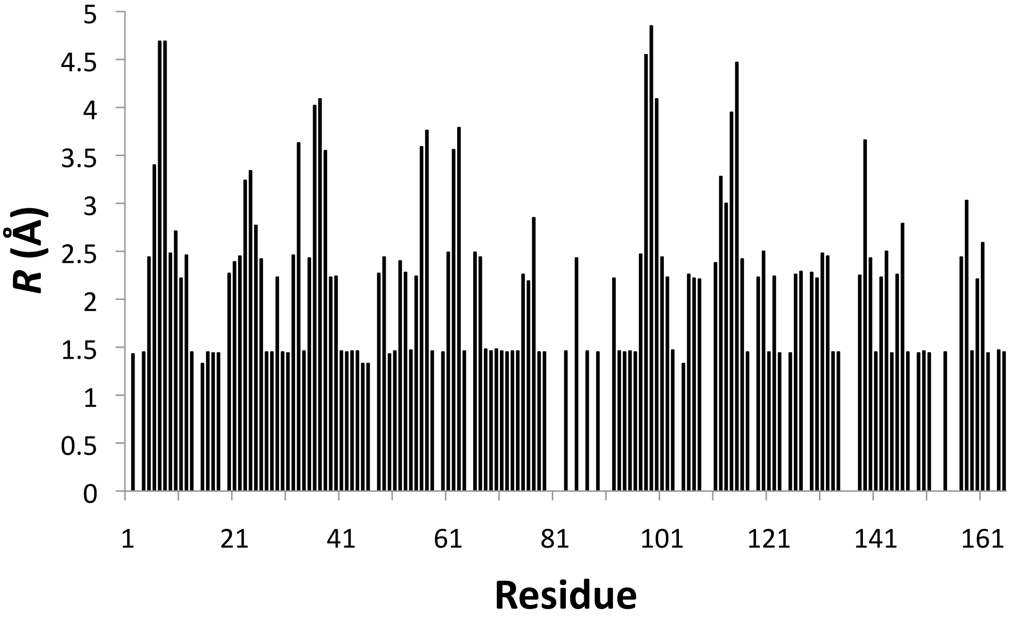


Figure 4.2: The burial distance $R$ for each residue of CypA. Exposure level is determined based on the amide nitrogen atom, as described in Section 4.2.1.

core. The combined FIRST and FRODA folding core is simply the intersection of the FIRST folding core and the FRODA folding core. The folding core predictions are plotted with the HDX folding core in Figure 4.3 by selectively colouring the protein backbone. The overlap between each of the theoretical folding cores and the HDX folding core is strong in each case. The match between the FIRST folding core is qualitatively similar to the folding core predictions previously published using FIRST [Hespenheide et al., 2002; Rader and Bahar, 2004]. There are large sections of the protein, for example residues 146-156, which are part of the FIRST folding core and not the HDX folding core. This effect appears to be eliminated for the secondary structure folding core, which is smaller. Since rigidity dilutions capture the secondary structure elements through persistent rigidity [Wells et al., 2009], the folding core is likely to consist predominantly of residues belonging to $\alpha$-helices and $\beta$-sheets. These residues appear to correlate well with HDX, so it may be that the FIRST folding core inherits this correlation. There are few sections of the backbone which are part of both the FIRST and HDX folding cores but not the secondary structure elements. There are a number of residues in the FRODA folding core which are not part of the HDX folding core. For example, there are eight of these between residues 64 and 95. When the FIRST and FRODA folding cores are combined, the prediction is improved. Figure 4.3 allows us to qualitatively assess the folding core

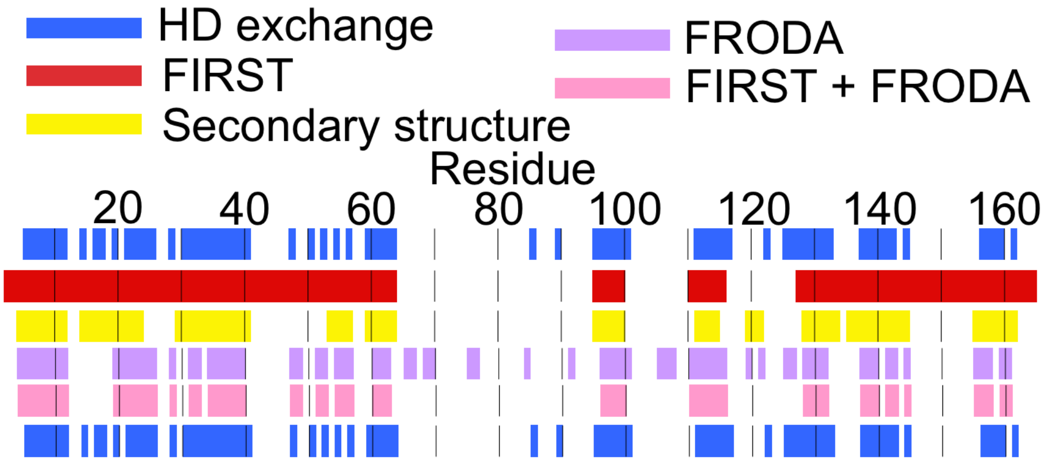


Figure 4.3: Computationally determined folding core residues as predictors of the experimental HDX folding core. The protein backbone is coloured according to the residues which make up the folding core. Residues which are not part of the folding core are not coloured. The top and bottom rows of the plot both show the experimentally determined HDX folding core [Shi et al., 2006]. The colour key for the computational methods in between is given above the plot.

predictions. The qualitative approach is made quantitative in Table 4.1, where for each method the measures $s(AB)$, $\%_C$ and $\%_A$ are calculated using $N_B$ and $c(A, B)$ as defined in Section 4.2.2. Since $s(AB) > 1$ for all of the theoretical methods, they are each improvements upon random selections [Rader and Bahar, 2004]. None of the methods have $s(AB) \geq 2$, implying that none are twice as good as random. The FIRST folding core has the lowest $s(AB)$ value, 1.33, suggesting the smallest improvement on a random selection. The $\%_C$ is the highest of all the methods at 90.41%, although the $\%_A$ of 58.93% is the lowest. The combined approach of FIRST and FRODA yields the highest $s(AB)$. Although this method has the lowest coverage ($\%_C$ = 69.86), the residues are identified with the greatest amount of accuracy ($\%_A$ = 86.44).

By these measures FIRST and FRODA perform poorly relative to the secondary structure. In combination the two methods perform comparably – there is a lower coverage but a greater accuracy. The advantage of using FIRST and FRODA together it may be possible to capture the effect of ligand binding through these methods. The secondary structure does not change when the ligand is removed from the crystal structure and therefore this method is not adaptable to such considerations.

Table 4.1: Agreement between theoretical methods and experimental HDX folding core. For each theoretical method, the number of residues in the folding core, $N_B$ is given along with $c(A, B)$, the size of the overlap between the theoretical folding core and the HDX folding core. These numbers are used to determine $s(AB)$, $\%_C$ and $\%_A$ as defined in Section 4.2.2. The final two methods will be discussed later in this chapter

| Method | $N_B$ | $c(A, B)$ | $s(AB)$ | $\%_C$ | $\%_A$ |
|---|---|---|---|---|---|
| FIRST | 112 | 66 | 1.33 | 90.41 | 58.93 |
| Secondary structure | 75 | 58 | 1.75 | 79.45 | 77.33 |
| FRODA | 76 | 54 | 1.61 | 73.97 | 71.05 |
| FIRST + FRODA | 59 | 51 | 1.95 | 69.86 | 86.44 |
| Pseudodihedrals | 95 | 45 | 1.13 | 67.12 | 51.58 |
| FIRST with HB0HP0 | 73 | 52 | 1.63 | 71.23 | 72.22 |

### 4.3.2 The mobility of CypA characterised by pseudodihedral angles.

We used pseudodihedral angles to calculate $M$ for each residue of CypA during FRODA simulations, plotted in Figure 4.4. There are two regions corresponding to $\alpha$-helices, residues 30–41 and 136–143, where the mobility is highly restricted. The residues central to these helices each have $M \simeq 0$. Residue W121, which is part of the $3_{10}$ helix (residues 121–123) also has $M \simeq 0$. Elsewhere peaks of relatively high mobility are distributed across the protein backbone. Average $M$ across all residues is 0.10. Residues with $M < 0.10$ were compared with the HDX folding core. There were 95 such residues, 49 of which were part of the HDX folding core. The $s(AB)$, $\%_C$ and $\%_A$ values were 1.17, 67.12% and 51.58% respectively, scoring lower in each case than the other theoretical methods presented in Table 4.1. With a small protein such as CypA with a large surface area to volume ratio, HDX is dominated by surface exposure. Average $M$ changes monotonically with $E_{\text{cut}}$. When $E_{\text{cut}}$ = -0.5, -1.0, -1.5, and -3.0 kcal/mol, average $M$ is 0.03, 0.04, 0.09 and 0.11 respectively. Lowering $E_{\text{cut}}$ removes bonds from the bond network and constrains the protein to a lesser extent. It is not just the overall magnitude of $M$ which changes with $E_{\text{cut}}$ but also the residues which are mobile. In Figure 4.5 we plot $M$ for different values of $E_{\text{cut}}$. At $E_{\text{cut}} = -1.5$ and $-3.0$ kcal/mol the protein is largely mobile, similar to the plot for $E_{\text{cut}} = -2$ kcal/mol. For $E_{\text{cut}} = -0.5$ and $-1.0$ kcal/mol the mobility is restricted, with $M \simeq 0$ for residues 4–61 and 128–165 in both cases. In the RD of 1CWA, there is a first-order transition [Wells et al., 2009] from rigid to flexible at $E_{\text{cut}}$ = -1.376 kcal/mol (data not shown). The mobility of the protein is clearly

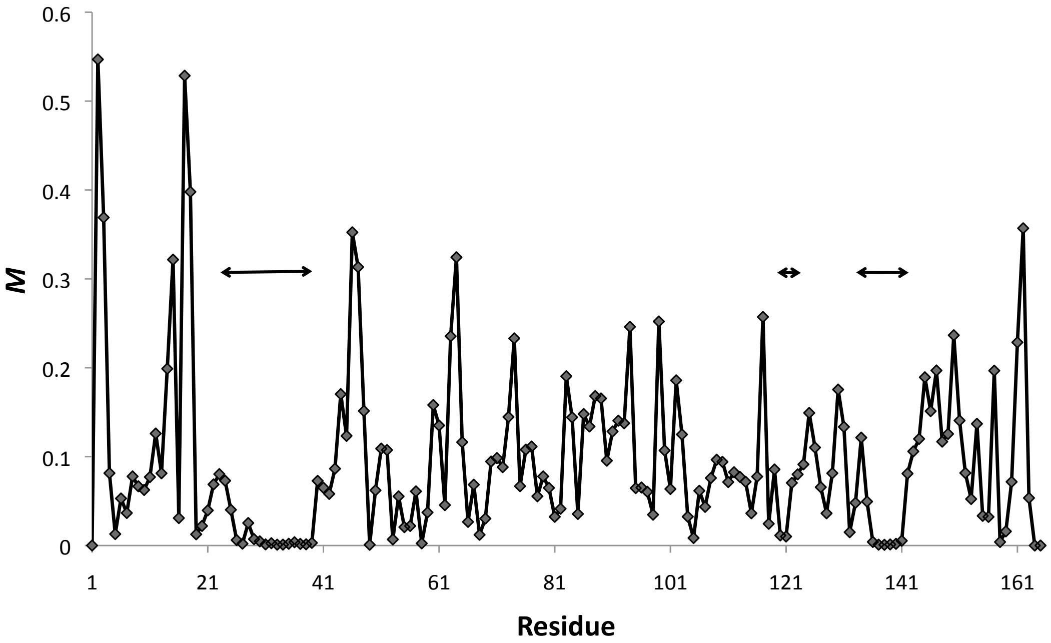


Figure 4.4: The mobility of each residue, $M$, determined from pseudodihedral angles generated in FRODA simulations. Horizontal black arrows indicate helical regions of the protein.

influenced by the change in rigidity, and therefore the choice of $E_{\mathrm{cut}}$.

### 4.3.3 The effect of ligand binding on $M$

We calculated $M$ for each residue in the structure before and after the removal of the ligand. We denote the absolute difference of these values $\Delta M$. Figure 4.6 shows $\Delta M$ for each residue. The $\alpha$ helix regions have low $M$ irrespective of the presence of CsA. The spikes at the termini in Figure 4.4 have been reduced relative to other areas of the protein, indicating high $M$ before and after ligand removal. The majority of the residues in Figure 4.6 have $\Delta M < 0.05$. The remaining set of eight residues (60, 63, 64, 94, 99, 117, 119 and 154) are affected to a greater extent by the presence of the ligand. Residues 64, 94 and 99 had particularly high $M$ in the unbound CypA, but the other residues did not. Only residues 60 and 63 are part of the CypA binding cleft discussed in Chapter 3. These binding site residues are indicated by the black arrows in Figure 4.6. The minimum distances between the heavy atoms of the ligand and the heavy atoms of residues 94 and 154 are 11.7 Å and 16.8 Å respectively, suggesting effects of ligand binding which are not proximal to the ligand itself. Raising $E_{\mathrm{cut}}$ restricts the mobility by introducing

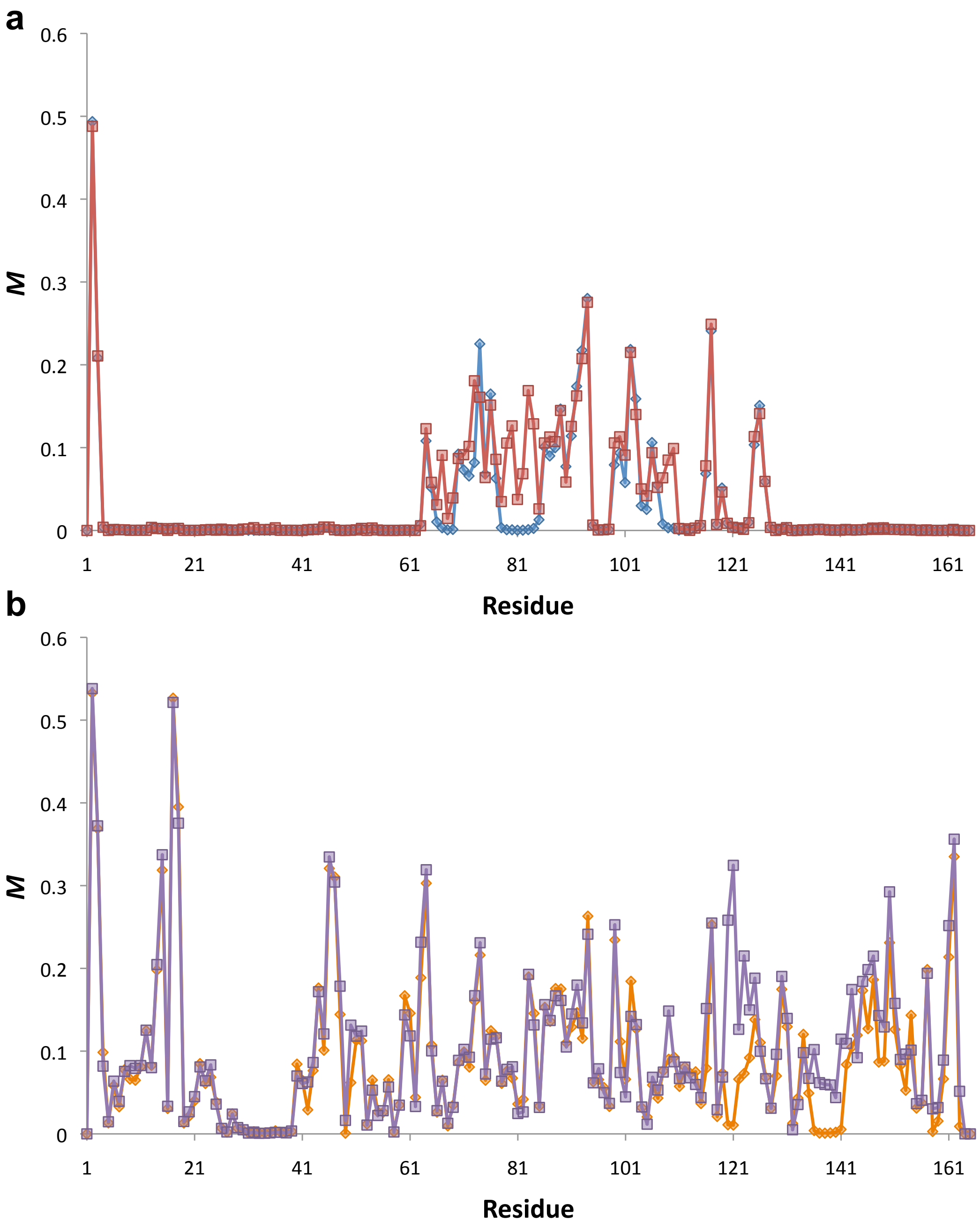


Figure 4.5: The effect of $E_{\text{cut}}$ on $M$. (a) $M$ evaluated at $E_{\text{cut}} = -0.5$ kcal/mol (blue), and at $E_{\text{cut}} = -1.0$ kcal/mol (red). (b) $M$ evaluated at $E_{\text{cut}} = -1.5$ kcal/mol (orange), and at $E_{\text{cut}} = -3.0$ kcal/mol (purple).

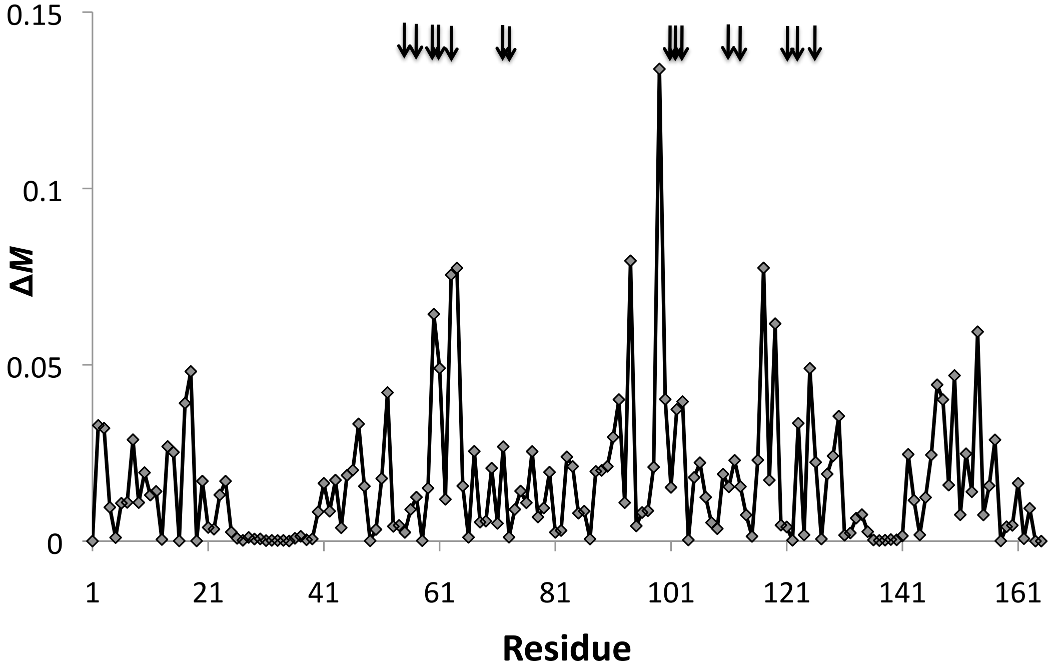


Figure 4.6: The $\Delta M$ value for each residue at $E_{\text{cut}} = -2$ kcal/mol. Black arrows indicate the 15 binding site residues.

more constraints into the system. At a higher value of $E_{\text{cut}}$, more protein-ligand interactions may be introduced and consequently $\Delta M$ increased. Figure 4.7 shows the $\Delta M$ for each residue, evaluated at $E_{\text{cut}}$ = -0.5, -1 and -1.5 kcal/mol. Clearly the choice of $E_{\text{cut}}$ alters the effect of ligand binding on $M$. In general, $\Delta M$ is greater when $E_{\text{cut}} = -0.5$ or $-1.5$ kcal/mol than when $E_{\text{cut}} = -1$ kcal/mol. Unlike with $M$, there is no monotonic relationship between $E_{\text{cut}}$ and $\Delta M$. This is surprising — from the way we have calculated $\Delta M$, we would expect the values to increase with decreasing $E_{\text{cut}}$, since at lower $E_{\text{cut}}$ we take the absolute difference of two $M$ larger values. We also observe that the identity of the residues with relatively large $\Delta M$ is also dependent on choice of $E_{\text{cut}}$.

In Chapter 3, we designed the RDM to display the effect of ligand binding on protein rigidity. The RDM for 1CWA is shown in Figure 4.8 (centre). We plot $\Delta M$ at $E_{\text{cut}}$ = -0.5 kcal/mol above, and at $E_{\text{cut}}$ = -1.5 kcal/mol below. Figure 4.8 demonstrates the direct impact of $E_{\text{cut}}$ on protein mobility. The choice of $E_{\text{cut}}$ affects the bond network and therefore the rigidity of the protein. Since the bond network for the complex includes protein-ligand interactions, $E_{\text{cut}}$ impacts directly the effect of ligand binding on protein rigidity. The RDM demonstrates this clearly,

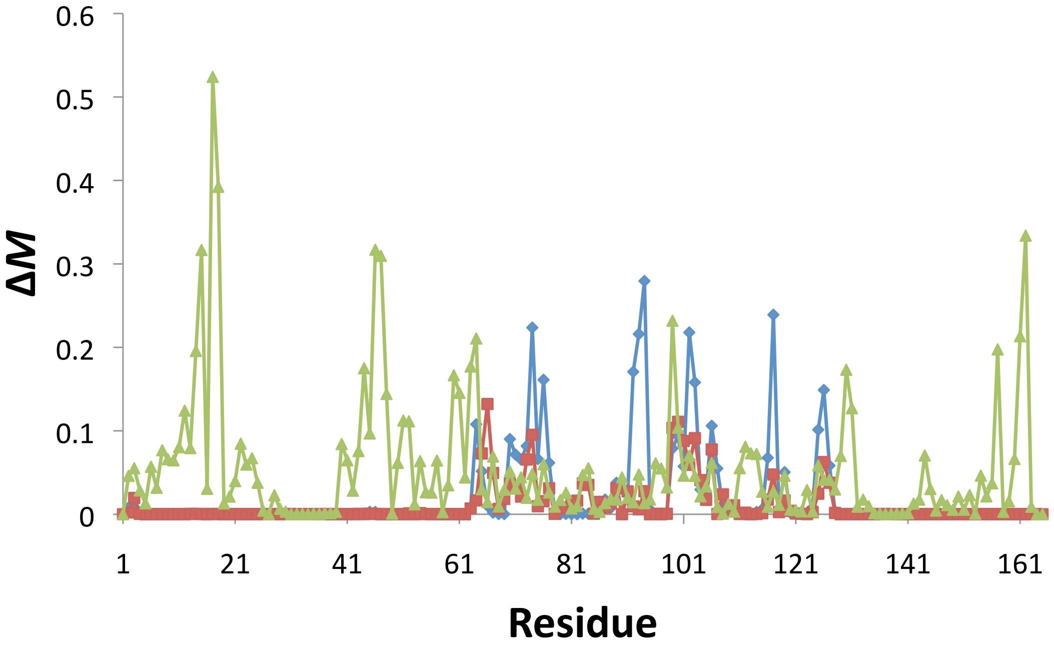


Figure 4.7: The $\Delta M$ value for each residue at $E_{\text{cut}} = -0.5$ kcal/mol (blue), $-1.0$ kcal/mol (red) and $-1.5$ kcal/mol (green).

and the accompanying plots of $\Delta M$ show how mobility inherits rigidity. The choice of $E_{\text{cut}}$ impacts both the mobility of the protein (Figure 4.5) and the effect of ligand binding on protein mobility (Figures 4.7 and 4.8). For this reason, as with rigidity analysis, selection of $E_{\text{cut}}$ cannot be made lightly when running FRODA simulations.

### 4.3.4 Burial distances before and after ligand removal

The effect of ligand binding on $R$ is shown in Figure 4.9. We calculated the average $R$ of each residue from the final conformers of the FRODA simulations. The absolute difference between the ligand-bound and ligand-free structures, $\Delta R$ is plotted in Figure 4.9. For the majority of residues, $\Delta R < 0.5$ Å. The remaining eleven residues are 62, 63, 99, 101, 102, 103, 104, 110, 112, 113 and 114. Residues 63, 101, 102, 103, and 113 are binding site residues as determined in Chapter 3. Of the other six, none are further than two residues along the backbone from a binding site residue. We also calculated $\Delta R$ for each residue at $E_{\text{cut}}$ = -0.5, -1.0, -1.5 and -3.0. In each case, the residues which are part of, or close to, the binding site have the largest $\Delta R$ (data not shown). Although ligand binding does affect the mobility of CypA (Figure 4.8), it does not appear to change the surface exposure levels of residues other than those proximal to the binding site.

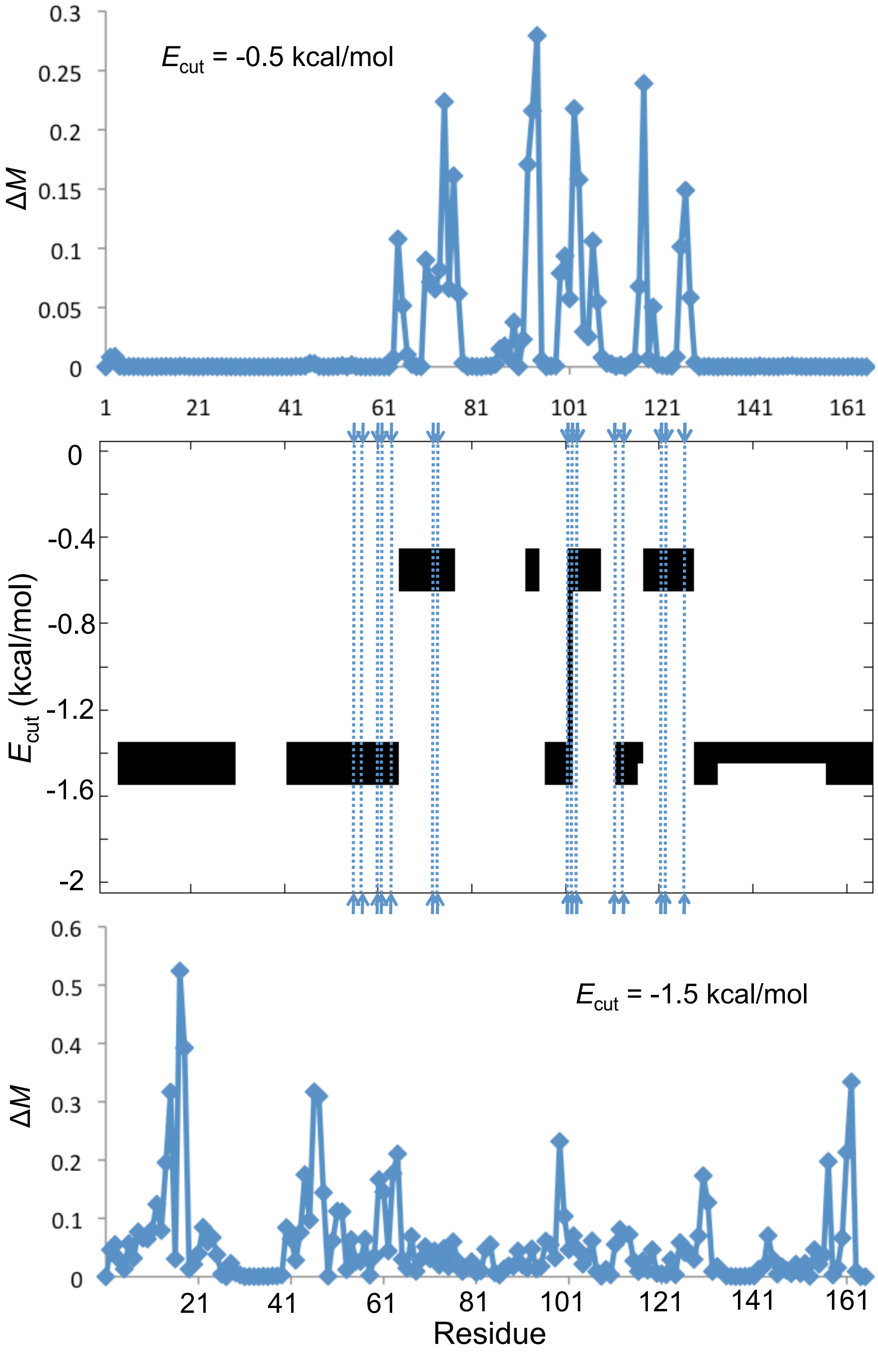


Figure 4.8: The RDM for 1CWA is shown in the centre. The $\Delta M$ value for each residue is shown at $E_{\text{cut}}$ = -0.5 kcal/mol (top) and at $E_{\text{cut}}$ = -1.5 kcal/mol (bottom). The arrows joined by dotted lines indicate the binding site residues.

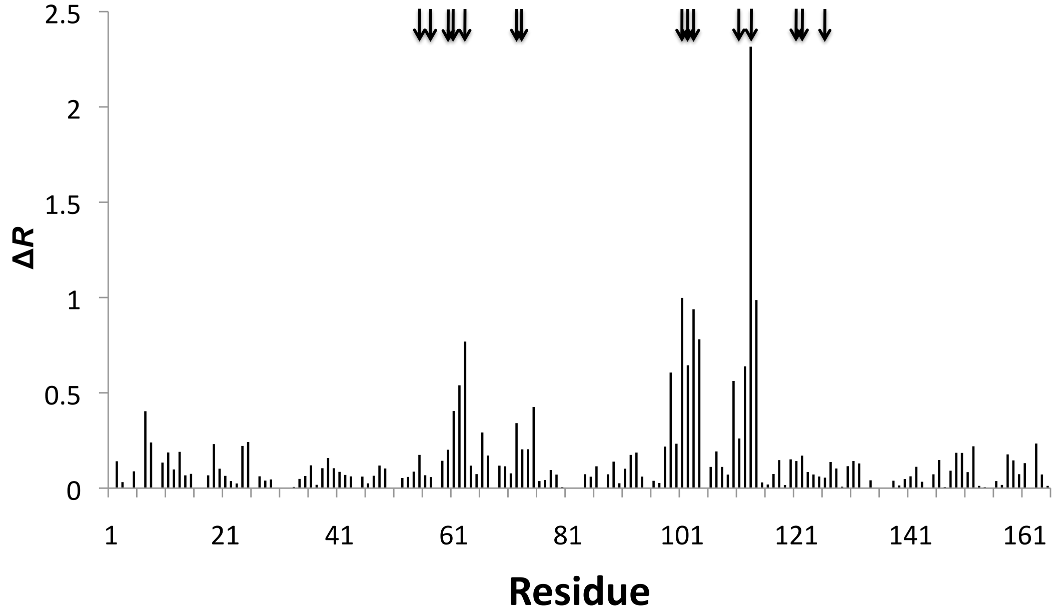


Figure 4.9: The change in burial distance, $\Delta R$, for each residue at $E_{\text{cut}} = -2$ kcal/mol. Black arrows indicate binding site residues.

