

IMPERIAL COLLEGE LONDON
DEPARTMENT OF CHEMISTRY

**Exploring protein dynamics using graph theory and
single-molecule spectroscopy**

Robert L. Peach

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence, if not stated otherwise. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Proteins are the workhorses of any living system. They are essential for day-to-day life and yet we still do not fully understand how they work. Proteins are intricate structures composed of thousands of atoms that interact and fluctuate over multiple temporal and spatial scales. These motions are encoded in the peptide sequence; fast and local vibrations develop collectively into the motion of secondary structures and catalytic domains. Despite the importance of dynamical fluctuations in mediating function and other biomolecular processes such as allostery, there is inadequate evidence supporting the various theories and calculational approaches which attempt to determine the relationships between structure, dynamics and function.

Graph theory provides a way of capturing and representing the 3-dimensional atomistic physico-chemical details of a protein in a reduced form. In this thesis, we use Markov Stability, a novel graph-theoretic method for analysing the hierarchy of protein motions. We begin by producing an energy-weighted atomistic graph representation of a protein. We then use the transient behaviour of a random-walk to identify regions of atoms/residues that form communities. Using mutagenesis we are able to identify mutations that result in a large change to the community structure. We then use experimental methods (primarily single-molecule FRET) to explore the effect of these predicted mutations and validate the predictive qualities of Markov Stability.

In this thesis, we demonstrate a strong correlation between theoretical and experimental measures. We focus our study on Adenylate Kinase; a protein that balances molecular stability and fast large domain motions in the pursuit of cellular energy homeostasis. Given the structural data taken from the protein data bank we use the Markov stability graph theoretical approach to identify a number of biologically relevant community structures at different timescales. We are able to illustrate agreement between Markov Stability and the rate of motions of ADK subdomains. We then use the graph theoretical analysis to predict the effect of mutations to global dynamics and validate these predictions using single-molecule FRET; we show excellent agreement between the calculated properties of the graph theoretical analysis and the variation in FRET dynamics. Finally, we explore the relationship between the predictive score attributed to mutants by Markov Stability and the associated change in molecular stability.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisors Professor Keith Willison, Professor David Klug, Professor Sophia Yaliraki and Professor Mauricio Barahona for giving me the continued support, guidance and patience that was essential for the completion of my PhD. I have come to enjoy the nuances of the different scientific disciplines which wouldn't have been possible without such a broad range of backgrounds across my supervisors. I would also like to thank Dr. Liming Ying who provided the equipment and much needed expertise to help me perform the single-molecule experiments.

Secondly, thank you to my friends and family that have put up with me over the past few years, including Marion Peach, David Infield, Isabel Peach, Michael Peach, Julia Schmidt, Tim Turdukulov, William Hinds, Thomas Higginson and Jonathon Knight-Fallows to name a few. Their constant support has enabled me to continue my research even when my motivation and confidence was lacking. In particular, I'd like to thank my father, David Infield, for inspiring me to study science and to pursue a career which truly interests me.

Finally, I would like to thank my colleagues, past and present, who have made everyday as enjoyable as possible. John Simpson, Alex Bond, Paul Girvan, Raquel Fraccari, Max Hodges, Francesca Vianello, Heng Zhang, Mike Ray, Hugh Sowley, Sophie Sim, Alastair Magness, Stelios Chatzimichail, Ryan Lu, Thomas Mickleburgh, and many others.

Declaration of Originality

I declare herewith, that the work shown within this thesis is my own unless referenced otherwise, including the data, images, tables and text. Furthermore, I confirm that all data and findings have not been falsified or embellished and that the work has not been used for alternative exam processes.

Contents

List of Tables	17
List of Figures	19
1 Introduction	23
1.1 Motivation and Objectives	23
1.2 Thesis outline	24
1.3 Publications	26
2 Experimental and computational analysis of protein dynamics	27
2.1 Thermodynamic fluctuations and the Energy Landscape	28
2.2 High resolution structural approaches to infer motions	30
2.3 Experimental methods to explore protein dynamics	31
2.3.1 Atomic resolution methods	31
2.3.2 Low resolution and local-site methods	33
2.4 Theoretical and computational approaches to protein dynamics	36
2.4.1 Molecular Dynamics (MD)	36
2.4.2 Monte-Carlo	38
2.4.3 Graph theoretical approaches to protein dynamics	38
2.5 Summary	40
3 A short review of Adenylate Kinase	42
3.1 A history of ADK	42
3.2 A biochemical perspective of ADK	43
3.3 Functional mechanisms	43
3.4 Dynamic equilibrium of the open and closed conformations	45
3.5 Rate-limiting steps and the relationship with catalysis	45
3.6 Transitional coupling, intermediate states and pathways of the subdomains	46
3.7 Free energy landscape of ADK	48

4	Computational Methods	49
4.1	Graph theoretical approaches	49
4.1.1	Network Construction	49
4.1.2	Community detection in networks	52
4.1.3	Random walks on graphs	53
4.1.4	Markov stability of a graph partition	54
4.1.5	Partition optimisation with the Louvain algorithm	56
4.1.6	Robustness of a partition	58
4.2	Molecular Dynamics and Computational Mutation	59
4.2.1	Residue substitution	59
4.2.2	Energy Minimisation	59
4.2.3	Molecular dynamics simulations	60
5	Experimental Materials and Methods	62
5.1	Experimental Methods	62
5.1.1	DNA Techniques	62
5.1.2	Protein Techniques	63
5.1.3	Ultraviolet-Visible (UV-Vis) Spectroscopy	65
5.1.4	Cleaning and Decontamination	66
5.1.5	Confocal Microscopy	66
5.2	Fluorescence	69
5.2.1	Förster Resonance Energy Transfer	70
5.3	Circular Dichroism	72
5.4	Analytical Methods	73
5.4.1	Single-molecule fluorescence analytics	73
5.4.2	Melting temperatures curves from circular dichroism	79
5.5	Materials and Sample Preparation	82
5.5.1	<i>Aquifex</i> Adenylate Kinase construction	82
5.5.2	Buffers	85

6	Conformational dynamics of Aquifex ADK	86
6.1	Using Markov stability to identify biologically relevant communities	86
6.2	Construction of three FRET pair mutants	92
6.3	Single-molecule FRET details population equilibrium	94
6.4	Comparison of FCS dynamical rates of sub-domains agrees with Markov time . .	97
6.5	Substrate bound ADK	99
6.5.1	Markov stability analysis of inhibitor bound ADK	99
6.5.2	Relative FRET efficiencies of ADK in the presence of substrate/inhibitor	100
6.5.3	FCS indicates slower dynamical rates in the presence of substrate	103
6.6	Discussion	104
7	Computational mutagenesis of ADK reveals functionally important residues	106
7.1	Identification of important mutations using alanine scanning mutagenesis	106
7.1.1	Mutations that alter the robustness of partitions (VI)	109
7.1.2	Mutations that alter the make-up of the optimal partitions (CV)	111
7.1.3	Combining VI and CV to create a final mutant score	113
7.1.4	Robustness of the scoring method	118
7.2	Experimental investigation of various alanine mutants	120
7.2.1	Construction of ADK lid mutants according to prediction scores	121
7.2.2	Single-molecule FRET measurements of ADK mutants	122
7.2.3	Catalytic rates measured using kinase assay	124
7.2.4	Molecular dynamics probes the flexibility of mutants	125
7.3	A summary of the predicted mutations	127
7.3.1	Control mutations (E115A/Q149A)	127
7.3.2	P141A	128
7.3.3	I118A	128
7.3.4	K154A	129
7.3.5	D152A & D153A	129
7.4	Discussion	130

8	Validation of mutant predictions during the 3-way partition in ADK	132
8.1	Substitution of any residue type <i>in silico</i>	133
8.2	Building a relationship between theoretical scores and smFRET measurements .	133
8.2.1	A Markov stability analysis of computational residue substitutions	133
8.2.2	Measuring the significance of Asp152 mutations using mutant scoring . .	134
8.2.3	Conformational dynamics of ADK D152 mutants measured by smFRET .	136
8.2.4	High resolution Markov stability analysis of chosen mutants	136
8.2.5	FRET efficiency histograms of D152 mutants	139
8.2.6	Enthalpic changes drive the correlation between mutant score and the conformational equilibrium	140
8.2.7	The role of the conformational equilibrium in catalysis.	143
8.2.8	Comparing the $\overline{\text{VI}}$ ensemble of partitions uncovers communication between subdomains in D152 mutants	143
8.2.9	Multidimensional scaling reveals the differences between wildtype and mutant partitions	144
8.3	Arginine scan reveals reveals changes to long timescale dynamics	145
8.3.1	Markov stability analysis of arginine mutant ensemble	146
8.3.2	Single-molecule FRET of V164R	147
8.4	Discussion	149
9	Thermostability analysis of ADK using Markov Stability and Circular Dichroism	151
9.1	Markov Stability and circular dichroism melting temperature study of ADK mutants	152
9.1.1	Computational mutagenesis identifies important residues in the core domain of ADK	152
9.1.2	Construction of ADK core mutants according to prediction scores	153
9.1.3	Melting temperature measurements using circular dichroism	154
9.1.4	A detailed deconstruction of the mutant community structures	156
9.1.5	Combining Markov Stability predictions and melting temperatures	158
9.1.6	Robustness of scoring across timescales	159
9.2	Revealing residues essential for thermophilic properties	161
9.3	Fuzzy community structures identify molecular stability hotspots	164
9.4	Comparison of ADK temperature homologues	166
9.5	Discussion	168

10 Conclusions	171
10.1 Summary of main biological results	172
10.2 A philosophical outlook	173
10.3 Open questions and future work	174
10.3.1 Derivation of combining two metrics into a single score.	174
10.3.2 Mutating V164 to all residue types	175
10.3.3 Amide hydrogen-deuterium exchange to identify local unfolding events . .	175
10.3.4 Prediction and validation in multimer systems	176
10.3.5 Edge based Markov stability	176
10.3.6 Markov State Models using Markov Stability	177
Bibliography	177
Appendices	200
A Graph construction and parameters	201
A.1 Covalent bonds	201
A.2 Hydrogen bonds	201
A.3 Electrostatic interactions	202
A.4 Hydrophobic interactions	203
B DNA Sequencing results	204
B.1 Wildtype (FP1)	204
B.2 L199C & C52Y mutant (FP2)	205
B.3 L199C & C145V mutant (FP3)	206
C FCS measurements	207
D Gaussian Process Regression	208
D.1 Using GPR on a simple test set	208
D.2 Using GPR on the ensemble of mutated VI trajectories	211
E Kullback-Leibler divergence	212
F 2-way community structure of D152 mutations	214

G	Multidimensional scaling	215
G.1	Details of MDS	215
G.2	Using MDS on the ensemble of Louvain partitions	216
H	Conservation scores of Adenylate Kinase	217

Glossary

Community Variation	A direct measure of dissimilarity between the optimal community structures of the wildtype and mutant at time t . This is calculated using the variation of information between the wildtype and mutant structure.
Gaussian process regression	A non-parametric kernel-based probabilistic model that avoids making assumptions regarding the relationship between variables. In this thesis we use GPR to compare VI trajectories that are produced by mutant Markov Stability calculations.
Variation of information	A metric that measures the information-theoretic distance based on the similarity/dissimilarity of information between two community partitions. References to VI in this thesis are referring to the average distance between the ensemble of partitions that are optimised by the Louvain algorithm (usually 100 optimisations).

Acronyms

ADK	Adenylate Kinase.
ADP	Adenosine diphosphate.
AIC	Akaike information criterion.
AMP	Adenosine monophosphate.
Ap5A	Bis(5'-adenosyl) pentaphosphate.
APD	Avalanche photodiodes.
ATP	Adenosine triphosphate.
BIC	Bayesian information criterion.
CD	Circular dichroism.
DNA	Deoxyribonucleic acid.
E.Coli	Escherichia coli.
EM	Electromagnetic.
ENM	Elastic network models.
FCS	Fluorescence correlation spectroscopy.
FIRST	Flexible inclusions and rigid substructure topography.
FP	FRET pair.
FRET	Förster resonance energy transfer.
GPR	Gaussian process regression.
IR	Infra-red.
KL	Kullback-Leibler.

MC	Monte-Carlo.
MD	Molecular dynamics.
MDS	Multi-dimensional scaling.
NMA	Normal mode analysis.
NMR	Nuclear magnetic resonance.
PDB	Protein database.
QM	Quantum Mechanical.
RSS	Residual sum of squares.
SAXS	Small angle X-ray scattering.
smFRET	Single molecule Förster resonance energy transfer.
VI	Variation of information.
WAXS	Wide angle X-ray scattering.

List of Tables

5.1	The quantum yields of the dyes bound to ADK were determined relative to the quantum efficiencies of free dye by Henzler-Wildman et al. (2007b)	72
6.1	A comparison of the FRET distances (calculated using the FRET efficiency of each state) with the geometric distances between the labeled residues in the x-ray crystallographic structure.	97
7.1	A ranking of the top 10 most disruptive alanine mutations during each relevant timescale according to the VI criterion.	112
7.2	The top ten highest scoring lid subdomain mutants (as a combined metric of VI and CV) during each relevant timescale.	117
7.3	The high scoring arginines that cause a large effect on the community structure of the lid subdomain. Here we list their most important interactions with neighbouring residues.	118
9.1	Core mutations chosen for molecular stability experiments and their assigned Markov Stability score.	153
9.2	The melting temperature of the wildtype and mutant ADK structures are determined from the circular dichroism melting curves. Two thermodynamic models are used, the first (I) assumes a change in heat capacity between the folded and unfolded states, whilst the second model (II) introduces corrections for pre- and post- transition linear changes to ellipticity.	155
9.3	The Markov Stability mutant score for chosen mutations during the 4-way partition.	161
9.4	The conservation score (Consurf (Ashkenazy et al., 2010)) for each mutant and their respective melting temperature.	162

A.1	The energies used to weight covalent interactions in the graph representation of a protein. These values were taken from Huheey et al. (1993)	201
-----	---	-----

List of Figures

2.1	An illustration of the free energy landscape which defines the hierarchy of protein motions.	29
3.1	An illustration of Adenylate Kinase and its distinctive conformational motions. . .	44
4.1	Weighted graph construction from an experimental PDB structure.	51
4.2	Schematic of a Markov stability random-walk on a protein network.	53
4.3	The variation of information represented and its use to explore similarity in optimised community structures.	57
5.1	Illustration of how confocal microscopy allows in-focus light to pass to the detector. .	67
5.2	Overview of the instrumentation and the data reduction in the confocal single-molecule spectroscopy setup.	68
5.3	Fundamentals of fluorescence described with a Jablonski diagram	69
5.4	The basic principles of FRET as a consequence of the spectral overlap between fluorophores.	71
5.5	A theoretical schematic of circular dichroism in protein systems and the characteristic ellipticity spectrums produced in practical experiments.	73
5.6	Schematic diagram of fluorescence correlation spectroscopy.	76
5.7	DNA gel electrophoresis of Quickchange mutagenesis.	82
5.8	SDS page of purified and labelled proteins visualised using coomassie staining and typhoon fluorescence.	84

6.1	Markov stability results for the open (blue) and closed (orange) conformations of <i>Aquifex</i> ADK	88
6.2	Multiscale partitioning of <i>Aquifex</i> ADK as a function of Markov time.	90
6.3	A heatmap showing the similarities between community structures optimised at different temporal scales.	91
6.4	Crystallographic structure of the open and closed conformations with the different fluorophore positions labeled.	93
6.5	FRET efficiency histograms of ADK exhibiting the interdomain dynamics.	94
6.6	The AIC identifies the optimal number of distributions for the FP3 FRET efficiency histogram.	96
6.7	FCS autocorrelation measurements for the three ADK FRET pairs.	98
6.8	A Markov stability analysis of the inhibitor (Ap5a) bound conformation of Adenylate Kinase.	100
6.9	FRET histograms of the three ADK FRET pairs in the presence of various inhibitors and substrate.	102
7.1	Illustration of mimicking alanine mutations <i>in silico</i>	107
7.2	Markov stability results for the closed conformation of <i>Aquifex</i> ADK and the timescales of interest to look for perturbative mutations.	108
7.3	Identifying perturbative mutations to the community structure at every timescale.	110
7.4	Mutations that cause the largest effects on VI.	111
7.5	Identifying mutants that causes large changes to the community structure using CV.	113
7.6	An illustration of the VI and CV metrics in partition space.	114
7.7	Markov stability predicts mutations that perturb different timescales of dynamics.	115
7.8	High scoring mutants are defined at each timescale relative to the ensemble of mutations.	116
7.9	A comparison of the Kullback-Liebler method and RMSD method of combining VI and CV into a mutant score.	119

7.10	The mutations chosen for experimental measurement are highlighted on the protein structures.	121
7.11	Single-molecule FRET measurements of ADK mutants that were predicted across various timescales.	123
7.12	The activity of the ADK mutants as measured using a luciferase assay.	124
7.13	MD simulations initiated from the wildtype and alanine mutant structures. . . .	126
8.1	Peaks in $\overline{VI}$ exhibit the different effects of D152 mutants.	134
8.2	Scoring D152 mutations according to $\overline{VI}$ and CV.	135
8.3	Markov stability analysis of D152 mutants.	137
8.4	Single-molecule FRET efficiency histograms of the chosen set of D152 mutants. .	138
8.5	A correlation between the Markov stability mutant score across the 3-way partition and the shift in the conformational equilibrium as measured by smFRET. .	140
8.6	A correlation between the shift in the conformational equilibrium as measured by smFRET and $\overline{VI}$ and CV.	141
8.7	Exploring the relationship between the mutant score, the total bond enthalpy of mutants and the change to the conformational equilibrium.	142
8.8	Activity assays for the D152 ADK variants.	143
8.9	A Louvain optimisation partition matrix shows the ensemble of partitions that appear upon mutation and reduce the robustness of the community structure. . .	145
8.10	A multidimensional scaling (MDS) of ADK mutants in partition space.	146
8.11	A Markov stability arginine scan to identify perturbative mutations.	147
8.12	Experimental investigation of ADK V164R using smFRET and luciferase assays. .	148
9.1	Mutations with the largest effect on 6-way partition of ADK.	152
9.2	Melting curves of ADK mutants measured by circular dichroism.	154
9.3	An analysis of the Markov Stability community structure of each core subdomain mutant.	157
9.4	A plot of computational scores against melting temperatures for the ADK mutants.	159

9.5	Comparing the melting temperature of each mutant with the scores across all Markov timescales.	160
9.6	Using the mutant score and conservation scores to identify thermophilic adaption residues.	164
9.7	Using fuzzy communities to identify residues/regions that have no stable community structure.	166
9.8	Markov stability calculations of wildtype ADK species.	167
A.1	Variables used to calculate hydrogen bond strength.	202
B.1	DNA sequencing results of FP1 from GENEWIZ.	204
B.2	DNA sequencing results of FP2 from GENEWIZ.	205
B.3	DNA sequencing results of FP3 from GENEWIZ.	206
C.1	Determining the triplet relaxation rate and the diffusion rates for ADK.	207
D.1	A Gaussian process regression example with 25 training-set points.	209
F.1	The variation between the optimal partitions of the 2-way community structures upon mutation.	214

1. Introduction

1.1. Motivation and Objectives

Proteins are complex biomolecules constructed from thousands of atoms that play many critical roles in any living organism. They are most often found in cells and are responsible for structural, functional and regulatory processes. For example, Rubisco is a large heteromeric enzyme that assimilates carbon in photosynthetic organisms, whilst MscL is a mechanosensitive membrane channel protein that aids the transport of ions into and out of E.Coli.