### 4.3.5 Rigidity analysis with different HB/HP settings

The results presented thus far have used the default settings of FIRST to establish the bond network for the simulations in FRODA. In this section we explore the impact of changing these default settings. We use the HB/HP settings which were introduced in Section 4.1.3. Table 4.2 shows the number of interactions resulting from each bond setting for the structure 1CWA with the ligand removed. In the default setting, HB2HP2, the ratio of HBs to HPs is approximately 3:1. As the HB and HP settings are changed together from 2 to 1 and from 1 to 0, the number of interactions is decreased by around 20% in both cases and the 3:1 ratio is maintained. During a rigidity dilution the HPs are maintained. HBs are removed as before, and so the number of HBs included depends on $E_{\text{cut}}$. The relative increments in the number of HBs in HB0, HB1 and HB2 remains approximately the same regardless of $E_{\text{cut}}$. For example, when $E_{\text{cut}} = -2.0$ kcal/mol the number of interactions in settings HB0, HB1 and HB2 are 78, 90 and 110 respectively. RD plots were generated for unbound CypA using each of the nine bond settings. The overall pattern of rigidity loss is independent of the bond network; a first order transition between $E_{\text{cut}} = -0.8$ and $-1.3$ kcal/mol is observed in each case (data not shown). This first order transition is observed in all RD plots for CypA as the protein is largely $\beta$-sheet in structure

Table 4.2: The number of interactions for each HB and HP setting in 1CWA with no ligand bound. HBs are subsequently removed during a rigidity dilution, whilst the number of HPs remains constant.

| Bond setting | Number of interactions |
|---|---|
| HB0 | 149 |
| HB1 | 202 |
| HB2 | 242 |
| HP0 | 47 |
| HP1 | 64 |
| HP2 | 78 |

Table 4.3: The folding core energy, $E_{\mathrm{fc}}$, and the number of residues in the folding core are given for each HB/HPbond setting.

| Bond setting | $E_{\mathrm{fc}}$ (kcal/mol) | Size of folding core |
|---|---|---|
| HB0HP0 | $-0.820$ | 73 |
| HB0HP1 | $-0.820$ | 73 |
| HB0HP2 | $-1.062$ | 73 |
| HB1HP0 | $-1.062$ | 79 |
| HB1HP1 | $-1.062$ | 79 |
| HB1HP2 | $-1.062$ | 79 |
| HB2HP0 | $-1.062$ | 112 |
| HB2HP1 | $-1.263$ | 112 |
| HB2HP2 | $-1.263$ | 112 |

[Wells et al., 2009]. In Table 4.3, we show the size of the FIRST folding core for the nine different bond settings, along with $E_{\mathrm{fc}}$, the folding core energy. The HB setting affects the size of the folding core but the HP setting does not. It is unsurprising that the default HB setting HB2 results in the largest folding core because this setting contains more hydrogen bonds. Although the HP setting does not change the size of the folding core it can alter the folding core energy. When this happens, an increase in the number of HPs results in the folding core energy being lowered. The folding core energy can also be affected by changing the HB settings. From these data it is clear that changing the bond settings can affect the rigidity of the whole protein: more bonds are likely to result in lower $E_{\mathrm{fc}}$. The change in the folding core is dominated by the HB setting and not the setting for HPs. The respective folding cores in Table 4.3 differ in that additional residues are added along with additional bonds. There are no residues in the folding core of size 73 which are not part of the

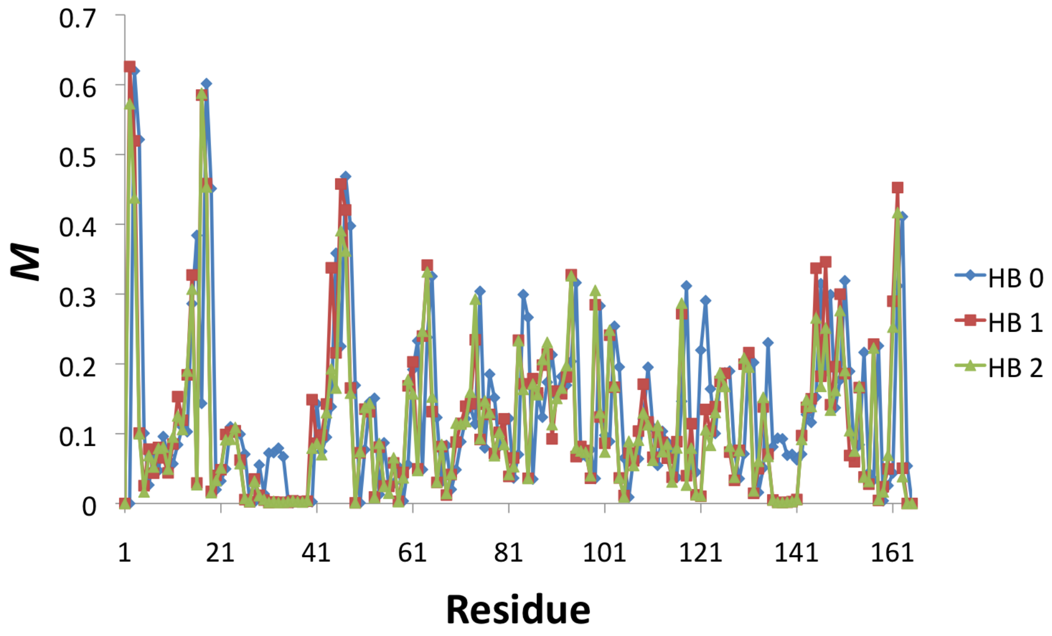


Figure 4.10: For each residue, $M$ is plotted for the three hydrogen bond settings.

folding core of size 79 or indeed that of size 112. Using the bond setting HB0HP0 the folding core has 73 residues, 52 of which are part of the HDX folding core. This gives $s(AB) = 1.63$, $\%_C = 71.23\%$ and $\%_A = 72.22\%$. This is an improvement on the initial FIRST folding core, the scores for which are given in Table 4.1. For setting HB0HP0, non-covalent interactions contacts only exist between buried residues of the protein. The FIRST folding core from this setting is in better agreement with the HDX folding core since surface exposure is taken into account during rigidity analysis.

### 4.3.6 The effect of HB/HP settings on $M$

Substantial changes to the folding core indicated the impact of HB setting on protein rigidity. In order to investigate whether these would translate into changes in mobility, we calculated $M$ using different HB settings, as shown in Figure 4.10. The plots are similar, with most of the peaks overlapping in location and height. In general, higher $M$ is observed with the HB0 setting which is expected due to the presence of fewer bonds constraining motion. At this $E_{\text{cut}}$ ($-2$ kcal/mol), changing the bond network involves removing bonds from a protein which is already highly flexible. Since the mainchain rigidity is unaffected, the same goes for $M$. Similarly, changing the HB setting does not affect $\Delta M$ at this $E_{\text{cut}}$ (data not shown). Figure

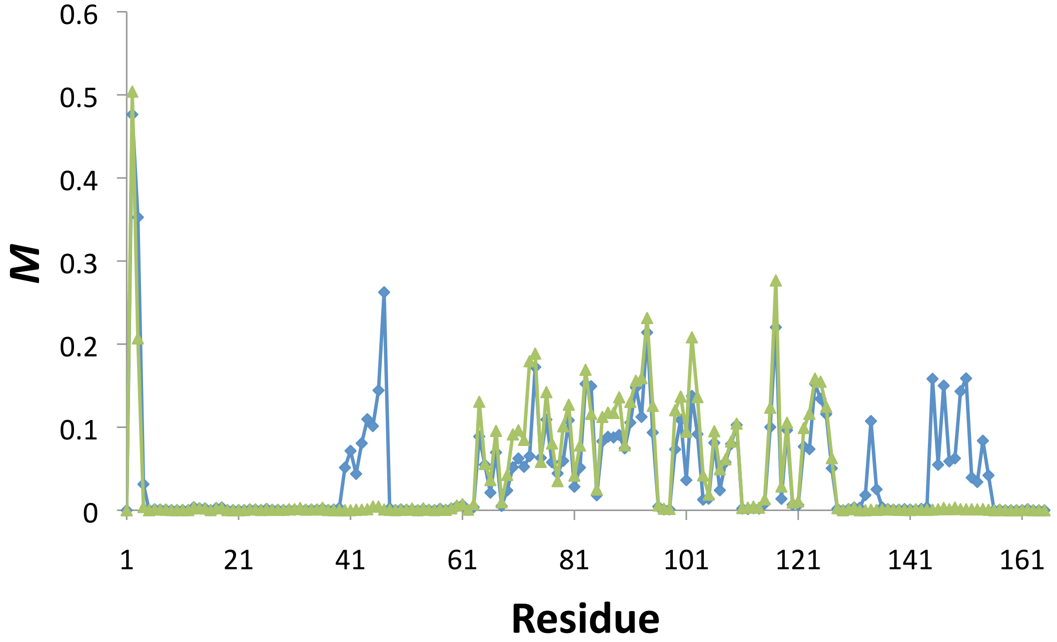


Figure 4.11: Mobility plots for the bond settings HB1HP0 (blue) and HB2HP0 (green) at $E_{\text{cut}}$ = -1.062 kcal/mol.

4.11 shows $M$ evaluated at $E_{\text{cut}} = -1.062$ kcal/mol, with two different bond settings, HB2HP0 and HB1HP0. From Table 4.3 we see that $E_{fc} = -1.062$ kcal/mol for both of these bond settings, but that the folding core is smaller for HB1HP0 than for HB2HP0. This difference in rigidity manifests itself in the $M$ plot. This underlines the findings of Section 4.3.3. Changes in rigidity confer changes in $M$. The bond network, determined in this case by both $E_{\text{cut}}$ and the HB/HP settings, directly affects protein mobility.

### 4.3.7 Further simulation data

When carrying out this work, more data was generated than presented here. For the three bond settings HB0HP0, HB1HP1, and HB2HP2, FRODA simulations were run at $E_{\text{cut}} = -0.5, -1, -1.5, -2$ and $-3$ kcal/mol. This gives 15 different bond setting and $E_{\text{cut}}$ combinations. FRODA was used to generate 2000 conformers along the positive and negative directions of ten normal modes. Ten conformers were recorded in both directions of each normal mode – 200 conformers for each of the 15 different bond/$E_{\text{cut}}$ settings. In total, 3000 conformers were recorded. This was repeated with the ligand-bound structure, yielding a total of 6000 conformers. The results here are representative of the data generated. We have necessarily been systematic in presenting these data.

We also simulated $m_{17}, ..., m_{26}$ for the ligand bound structure, but the distribution of average $M$ across the protein backbone did not change dramatically. The decision was made that the motion could be captured using the first ten normal modes. Based on previous studies, the majority of the large-scale dynamics of a protein can be captured using just a small number of normal modes [Alexandrov et al., 2005; Jimenez-Roldan et al., 2012]. Since storage was not a problem, we chose to use ten normal modes here. Initially, instead of using the mean when calculating $M$ as defined in Section 4.2.1, the maximum value was used. The resulting values were summed across the modes rather than averaged. It was found that the first few modes, involving movement of one flexible protein terminus, dominated these measures to a greater extent, thus failing to capture movement throughout the rest of the protein backbone. When summing the mobility values, the number of modes involved affected results more dramatically.

The pseudodihedral measure itself was preceded by the d4 measure, defined as the Euclidean distance between the $C_\alpha$ atoms of residues $i$ and $i+3$. Repeating this calculation for all residues and all conformers generated during the FRODA analysis gave a measure of the mobility of the residue relative to its neighbours. The pseudodihedral angle measure better captures the twisting and rotating movements of residues in the protein, and is also a more familiar method for capturing protein motion due to the ubiquity of $\phi$ and $\psi$ angles. Initially, the proportion of the surface area of the residue in contact with the surface was used as a measure of exposure. In HDX the signals disappear when the amide protons exchange with the solvent deuterons, and so the burial distance of the amide proton is more representative of the process of HDX. Data for these simulations are available but not presented in this thesis.

## 4.4 Discussion

### 4.4.1 Predicting the HDX folding core

The FIRST + FRODA folding core is a good match for the HDX folding core. With FRODA we account for surface exposure and with FIRST for rigidity. The intersection consists of residues which remain both buried and rigid. This selection factors in both hydrogen bonding and surface exposure, the two key parameters which dominate HDX [Woodward et al., 1982].

We quantified the agreement between the theoretical folding cores and the HDX folding core. The problem with the $s(AB)$ measure is that it is not comparable across different proteins, due to the variation in $N$ and $N_A$. As such there is no definitive value of $s(AB)$ which can be used as a benchmark, despite the use of $s(AB) = 1.5$ as a benchmark in previous work [Rader and Bahar, 2004]. As $N$ and $N_A$ change, so does the value of $s(AB)$ for the *perfect* theoretical model, $P$, which matches the HDX folding core exactly, so that $N_P = c(A, P)$. In the case of unbound CypA the perfect model has $N_P = c(A, P) = 73$, meaning that $s(AP) = 165/73 = 2.26$. With this in mind, we normalise $s(AB)$ by the score achieved by the perfect model, $s(AP)$. Now, $\frac{s(AB)}{s(AP)} = \frac{c(A,B)}{N_B}$. The measure $\%_A$ is exactly this, expressed as a percentage. Using $\%_A$ instead of $s(AB)$ ensures that the measure can be fairly compared with results using other proteins. The results have been normalised against a perfect score; the resulting percentage is a measure of how close to ideal the predictions are. There is a further problem with $s(AB)$, which is inherited by $\%_A$. It is for this reason that we introduce the additional measure of $\%_C$. A theoretical measure which predicts a folding core consisting of just one residue gets a high $\%_A$ score if the residue is also part of the HDX folding core. In fact, since $\%_A = 100 \times (\frac{c(A,B)}{N_B})$, in this case we have $\%_A = 100 \times \frac{1}{1} = 100\%$. The method is 100% accurate, but not very impressive! Similarly, since $c(AB) = N_B$, we get that $s(AB) = N/N_A$, which is the maximum value for $s(AB)$. In order to correct for this, we use the $\%_C$, which measures the *coverage* of the method – how close it comes to predicting all of the residues in the HDX folding core. The above method which predicted a single-residue folding core for CypA would have $\%_C = 100 \times (\frac{c(A,B)}{N_A}) \simeq 0.01\%$. Method $P$ would score 100% for both $\%_A$ and $\%_C$. When taken together, the measures $\%_A$ and $\%_C$ give a fuller picture of the quality of agreement than the single measure $s(AB)$.

### 4.4.2 Simulations of protein motion

The data from FRODA simulations was evaluated using pseudodihedral angles to track the motion. In the past, simulations with FRODA have involved pushing the protein as far as possible along the direction of a normal mode and calculating the root-mean-square deviation (RMSD) or fitted RMSD of the $C_\alpha$ atoms with respect to the original structure [Jimenez-Roldan et al., 2012]. These were plotted against the mode number to characterise the extent of motion along a particular mode. The pseudodihedral measure has allowed us to model the mobility of the protein across ten different normal modes, and extract a single mobility value for each residue. The mobility is influenced by the rigidity of the protein, and therefore the choice of $E_{\text{cut}}$. These simulations show that the input from the rigidity analysis is critical and great care must be taken in selecting $E_{\text{cut}}$ values for running simulations. The effect of ligand binding on protein rigidity is also reflected by the mobility data. This can be seen in Figure 4.8. The most common $E_{\mathrm{cut}}$ value for running FRODA simulations is $-1$ kcal/mol [Wells et al., 2005; Jolley et al., 2006; Macchiarulo et al., 2007]. For CypA, we note that at $E_{\mathrm{cut}}$ = -1 kcal/mol, the effect of ligand binding on protein mobility is not captured. As shown in Figure 4.7, the effect is greater at $E_{\mathrm{cut}}$ = -0.5 and -1.5 kcal mol. The RDM suggests that most $E_{\mathrm{cut}}$ values do not show an effect of ligand binding on mainchain protein rigidity. We recommend that when studying the impact of ligand binding on protein mobility, it is wise to eschew the default $E_{\mathrm{cut}}$ choice. Instead, running an RDM establishes interesting values of $E_{\mathrm{cut}}$. Rigidity analysis can be used in this way to predict a set of $E_{\mathrm{cut}}$ values with corresponding sets of residues which will be affected in terms of mobility upon ligand binding. The FRODA simulations establish the extent of the impact of ligand binding on mobility. Only the residues in or surrounding the binding site were affected in terms of surface exposure in the FRODA simulations. The set of buried residues did not change substantially with $E_{\mathrm{cut}}$. The buried residue measure was useful in improving the prediction of the HDX folding core.

The HB and HP settings did not alter the nature of the rigid-flexible transition for CypA. The HB setting did affect the folding core. The most stringent HB setting, HB0, demands that both bonding atoms be buried. The number of HBs at the start of the dilution is reduced to 149 as a result; the default setting has 242 HBs. The size of the folding core is reduced from 112 residues at the default setting to 73 residues. The smaller folding core is a better predictor of the HDX folding core as determined by the measures $s(AB)$, $\%_C$ and $\%_A$. The setting HB0 reduces the rigidity of the surface residues. This takes into account the influence of surface exposure on HDX. Protein mobility is influenced by the bond setting. The bond setting, along with $E_{\mathrm{cut}}$, determines the bond network and therefore the constraints placed on the protein during motion. Where a change in bond setting influences the rigidity of the protein, this will be reflected in the mobility data. An example of this is given in Figure 4.11. When the change in bond setting does not affect the rigidity of the protein the mobility will also remain unchanged, as in Figure 4.11.

# Chapter 5

# Expression, purification and characterisation CypA

## 5.1 Protein Expression

In order to study CypA experimentally, we first had to express the protein in competent cells and then isolate it from other proteins and cellular material. The protein was expressed in two different strains of E. coli competent cells, JM109 and Top10. For our purposes, we did not detect a difference between the two strains and so they were used interchangeably throughout. The DNA plasmid used for expression was kindly provided by Professor G. Fischer from the Max Planck research unit for enzymology of protein folding, Germany. The purpose of the plasmid is to take over the protein synthesis mechanism in the host cell, so that the protein coded for by the DNA sequence on the vector is overexpressed [Baneyx, 1999]. A diagram of the plasmid, which is a derivation of the expression vector pQE-70, is given in Figure 5.1. The origin of replication ensures that the plasmid can be replicated in E. coli cells. The plasmid encodes for proteins which confer resistance to the antibiotic ampicilin upon the bacteria which take it up. The T5 promoter binds RNA polymerase so that transcription can begin. The lac operons act as a switch for gene expression. A repressor binding to the lac operon prevents RNA polymerase from binding to the promoter. We use isopropyl-$\beta$-D-thiogalactopyranoside (IPTG) to control this interaction — IPTG binds the repressor and therefore allows transcription to proceed [Baneyx, 1999]. The plasmid encodes for the mRNA sequence which binds to the ribosome during translation. This is labelled RBS in Figure 5.1. The multiple cloning site allows for insertion of a DNA sequence encoding the desired protein. The plasmid we received contained the gene for CypA inserted into the multiple

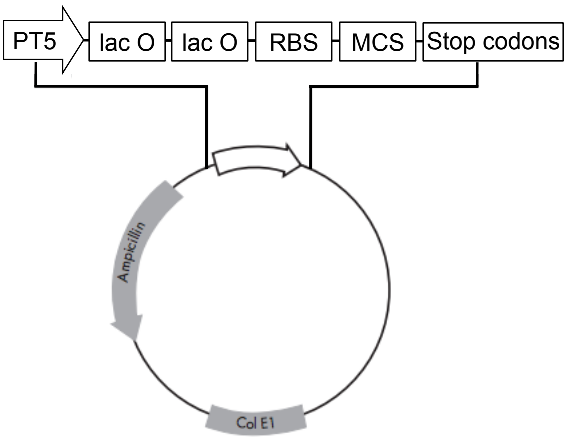


Figure 5.1: The anatomy of the expression vector derived from pQE-70 which was used for making recombinant CypA. The arrows on the plasmid indicate the direction of transcription. The section of the plasmid encoding for proteins which confer ampicilin resistance is labelled. Col E1 is the origin of replication. The clockwise arrow contains the T5 promoter (PT5), the lac operons (lac O), the section encoding for the ribosomal binding site (RBS), the multiple cloning site (MCS) and the stop codons.

cloning site. The stop codons are transcribed into mRNA stop codons which then terminate translation.

The concentration of the purified DNA plasmid was estimated from its absorbance peak at 260 nm to be 316.7 ng/$\mu$L. A sample of the plasmid solution was sent for DNA sequencing within the department. The results verified that the DNA in the plasmid indeed coded for the amino acid sequence of CypA (data not shown).

Competent cells were purchased from New England BioLabs. Stocks of 100 $\mu$L were defrosted and stored on ice before adding 1 $\mu$L of the CypA plasmid solution. After 15 minutes, that mixture was submerged in a water bath at 42°C for 30 seconds to induce a heat shock and encourage uptake of the plasmid, before being returned to ice for a further 15 minutes. Bacteria were grown in lysogeny broth (LB), a nutritionally rich medium. Initially bacterial colonies were grown overnight at 37°C on an LB agar plate, with ampicilin added so that only bacteria which had taken up the plasmid would grow. A single colony from the plate was then used to inoculate 50 mL of LB in a 250 mL conical flask containing 50 $\mu$L of 100 mg/mL ampicillin. This was incubated overnight at 37°C shaking at 180 rpm. In order to track bacterial growth, the optical density at 600 nm ($OD_{600}$) of the sample was measured compared to an LB-only reference, using a Jenway 6305 spectrophotometer. The appropriate volume was used to inoculate a larger culture such that the $OD_{600}$ of the larger culture was 0.05. Typically, 10 mL of the overnight culture would be enough to inoculate 400 mL of LB. Two-litre flasks were used, each containing 400 mL of culture. The selection pressure was kept up by adding 400 $\mu$L of 100 mg/mL ampicillin. The cultures were incubated at 37°C shaking at 180 rpm until $OD_{600} = 0.5$, which typically took 1.5 – 2 hours. At this point 400 $\mu$L of 1 M IPTG was added to induce the overexpression of the recombinant protein. The cultures were then returned to the incubator at 37°C shaking at 180 rpm for six hours. The process of optimisation included varying the growth time after adding IPTG, with time periods of four, six, twelve and sixteen hours tested, and six hours found to be optimal (data not shown). After six hours had elapsed the cells were harvested by centrifugation at 36,000 $\times g$ for 15 minutes at 4 °C. Each cell pellet was resuspended in 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffer at pH = 6.5. The cell pellets were resuspended in a volume equal to 5 % of the culture volume. Cells were then frozen to lyse them. At this point the cells could be stored in the freezer at $-20$°C until it was convenient to begin the purification procedure. Typically they were stored overnight, but could be stored in this way for up to two months.

### 5.1.1 Expression of $^{15}$N-labelled CypA

We needed to prepare $^{15}$N-labelled CypA for NMR experiments, so adapted the above expression protocol accordingly. LB could no longer be used as the growth medium since it contains peptides and therefore $^{14}$N nuclei. Instead, a growth medium containing the minimal nutrients required for bacterial growth, referred to as minimal medium (MM), was used. Table 5.1 lists the ingredients used to prepare MM. Stock solutions for all of the ingredients were prepared and sterilised either by autoclaving or by filter sterilisation. The medium contains no amino acids, and a single nitrogen source, $(NH_4)_2SO_4$. The salts provide essential elements for synthesis of nucleic acids and amino acids. Initially MM was prepared with $(^{14}NH_4)_2SO_4$ as the nitrogen source, allowing us to verify that protein could be successfully expressed in the MM before the more expensive $(^{15}NH_4)_2SO_4$ was used to produce $^{15}$N-labelled protein. The nitrogen source was restricted to ensure that the amino acids synthesised by bacteria grown in the medium contained $^{15}$N in place of $^{14}$N, and so $^{15}$N-labelled CypA was expressed. The $(^{15}NH_4)_2SO_4$ was obtained from Cambridge Isotope Laboratories, and had purity of $> 99\%$.

Bacterial colonies were grown on agar LB plates with ampicilin as before, and then added to 50 mL MM overnight. To encourage growth in the MM the inoculation volume was increased so that the large cultures (400 mL) had an initial $OD_{600}$ of

Table 5.1: The nutrients added to water to make 1 L of MM.

| Nutrient | Mass (mg) | Nutrient | Mass (mg) |
|---|---|---|---|
| $(NH_4)_2SO_4$ | 1000 | $ZnCl_2$ | 0.5 |
| Glucose | 4000 | $CuCl_2$ | 0.1 |
| $Na_2HPO_4$ | 6800 | $CoCl_2$ | 0.1 |
| $KH_2PO_4$ | 3000 | $H_3BO_3$ | 0.1 |
| NaCl | 500 | $MgSO_4$ | 120 |
| $Na_2SO_4$ | 43 | $CaCl_2$ | 33 |
| EDTA | 50 | d-Biotin | 1 |
| $MnCl_2$ | 16 | Thiamine | 1 |
| $FeCl_3$ | 5 | | |

0.1 instead of 0.05. In MM the growth cultures typically reached $OD_{600} = 0.5$ after 2 – 2.5 hours. Overexpression of CypA was induced with 400 $\mu$L 1 M IPTG. During the preparatory protein expressions carried out with $(^{14}NH_4)_2SO_4$ MM we found that the yield improved with a longer growth time following induction. In subsequent MM preparations with $(^{15}NH_4)_2SO_4$, cultures were grown overnight at 37°C, shaking at 180 rpm. The cells were then harvested and the resulting pellets resuspended as before.

### 5.1.2 Electrophoresis analysis: SDS-PAGE

We used sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) to analyse protein samples. SDS is a detergent which causes proteins to unfold and also imparts an even negative charge on the protein. A mixed sample of negatively charged proteins is placed in a well at the top of a polyacrylamide gel. The gel is submerged in a buffer and an electric field is applied across the gel. The negatively charged particles travel away from the negative electrode at the top of the gel and towards the positive electrode at the bottom. The speed of proteins through the gel is dependent upon their mass to charge ratio. In native gels, the proteins are not denatured and so their passage through the gel is also dependent on cross sectional area. Since SDS denatures the proteins and distributes charge evenly, two proteins of the same size will have the same cross sectional area and also the same mass-to-charge ratios. the speed of proteins through the gel in SDS-PAGE is therefore determined by the mass of the protein. The electric field is switched off when the proteins in the mixture have partially traversed the gel. A dye which binds to proteins is then added to the gel so that one can see how far the proteins travelled before the field was switched off. Typically one of the wells in the gel is reserved for a sample containing a mix of proteins, each of known mass. The separation on the gel of the proteins from this sample acts as a molecular ruler against which one can estimate the protein masses in the other wells.

We prepared discontinuous gels with a stacking gel of 7 % acrylamide and pH 6.8 on top of a resolving gel of 16 % acrylamide and pH 8.8. Proteins move through the stacking gel and become concentrated in a sharp band at the top of the resolving gel, with the more mobile proteins a the bottom of the band. The dense resolving gel has narrow pores, allowing small proteins to travel faster than large proteins. This encourages separation of the smaller proteins on the gel. The gels were cast using the Mini Protean II gel kit (Bio-Rad). The loading buffer we used contained 1 M TRIS-HCl pH 6.8, 40 % glycerol, 0.8 % SDS and 0.1 % w/v bromophenol blue, to which was added 10 % $\beta$-mercaptoethanol. The samples in loading buffer we heated to 90°C for 5 minutes before being cooled to 4°C for two minutes. This further denatures the proteins by reducing any disulfide bonds. The samples were run through the gel at the same time as a low molecular weight marker, which was obtained from GE Healthcare. The low molecular weight marker contained a mix of proteins sized 94, 67, 43, 30, 20 and 14 kDa. The band for CypA is therefore expected to be between the bottom two bands of the marker. A voltage of 80 V was applied for ten minutes so that the proteins travelled through the wells to the base of the stacking gel. The voltage was then increased to 180 V for 55 minutes. Once completed, the gels were stained using the dye Coomassie brilliant blue R250. A destaining solution consisting of water, methanol and acetic acid in a volume ratio of 5 : 4 : 1, was added to remove any dye which had not bound to a protein.

Figure 5.2 shows a section of an SDS-PAGE gel demonstrating that CypA has been overexpressed in bacterial cells. The sample in the right-hand well (B) was taken prior to purification. There are many proteins in this sample with a variety of masses, the most abundant of which is CypA with mass 18 kDa. The relative abundance is suggested by the thick band, labeled as CypA in Figure 5.2.

## 5.2 Purification of CypA

We needed to isolate CypA from the other proteins produced by the bacteria during culture growth. The cell pellets were defrosted in a room temperature water bath. The cells were further lysed by two 30 second periods of sonication using a Bandelin Sonoplus HD 2070 sonicator at 70 % power. There was a 90 second interval between the sonication periods whilst the cells were stored on ice to prevent over-

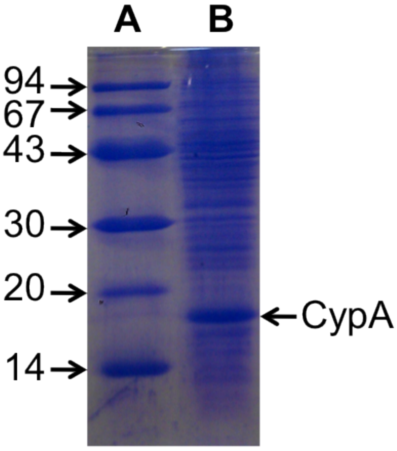


Figure 5.2: A section of an SDS-PAGE protein gel showing that CypA has been expressed. Column A shows the low molecular weight marker, containing a mixture of proteins of known mass. The bands corresponding these proteins are labelled with the respective masses in kDa. Column B shows the mixture of proteins expressed in the bacterial culture. The thick band indicates that CypA, of mass 18 kDa, has been overexpressed.

heating. Following sonication, the broken cells were centrifuged for 15 minutes at 4°C and 17,500 $\times g$ to pellet all insoluble cell debris. The pellet was discarded and the supernatant, containing a mixture of proteins including CypA, was retained. A demonstration of the typical composition of this mixture is given in column B of Figure 5.2. In this mixture CypA is abundant, although far from pure. We developed a method for protein purification, established over a period of months. Initially the protocol used was similar to the method outlined in [Liu et al., 1990], but this was adapted to suit the resources available to us and overcome problems we encountered when using that method. The protocol we developed involves ion exchange chromatography (IEC) and size exclusion chromatography (SEC). Earlier versions of our protocol incorporated diethylaminoethanol-sepharose anion exchange and ammonium sulphate precipitation. These steps were later deleted when results from SDS-PAGE suggested that purification was not improved by their inclusion (data not shown). Removing these steps shortened the purification procedure, and therefore also limited the number of steps at which the yield may be reduced.

### 5.2.1 Ion exchange chromatography

IEC can be used to separate a mixture of proteins species based on their overall net charge. The protein sample is loaded onto an ion exchange column which contains a charged resin. As the protein solution flows through the column, proteins bind to the resin on the basis of their net charge. We used cation exchange chromatography, where the resin is negatively charged and so binds proteins with positive net charge. The stronger the net positive charge, the stronger the protein binds to the resin. Once the sample has passed through the column, a steady gradient of increasing NaCl concentration was applied. The $Na^+$ ions displace the proteins, binding to the negatively charged resin more strongly. The first proteins to be displaced are those which have the least affinity for the column resin. Proteins are therefore eluted from the column in order of increasing positive net charge. The material eluted from the column is collected in small volumes (fractions) so that the proteins are separated. As the material is eluted, the presence of protein is detected by measuring absorbance at 280 nm ($A_{280}$).

Initially, 20 mM tris(hydroxymethyl)aminomethane (TRIS) buffer at pH = 7.8 was used during the expression and purification [Liu et al., 1990]. The isoelectric point of CypA — the pH value at which the net charge is zero — is estimated to be 7.68 [ProtParam tool, 2013]. At pH = 7.8 therefore, CypA has a small net negative charge. We therefore tried using an anion exchange column. Under these conditions, a lot of protein was lost during ion exchange due to most of the CypA not binding to the column. In response, the pH of the system was lowered to below 7.68 and cation exchange was used. At lower pH we observed improved binding of CypA to the ion exchange column and hence a better yield during purification. In order to lower the pH further from the isoelectric point the buffer system was changed to HEPES which is more effective at lower pH than TRIS. Although some material was still lost by not binding to the ion exchange column, the yield became large enough for us to continue. The pH was not lowered further since yields were satisfactory and CypA remained stable in HEPES at pH = 6.5. We used a 10 mL Source 30S column (from GE Healthcare) containing a resin with a negative charge for cation exchange. The column was connected to an ÄKTA purifier 100 fast protein liquid chromatography (FPLC) system (from GE Healthcare). The column, which was stored in ethanol, was first equilibrated in 20 mM HEPES buffer, pH 6.5. This was done by running at least 30 mL (three column volumes) of buffer through the column. The protein sample was loaded onto the column using an injection loop. A gradient from 0 –100 % of a second buffer of 20 mM HEPES with 150 mM NaCl at pH 6.5 was added over 40 minutes. The buffers were run through the ÄKTA purifier at 5.0 mL/min, and so a total of 200 mL was passed through the column — equivalent to 20 column volumes. Eluted fractions were collected in volumes of 2.5 mL throughout, with protein elution monitored using UV absorbance at 280 nm

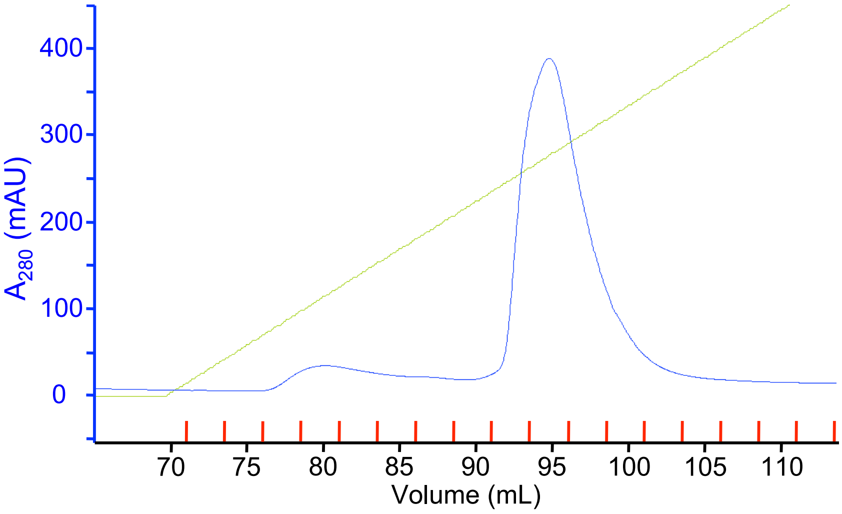


Figure 5.3: An ion exchange chromatogram showing elution of CypA. The $A_{280}$ in mAU, shown in blue, is recorded as material is eluted from the column. The space between two red dashes indicates one 2.5 mL fraction. The green line represents the gradient of [NaCl], which increases from 0.0 mM at 70 mL to 18.75 mM at 95 mL.

($A_{280}$). Figure 5.3 shows an example of an ion exchange chromatogram for CypA. The blue absorbance peak shows where the majority of protein has been eluted from the column.

The composition of the fractions corresponding to large $A_{280}$ peaks in the ion exchange chromatogram were analysed using SDS-PAGE. Figure 5.4 is an example of such an SDS-PAGE gel. These samples are far purer than the sample before ion exchange (Figure 5.2), yet there remains a protein of higher molecular weight in the sample along with CypA. The bands in the gel corresponding to this unwanted protein are labelled with an X in Figure 5.4. The columns labelled B in Figure 5.4 show the protein in the sample which passed through the ion exchange column without binding to the resin. Clearly there is some CypA which has not bound to the column. This could be because of the column is overloaded and so some CypA cannot bind. The affinity of CypA for the column is not particularly strong, since CypA is eluted at a low concentration of NaCl. To counteract this problem, the sample was split in half prior to loading and then ran two IEC experiments were run in sequence — one for each half of the protein sample. This appeared to decrease the relative abundance of CypA in the fractions which did not bind to the columns (data not shown). There was not a notable increase in final protein yield, however, and the problem of some protein not binding to the column still remained. Since IEC separates proteins according to their net overall charge, it is possible for

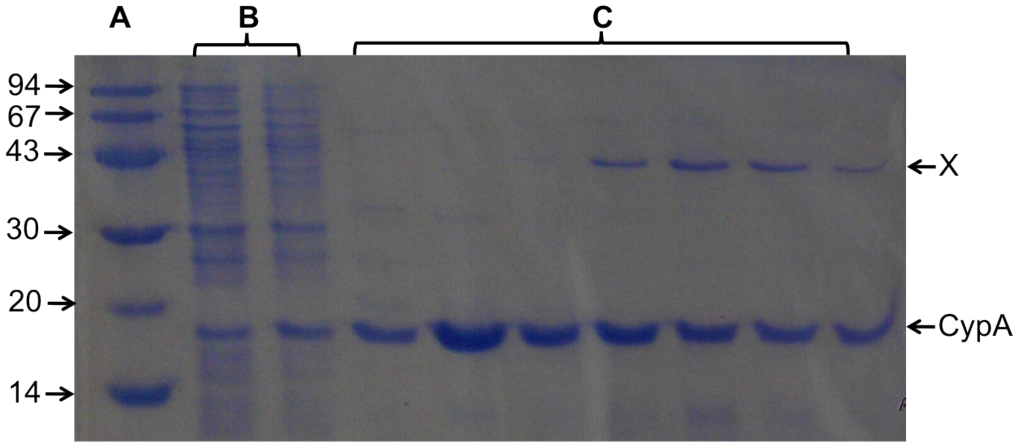


Figure 5.4: An SDS-PAGE gel showing incomplete purification as a result of IEC. Column A is the low molecular weight marker. The mass in kDa for each protein component of the marker is labelled. The columns B show the material which did not bind to the column. Columns C are selected fractions following an IEC purification step. The bands corresponding to CypA are labelled. The bands for the unwanted higher molecular weight protein are labelled with an X.

two or more different proteins with different masses to be eluted from the column at the same time. The SDS-PAGE gel in Figure 5.4 demonstrates that this has indeed happened. Nevertheless, as shown by the columns labelled C in Figure 5.4, IEC succeeded isolating CypA from most of the other proteins in the initial sample.

### 5.2.2 Size exclusion chromatography

After IEC, CypA had not been completely isolated and so we required a further purification step. We used SEC, which separates proteins based on size rather than charge. The SEC column is packed with material composed of inert porous spheres. The pores and the space between the spheres fill with the buffer being used. Small molecules enter the spheres and therefore travel through a greater volume of the column material than larger molecules which do not enter the spheres. Larger proteins therefore traverse the column more rapidly and are eluted first. We used a 23.56 mL Superdex 75 gel filtration column (GE Healthcare), connected to an ÄKTA purifier 100 FPLC (GE Healthcare). We prepared an SEC buffer of 4.2 mM $NaH_2PO_4$, 15.8 mM $Na_2HPO_4$ and 150 mM NaCl at pH 6.5. The column, stored in ethanol, was equilibrated first into water and then into the buffer, each running at 0.2 mL/min for 40 mL. The protein sample loaded onto the column consisted of pooled fractions eluted during IEC. The fractions were pooled on the basis of

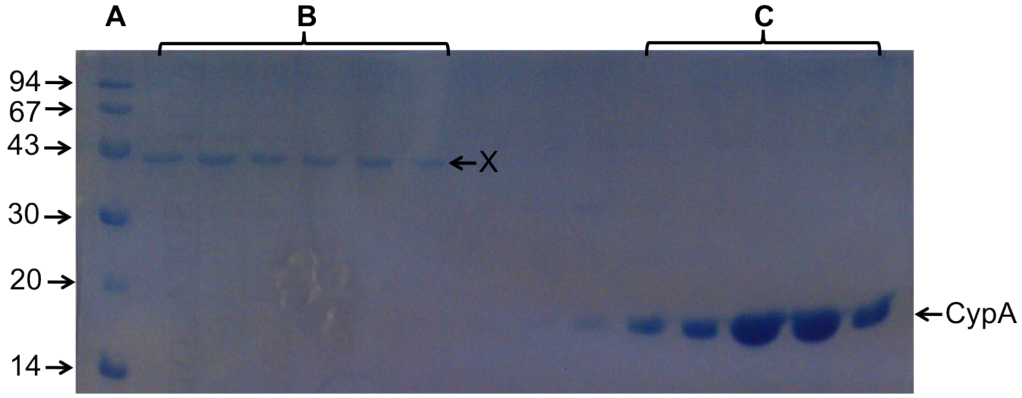


Figure 5.5: An SDS-PAGE gel showing purified CypA as a result of size exclusion chromatography. Column A is the low molecular weight marker. The mass in kDa for each protein component of the marker is labelled. The columns B show fractions containing the isolated unwanted higher molecular weight protein, labelled with an X. Columns C show the fractions containing purified CypA.

the IEC chromatogram and SDS-PAGE analysis. To ensure that molecules of the same size are eluted from the SEC column at the same time, it is important to load them on to the column approximately simultaneously. The pooled fractions were therefore concentrated to 0.5 mL before loading onto the column. We used a 5000 molecular weight cut-off vivaspin column (from Sarstedt) to concentrate the fractions, centrifuged at 3220 $\times g$ and 4°C. The sample was loaded onto the SEC column using a 1 mL injection loop, and passed through the column at a flow rate of 0.5 mL/min. The dead volume was 8 mL, after which 0.3 mL fractions were collected for the following 13 mL of eluting material. Protein elution was monitored at $\mathrm{A}_{280}$. The fractions corresponding to absorbance peaks were analysed by SDS PAGE to determine which contained purified CypA. SEC was successful in isolating CypA from the higher molecular weight protein, as shown in Figure 5.5, which is an SDS-PAGE of fractions collected from SEC.

### 5.2.3 Estimating protein concentration

The fractions corresponding to purified CypA as determined by the SEC chromatogram and SDS-PAGE were pooled, and the $\mathrm{A}_{280}$ was measured with a Jenway 6305 spectrophotometer. SEC buffer was used as a reference. Protein concentration $c$ is related to $\mathrm{A}_{280}$ by the Beer-Lambert law

$$c = \mathrm{A}_{280}/(\epsilon_{280} \times l), \tag{5.1}$$

where $\epsilon_{280}$ is the molar extinction coefficient of the protein at 280 nm and $l$ is the path length of the cuvette in cm. The units for $c$ are M and for $\epsilon_{280}$ they are $M^{-1}cm^{-1}$. For CypA, $\epsilon_{280} = 8480\ M^{-1}cm^{-1}$ [ProtParam tool, 2013]. From the estimated concentration, the protein yield was calculated. For the unlabelled protein, a growth volume of $6x400$ mL typically produced around 20 mg CypA — a yield of 8.34 mg per litre of culture. This yield is lower than the yield stated in published work, which was 17.5 mg per litre of culture [Liu et al., 1990]. When growing $^{15}N$ labelled protein, the yield decreased by approximately one half. Since our yields were sufficient for performing CD, fluorescence spectroscopy and NMR, the protocol was not further optimised. At this stage, unlabelled protein was either used immediately or concentrated as before using SEC and then divided into small aliquots and stored in the freezer at $-20$°C. $^{15}N$-labelled protein for NMR experiments was concentrated and used immediately if possible, or stored at 5°C until required. The $^{15}N$-labelled protein used for the HDX experiments was dialysed and freeze-dried as described in Chapter 6.

## 5.3 Characterisation of CypA

### 5.3.1 Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy can be used to study the secondary structure of proteins. In a CD spectrophotometer, plane polarised light is generated and then split into its two circularly polarised components (anti-clockwise and clockwise, referred to as $L$ and $R$ respectively). When this light is passed through the sample, a chiral molecule such as a protein will absorb the $L$ and $R$ components to different extents. CD spectrophotometers measure this difference in absorbance, $\Delta A = A_L - A_R$ [Kelly et al., 2005]. If $\Delta A \neq 0$, then the light which reaches the detection unit, the photomultiplier, is elliptically polarised. CD signal is reported in terms of the ellipticity of the polarised light detected, in degrees. The CD effect is typically a very small difference between relatively large $A_L$ and $A_R$; the CD signals shown in Figures 5.6 – 5.9 are on the order of 10 mdeg.

The CD signal depends on the wavelength of the polarised light. Absorption in far-UV region of the spectrum (240 nm and below) is due principally to the peptide bond [Kelly et al., 2005]. Sections of the protein with well-defined secondary structure will exhibit recognisable absorbance patterns at certain wavelengths within the UV spectrum [Greenfield and Fasman, 1969]. For example, a large positive peak at 198 nm corresponds to $\beta$-sheet structure and a negative peak at 222 nm corresponds to $\alpha$-helix [Corrêa and Ramos, 2009]. The CD signal of a protein

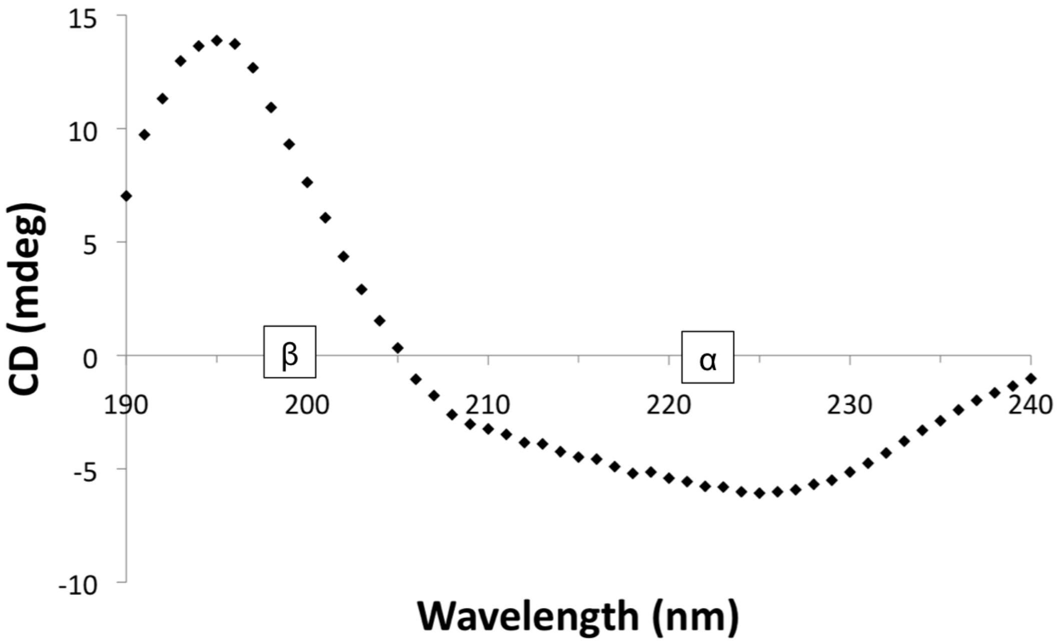


Figure 5.6: Far-UV CD spectrum of CypA at 25°C. The average CD signal from 16 scans is plotted at 1 nm intervals between 190 nm and 240 nm. The wavelengths corresponding to signature $\alpha$-helix and $\beta$-sheet signals are labelled $\alpha$ and $\beta$ respectively on the $x$ axis.

is approximately a proportional sum of the spectra resulting from its secondary structural units [Nordén et al., 2010]. The CD spectrum is known for many proteins and there are large databases available which which to compare new spectral data [Kelly et al., 2005]. In this way, the percentage composition of $\alpha$-helix and $\beta$-sheet components can be estimated. To estimate the secondary structure composition for purified CypA in order to verify that it was correctly folded, the far-UV CD spectrum of CypA was measured at 25°C using a Jasco J-815 CD spectropolarimeter. The resulting spectrum is shown in Figure 5.6. All CD spectra presented in this section were recorded for protein samples of 0.1 mg/mL in 5 mM sodium phosphate, pH 7.3 unless otherwise stated. Small volumes of concentrated purified protein were diluted in sodium phosphate buffer and the resulting concentration confirmed by measuring $A_{280}$. Data points were collected during far-UV scans between 180 nm and 260 nm at 1 nm intervals. A baseline spectrum was recorded in the same way for a sample containing buffer only. To generate Figure 5.6, 16 scans were collected at 100 nm/min in continuous scanning mode, and the average CD signal minus the average baseline signal was plotted. We show the data collected between 190 nm and 240 nm at 1 nm intervals. The data was analysed with Dichroweb [Whitmore

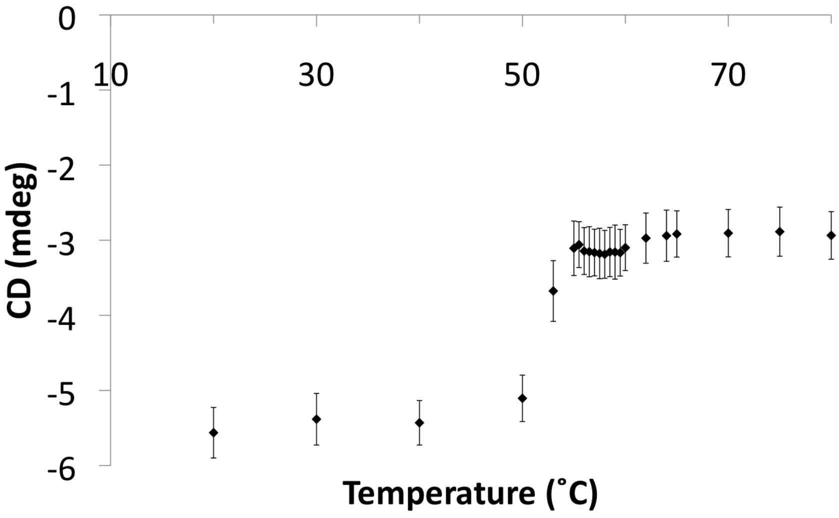


Figure 5.7: Thermal denaturation of CypA as measured by CD signal at 222 nm. At each temperature increment, the sample was equilibrated for ten minutes and then data was recorded for two minutes. The average CD signal from 240 data points is plotted, with error bars showing standard deviation.

and Wallace, 2004], using the reference database SP175 to calculate the proportional secondary structure composition of CypA. Our protein sample was determined to be 18 % helix and 33 % sheet. We compared this with the results of the DSSP algorithm applied to five X-ray crystal structures in the PDB. The average composition of these structures was 13 % $\alpha$-helix and 32 % $\beta$-sheet.

We also used CD spectrometry to assess the stability of the folded protein. To do this, the CD signal at 222 nm was measured at a series of strictly increasing temperatures between 20 °C and 80 °C. The CD signal at 222 nm is the signature signal of the $\alpha$-helical regions of the protein [Corrêa and Ramos, 2009], and so as the protein unfolds with increasing temperature, the strength of the signal is reduced. At each temperature point in the series, the sample was equilibrated for ten minutes. The CD signal at 222 nm was then recorded for two minutes at 0.5 second intervals and the average recorded value was determined. Figure 5.7 shows the average recorded CD signal at each temperature, along with the standard deviation of the signal during the two minute recording period. Initially the sample was held for three minutes at each temperature prior to recording. This time was increased to ten minutes in order to reduce the standard deviation of the recorded CD signals during the two subsequent minutes. The CD signal at 222 nm decreases rapidly between 50°C and 55°C and then remains approximately constant between

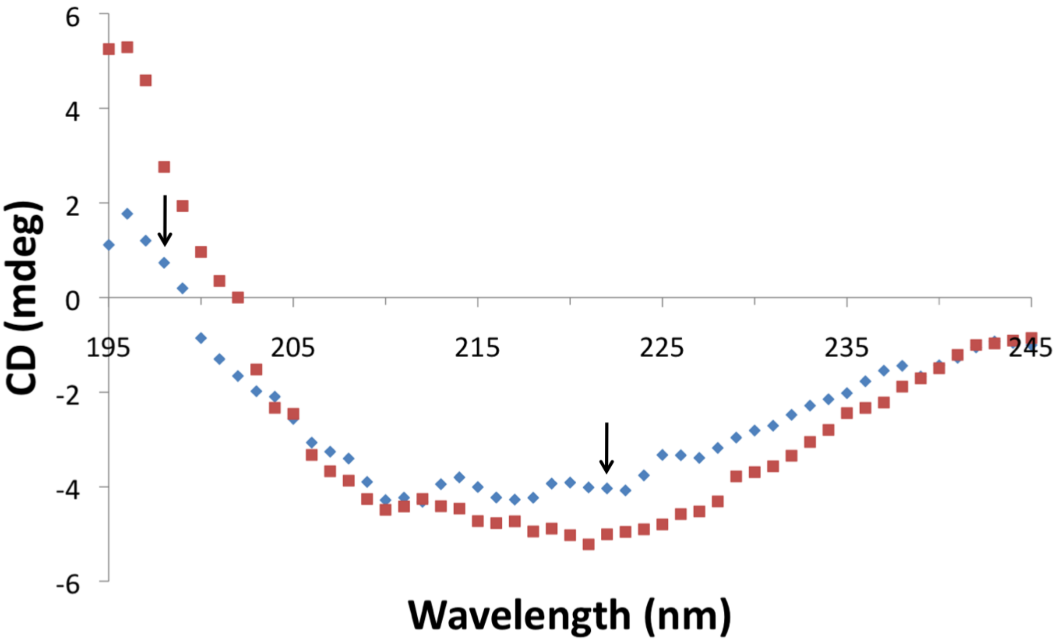


Figure 5.8: Red: the CD spectrum of folded CypA at 20°C. Blue: the spectrum of CypA after heating to 80°C and then cooling to 20°C. Black arrows at 198 nm ($\beta$-sheet) and 222 nm $\alpha$-helix indicate the location of signature CD signals from secondary structure motifs.

55-80°C. This suggested that the protein unfolds between 50°C and 55°C. The CD signal is approximately constant between 20-50°C, suggesting that CypA is stable in this temperature range for this pH and buffer system.

The refolding of CypA was investigated by cooling the 80°C sample to 20°C and recording the far-UV CD spectrum. This is shown in Figure 5.8 along with the initial spectrum for the sample recorded before thermal denaturation, at 20°C. CypA has not successfully refolded, although the signal at 222 nm is stronger than at 80°C (Figure 5.7), suggesting that the protein has regained some of its secondary structure.

The structural stability of CypA at different pH levels was studied using CD. At pH $= 7.3, 6.8$ and 6.3, 5 mM sodium phosphate buffer was used. In order to lower the pH further, a 5mM acetate buffer was used at pH $= 6.3, 5.8$ and 5.3. A thermal denaturation experiment was conducted for the two buffer systems at each pH. The temperature was increased from 40-60°C at 5°C intervals. At each temperature point the sample was equilibrated for ten minutes, and then CD signal was recorded every 0.5 seconds for two minutes. In Figure 5.9, the average and standard deviation of the recorded CD signal at each temperature is plotted for

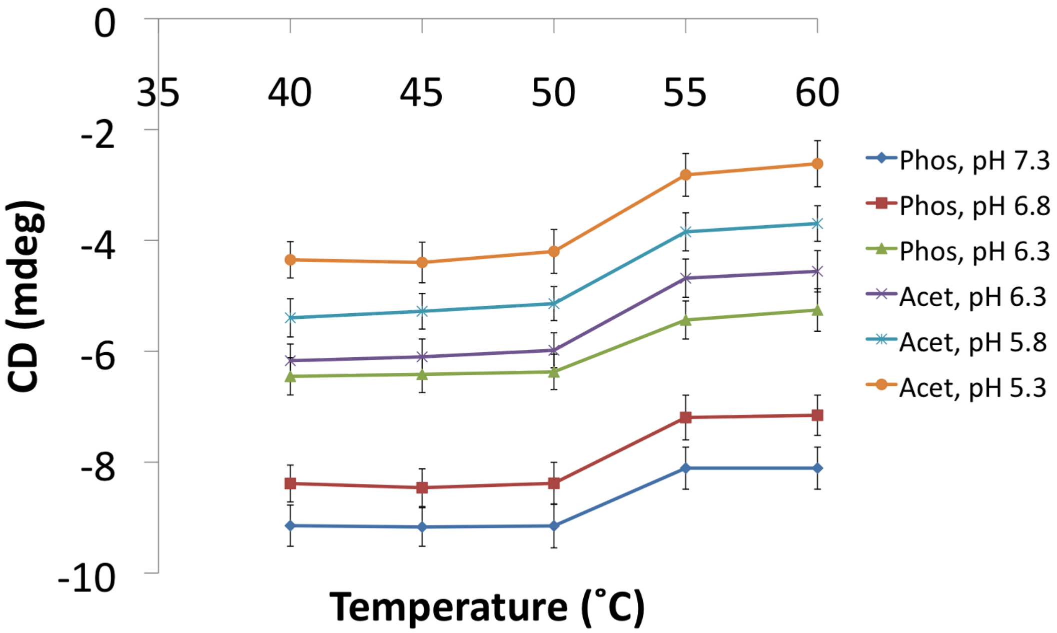


Figure 5.9: Thermal denaturation of CypA under different buffer systems and pH. At each temperature increment, the sample was equilibrated for ten minutes and then data was recorded for two minutes. The average CD signal from 240 data points is plotted, with error bars showing standard deviation. 5 mM sodium phosphate buffer (Phos) and 5 mM acetate buffer (Acet) were used at the pH levels indicated.

the six experiments. Each time we observe a clear change in CD signal at 222 nm between 50°C and 55°C, showing that the unfolding transition for CypA remains in this temperature range for pH between 5.3 and 7.3. We note that in Figure 5.9 the CD signal at 40°C decreases with increasing pH, and that this separation remains consistent at each temperature. We are unsure of the origins of this effect; we did not investigate this further.

### 5.3.2 Fluorescence spectroscopy

Fluorescence occurs when electrons in a sample are excited by an incident beam of electromagnetic radiation such as UV light. The energy absorbed is subsequently emitted at a lower wavelength, typically as visible light. The fluorescence of a protein in its native state is due to the fluorescence of its aromatic residues. At an excitation wavelength of 290 nm, the fluorescence emission is dominated by tryptophan fluorescence. The fluorescence of Trp is highly dependent on its environment — the fluorescence emission intensity decreases as the polarity of its environment increases. We use this effect to study the binding of CsA to CypA. CypA contains a single Trp residue, Trp121, which is located near the binding site. When CypA is in its unbound state, Trp121 is exposed to the polar solvent. Upon ligand binding, Trp121 becomes shielded from the solvent and its fluorescence is expected to increase as a result [Liu et al., 1990; Husi and Zurini, 1994]. CsA was titrated into a solution of CypA and fluorescence spectroscopy was used to monitor the change in tryptophan fluorescence. The experiment was conducted using a Photon Technology International fluorimeter. A stock solution of CypA (65 $\mu$M) was diluted 1 mL in 50 mM TRIS buffer at pH 7.3 so that the final CypA concentration was 4.5 $\mu$M. CsA was stored in ethanol at a concentration of 1.0 mM. A stock solution of 0.1 mM CsA for the titration was made by diluting this ten-fold in TRIS buffer. Each titre consisted of 10 $\mu$L of CsA stock, and so increased the total concentration of CsA in the fluorescence sample by approximately 1 $\mu$M. After each increment in [CsA] the sample was mixed using a pipette. The emission spectra were recorded at 1 nm intervals between 300 nm and 400 nm, using an excitation wavelength of 290 nm. The average fluorescence emission of three scans was recorded for each wavelength, using a slit width of 2 nm. Two baselines were recorded, for samples containing TRIS buffer only and 5 $\mu$M CsA in TRIS buffer. There was no significant difference between the baselines (data not shown). Baseline adjusted fluorescence emission spectra for CypA with various concentrations of CsA are shown in Figure 5.10. The key shows the concentration fraction [CsA]/[CypA]. We observe a steady increase in fluorescence with increasing [CsA] until [CsA]/[CypA] = 1. After this point is reached, adding more CsA does not enhance fluorescence.

To compare our data with previously published work, [CsA]/[CypA] was plotted against the fractional change in fluorescence [Liu et al., 1990; Husi and Zurini, 1994]. The is shown in Figure 5.11, along with the same plot using data from [Liu et al., 1990]. The peak of fluorescence emission was taken to be 340 nm. The fluorescence at this wavelength was taken from each spectrum in Figure 5.10 and the baseline measurement was subtracted. Results were then normalised by the largest of these values so that the fractional fluorescence change could be plotted on a scale from 0 to 1. Each addition of 10 $\mu$L of the CsA stock increased the concentration ratio by approximately 0.22. The datasets in Figure 5.11 both show a rapid increase in fractional fluorescence change with increasing relative concentration of CsA before it levels off when the concentration of CsA and CypA are equal ([CsA]/[CypA] = 1). This shows that the protein is indeed binding to CsA. There are differences between the datasets, in particular around [CsA]/[CypA] = 0.2, where our data shows a smaller initial increase in fluorescence. In both datasets, however,

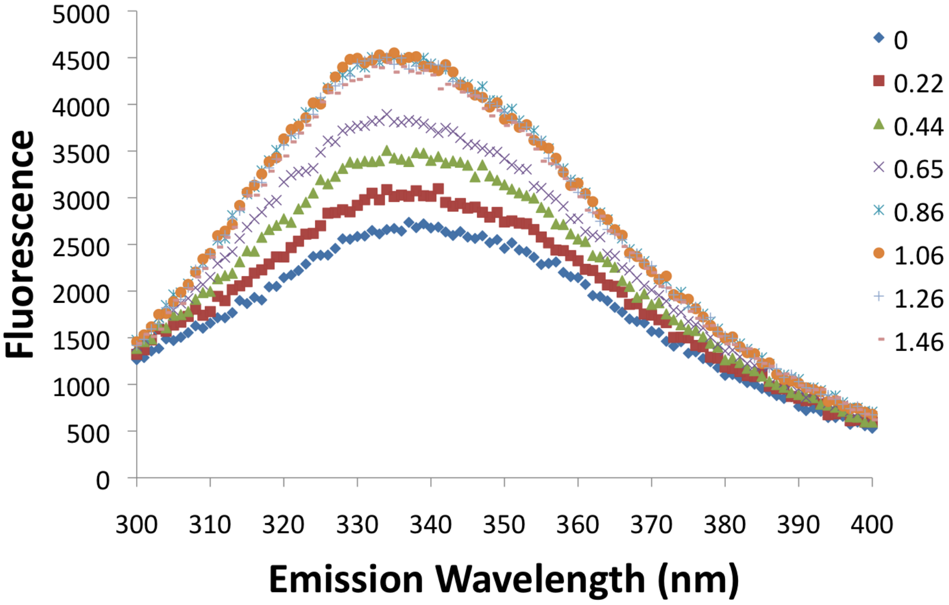


Figure 5.10: Fluorescence emission spectra of CypA with increasing CsA concentration. The key shows the concentration ratio [CsA]/[CypA]. Emission spectra were recorded at 1 nm intervals between 300 nm and 400 nm with excitation wavelength of 290 nm. The average emission value of three scans is plotted.

we observe fluorescence enhancement followed by saturation at equal concentrations of protein and ligand.

## 5.4 Discussion

In this chapter we established protocols for producing purified both unlabelled CypA and $^{15}$N-labelled CypA. We used SDS-PAGE to verify expression and analyse protein samples during the purification. The concentration of the purified protein, and therefore the yield, was estimated using $A_{280}$. CD spectroscopy was used to assess the fold of the protein, which compared well with the secondary structure composition in a selection of X-ray structures. We also used CD to investigate the thermal stability of the protein and found a temperature interval, 50-55°C, in which CypA becomes denatured. This temperature interval remained consistent at six different pH levels, measured in two different buffer systems. We observed incomplete refolding of CypA following thermal denaturation. We studied the binding of CsA to CypA using fluorescence spectroscopy. The fluorescence enhancement of Trp121 upon CsA binding was observed, reaching a plateau at [CsA]/[CypA]= 1, suggesting that all CypA molecules had bound at this concentration ratio. These

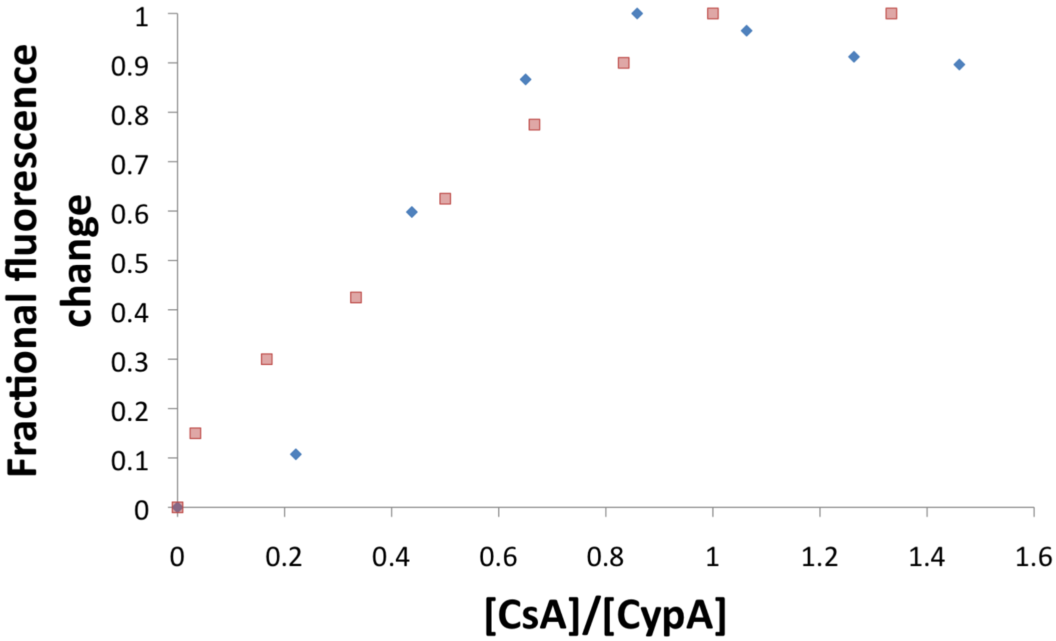


Figure 5.11: Fractional fluoresence change during the CypA-CsA titration. Data points taken from Figure 5.10 are plotted in blue, with data from [Liu et al., 1990] in red.

characterisations represent critical steps prior to performing NMR spectroscopy on the $^{15}$N-labelled CypA in its unbound state and in complex with CsA.

# Chapter 6

# Hydrogen-deuterium exchange NMR experiments

## 6.1 Introduction

In this chapter we present and analyse our NMR experiments on CypA in the presence and absence of the ligand CsA. First we assign $^{15}$N-$^{1}$H-HSQC spectra for unbound CypA and the CypA-CsA complex using data from 3D $^{15}$N-$^{1}$H TOCSY-HSQC and $^{15}$N-$^{1}$H NOESY-HSQC NMR data. The TOCSY-HSQC data enable us to identify the type of amino acid associated with each crosspeak in the HSQC spectrum. We then use sequential assignment of the NOESY-HSQC data to determine the position of each amino acid in the primary structure of the protein. Assigned HSQC spectra are essential for the interpretation of the HDX experiments which form the second part of this chapter. From these HDX experiments we determine the HDX folding core — as discussed in Chapter 4 — for both CypA and the CypA-CsA complex.

CypA has been studied using NMR spectroscopy previously. There are two sets of assigned NMR data available, both of which are studies of the CypA-CsA complex [Neri et al., 1991; Spitzfaden et al., 1994]. Although unbound CypA has also been studied using NMR, a publicly accessible assigned dataset was not available. There are HSQC spectra of unbound CypA published [Neri et al., 1991; Shi et al., 2006]. We expect some discrepancies between our data and those already published, since chemical shifts are dependent on pH, temperature and buffer composition. This means that the assignment procedure is not trivial for either sample. We compare the spectra for CypA-CsA and for unbound CypA in order to determine the impact of ligand binding on the chemical environment of each residue. In

Appendix C (Table C.1) we list all the backbone N-H assignments we have made for both samples. HDX experiments on unbound CypA have been previously reported [Shi et al., 2006]. To our knowledge the present work is the first time that HDX experiments have been carried out on the CypA-CsA complex. We compare the HDX folding cores determined in our experiments with each other as well as with data for unbound CypA with the previously published experiments. We also discuss the HDX folding cores in the context of our computationally determined folding cores using the methods of Chapter 4.

## 6.2 Methods

### 6.2.1 Sample preparation

$^{15}$N-labelled CypA was expressed and purified according to the protocol given in Chapter 5. NMR samples for the 3D $^{15}$N-$^{1}$H TOCSY-HSQC and $^{15}$N-$^{1}$H NOESY-HSQC experiments consisted of purified $^{15}$N-labelled CypA in SEC buffer (4.2 mM $NaH_2PO_4$, 15.8 mM $Na_2HPO_4$ and 150 mM NaCl at pH 6.5) with 10% $D_2O$ (99.9 % isotopic purity, from Sigma Aldrich). For experiments on the CypA-CsA complex, CsA was added in excess so that the [CypA]:[CsA] concentration ratio was between 1:1 and 1:1.5. The pH of the solution was verified with a pH meter and then 180 $\mu$L of the solution was added to a 3 mm Bruker Match NMR tube. For the HDX experiments, purified $^{15}$N-labelled CypA in gel filtration buffer was dialysed overnight into 15 mM ammonium acetate at pH 6.5. Ammonium acetate is a volatile buffer and so after the freeze-drying process only protein sample remains as lyophilised powder. The dialysed sample was concentrated to below 2 mL using a 5000 molecular weight cut-off vivaspin column (from Sarstedt), centrifuged at 4500 $\times g$ and 4 °C. The sample was then frozen using liquid nitrogen and freeze-dried overnight. The buffer for the HDX experiments was prepared as above, using $D_2O$ instead of $H_2O$. This buffer was added to the lyophilised powder immediately before recording the first spectrum.