Enzymes are a subset of proteins that perform their ‘job’ through interactions with other molecules, to do so they require conformational motions. For example, the subdomains of a enzymes must open to allow a substrate to enter the active site for catalysis, or the conformation of a enzyme must adapt such that its ‘new’ surface can bind to another protein. Several impressive studies have demonstrated the importance of protein dynamics in catalysis ([Boehr et al., 2006](#)) and the evolutionary conservation of dynamical features among protein families ([Gagné et al., 2012](#)). Deriving the relationship between structure, dynamics and function is thus of great importance in biology, it opens up an enormous range of opportunities to design and manipulate enzymes. Moreover, from a purely intellectual standpoint, understanding how these molecular machines work will provide an insight into the evolution of life as it exists.

The increased understanding between structure, dynamics and function offers opportunities for protein engineering. Proteins such as Rubisco can be engineered to improve catalytic efficiency such that photosynthesis can occur at an accelerated rate thus improving crop yields, others can be designed to act as therapeutics for illnesses or manipulated to act as sensors. However, without understanding the dynamics that links structural and functional features, we cannot build a framework for protein engineering.

At present we understand that fluctuations of protein structures are intrinsic thermodynamic phenomena, but how do the small ‘wiggings and jiggings’ of atoms build into large collective motions that provide functional properties? Despite recent experimental and computational advances, we are still unable to predict the effect of single mutations on dynamics let alone

design proteins from scratch. The primary contribution of this thesis is to develop, use and validate a graph theoretic approach to investigate the hierarchy of dynamics and ‘predict’ the effect of mutations on functional dynamics. Graph theory provides a common framework to represent and analyse complex biological, sociological and physical systems. Proteins have a natural representation as a network where atoms and their interactions can be modeled as nodes and edges respectively. Previous work has used the transient behaviour of a dynamical process on biomolecular systems to reveal hierarchical organisation of ATCase and Rubisco (Delmotte et al., 2011; Delmotte, 2014), and to reveal the allosteric nature of proteins (Amor et al., 2014). Here, we extend and adapt this work to predict and validate the conformational fluctuations of proteins.

1.2. Thesis outline

A range of theoretical, computational and experimental techniques have been developed in the past few decades to help us understand the dynamical nature of proteins. Chapter 2 takes a detailed look at the methodological advances that have provided key insights into the relationship between structure, dynamics and function of protein molecules. We take a look at the energy landscape and its importance in describing the hierarchy of protein motions, from the bond vibrations through to the collective domains motions. To put our work into context and to justify the techniques used throughout this thesis, we give an overview of theoretical and experimental techniques used to explore protein dynamics. We detail the successes of each method, their specific applicability (i.e. what can they tell us) and most importantly we discuss their limitations. Finally, we justify how the methods developed and used within this thesis improve upon and provide novel insights into the structure-dynamics-function relationship.

This thesis provides a novel theoretical investigation that is tested and validated using experimental methods, as such we have focussed on Adenylate Kinase (ADK), a well studied protein system to provide assurance and comparison of our methodologies. In Chapter 3, we provide a review of ADK, a protein that balances molecular stability and fast large domain motions in the pursuit of cellular energy homeostasis. We describe the history of ADK and provide a critical analysis of current studies that have contributed towards our understanding of the molecular mechanisms that underly its functional dynamics.

Chapters 4 and 5 provide a detailed deconstruction of the computational and experimental methodologies used throughout this thesis respectively. We describe how we construct an energy-weighted network representation of a protein structure and how this forms the basis of the network based analysis performed throughout this thesis. We then provide a detailed description of Markov stability, a multi-scale graph community detection algorithm that can identify meaningful communities across all timescales. Finally, we detail the experimental procedures and

analysis methods such that any studies can be accurately replicated.

In Chapter 6, we use Markov stability to reveal the hierarchical biomolecular structure of ADK and infer that these robust communities correspond to dynamically coupled regions. We first find that the subdomains appear as a robust community structure, suggesting that the energy-weighted network structure encodes the dynamical topology. A comparison of the Markov timescale at which the separate subdomains appeared as single communities suggested different rates of dynamical motion, this was confirmed using single-molecule FRET and fluorescence correlation spectroscopy which showed the AMP_{bind} subdomain to move faster than the lid subdomain. We also analysed the inhibitor bound conformation of ADK and showed that the community structure becomes less modular which agreed with inhibitor bound smFRET which showed a shift towards a single-state conformation.

To identify mutations that are critical to the dynamical nature of ADK we perform a full computational mutagenesis in Chapter 7. We use Markov stability to introduce a novel method to measure/rank mutations according to their perturbation to the community structure. As a consequence of the multiscale methodology, we can identify mutations according to the timescale at which they become significant. We use two metrics to predict the effect of mutations, we use them as separate measures and also combine them into a single mutant *score*. We then experimentally investigate the physical effect of a set of mutations on the dynamics and function of ADK.

In Chapter 8, we build upon the mutant scoring methodology and focus on the long timescale global motions that we can access using smFRET. As such we look at Asp152, the highest scoring residue when mutated within the 3-subdomain community structure. We computationally substitute Asp152 with all potential residues and measure the ranging effect each has on the community structure. We then corroborate the theoretical predictions with an experimental investigation of each mutation, we find that the higher ‘scoring’ substitutions result in a larger shift in the population equilibrium measured by smFRET. We also carry out a full arginine scan across the entire protein structure and validate the effect of another mutation (V164R) that also causes a shift in the population equilibrium.

To further understand and validate the ‘scoring’ methodology of mutations we look at the relationship between our mutant predictions and molecular stability. In Chapter 9, we switch our focus towards the core subdomain. We score mutations within the core subdomain and measure the melting temperature of a selection of high scoring (and control) mutations. We find a correlation between an increased effect of a mutation on the community structure and an increased reduction in the melting temperature. We then combine the mutant scores of each residue with their respective conservation score across all ADK species, using this we identify residues that have been evolutionarily introduced to provide thermophilic adaption.

1.3. Publications

The results presented in Chapters 7 and 8 are in a final draft for publication: Intended for PNAS ‘Controlling the conformational dynamics of Adenylate Kinase using graph theory’.

Chapter 9 is currently being drafted for publication. Intended for PloS ‘A graph theoretical approach to identifying critical residues for molecular stability’

2. Experimental and computational analysis of protein dynamics

“Everything that living things do can be understood in terms of the jiggings and wiggings of atoms”

Feynman, R

Life is governed by changes over time, and biologists explore this phenomena by watching, for example, neuron signalling through the body of a zebra fish, mice running around a maze, or the growth of bacteria on an agar plate. Studies like this aim to explain why and how biology ended up at the molecular level. The dawn of X-ray crystallography revolutionised our understanding of proteins, providing high resolution structures that enabled a surge of studies that sought to understand the structure-function relationship. Despite this, X-ray crystal structures can be misleading, they provide a snapshot of the protein in its ‘ground state’ suggesting that there is only one unique structure. Quite the contrary: proteins are composed of movable parts that work together to deliver a wide range of physiological functions. Ironically, to facilitate a higher resolution structure we must freeze the protein, obstructing our view of the dynamic nature of biology at the microscopic level.

Proteins are large complex macromolecules present in all living organisms, they are fascinating molecular machines that perform and aid sophisticated, and often convoluted, processes. In the cell, proteins are able to accommodate structural and dynamical alterations in response stimuli, including changes to the molecular environment (i.e. temperature or pH), binding to molecules or chemical modifications (i.e. phosphorylation).

From the perspective of a physicist, its justifiable to reject the colourful static crystal structures that we see everyday in theses and articles. At the molecular level, proteins are soft squishy structures that are governed by statistical thermodynamics, populating multiple conformational states that together constitute the native state. The free energy landscape provides a complete description of a protein, governing the relative probabilities of each state and the kinetic barriers

separating them.

Conformational plasticity is essential for protein function and in the past decade biologists have extended the structure-function paradigm into the fourth dimension to understand proteins in action. However, we can't simply watch the atoms of a protein moving around in real-time. Instead, we must employ sophisticated biophysical experimental and computational techniques which can probe the physical properties of a protein from which we can derive dynamics.

In this introduction, we first describe the physical principle governing protein conformational behaviour: the energy landscape. We then review high resolution structural approaches, including X-ray crystallography and NMR, from which we can infer dynamical information. We then look at experimental and computational methods that are used to characterise the ensemble of conformations and the motions that separate them. A particular focus is made on graph theoretical and single-molecule FRET approaches to protein dynamics, since they constitute a significant role within this thesis.

Our understanding of protein dynamics, and its intrinsic relationship with structural and functional properties, is still incomplete. In this thesis we have applied a novel graph-theoretic method to elucidate the spatial and temporal scales that cloud the dynamic characteristics of proteins. This chapter places our work in perspective by giving an overview of the physical principles, experimental, theoretical and computational methods that have contributed to our understanding of protein dynamics.

2.1. Thermodynamic fluctuations and the Energy Landscape

Proteins structures, like any molecular system, fluctuate as a consequence of thermodynamics. Fluctuations in energy and volume can reveal measurable macroscopic properties which help us build a quantitative picture. At thermodynamic equilibrium, the fluctuations are not only induced by intra-protein interactions but also by fluctuations of the energy of surrounding water molecules (Kim and Hirata, 2013).

Despite the free energy landscape being most familiar in the context of protein folding (Onuchic et al., 1995), where visualising the landscape as a funnel elegantly resolves Levinthal's paradox (Karplus, 1997), the concept had already been applied to explain the ensemble nature of proteins more than 40 years ago by Austin et al. (1975). Statistical information regarding the conformational ensemble and their thermodynamic fluctuations could be extracted from the free energy landscape. A statistical mechanics approach, including atomic level simulations, can probe the landscape. Using methods as simple as a 2D lattice model are capable of providing insight into the ensemble of transition states (Cieplak and Banavar, 2013). The free energy landscape isn't

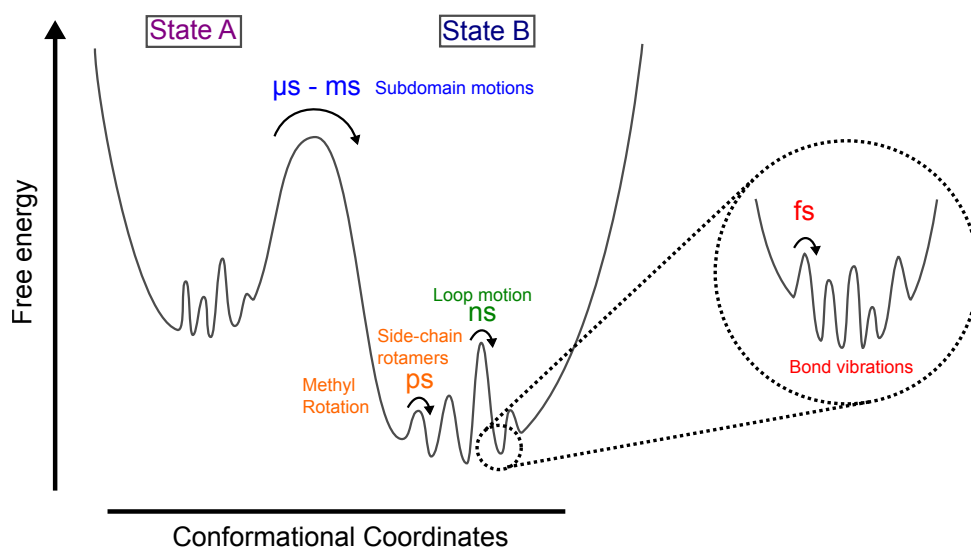


Figure 2.1: An illustration of the free energy landscape which defines the hierarchy of protein motions. The amplitude and timescale of protein fluctuations is defined by the energy landscape. This illustration shows a one-dimensional cross-section of the free energy landscape. A state is a minima in the energy landscape whilst the peaks correspond to energy barriers. Here we illustrate two major conformational states (A and B) which are each made up of an ensemble of minor conformational states. The small barriers describe faster fluctuations.

to be confused with the potential energy surface which is a mathematical function that gives the energy of a molecule as a function of its geometry.

The free energy landscape contains multiple minima, bridged by complex transition states, which define the amplitude and timescale of protein motions. A one-dimensional cross section (Figure 2.1) through the free energy landscape of a protein showcases the ensemble and hierarchical nature of protein dynamics. A state is defined by a minimum in the free energy surface, whilst transition states are described by the maxima between states. The ‘slow’ timescale dynamics associated with larger domain motions are separated by barriers of several kT which occur on the order of microseconds or longer at physiological temperatures. Typically, in a folded protein, there are only a few of these major states. However, these large amplitude motions are a consequence of collective smaller motions. The protein is not static within a major state, instead it samples the average structure at a faster timescale, exploring an ensemble of similar structures. We can denote the width and frustration of a unique global energy minimum as a measure of conformational entropy of the system.

The transition events associated with slow timescales are relatively rare, owing to the low probability of overcoming the kinetic barriers between states. However, dynamics at this timescale have been implicated in many biological processes, including signal transduction, enzyme catalysis and protein-protein interactions. In contrast to the slow timescales, a large ensemble of structurally similar sub-states are found within each major state. These energy barriers are often separated by less than $1 kT$ and result in small-amplitude picosecond to nanosecond motions.

These fast timescale motions can be broken further into fluctuations of groups of atoms (such as loop motions), local atomic fluctuations (such as sidechain rotations) and bond vibrations. In fact, the hierarchy of timescales associated with protein dynamics becomes a more continuous distribution where energy barriers of any size and roughness are possible. Naturally, the amino-acid sequence encodes the features of atomic motion, where the extent of constraint placed upon atoms is a function of their spatial position within the protein.

2.2. High resolution structural approaches to infer motions

High-resolution biomolecular structure determination techniques indirectly provide information regarding dynamical properties of proteins. For example, parts of the protein missing from a PDB structure file are often a consequence of flexible regions that cannot be fitted to an electron density map. In a similar manner, residues that adopt different conformations can appear with partial occupancy in crystal structures. A commonly used metric available from the Protein Data Bank is the X-ray B-factor which provides a rough measure of flexibility. The B-factor is calculated from the mean fluctuation of a residue from its average position, and is a useful method to distinguish between rigid and flexible regions. Whilst a useful measure, any X-ray derived descriptor should be taken with a pinch-of-salt; the B-factors generally only account for harmonic motions systematically underestimating the magnitude of protein fluctuations (Kuzmanic et al., 2014) and when a residue is spatially undefined the B-factor will be infinitely large.

More recently, a number of studies have illustrated the potential to mix X-ray structures with co-evolutionary signals derived from sequence analysis in an effort to explore the long-timescale dynamic conformational space accessible to proteins. This idea has been successful in determining alternate conformations of protein when combined with simple simulation engines (Hopf et al., 2012; Morcos et al., 2013).

A richer source of dynamical information is NMR spectroscopic high resolution structural analysis (Fenwick et al., 2011b; Mittermaier and Kay, 2006). An NMR experiment can provide information on a variety of ‘ensemble properties’ including the Nuclear Overhauser Effects (NOE), three bond scalar couplings (^3J), chemical shifts (CS), residual dipolar coupling (RDC), transhydrogen bond scalar couplings (^3HJ) and paramagnetic relaxation enhancement (PRE) (Lindorff-Larsen et al., 2005; Fenwick et al., 2011a). These parameters can be used to evaluate the ensemble of conformations sampled by a protein structure. NMR experiments can also provide information on dynamical motions across multiple temporal scales which we will see in section 2.3.1.

The final, and possibly most ‘simple’ method to explore flexibility, is a comparison of the protein structures deposited in the Protein Data Bank (PDB (Berman et al., 2000)). Whilst potentially

a consequence of crystallisation conditions, an analysis of the diversity between crystal structures can yield important dynamical information. In particular, NMR produces multiple structures that are solutions to the NMR derived constraints e.g. NOE and ^3J 's. Whilst the structural diversity between NMR structures could be attributed to technical noise, studies have shown good agreement between dynamic patterns derived from the ensemble of structures and other measures of flexibility (Orellana et al., 2010). In addition, we can explore the structural diversity of a protein through multiple static X-ray structures of the same protein crystallised under different conditions (i.e. pH, mutants, apo or holo states). Although it is important to note that differences between structures can simply be artefacts of crystallisation or other spurious reasons. Moreover, the reaction coordinates associated with the transition between x-ray structures may not follow a direct path. Despite this, several studies have used static X-ray structures to build a pseudo-trajectory between conformational states, agreeing with molecular dynamics simulations (Velázquez-Muriel et al., 2009; Leo-Macias et al., 2005).

2.3. Experimental methods to explore protein dynamics

Various experimental ensemble approaches have been developed to study protein conformations and the motions that connect them. Major experimental methods include, but are not limited to, single molecule spectroscopy, small-angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR) and cryo-electron microscopy. These methods are often combined with significant data analysis techniques that use the experimental data as structural constraints. In this section we choose a selection of widely used experimental methods and we discuss both their successes, and more importantly, their limitations. We split the experimental approaches into high-resolution (i.e. atomistic information regarding motions) and, low-resolution and local-site methods.

2.3.1. Atomic resolution methods

Nuclear magnetic resonance (NMR)

The hierarchical nature of dynamics in proteins makes it difficult to probe the mechanisms of catalysis, signaling and allostery to name but a few. NMR is an exquisite technique for measuring the dynamical properties of proteins at high resolution and across multiple temporal scales. NMR exploits the magnetic properties of nuclei (Rule and Hitchens, 2006). Spin 1/2 isotopes, which exhibit magnetic dipole moments, are introduced at desired locations throughout the protein. A strong magnetic field is applied to align the NMR-active isotopes. After which, a weaker magnetic field is temporarily applied in a perpendicular manner to rotate the bulk magnetic moment into the transverse plane. In a similar fashion to the precession of a spinning top's angular momentum, the bulk moment will undergo Larmor precession about the initial

magnetic field. The nuclei will precess at rates that are consequential of their respective local magnetic fields.

NMR provides unmatched coverage across almost every atom in the protein, allowing complete access to structure and dynamics (Ohki and Kainosho, 2008). In contrast to the solid-state required for X-ray crystallography, proteins can be studied in a solution by NMR allowing native conditions and binding studies (Felli and Brutscher, 2009). To date NMR studies have revealed protein folding pathways at atomic resolution (Neudecker et al., 2009), catalysis (Boehr et al., 2006) and allostery (Kern and Zuiderweg, 2003). NMR has even shown that dynamics of proteins can enable them to perform multiple distinct catalytic functions (Tawfik and Khersonsky, 2010).

Whilst able to access multiple temporal scales, NMR is limited by the poor sensitivity of signal detection as a result of the low energy magnetic spin transitions. As a consequence, NMR experiments can take a long time, require a lot of material and also are limited as to the size of protein they can study (typically $< 50 - 100$ kDa) (Ruschak and Kay, 2010). Additionally, NMR technology is expensive due to the high magnetic field instrumentation and isotopically labeled samples. Another notable limitation is spectral crowding, which makes analysis of NMR spectra statistically difficult. However, this is also a result of its ability to detect tens to hundreds of probes per protein.

Cryo-EM

Whilst photons (X-ray) and magnetic spin transitions (NMR) are the most widely used techniques for atomistic structural explorations, we can also use electrons to look at protein structures. Electrons offer many advantages including their ability to be manipulated and accelerated by strong electric fields over very short distances and they can be focussed easily with an electromagnetic lens. Cryo-EM can produce atomistic 2D images by bombarding a protein with electrons, the relative orientations of each 2D image is defined *a priori* using a computer. A 3D model of the protein structure can be produced from a large collection of 2D images (Frank et al., 1981).

As described throughout this introduction, proteins adopt an ensemble of conformations in solution. As such, the samples prepared for cryo-EM will often contain a variety of structures. This initially created a problem; the different protein structures needed to be separated using maximum-likelihood classification algorithms (Scheres et al., 2007). What was initially considered an impediment for cryo-EM quickly became an opportunity; the presence of multiple structures provides a valuable insight into protein dynamics and the underlying ensemble. Cryo-EM has unveiled molecular mechanisms on DNA replication (Fernandez-Leiro et al., 2015), transcription (Hoffmann et al., 2015) and CRISPR systems (Jiang et al., 2016).

However, the mechanistic motions are still difficult to capture and significant computational analysis is required to separate the conformational states. Cryo-EM also breaks down for smaller proteins where the details of a small protein can often be lost. Moreover, storing and processing large cryo-EM data volumes is costly (Fernandez-Leiro and Scheres, 2016).

Small angle X-ray scattering (SAXS)

Small angle X-ray scattering is another structural characterisation methodology. It is a rapid method that yields time-resolved information regarding kinetic processes of proteins. A protein solution is illuminated by X-ray photons and the scattered photons are recorded by an X-ray detector. The scattering pattern is related to the shape and size of the particles from which we can infer dynamical properties.

Recent advances in SAXS instrumentation has explained the influence of disordered and flexible regions of mitochondrial glutaminase on enzymatic activity and the oligomeric state (Møller et al., 2013). Additionally, SAXS has been useful in identifying the effect of point mutations on structure and dynamics (Kim et al., 2011). In particular SAXS has had significant success in the exploration and detection of structural transitions in intrinsically disordered proteins (IDPs). Using SAXS, Uversky et al. (1999) found the decreased compactness of two IDPs, human NHE1 and ACTR, increased with temperature. However, despite numerous advances and victories, SAXS is unable to characterise the timescales of interconversion.

Wide angle X-ray scattering (WAXS)

An extension of the small angle is the wide angle X-ray scattering, this takes advantage of the scattering bands outside of the first scattering peak. WAXS is able to provide more detailed information about the fold of α -helices and β -sheets. Successful studies have used WAXS to explore the refolding of cytochrome *c* and to observe the tertiary dynamical relaxations of myoglobin (Cammarata et al., 2008). The time resolution of WAXS is limited to about 100 μ s (Menk et al., 2000) making it difficult to measure mid-to-slow transition rates.

2.3.2. Low resolution and local-site methods

Circular Dichroism (CD)

Circular dichroism (CD) takes advantage of the differential absorption of left and right circularly polarised light by the chiral structures or asymmetry of a protein structure. Whilst CD is a structural method, we can infer information regarding the conformational flexibility of a protein

structure. A study [Celej et al. \(2003\)](#) was able to explore the effect of ligand binding on protein flexibility of BSA. Despite this, detailed mechanistic information is unavailable using CD.

Infra-Red (IR)

IR spectroscopy provides a bond specific approach to study proteins. An IR label is incorporated into a protein, the label is sensitive to its surroundings such as electric field or interactions. Changes to the local environment of the label can produce an inhomogeneous line width reflecting the distribution of frequencies. Using a time-dependent evolution of the vibrational frequency, spectral changes can be associated with structural changes. In particular, ultra-fast IR methods can resolve fast timescale dynamics associated with protein function.

IR studies of the amide backbone vibrations have yielded many protein dynamics studies including local hydration dynamics in beta amyloid fibrils ([Kim et al., 2009](#)) and transitional motions of a synthetic alpha-helical peptide ([Huang et al., 2001](#)). However, overlapping of bands in the IR spectra can complicate interpretation of protein dynamics and the choice of IR label is difficult due to environmental, vibrational coupling and spectral overlap effects.

Raman Spectroscopy

Raman spectroscopy is similar to IR, it is a form of vibrational spectroscopy that is based on inelastic scattering of light. Raman spectroscopy can be used in a steady-state or time-resolved manner to probe the vibrational spectra of chromophores. Raman spectroscopy has proved itself as a sensitive, selective and practical method for studying the structure of biological molecules, however, its ability to provide mechanistic insights into protein dynamics is limited by the instrumentation available. Despite this, a study by [Balakrishnan et al. \(2004\)](#) used UVR to investigate dynamics of heme proteins and [Sanchez et al. \(2011\)](#) was able to gain insight into the lipid interaction dynamics of outer membrane protein A.

Single-molecule Fluorescence

Single-molecule fluorescence methods have seen a sort of renaissance in recent years. They offer the ability to observe rare events, heterogeneity, transient intermediate states and all whilst accessing multiple timescales. The key advantage of single-molecule fluorescence experiments is the potential to avoid ensemble averaging, instead information is extracted from proteins one-by-one.

At a very basic level, fluorescence studies employ fluorophores that can emit photons after illumination (or excitation) from a slightly higher energy photon. These fluorophores can

be bound to proteins to derive site specific information and dynamical motions. The emitted photons can be collected using microscopy techniques and counted using cameras or single photon detectors.

There are a variety of fluorescence techniques including fluorescence lifetime measurements, fluorescence polarisation anisotropy and FRET. Fluorescence lifetime measurements take advantage of changes to the time that a fluorophore spends in the excited state as a consequence of its surroundings, it can offer dynamical information that occurs on the microsecond timescale. A study by [Otosu et al. \(2015\)](#) used 2-dimensional fluorescence lifetime correlation spectroscopy to clarify the microsecond conformational dynamics of cytochrome *c* and revealed its heterogeneous nature. Time-resolved fluorescence polarisation anisotropy provide details regarding mobility and dynamics of a fluorophore and has successfully studied the short timescale (nanosecond) motions of tryptophan residues ([Munro et al., 1979](#)) and to observe the loop motion of rhodopsin and bacteriorhodopsin ([Alexiev et al., 2003](#)).

FRET

Single-molecule FRET (smFRET) is quite likely the most popular fluorescence technique for studying protein dynamics. The resonant energy transfer between two fluorophores acts as a nanometer ruler from which we can infer the distances between parts of a protein and the rate of their relative motions (a detailed description of the technique is available in section 6.3). Continuous observation of multiple protein instances allows us to statistically build a histogram that describes the distance populations.

FRET was first established as a method for investigating protein dynamics in 1967 by [Stryer \(1978\)](#), however, the first smFRET measurements weren't established until 30 years later ([Ha et al., 1996](#)). Dynamical information can be extracted from the broadening and exchange of subpopulations in FRET histograms ([Schuler et al., 2002](#)) or even directly from the fluorescence trajectories over time of proteins immobilised on a surface ([Chung et al., 2013](#)). For a more in-depth description of technical and conceptual details of smFRET see [Enderlein and Keller \(2002\)](#), or for a quick primer in the subject, [Schuler \(2013\)](#) offers an simple yet comprehensive introduction to the topic.

Single-molecule FRET has seen a number of landmark experiments that have improved our understanding of protein dynamics. A smFRET study by [Seo et al. \(2014\)](#) revealed that the dynamical rates dictated the ligand dissociation rates in the maltose-binding proteins, despite not affecting the ligand binding step ([Kim et al., 2013](#)). The open and closed conformations of Adenylate Kinase were found to be sampled equally at equilibrium by [Henzler-Wildman et al. \(2007b\)](#) and the respective population of the open and closed states would shift towards closed upon ligand binding.

In more recent years, many variations of smFRET have been introduced. Multicolour-FRET employs multiple fluorophores which, in principle, allows multiple distances to be measured simultaneously (Clamme, 2005; Lee et al., 2010). Pulsed laser excitation can allow us to distinguish proteins that are labelled with only donor fluorophores which can improve low transfer efficiency studies (Milles et al., 2012) and also allows measurement of fluorescence lifetimes.

Despite smFRET being a local-site measure of dynamics and its inability to access short timescales, it is clear that smFRET allows us to access the long timescale motions that contribute to functional activity.

2.4. Theoretical and computational approaches to protein dynamics

2.4.1. Molecular Dynamics (MD)

Computation provides an unbeatable edge: it can completely describe protein dynamics within the realms of its initial structural representation and the equations we provide it. We can quantify the exact position and velocity of each atom at any instance in time. Experimental methods are unable to achieve an atomic resolution description of the energy landscape, in other words they are unable to follow the protein between energy wells owing to the low probability and fast transient timescales associated with each sub-state. Significant progress has been made regarding both the descriptive forcefields and also the computational efficiency of molecular dynamics simulations. In particular, supercomputers are now capable of producing 10-20 μ s of simulated protein data per day (Anton Supercomputer, unpublished data, Biophysical Meeting Berlin, August 2017) on moderately sized proteins.

There are a wide range of Hamiltonians (i.e. forcefields) used to represent molecular interactions. We will lightly touch upon a handful of these in the following sections, however, irrespective of the Hamiltonian we must employ a sampling method that is able to model a transition pathway and/or generate the ensemble of protein structures given the physiological conditions. Molecular dynamics achieves this by integrating the equations of motion. We can employ Newtonian physics which offers a classical Hamiltonian where integration of the equations of motion is simple, or we can include the electron degrees of freedom which requires consideration of quantum physics and makes integration more complex. This section gives a short overview of molecular dynamic sampling methods and their achievements.

Classical MD

Classical MD is probably the better known and most widespread method for studying protein dynamics. Both the protein and solvent are treated at an atomic level of resolution using classical

forcefields that have been parameterised against quantum mechanical calculations. The forcefield energy can be differentiated to extract the respective forces on each atom, the corresponding accelerations can then be integrated numerically (usually on the order of a femtosecond) to form new positions and velocities. The trajectory of the protein contains all the necessary information regarding dynamics and flexibility given the simulation conditions.

The first MD simulations of proteins were performed in 70's (McCammon et al., 1977; Levitt and Warshel, 1975) and refined throughout the 80's. More recently Shaw et al. (2010) managed to reach millisecond timescales for classical MD simulations of folded BPTI, they revealed a small number of distinct conformational states and found that the local relaxations within each state was a factor of 1000 faster than the reversible interconversion rates. Lindorff-Larsen et al. (2016) was able to show that the various conformations visited by free ubiquitin was similar to the ensemble of crystal structures of ubiquitin in complex with other proteins. They also successfully folded a number of small proteins in the sub-millisecond timescale in an unbiased simulation (Lindorff-Larsen et al., 2016). Classical MD has also been used for more complex processes including ligand binding (Pan et al., 2013), allosteric transitions (Arkhipov et al., 2013) and the mechanisms underlying membrane ion channels (Jensen et al., 2012; Ostmeier et al., 2013).

Whilst increasingly fast, the slow and large amplitude motions are still out of reach for MD simulations on the majority of computers. Therefore, in order to gain mechanistic insight into the dynamics of long timescale motions, we must resort to a time independent approach.

Quantum molecular dynamics

Quantum molecular dynamics can be split into two major frameworks:

1. **Born Oppenheimer molecular dynamics (BOMD).** This method considers each electron at equilibrium for each nuclei motion which simplifies the integration of electron degrees of freedom.
2. **Carr-Parrinello molecular dynamics (CPMD).** Likely the most popular approach of the two, it uses DFT planewaves to describe the electron distribution and explicitly considers every electron degree of freedom.

These methods are very powerful and the application to protein systems is increasing, in particular hybridising QM methods with classical MD methods has been a highly fruitful area due to the increased computation time with the associated accuracy of modelling ligand binding and bond breakage accurately. In particular QM/MD methods have revealed transition oxacarbenium

ion-like transition states in 1,3-1,4- β -glucanase (Biarnés et al., 2011) and lipopolysaccharyl- α -1,4-galactosyltransferase (Gómez et al., 2012). In addition QM/MD predicted an intermediate formation of a covalent interaction in *Trypanosoma cruzi* trans-sialidase (Pierdominici-Sottile et al., 2011) and revealed the reaction mechanism of HIV-1 protease (Piana et al., 2004).

However, QM molecular dynamics (including hybrids with classical MD) are computationally expensive relative to classical MD simulations. This makes them unfeasible for modelling the slow timescale fluctuations of proteins associated with function.

Enhanced sampling methods

Molecular dynamics is limited by its computational inefficiency, particularly through its inability to efficiently explore rough energy landscapes. To overcome this, various enhanced sampling methods have been introduced. One particular method is replica exchange simulations (RExMD) that uses multiple trajectories at various temperatures combined with a Metropolis algorithm to accept or reject configurations (Sugita and Okamoto, 1999).

2.4.2. Monte-Carlo

Monte-carlo (MC) methods are a set of algorithms that can generate an ensemble of conformational states through repeated sampling. The most common Monte Carlo method is the Metropolis implementation named after Nicholas Metropolis in 1953, a stochastic approach which generates random movements on a given structure and then accepts or rejects the new structure according to the relative energy of the seed structure. Given an infinite set of attempts, we are guaranteed an Boltzmann ensemble, however in reality the method is only efficient when provided with suitable sampling variables. For example, sampling side chain motions (Jorgensen and Tirado-Rives, 2005) or fluctuations of significantly coarse grained structures are efficient (Dill et al., 1995), whilst protein motion in cartesian coordinates is met with a high rejection rate at each iteration. Success has been found with Monte-Carlo methods to refine binding modes for ultra-affinity drugs (Lee et al., 2013). MC algorithms have also been combined with elastic network models and normal mode analysis (see section 2.4.3) to sample dynamical motions along the normal modes (Pesonen et al., 2012).

2.4.3. Graph theoretical approaches to protein dynamics

In mathematics, graph theory is the study of a network of points (referred to as *nodes*) connected by lines (referred to as *edges*). It provides an abstract representation of a set of objects with some form of connection between them. Any system that can be modelled in this way (e.g.

airports and the flights that connect them, protein interaction pathways, electrical power grids) is subject to a universal framework of graph-theoretical tools. Protein structures can also be represented as a network, atoms or residues become nodes whilst the interactions such as forces or bond strengths form the edges of the graph. The network paradigm has increasingly been used to study complex biomolecular systems; the graph theoretical approach has revealed protein-protein interaction mechanisms (del Sol et al., 2005), key residues in protein folding (Vendruscolo et al., 2002), allosteric pathways (del Sol et al., 2006) and protein dynamics (Atilgan et al., 2004).

Topological graphs reveal dynamic information

Protein structure networks represent segments of the protein while their weighted edges account for some physical distance between these segments. The network can be atomistic (each atom is represented by a node) or coarse grained (each residue is represented by a node). The atomistic networks usually only include the backbone atoms (α C β C). Whilst the edges can be weighted according to a force or interaction energy between the nodes, unweighted protein structures are quite common and are derived from distance thresholds. It is important to note that protein structure networks should not be confused with protein-protein interaction networks that are often dubbed ‘protein networks’.

Initial studies looked at the topology of the network to ascertain structural information from which dynamical information could be inferred. Atilgan et al. (2004) used a coarse grained network construction on 595 proteins to show that the average shortest path lengths calculated from the network structure are highly correlated with residue fluctuation. This was reflected in the restricted motions of central amino acids. However, this approach only includes side chain interactions and neglects backbone interactions.

The degree of nodes can be used to ascertain protein properties. Protein structure graphs have been successful in identifying the key role of hydrophobic interactions in the core-structure of proteins (Aftabuddin and Kundu, 2007) and in determining folding probability (Dokholyan et al., 2002). Protein structure networks often contain modules or clusters (i.e. communities of atoms or residues with a relationship such as intra-modular density). Kannan and Vishveshwara (1999) used graph spectral theory to accurately reveal the domain structure of several proteins.

Normal mode analysis (NMA)

Whilst the topology of a protein network structure can reveal information regarding dynamics, the structure is inherently static. Normal mode analysis provides a simple and rapid method for calculating vibrational modes and protein flexibility. The atoms are modelled as point masses which are connected by springs that represent the inter-atomic potentials. A well known type

of NMA is the elastic network model, in which the springs are all equally strong and are only included based on a distance cut-off.

Comparisons of low frequency normal modes and the large amplitude conformational motions observed in molecular dynamics indicate strong similarities (Amadei et al., 1993; Hayward et al., 1997). In particular the first modes were found to be overwhelmingly closure axes (Hayward et al., 1997). Bahar and Jernigan (1998) used ENMs to successfully model the dynamics of transfer RNAs both in the apo and bound state with cognate synthetase. The three lowest-frequency modes was used by Valadié et al. (2003) to analyse the first step of the gating mechanism in MscL.

NMA provides a atomistic description of the slow large-amplitude motions that are currently out-of-reach for molecular dynamics simulations. In the case of ENMs, the calculations are usually constrained to only backbone atoms and thus can ignore important inter-subdomain interactions (Alexandrov, 2005) making ENMs unusable for identifying key dynamical residues and thus limits protein engineering. However, in general the full atomistic description of a protein is used in NMA. Despite this, the NMA approach is limited as a result of the harmonic approximation, limited information regarding multiple minima and energy barriers in the energy landscape, and the neglect of solvents (Elber and Karplus, 1987).