### 6.2.2 Data acquisition, processing and analysis.

The initial 3D $^{15}$N-$^{1}$H TOCSY-HSQC NMR experiments were conducted on unbound CypA. The experiment was performed on a Bruker AV II 700 spectrometer with a 16.45 Tesla ultra-shielded superconducting magnet. A Bruker triple resonance cryoprobe was used. The $^{1}$H resonance frequency was 700.24 MHz. All other 3D NMR data were collected using a Bruker AV III 600 spectrometer, with a 14.10

Tesla ulta shielded superconducting magnet with $^{1}$H resonance frequency of 600.13 MHz. The operating temperature for each experiment was 298 K. The 3D $^{15}$N-$^{1}$H TOCSY-HSQC experiments on unbound CypA were repeated on this spectrometer for use during the assignment process. HDX experiments were also carried out on this spectrometer. Results were acquired using TopSpin version 2.1 (from Bruker). The sample was locked against the deuterium signal. Shimming, matching and tuning were performed automatically; longer experiments used auto-shimming to maintain an homogenous magnetic field throughout. Mixing times for 3D TOCSY-HSQC and NOESY-HSQC experiments were 60 ms. The $^{1}J_{^{1}H,^{15}N}$ coupling constant for all datasets was 90 Hz. The raw NMR data were processed using TopSpin. The FID signal was truncated using the qsin apodisation function. This is a squared shifted sine-bell function with a sine-bell shift of 2.1. We did this to optimise both the signal-to-noise ratio and the shape of the peaks. The digital resolution of the spectrum was enhanced by zero filling of the truncated signal. A Fourier transform was subsequently applied to the FID to change the signal from the time domain to the frequency domain. Following the Fourier transform the data were phased and baseline-corrected.

For the 3D data, the number of data points in the direct dimension F3 ($^{1}$H) was 2048. There were 40 increments in dimension F2 ($^{15}$N) and 160 in F1 ($^{1}$H), zero filled to 512.

The 2D HSQC spectrum for the CypA-CsA complex (Figure 6.1) was recorded on the 600 spectrometer. There were 2048 data points in the direct dimension F2 ($^{1}$H). The number of scans was 16, each with 128 data points in F1 ($^{15}$N), zero-filled to 512 datapoints. The data for the HSQC spectrum of unbound CypA (Figure 6.2) was collected on the 700 spectrometer. The same number of data points in F2 were collected; 256 data points in F1 ($^{15}$N) were zero-filled to 512, and the number of scans was 8. The spectral width in the $^{1}$H dimension was 16 ppm. The $^{15}$N dimension was centred at 118 ppm, with a spectral width of 38 for the 3D experiments and 42 ppm for the HSQC experiments. The bulk water signal had a chemical shift of 4.692. This was suppressed in the 1D spectrum using an excitation sculpting pulse sequence. The centre of each spectrum was set to the frequency corresponding to this peak. Chemical shifts were referenced to a standard to allow comparison with previously published datasets [Wishart et al., 1995].

All NMR data were analysed using the Sparky program, version 3.1. [Goddard and Kneller].

### 6.2.3 HDX experiments

We used a dummy sample of $^{15}$N-labelled protein for locking, shimming, matching, tuning and optimisation of parameters. This was done to minimise the amount of time between solubilising the protein in $D_2O$ and recording the first HSQC spectrum. After optimisation we resuspended the lyophilised protein in the $D_2O$ phosphate buffer and a HSQC spectrum was recorded. There were 4096 points in the $^1$H time domain, 2048 in the frequency domain and 64 increments in the $^{15}$N dimension. Zero-filling was carried out prior to the Fourier transform to increase the number of points in the $^{15}$N dimension to 256 in the frequency domain. For CypA-CsA, each HSQC spectrum was generated following four scans. We increased the number of scans to 16 for the unbound CypA to improve the signal-to-noise ratio. HSQC spectra were recorded at set time intervals, as detailed in Tables 6.2 and 6.3.

NMR assignment data for the CypA-CsA complex were retrieved from the Biological Magnetic Resonance Data Bank (BMRB) entry 2208 [Neri et al., 1991]. The PDB entry 3CYS, also used to help with assignments, is an NMR structure of the CypA-CsA complex which is not listed in the BMRB [Spitzfaden et al., 1994]. We will refer to these structures as 2208 and 3CYS respectively.

## 6.3 Results

### 6.3.1 The $^1$H-$^{15}$N HSQC spectra for unbound CypA and the CypA-CsA complex

Figures 6.1 and 6.2 show the HSQC spectra for unbound CypA and the CypA-CsA complex, respectively. Figure 6.3 is an overlay of these two spectra. Each cross peak in the HSQC spectra is shown as a series of contour lines, where the lowest contour is selected so as to eliminate noise. For both samples, we observe clearly defined, well-spread peaks, which indicate abundant and folded protein. The spectra look similar to previously published images [Neri et al., 1991; Shi et al., 2006]. Many of the peaks in Figure 6.3 overlap, indicating that some residues do not experience a change in chemical environment upon binding. Some of the peaks do not overlap, showing an effect of ligand binding on the chemical environment of the corresponding residues. This is in accordance with published NMR data on CypA [Neri et al., 1991].

### 6.3.2 Assignment of the HSQC spectra

For the CypA-CsA complex, some of the assignment could be done tentatively by comparison with the previously published NMR data, 2208 and 3CYS [Neri et al.,

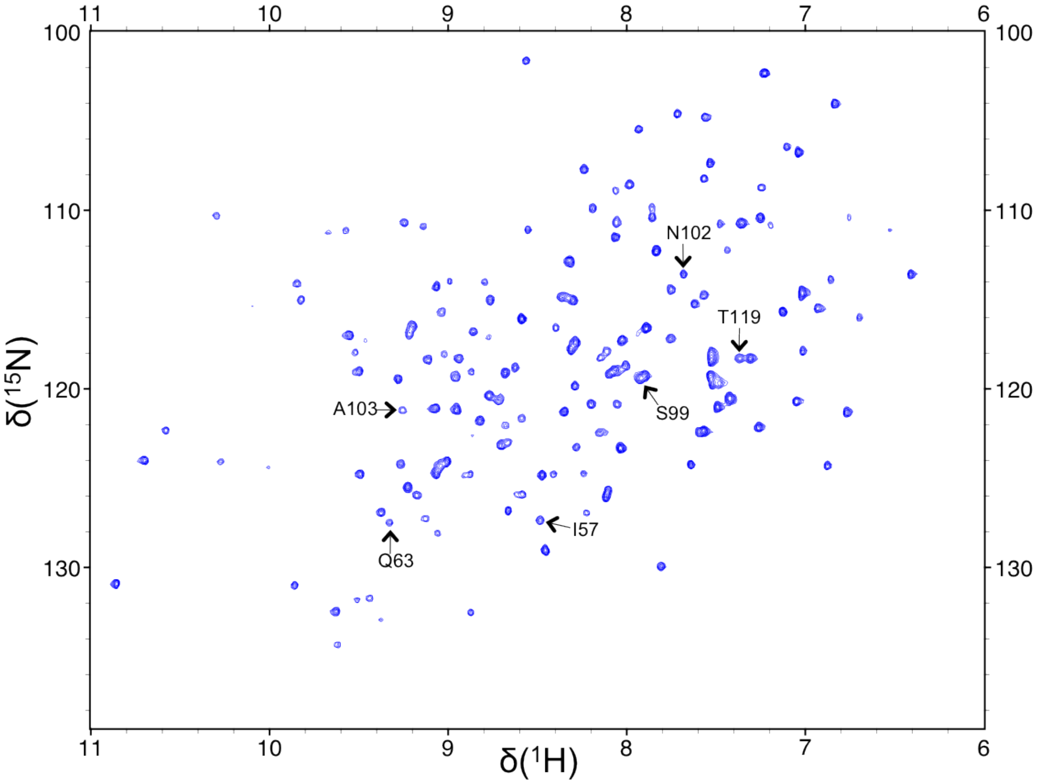


Figure 6.1: The HSQC spectrum the CypA-CsA complex, recorded on the 600 MHz spectrometer. Chemical shifts $\delta(^{15}\mathrm{N})$ and $\delta(^{1}\mathrm{H})$ are given in ppm. Six of the cross-peaks are labelled — those corresponding to I57, Q63, S99, N102, A103 and T119.

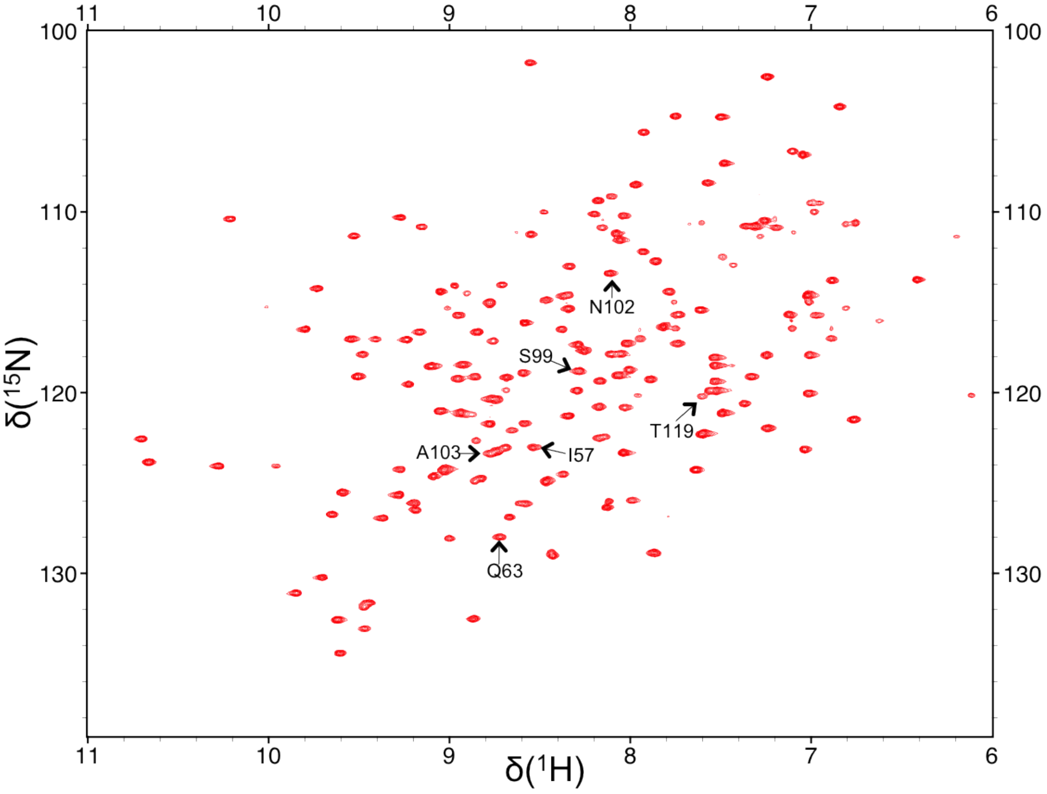


Figure 6.2: The HSQC spectrum of unbound CypA, recorded on the 700 MHz spectrometer. Chemical shifts $\delta(^{15}\text{N})$ and $\delta(^{1}\text{H})$ are given in ppm. The six cross-peaks corresponding to I57, Q63, S99, N102, A103 and T119 are labelled.

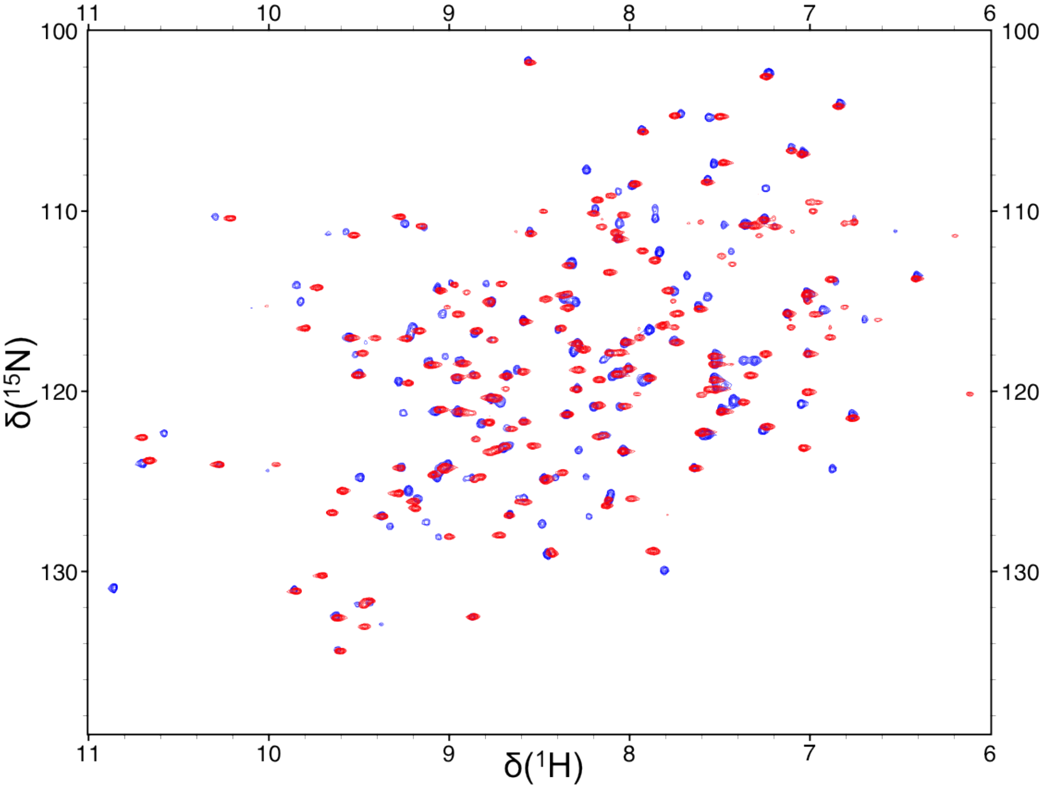


Figure 6.3: HSQC spectra of unbound CypA (red) and the CypA-CsA complex (blue). Chemical shifts $\delta(^{15}\mathrm{N})$ and $\delta(^{1}\mathrm{H})$ are given in ppm.

1991; Spitzfaden et al., 1994] and confirmed using the 3D NMR data. We used the 3D TOCSY-HSQC data to identify the type of amino acid with a particular spin system, and the 3D NOESY-HSQC data to connect the residues together through sequential assignment and clear up any remaining ambiguities. An example of the sequential assignment procedure is shown in Figure 6.4, with three NOESY cross-peaks identified between adjacent residues from Lys31 to Glu34. NOESY strips (blue) are given to the left of the TOCSY strips (red). For each residue $i$, there is a NOESY peak corresponding to the NH peak visible in both TOCSY and NOESY strips for residue $i-1$. We give a list of assignments for the $^{15}$N-$^{1}$H backbone pairs Appendix C, Table C.1. We have assigned 148 of the 159 non-proline residues for both the liganded and the unbound protein. Three of the residues which remain unassigned in both cases are part of the flexible N-terminus of the protein, of which to our knowledge only Asn3 has been previously identified [Spitzfaden et al., 1994]. There are four other residues — Glu81, Gly135, Arg148 and Asn149 — which we have been unable to assign in both HSQC spectra. Glu81 and Arg148 were also not assigned in the previous HDX study of CypA. In addition, His70, Met100, Ile114 and Val127 were not assigned for the unbound CypA. For the CypA-CsA complex, Cys62, Lys118, His126 and Val132 were not assigned. We have assigned 95.6 % of the non-proline residues in at least one of the spectra and 90.57% in both.

### 6.3.3 Chemical shift perturbation

The differences between the HSQC spectra can be seen in Figure 6.3 as well as in Table C.1. In Figure 6.5, we show overlaid sections of the HSQC spectra for the unbound CypA and the CypA-CsA complex, with assigned residues labeled. The $^{15}$N chemical shift of Ile57, and therefore its chemical environment, changes upon ligand binding. This effect is chemical shift perturbation, $\Delta_{\mathrm{S}}$, measured in ppm. Equation (6.1) is used to calculate $\Delta_{\mathrm{S}}$ for each peak as a function of $\Delta^{1}$H and $\Delta^{15}$N, the chemical shift differences in the $^{1}$H and the $^{15}$N dimension, respectively,

$$\Delta_{\mathrm{S}} = \sqrt{(\Delta^{1}\mathrm{H})^2 + (\frac{1}{6}\Delta^{15}\mathrm{N})^2}. \tag{6.1}$$

The $^{15}$N shift difference is divided by six to account for the chemical shift range of $^{15}$N being approximately six times larger than for $^{1}$H [Amin et al., 2013]. Ile57 has ($\delta^{1}$H, $\delta^{15}$N) = (8.54,123.02) and (8.49,127.36) in the unbound CypA and the CypA-CsA complex, respectively. Therefore $\Delta_{\mathrm{S}}$(Ile57) = 0.73 ppm. Figure 6.6 shows in black the $\Delta_{\mathrm{S}}$ values for all of the residues which were assigned in both spectra.

We also make a quantitative comparison between our dataset for the CypA-

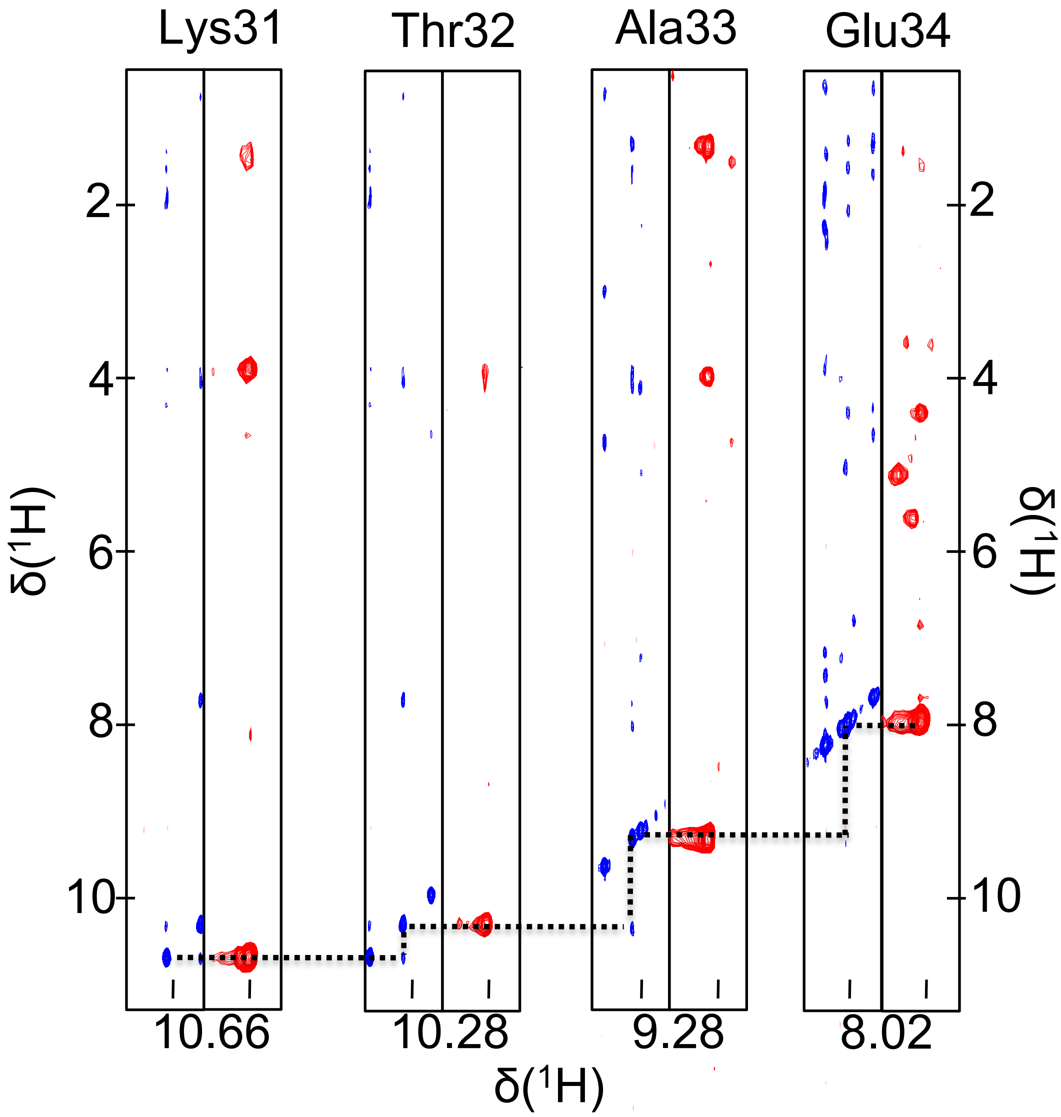


Figure 6.4: Strip plots the 3D NOESY-HSQC and TOCSY-HSQC NMR data for the unbound CypA, demonstrating sequential assignment. Residues Lys31, Thr32, Ala33 and Glu34 are shown along with three NOESY cross-peaks between them. For each residue, the left-hand strip shows the NOESY peaks in blue and the right-hand strip shows the TOCSY peaks in red. The dotted line connects the NH crosspeak of residue $i$, visible in both TOCSY and NOESY strips, with the corresponding NOE of residue $i+1$.

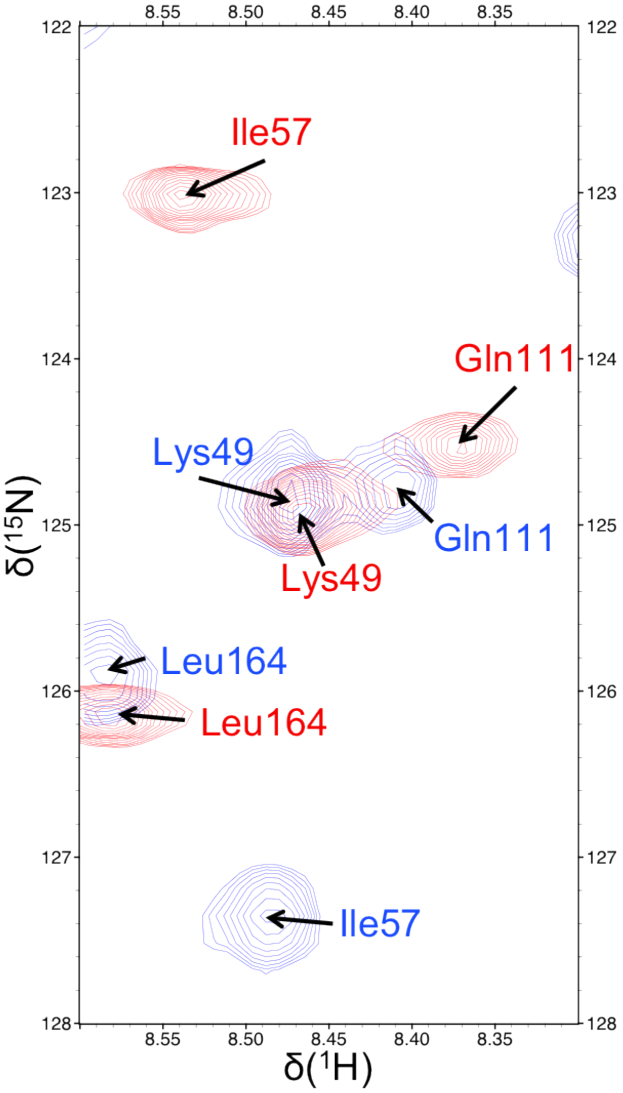


Figure 6.5: A section of the overlaid HSQC spectra showing chemical shift perturbation. The cross-peaks and corresponding assignments for the unbound CypA are shown in red and for the CypA-CsA complex in blue. Chemical shifts $\delta(^{15}\mathrm{N})$ and $\delta(^{1}\mathrm{H})$ are given in ppm.

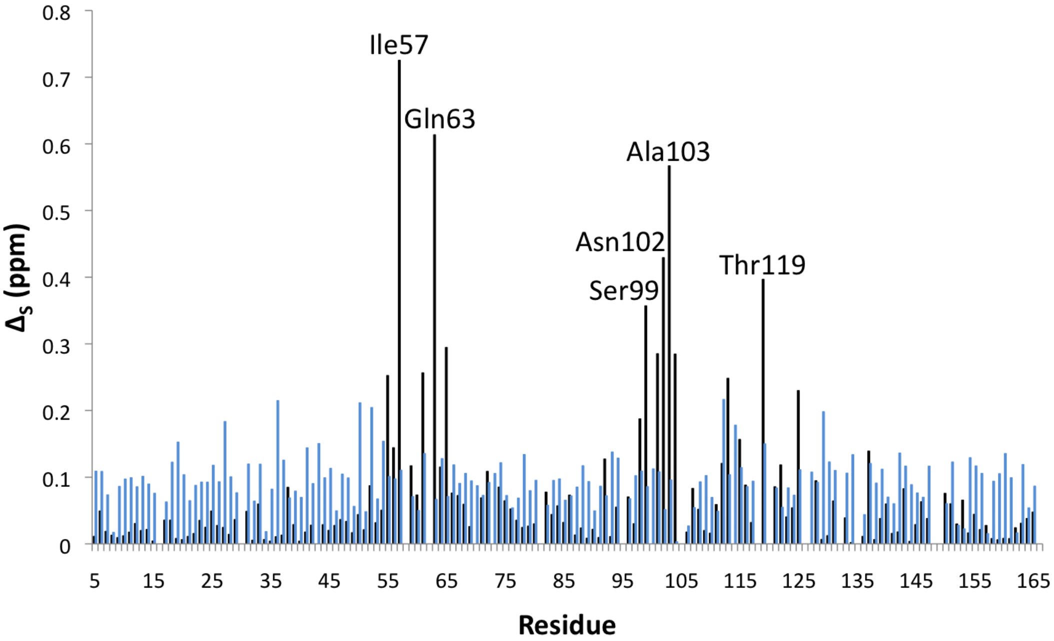


Figure 6.6: Black columns show chemical shift perturbations as a result of CypA-CsA complex formation. Blue columns show differences between the CypA-CsA complex chemical shifts and previously published data [Neri et al., 1991].

CsA complex and the previously published dataset 2208 [Neri et al., 1991]. We use Equation (6.1) to make this comparison. It must be noted, however, that in this case we are not technically calculating the chemical shift perturbation. Instead, we use $\Delta_S$ as a way of comparing assignments. These values are plotted in blue in Figure 6.6, and form a baseline variation for the comparison between the unbound CypA and the CypA-CsA complex. The high values of $\Delta_S$ seen for some residues in black in Figure 6.6 stand out against the baseline and emphasise the large impact of ligand binding on the chemical environment of these residues. For example, Ile57 has a large value of $\Delta_S$ relative to the rest of the protein. As this residue forms part of the binding site of CypA, a dramatic change in chemical environment and therefore a large $\Delta_S$ is not unexpected.

In Table 6.1, we summarise the data in Figure 6.6, along with a similar comparison between the two previously published CypA-CsA datasets, 2208 and 3CYS. The unbound CypA and the CypA-CsA complex have similar HSQC spectra, as seen in Figure 6.3. The majority of peaks overlap or are close to doing so, and so have small values of $\Delta_S$. There are 19 residues with $\Delta_S > 0.2$, six of which have $\Delta_S > 0.3$. These six residues are labelled in Figure 6.6, as are their respective

Table 6.1: The data for the CypA-CsA complex is compared with the data for the unbound CypA as well as the published dataset 2208. The published datasets 2208 and 3CYS are also compared. The first comparison shows the impact of ligand binding, the other two compare assignments. Each time the average, standard deviation and maximum values of $\Delta_S$ are given, along with the number of peaks for which $\Delta_S$ is greater than 0.1, 0.2 and 0.3. All values of $\Delta_S$ are in ppm, calculated using Equation (6.1).

| Spectrum 1 | CypA-CsA | CypA-CsA | 2208 |
|---|---|---|---|
| Spectrum 2 | CypA only | 2208 | 3CYS |
| Average $\Delta_S$ | 0.06 | 0.09 | 0.09 |
| Standard deviation | 0.11 | 0.04 | 0.04 |
| Maximum $\Delta_S$ | 0.73 | 0.22 | 0.24 |
| Peaks with $\Delta_S > 0.1$ | 23 | 61 | 68 |
| Peaks with $\Delta_S > 0.2$ | 13 | 4 | 4 |
| Peaks with $\Delta_S > 0.3$ | 6 | 0 | 0 |

cross-peaks in Figures 6.1 and 6.2. The effect of ligand binding on the chemical environment appears significant for a small number of residues and negligible for the majority. The other two comparisons concern assignments. The average $\Delta_S$ values are larger but in both cases the standard deviation is smaller and there are fewer residues with unusually large $\Delta_S$ and none with $\Delta_S > 0.3$ ppm. The comparison, for each residue, between our CypA-CsA data and 2208 is plotted in blue in Figure 6.6. These results show that the three datasets for the CypA-CsA complex are largely consistent with each other. The variation which does exist may be due to differences in experimental conditions such as temperature and buffer composition. Dataset 2208 was collected at 20°C in 50 mM phosphate buffer containing 100 mM NaCl and 5 mM DTT. Dataset 3CYS was collected at 26°C in 10 mM phosphate buffer containing 100 mM $NaNO_3$, 5 mM DTT and 50 $\mu$M $NaN_3$. In Figure 6.7, the residues with large chemical shift perturbations are shown on the 3D structure of CypA, represented using the crystal structure 1CWA. The majority of these residues are clustered around the ligand binding site.

### 6.3.4 HDX experiments on the CypA-CsA complex

During the HDX experiments, HSQC spectra were completed according to the time-frame given in Table 6.2. Figure 6.8 shows the HSQC spectrum of the CypA-CsA complex from Figure 6.1 alongside spectra recorded 10, 110 and 4270 minutes (71 hours and 10 minutes) after initiating HDX. Figure 6.9 shows close-ups of the same spectra, focussing on residues 55, 97, 122 and 139. In these two figures, we see

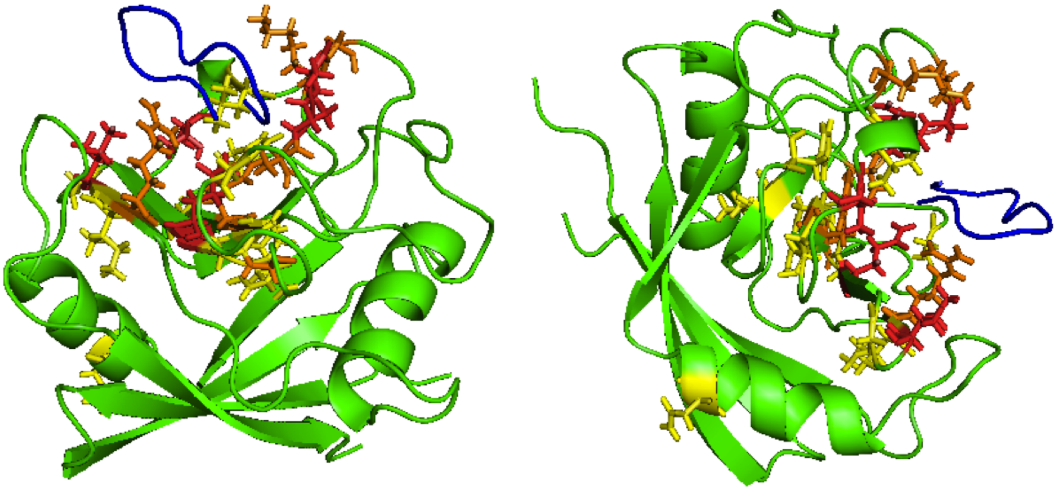


Figure 6.7: Two images the CypA-CsA complex are shown. The backbone of the crystal structure 1CWA is depicted using PYMOL. The image on the right is a rotation of the image on the left. CsA is coloured blue. From the comparison between the complex and unbound CypA, residues with $\Delta_S > 0.1$ are shown as sticks. Residues with $0.1 < \Delta_S \leq 0.2$ are coloured in yellow, those with $0.2 < \Delta_S \leq 0.3$ in orange and those with $\Delta_S > 0.3$ in red.

Table 6.2: Sequence of time intervals for recording HSQC spectra during HDX experiments on CypA-CsA. The spectrum number is given along with the elapsed time in minutes.

| Spectrum | Time (min) | Spectrum | Time (min) |
|---|---|---|---|
| 1 | 10 | 17 | 160 |
| 2 | 15 | 18 | 190 |
| 3 | 20 | 19 | 220 |
| 4 | 25 | 20 | 250 |
| 5 | 30 | 21 | 310 |
| 6 | 35 | 22 | 370 |
| 7 | 40 | 23 | 430 |
| 8 | 45 | 24 | 490 |
| 9 | 50 | 25 | 610 |
| 10 | 55 | 26 | 730 |
| 11 | 60 | 27 | 850 |
| 12 | 65 | 28 | 1030 |
| 13 | 70 | 29 | 1210 |
| 14 | 90 | 30 | 1390 |
| 15 | 110 | 31 | 2830 |
| 16 | 130 | 32 | 4270 |

Table 6.3: Sequence of time intervals for recording HSQC spectra during HDX experiments on unbound CypA. The spectrum number is given along with the elapsed time in minutes.

| Spectrum | Time (min) | Spectrum | Time (min) |
|---|---|---|---|
| 1 | 53 | 13 | 449 |
| 2 | 73 | 14 | 509 |
| 3 | 93 | 15 | 629 |
| 4 | 113 | 16 | 749 |
| 5 | 133 | 17 | 929 |
| 6 | 153 | 18 | 1109 |
| 7 | 179 | 19 | 1289 |
| 8 | 199 | 20 | 1469 |
| 9 | 219 | 21 | 2909 |
| 10 | 269 | 22 | 4349 |
| 11 | 329 | 23 | 5789 |
| 12 | 389 | | |

clearly the gradual loss of signal over time, demonstrating the process of HDX. The residues of CypA can be categorised according to the disappearance of their corresponding peaks in the HSQC spectra as time elapses. A large number of peaks have disappeared before the first spectrum was recorded. Subsequently there was a steady decline in the number and intensity of remaining peaks. Figures 6.8 and 6.9 illustrate this. The intensity of the peak corresponding to Arg55 has reduced to the extent that the peak has almost disappeared after ten minutes. After 4270 minutes, the peak corresponding to Leu122 has disappeared; those corresponding to Val139 and Ile97 are still present, but with reduced intensity.