2.5. Summary

Developments in experimental (primarily NMR, but also recent advances in smFRET and cryo-EM) and computational (MD) techniques have provided crucial insights into protein dynamics, strengthening our understanding of the relationship between structure, function and motions. The wide range of techniques are making it possible to characterise the entire ballet of ‘jiggings and wiggings’ of proteins. We are finally able to start mapping the energy landscapes of proteins during catalysis (Kerns et al., 2015).

Whilst molecular dynamics has advanced significantly in recent years, it is still wholly inefficient to sample the long timescale motions without biasing or directing the simulation (especially for large proteins). Indeed, performing any perturbation to a protein structure would require an equally long simulation to analyse its effect which is out of the question with current technology available to the average researcher.

An obvious question to ask is whether we can associate the short timescale motions with the long timescale motions that are associated with functional activity in an efficient manner. Can we access the hierarchy of timescales associated with dynamics? And can we measure the effect of perturbations to these proteins (such as mutation) on the long timescale motions? The graph theoretic methods offer a natural and computationally efficient way of accessing global motions

in proteins. Significant work has been completed on identifying the large amplitude motions using graphs and spring models.

The work in this thesis builds upon and uses a number of the approaches discussed in this chapter. We begin with building a fully atomistic graph representation of a protein structure that includes all physio-chemical interactions (covalent and non-covalent) and their respective weights according to the strength of each interaction. Using a multiscale dynamics-based community detection algorithm we are able to explore the hierarchical structure of a protein. Since our graph theoretic approach offers information on all timescales we are able to associate that with the corresponding timescales of motion accessible by smFRET. The computational efficiency of our approach allows us to obtain mutagenesis information which is considered too costly for MD simulations, whilst retaining atomistic information that is lost by coarse-grained non-energy-weighted network approaches. In the following chapter, we describe the protein system (Adenylate Kinase) which is studied throughout this thesis.

3. A short review of Adenylate Kinase

This thesis primarily explores Adenylate Kinase, an enzyme that has been explored extensively by both experimental and theoretical groups to understand protein dynamics. We chose this protein such that we could compare our experimental and theoretical approaches with available literature. In particular we are exploring *Aquifex* Adenylate Kinase which is derived from the hyperthermophile *Aquifex* Aeolicus. We chose *Aquifex* Adenylate Kinase because its dynamical fluctuations are slower at room temperature relative to its mesophilic analogues. The slower fluctuations allow us to more accurately measure the FRET dynamics. Additionally, the results of *Aquifex* Adenylate Kinase FRET experiments are available in literature, providing us with a framework with which to compare our measurements.

3.1. A history of ADK

The human body is dependent upon metabolic signals to regulate and integrate many vital functions such as energy homeostasis, hormonal status and brain performance. Metabolic monitors are essential to measure the cellular energy state and respond to imbalances through generation and delivery of signaling molecules to metabolic sensors that can produce a regulatory response. This thesis focuses on Adenylate Kinase, a critical enzyme for metabolic monitoring and the systemic integration of various signalling pathways. It ensures cellular energy homeostasis and provides applicable responses to environmental, functional and stress based challenges. We look specifically at Adenylate Kinase sourced from the hyperthermophile *Aquifex* Aeolicus due to its slower dynamical rates at ambient temperatures.

Adenylate Kinase has been widely studied, from its gene polymorphism, expression and intracellular distribution on cellular scales through to its inherent structure-function relationship at atomistic scales. This broad range of studies make ADK a good testbed for developing new theoretical and experimental hypotheses and methodologies. In this chapter we review the current understanding of dynamical mechanisms and their relationship with catalysis in ADK.

3.2. A biochemical perspective of ADK

The structure of ADK is well characterised with more than 100 structures of various ADK species in a range of open to closed conformations. Despite the myriad of structures, only four are available in the apo state (PDB ID: 4AKE, 2RH5, 3UMF and 3GMT) which takes the form of the open conformation.

ADK regulates the ratio of AMP and ADP/ATP by catalysing the reversible nucleotide phosphoryl exchange reaction of



Kerns et al. (2015) quoted ADK to accelerate the *p*-transfer rate by 14 orders of magnitude relative to the uncatalysed reaction. Adenylate Kinase has three sub-domains: the core, lid and AMP_{bind}, with two distinct binding sites. The ATP binding site is sandwiched between the lid and core subdomains, whilst AMP binds between the AMP_{bind} and core. During binding and catalysis the core domain remains stable whilst the lid and AMP_{bind} both undergo large conformational transitions as revealed by crystallographic structures (Müller and Schulz, 1992; Müller et al., 1996; Berry et al., 1994; Bilderback et al., 1996) and FRET (Sinev et al., 1996). After *p*-transfer the lid and AMP_{bind} subdomains open and release the products. The role of magnesium in catalysis is crucial for ADKs high catalytic rate (Bellinzoni et al., 2006; Krishnamurthy et al., 2005); it both accelerates lid opening and reduces nonproductive active site fluctuations (Kerns et al., 2015).

3.3. Functional mechanisms

The functional mechanism of ADK is intrinsically linked with the large-scale conformational motions of its sub-domains. However, these large scale motions are the end-product of a hierarchy of folding (Rhoades et al., 2003; Pirchi et al., 2011) and dynamical (Henzler-Wildman et al., 2007a) timescales. Elucidating the underlying mechanistic details of ADK dynamical fluctuations is key to guiding future rational drug and medicine development. Moreover, understanding the structure-function relationship will help build a platform for both exploring other protein systems and enabling accurate engineering of enzymes. There are currently two contradictory mechanistic models that govern the conformational transitions (Whitford et al., 2008):

- (i) **Induced fit.** Analysis of crystallographic structures posits the idea of locally rigid segments spatially translating around hinges. The lid-connector helices rotate around the core whilst the floppy lid insert rotates around the lid connector helices.

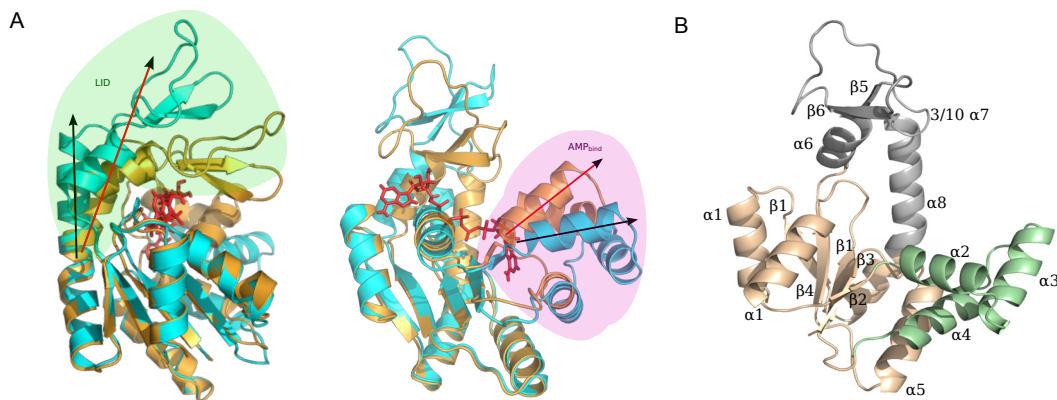


Figure 3.1: An illustration of Adenylate Kinase and its distinctive conformational motions. (a) An overlay of the open (PDB ID: 2RH5) and closed (PDB ID: 2RGX) conformations of *Aquifex* Adenylate Kinase. The two different perspectives highlight the large subdomain motions of the lid and AMP_{bind}. (b) The α -helix and β -sheet numbering for *Aquifex* ADK used throughout this thesis.

- (ii) **Cracking model.** Potential energy is built up through strain in the lid connector helices. Once high enough, a kinetic activation barrier is overcome and this energy is released through a local unfolding event during the transition state. The fall down the other side of the activation barrier leads to local refolding as the protein reaches the thermodynamic minimum corresponding to the closed state (Whitford et al., 2007; Maragakis and Karplus, 2005).

A number of experimental and theoretical approaches have been used to study the molecular mechanism of ADK and the general consensus is pushing towards a ‘cracking’ mechanism. Coarse-grained models were first to propose the ‘cracking’ mechanism and the respective ‘cracking’ regions (Whitford et al., 2007; Maragakis and Karplus, 2005; Miyashita et al., 2003, 2005). NMR and thermal denaturation experiments provided further evidence for the ‘cracking’ model; Olsson and Wolf-Watz (2010) showed that the lid connector helices unfolded and refolded during the conformational transition. This was further supported by Rundqvist et al. (2009) who induced locally unfolded regions of the lid using site-directed mutagenesis. Short MD simulations showed local unfolding of an important hinge during the closed to open transition (Brokaw and Chu, 2010). A multi-time correlation function of MD simulations and single-molecule FRET experiments revealed locally unfolded structures of the lid subdomain (Ono et al., 2015). A 1000 ns molecular dynamics and metadynamics simulation by Li et al. (2015) was able to map the transition pathways and intermediate states of ADK during a full conformational transition. They clearly identified the lid connector helices as cracking regions and that multiple transition states facilitate the rapid conformational change. However, this long MD simulation did not reach the fully closed state and therefore the complete mechanism is still elusive.

The induced fit and cracking models are pushed in ADK literature, however, an alternative to both of these is the conformational capture model. ADK could transiently be accessing each state and if the ligand recognises a conformation a specific conformation then the occupancy of that state will progressively increase. Indeed, these models aren't mutually exclusive, for example the cracking model may describe the transition between a non-binding conformation to a specific conformation that can be recognised by the ligand.

3.4. Dynamic equilibrium of the open and closed conformations

Molecular dynamics studies have shown the rapid dynamic equilibrium between the open, semi-open and semi-closed states (Brokaw and Chu, 2010; Pontiggia et al., 2008; Song and Zhu, 2013). The presence of a dynamic equilibrium agreed with experimental studies; Ådén and Wolf-Watz (2007) used NMR to show that the lid populates both the open and closed states in almost equal populations which was further evidenced by a FRET study by Henzler-Wildman et al. (2007b). Another FRET study by Hanson et al. (2007) found the closed state was favoured even in the absence of ligands. However, the disagreement in the population equilibrium may be due to fluorophore positioning (Li et al., 2015). The free energy landscape computed through MD simulations indicate heavily populated open conformation (Li et al., 2015) which agrees with a study by Potoyan et al. (2012) that argue that the free energy basin of the open conformation may be larger than first thought. The idea of a major open conformation would corroborate with the high catalytic rates where a shift towards an open conformation is correlated with an increase in catalytic activity (Ådén et al., 2012). However, a recent study puts this paradigm in doubt; Kovermann et al. (2017) produced a high-energy state ADK structure using a salt bridge to trap the closed conformation and showed that substrate binding still occurred in the closed conformation suggesting that steric hindrance wasn't an issue.

ADK exhibits a dynamic equilibrium in the absence of substrate, however, upon addition of substrate the equilibrium is shifted towards a closed conformation with a concomitant change in the transition rates (Henzler-Wildman et al., 2007b; Ådén and Wolf-Watz, 2007; Li et al., 2015).

3.5. Rate-limiting steps and the relationship with catalysis

The dynamic coupling between conformation and enzymatic function has drawn both experimental and theoretical interest. The seminal work by Henzler-Wildman et al. (2007a) suggested that the hierarchical dynamics contribute to the catalytic function. However, Pislakov et al. (2009) and Kamerlin and Warshel (2010) have proposed that conformational change does not necessarily contribute significantly to catalysis.

In the ligand-free state the domain motion does not constitute the rate-determining step for ADK catalysis (Shapiro and Meirovitch, 2006) where opening/closing transitions occur rapidly at rates of $2 \times 10^7 s^{-1}$. However, an ^{15}N -NMR experiment showed that the opening of the lid and AMP_{bind} domain in the ligand-bound state determined the rate-limiting step for overall catalytic turnover (Wolf-Watz et al., 2004; Shapiro and Meirovitch, 2006; Kern et al., 2005) (at rates of approximately $300 s^{-1}$). Long timescale molecular dynamics (Li et al., 2015; Formoso et al., 2015) and theoretical modelling of experimental results (Wang et al., 2013) found that the lid domain is much more flexible than the AMP_{bind} domain, indicating that the conformational transition of the AMP_{bind} is the rate-limiting step for ADK activity. Interestingly, Kern et al. (Kerns et al., 2015) found that the Mg^{2+} ion specifically accelerates p -transfer and also accelerated lid opening by weakening strong electrostatic interactions in ADK.

3.6. Transitional coupling, intermediate states and pathways of the subdomains

The independent binding of ATP and AMP to the lid and AMP_{bind} subdomains respectively, lends itself towards the idea of independent motion of the separate subdomains. However, a number of studies have found evidence for, or support the hypothesis that, the subdomain closure is coupled. Residue dipolar coupling extracted from NMR experiments by Esteban-Martín et al. (2014) suggested a step-wise mechanism where the lid subdomain closes before the AMP_{bind} subdomain during the closing transition. This sequential ordering was confirmed by long (Li et al., 2015) and short (Arora and Brooks, 2007) timescale MD simulations, normal mode analysis (Maragakis and Karplus, 2005) and coarse-grained simulations (Whitford et al., 2007; Lu and Wang, 2008; Chu and Voth, 2007). However, this lid-first closure requires the presence of a ligand (Daily et al., 2010). Further MD simulations by Jana et al. (2011) suggest a dynamical coupling between the lid and AMP_{bind} subdomains during the open transition and Zeller and Zacharias (2015) propose that opening of the lid is most likely the first step in substrate release. However, this contrasts with simulations that suggest that the lid can only reach the open state after the AMP_{bind} has opened (Kubitzki and de Groot, 2008; Li et al., 2015; Formoso et al., 2015) due to strong salt bridges between the lid and core that introduce a significant enthalpic penalty that obstructs lid opening before the AMP_{bind} (Kubitzki and de Groot, 2008; Beckstein et al., 2009).

The high catalytic rate of ADK owes itself to a variety of features; its ability to prevent non-functional binding is a key attribute, therefore, it is important to explore the coupling of subdomains during the possible binding of substrates to their non-designated subdomain. ATP binds strongly at the ATP lid site $K_d^{ATP_{lid}}(ATP) = 50 \mu M$ and at the AMP binding site $K_d^{AMP_{bd}}(ATP) = 53 \mu M$ suggesting that a significant fraction of ADK molecules may bind

ATP in an incorrect manner. However, whilst AMP binds strongly at the AMP binding site $K_d^{AMP_{bd}}(AMP) = 210\mu M$ it doesn't bind well at the ATP site $K_d^{ATP_{lid}}(AMP) = 1700\mu M$ (Ådén and Wolf-Watz, 2007). A study by Zeller and Zacharias (2015) suggests that ATP binding to the AMP site induces a global domain closing that sterically prevents substrate binding to the ATP site to prevent a stable yet enzymatically unproductive state.

The 'cracking' mechanism requires a local unfolding and refolding of the lid domain, therefore, it is important to explore the folding mechanism of separate subdomains. The core has generally been assumed as a thermodynamically stable relative to the lid and AMP_{bind} domains (Rundqvist et al., 2009), and chimeric analysis of temperature divergent ADK homologues, has shown the melting temperature to be dependent on the core region (Bae and Phillips, 2006). Moreover, the role of ion-pairs in the core has also been posited as important in determining ADK thermal stability (Bae and Phillips, 2005). NMR relaxation experiments revealed non-cooperative folding of the subdomains (Rundqvist et al., 2009) in contrast to simulations that suggest that the lid and AMP_{bind} cooperatively fold with the core domain respectively (Giri Rao and Gosavi, 2014).

Intermediate states that are found along the transition path between major conformations have been identified in a number of proteins (Orellana et al., 2016). The NMR experiments by Esteban-Martín et al. (2014) provided evidence for a closed-like state in the absence of substrate. This agreed with crystal structures (PDB ID: 2AK2 and 2AK1) that appear to exhibit a closed-like state (Schlauderer and Schulz, 1996). The presence of intermediate/sub-states was further evidenced by MD simulations (Li et al., 2015; Feng et al., 2009; Lee et al., 2015; Ping et al., 2013), normal-mode analysis (Ahmed et al., 2011; Peng et al., 2010) and principal component analysis (Cukier, 2009; Lou and Cukier, 2006) that observed a number of intermediate conformational states. Moreover, a multi-time correlation function of MD simulations and single-molecule FRET experiments revealed locally unfolded structures of the lid subdomain (Ono et al., 2015). These intermediate states are a key step in the motion of the AMP_{bind} subdomain. After forming interactions with the lid, the AMP_{bind} moves from the open conformation towards the closed. There is evidence to suggest that the intermediate steps are critical for the opening/closing mechanism in ADK. So does ligand binding occur along the pathway towards closing? And if so, are these intermediate states the critical binding states?

Until now we have used the term pathway to describe the assumed singular path between the open and closed states. However, there is evidence to suggest that there are multiple pathways through which the opening and closing transitions can take, indeed, the path for opening may not be the same as that for closing. A combination of theory and experiment allowed Wang et al. (2013) to theoretically predict the existence of two major parallel stepwise pathways. MD simulations have shown a diversity of pathways from the closed to open state (Peng et al., 2010; Song and Zhu, 2013), however, due to short simulation times its difficult to make any conclusions

regarding pathway symmetry.

3.7. Free energy landscape of ADK

A range of studies have produced one-dimensional representations of the free energy landscape for conformational changes in ADK (Song and Zhu, 2013; Arora and Brooks, 2007; Lou and Cukier, 2006; Potoyan et al., 2012). Impressively, Kerns et al. (2015) have produced a one-dimensional model of the energy landscape during catalysis. However, whilst these provide a reaction trajectory they fail to capture the full conformational changes that are outside a one-dimensional perspective. The coarse grained models have provided good estimates of the change in free energy, however, they still lack the atomistic details of interactions despite the interaction parameters being fitted by all-atom simulations. The dynamical angles and distances of the lid and AMP_{bind} already introduce a complex multidimensional free energy landscape.

According to a number of studies the free energy difference between the open and closed conformations in the absence of ligands is in the range of 1-2 k_BT (Song and Zhu, 2013; Arora and Brooks, 2007; Lou and Cukier, 2006; Potoyan et al., 2012; Li et al., 2015) disagreeing significantly with the 30 kJ/mol determined by (Shapiro and Meirovitch, 2006). A study by Gur et al. (2013) found that an energy barrier is required to be overcome for AMP_{bind} motions but not lid motions. The low kinetic barrier provides further evidence for the dynamic equilibrium in the absence of ligands and also provides the susceptibility to perturbations that is essential for measuring the environment and relaying signals (Potoyan et al., 2012).