### 6.3.5 HDX experiments on unbound CypA

An unexpectedly low yield meant that the sample of unbound CypA we used for the HDX experiments was less concentrated than the sample of the CypA-CsA complex. As a result the NMR signals for these experiments were weaker. This created difficulty with data collection and processing. We increased the number of scans recorded for each spectrum from 4 to 16 in response in order to maximise the signal to noise ratio. The earliest usable HSQC spectrum was completed 53 minutes after adding $D_2O$ to the protein sample, and the subsequent spectra were therefore recorded less frequently than with the CypA-CsA experiments. Nevertheless, the results that were obtained are sufficient to define the HDX folding core (see Section 6.3.6, below). Table 6.3 lists the time after the addition of $D_2O$ at which each spectrum was completed. Figure 6.10 shows the full HSQC spectrum, along with

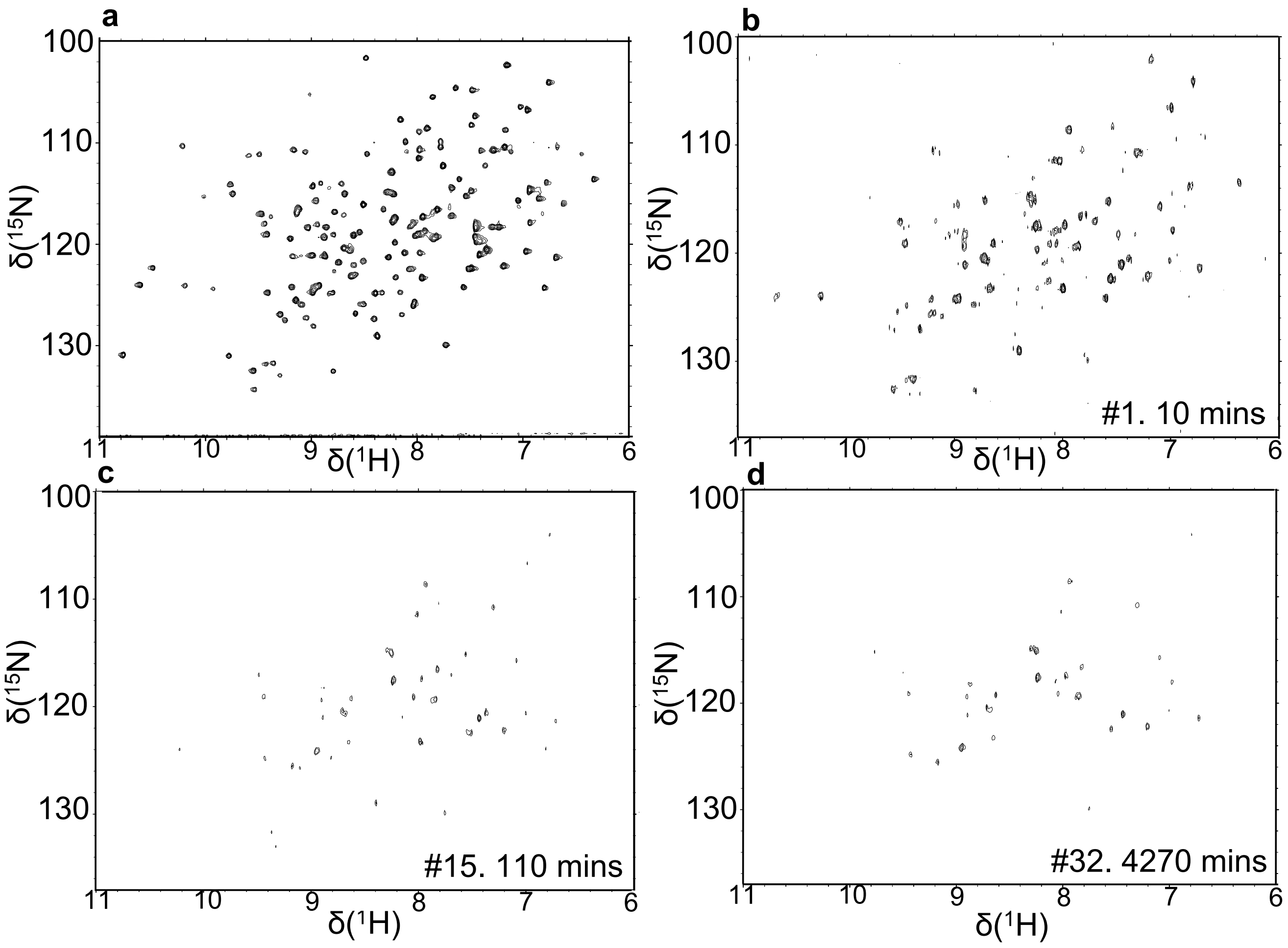


Figure 6.8: (a) HSQC spectrum of the CypA-CsA complex. Following initiation of HDX, spectra were recorded after an elapsed time of (b) 10 minutes, (c) 110 minutes and (d) 4270 minutes. With reference to Table 6.2, spectra (b), (c) and (d) are spectra numbers 1, 15 and 32 respectively.

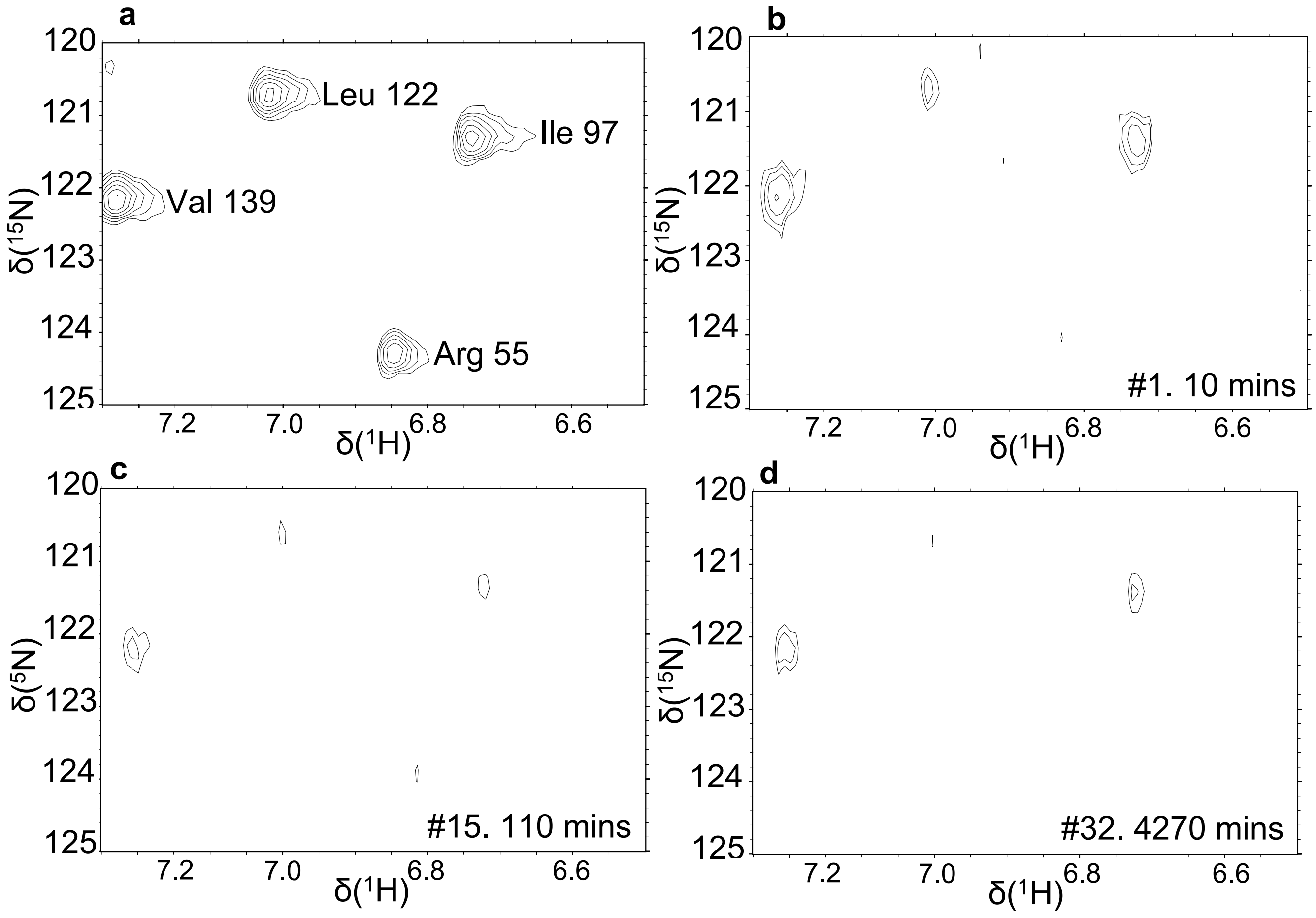


Figure 6.9: (a) Peaks for residues 55, 97, 122 and 139 in the HSQC spectra of the CypA-CsA complex. The same peaks are shown following initiation of HDX, after elapsed times of (b) 10 minutes, (c) 110 minutes and (d) 4270 minutes. With reference to Table 6.2, spectra (b), (c) and (d) are spectra numbers 1, 15 and 32 respectively. These are close-ups of the spectra in Figure 6.8.

spectra taken at different time intervals following the initiation of HDX. Spectrum numbers 1, 4 and 22 are shown, recorded after 53, 113 and 4349 minutes respectively. In Figure 6.11 a close up of each of the spectra is shown. The section of the HQSC plot shown is as in Figure 6.9. The same four peaks are visible initially, although they have shifted slightly from their positions: Arg55, Ile97 and Val139 all have $\Delta_S > 0.1$ ppm upon ligand binding and Leu122 has $\Delta_S = 0.08$. After 53 minutes the peaks corresponding to Leu122 and Arg55 have disappeared. Those corresponding to Val139 and Ile97 remain after 4349 (72 hours and 29 minutes) minutes, although the signal for the Ile97 peak is very weak.

### 6.3.6 Comparison of HDX folding cores

From the HDX data we can define the HDX folding core for both the CypA-CsA complex and the unbound CypA. We select the residues for which peaks are present 100 minutes after adding $D_2O$ to form the respective folding cores. This was chosen to facilitate comparisons between our data and the previously published study, where we took the folding core to be the slow and medium exchanging residues, for which $k_{\text{ex}} > 10^{-2}$ min$^{-1}$. For CypA-CsA the residues whose peaks are present in spectrum 15 form the folding core. Similarly, the folding core for the unbound CypA corresponds to the peaks present in spectrum 4. These spectra were collected respectively at 110 minutes and 113 minutes following the initiation of HDX. In Figure 6.12 both of these spectra are shown, with the unbound CypA in red and the CypA-CsA complex in blue. Assignments for the residues in the CypA-CsA folding core are labelled. The lower signal-to-noise ratio in the unbound CypA experiment makes determining the folding core more difficult. The peaks for this spectrum are shown from a lower contour than for the CypA-CsA spectrum, thus inviting more noise into the image.

Table 6.4 lists the folding core residues for both experiments. Highlighted in blue are the residues only present in the folding core of the CypA-CsA complex. The residues which are only present in the unbound CypA are listed in red. All of the residues listed in black are part of both of the folding cores. The residues in Table 6.4 are illustrated on the 3D structure of the protein in Figure 6.13. The size of the two folding cores is strikingly different, with 80 residues in the CypA-CsA folding core and 48 in the CypA folding core. The presence of the ligand shields some residues from the surrounding solvent and therefore may prevent HDX, resulting in a larger folding core. A number of the residues in the CypA-CsA folding core only, coloured blue in Table 6.4 and Figure 6.13, are proximal to the binding site if not directly part of it. In addition to these residues, however, there are others such as Ala26 and

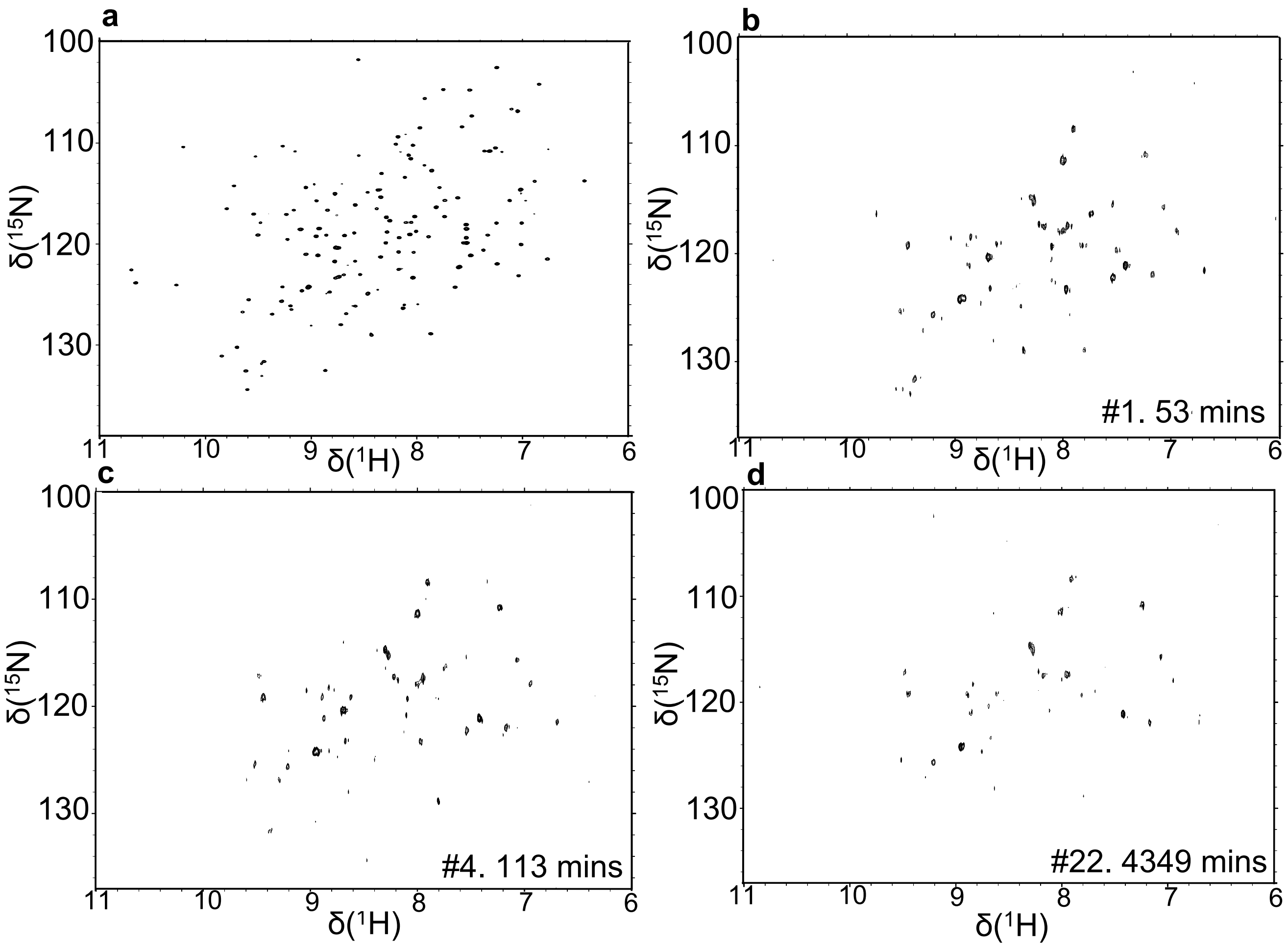


Figure 6.10: (a) HSQC spectrum of unbound CypA. HSQC spectra are also shown at different time intervals following initiation of HDX. These are (b) 53 minutes, (c) 113 minutes and (d) 4349 minutes. With reference to Table 6.3, spectra (b), (c) and (d) are spectra numbers 1, 4 and 22 respectively.

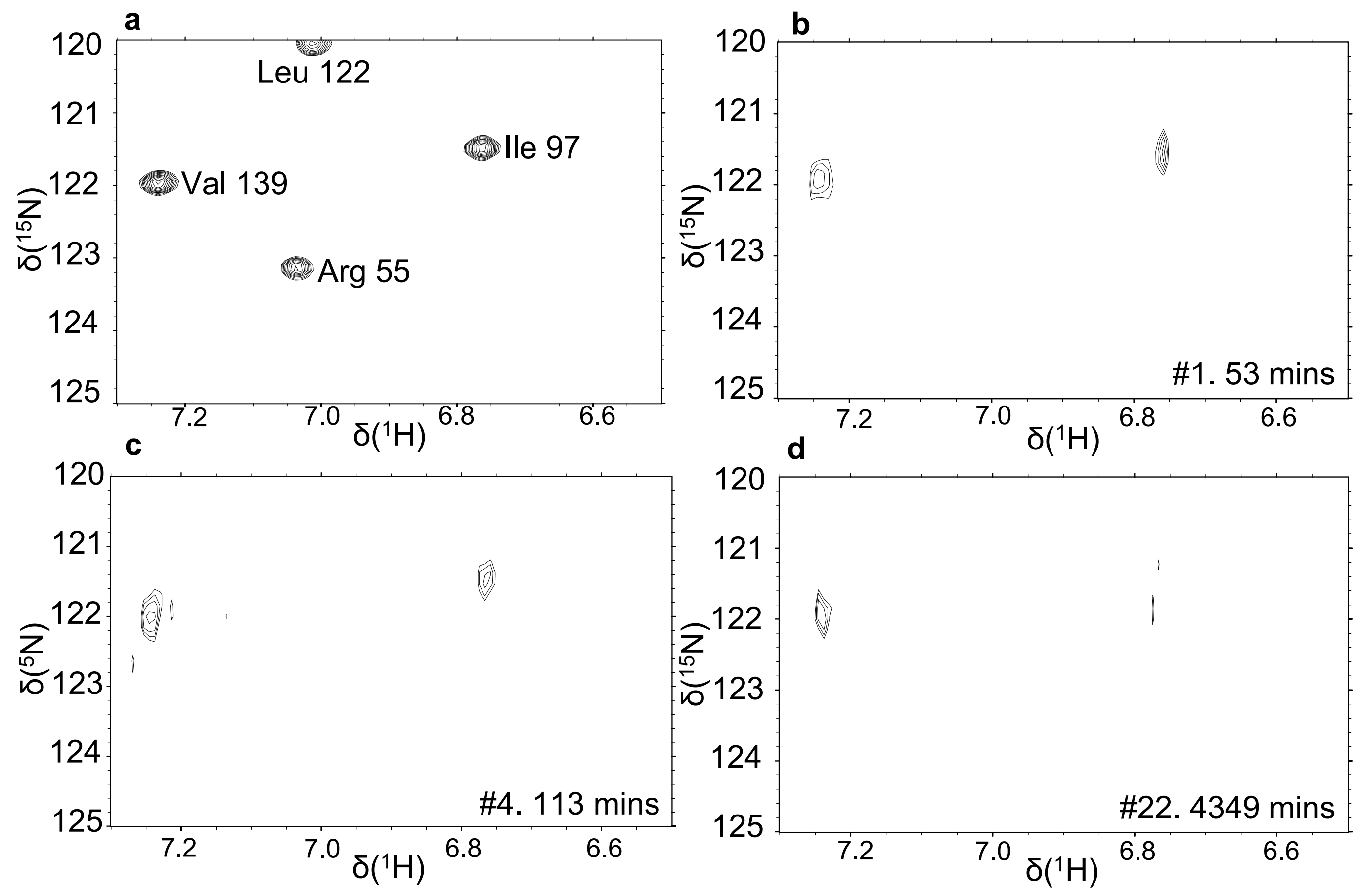


Figure 6.11: (a) The section of the HSQC spectrum of unbound CypA containing the peaks for residues 55, 97, 122 and 139. The same sections are also shown at different time intervals following initiation of HDX. These are (b) 53 minutes, (c) 113 minutes and (d) 4349 minutes. With reference to Table 6.3, spectra (b), (c) and (d) are spectra numbers 1, 4 and 22 respectively

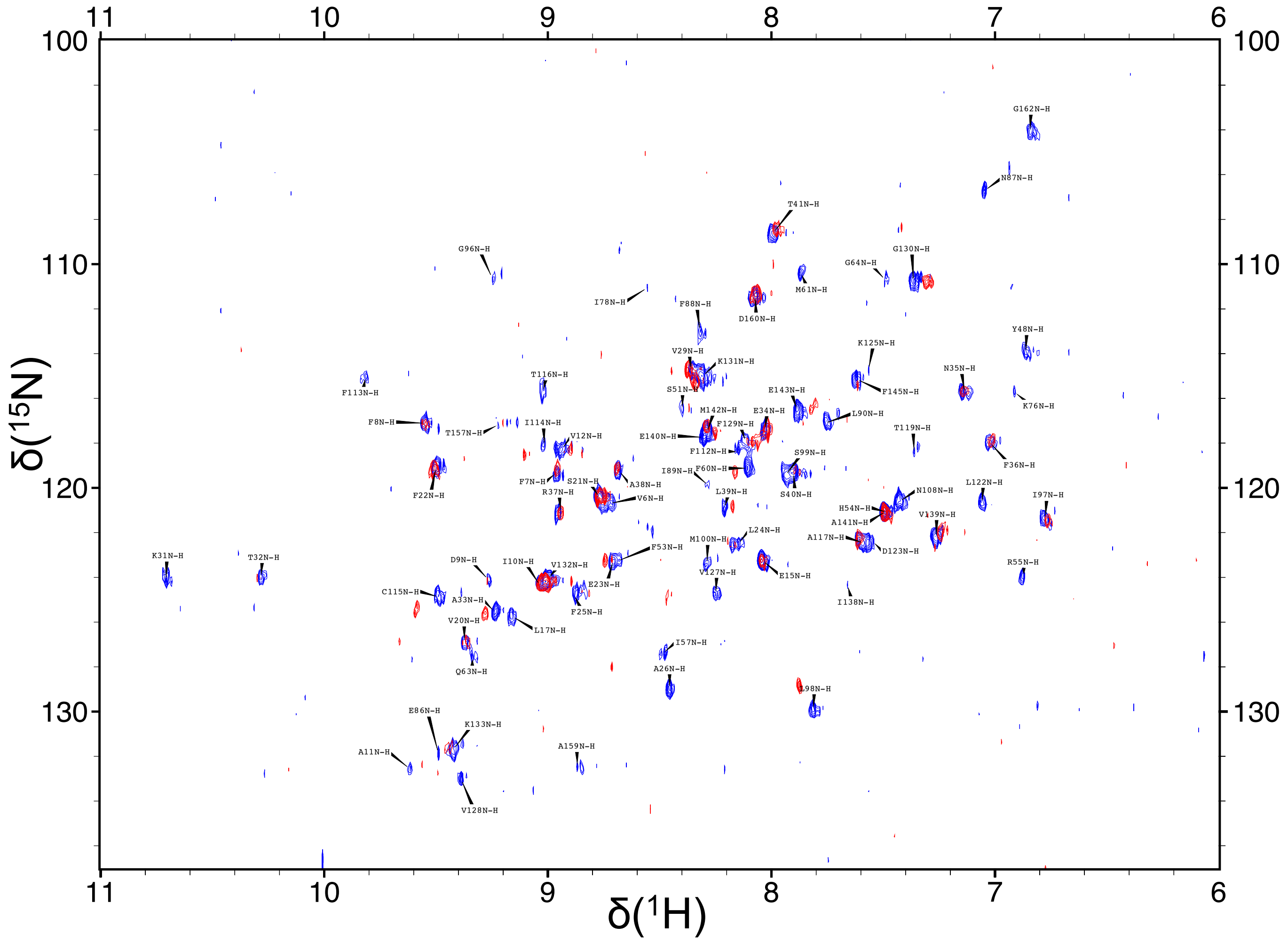


Figure 6.12: HSQC plot showing the HDX folding cores of the CypA-CsA complex (blue) and the unbound CypA (red). Assignments for the remaining residues in the CypA-CsA complex are labelled. Chemical shifts are in ppm.

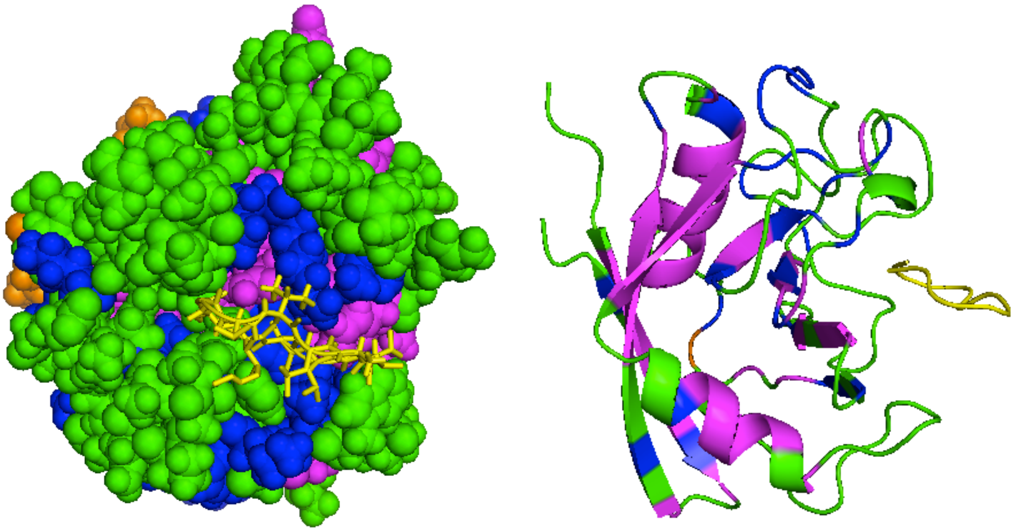


Figure 6.13: HDX folding cores are shown on the crystal structure 1CWA using PYMOL. The left-hand image is formed looking down on to the binding site, with protein atoms shown as spheres and the ligand as yellow sticks. Residues which form the folding core of the CypA-CsA complex but not the unbound protein are shown in blue. Those which are in the folding core of the unbound protein but not the CypA-CsA complex are shown in orange. The intersection of the two folding cores is coloured magenta. The right-hand image shows the secondary structural motifs using the same colour scheme.

Table 6.4: Folding core residues for the CypA-CsA complex and for unbound CypA. The residues highlighted in blue are present in the CypA-CsA folding core but not for the unbound CypA. Those in red are present in the folding core of unbound CypA but not the CypA-CsA complex. Underlined residues are not part of the previously determined HDX folding core for CypA [Shi et al., 2006]. All residues in black are present in the folding cores for both of our HDX experiments.

| Experiment | Folding Core Residues |
|---|---|
| CypA-CsA complex | V6, F7, F8, D9, I10, A11, V12, E15, L17, V20, S21, F22, E23, L24, F25, A26, V29, K31, T32, A33, E34, N35, F36, R37, A38, L39, S40, T41, Y48, S51, F53, H54, R55, I57, F60, M61, Q63, G64, K76, I78, E86, N87, F88, I89, L90, G96, I97, L98, S99, M100, N108, F112, F113, I114, C115, T116, A117, T119, L122, D123, K125, V127, V128, F129, G130, K131, V132, K133, I138, V139, E140, A141, M142, E143, F145, T157, A159, D160, G162 |
| Unbound CypA | V6, F7, F8, D9, I10, V12, E15, V20, S21, F22, E23, L24, F25, V29, T32, A33, E34, N35, F36, R37, A38, L39, S40, T41, K44, K49, S51, F53, F60, M61, Q63, I97, L98, F112, C115, A117, D123, F129, G130, K131, V132, K133, V139, E140, A141, M142, E143, F145, T157, D160 |

Gly162 which are part of this set but distant from the ligand. The smaller folding core for the unbound protein may be due to the lower signal-to-noise ratio. A small loss in intensity may have caused a peak to disappear amongst the noise, whereas the same relative loss in intensity may not have had this effect in the experiment on the complex. It is therefore important to treat the CypA dataset with caution. If the folding core were indeed larger than detected in this instance, the set of residues unique to the folding core of the complex (coloured blue in Figure 6.13) would be an overestimate as some of them would in fact be part of both folding cores (coloured magenta in Figure 6.13).

The previously published HDX data for unbound CypA [Shi et al., 2006] yielded a folding core of 73 residues, as discussed in Chapter 4. The overlap between these 73 residues and the 48 which form the folding core for CypA in Table 6.4 is considerable. In fact, only S21 and the residues highlighted in red in Table 6.4 are not part of the Shi folding core. The remaining 45 residues are therefore part of all three folding cores. The CypA-CsA folding core also largely overlaps with the Shi folding core. In Section 6.3.3, we noted that the binding of CsA has little effect on the chemical environment of the majority of the residues. We know also that there is not a large conformational change upon ligand binding [Ottiger et al., 1997]. It is therefore unsurprising that there is a large overlap between the Shi folding core

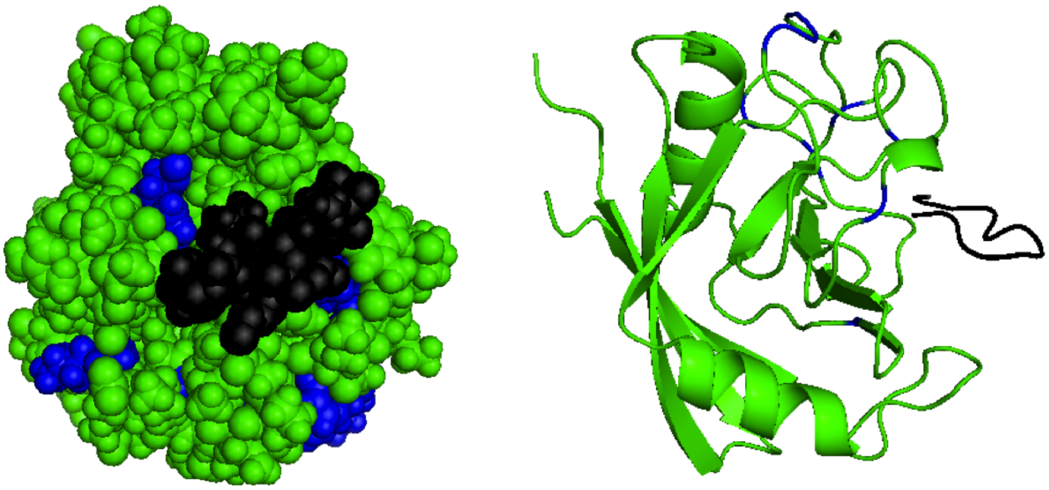


Figure 6.14: The effect of CsA binding on the HDX folding core is shown on the crystal structure 1CWA using PYMOL. The nine residues which are in the CypA-CsA folding core but not in either of the unbound CypA folding cores are coloured blue, with the remaining residues in green. CsA is coloured black. In the left-hand image all atoms are represented as spheres, with the view down onto the binding site. The secondary structure is shown in the right-hand image, with the binding site on the right-hand side.

and that of CypA-CsA. There are ten residues which are in the CypA-CsA folding core but not the Shi folding core. These are underlined in Table 6.4. S21 is also present in the folding core for of the unbound CypA; the other nine residues are all of interest since of the three datasets under comparison, these residues only appear in the CypA-CsA folding core. Figure 6.14 shows the location of these nine residues on the 3D protein structure. Residues H54, T119 and L122 are close to the ligand. None of these residues form part of the secondary structure motifs. Residues E86 – I89 are further from the ligand, and so it is surprising that these should be affected by ligand binding in the HDX experiments. There are also a small number of residues which are part of the Shi folding core but not the CypA-CsA folding core. Two of these, C62 and H126, have not been assigned in the CypA-CsA HSQC spectrum and may be part of the CypA-CsA folding core but remain undetected because we do not know the location of the corresponding peaks. The other such residues are G18, A101 and I158, none of which appear in the folding core for CypA.

Instead of taking the peaks which remain in the HSQC spectra 100 minutes after the addition of $D_2O$, one could examine later spectra to determine the folding core. Spectra 32 and 22 for the CypA-CsA and CypA experiments respectively were taken after 4270 and 4349 minutes. The remaining peaks at these time points form

Table 6.5: Agreement between the theoretical methods introduced in Chapter 4 and the HDX folding core of CypA-CsA determined experimentally in this chapter. For each theoretical method, the number of residues in the folding core, $N_B$ is given along with $c(A,B)$, the size of the overlap between the theoretical folding core and the HDX folding core. These numbers are used to determine $s(AB)$, $\%_C$ and $\%_A$ as defined in Section 4.2.2.

| Method | $N_B$ | $c(A,B)$ | $s(AB)$ | $\%_C$ | $\%_A$ |
|---|---|---|---|---|---|
| FIRST | 86 | 57 | 1.37 | 71.25 | 66.28 |
| Secondary structure | 75 | 59 | 1.62 | 73.75 | 78.67 |
| FRODA | 80 | 56 | 1.44 | 70.00 | 70.00 |
| FIRST + FRODA | 54 | 44 | 1.68 | 55.00 | 81.48 |
| Pseudodihedrals | 99 | 56 | 1.17 | 70.00 | 56.57 |
| FIRST with HB0HP0 | 72 | 53 | 1.52 | 66.25 | 73.61 |

subsets of the folding cores listed above. Comparing these subsets is similar to the above analysis. Instead of the nine residues determined above which are present in the CypA-CsA folding core but not in either of the CypA folding cores, we get a set of four residues: H54, F88, N108 and L122.