Arora and Brooks (2007) predicted that the ligand does not bind to the fully open conformation, instead it binds along the pathway towards closing. Despite the multiwell free energy landscape suggesting a population shift mechanism, it has been proposed that lid-gated enzymes such as ADK are governed by an induced fit mechanism (Sullivan and Holyoak, 2008) because whilst the ADK samples the closed conformation, it is unable to bind ligands in that conformation. A more sensible alternative, proposed by Hammes et al. (2009), is a hybrid of population shift and induced fit mechanisms that is dependent on the protein and substrate concentrations. In fact, this mechanism has already been suggested by Adén et al. (2013) for ADK and MD simulations showcase the importance of ligand binding to reach the fully closed conformation (Brokaw and Chu, 2010).

4. Computational Methods

We saw in Chapter 2 that dynamics across a hierarchy of timescales is essential for functionality. We highlighted the difficulty associated with probing these temporal scales and understanding the mechanisms that drive catalytic dynamics. These difficulties are manifested in our rather limited experimental and theoretical understanding of protein dynamics. In this chapter, we introduce the computational methods used to explore the conformational dynamics of ADK throughout this thesis.

The first methodology, and the focus of this thesis, is Markov stability. An unbiased graph theoretical community detection framework that can access all temporal and spatial scales. The second computational method is molecular dynamics (MD) simulations using classical forcefield's to explore protein flexibility.

4.1. Graph theoretical approaches

4.1.1. Network Construction

The graph theoretical studies discussed in Chapter 2 used various approaches to construct a graph from a protein structure. However, the overwhelming number used coarse grained residue-residue interaction networks (RRINs) to capture inter-residue interactions (Chennubhotla and Bahar, 2007). Amino acids are considered as nodes and the edges between them are either binary defined by a euclidean distance cut-off or weighted by the number of atom-atom contacts between residues (Chennubhotla and Bahar, 2007). Some studies have used chemical groups of similar atoms to granulate the graph structure. This approach showed significant advantages over residue-based graphs because they were able to encapsulate a portion of chemical information in the graph (Benson and Daggett, 2012). Another technique used the cross-correlation in MD simulations to derive edge weights (Sethi et al., 2009) but this was restrained by long simulation times and inherently still produced a coarse-grained graph representation. Whilst RRINs have produced insight into areas such as protein stability and dynamics (Swint-Kruse, 2004), allosteric regulation (Süel et al., 2003) and protein folding kinetics (Jung et al., 2005),

it has been shown that coarse-graining removes important local chemistry that contributes to allosteric communication (Ribeiro and Ortiz, 2014).

Common graph theoretical properties such as community clustering, betweenness, shortest paths and centrality are derived directly from the network, however, they are highly dependent on the assignment of edge weights. A number of studies have highlighted the importance of weighting edges with interaction energies. Vijayabaskar and Vishveshwara (2010) presented a coarse-grained structural network based on the atomic interaction energies derived from MD simulations. Unfortunately, this method only included non-covalent interactions when it is reasonable to expect all physico-chemical interactions to contribute to protein properties. Indeed, a similar method developed by Ribeiro and Ortiz (2014) used the same MD simulation approach but included covalent interactions in their graph and found they were crucial for identifying signal propagation. Their method was based on the assumption that signal propagation preferentially occurs through covalent or strong non-covalent interactions and emphasises the need for accurately assigning edge weights.

The graph construction method we use addresses these previous limitations of granularity and edge weighting. The graph construction method was first developed in Meliga (2009) and Delmotte et al. (2011), building upon the works of Jacobs et al. (2001) that defined a sparse graph based on atomic details. The method was subsequently further developed by Delmotte (2014). The graph construction method used here provides an *atomistic* description in contrast to the coarse-grained methods described above. Furthermore, it includes both covalent and non-covalent interactions that are weighted in proportion to their interaction energy. There is significant heterogeneity in the size, shape and chemical properties of different amino acids and whilst an atomistic description increases the number of nodes and edges, the graph-theoretical methods used are of low-computational cost and, therefore, the increased network size is an obvious trade-off for a more accurate protein-graph representation. Moreover, using an atomistic granularity also allows us to explore the effect of single covalent or non-covalent interactions.

The graph construction process begins with the atomic spatial coordinates obtained from x-ray crystallography or NMR spectroscopy. The 3D structures of proteins are obtained from the on-line Protein Data Bank (PDB) archive (Berman et al., 2000). Covalent and non-covalent interactions are identified through a program called FIRST which was developed by Jacobs et al. (2001) to explore protein unfolding (Rader et al., 2002) using a graph theoretic approach. This provides the initial unweighted protein-graph representation. Edge weights are then derived from the specific potentials associated with the type of interaction. The graph construction process is illustrated in Figure 4.1 and summarised below.

- **Adding hydrogen atoms to the PDB molecular structure.** The protonation software package Reduce (Word and Richardson, 2012) is used to add missing hydrogens in

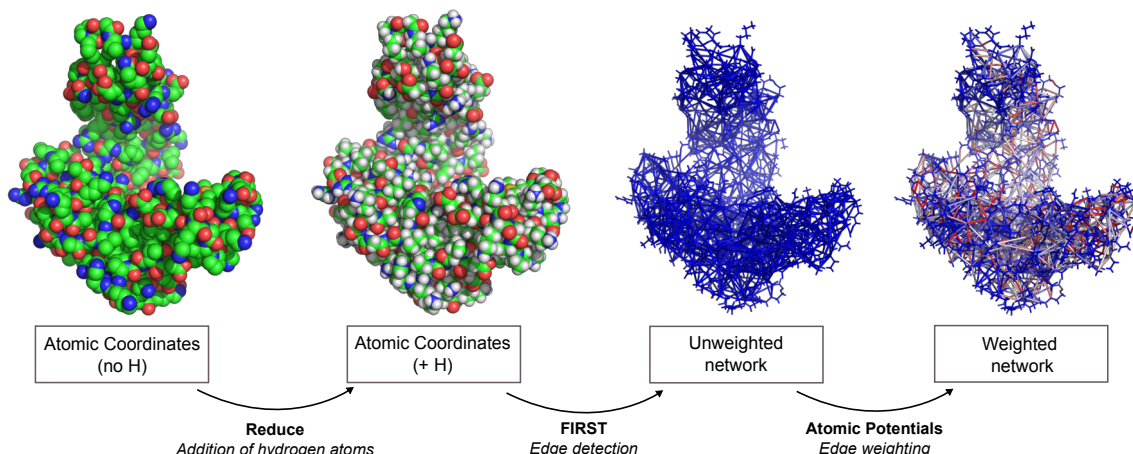


Figure 4.1: Weighted graph construction from an experimental PDB structure. Starting with a PDB file (2RH5 in this example) the structure is first protonated, then every interaction is identified using FIRST and a network is constructed. Using atomic potentials we are able to weight the edges accordingly.

x-ray crystallographic files before graph construction begins. Energy minimisation of mutant structures through Gromacs require different notation and therefore an alternative inbuilt protonation function `pdb2gmh` is used to add hydrogens.

- **Identifying covalent and non-covalent interactions.** The command line executable program FIRST uses geometric and energetic criteria to identify hydrogen bonds, salt bridges and hydrophobic contacts (Jacobs et al., 2001). An unweighted edge is assigned between two atoms based on the type of interaction.
- **Assigning edges weights proportional to interaction energies.** Covalent bond energies are extracted from the standard tabulated bond dissociation energies (Huheey et al., 1993). Dahiyat et al. (1997) and Rader et al. (2002) developed the Mayo potential which is used here to weight the hydrogen bonds and salt bridges, and Lin et al. (2007) proposed a hydrophobic potential of mean force which is used to weight the hydrophobic interactions. π stacking is assigned a fixed weight of 10 kcal/mol (Hobza, 2008). Electrostatic interactions are ignored except when considering ions during which the interactions are calculated using the Coulomb potential. A detailed description of the interaction energies can be found in Appendix A.

Despite edge weighting being crucial for analysis, numerics have shown the final results to generally be insensitive to small variations in the edge weights. The overall framework is robust to changes in forcefields or potentials making the analysis and conclusions universally similar across all fully energy-weighted networks.

4.1.2. Community detection in networks

Global features of a network can be difficult to derive by looking solely at single elements located within a web of non-trivial relationships. An interesting feature of networks is the ability to group nodes into clusters, where the graph topology is organised into separate communities. Clustering nodes provides a mesoscopic description to simplify the overall organisation and derive function and properties of a complex data set. The general principle here is that nodes within a community are highly similar whilst nodes between communities present low similarity.

A variety of graph clustering methods have been proposed and a wide selection of these are reviewed in detail by [Fortunato \(2010\)](#). A more ‘light-hearted’, yet detailed, review of the various limitations of graph clustering methods has been published by [Moore \(2017\)](#). The difficulty underlying graph clustering is that there remains no agreed definition of what defines a ‘good partition’. To assess the goodness of a partition it is useful to introduce a quality function that provides a quantitative criterion to each partition of a graph. The most common notion for defining the goodness of partition is to look at the edge density; comparing the number of intra-community edges with the inter-community edges ([Fortunato and Castellano, 2012](#)). This notion underlies modularity, the most popular quality function devised by [Newman and Girvan \(2004\)](#), and the multiscale-Potts models ([Dorogovtsev et al., 2004](#)). Unfortunately these methods are limited; they have been found to be unreliable for community detection in sparse graphs ([Fortunato, 2010](#)), they exhibit a resolution limit ([Fortunato and Barthélemy, 2007](#)) and they tend to group communities into clique like communities ([Schaub et al., 2012](#)). These methods are therefore unable to access all the relevant dynamical scales associated with hierarchical systems such as proteins.

In contrast to the structure-based approaches, methods have been proposed that detect communities from a dynamical perspective such a random walk or through oscillation synchronisation ([Boccaletti et al., 2007](#)). The main advantages include the ability to identify non-clique communities and explore the hierarchy of scales ([Schaub et al., 2012](#)). Indeed, a diffusive random walk based on a Markov process underpins the Markov Cluster Algorithm (MCL) ([Dongen, 2000](#)), Infomap ([Rosvall and Bergstrom, 2007](#)) and Walktrap ([Pons and Latapy, 2005](#)), and has been used as a method of coarse graining ([Gfeller and De Los Rios, 2007](#)). However, these methods are limited; they do not elucidate the entire spectrum of scales that can be accessed using a diffusive process. The method used here, Markov stability, uses the time evolution of a Markov process on a graph to act as a zooming lens that can identify communities as small as one node through to the global two-community partitions ([Delvenne et al., 2010](#)).

Within the Markov stability framework, the dynamical process can be considered as a random-walker that jumps between nodes that are bound by an edge. Given a short time the random-walker will only explore a small subset of densely connected nodes, however, given longer times

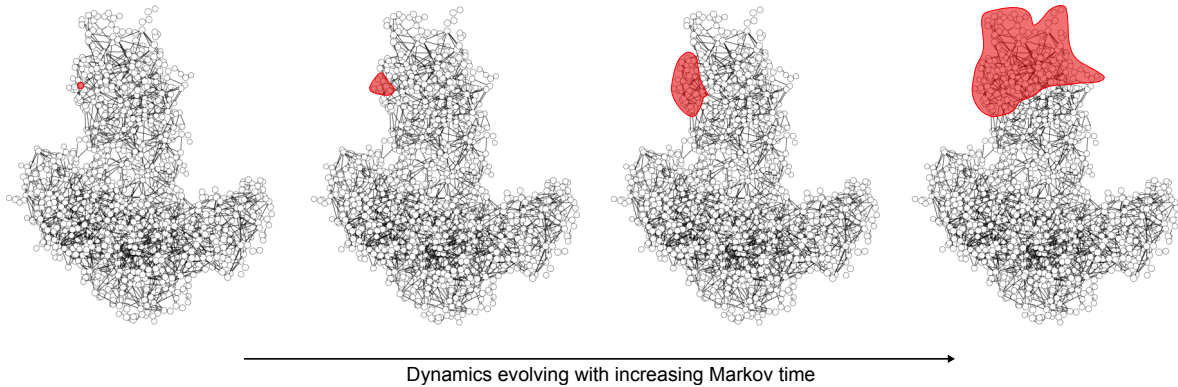


Figure 4.2: Schematic of a Markov stability random-walk on a protein network. A diffusion process can be used to reveal communities in which the random-walker becomes trapped. Longer timescales allows the diffusion of the random-walker to explore larger areas of the protein graph and reveal relevant partitions at different levels of granularity.

the probability of the random walker escaping this subset and exploring further regions of the graph increases. More precisely, a partition will be significant if the communities correspond to subgraphs in which the random-walker becomes trapped at a particular timescale. Using the time of the random-walk as a zooming lens we can uncover a hierarchy of significant partitions. This is an essential component for exploring biomolecules where local bond fluctuations and residue side-chain motions will contribute to the global conformational fluctuations. Indeed, it seems inherently justified to use a dynamical perspective to explore a dynamical system.

4.1.3. Random walks on graphs

The underlying component to the Markov stability framework is the random-walk. This can be considered as a Markovian process that acts on the graph. The topology of a graph with N nodes is encoded in an $N \times N$ adjacency matrix A where the components A_{ij} define the weight of an edge w_{ij} between node i and node j . If no edge is present then $A_{ij} = 0$. The degree of each node d_i can be found from the total number of edges adjacent to that node $d = A1_N$. Diagonalising the degree of nodes vector $D = \text{diag}(d)$ we can define the Markov transition matrix $M = DA^{-1}$ which gives the probability or likelihood M_{ij} of transitioning between node i and node j upon a single discrete timestep. The time evolution of a discrete random-walk can then be defined as

$$p_{t+1} = Mp_t \quad (4.1)$$

where p_t is the $1 \times N$ probability distribution of the random-walker after t timesteps. A fully connected graph under aperiodic conditions will exhibit an ergodic random-walk that will reach a stationary distribution defined as the eigenvector of M , $\pi = \pi M$.

A more practical method is to use a continuous time Markovian process where we define a transition rate matrix Q that defines the rate Q_{ij} at which a random walker jumps between nodes i and j . We can then define a continuously evolving probability distribution

$$\dot{p}(t) = p(t)Q, \quad (4.2)$$

where the walker has a continuous exponential waiting time at each node essentially defining a Poisson process on each node. The solution to equation 4.2 is

$$p(t) = p(0)\exp(Qt), \quad (4.3)$$

where the components of the matrix exponential $(\exp(Qt))_{ij}$ define the probability of starting at node i and ending at node j after time t .

4.1.4. Markov stability of a graph partition

The time evolution of the Markov process can now be used to detect communities. We first encode the binary $N \times c$ community indicator matrix H where the value H_{ij} is 1 if node i falls into community j and 0 otherwise. The location of the random walker at time t is encoded in the $N \times 1$ random walk indicator matrix $X(t)$ where $X_k(t) = 1$ if the random-walker is located on node k and zero elsewhere. This then allows us to define the community indicator matrix $Y(t) = H^T X(t)$ where $Y_k(t) = 1$ if the random-walker is located in community k at time t .

A set of sub-graphs in which the random-walker becomes transiently trapped is assumed to be a good community partition. Using the autocorrelation of the community indicator matrix with itself we can define an autocovariance matrix to reveal these regions

$$\begin{aligned} R(t, H) &= \text{Cov}(Y(t), Y(t + \tau)) = \text{Cov}(H^T X(t), H^T X(t + \tau)) \\ &= H^T \text{Cov}(X(t), X(t + \tau)) H = H^T [\text{E}(X(t), X(t + \tau)) - \text{E}(X(t + \tau))^2] H \\ &= H^T [\Pi \exp(Qt) - \pi^T \pi] H, \end{aligned} \quad (4.4)$$

where E is the expectation and $\Pi = \text{diag}(\pi)$. It is reasonably assumed that the stationary distribution π is the initial condition of the Markov process in the absence of any *a priori* information. The elements $R(t, H)_{ij}$ have an intuitive interpretation; they provide the probability of the random-walker starting in community i and ending in community j after time t

minus the expectation of the random-walker being on that node by chance. The diagonal elements $R(t, H)_{ii}$ give the probability of the random-walker starting in community i and ending in community i . Therefore, we can define the trace of $R(t, H)$ as a quality function of a graph partition

$$r(t, H) = \text{trace} R(t, H), \quad (4.5)$$

this is the *Markov stability* of a graph partition. Markov stability can be used to assess the quality of a single existing community partition but its most powerful application is during an community optimisation process across partition space. We can look to maximise the Markov stability by iterating through community partitions.

The transition rate matrix Q encodes the dynamics of the system and can be defined in multiple ways such that *a priori* information about the type of system can be included (Lambiotte et al., 2014). Here we use the combinatorial Laplacian matrix L which has been used previously to describe the dynamics of electrical networks (Schaub et al., 2014; Wu and Huberman, 2004), linear synchronised oscillation systems (Barahona and Pecora, 2002; Arenas et al., 2006), heat diffusion (Lafferty and Kondor, 2002) and appropriately it has been applied in the context of vibrations (Reuveni et al., 2012) and signal propagation (Chennubhotla and Bahar, 2007) in protein systems. The Laplacian is defined by

$$Q = A - D = -L, \quad (4.6)$$

where A and D are the pre-defined adjacency matrix and diagonal degree of nodes respectively. The combinatorial Laplacian used throughout this work is defined as

$$L_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ \sum_i -w_{ij} & \text{if } i = j \end{cases} \quad (4.7)$$

where w_{ij} is the weight of the edges between node i and node j . The random-walk defined by this Laplacian has waiting times on each node that are inversely proportional to the degree of that node. The probability of transition from node i to node j is proportional to weight of the interaction between the nodes w_{ij} and therefore we see propagation of the random-walk preferentially through stronger interactions.

Using the Laplacian matrix the stationary distribution becomes uniform over all nodes $\pi = 1_N/N$ and therefore we can rewrite Markov stability as

$$r(t, H) = \text{trace} \left(H^T \left[\frac{1}{N} e^{-Lt} - \frac{1_T 1_N}{N^2} \right] H \right). \quad (4.8)$$

We are now in a position to use Markov stability to optimise the community partition at any time t over a chosen time range. The time evolution of the probabilistic flow acts as a resolution parameter uncovering different optimal community partitions. At shorter times the Markov process will become trapped inside local subgraphs but as time increases the process will reveal coarser partitions. In contrast to traditional community detection methods, Markov stability provides a multiscale view of the graph that is not limited by an implicit scale. This framework is, therefore, well suited to analyse protein structures which exhibit levels of organisation that span various temporal and spatial scales. Moreover, using the retention of a probabilistic flow allows the detection of non-clique-like communities. This is important when dealing with natural communities such as amino acids where not every atom interacts with every other.