### 6.3.7 Theoretical folding cores for the CypA-CsA complex

We used the computational methods of Chapter 4 to generate folding cores for the CypA-CsA complex from the structure 1CWA. These were compared with the experimental HDX folding core for the complex. Table 6.5 shows $s(AB)$, $\%_C$ and $\%_A$ for each comparison. Notably, the FIRST folding core for for the CypA-CsA complex is smaller than for the unbound protein. As a result, there are fewer false negatives in the FIRST folding core, and $\%_A$ increases from 58.93% to 66.28%. The smaller FIRST folding core in turn shrinks the FIRST + FRODA folding core, which remains the most accurate method. For the complex, $E_{fc} = -1.452$ whereas $E_{fc} = -1.263$ for the unbound protein. The complex is rigidified by the ligand despite the smaller folding core. The smaller folding cores in the presence of the ligand goes against the experimental evidence which shows that the size of the folding core increases upon ligand binding. Clearly, the theoretical measures are not yet accurate or indeed sensitive enough to capture the impact of ligand binding for CypA.

## 6.4 Discussion

3D $^{15}$N-$^{1}$H TOCSY-HSQC and $^{15}$N-$^{1}$H NOESY-HSQC NMR experiments were performed on both unbound CypA and the CypA-CsA complex. The data from these experiments were used to assign the $^{15}$N-$^{1}$H HSQC plots, shown in Figures 6.1 – 6.3. An assignment table detailing the assignments for the backbone N-H pairs is given in Appendix C (Table C.1). The two HSQC plots were highly similar with subtle changes indicating the effects of ligand binding. This result is not unexpected, since CypA is known not to undergo a large conformational change upon ligand binding [Fanghänel and Fischer, 2003] but rather small variations around the stable scaffold of the native protein structure [Ottiger et al., 1997]. The CypA-CsA NMR data were compared to 2208 and 3CYS, the published datasets on the same complex. The similarities between the CypA-CsA data acquired here and these two datasets underlined the differences observed when comparing the CypA-CsA complex with the unbound CypA. Although the structure of CypA has been solved using solution NMR, the resulting NMR assignments were not available to us when compiling this study [Ottiger et al., 1997; Shi et al., 2006]. The HDX experiments were performed on both the CypA-CsA complex and the unbound CypA. Figures 6.8 – 6.11 demonstrate the process of signal loss over time. In these figures it is clear that although many peaks disappear before the first spectrum is recorded, others remain visible 72 hours after the initiation of HDX, demonstrating the structural stability of the protein. The HDX experiments for the unbound protein were more difficult to evaluate than for the CypA-CsA complex, as discussed in the text. Nevertheless, we have been able to use the published HDX data on the unbound CypA to inform our comparison of the HDX folding cores. We were able to isolate a small number of residues which are part of the folding core for CypA-CsA but not for CypA.

# Chapter 7

# Conclusions

## 7.1 Summary and outcomes

We have studied the effects of ligand binding on the two proteins, HIV-1 protease and CypA. Rigidity analysis in FIRST, mobility simulations in FRODA and biophysical techniques including HDX NMR experiments were used to assess these effects. In Chapter 2 we studied 212 structures of HIV-1 protease using FIRST. We defined the measure $\Delta\Phi$ to extract information about the impact of ligand binding on rigidity. For each structure, a single value was determined to quantify the impact of the inhibitor on the rigidity of the $\beta$-sheet flaps which shield the active site of the enzyme from the surrounding solvent. A full list of these structures and their $\Delta\Phi$ values is given in Appendix B.3, Table B.2. We categorised the inhibitors based on the $\Delta\Phi$ value, and broadly these inhibitors fell in to two different categories — those which rigidified the flaps ($\Delta\Phi \simeq 1$) and those which did not ($\Delta\Phi \simeq 0$). Of the FDA approved inhibitors, tipranavir consistently rigidified the flaps whereas darunavir did not. This may account for their particular efficacy in combination therapies [Hicks et al., 2006]. The study showed that rigidity analysis on as many as 212 structures is feasible and informative. Differences in the bonding networks inferred from such a large range of structures can be simply interpreted in terms of the effect on the rigidity of a specific region of the protein. These data have been used in a subsequent study on computational predictions of binding affinities [Greenidge et al., 2013].

We applied a similar approach to CypA, using rigidity analysis on many structures of the protein. Whereas RD plots of HIV-1 protease demonstrated a clear effect of some ligands on protein rigidity, namely enhanced flap rigidity, there was no such clear pattern in the RD plots of CypA. We examined established global and local rigidity parameters $X_{\text{LRC}}$ and $Z_{\text{LRC}}$ and observed their dependence on $E_{\text{cut}}$. We adapted $\Delta X_{\text{LRC}}$ and $\Delta Z_{\text{LRC}}$ so that the effect of ligand binding could be assessed. We showed that these difference measures were also dependent on $E_{\text{cut}}$, but in a non-monotonic fashion. This was observed for a number of crystal structures. Responding to this, we designed the RDM, which maintains residue-specific information and captures the impact of ligand binding on the rigidity of a protein at a range of different $E_{\text{cut}}$ values. The clear impact of ligand binding on the rigidity of certain HIV-1 protease structures led us to design a single parameter $\Delta\Phi$ to describe this effect. Ligand binding affects the rigidity of CypA in a more subtle manner and hence we designed the RDM to highlight such changes. Whilst $\Delta\Phi$ was tailored to HIV-1 protease and would need to be adapted for other proteins, the RDM could be applied to any protein and may prove to be a useful construct for determining appropriate $E_{\text{cut}}$ values at which to run mobility simulations. We used the RDM to compare structures before and after ligand removal. A similar approach could be used to compare the impact on rigidity of other changes to similar structures, for example point mutations.

In addition to rigidity analysis, we used the FRODA module in FIRST to simulate protein motion. We directed large-scale motion using normal modes for unbound CypA and for the CypA-CsA complex. The simulations were evaluated using different bond networks. Five different $E_{\text{cut}}$ values were used to determine the bond network. Our results show that altering the $E_{\text{cut}}$ value can change the results of mobility simulations. Subtle changes in structural geometry can change the calculation of hydrogen bond strengths dramatically, and therefore also the potential impact of $E_{\text{cut}}$ on rigidity [Wells et al., 2009]. The fact that these subtle changes may have a large impact on the results of protein mobility simulations in FRODA underlines the need for caution when selecting the $E_{\text{cut}}$ value. We also varied the hydrogen bond (HB) and hydrophobic tether (HP) settings. Whereas $E_{\text{cut}}$ dictates the minimum strength for a hydrogen bond to be included in the network, the HB and HP settings place restrictions on the location of these bonds based on the surface exposure of the bonding atoms. Changing the HB/HP settings alters the bond network, and therefore potentially the rigidity, of the protein surface. The results of HDX are dominated by the exposure of atoms in the protein to the surrounding solvent, and so changing the bond network around the surface of the protein may bring the results of rigidity analysis in line with experimental data.

The nine different combinations of HB and HP settings and five $E_{\text{cut}}$ values means that there are 45 different bond networks for which we simulated protein motion. We characterised protein mobility in two ways. Burial distance measurements were used to track the surface of the protein throughout the simulations and pseudodihedral angles to assess its backbone motion. The pseudodihedral measurements were affected by the choice of $E_{\text{cut}}$, with the rigidity of the protein feeding directly into its dynamics. The RDM for CypA demonstrated the $E_{\text{cut}}$-dependent effect of ligand binding on protein rigidity. The impact of ligand binding on the mobility of CypA is similarly dependent on $E_{\text{cut}}$. As the HB/HP setting affected rigidity, so it also affected mobility. The combination of bond setting and $E_{\text{cut}}$ determines the bond network which in turn determines the mobility of the protein via its rigidity. The same simulations characterised in terms of surface exposure rather than change in pseudodihedral angles did not reflect this dependence. Changes in the surface of CypA were subtle even for large amplitude protein motion. The initial protein surface remains the same after motion along the low-frequency normal modes. For this reason the impact of ligand binding on the mobility as characterised by burial distance is restricted to the binding site and its immediate surroundings.

In terms of relating mobility data to HDX experiments, the burial distance measure was useful. We predicted the HDX folding core using the established FIRST folding core [Hespenheide et al., 2002]. Since the FIRST folding core for unbound CypA is large relative to the protein, there were many residues which were part of the FIRST folding core but not the HDX folding core. The combined approach, considering residues which are both buried and rigid, improved the predictions. Another way of taking the level of surface exposure into account during FIRST simulations was to change the HB/HP settings, only permitting HB and HP interactions when both of the atoms involved are buried. This makes the surface residues of the protein more flexible, reducing the bond network and also the size of the FIRST folding core. In turn, the FIRST folding core improved as a predictor of the HDX folding core, containing fewer false positive results. We also observed that the secondary structure motifs strongly match the HDX folding core of unbound CypA.

With the goal of exploring the link between computational measures of protein rigidity and mobility and experimental data, we carried out our own HDX NMR experiments. We probed the effects of ligand binding by running experiments on both unbound CypA and the CypA-CsA complex. In order to do this, we first purified and characterised the protein using standard biochemical techniques. We adapted an established purification protocol [Liu et al., 1990] using the resources available. Circular dichroism (CD) was applied to measure the percentage composition of the secondary structure and it was found to be in agreement with the available structural data. The stability of the folded protein with respect to temperature was assessed using CD under different buffer conditions (phosphate and acetate buffers, various pH). The binding of CsA to CypA alters the environment of residue W121, resulting in enhanced fluorescence [Liu et al., 1990]. We observed the formation of the CypA-CsA by measuring this effect using fluorescence spectroscopy.

Having established these characteristics, we adapted our expression protocol so that we could make $^{15}$N-labelled CypA for use in NMR. The 3D $^{15}$N-$^{1}$H TOCSY-HSQC and $^{15}$N-$^{1}$H NOESY-HSQC experiments were used along with the available published structural data to assign the $^{15}$N-$^{1}$H HSQC spectra for unbound CypA as well as for the CypA-CsA complex. In agreement with previously published data [Neri et al., 1991; Spitzfaden et al., 1994] these spectra are highly similar, indicating subtle changes upon ligand binding. We collected HDX data for the unbound CypA and the CypA-CsA complex and consequently established the HDX folding core in both cases. The folding core for the complex was compared with the HDX folding core for the unbound protein as determined by both our own dataset and that of a previous publication [Shi et al., 2006].

The mobility simulations, and in particular the predictions of HDX folding cores, were revisited in the light of our new experimental data. Theoretical predictions of the folding core of the CypA-CsA complex match well with the HDX folding core. The HDX folding cores for the unbound protein and the CypA-CsA complex are highly similar. The theoretical methods of folding core prediction are not currently accurate enough to correctly predict small changes caused by ligand binding.

## 7.2 Context and outlook

A natural extension of our experimental work would be to study the HIV-1 capsid protein CA and characterise the CypA-CA complex through HDX. We would then be able to use the three HDX datasets and compare the binding of both CsA and CA. The exact role played by CypA during HIV-1 infection is not fully understood and is currently being investigated [Schaller et al., 2011; Takemura et al., 2013]. Some CA mutants such as N121K appear to alter the function of CypA so that it becomes inhibitory to HIV-1 replication, in certain cell lines [Takemura et al., 2013]. Performing HDX experiments in the presence of CA and the mutated CA may elucidate differences in ligand binding which help establish the reasons for this change of function. The binding of the CypA-CsA complex to calcineurin (CN) is also of interest, since it is the formation of this complex which results in the realisation of the function of CsA as an immunosuppressant. There is one crystal structure of this complex in the PDB (1MF8), in which it appears that CN wraps around the CypA-CsA complex at the CypA-CsA binding site. This would likely affect the chemical environment of the residues in the flexible loop regions which surround this site. The HDX folding core of CypA in the CypA-CsA-CN complex may therefore be substantially larger than its HDX folding cores when unbound or in the CypA-CsA complex. It would be interesting to see whether our theoretical methods are sensitive enough to predict this change if indeed it does occur.

Improving the sensitivity of these methods remains a goal for further study. The prediction of HDX folding cores through computational methods is of ongoing interest [Tartaglia et al., 2007; Liu et al., 2011; Lobanov et al., 2013; Thomas et al., 2013]. A recent study applied FIRST to a set of NMR refinement structures and used surface exposure information to refine HDX predictions [Sljoka and Wilson, 2013]. We feel that pairing surface exposure with rigidity will prove to be the way to accurately predict HDX folding cores. Methods for predicting the HDX folding core from the amino acid sequence of the protein have also been developed [Tartaglia et al., 2007; Lobanov et al., 2013]. We used the web server in [Liu et al., 2011] to predict the folding core of CypA, but the results did not compare favourably with the theoretical methods discussed in Chapter 4 ($s(AB) = 1.49$, $\%_C = 53.42\%$, $\%_A = 66.10\%$). The advantage of such a method is clearly that the input data is simple and no crystal structure needs to have been resolved.

The DXCOREX algorithm has been used to simulate the results of HDX coupled with mass spectrometry (HDX MS) [Liu et al., 2011]. To our knowledge, FIRST and FRODA have not been used to try to predict the results of HDX MS. We focused on predicting the HDX of the protein backbone in our simulations. This approach was tailored towards HDX NMR, where only the exchange of N-H groups in the protein is observed. In HDX MS, all proton-deuteron exchange events in the protein or peptide fragment are detected as an increase in mass. We would need to process the results of our simulations differently in order to consider the surface exposure of all exchangeable hydrogen atoms in the protein. Such modifications are feasible. Proteins such as integral membrane proteins can be studied using HDX MS [Busenlehner et al., 2006, 2007; Zhang et al., 2010], expanding the potential application of folding core predictions through FIRST and FRODA. These computational methods are applicable to large system sizes [Rader et al., 2004; Wang et al., 2004] and so using them to try to predict the results of HDX MS on integral membrane proteins is an excellent challenge.

A major advantage for theoreticians addressing these kinds of problems is the large, and ever-increasing, amount of structural data available. Using the example of HIV-1 protease, at the time of writing (September 2013) there have been 36

X-ray crystal structures deposited in the PDB since the beginning of the year, with a further 21 structures currently unreleased. This wealth of structural data facilitates the kind of large-scale rigidity analysis study, using a diverse set of input structures, that we have presented in this thesis. We are currently investigating the use of measures extracted from the RDMs to test for a correlation between the binding affinity of a ligand and the effect that ligand binding has on the rigidity of the protein. The development of computational techniques which can predict these experimentally determined quantities is a prevalent concern [Greenidge et al., 2013]. Within this investigation lies the challenge of implementing FIRST to produce results which are comparable not just for a set of structures of a single protein or protein family but across all proteins. To achieve this, the dependence on the input structure or structures needs to be reduced.

Other recent investigations have addressed the dependence on input data by incorporating the notion of an ensemble of structures for rigidity analysis [Sljoka and Wilson, 2013; Pfleger and Gohlke, 2013]. FIRST has been applied to a set of refinement structures generated from NMR data [Sljoka and Wilson, 2013]. This could also be done with X-ray crystallography, but there are few X-ray crystal structures which have been reported as an ensemble in this way [Levin et al., 2007; Kondrashov et al., 2008]. The method reported involves averaging HB strengths from 20 input structures and accounting for the absence of some of the HBs from some of the structures with a persistence penalty. The penalty systematically weakens the bonds according to their prevalence in the input ensemble. A highly persistent bond removes more degrees of freedom than a bond which is mostly absent. Through this method the HDX folding core of the acylphosphase enzyme from the hyperthermophile *Sulfolobus solfataricus* is accurately predicted ($s(AB) = 2.64$, $\%_C = 96.55$, $\%_A = 75.68$). The conditions placed on the average bond strengths and persistence involve selecting an $E_{\text{cut}}$ value, which is taken to be $-1$ kcal/mol. The impact of this choice is not discussed, and this may cause problems when generalising the method to other proteins. We would be keen to see a repeat of the study on the same structure with a ligand bound. The high quality of the prediction also comes at the price of requiring sophisticated structural input. There are far fewer NMR ensembles available to model than there are X-ray crystal structures [The Protein Data Bank, 2013]. The authors comment that the persistence data could be incorporated into the RCDs generated by FIRST. We would be interested in integrating the idea of HB persistence in to a study involving many X-ray crystal structures collected from diverse sources.

Another recent development is the use of constraint network analysis to model the flickering of hydrogen bonds of a protein in solution, whereby the HBs are continually breaking and re-forming [Krüger et al., 2013; Pfleger and Gohlke, 2013]. In this way, a single input structure can be represented by an ensemble of networks. The use of an ensemble network ensures that the results of rigidity analysis are independent of the input structure. Results from such ensemble networks have been compared favourably with those from MD simulations. These networks have not been applied directly to the problems of ligand binding or HDX and so this presents another avenue for further investigation. This may require statistically weighting the influence of the surface exposure of the bonding atoms in the network.

It appears that the application of rigidity analysis in FIRST is in a state of evolution towards a more robust approach which captures better the rigidity of the *protein* rather than merely the input protein *structure*. This is being done through different methods which endeavour to capture the nature of the protein structural ensemble and reduce the dependence on the fixed geometry of a single input structure [Heal et al., 2012; Sljoka and Wilson, 2013; Pfleger and Gohlke, 2013]. We hope that these approaches will meet the challenge of predicting HDX accurately and consistently for a large number of proteins, with sufficient sensitivity to capture the impact of ligand binding. The work in this thesis, where we observe subtle ligand binding effects, underlines the magnitude of the challenge.

# Appendix A

# Amino acid codes

Table A.1: The three-letter and one-letter amino acid codes used throughout the thesis.

| **Amino Acid** | **Three-Letter Code** | **One-Letter Code** |
|---|---|---|
| Alanine | Ala | A |
| Arginine | Arg | R |
| Asparagine | Asn | N |
| Aspartate | Asp | D |
| Cysteine | Cys | C |
| Glutamate | Glu | E |
| Glutamine | Gln | Q |
| Glycine | Gly | G |
| Histidine | His | H |
| Isoleucine | Ile | I |
| Leucine | Leu | L |
| Lysine | Lys | K |
| Methionine | Met | M |
| Phenylalanine | Phe | F |
| Proline | Pro | P |
| Serine | Ser | S |
| Threonine | Thr | T |
| Tryptophan | Trp | W |
| Tyrosine | Tyr | Y |
| Valine | Val | V |

# Appendix B

# Further rigidity analysis of HIV-1 protease

## B.1 Adding hydrogen atoms to PDB files

Table B.1 gives the results for the eight structures that were analysed twice, varying the order of water removal and hydrogenation. In the main study, REDUCE was applied after the water molecules had been removed. As the table demonstrates, reversing the order of these two processes does not affect the $\Delta\Phi$ values. In fact, for six of the eight structures there is no difference in the $\Delta\Phi$ values and for the other two - 2F8G and 2HS1 - the difference is smaller than 0.002. The eight structures were each crystallised with one of the 'main inhibitors' and a range of $\Phi(I)$, $\Phi(U)$ and $\Delta\Phi$ values are represented within the selection.

Table B.1: A selection of eight structures were analysed (A) when REDUCE was applied *after* the removal of water molecules, as in the main study, and (B) when this order was reversed.

| PDB Code (Inhibitor) | A | | | B | | |
|---|---|---|---|---|---|---|
| | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
| 1D4Y (TPV) | 0.787 | 0.000 | 0.787 | 0.787 | 0.000 | 0.787 |
| 2F8G (DRV) | 0.011 | 0.009 | 0.001 | 0.011 | 0.009 | 0.001 |
| 2FGU (SQV) | 0.462 | 0.010 | 0.452 | 0.462 | 0.010 | 0.452 |
| 2FGV (SQV) | 0.726 | 0.080 | 0.646 | 0.726 | 0.080 | 0.646 |
| 2FXD (ATZ) | 0.027 | 0.128 | -0.100 | 0.027 | 0.128 | -0.100 |
| 2HS1 (DRV) | 0.021 | 0.077 | -0.055 | 0.023 | 0.077 | -0.053 |
| 3NU3 (APV) | 0.009 | 0.042 | -0.033 | 0.009 | 0.042 | -0.033 |
| 2NU9 (APV) | 0.013 | 0.025 | -0.012 | 0.013 | 0.025 | -0.012 |

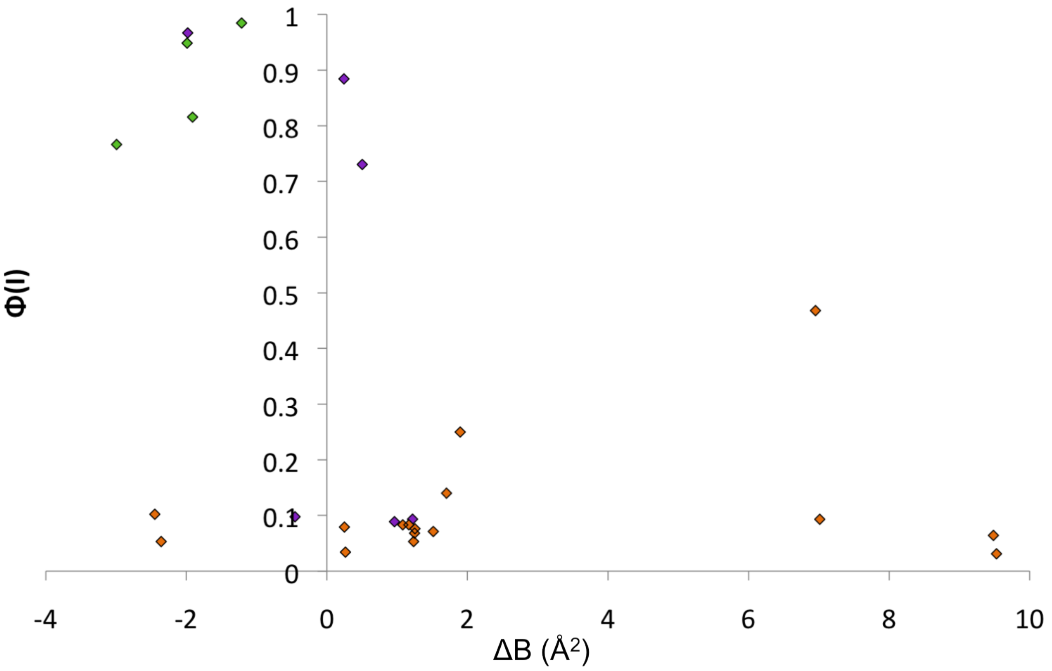


Figure B.1: Flap rigidity Φ(I) against relative flap B-factors ΔB for the structures crystallised with TPV (green), DRV (orange) and SQV (purple).

## B.2 X-ray crystal B-factors and their influence on ΔΦ values

In X-ray crystallographic structures, the B-factor in $\text{Å}^2$ is a measure of fit between the structure and the diffraction pattern from the crystal and can be used as an indication of the thermal motion [Rhodes, 2013]. We compared the B-factor values in the flap region with those for the rest of the protein, for the structures crystallised with DRV, TPV or SQV. The average B-factor for the entire structure was subtracted from the average B-factor for the flap region. The resulting value, ΔB, is negative when the flaps exhibit less thermal motion than the protein. In Figure B.1, Φ(I) is plotted against ΔB for the 26 structures crystallised with DRV, TPV or SQV.

There is no strong correlation between ΔB and Φ(I). For 22 of the structures, $-3\,\text{Å}^2 < \Delta\text{B} < 2\,\text{Å}^2$, and in this range values of $\Phi(\text{I}) = 0.034$ and $\Phi(\text{I}) = 0.985$ are observed.

## B.3 Complete data list for rigidity of HIV-1 protease

In Chapter 2, Figure 2.3 shows all of the 206 high resolution structures in complex with various ligands which have been analysed using FIRST. In Table B.2, the results of that rigidity analysis are shown. In addition, six high resolution structures which were uninhibited when crystallised are presented, taking the total number of structures presented to 212. It is interesting to note that in terms of rigidity, these six uninhibited crystal structures exhibited the same properties as the structures which have had their inhibitors deleted, as demonstrated by generally low values of Φ(U). This provides further justification that the removal of the inhibitor from the crystal structure does, in terms of rigidity analysis, offer a suitable representation of the unbound enzyme.

Table B.2: All of the 212 PDB structures analysed using First. The inhibitor names are in general taken from the PDB entries themselves. In some cases, the names are altered so as to make the table as uniform as possible. The 6 structures of the protease which were crystallised without an inhibitor are given at the base of the table; the structures are otherwise ordered first alphabetically according to inhibitor and subsequently alphabetically by PDB code. The resolution (Res) in Å of each structure is given along with the Φ(I), Φ(U) and ΔΦ values.

| PDB Code | Inhibitor | Mutation | Res | Φ(I) | Φ(U) | ΔΦ |
|---|---|---|---|---|---|---|
| 1F7A | 10 residue peptide | Wild-type | 2.00 | 0.000 | 0.000 | 0.000 |
| 1MT7 | 10 residue peptide | Various | 1.90 | 0.642 | 0.065 | 0.577 |
| 1TSQ | 10 residue peptide | Various | 2.00 | 1.000 | 0.071 | 0.930 |
| 2NXD | 10 residue peptide | Active site | 2.00 | 0.826 | 0.027 | 0.799 |
| 2NXL | 10 residue peptide | Active site | 2.00 | 0.593 | 0.035 | 0.559 |
| 3M9F | 10b | Wild-type | 1.80 | 0.017 | 0.013 | 0.004 |
| 3HDK | 2NC | Various | 1.80 | 0.043 | 0.043 | 0.000 |
| 3DOX | 3 and 4 residue peptides | Terminal | 2.00 | 0.086 | 0.086 | 0.000 |
| 2NPH | 4 and 5 residue peptides | Wild-type | 1.65 | 0.949 | 0.949 | 0.000 |
| 1SP5 | 5 residue peptide | Wild-type | 1.80 | 1.000 | 0.060 | 0.940 |
| 1NH0 | 5 residue peptidomimetic | Wild-type | 1.03 | 1.000 | 1.000 | 0.000 |
| 1IIQ | 6 residue peptide | Wild-type | 1.83 | 1.000 | 0.026 | 0.974 |

| PDB Code | Inhibitor | Mutation | Res | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 3B7V | 6 residue peptide | Wild-type | 1.46 | 0.091 | 0.091 | 0.000 |
| 3B80 | 6 residue peptide | Flaps | 1.50 | 0.417 | 0.087 | 0.330 |
| 1ZBG | 6 residue peptidomimetic | Core | 2.00 | 1.000 | 0.046 | 0.954 |
| 1ZJ7 | 6 residue peptidomimetic | Core | 1.93 | 1.000 | 1.000 | 0.000 |
| 1ZPK | 6 residue peptidomimetic | Core | 1.65 | 1.000 | 0.070 | 0.930 |
| 1ZSF | 6 residue peptidomimetic | Wild-type | 1.98 | 0.951 | 0.408 | 0.544 |
| 1DAZ | 7 residue peptide | Various | 1.55 | 1.000 | 0.114 | 0.886 |
| 1DW6 | 7 residue peptide | Various | 1.88 | 0.514 | 0.054 | 0.460 |
| 1PRO | A-98881 | Wild-type | 1.80 | 1.000 | 0.034 | 0.966 |
| 1HVL | A76 | Wild-type | 1.80 | 0.101 | 0.024 | 0.077 |
| 1HVI | A77 | Wild-type | 2.00 | 0.073 | 0.073 | 0.000 |
| 1HVJ | A78 | Wild-type | 2.00 | 0.234 | 0.012 | 0.223 |
| 1HVK | A79 | Wild-type | 1.80 | 0.334 | 0.030 | 0.304 |
| 1DIF | A79285 | Wild-type | 1.70 | 0.345 | 0.085 | 0.260 |
| 2HC0 | AB-2 | Various | 1.30 | 0.000 | 0.000 | 0.000 |
| 2I0D | AD-81 | Wild-type | 1.95 | 0.070 | 0.063 | 0.007 |
| 3GI6 | AD78 | Wild-type | 1.84 | 0.032 | 0.051 | -0.020 |
| 1AJX | AHA 001 | Wild-type | 2.00 | 0.805 | 0.012 | 0.793 |
| 1AJV | AHA 006 | Wild-type | 2.00 | 0.879 | 0.023 | 0.856 |
| 1G35 | AHA 024 | Wild-type | 1.80 | 1.000 | 0.021 | 0.979 |
| 1G2K | AHA 047 | Wild-type | 1.95 | 0.771 | 0.295 | 0.476 |
| 2CEN | AHA 478 | Wild-type | 1.70 | 1.000 | 0.074 | 0.926 |
| 2CEM | AHA 480 | Wild-type | 1.80 | 0.902 | 0.009 | 0.893 |
| 2UY0 | AHA 722 | Wild-type | 1.76 | 0.242 | 0.071 | 0.171 |
| 2FLE | AI | Core | 1.90 | 0.049 | 0.262 | -0.213 |
| 2BB9 | AKC4p_133a | Wild-type | 1.35 | 0.803 | 0.031 | 0.772 |
| 1XL5 | Amidhyroxysulfone | Wild-type | 1.73 | 0.068 | 0.012 | 0.056 |
| 3EKV | DRV | Wild-type | 1.75 | 0.100 | 0.020 | 0.080 |
| 3NU3 | DRV | Wild-type | 1.02 | 0.033 | 0.033 | 0.000 |
| 3NU4 | DRV | Core | 1.20 | 0.064 | 0.029 | 0.035 |
| 3NU5 | DRV | Flaps | 1.29 | 0.072 | 0.027 | 0.045 |
| 3NU6 | DRV | Flaps | 1.16 | 0.098 | 0.098 | 0.000 |

| PDB Code | Inhibitor | Mutation | Res | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 3NU9 | DRV | Core | 1.85 | 0.013 | 0.025 | -0.012 |
| 3NUJ | DRV | Flaps | 1.50 | 0.075 | 0.075 | 0.000 |
| 3NUO | DRV | Terminal | 1.35 | 0.073 | 0.073 | 0.000 |
| 2AQU | DRV | WT | 2.00 | 0.084 | 0.049 | 0.035 |
| 2FXD | DRV | Various | 1.60 | 0.953 | 0.122 | 0.831 |
| 3EKY | DRV | Wild-type | 1.80 | 0.218 | 0.099 | 0.119 |
| 3EL1 | DRV | Wild-type | 1.70 | 0.112 | 0.009 | 0.103 |
| 2A4F | Azacyclic urea inhibitor | Various | 1.90 | 0.591 | 0.090 | 0.501 |
| 1EBW | BEA 322 | Wild-type | 1.81 | 0.403 | 0.012 | 0.392 |
| 1EC0 | BEA 403 | Wild-type | 1.79 | 0.299 | 0.030 | 0.269 |
| 1D4I | BEA 425 | Wild-type | 1.81 | 0.182 | 0.232 | -0.050 |
| 1EC2 | BEA 428 | Wild-type | 2.00 | 0.897 | 0.109 | 0.788 |
| 1D4H | BEA 435 | Wild-type | 1.81 | 0.157 | 0.162 | -0.005 |
| 3BGC | Benzyl-peptide | Wild-type | 1.80 | 0.015 | 0.008 | 0.007 |
| 1W5V | C2-symmetric diol-based inhibitor | Wild-type | 1.80 | 0.648 | 0.008 | 0.640 |
| 1W5W | C2-symmetric diol-based inhibitor | Wild-type | 1.80 | 0.463 | 0.016 | 0.447 |
| 1W5X | C2-symmetric diol-based inhibitor | Wild-type | 1.90 | 0.682 | 0.036 | 0.646 |
| 1W5Y | C2-symmetric diol-based inhibitor | Wild-type | 1.90 | 0.262 | 0.024 | 0.238 |
| 1MTR | Cyclic peptidomimetic | Wild-type | 1.75 | 0.050 | 0.050 | 0.000 |
| 1MES | DMP | Core | 1.90 | 1.000 | 0.083 | 0.917 |
| 1MET | DMP | Core | 1.90 | 0.157 | 0.063 | 0.094 |
| 1MEU | DMP | Core | 1.90 | 0.608 | 0.033 | 0.575 |
| 1QBS | DMP | Wild-type | 1.80 | 0.265 | 0.042 | 0.223 |
| 3JVW | DMP | Terminal | 1.80 | 0.308 | 0.108 | 0.200 |
| 1DMP | DMP 450 | Wild-type | 2.00 | 0.054 | 0.036 | 0.018 |
| 1MER | DMP 450 | Core | 1.90 | 0.559 | 0.036 | 0.523 |
| 1T3R | DRV | Wild-type | 1.20 | 0.053 | 0.037 | 0.016 |
| 1T7I | DRV | Wild-type | 1.35 | 0.102 | 0.053 | 0.049 |
| 2F80 | DRV | Core | 1.45 | 0.083 | 0.014 | 0.070 |
| 2F81 | DRV | Terminal | 1.25 | 0.076 | 0.008 | 0.068 |
| 2F8G | DRV | Flaps | 1.22 | 0.071 | 0.009 | 0.063 |
| 2HS1 | DRV | Core | 0.84 | 0.140 | 0.065 | 0.076 |