Appropriately, the dynamical framework of Markov stability is a generalisation of classical graph partitioning heuristics (Delvenne et al., 2013). In the limit $t \rightarrow \infty$, the optimal Markov stability partition is equal to the Fiedler partition (Fiedler, 1973, 1975) and at $t = 0$ each node is its own community (associating itself with entropy) (Delvenne et al., 2010). In the case of using discrete dynamics we find that modularity is equivalent to $t = 1$ (Delvenne et al., 2013).

4.1.5. Partition optimisation with the Louvain algorithm

Markov stability is a quality function that assesses the goodness of a predefined partition. However, similar to other quality functions (Fortunato, 2010), optimising Markov stability exhaustively across partition space is NP-hard (Brandes et al., 2008). More simply, it would take too long to calculate the Markov stability of every possible partition in a large graph to find the exact solution. In practice, a number of heuristic methods have been developed that generally use agglomeration (Newman, 2004; Clauset et al., 2004), divisiveness (Newman, 2006; Smyth and White, 2005), extremal optimisation (Duch and Arenas, 2005) or simulated annealing (Guimerà et al., 2004; Reichardt and Bornholdt, 2006). These strategies can be used to identify good partitions that can subsequently be ranked by Markov stability to identify near-optimal partitions. Here we use the Louvain algorithm (Blondel et al., 2008), a greedy agglomerative method that is a highly efficient at optimising Markov stability.

The Louvain method can be broken into two stages. The first stage is the reassignment of communities where we first assign each node to its own community and then randomly iterate through each node. We then remove each node from its current community and consider the change in Stability upon placing that node in the community of a neighbour. If Markov stability is larger after the community reassignment then we move that node into the community asso-

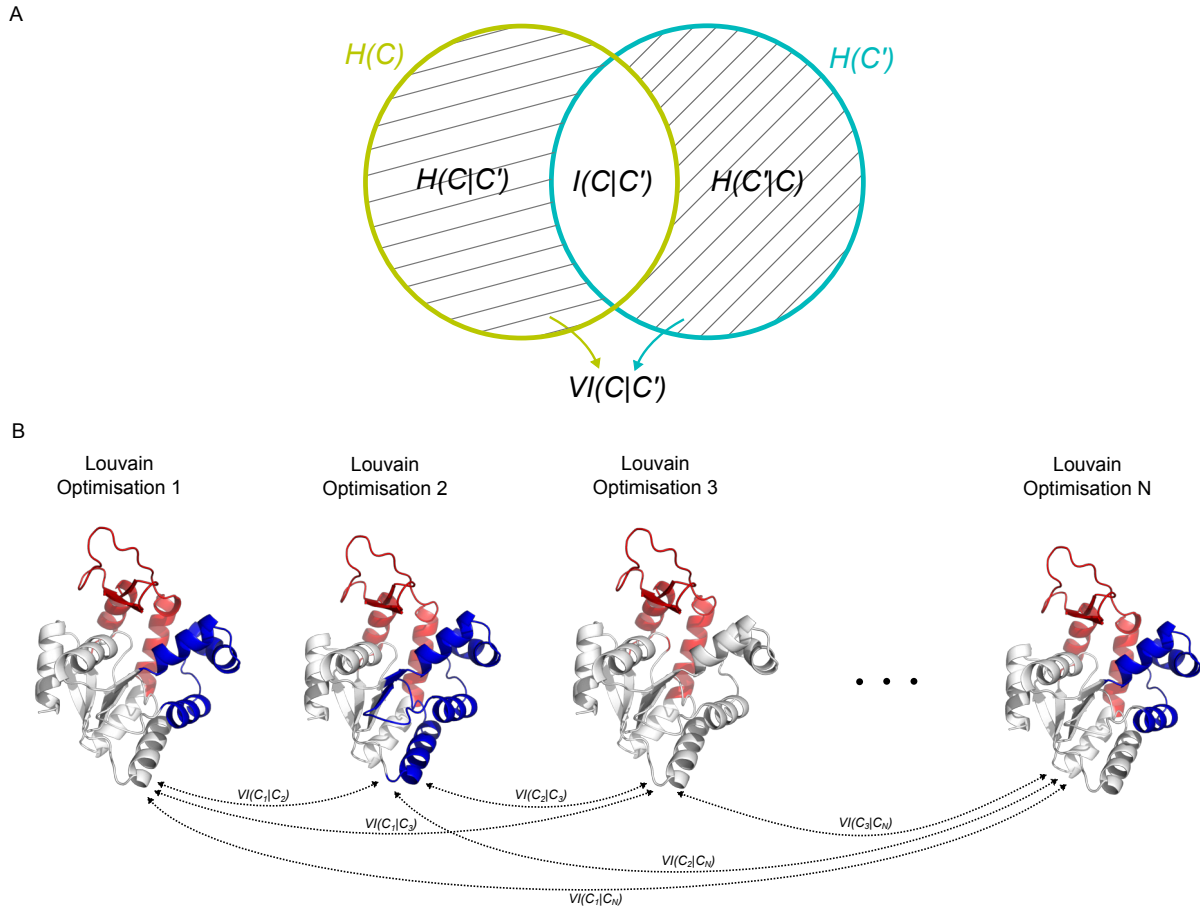


Figure 4.3: The variation of information represented and its use to explore similarity in optimised community structures. (A) The variation of information is represented by the sum of lined areas. (B) The variation of information is calculated across N louvain optimisations of Markov stability, which is then averaged to give the $\overline{VI}$.

ciated with the positive change, else we keep the node in its current community. This process is repeated until node reassignment yields no positive changes to Markov stability. The second stage, coarse graining, uses the newly defined communities from the community reassignment stage as nodes giving a meta-graph - a coarse-grained representation of the graph. The edge weights between community nodes are the sum of edge weights between the constituent nodes of each community. The first stage is then repeated on the newly produced meta-graph until no further reassignment can further improve the Markov stability value $r(t, H)$ at time t . Whilst the Louvain method is deterministic, it is dependent on the initial conditions. The final solution will vary according to the order in which the nodes are reassigned. The initial ordering is random, therefore, running the louvain algorithm multiple times will produce variability in the observed solution. This variability can be used to calculate the statistical relevance or robustness of a partition.

4.1.6. Robustness of a partition

We are now able to find the optimal partition for each value of Markov time, however, this is not necessarily indicative of a meaningful partition. There is no *a priori* reason that an optimal partition at time t has any significance, indeed there would only be a finite number of meaningful partitions present, if any at all. When considering a multiscale structure it is important to distinguish a meaningful partition from a transient solution. We can then define a significant partition as one that is robust to perturbations as suggested by Karrer et al. (2008). So what perturbations do we use? Here we employ two perturbative strategies for measuring the robustness of a community partition.

The first strategy is to look at the temporal lifetime of a partition where the time is itself a perturbation (Ronhovde and Nussinov, 2009). Simply put, the robustness is measured by how long a particular partition exists across Markov time. We can therefore look for plateaus in the number of communities across Markov time.

The second strategy for measuring robustness is to quantify the difference in a set of near-optimal partitions that are found using various instances of the Louvain algorithm. There are a variety of partition-distance measures such as the Rand index (Rand, 1971) and Jaccard coefficient (Ben-Hur et al., 2002). Here we use a true metric known as the *variation of information* which is an information-theoretic distance based on the similarity/dissimilarity of information between two partitions (Meila, 2007). Using the probability $P(k) = \frac{n_k}{n}$ of a node being inside cluster C_K we can calculate the Shannon entropy associated with clustering C as

$$H(C) = - \sum_{k=1}^K P(k) \log P(k). \quad (4.9)$$

We can then use the Shannon entropy to define the variation of information between two clusterings C and C'

$$VI(C, C') = H(C|C') + H(C'|C), \quad (4.10)$$

where $H(C|C')$ is the conditional entropy of C given C' . The conditional entropy $H(C|C')$ uses the probability $P(k, k')$ that a node belongs to a particular community k in clustering C and k' in clustering C' . Effectively, the variation of information calculates the amount of uncertainty left about C when you know C' .

The variation of information is dependent on the number of nodes where $\max(VI(C, C')) = \log(N)$ and $C = C'$. Therefore, here we normalise the variation of information by dividing

through by $\log(N)$. At each Markov time L Louvain optimisations finds an ensemble of L near-optimal partitions where the average pairwise variation of information can be calculated.

$$\bar{VI} = \frac{1}{L(L-1)} \sum_{1 \leq i < j \leq L} VI(P_i, P_j) \quad (4.11)$$

The variation of information tells us whether a partition of the protein is meaningful. A low value suggests a meaningful or biophysically relevant partition whilst a high value suggests the partition isn't biophysically relevant. In more detail, a high VI means that each time we run Markov Stability we are finding distinctly different partitions of the protein suggesting that there is no biophysically meaningful partition at that timescale. Whilst a drop in VI means that each time we run Markov Stability we are finding very similar or the same partition each time, therefore, the partition at that timescale is likely biophysically relevant.

4.2. Molecular Dynamics and Computational Mutation

This section gives a basic introduction to the molecular dynamics (MD) theory and implementation applied to biomolecular structures. This methodology provides two benefits to these studies. Firstly, we use molecular dynamics to produce a realistically packed structure after computational mutation. Substitution of smaller residues with larger residues can introduce steric clashes, therefore using energy minimisation we can eliminate any residual strain. Secondly, we can probe the atomistic changes in flexibility associated with the mutations predicted by Markov stability.

4.2.1. Residue substitution

Residue mutations were introduced directly to the 3-Dimensional PDB structure using PyRosetta ([Chaudhury et al., 2010](#)). The *mutate_residue* set of functions was imported from the PyRosetta toolbox. A *Pose* object is created and filled with information from the PDB file, including the residue numbering. The *mutants.py* functions are used to introduce a mutation and output (dump) a new mutant PDB file.

4.2.2. Energy Minimisation

Before an MD simulation, or before constructing a graph from a mutated structure, we must energy minimise the system to produce realistic side-chain packing. Energy minimisation is a process of spatially rearranging atoms to minimise the net inter-atomic force on each atom

according to chemical bonds. Here we are not looking to identify a global energy minimum but a local one that will allow suitable side-chain packing.

Energy minimisation was performed using the GROMACS 5.0 package¹. There are a number of steps required to energy minimise a protein structure:

- **Preprocessing the PDB file.** A protein structure file is downloaded from the RCSB website and any crystallised water molecules must be removed. The protein structure is carefully assessed to identify any necessary missing regions or incomplete sequences.
- **Generating the topology information using a forcefield.** The ‘*pdb2gmx*’ GROMACS module is used to generate the topology of the protein such that it includes all the necessary information for simulation such as atom types, charges, bonds etc. The information stored will be dependent upon the chosen forcefield, here we use the all-atom OPLS-AA forcefield (Jorgensen and Tirado-rives, 1988) such that every atom is included explicitly.
- **Protein Solvation.** The protein must be minimised in respect to a solvent, in this thesis we have used water. A box with dimensions suitably larger than the protein is filled with water using the ‘*solvate*’ module.
- **Energy minimisation.** The structure is relaxed using the ‘*mdrun*’ module which is part of the GROMACS MD engine. To assess whether energy minimisation was successful we look for a negative potential energy on the order of $10^5 - 10^6$.

This will produce a PDB structure where the atom coordinates have been locally altered and any residual strain is eliminated.

4.2.3. Molecular dynamics simulations

Since the first introduction in the late 1950’s by Alder and Wainwright (1957), molecular dynamics simulations are now commonplace to investigate the thermodynamics, structure and conformational fluctuations of proteins. They provide detailed information on time-dependent behaviour of a protein system and are often used in conjunction with experimental procedures. Here we use short molecular dynamics simulation to assess the changes in flexibility associated with the different mutant structures.

An energy minimised structure is first be produced as described in Section 4.2.2 to ensure that we have a reasonable starting structure in terms of geometry and solvation. Using GROMACS

¹GROMACS 5.0 package freely available at <http://www.gromacs.org/> (Abraham et al., 2015)

5.0 package (Abraham et al., 2015) we equilibrated the solvent around the protein to prevent system collapse. The protein coordinates are restrained and the solvent is increased/decreased to the temperature at which we wish to simulate (NVT equilibration). Once the system has reached the correct temperature based on kinetic energies we implement a pressure to bring the system to the proper density (NPT equilibration). The system should now be well-equilibrated and at the desired temperature and pressure. The position restraints are removed and a full production run MD simulation is performed for 2 ns. Due to the stochastic nature of an MD simulation, 10 repeat MD simulations are calculated for each protein.

The root mean square distance (RMSD) provides a measure of the average distance of an atom between two structures. The RMSD is computed relative to a ‘reference’ conformation, usually defined as the original protein structure. Atom-positional mean-square fluctuations (RMSF’s) were calculated between the input reference structure and the trajectory structures using GRO-MACS. The mean and standard deviation of RMSF for each residue was calculated across the 10 repeats.

LINCS constraints algorithm which uses Lagrange multipliers was used to constrain bonds. A leap-frog integrator was used with 2 fs timesteps. A Particle Mesh Ewald (PME) coulomb type was used for long-range electrostatics with a 4-order cubic interpolation and a 0.16 grid spacing for fast fourier transform. A modified Berendsen thermostat was used for temperature coupling with a 0.1 ps time constant and 300K reference temperature. Parrinello-Rahman was used for isotropic pressure coupling with a 2ps time constant and a 1bar reference pressure. Isothermal compressibility was defined as that of water: $4.5^{-5}bar^{-1}$.

5. Experimental Materials and Methods

This chapter outlines the experimental materials, methods and biophysical theory presented in this thesis. The methods have been split into 4 sections. Firstly, we provide a description of the core experimental techniques, including manipulation and visualisation of DNA and protein, through to the confocal microscopy setups. This is followed by a description of the fluorescence theory that underlies the majority of experiments throughout this thesis. Thirdly, we describe the analytical methods used to extract information from fluorescent experiments such as single-molecule FRET, fluorescence correlation spectroscopy and circular dichroism. The final section details the preparation of protein samples and the associated materials.

5.1. Experimental Methods

5.1.1. DNA Techniques

Quick-change mutagenesis

Point mutations were introduced to the DNA plasmid using the QuikChange II site-directed mutagenesis kit (Agilent Technologies, Santa Clara, USA). The basic procedure employs two complimentary synthetic oligonucleotide primers containing the desired mutation which are extended during temperature cycling using a DNA polymerase. A sample reaction consisted of 5 μ l of 10 $\times$ reaction buffer, 1 μ l of 50ng/ μ l DNA template, 125ng of each oligonucleotide, 1 μ l of dNTP mix (ThermoFisher Scientific), 1 μ l of PfuUltra HF DNA polymerase (2.5 U/ μ l) and UltraPureTM DNase/RNase-Free Distilled Water (ThermoFisher Scientific) to a final volume of 50 μ l.

The primers were designed such that the mutation is located in the middle of the primer with 10-15 bases of correct sequence either side. The melting temperatures of the primers are calculated using

$$T_m = 81.5 + 0.41(\%GC) - (675/N) - \%mismatch, \quad (5.1)$$

where N is the primer length, %GC is the percentage of C or G bases and %mismatch is the percentage of bases that are non-complimentary. The primer is optimised such that $T_m \geq 78^\circ\text{C}$.

The heating cycle was performed on a PTC-200 Peltier Thermal Cycler (GMI, Minnesota, USA) where an initial 95°C for 30 seconds was followed by 18 cycles of 95°C 30 seconds, 55°C 1 minute, 68°C 1 min/kb of plasmid length. Digestion of the amplification products was performed using $1\mu\text{l}$ of Dpn I ($10\text{ U}/\mu\text{l}$) restriction enzyme (ThermoFisher Scientific) at 37°C for 1 hour.

Heat-Shock transformation

Either XL1-Blue supercompetent cells (post-mutagenesis), DH5 α (for DNA extraction through mini prep methods) or BL21 (DE3) cells (prior to protein expression) were thawed on ice. For each sample reaction to be transformed, $50\mu\text{l}$ was aliquotted into a prechilled 14ml BD Falcon polypropylene round-bottom tube. To each bacterial aliquot, $1\mu\text{l}$ of DNA sample was added and the transformation reaction was swirled and incubated on ice for 30 minutes. A heatbath was set to precisely 42°C and the sample was heat pulsed for 45 seconds before being placed on ice for 2 minutes. The bacteria was then added to 0.5ml of preheated 42°C NZY broth without any antibiotic, the sample was then shaking incubated for 1 hour at 37°C with a rate of 250rpm.

DNA gel-electrophoresis

Depending on the size of the DNA sample, a 0.5-3% agarose gel (Sigma-Aldrich Company Ltd., Dorset, UK) pre-stained with $3\mu\text{l}$ of GelRed dye (Biotium Inc. California, USA) was molded into a 13 x 16 cm casting tray. Small DNA samples ($<100\text{ bp}$) required a higher percentage agarose to resolve accurately. The wells were loaded with approximately 50ng of DNA sample alongside an appropriately chosen DNA ladder relative to the size of the DNA sample (either a '1 kb DNA ladder' or a '50 bp DNA ladder' from ThermoFisher Scientific).

The electrophoresis buffer was 50 $\times$ Tris-acetate-EDTA 'TAE' (ThermoFisher Scientific) diluted with MilliQ water accordingly and supplemented with $2\mu\text{l}$ of GelRed. A potential of 5 V/cm was applied at room temperature. Images of the gel were produced using G:BOX Chemi XR5 (SYNGENE, Cambridge, UK).

5.1.2. Protein Techniques

SDS-PAGE electrophoresis

Measuring the protein size and identifying fluorophore labelling can easily be visualised using SDS-PAGE electrophoresis. Soluble protein was diluted to a concentration of 0.5 mg/ml and a

6.5 μl volume was unfolded using 1 μl of 10 $\times$ NuPAGE sample reducing agent (ThermoFisher Scientific) in the presence of 2.5 μl of 4 $\times$ NuPAGE LDS sample buffer (ThermoFisher Scientific) was added to optimise activity of the reducing agent. Protein length was approximated using a 10 μl Novex Sharp Pre-Stained protein standard. The samples and protein ladder were loaded onto a 4 – 12% Bis-Tris gel inside an X-Cell SureLock mini-cell electrophoresis system (ThermoFisher Scientific). The gel tank was filled with either 1% MES or MOPS running buffer (ThermoFisher Scientific).

Coomassie Brilliant Blue G-250 (Bio-Rad Laboratories, Hercules, USA) was used to stain and visualise unlabelled proteins (45% methanol, 8% acetic acid, 0.08% (w/v) Coomassie Blue). A washing solution (10% acetic acid, 10% methanol) was used to destain by removing excess unbound coomassie such that the background was reduced and protein bands were easily visualised.

Visualisation of labelled proteins was attained using a Typhoon FLA 9500 (Amersham) in fluorescence imaging mode. Excitation of Alexa 488 and Alexa 633 was performed using a 488 nm and 633 nm laser respectively. The system was set to use Cy2 (510 nm) and Cy5 (670 nm) emission filters to measure the emission of Alexa 488 and Alexa 633 respectively. The Typhoon was linked to a PC running ImageQuantTM software package (Amersham), allowing image quantification. Noise filtering and protein band selection could be identified using the softwares contour fitting routines (user-guided application) and utilised a ‘rolling ball’ method to subtract background.

Native-PAGE electrophoresis

Assessing the native state of the protein is important to check for aggregation or oligomerisation which would give erroneous conclusions from the single-molecule measurements. Here we used the NativePAGETM Bis-Tris Gel system from Thermofisher. This method is based on the Blue Native Polyacrylamide Gel Electrophoresis technique (BN PAGE) that was developed by Schagger and von Jagow that uses Coomassie G-250 as a charge shift molecule (whereas SDS PAGE uses SDS as a charge shift molecule). The Coomassie G-250 is used to confer a net negative charge to a folded protein to maintain molecular stability. The cathode buffer and sample buffer contains G-250 to continuously provide a supply of G-250 and maintain a negative charge on the protein.

Native PAGE was performed in a X-Cell Surelock protein gel tank with a 15-well 4-16% NativePage Novex Bis-Tris gel. Samples were prepared with 6.5 μl of 0.5 mg/ml soluble protein, 2.5 μl NativePAGE Sample Buffer (4 $\times$), 0.5 μl NativePage 5% G-250 sample additive. In Native Page the separation of buffers is important to retain a negative charge on the protein at all times and, therefore, the upper cathode buffer chamber was filled with 200 ml of 1 $\times$ dark blue cathode

buffer and the lower anode buffer chamber was filled with 550 ml of anode buffer.

Luciferase activity assays

An ADP substrate assay was used to determine the rate of production of ATP to assess the enzymatic activity. A Kinase-Glo® Luminescent Kinase Assay Luminescence (Promega, Wisconsin, USA) was used to accurately quantify the amount of ATP present. The kinase reaction was made to include an excess of substrate, 500 μM ADP, and 5 nM protein final concentrations such that the rate-limiting step was enzymatic turnover and not available ADP. A final reaction solution was made with a 50 μl kinase reaction and 50 μl Kinase-Glo® reagent in a 96 well-plate. The ADP was added last to the sample reaction such that ATP production wouldn't begin before monitoring. The production of ATP was measured via a luciferase reaction using a time-resolved luminometer with no filter over a 4 hour period at constant temperature of 22°C .

5.1.3. Ultraviolet-Visible (UV-Vis) Spectroscopy

UV-Vis spectroscopy is a useful method for measuring the concentration of either DNA or protein. It measures the absorption of light at wavelengths that correspond to the particular biomolecule.

Prior to DNA transformation or mutagenesis, a Nanodrop™ 2000 spectrophotometer (ThermoFisher Scientific) was used to measure the concentration and purity of DNA samples. The 260nm/280nm absorbance ratio was the primary measure of purity where a ratio of 1.8 is deemed 'pure' for DNA according to the Nanodrop manual (Nanodrop, 2009). A value of 1.8-2.2 using the 260nm/230nm ratio is a secondary measure of high purity. The pedestal was cleaned using MilliQ water followed by a blank measurement using sample buffer. A lint-free laboratory wipe was used to gently remove any sample buffer before 2 μl of sample was loaded onto the pedestal for a UV-VIS measurement.

Whilst the concentration required for single-molecule measurements was a process of trial and error, circular dichroism and activity assays required accurate measurement of protein concentration. This was measured using the Nanodrop™ 2000 spectrophotometer with two different methods:

- **Bradford Assay.** A coomassie G-250 dye binds to the protein resulting in a spectral shift from 465nm to 610nm. Comparison of the sample absorbance against a known BSA calibration curve determines to concentration of protein.
- **Absorbance at 280nm.** Quick measurements of protein concentration can be obtained using the absorbance at 280nm. Its accuracy is reduced by non-protein UV absorbing

components in the sample mixture. Suitably accurate measurements require knowledge of the protein extinction coefficient.

5.1.4. Cleaning and Decontamination

In order to produce reliable and accurate measurements it is essential to thoroughly clean and decontaminate equipment. This is particularly important during the biochemistry stages where contamination can result in unwanted species. General equipment was cleaned with ethanol and rinsed with MQ, whilst culturing equipment was decontaminated with Virkon (Sigma-Aldrich Company Ltd., Dorset, UK) and sent for departmental cleaning. Particular equipment, such as quartz cuvettes, were soaked in 1 mM EDTA solution to remove surface adsorbed metal ions and subsequently left in 1% decon 90 overnight to remove proteins adsorbed to the surfaces.

5.1.5. Confocal Microscopy

Conventional microscopy (i.e. wide-field microscopy) suffers from a number of limitations, the principle issue being the large unfocused background that is a consequence of flooding the entire specimen with light. To overcome this, there are a number of advanced techniques such as total internal reflection and super-resolution imaging. Here we use confocal microscopy, developed in 1955 by [Minsky \(1988\)](#), which uses spatial filtering to eliminate out-of-focus light and accomplish a true 3-dimensional optical resolution.

A pinhole is placed in front of the detector as illustrated in [Figure 5.1](#). The pinhole geometry is used to remove any scattering or emissions produced outside of the nano-litre excitation volume ([Moerner and Fromm, 2003](#)). A laser beam produces the exciting radiation which is subsequently directed into the microscope objective, and thus through to the sample, via a dichroic mirror. The same objective then collects the fluorescence from the sample and then the light passes through the dichroic and emission filter. The pinhole in the field aperture (image-plane) prevents any fluorescence from any region outside of the focal nano-volume, providing axial resolution.

Microscopy setup

Freely diffusing single-molecules were detected using a home-built dual-channel confocal fluorescence microscope, further details on this apparatus can be found in [Lenn et al. \(2011\)](#) and a schematic of the instrumentation can be seen in [Figure 5.2](#). A tunable wavelength argon ion laser (model 35LAP321-230, Melles Griot, Carlsbad, CA) was set to 514.5 nm to excite Alexa 488. Referencing [Figure 5.2](#), coming out of the laser the beam is redirected and lifted to the height

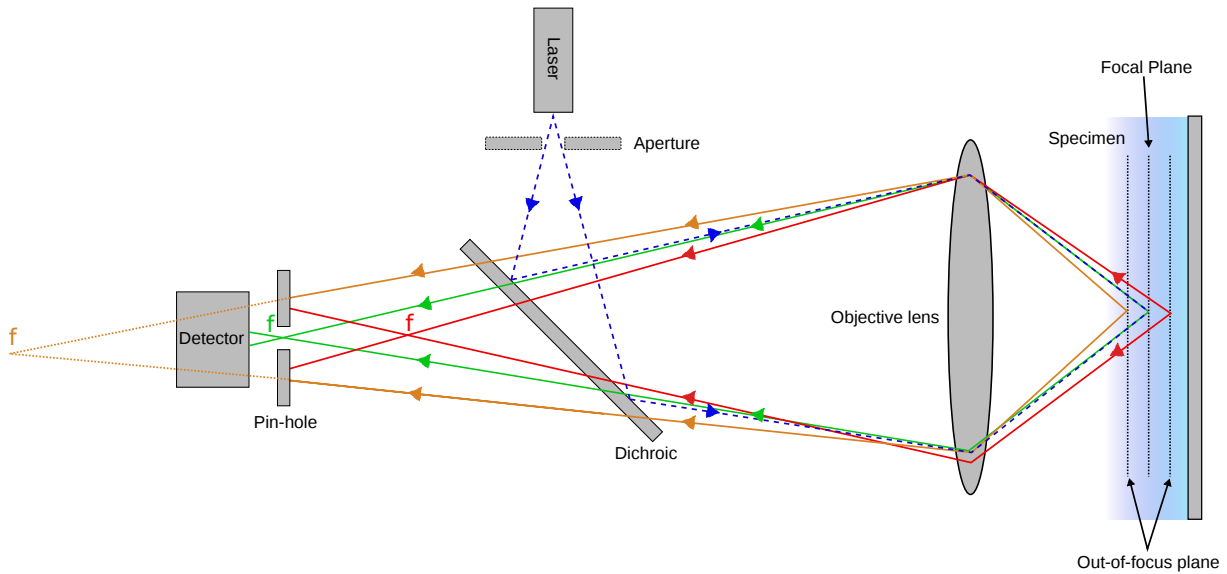


Figure 5.1: Illustration of how confocal microscopy allows in-focus light to pass to the detector. The light rays coloured red represent fluorescent emission from outside the focal plane and upon contact with the confocal pinhole these rays do not reach the detector. Only the green emitted light rays from the in-focus plane can reach the detector. Adapted from Gramatikov (2014).

of the microscope by a mirror. A neutral density filter is used to reduce the beam intensity by 10^{-D} where D is the optical density (Omega, 2017). Plasma lines are then removed using a Dove prism which totally internally reflects the beam (OptoSigma, 2017). A beam expander is employed to focus the beam to a single point such that the collimating lens can re-collimate the expanded beam. Aberrations can be removed from the beam using an aperture (spatial filter) which is then followed by two positional mirrors for entry to the microscope. Prior to entering the inverted microscope an aperture is placed to reduce the beam size such that it fills the back-aperture of the objective. This acts to elongate and enlarge the confocal volume, reduce diffraction fringes and produce a near Gaussian illumination profile (Hess and Webb, 2002).

The laser power is measured after the aperture and before entering the inverted microscope. Inside the inverted microscope an excitation filter (band-pass filter) is used to isolate light at the laser wavelength. The beam was focused into the sample solution to a diffraction-limited spot with a high numerical aperture oil-immersion objective (Nikon Plan Apo TIRF 60x, NA 1.45). The closer refractive indexes of oil and glass relative to water and glass make oil immersion preferable due to reduced light reflection (Lichtman and Conchello, 2005; Moerner and Fromm, 2003). Type FF immersion oil (Cargille, USA) was used due to its negligible fluorescent properties.

The objective-sample distance is adjusted such that the focus of the beam is located $6 \mu\text{m}$ from the slide surface to create a consistent confocal intensity profile across all measurements. The excitation volume together with the pinhole (see Figure 5.2) provides an observation volume of the order $\sim 1 \text{ fl}$. The same objective was used to collect the emitted photons (epifluorescence) which

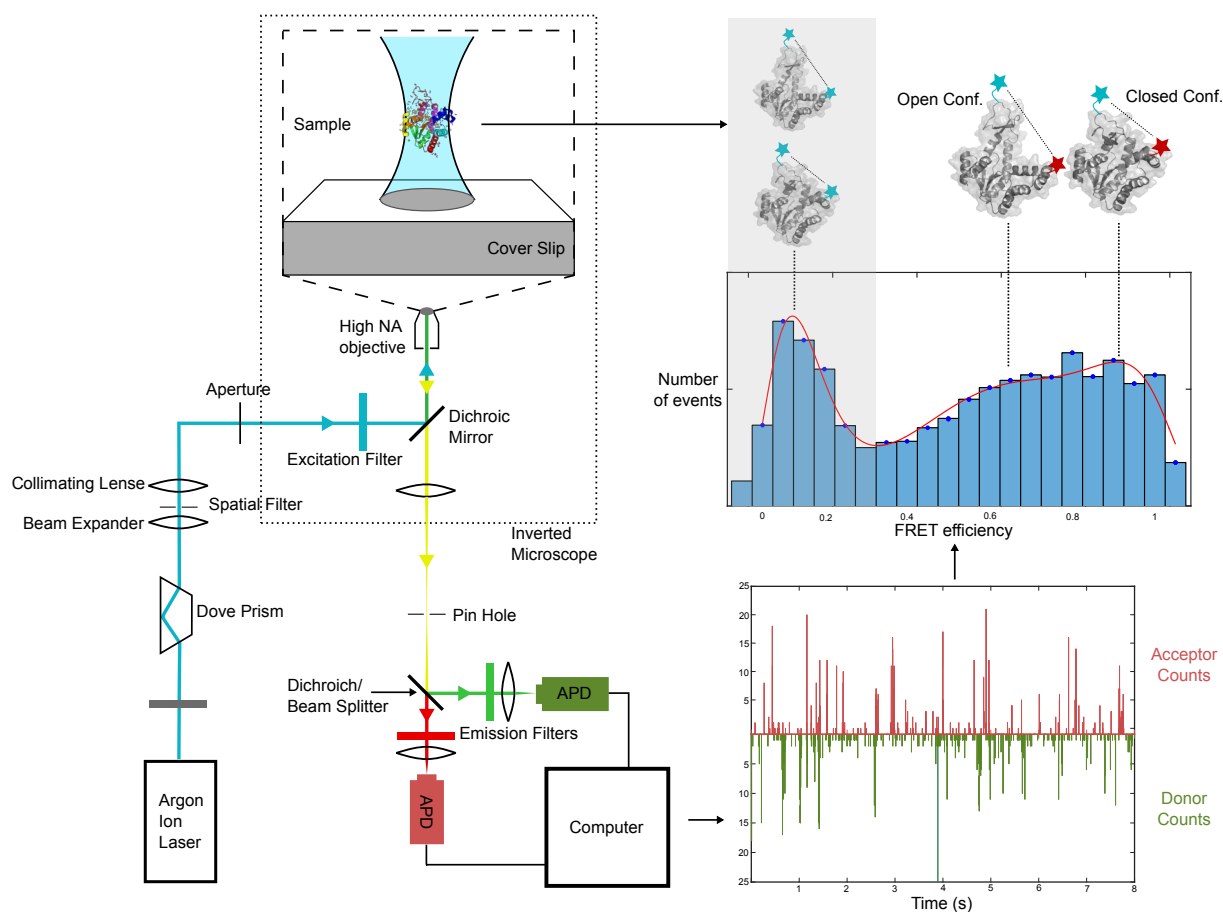


Figure 5.2: Overview of the instrumentation and the data reduction in the confocal single-molecule spectroscopy setup. The schematic on the left illustrates the main components of the 2-APD confocal microscopy setup. The dichroic/beam splitter is alternated for smFRET and FCS respectively. In the case of FCS, the emission filters are the same. The data shown on the right is a sample trajectory of detected photons recorded by freely diffusing molecules over 8 seconds. Each burst corresponds to a single molecule floating through the confocal volume. The time trajectory of photons is reduced into a FRET efficiency histogram, where subpopulations appear that correspond to conformational states of the protein structure.

were then directed through a dichroic mirror that separates the excitation and emission beams. The emission beam is then focused to a point at which the pin-hole is positioned. A second dichroic is used to separate the donor and acceptor fluorescence signals which are then separately detected by two Avalanche photodiodes ‘APDs’ photon-counting modules (SPCMAQR14, Perkin-Elmer) after passing through their respective emission filters. The sensors are approximately $180\ \mu\text{m}$ in diameter, however, the edges are less sensitive (Optoelectronics, 2002) so the beam is narrowed to an $80\ \mu\text{m}$ diameter spot. The sensors have a dead time of 50 ns between detection events (Optoelectronics, 2002). Photon detection was recorded by two computer-implemented multichannel scalar cards (MCS-PCI, ORTEC, Canada) with a sample rate of up to 150 MHz. For FCS measurements the APDs are connected to a Flex02-01d/c (MCS-pci. ORTEC, Tennessee) digital hardware correlator with a sample rate of 640 MHz.

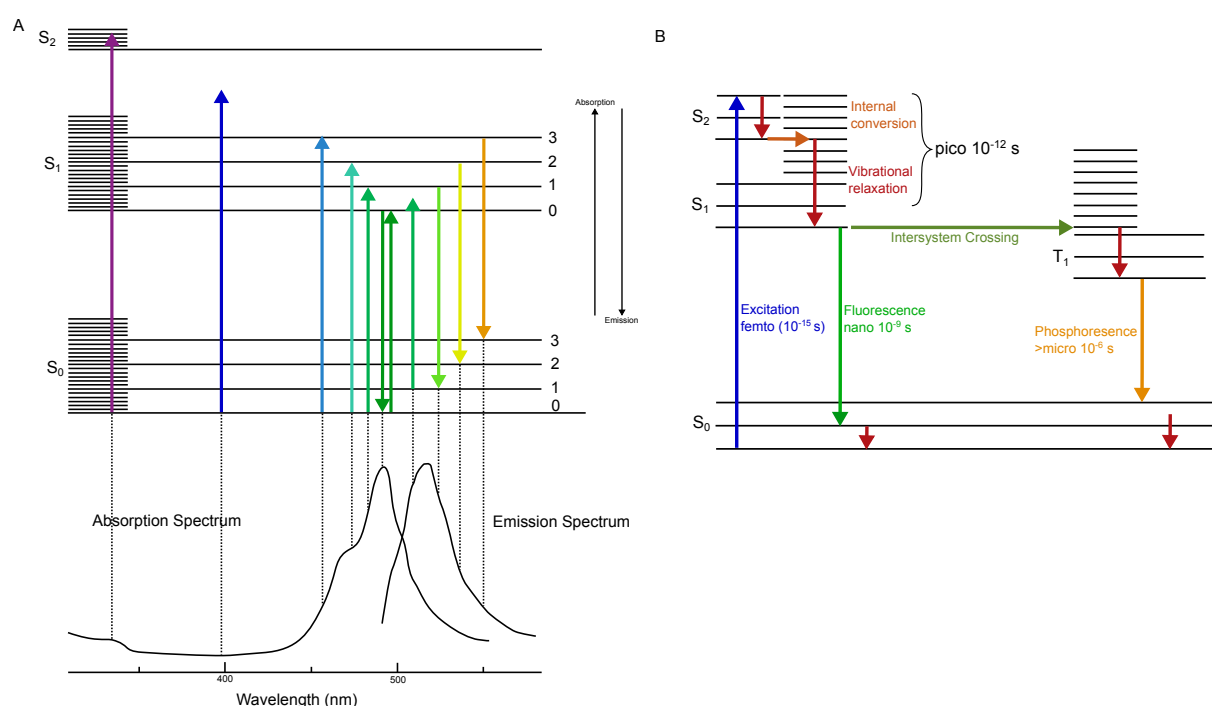


Figure 5.3: Fundamentals of fluorescence described with a Jablonski diagram. (a) The spectral characteristics of absorbance and emission is related to the size of the energy steps between the various states. Each vertical line corresponds to absorbed (arrow upwards) or emitted (arrow downwards) photons. The orange arrow on the right corresponds to the lowest energy photon that can be emitted whilst the purple arrow on the left corresponds to a high energy UV photon that can excite an electron from the ground state through to the second singlet excited state. It is important to note that overlap between the absorption and emission spectrum's where photons are absorbed when the electron is not at the lowest vibrational level. (b) The timescales that the various excitations and relaxations occur over.