| PDB Code | Inhibitor | Mutation | Res | Φ(I) | Φ(U) | ΔΦ |
|---|---|---|---|---|---|---|
| 2HS2 | DRV | Flaps | 1.22 | 0.064 | 0.070 | -0.006 |
| 3BVB | DRV | Active-site | 1.30 | 0.083 | 0.011 | 0.072 |
| 3CYW | DRV | Flaps | 1.40 | 0.068 | 0.013 | 0.054 |
| 3D1Z | DRV | Flaps | 1.30 | 0.250 | 0.232 | 0.018 |
| 3D20 | DRV | Flaps | 1.05 | 0.053 | 0.011 | 0.042 |
| 3EKT | DRV | Various | 1.97 | 0.093 | 0.020 | 0.072 |
| 3JVY | DRV | Terminal | 1.60 | 0.079 | 0.013 | 0.066 |
| 3JW2 | DRV | Terminal | 1.80 | 0.034 | 0.034 | 0.000 |
| 3LZS | DRV | Terminal | 1.95 | 0.031 | 0.015 | 0.016 |
| 3LZU | DRV | Terminal | 1.76 | 0.468 | 0.013 | 0.455 |
| 1A30 | Glu-Asp-Leu | Wild-type | 2.00 | 0.218 | 0.218 | 0.000 |
| 2WHH | GLU-PPN | Wild-type | 1.69 | 0.851 | 0.040 | 0.811 |
| 3DK1 | GRL-0105A | Wild-type | 1.07 | 0.163 | 0.058 | 0.105 |
| 3H5B | GRL-02031 | Wild-type | 1.29 | 0.092 | 0.018 | 0.074 |
| 3I6O | GRL-0216A | Wild-type | 1.17 | 0.019 | 0.013 | 0.006 |
| 3DJK | GRL-0255A | Wild-type | 1.00 | 0.101 | 0.021 | 0.081 |
| 3OK9 | GRL-0519A | Wild-type | 1.27 | 0.073 | 0.066 | 0.007 |
| 2HB3 | GRL06579 | Wild-type | 1.35 | 0.049 | 0.049 | 0.000 |
| 2QCI | GRL-98065 | Core | 1.20 | 0.079 | 0.009 | 0.070 |
| 2QD6 | GRL-98065 | Flaps | 1.28 | 0.029 | 0.068 | -0.039 |
| 2QD7 | GRL-98065 | Core | 1.11 | 0.029 | 0.036 | -0.007 |
| 2QD8 | GRL-98065 | Core | 1.35 | 0.057 | 0.019 | 0.038 |
| 2BBB | hh1_173_3a | Wild-type | 1.70 | 0.539 | 0.048 | 0.491 |
| 1SDT | IDV | Various | 1.30 | 0.065 | 0.057 | 0.008 |
| 1SDU | IDV | Various | 1.25 | 0.180 | 0.015 | 0.165 |
| 1SDV | IDV | Various | 1.40 | 0.123 | 0.023 | 0.100 |
| 2AVS | IDV | Flaps | 1.10 | 0.193 | 0.056 | 0.137 |
| 2AVV | IDV | Wild-type | 1.50 | 0.231 | 0.070 | 0.161 |
| 2J9J | JG-365 | Wild-type | 1.04 | 0.000 | 0.000 | 0.000 |
| 2JE4 | JG-365 | Wild-type | 1.07 | 1.000 | 0.000 | 1.000 |
| 2I0A | KB-19 | Wild-type | 1.80 | 0.068 | 0.040 | 0.028 |
| 3GI4 | KB60 | Wild-type | 1.85 | 0.069 | 0.064 | 0.005 |
| 3GI5 | KB62 | Wild-type | 1.80 | 0.036 | 0.084 | -0.048 |
| 1HPX | KNI-272 | Wild-type | 2.00 | 0.007 | 0.008 | -0.001 |
| 3FX5 | KNI-272 | Wild-type | 0.93 | 0.097 | 0.010 | 0.088 |
| 1MRW | KNI-577 | Various | 2.00 | 0.061 | 0.000 | 0.061 |

| PDB Code | Inhibitor | Mutation | Res | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 1MRX | KNI-577 | Various | 2.00 | 0.726 | 0.052 | 0.674 |
| 1MSM | KNI-764 | Various | 2.00 | 0.578 | 0.041 | 0.538 |
| 1MSN | KNI-764 | Various | 2.00 | 0.835 | 0.066 | 0.769 |
| 2PK5 | KNI-10075 | Various | 1.90 | 0.620 | 0.007 | 0.613 |
| 2PK6 | KNI-10033 | Various | 1.45 | 0.079 | 0.012 | 0.066 |
| 3KDD | KNI-10265 | Various | 1.80 | 0.287 | 0.051 | 0.236 |
| 3A2O | KNI-1689 | Wild-type | 0.88 | 0.223 | 0.020 | 0.202 |
| 2BPV | L-738-317 | Wild-type | 1.90 | 0.083 | 0.083 | 0.000 |
| 2BPY | L-739-622 | Wild-type | 1.90 | 0.387 | 0.387 | 0.000 |
| 1NPV | LDC271 | Wild-type | 2.00 | 0.808 | 0.447 | 0.361 |
| 1NPW | LGZ479 | Wild-type | 2.00 | 0.366 | 0.055 | 0.311 |
| 1ODY | LP-130 | Wild-type | 2.00 | 0.081 | 0.000 | 0.081 |
| 1B6J | Peptidomimetic | Wild-type | 1.85 | 1.000 | 0.166 | 0.834 |
| 1B6L | Peptidomimetic | Wild-type | 1.75 | 0.240 | 0.015 | 0.224 |
| 1B6M | Peptidomimetic | Wild-type | 1.85 | 0.519 | 0.023 | 0.497 |
| 1B6P | Peptidomimetic | Wild-type | 2.00 | 0.438 | 0.036 | 0.402 |
| 1D4K | Peptidomimetic | Wild-type | 1.85 | 0.163 | 0.017 | 0.147 |
| 1D4L | Peptidomimetic | Wild-type | 1.75 | 0.045 | 0.000 | 0.045 |
| 1Z1H | Peptidomimetic | Wild-type | 1.85 | 0.150 | 0.013 | 0.137 |
| 1Z1R | Peptidomimetic | Wild-type | 1.85 | 0.452 | 0.018 | 0.434 |
| 1SGU | MK1 | Various | 1.90 | 0.162 | 0.162 | 0.000 |
| 2AVO | MK1 | Core | 1.10 | 0.459 | 0.010 | 0.450 |
| 1EC3 | MSA367 | Wild-type | 1.80 | 1.000 | 0.016 | 0.984 |
| 1D4J | MSL370 | Wild-type | 1.81 | 0.155 | 0.013 | 0.143 |
| 2J9K | MVT-101 | Wild-type | 1.20 | 1.000 | 0.016 | 0.984 |
| 2O40 | MVT-101 | Various | 1.65 | 0.708 | 0.015 | 0.693 |
| 2PYM | NFV | Various | 1.90 | 0.171 | 0.225 | -0.054 |
| 2PYN | NFV | Wild-type | 1.85 | 0.033 | 0.021 | 0.012 |
| 3EKX | NFV | Wild-type | 1.97 | 0.045 | 0.091 | -0.046 |
| 3EL0 | NFV | Various | 2.00 | 0.013 | 0.000 | 0.013 |
| 2AID | Non-peptide. THK | Wild-type | 1.90 | 0.449 | 0.131 | 0.318 |
| 1FEJ | PEP | Terminal | 1.78 | 0.477 | 0.048 | 0.429 |
| 1FF0 | PEP | Flaps | 1.85 | 0.281 | 0.101 | 0.180 |
| 1FFI | PEP | Core | 1.70 | 0.780 | 0.043 | 0.737 |
| 1FG6 | PEP | Terminal | 1.80 | 0.732 | 0.107 | 0.626 |
| 1FGC | PEP | WT | 1.90 | 0.263 | 0.145 | 0.119 |

| PDB Code | Inhibitor | Mutation | Res | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 2AVQ | PEP | Flaps | 1.30 | 0.876 | 0.189 | 0.687 |
| 1FFF | RV(DCL)FEA(NLE) | Core | 1.90 | 0.862 | 0.080 | 0.782 |
| 1FG8 | RV(DCL)FEA(NLE) | Terminal | 1.85 | 0.767 | 0.158 | 0.609 |
| 1K1T | RV(DCL)FEA(NLE) | Various | 1.20 | 0.280 | 0.070 | 0.210 |
| 1K1U | RV(DCL)FEA(NLE) | Various | 1.55 | 0.804 | 0.018 | 0.786 |
| 1K2B | RV(DCL)FEA(NLE) | Terminal | 1.70 | 1.000 | 0.069 | 0.932 |
| 3BHE | Pyrrolidine derivative | Wild-type | 1.75 | 0.000 | 0.000 | 0.000 |
| 3BC4 | Pyrrolidine diester | Core | 1.82 | 0.072 | 0.072 | 0.000 |
| 3CKT | Pyrrolidine-based | Wild-type | 1.65 | 0.054 | 0.054 | 0.000 |
| 1XL2 | Pyrrolidinmethanamine | Wild-type | 1.50 | 0.671 | 0.000 | 0.671 |
| 1IZH | Q50 | Wild-type | 1.90 | 0.365 | 0.011 | 0.355 |
| 1HXW | RTV | Wild-type | 1.80 | 0.195 | 0.011 | 0.184 |
| 1RL8 | RTV | Core | 2.00 | 0.298 | 0.024 | 0.274 |
| 2F3K | RO1 | Wild-type | 1.60 | 0.048 | 0.048 | 0.000 |
| 2FGU | SQV | Core | 2.00 | 0.884 | 0.009 | 0.876 |
| 2FGV | SQV | Core | 1.50 | 0.967 | 0.091 | 0.876 |
| 2NMY | SQV | Core | 1.10 | 0.098 | 0.020 | 0.077 |
| 2NNK | SQV | Core | 1.25 | 0.093 | 0.080 | 0.014 |
| 3D1X | SQV | Flaps | 1.05 | 0.730 | 0.730 | 0.000 |
| 3D1Y | SQV | Flaps | 1.05 | 0.087 | 0.008 | 0.080 |
| 1BWB | SD146 | Core | 1.80 | 1.000 | 0.053 | 0.947 |
| 1QBR | SD146 | Wild-type | 1.80 | 0.588 | 0.038 | 0.550 |
| 1QBU | SD146 | Wild-type | 1.80 | 0.396 | 0.062 | 0.334 |
| 2AOE | Substrate analog CA-P2 | Core | 1.54 | 1.000 | 0.090 | 0.910 |
| 2FNS | Substrate analog NC-P1 | Active site | 1.85 | 0.600 | 0.000 | 0.600 |
| 2FNT | Substrate analog NC-P1 | Various | 1.44 | 1.000 | 0.701 | 0.300 |
| 2AOF | Substrate analog P1-P6 | Core | 1.32 | 0.845 | 0.344 | 0.501 |
| 2AOI | Substrate analog P1-P6 | Wild-type | 1.40 | 0.753 | 0.000 | 0.753 |
| 2AOC | Substrate analog P2-NC | Core | 1.30 | 0.773 | 0.024 | 0.749 |
| 2AOD | Substrate analog P2-NC | Wild-type | 1.40 | 0.820 | 0.948 | -0.128 |
| 2AOG | Substrate analog P2-NC | Core | 1.10 | 1.000 | 0.037 | 0.963 |
| 3BVA | Substrate analog P2-NC | Active site | 1.05 | 0.039 | 0.069 | -0.030 |
| 3FSM | Substrate analog P2-NC | Wild-type | 1.60 | 0.069 | 0.069 | 0.000 |
| 2AOH | Substrate analog P6-PR | Core | 1.42 | 0.605 | 0.000 | 0.605 |
| 2AOJ | Substrate analog P6-PR | Wild-type | 1.60 | 1.000 | 0.000 | 1.000 |
| 1D4Y | TPV | Various | 1.97 | 0.766 | 0.000 | 0.766 |

| PDB Code | Inhibitor | Mutation | Res | $\Phi(I)$ | $\Phi(U)$ | $\Delta\Phi$ |
|---|---|---|---|---|---|---|
| 2O4L | TPV | Various | 1.33 | 0.949 | 0.009 | 0.940 |
| 2O4N | TPV | Various | 2.00 | 0.816 | 0.018 | 0.798 |
| 2O4P | TPV | Terminal | 1.80 | 0.949 | 0.011 | 0.938 |
| 2AZ8 | TL-3 | Wild-type | 2.00 | 0.451 | 0.233 | 0.218 |
| 3KFR | TL-3 and 1F1 | Wild-type | 1.30 | 1.000 | 1.000 | 0.000 |
| 3KFS | TL-3 and 2F4 | Wild-type | 1.80 | 0.561 | 0.840 | -0.279 |
| 3KF0 | TL-3 and 4D9 | Wild-type | 1.80 | 0.481 | 0.038 | 0.443 |
| 3KFN | TL-3 and 4D9 | Wild-type | 1.77 | 0.079 | 0.008 | 0.071 |
| 3KFP | TL-3 and DMSOs | Wild-type | 1.77 | 0.029 | 0.000 | 0.029 |
| 3TLH | TL-3-093 | Wild-type | 2.00 | 0.574 | 0.000 | 0.574 |
| 2I4U | TMC-126 | Wild-type | 1.50 | 0.186 | 0.186 | 0.000 |
| 2A1E | TS-126 | Wild-type | 1.30 | 0.685 | 0.624 | 0.061 |
| 1HIV | U75875 | Wild-type | 2.00 | 0.297 | 0.062 | 0.235 |
| 2IDW | UIC-94017 | Core | 1.10 | 0.019 | 0.008 | 0.011 |
| 2IEN | UIC-94017 | Wild-type | 1.30 | 0.064 | 0.063 | 0.001 |
| 2IEO | UIC-94017 | Core | 1.53 | 0.049 | 0.062 | -0.012 |
| 3I7E | UIC-98038 | Wild-type | 1.70 | 0.091 | 0.091 | 0.000 |
| 1HPV | VX-478 | Wild-type | 1.90 | 0.022 | 0.020 | 0.002 |
| 1HVR | XK2 | Wild-type | 1.80 | 1.000 | 0.023 | 0.977 |
| 1BV7 | XV638 | Core | 2.00 | 1.000 | 0.013 | 0.987 |
| 1BV9 | XV638 | Core | 2.00 | 1.000 | 0.039 | 0.961 |
| 1BWA | XV638 | Core | 1.90 | 0.031 | 0.014 | 0.017 |
| 1ODX | Uninhibited | Core | 2.00 | N/A | 0.013 | N/A |
| 1RPI | Uninhibited | Various | 1.86 | N/A | 0.000 | N/A |
| 1TW7 | Uninhibited | Various | 1.30 | N/A | 0.000 | N/A |
| 3HBO | Uninhibited | Flaps | 1.71 | N/A | 0.359 | N/A |
| 3KT2 | Uninhibited | Terminal | 1.65 | N/A | 0.036 | N/A |
| 3KT5 | Uninhibited | Terminal | 1.80 | N/A | 0.734 | N/A |

# Appendix C

# NMR chemical shift assignment tables

Table C.1: Chemical shift assignment of CypA backbone $^{15}$N-$^{1}$H pairs, for the unbound protein as well as the CypA-CsA complex. Chemical shifts are given in ppm

| Residue | CypA only | | CypA-CsA | | Residue | CypA only | | CypA-CsA | |
|---|---|---|---|---|---|---|---|---|---|
| | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N | | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N |
| Thr5 | 8.78 | 115.02 | 8.76 | 115.03 | Phe22 | 9.51 | 119.11 | 9.50 | 119.02 |
| Val6 | 8.74 | 120.40 | 8.71 | 120.65 | Glu23 | 8.74 | 123.23 | 8.70 | 123.15 |
| Phe7 | 8.95 | 119.23 | 8.96 | 119.34 | Leu24 | 8.17 | 122.50 | 8.14 | 122.45 |
| Phe8 | 9.55 | 117.03 | 9.56 | 117.01 | Phe25 | 8.82 | 124.75 | 8.87 | 124.77 |
| Asp9 | 9.27 | 124.24 | 9.27 | 124.22 | Ala26 | 8.43 | 129.00 | 8.46 | 129.05 |
| Ile10 | 9.03 | 124.31 | 9.04 | 124.27 | Asp27 | 8.97 | 114.09 | 8.99 | 113.99 |
| Ala11 | 9.62 | 132.57 | 9.63 | 132.48 | Lys28 | 7.53 | 118.08 | 7.53 | 118.00 |
| Val12 | 8.93 | 118.47 | 8.94 | 118.31 | Val29 | 8.36 | 114.64 | 8.36 | 114.86 |
| Asp13 | 9.85 | 131.09 | 9.86 | 131.00 | Pro30 | N/A | N/A | N/A | N/A |
| Gly14 | 8.55 | 101.75 | 8.56 | 101.63 | Lys31 | 10.66 | 123.84 | 10.70 | 124.02 |
| Glu15 | 8.04 | 123.34 | 8.03 | 123.33 | Thr32 | 10.28 | 124.06 | 10.27 | 124.08 |
| Pro16 | N/A | N/A | N/A | N/A | Ala33 | 9.28 | 125.67 | 9.23 | 125.52 |
| Leu17 | 9.20 | 126.13 | 9.17 | 125.96 | Glu34 | 8.02 | 117.28 | 8.03 | 117.29 |
| Gly18 | 7.24 | 102.52 | 7.23 | 102.33 | Asn35 | 7.12 | 115.69 | 7.13 | 115.69 |
| Arg19 | 8.34 | 121.29 | 8.35 | 121.28 | Phe36 | 7.01 | 117.92 | 7.02 | 117.89 |
| Val20 | 9.37 | 126.94 | 9.38 | 126.92 | Arg37 | 8.94 | 121.10 | 8.95 | 121.15 |
| Ser21 | 8.76 | 120.36 | 8.77 | 120.39 | Ala38 | 8.69 | 119.16 | 8.62 | 118.81 |

| Residue | CypA only | | CypA-CsA | | Residue | CypA only | | CypA-CsA | |
|---|---|---|---|---|---|---|---|---|---|
| | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N | | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N |
| Leu39 | 8.17 | 120.78 | 8.20 | 120.86 | Gly74 | 8.71 | 114.04 | 8.80 | 114.03 |
| Ser40 | 7.89 | 119.28 | 7.89 | 119.30 | Gly75 | 8.11 | 109.15 | 8.06 | 108.90 |
| Thr41 | 7.97 | 108.49 | 7.98 | 108.56 | Lys76 | 6.97 | 115.71 | 6.93 | 115.50 |
| Gly42 | 7.57 | 108.40 | 7.57 | 108.24 | Ser77 | 7.79 | 114.41 | 7.75 | 114.44 |
| Glu43 | 8.01 | 118.7 | 8.01 | 118.70 | Ile78 | 8.55 | 111.24 | 8.55 | 111.10 |
| Lys44 | 9.10 | 118.55 | 9.11 | 118.38 | Tyr79 | 8.03 | 120.83 | 8.06 | 120.86 |
| Gly45 | 7.93 | 105.59 | 7.93 | 105.48 | Gly80 | 7.10 | 106.64 | 7.10 | 106.46 |
| Phe46 | 6.41 | 113.75 | 6.41 | 113.59 | Glu81 | | | | |
| Gly47 | 7.75 | 104.70 | 7.72 | 104.61 | Lys82 | 7.86 | 112.72 | 7.83 | 112.28 |
| Tyr48 | 6.88 | 113.79 | 6.86 | 113.89 | Phe83 | 9.17 | 116.64 | 9.20 | 116.47 |
| Lys49 | 8.47 | 124.92 | 8.48 | 124.84 | Glu84 | 9.23 | 119.54 | 9.28 | 119.46 |
| Gly50 | 9.48 | 117.90 | 9.52 | 117.97 | Asp85 | 8.59 | 118.90 | 8.62 | 118.81 |
| Ser51 | 8.38 | 116.50 | 8.40 | 116.58 | Glu86 | 9.45 | 131.63 | 9.51 | 131.82 |
| Cys52 | 10.01 | 115.30 | 10.10 | 115.31 | Asn87 | 7.05 | 106.84 | 7.04 | 106.78 |
| Phe53 | 8.69 | 123.05 | 8.66 | 122.99 | Phe88 | 8.33 | 113.02 | 8.32 | 112.90 |
| His54 | 7.56 | 119.88 | 7.52 | 119.70 | Ile89 | 8.30 | 119.89 | 8.29 | 119.85 |
| Arg55 | 7.04 | 123.14 | 6.87 | 124.30 | Leu90 | 7.74 | 117.28 | 7.75 | 117.20 |
| Ile56 | 9.19 | 126.48 | 9.13 | 127.27 | Lys91 | 8.07 | 119.05 | 8.07 | 119.00 |
| Ile57 | 8.54 | 123.02 | 8.49 | 127.36 | His92 | 10.70 | 122.57 | 10.58 | 122.33 |
| Pro58 | N/A | N/A | N/A | N/A | Thr93 | 7.26 | 110.48 | 7.25 | 110.42 |
| Gly59 | 9.73 | 114.24 | 9.85 | 114.11 | Gly94 | 7.48 | 107.31 | 7.53 | 107.37 |
| Phe60 | 8.17 | 119.37 | 8.10 | 119.19 | Pro95 | N/A | N/A | N/A | N/A |
| Met61 | 8.08 | 111.20 | 7.86 | 110.41 | Gly96 | 9.27 | 110.31 | 9.25 | 110.70 |
| Cys62 | 8.47 | 114.88 | | | Ile97 | 6.76 | 121.49 | 6.77 | 121.31 |
| Gln63 | 8.72 | 127.99 | 9.33 | 127.50 | Leu98 | 7.87 | 128.89 | 7.81 | 129.95 |
| Gly64 | 7.36 | 110.79 | 7.48 | 110.77 | Ser99 | 8.28 | 118.81 | 7.94 | 119.40 |
| Gly65 | 9.36 | 106.00 | 9.10 | 105.22 | Met100 | | | 8.28 | 123.27 |
| Asp66 | 9.96 | 124.06 | 10.01 | 124.40 | Ala101 | 7.99 | 125.97 | 8.22 | 126.95 |
| Phe67 | 6.62 | 116.04 | 6.69 | 116.00 | Asn102 | 8.11 | 113.40 | 7.68 | 113.59 |
| Thr68 | 7.28 | 109.02 | 7.25 | 108.73 | Ala103 | 8.77 | 123.37 | 9.18 | 121.03 |
| Arg69 | 8.65 | 122.09 | 8.68 | 122.05 | Gly104 | 8.18 | 109.38 | 8.24 | 107.71 |
| His70 | | | 6.53 | 111.11 | Pro105 | N/A | N/A | N/A | N/A |
| Asn71 | 7.49 | 112.49 | 7.44 | 112.24 | Asn106 | 8.86 | 119.13 | 8.87 | 119.05 |
| Gly72 | 9.66 | 110.60 | 9.67 | 111.25 | Thr107 | 10.21 | 110.39 | 10.30 | 110.31 |
| Thr73 | 7.93 | 112.20 | 7.86 | 109.83 | Asn108 | 7.37 | 120.61 | 7.42 | 120.56 |

| Residue | CypA only | | CypA-CsA | | Residue | CypA only | | CypA-CsA | |
|---|---|---|---|---|---|---|---|---|---|
| | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N | | $^{1}$H | $^{15}$N | $^{1}$H | $^{15}$N |
| Gly109 | 9.15 | 110.83 | 9.14 | 110.91 | Ile138 | 7.64 | 124.28 | 7.64 | 124.24 |
| Ser110 | 8.76 | 117.15 | 8.77 | 117.12 | Val139 | 7.24 | 121.97 | 7.26 | 122.16 |
| Gln111 | 8.37 | 124.52 | 8.41 | 124.79 | Glu140 | 8.26 | 117.69 | 8.31 | 117.82 |
| Phe112 | 8.05 | 117.86 | 8.15 | 118.29 | Ala141 | 7.49 | 121.13 | 7.49 | 121.04 |
| Phe113 | 9.80 | 116.50 | 9.83 | 115.02 | Met142 | 8.29 | 117.30 | 8.29 | 117.41 |
| Ile114 | | | 9.02 | 118.07 | Glu143 | 7.82 | 116.36 | 7.89 | 116.60 |
| Cys115 | 9.59 | 125.52 | 9.49 | 124.78 | Arg144 | 7.02 | 114.61 | 7.02 | 114.63 |
| Thr116 | 8.95 | 115.72 | 9.04 | 115.71 | Phe145 | 7.61 | 115.42 | 7.62 | 115.25 |
| Ala117 | 7.61 | 122.32 | 7.58 | 122.41 | Gly146 | 7.50 | 104.75 | 7.56 | 104.82 |
| Lys118 | 8.69 | 119.87 | | | Ser147 | 8.20 | 110.11 | 8.19 | 109.89 |
| Thr119 | 7.61 | 120.21 | 7.37 | 118.30 | Arg148 | | | | |
| Glu120 | 9.09 | 124.63 | 9.06 | 124.65 | Asn149 | | | | |
| Trp121 | 7.25 | 117.92 | 7.30 | 118.30 | Gly150 | 8.04 | 110.22 | 8.06 | 110.65 |
| Leu122 | 7.01 | 120.05 | 7.05 | 120.72 | Lys151 | 7.53 | 119.89 | 7.47 | 119.80 |
| Asp123 | 7.59 | 122.24 | 7.56 | 122.37 | Thr152 | 8.85 | 116.65 | 8.86 | 116.81 |
| Gly124 | 9.53 | 111.33 | 9.57 | 111.14 | Ser153 | 9.41 | 117.05 | 9.46 | 117.28 |
| Lys125 | 7.74 | 115.68 | 7.57 | 114.75 | Lys154 | 7.53 | 119.37 | 7.53 | 119.27 |
| His126 | 7.61 | 120.21 | | | Lys155 | 8.78 | 121.73 | 8.82 | 121.79 |
| Val127 | | | 8.24 | 124.74 | Ile156 | 9.61 | 134.42 | 9.62 | 134.33 |
| Val128 | 9.47 | 133.07 | 9.38 | 132.94 | Thr157 | 9.24 | 117.07 | 9.22 | 116.94 |
| Phe129 | 8.10 | 117.87 | 8.11 | 117.89 | Ile158 | 8.58 | 121.70 | 8.59 | 121.66 |
| Gly130 | 7.36 | 110.79 | 7.36 | 110.73 | Ala159 | 8.87 | 132.53 | 8.87 | 132.53 |
| Lys131 | 8.34 | 115.36 | 8.30 | 115.06 | Asp160 | 8.06 | 111.55 | 8.06 | 111.51 |
| Val132 | 9.03 | 124.26 | | | Cys161 | 8.58 | 116.13 | 8.59 | 116.09 |
| Lys133 | 9.47 | 131.87 | 9.44 | 131.73 | Gly162 | 6.84 | 104.18 | 6.84 | 104.04 |
| Glu134 | 7.53 | 118.49 | 7.53 | 118.50 | Gln163 | 9.05 | 121.01 | 9.08 | 121.11 |
| Gly135 | | | | | Leu164 | 8.59 | 126.15 | 8.59 | 125.92 |
| Met136 | 8.85 | 122.66 | 8.86 | 122.66 | Glu165 | 8.13 | 126.36 | 8.12 | 126.08 |
| Asn137 | 9.05 | 114.40 | 8.91 | 114.42 | | | | | |

# Bibliography

S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106:1589–1615, 2006.

V. Alexandrov, U. Lehnert, N. Echols, D. Milburn, D. Engelman, and M. Gerstein. Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. *Prot. Sci.*, 14:633–643, 2005.

J. B. Alimonti, T. B. Ball, and K. R. Fowke. Mechanisms of CD4$^+$ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. *J. Gen. Virol.*, 84:1649–1661, 2003.

N. T. Amin, A. K. Wallis, S. A. Wells, M. L. Rowe, R. A. Williamson, M. J. Howard, and R. B. Freedman. High-resolution NMR studies of structure and dynamics of human ERp27 indicate extensive interdomain flexibility. *Biochem. J.*, 450:321–332, 2013.

E. Arnold, K. Das, J. Ding, P. N. Yadov, Y. Hsiou, P. L. Boyer, and S. H. Hughes. Targeting HIV reverse transcriptase for anti-AIDS drug design: structural and biological considerations for chemotherapeutic strategies. *Drug Des. Discov.*, 13(3-4):29–47, 1996.

I. Bahar and A. J. Rader. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, 15(5):586–592, 2005.

R. L. Baldwin. The nature of protein folding pathways: the classical versus the new view. *J. Biomol. NMR*, 5:103–109, 1995.

F. Baneyx. Recombinant protein expression in *Escherichia coli*. *Curr. Opin. Biotech.*, 10: 411–421, 1999.

F. N. Barrera, M. del Álamo, M. G. Mateu, and J. L. Neira. Envelope lipids regulate the in vitro assembly of the HIV-1 capsid. *Biophys J: Biophys. Lett.*, pages L08–L10, 2007.

D. J. Bergsma, C. Eder, M. Gross, H. Kersten, D. Sylvester, E. Appelbaum, D. Cusimano, G. P. Livi, M. M. McLaughlin, K. Kasyan, T. G. Porter, C. Silverman, D. Dunnington, A. Hand, W. P. Prichett, M. J. Bossard, M. Brandt, and M. A. Levy. The cyclophilin multigene family of peptidyl-prolyl isomerases. *J. Biol. Chem.*, 266(34):23204–23214, 1991.

C. A. Bewley and S. Shahzad ul Hussan. Characterizing carbohydrate–protein interactions by nuclear magnetic resonance spectroscopy. *Biopolymers*, 99(10):796–896, 2013.

M. Bienkowska-Haba, H. D. Patel, and M. Sapp. Target cell cyclophilins facilitate human papillomavirus type 16 infection. *PLoS Pathogens*, 5(7):e1000524, 2009.

D. A. Bosco and D. Kern. Catalysis and binding of cyclophilin A with different HIV-1 capsid constructs. *Biochem.*, 43:6110–6119, 2004.

S. Bose, M. Mathur, P. Bates, N. Joshi, and A. K. Banerjee. Requirement for cyclophilin A for the replication of vesicular stomatitis virus New Jersey serotype. *J. Gen. Virol.*, 84: 1687–1699, 2003.

B. M. Branson and J. D. Stekler. Detection of acute HIV infection: we can't close the window. *J. Infect. Dis.*, 205:521–524, 2012.

D. P. Brazil and B. A. Hemmings. Ten years of protein kinase B signalling: a hard Akt to follow. *Trends Biochem. Sci.*, 26(11):657–664, 2001.

J. M. Brenchley, T. W. Schacker, L. E. Ruff, D. A. Price, J. H.Taylor, G. J. Beilman, P. L. Nguyen, A. Khoruts, M. Larson, A. T. Haase, and D. C. Douek. $CD4^+$ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J. Exp. Med.*, 200(6):749–759, 2004.

D. Bucher, B. J. Grant, and J. A. McCammon. Induced fit of conformational selection? the role of the semi-closed state in the maltose binding protein. *Biochemistry*, 50:10530–10539, 2011.

S. G. Burton, D. A. Cowan, and J. M. Woodley. The search for the ideal biocatalyst. *Nat. Biotechnol.*, 20:37–45, 2002.

L. S. Busenlehner, L. Salomonsson, P. Brzezinski, and R. N. Armstrong. Mapping protein dynamics in catalytic intermediates of redox-driven proton pump cytochrome *c* oxidase. *P. Natl. A. Sci. USA*, 103(42):15398–15403, 2006.

L. S. Busenlehner, J. Ålander, C. Jegerscöhld, P. J. Holm, P. Bhakat, H. Herbert, R. Morgenstern, and R. N. Armstrong. Location of substrate binding sites within the integral membrane protein microsomal glutathione transferase-1. *Biochem.*, 46:2812–2822, 2007.

P. A. Cane. New developments in HIV drug resistance. *J. Antimicrob. Chemoth.*, 64(1): i37–i40, 2009.

D. C. Chan, D. Fass, and J. M. Berger. Core structure of gp41 from the HIV envelope glycoprotein. *Cell*, 89:263–273, 1997.

C-E. Chang, T. Shen, J. Trylska, V. Tozzini, and J. A. McCammon. Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophys. J.*, 90:3880–3885, 2006.

T-W. Chun and A. S. Fauci. HIV reservoirs: pathogenesis and obstacles to viral eradication and cure. *AIDS*, 26:1261–1268, 2012.

H. J. Cleaves II. The origin of the biologically coded amino acids. *J. Theor. Biol*, 263: 490–498, 2010.

C. Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.*, 18:10–15, 2008.

Jack R. Collins, Stanley K. Burt, and John W. Erickson. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nat. Struct. Mol. Biol.*, 2(4):334–338, 1995.

P. M. Colman and B. J. Smith. The trypanosomal trans-sialidase: two catalytic functions associated with one catalytic site. *Structure*, 10:1466–1468, 2002.

S. D. Copley. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.*, 7:265–272, 2003.

D. H. A. Corrêa and C. H. I. Ramos. The use of circular dichroism spectroscopy to study protein folding, form and function. *Afr. J. Biochem. Res.*, 3(5):164–173, 2009.

B. R. Cullen. HIV-1 infection: fooling the gatekeeper. *Nat. Med.*, 9(9):1112–1113, 2003.

B. I. Dahiyat, D. B. Gordon, and S. L. Mayo. Automated design of the surface positions of protein helices. *Prot. Sci.*, 6:1333–1337, 1997.

P. L. Darke, R. F. Nutt, S. F. Brady, V. M. Garsky, T. M. Ciccarone, C-T. Leu, P. K. Lumma, R. M. Freidinger, D. F. Veber, and I. S. Sigal. HIV-1 protease specificity of peptide cleavage is sufficient for processing of Gag and Pol polyproteins. *Biochem. Bioph. Res. Co.*, 156(1):297–303, 1998.

R. S. DeWitte and E. I. Shakhnovich. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Prot. Sci.*, 3:1570 – 1581, 1994.

R. Diamond. On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Cryst.*, A46:425–435, 1990.

K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

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

K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338 (6110):1042–1046, 2012.

K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

S. E. Dobbins, V. I. Lesk, and M. J. E. Sternberg. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking. *P. Natl. A. Sci. USA*, 105(30):10390–10395, 2008.

M. Doucleff, M. Hatcher-Skeers, and N. J. Crane. *Pocket Guide to Biomolecular NMR*. Springer, 2011.

E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(3):117–121, 2005.

J. Fanghänel and G. Fischer. Thermodynamic characterization of the interaction of human cyclophilin 18 with cyclosporin A. *Biophys. Chem.*, 100:351–366, 2003.

E. Fayard, L. A. Tintignac, A. Baudry, and B. A. Hemmings. Protein kinase B/Akt at a glance. *J. Cell. Sci.*, 118(24):5675–5678, 2005.

A. R. Fersht, A. Matouschek, and L. Serrano. The folding of an enzyme: 1. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, 224:771–782, 1992.

L. Fielding. NMR methods for the determination of protein-ligand dissociation constants. *Prog. Nucl. Mag. Res. Sp.*, 51:219–242, 2007.

G. Fischer, B. Wittmann-Liebold, K. Lang, T. Kiefhaber, and F. X. Schmid. Cyclophilin and peptidyl-prolyl cis-trans isomerase are probably identical proteins. *Nature*, 337:476–478, 1989.

G. Fischer, P. Gallay, and S. Hopkins. Cyclophilin inhibitors for the treatment of HCV infection. *Curr. Opin. Invest. Dr.*, 11(8):911–918, 2010.

Flexweb. Analysis of flexibility in biomolecules and networks, 2013. URL `http://flexweb.asu.edu/software/first/`.

B. Folch, M. Rooman, and Y. Dechouck. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. *J. Chem. Inf. Model.*, 48:119–127, 2008.

A. D. Frankel and J. A. T. Young. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.*, 67:1–25, 1998.

E. O. Freed. HIV-1 Gag proteins: diverse functions in the virus life cycle. *Virology*, 251: 1–15, 1998.

Darón I. Freedberg, Rieko Ishima, Jaison Jacob, Yun-Xing Wang, Irina Kustanovich, John M. Louis, and Dennis A. Torchia. Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Prot. Sci.*, 11(2):221–232, 2001.

M. Fuxreiter, C. Magyar, T. Juhász, Z. Szeltner, L. Polgár, and I. Simon. Flexibility of prolyl oligopeptidase: Molecular dynamics and molecular framework analysis of the potential substrate pathways. *Prot: Struct. Func. Bioinf.*, 60:504–512, 2005.

M. Gastmans, G. Volckaert, and Y. Engelborghs. Tryptophan microstate reshuffling upon the binding of cyclosporin A to human cyclophilin A. *Prot: Struct. Func. Gen.*, 35: 464–474, 1999.

Arun K. Ghosh, Z. L. Dawson, and H. Mitsuya. Darunavir, a conceptually new HIV-1 protease inhibitor for the treatment of drug-resistant HIV. *Bioorganic and Med. Chem.*, 15(24):7576–7580, December 2007.

C. F. Gilks, S. Crowley, R. Ekpini, S. Gove, J. Perriens, Y. Souteyrand, D. Sutherland, M. Vitoria, T. Guerma, and K. DeCock. The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *Public Health*, 368:505–510, 2006.

T. D. Goddard and D. G. Kneller. *SPARKY 3*. University of California, San Francisco.

H. Gohlke and M. F. Thorpe. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.*, 91:2115–2120, 2006.

H. Gohlke, L. A. Kuhn, and D. A. Case. Change in protein flexibility upon complex formation: analysis of ras-raf using molecular dynamics and a molecular framework approach. *Prot: Struct. Func. Bioinf.*, 56:332–337, 2004.

N. Greenfield and G. D. Fasman. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochem.*, 8(10):4108–4116, 1969.

P. A. Greenidge, C. Kramer, J-C. Mozziconacci, and R. M. Wolf. MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement. *J. Chem. Inf. Model.*, 53(1):201–209, 2013.

S. Hallenberger, V. Bosch, H. Angliker, E. Shaw, H-D. Klenk, and W. Garten. Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature*, 360:358–361, 1992.

R. E. Handschumacher, M. W. Harding, J. Rice, R. J. Drugge, and D. W. Speicher. Cyclophilin: a specific cytosolic binding protein for cyclosporin A. *Science*, 226(4674): 544–547, 1984.

X. Hanoulle, A. Badillo, J-M. Wieruszeski, D. Verdegem, I. Landrieu, R. Bartenschlager, F. Penin, and G. Lippens. Hepatitis C virus NS5A protein is a substrate for the peptidyl-prolyl cis/trans isomerase activity of cyclophilins A and B. *J. Biol. Chem.*, 284(20): 13589–13601, 2009.

M. W. Harding, R. E. Handschumacher, and D. W. Speicher. Isolation and amino acid sequence of cyclophilin. *J. Biol. Chem.*, 261(18):8547–8555, 1986.

S. E. Harding and B. Chowdhry, editors. *Protein-ligand interactions: hydrodynamics and calorimetry.* Oxford Univ. Press, 2001.

William E. Harte and David L. Beveridge. Mechanism for the destabilization of the dimer interface in a mutant HIV-1 protease: a molecular dynamics study. *J. Am. Chem. Soc.*, 115(4):1231–1234, 1993.

William E. Harte, S. Swaminathan, and David L. Beveridge. Molecular dynamics of HIV-1 protease. *Prot: Struct. Func. Gen.*, 13(3):175–194, 1992.

Harvard Medical School. Hyperpolarized noble gas MRI laboratory, 2013. URL `http://www.spl.harvard.edu/archive/HypX/theory2.html`.

D. J. Harvey, F. Sobott, M. Crispin, A. Wrobel, C. Bonomelli, S. Vasiljevic, C. N. Scanlan, C. A. Scarff, K. Thalassinos, and J. H. Scrivens. Ion mobility mass spectrometry for extracting spectra of N-glycans directly from incubation mixtures following glycan release: Application to glycans from engineered glycoforms of intact, folded HIV gp120. *Am. Soc. Mass Spectrom.*, 22:568–581, 2011.

J. W. Heal, S. A. Wells, J. E. Jimenez-Roldan, R. B. Freedman, and R. A. Römer. Rigidity analysis of HIV-1 protease. *J. Phys. Conf. Series*, 286:01206, 2011.

J. W. Heal, J. E. Jimenez-Roldan, S. A. Wells, R. B. Freedman, and R. A. Römer. Inhibition of HIV-1 protease: the rigidity perspective. *Bioinformatics*, 28(3):350–357, 2012.

K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(6):913–918, 2007.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(13):964–972, December 2007.

B. M. Hespenheide, A. J. Rader, M. F. Thorpe, and L. A. Kuhn. Identifying protein folding cores: Observing the evolution of rigid and flexible regions during unfolding. *J. Mol. Graph. & Model.*, 21:195–207, 2002.

B. M. Hespenheide, D. J. Jacobs, and M.F. Thorpe. Structural rigidity and the capsid assembly of cowpea chlorotic mottle virus. *J. Phys.: Condens. Matter*, 16:S5055–S5064, 2004.

C. B. Hicks, P. Cahn, D. A Cooper, S. L. Walmsley, C. Katlama, B. Clotet, A. Lazzarin, M. A. Johnson, D. Neubacher, D. Mayers, and H. Valdez. Durable efficacy of tipranavir-ritonavir in combination with an optimised background regimen of antiretroviral drugs for treatment-experienced HIV-1-infected patients at 48 weeks in the Randomized Evaluation of Strategic Intervention in multi-drug reSistant patients with Tipranavir (RESIST) studies: an analysis of combined data from two randomised open-label trials. *The Lancet*, 368(9534):466–475, 2006.

D. M. Himmel, K. Das, A. D. Clark, Jr., S. H. Hughes, A. Benjahad, S. Oumouch, J. Guillemont, S. Coupa, A. Poncelet, I. Csoka, C. Meyer, K. Andries, C. H. Nguyen, D. S. Grierson, and E. Arnold. Crystal structures for HIV-1 reverse transcriptase in complexes with three pyridinone derivatives: A new class of non-nucleoside inhibitors effective against a broad range of drug-resistant strains. *J. Med. Chem.*, 48:7582–7591, 2005.

K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

J. Hiscott, H. Kwon, and P. Génin. Hostile takeovers: viral appropriation of the NF-$\kappa$B pathway. *J. Clin. Invest.*, 107(2):143–151, 2001.

L. Hong, X-J. Zhang, S. Foundling, J. A. Hartsuck, and J. Tang. Structure of a G48H mutant of HIV-1 protease explains how glycine-48 replacements produce mutants resistant to inhibitor drugs. *FEBS J.*, 420(1):11–16, 1997.

P. J. Hore. *Nuclear magnetic resonance.* Oxford Chemistry Primers. Oxford University Press, 1995.

J. P. Hornak. The basics of nmr, 2011. URL `http://www.cis.rit.edu/htbooks/nmr/index.html`.

Viktor Hornak and Carlos Simmerling. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discovery Today*, 12(3-4):132–138, 2007.

Viktor Hornak, Asim Okur, Robert C. Rizzo, and Carlos Simmerling. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Nat. Acad. Sci.*, 103(4):915 –920, 2006.

H. Husi and M. G. M. Zurini. Comparative binding studies of cyclophilins to cycosporin A and derivatives by fluorescence measurements. *Anal. Biochem.*, 222:251–255, 1994.

Rieko Ishima, Darón I Freedberg, Yun-Xing Wang, John M Louis, and Dennis A Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure*, 7(9):1047–1055, S1–S12, 1999.

L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for the structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, 254:260–288, 1995.

D. J. Jacobs. Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.*, 31:6653–6668, 1998.

D. J. Jacobs and S. Dallakyan. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.*, 88(2):903 – 915, 2005.

D. J. Jacobs and B. Hendrickson. An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comp. Biol.*, 137:346–365, 1997.

D. J. Jacobs and M. F. Thorpe. Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.*, 75:4051–4054, 1995.

D. J. Jacobs, L. A. Kuhn, and M. F. Thorpe. Flexible and rigid regions in proteins. *Rigidity Theory and Applications*, 1999.

D. J. Jacobs, S. Dallakyan, G. G. Wood, and A. Heckathorne. Network rigidity at finite temperature: relationships between thermodynamic stability, the non-additivity of entropy and cooperativity in molecular systems. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 68(6.1):061109, 2003.

D. J. Jacobs, D. R. Livesay, J. Hules, and M. L. Tasayco. Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model. *J. Mol. Biol.*, 358:882–904, 2006.

D.J. Jacobs and M.F. Thorpe. Generic rigidity percolation in two dimensions. *Phys. Rev. Lett.*, 75:4051–4054, 1996.

D.J. Jacobs, A.J. Rader, L.A. Kuhn, and M.F. Thorpe. Protein flexibility predictions using graph theory. *Prot: Struct. Func. Gen.*, 44:150–165, 2001.

L. C. James and D. S. Tawfik. Conformational diversity and protein evolution — a 60-year-old hypothesis revisited. *Trends Biochem. Sci.*, 28(7):361–368, 2003.

T. James. *Biophysics Textbook Online, Chapter 1: Fundamentals of NMR.* The Biophysical Society, 1998. URL `http://qudev.ethz.ch/content/courses/phys4/studentspresentations/nmr/James_Fundamentals_of_NMR.pdf`.

J. E. Jimenez-Roldan, S. A. Wells, R. B. Freedman, and R. Römer. Integration of FIRST, FRODA and nmm in a coarse grained method to study protein disulphide isomerase conformational change. *J. Phys. Conf. Series*, 286:012002, 2011.

J. E. Jimenez-Roldan, R. B. Freedman, R. A. Römer, and S. A. Wells. Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses. *Phys. Biol.*, 9: 016008, 2012.

C. C. Jolley, S. A. Wells, P. Fromm, and M. F. Thorpe. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophys. J.*, 94:1613–1621, 2008.

Craig C. Jolley, Stephen A. Wells, Brandon M. Hespenheide, Michael F. Thorpe, and Petra Fromme. Docking of photosystem I subunit C using a constrained geometric simulation. *J. Am. Chem. Soc.*, 128(27):8803–8812, 2006.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

J. O. Kahn and B. D. Walker. Acute human immunodeficiency virus type 1 infection. *New Engl. J. Med.*, 229(1):33–39, 1998.

J. Kallen, V. Mikol, P. Taylor, and M. Walkinshaw. X-ray structures and analysis of 11 cyclosporin derivatives complexed with cyclophilin A. *J. Mol. Biol.*, 283:435–449, 1998.

M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–652, 2002.

M. Karplus and G. A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347:631 – 639, 1990.

M. Karplus and D. L. Weaver. Protein folding dynamics: the diffusion-collision model and experimental data. *Prot. Sci.*, 3:650–668, 1994.

Etsuko Katoh, John M. Louis, Toshimasa Yamazaki, Angela M. Gronenborn, Dennis A. Torchia, and Rieko Ishima. A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex. *Prot. Sci.*, 12(7):1376–1385, 2003.

A. Kaul, S. Stauffer, C. Berger, T. Pertel, J. Schmitt, S. Kallis, M. Z. Lopez, V. Lohmann, J. Luban, and R. Bartenschlager. Essential role of cyclophilin A for hepatitis C virus replication and virus production and possible link to polyprotein cleavage kinetics. *PLoS Pathogens*, 5(8):e1000546, 2009.

H. Kawasaki, E. S. Mocarski, I. Kosugi, and Y. Tsutsui. Cyclosporine inhibits mouse cytomegolavirus infection via a cyclophilin-dependent patheway specifically in neural stem/progenitor cells. *J. Virol.*, 81(17):9013–9023, 2007.

H. Ke and Q. Huai. Crystal structures of cyclophilin and its partners. *Front. Biosci.*, 9: 2285–2296, 2004.

S. M. Kelly, T. J. Jess, and N. C. Price. How to study proteins by circular dichroism. *Biochim. Biophys. Acta*, 1751:119–139, 2005.

G. Kern, D. Kern, F. X. Schmid, and G. Fischer. A kinetic analysis of the folding of human carbonic anhydrase II and its catalysis by cyclophilin. *J. Biol. Chem.*, 270(2):740–745, 1995.

J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 19: 120–127, 2009.

J. E. Kohn, P. V. Afonine, J. Z. Ruscio, P. D. Adams, and T. Head-Gordon. Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLoS Comput. Biol.*, 6(8):e1000911, 2010.

D. A. Kondrashov, W. Zhang, R. Aranda IV, B. Stec, and G. N. Phillips Jr. Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Prot: Struct. Func. Bioinf.*, 70:353–362, 2008.

W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, 48:682–695, 2002.

D. M. Krüger, P. C. Rathi, C. Pfleger, and H. Gohlke. CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucl. Acids Res.*, pages 1–9, 2013.

J. Kubelka, J. Hofrichter, and W. A. Eaton. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.*, 14:76–88, 2004.

E. J. Levin, D. A. Kondrashov, G. E. Wesenberg, and G. N. Phillips Jr. Ensemble refinement of protein crystal structures. *Structure*, 15(9):1040–1052, 2007.

C. Levinthal. *Mössbauer Spectroscopy in Biological Systems (chapter: "How to fold graciously")*. University of Illinois Press, 1969.

M. Levitt. Molecular dynamics of native protein: I. computer simulation of trajectories. *J. Mol. Biol.*, 168:595–620, 1983.

J. A. Levy. HIV pathogenesis: 25 years of progress and persistent challenges. *AIDS*, 23: 147–160, 2009.

R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Prot. Sci.*, 8: 1571–1591, 1999.

S. Lichtiger, D. H. Present, A. Kornbluth, I. Gelernt, J. Bauer, G. Galler, F. Michelassi, and S. Hanauer. Cyclosporine in severe uclerative coliltis refractory to steroid therapy. *New. Eng. J. Med.*, 330(26):1841–1845, 1994.

H. Lilie, K. Lang, R. Rudolph, and J. Buchner. Prolyl isomerases catalyze antibody folding *in vitro*. *Prot. Sci.*, 2:1490–1496, 1993.

K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011.

J. Liu, M. W. Albers, C. Chen, S. L. Schreiber, and C. T. Walsh. Cloning, expression, and purification of human cyclophilin in Escherichia coli and assessment of the catalytic role of cysteines by site-directed mutagenesis. *P. Natl. A. Sci. USA*, 87:2304–2308, 1990.

T. Liu, D. Pantazatos, S. Li ans Y. Hamuro, V. J. Hilser, and V. L. Woods Jr. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *Am. Soc. Mass Spectrom.*, 23: 43–56, 2011.

Y. Liu, J. Jiang, P. L. Richardson, and R. D. Reddy. A fluorescence polarization-based assay for peptidyl prolyl cis/trans isomerase cyclophilin A. *Anal. Biochem.*, 356:100–107, 2006.

D. R. Livesay and D. J. Jacobs. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Prot: Struct. Func. Bioinf.*, 62:130–143, 2006.

M. Y. Lobanov, M. Y. Suvorina, N. V. Dovidchenko, I. V. Sokolovskiy, A. K. Surin, and O. V. Galzitskaya. A novel web server predicts amino acid residue protection against hydrogen-deuterium exchange. *Bioinformatics*, 29(11):1375–1381, 2013.

J. Luban, K. L. Bossolt, E. K. Franke, G. V. Kalpana, and S. P. Goff. Human immunodeficiency virus type 1 gag protein binds to cyclophilins A and B. *Cell*, 73:1067–1078, 1993.

A. Macchiarulo, R. Nuti, D. Bellochi, E. Camaioni, and R. Pellicciari. Molecular docking and spatial coarse graining simulations as tools to investigate substrate recognition, enhancer binding and conformational transitions in indoleamine-2,3-dioxygenase (IDO). *BBA-Proteins Proteom*, 1774:1058–1068, 2007.

T. Mamonova, B. Hespenheide, R. Straub, M.F. Thorpe, and M. Kurnikova. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.*, 2:S137–S147, 2005.

G. S. Martin. The hunting of the Src. *Nat. Rev. Mol. Cell Bio.*, 2:467–475, 2001.

M. Matsumoto, S. Saito, and I. Ohmine. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature*, 416:409–413, 2002.

V. Mikol, C. Papageorgiou, and X. Borer. The role of water molecules in the structure-based design of (5-hydroxynorvaline)-2-cyclosporin: Synthesis, biological activity, and crystallographic analysis with cyclophilin A. *J. Med. Chem.*, 38:3361–3367, 1995.

V. Mikol, P. Taylor, J. Kallen, and D. Walkinshaw. Conformational differences of an immunosuppressant peptolide in a single crystal and in a crystal complex with human cyclophilin A. *J. Mol. Biol.*, 283:451–461, 1998.

D. Morgan, C. Mahe, B. Mayanja, J. M. Okongo, R. Lubega, and J. A. G. Whitworth. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS*, 16(4):597–603, 2002.

J. L. Mott, D. Zhang, J. C. Freeman, P. Mikolajczak, S. W. Chang, and H. P. Zassenhaus. Cardiac disease due to random mitochondrial DNA mutations is prevented by cyclosporin A. *Biochem. Bioph. Res. Co.*, 319:1210–1215, 2004.

P. Neri, R. Meadows, G. Gemmecker, E. Olejniczak, D. Nettesheim, T. Logan, R. Simmer, R. Helfrich, T. Holzman, J. Severin, and S. Fesik. $^{1}$H, $^{13}$C and $^{15}$N backbone assignments of cyclophilin when bound to cyclosporin A (CsA) and preliminary structural characterization of the CsA binding site. *FEBS J.*, 294(1):81–88, 1991.

Linda K. Nicholson, Toshimasa Yamazaki, Dennis A. Torchia, Stephan Grzesiek, Ad Bax, Stephen J. Stahl, Joshua D. Kaufman, Paul T. Wingfield, Patrick Y.S Lam, Prabhakar K. Jadhav, C. Nicholas Hodge, Peter J. Domaille, and Chong-Hwan Chang. Flexibility and function in HIV-1 protease. *Nat. Struct. Mol. Biol.*, 2(4):274–280, 1995.

B. Nölting and D. A. Agard. How general is the nucleation–condensation mechanism? *Proteins*, 73(754-764), 2008.

B. Nordén, A. Rodger, and T. Dafforn. *Linear dichroism and circular dichroism: a textbook on polarized-light spectroscopy.* RSC Publishing, 2010.

R. B. Nussenblatt and A. G. Palestine. Cyclosporine: immunology, pharmacology and therapeutic uses. *Surv. Ophthalmol.*, 31(3):159–169, 1986.

M. Oliveberg and A. R. Fersht. Thermodynamics of transient conformations in the folding pathway of barnase: reorganization of the folding intermediate at low pH. *Biochemistry*, 35:2738–2749, 1996.

M. Ottiger, O. Zerbe, P. Güntert, and K. Wütrich. The NMR solution conformation of unligated human cyclophilin A. *J. Mol. Biol.*, 272:64–81, 1997.

C. Papageorgiou, A. Florineth, and V. Mikol. Improved binding affinity for cyclophilin A by a cyclosporin derivative singly modified at its effector domain. *J. Med. Chem.*, 37: 3674–3676, 1994.

L. R. Pearce, D. Komander, and D. R. Alessi. The nuts and bolts of ACG protein kinases. *Nat. Rev. Mol. Cell Bio.*, 11:9–22, 2010.

P. Petrone and V. S. Pande. Can conformational change be described by only a few normal modes? *Biophys J.*, 90:1583–1593, 2006.

C. Pfleger and H. Gohlke. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure*, 2013.

C. Pfleger, S. Radestock, E. Schmidt, and H. Gohlke. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.*, 34:220–233, 2013a.

C. Pfleger, P. C. Rathi, D. L. Klein, S. Radestock, and H. Gohlke. Constraint Network Analysis (CNA): a python software package for effciently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.*, 53(4):1007–1015, 2013b.

S. Piana, P. Carloni, and U. Rothlisberger. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Prot. Sci.*, 11:2393–2402, 2002.

S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomistic description of the folding of a dimeric protein. *J. Phys. Chem. B*, 2013.

A. V. Pisliakov, J. Cao, S. C. L. Kamerlin, and A. Warshel. Enzyme millisecond conformational dynamics do not catalyze the chemical step. *P. Natl. A. Sci. USA*, 106(41): 17359–17364, 2009.

J. Pokorná, L. Machala, P. Řezáčová, and J. Konvalinka. Current and novel inhibitors of HIV protease. *Viruses*, 1:1209–1239, 2009.

E. R. Price, L. D. Zydowsky, M. Jin, C. H. Baker, F. D. McKeon, and C. T. Walsh. Human cyclophilin B: a second cyclophilin gene encodes a peptidyl-prolyl isomerase with a signal sequence. *P. Natl. A. Sci. USA*, 88:1903–1907, 1991.

ExPASy ProtParam tool, 2013. URL `http://web.expasy.org/protparam/`.

A. J. Rader. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.*, page 016002, 2010.

A. J. Rader and I. Bahar. Folding core predictions from network models of proteins. *Polymer*, 45:659–668, 2004.

A. J. Rader and S. M. Brown. Correlating allostery with rigidity. *Mol. Biosyst.*, 7:464–471, 2011.

A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe. Protein unfolding: rigidity lost. *P. Natl. A. Sci. USA*, 99:3540–3545, 1999.

A. J. Rader, G. Anderson, B. Isin, H. G. Khorana, I. Bahar, and J. Klein-Seetharaman. Identification of core amino acids stabilizing rhodopsin. *Proc. Nat. Acad. Sci.*, 101:7246–7252, 2004.

S. Radestock and H. Gohlke. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.*, 8(5):507–522, 2008.

S. Radestock and H. Gohlke. Protein rigidity and thermophilic adaptation. *Prot: Struct. Func. Gen.*, 79:1089–1108, 2011.

P. C. Rathi, S. Radestock, and H. Gohlke. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.*, 159:135–144, 2012.

G. Rhodes. Judging the quality of macromolecular models: A glossary of terms from crystallography, NMR, and homology modeling, 2013. URL `http://spdbv.vital-it.ch/TheMolecularLevel/ModQual/`.

E J Rodríguez, C Debouck, I C Deckman, H Abu-Soud, F M Raushel, and T D Meek. Inhibitor binding to the Phe53Trp mutant of HIV-1 protease promotes conformational changes detectable by spectrofluorometry. *Biochem.*, 32(14):3557–3563, 1993.

E. Le Rouzic and S. Benichou. The Vpr protein from HIV-1: distinct roles along the viral life cycle. *Retrovirology*, 2:11, 2005.

K. Satoh, P. Nigro, and B. C. Berk. Oxidative stress and vascular smooth muscle cell growth: a mechanistic linkage by cyclophilin A. *Antioxid Redox Sign*, 12(5):675–682, 2010a.

K. Satoh, H. Shimokawa, and B. C. Berk. Cyclophilin A: promising new target in cardiovascular therapy. *Circ J.*, 74(11):2249–2256, 2010b.

K. Sayasith, G. Sauvé, and J. Yelle. Targeting HIV-1 integrase. *Expert Opin. Ther. Tar.*, 5 (4):443–464, 2001.

C. A. Scarff, K. Thalassinos, G. R. Hilton, and J. H. Scrivens. Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements. *Rapid Commun. Mass Sp.*, 22:3297–3304, 2008.

T. Schaller, K. E. Ocwieja, J. Rasaiyaah, A. J. Price, T. L. Brady, S. L. Roth, S. Hué, A. J. Fletcher, K. Lee, V. N. KewalRamani, M. Noursadeghi, R. G. Jenner, L. C. James, F. D. Bushman, and G. J. Towers. HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathogens*, 7(12): e1002439, 2011.

J. A. Schellman. Temperature, stability, and the hydrophobic interaction. *Biophys. J.*, 73: 2960–2964, 1997.

H. J. Schirra. Two dimensional NMR spectroscopy, 2013. URL `http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm`.

H. Schneider, N. Charara, R. Schmitz, S. Wehrli, V. Mikol, M. G. M. Zurini, V. F. J. Quesniaux, and N. R. Movva. Human cyclophilin C: Primary structure, tissue distribution, and determination of binding specificity for cyclosporins. *Biochemistry*, 33:8218–8224, 1994.

M. F. Schneider, S. J. Gange, C. M. Williams, K. Anastos, R. M. Greenblatt, L. Kingsley, R. Detels, and A. Muñoz. Patterns of the hazard of death after aids through the evolution of antiretroviral therapy: 1984–2004. *AIDS*, 19(17):2009–2018, 2005.

A. Schreyer and T. Blundell. CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug Des.*, 73:157–167, 2009.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. PyMOL, 2013. URL `http://www.pymol.org/`.

P. M. Sharp and B. H. Hahn. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med*, 1(a006841), 2011.

Y-H. Shi, D-H. Lin, J-Y. Huang, and X. Shen. Study of stability of cyclophilin A by NMR and circular dichroism spectra. *Chinese J. Chem.*, 24:973–979, 2006.

A. Sljoka and D. Wilson. Probing protein ensemble rigidity and hydrogen-deuterium exchange. pre-print, 2013.

C. Spitzfaden, W. Braun, G. Wider, H. Widmer, and K. Wüthrich. Determination of the NMR solution structure of the cyclophilin A-cyclosporin A complex. *J. Biomol. NMR*, 4:463–482, 1994.

H. F. Stähelin. The history of cyclosporin A (Sandimmune®) revisited: another point of view. *Experientia*, 52:5–13, 1996.

K. Suhre and Y-H. Sanejouand. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res. (Web Issue)*, 32:610–614, 2004.

S. Swaminathan, W. E. Harte, and David L. Beveridge. Investigation of domain structure in proteins via molecular dynamics simulation: application to HIV-1 protease dimer. *J. Am. Chem. Soc.*, 113(7):2717–2721, 1991.

T. Takemura, M. Kawamata, M. Urabe, and T. Murakami. Cyclophilin A-dependent restriction to capsid N121K mutant human immunodeficiency virus type 1 in a broad range of cell lines. *J. Virol.*, 87(7):4086–4090, 2013.

H. Tan and A. J. Rader. Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins*, 74:881–894, 2009.

C. Tanford. *The hydrophibic effect: formation of micelles and biological membranes.* Wiley, 1980.

C. Tang, E. Loeliger, I. Kinde, S. Kyere, K. Mayo, E. Barklis, Y. Sun, M. Huang, and M. F. Summers. Antiviral inhibition of the HIV-1 capsid protein. *J. Mol. Biol.*, 327:1013–1020, 2003.

G. G. Tartaglia, A. Cavalli, and M. Vendruscolo. Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure*, 15:139–143, 2007.

O. Tastan, E. Yu, M. Ganapathiraju, A. Aref, A. J. Rader, and J. Klein-Seetharaman. Comparison of stability predictions and simulated unfolding of rhodopsin structures. *Photochem. and Photobiol.*, 83:351–362, 2007.

The Protein Data Bank, 2013. URL `http://www.rcsb.org/pdb/`.

S. Thomas, L. Tapia, C. Ekenna, H-Y. Yeh, and N. M. Amato. Rigidity analysis for protein motion and folding core identification. Paper from the AAAI 2013 Workshop, 2013.

M. F. Thorpe, B. M. Hespenheide, Y. Yang, and L. A. Kuhn. Flexibility and critical hydrogen bonds in cytochrome c. *Pac. Symp. Biocomput.*, pages 191–202, 2000.

M. F. Thorpe, M. Lei, A. J. Rader, D. J. Jacobs, and L. A. Kuhn. Flexible and rigid regions in proteins. *J. Mol. Graph. Model*, pages 60–69, 2001.

M. M. Tirion. Large amplitude elastic motions in proteins from single-parameter atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.

H. Toh, M. Ono, K. Saigo, and T. Miyata. Retroviral protease-like sequence in the yeast transposon ty 1. *Nature*, 315:691, 1985.

N. Tokuriki and D. S. Tawfik. Protein dynamism and evolvability. *Science*, 324:203–207, 2009.

P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, 37(12): 509–516, 2012.

G. J. Towers, T. Hatziioannou, S. Cowan, S. P. Goff, J. Luban, and P. D. Bieniasz. Cyclophilin A modulates the sensitivity of HIV-1 to host restriction factors. *Nat. Med.*, 9 (9):1138–1143, 2003.

V. Tozzini and J. A. McCammon. A coarse grained model for the dynamics of flap opening in HIV-1 protease. *Chem. Phys. Lett.*, 413:123–128, 2005.

UniProt knowledge base, 2013. URL `http://www.uniprot.org/`.

A. Velazquez-Campoy, Y. Kiso, and E. Freire. The binding energetics of first- and second-generation HIV-1 protease inhibitors: implications for drug design. *Arch. Biochem. Biophys.*, 390(2):169–175, 2001.

P. Wang and J. Heitman. The cyclophilins. *Genome Biol.*, 6(7):226, 2005.

Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.*, 147:302–314, 2004.

S. A. Wells, S. Menor, B. M. Hespenheide, and M. F. Thorpe. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.*, 2:S127–S136, 2005.

S.A. Wells, J.E. Jimenez-Roldan, and R.A. Römer. Comparative analysis of rigidity across protein families. *Phys. Biol.*, 6(4):046005–11, 2009.

L. Whitmore and B. A. Wallace. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.*, 32: W668–W673, 2004.

D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, and B. D. Sykes. $^{1}$H, $^{13}$C and $^{15}$N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, 6:135–140, 1995.

A. Wlodawer and J. Vondrasek. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Bioph. Biom.*, 27:249–284, 1998.

C. Woodward. Is the slow-exchange core the protein folding core? *Trends Biochem. Sci.*, 18(10):359–360, 1993.

C. Woodward, I. Simon, and E. Tüchsen. Hydrogen exchange and the dynamic structure of proteins. *Mol. Cell Biochem.*, 48:135–160, 1982.

J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: Using hydrogen atoms contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285:1735–1747, 1999.

D. M. York, T. A. Darden, L. G. Pedersen, and M. W. Anderson. Molecular dynamics simulation of HIV-1 protease in a crystalline environment and in solution. *Biochem.*, 32 (6):1443–1453, 1993.

M. I. Zavodszky, M. Lei, M. F. Thorpe, A. R. Day, and L. A. Kuhn. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Prot: Struct. Func. Gen.*, 57:243–261, 2004.

X. Zhang, E. Y. T. Chien, M. J. Chalmers, B. D. Pascal, J. Gatchalian, R. C. Stevens, and P. R. Griffin. Dynamics of the $\beta_2$-adrenergic G-protein coupled receptor revealed by hydrogen-deuterium exchange. *Anal. Chem.*, 82:1100–1108, 2010.

Zhongwei Zhu, David I. Schuster, and Mark E. Tuckerman. Molecular dynamics study of the connection between flap closing and binding of Fullerene-Based inhibitors of the HIV-1 protease. *Biochemistry*, 42(5):1326–1333, 2003.

L. D. Zydowsky, F. A. Etzkorn, H. Y. Chang, S. B. Ferguson, L. A. Stolz, S. I. Ho, and C. T. Walsh. Active site mutants of human cyclophilin A separate peptidyl-prolyl isomerase activity from cyclosporin A binding and calcineurin inhibition. *Prot. Sci.*, 1:1092–1099, 1992.