5.2. Fluorescence

Fluorescence and its applications in fluorescence microscopy have seen two decades of rapid evolution, with new equipment, techniques and probes available (Combs, 2010). Fluorescence commences with the absorbance of light and is followed nanoseconds later with emission of light that is typically of a longer wavelength. The difference between the absorption and emission wavelengths is called the Stokes shift. This is a critical component of fluorescent techniques; the emission photons can be separated from the excitation photons using filters such that we only observe the fluorescent objects.

Fluorescence first requires objects to possess the ability to fluoresce, such molecules that are used by virtue of their fluorescent properties are called fluorophores and are referred to as such throughout this thesis. The electrons in a molecule are separated by different energy levels as a consequence of the electrostatic field from the nuclei. These energy levels are further sub-divided as a consequence of vibrational and rotational modes accessible to the molecule.

Fluorophores in the *ground-state* (no excited electrons) can absorb photons that induce alter-

ations in the electronic, vibrational and rotational states of the fluorophore. Given a photon that matches the difference in energy between two states, the absorbed energy can excite a system such that an electron occupies a higher energy orbital at a timescale on the order of femtoseconds. Once the system is in the *excited state* the absorbed energy is shed through multiple pathways of vibrational relaxation and fluorescence emission as the fluorophore returns to its *ground-state*. Typically, an electron will drop into the lowest energy vibrational band through vibrational relaxation in a matter of picoseconds. After about a nanosecond, a photon is emitted and the electron will drop into a vibrational band of the *ground state* before repeating the vibrational relaxation to reach the lowest vibrational mode such that the fluorophore is in its ground state. The variation in relaxation modes and the outermost electron orbitals of the fluorophore define fluorescent properties including the quantum efficiency and the characteristic absorption and emission spectrums.

A useful approach to understand the excitation and emission process is using a Jablonski diagram (Figure 5.3). However, for a more in-depth and rigorous treatment there are several excellent books (Valeur, 2002; Lakowicz, 1999). The Jablonski diagrams in Figure 5.3 present the ground state (S_0) and two excited singlet states (S_1 and S_2) with the timescales of the various excitation and relaxation modes.

The Jablonski diagram also includes an alternative to fluorescence known as internal conversion. Internal conversion is the direct movement of the system between excited states (picosecond), occurring much faster than fluorescent emission.

Another significant relaxation pathway is through inter-system crossing, also presented in Figure 5.3b. This is a ‘forbidden transition’ where the electron moves from a singlet state into a triplet state whilst undergoing a change in spin state. Whilst in the triplet state, the electron may emit a photon through phosphorescence (microseconds) or undergo inter-system crossing bringing the fluorophore into one of the vibrational modes of the ground state. These processes are important to consider during fluorescence correlation spectroscopy measurements.

5.2.1. Förster Resonance Energy Transfer

Förster resonance energy transfer (FRET) is a non-radiative energy transfer between two dye molecules - a donor and acceptor. The extent of non-radiative energy transfer is dependent upon the proximity of the fluorophores and, therefore, gives a measure of distance between them that can be estimated from the ratio of acceptor intensity to total emission intensity.

The quantitative relationship between the probability of energy transfer (FRET efficiency) and the inter-dye distance was first developed by Theodore Förster (Förster et al., 1939). According to Förster’s theory, the transfer efficiency, E , is inversely dependent on the sixth power of the inter-dye distance, r ,

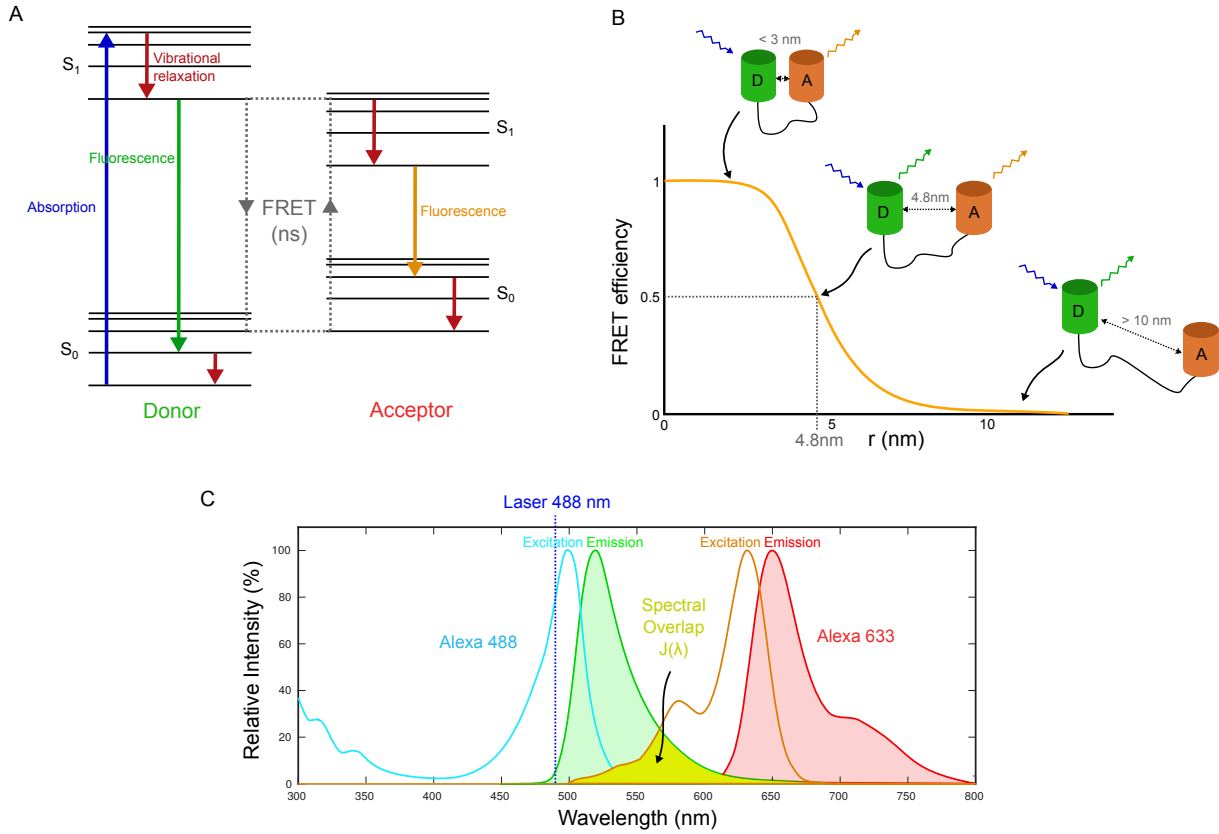


Figure 5.4: The basic principles of FRET as a consequence of the spectral overlap between fluorophores. (a) A Jablonski diagram explaining the effect of FRET when the acceptor and donor fluorophores are within close proximity. (b) The FRET efficiency is dependent on the inverse sixth order of the distance. (c) The overlap of Alexa 488 emission and Alexa 633 absorption spectrums allow for FRET.

$$E = \frac{R_0^6}{R_0^6 + r^6}, \quad (5.2)$$

where R_0 is the Förster radius, a characteristic distance corresponding to a 50% probability of energy transfer. As a result of the strong distance dependence, FRET can be used as a ‘spectroscopic ruler’ across the 2 - 10 nm range. Calculating R_0 is not simple, it is highly dependent on a variety of molecular and geometric properties,

$$R_0^6 = \frac{9000 \ln(10) \kappa^2 Q_D J}{128 \pi^5 n^4 N_A} \quad (5.3)$$

where Q_D is quantum yield of the donor, J is the spectral overlap integral and n is the refractive index of the inter-dye solution. The orientational factor is dependent on the angle between the donor emission and acceptor absorbance transition dipoles, Θ_T , and the angles between the donor-acceptor connection line, Θ_D and the donor emission and acceptor absorption transition moments, Θ_A .

Table 5.1: The quantum yields of the dyes bound to ADK were determined relative to the quantum efficiencies of free dye by [Henzler-Wildman et al. \(2007b\)](#)

Binding Spot	Fluorophore	Quantum yield (Y)
LID	Alexa 488	0.77
	Alexa 633	0.53
AMPbind	Alexa 488	0.74
	Alexa 633	0.55

$$\kappa^2 = (\cos(\Theta_T) - 3\cos(\Theta_D)\cos(\Theta_A))^2 \quad (5.4)$$

Whilst the value of κ^2 can vary anywhere between 0-4 in reality $\kappa^2 = 2/3$ is most frequently used in practice since it assumes a complete averaging of the relative orientation of the chromophores during the donor excitation lifetime. A value of $\kappa^2 = 1$ corresponds to parallel fluorophores, a value of $\kappa^2 = 0$ corresponds to orthogonal fluorophores and a value of $\kappa^2 = 4$ corresponds to collinear fluorophores.

The FRET pair Alexa 488 (donor) and Alexa 633 (acceptor) were used throughout these studies owing to their high photostability and quantum yield. The fluorescence quantum yields of the dye bound to ADK has been determined in a previous study, Table 5.1 ([Henzler-Wildman et al., 2007b](#)). Using Equ. 5.3 we can determine the Förster radius as $R_0 = 48 \text{ Å}$ in water. The spectral overlap J was computed from the emission and absorption spectrums of the donor and acceptor fluorophores respectively,

$$J = \int_0^\infty F_D(\lambda)\epsilon_A(\lambda)\lambda^4 d\lambda \quad (5.5)$$

and subtle changes in the site dependence of the quantum yields made a negligible difference to R_0 owing to the inverse sixth power dependence.

5.3. Circular Dichroism

Circular dichroism (CD) has proved itself as an excellent technique to explore the molecular stability of proteins as a function of temperature. It uses the difference in absorbance of left-handed and right-handed circularly polarised light and gives a measurement in degrees of ellipticity. The highly asymmetric secondary structural elements of folded proteins absorb the circularly polarised light making it a very sensitive technique to probe the structure of proteins ([Greenfield and Fasman, 1969](#)).

Left and right circularly-polarised (CP) light will travel through a medium at different velocities

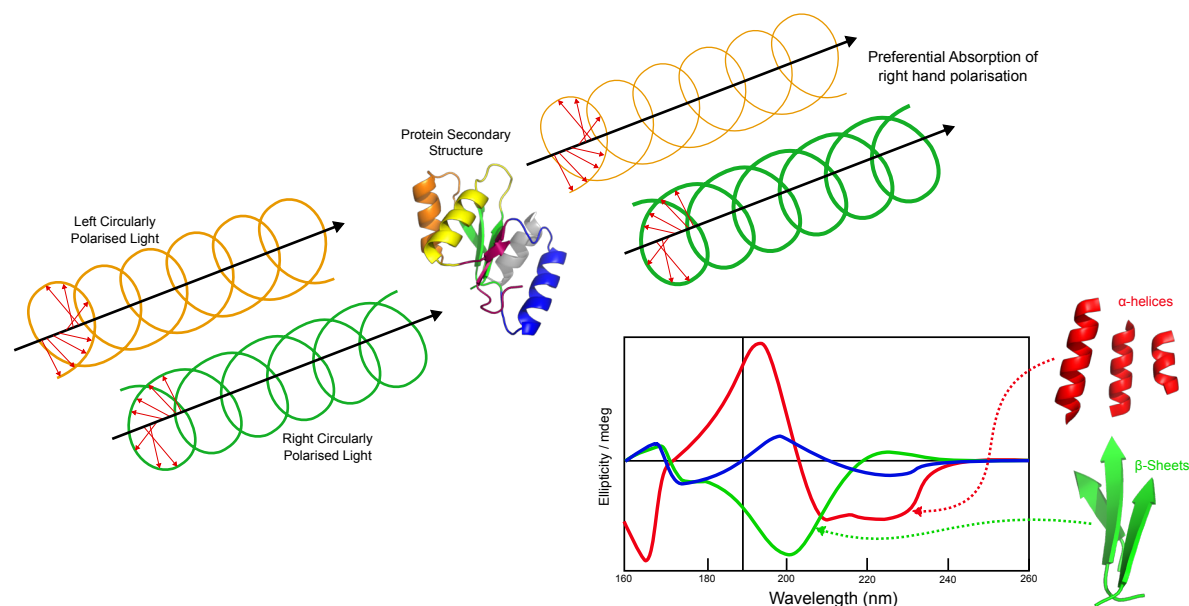


Figure 5.5: A theoretical schematic of circular dichroism in protein systems and the characteristic ellipticity spectra produced in practical experiments. When circularly polarised light passes through an optically active medium (such as a protein double helix) there is a difference in the absorbance of the left and right polarised light. The different secondary structure conformations contribute their own characteristic circular dichroism spectrum.

due to circular birefringence (differing refractive indices for left and right CP light). At particular wavelengths, due to the extinction coefficients for the two polarised rays, there will be differing absorption of the left and right CP light called *circular dichroism*.

5.4. Analytical Methods

5.4.1. Single-molecule fluorescence analytics

Single-molecule spectroscopy is an integral part of nano-biotechnology and biophysical research, addressing the previously unanswered questions pertaining to the structure (Roy et al., 2008), dynamics (Borgia et al., 2008) and mechanisms (König et al., 2015) of biological systems. Single-molecule techniques offer the ability to avoid ensemble averaging which is common in the majority of biochemical techniques. Instead of observing the average across a large number of molecules, information from each molecule can be separately recorded. Allowing a direct measure of each proteins properties.

This section describes single-molecule FRET (smFRET) and fluorescence correlation spectroscopy (FCS) which are used in this study. Both single-molecule methods use the confocal

microscopy setup described in Section 5.1.5 and illustrated in Figure 5.2.

Single-molecule FRET

Single-molecule FRET (smFRET) is a general and adaptable single-molecule technique. Its rapid development and widespread use has helped answer fundamental questions about transcription, translation, protein folding, motor proteins and signal transduction to name but a few. For further reading about implementation, Roy et al. (2008) have put together a short practical guide to smFRET.

SmFRET measurements were performed on the confocal microscope described in Section 5.1.5. Samples were aliquotted for FRET at 50 pM in 200 μ l of pH 7.5 FRET buffer (20 mM TRIS, 50 mM NaCl) with 0.3 mg/ml BSA to prevent surface adsorption. The low concentration results in a negligible probability of two fluorescent molecules residing in the confocal volume at the same time.

FRET efficiency histograms

Fluorescence bursts from individual molecules were identified by combining successive photons that were separated by inter-photon times less than 500 μ s and examining the total number of photons. A threshold of 30 - 50 counts per 500 μ s bin for the sum of the donor and acceptor fluorescence signals was used to differentiate single molecule bursts from the background, depending on the background intensity. Background measurements were calculated using independent measurements of buffer solutions without labeled samples, these were subsequently subtracted from each burst. Background measurements were found to be $\sim$ 0.6 - 0.8 counts and $\sim$ 1.1 - 1.3 counts per 500 μ s in the donor and acceptor channels respectively.

Apparent FRET efficiencies, E_{app} , of each burst were calculated according to $E_{app} = n_A / (n_A + n_D)$, where n_A and n_D are the acceptor and donor photon counts corrected for background respectively. Direct excitation of the acceptor was considered but relatively negligible comparatively. In some cases the wavelength of a photon from a fluorophore is at the extremity of its spectrum and is counted as a photon from its FRET partner instead, this is called cross-talk and introduces a value of γ into the apparent FRET efficiency calculation. In these experiments we assumed γ to be negligibly small given the emission spectrums of Alexa488 and Alexa633.

The individual FRET efficiency bursts were histogrammed to reveal FRET efficiency sub-populations (Figure 5.2). It is important to note that owing to these corrections, E_{app} values of < 0 and > 1 are possible on the FRET efficiency histograms.

The random fluorophore labelling resulted in 3 labelled species: (i) a protein with an acceptor and donor, (ii) and protein with two donors and (iii) and protein with two acceptors. The

two acceptor species would not be excited by the laser, however, the donor-donor species would introduce a zero FRET efficiency peak.

For an acceptor and donor pair at a fixed distance the resulting FRET efficiency histogram is broadened by shot noise which is the variation of count rates around their mean value to the discrete nature of photon counting. However, there are other contributions to distribution broadening which can be difficult to establish such as variations in quantum efficiencies, imperfect alignment of donor and acceptor channels and labelling permutations (Schuler, 2013; Gopich and Szabo, 2007).

Fitting of FRET efficiency histograms

To analyse subpopulations within the FRET efficiency histograms we approximated peaks with Gaussian (G) and four-parameter log-normal (L) probability distributions for symmetric and asymmetric peaks, respectively.

$$G(E) = Ae^{-\frac{(E-E_0)^2}{2w^2}} \quad \text{and} \quad L(E) = A \exp \left(-\frac{\ln(2)}{\ln(a)^2} \ln \left(1 + \frac{a^2 - 1}{aw} (E - E_0) \right) \right) \quad (5.6)$$

where A is the peak amplitude, E is the FRET efficiency, E_0 is the peak position (mean), w is the peak width and a is the asymmetry factor of the peak (Schuler et al., 2002). When fitting multiple peaks, the histogram was analysed as a sum of peak functions. Where necessary, the choice of fixed parameter values were based on the apo, substrate bound and osmolyte driven systems where the respective peaks were as best defined as possible.

Log-normal peak functions were used primarily for the donor only population with a fixed width ($w = 0.15$) and asymmetry ($a = 1.2$), however, high FRET efficiency peaks ($E_0 > 0.9$) were similarly modelled with a log-normal peak. The remaining FRET-labelled population was fitted with a Gaussian peak function.

Adenylate Kinase is known to be a two-state system, therefore, a two peak distribution (+ the log-normal zero peak) is assumed. The peak positions were identified using the apo, substrate-bound and TMAO osmolyte perturbed structures. The mean and standard deviation was then fixed across all mutant species of the same FRET pair.

However, in particular cases the number of distribution components was confirmed using the Akaike and Bayesian information criterion (AIC & BIC). By fitting the FRET efficiency histograms with a varying number of components we can compare the model quality without over-fitting the data. The fit will always improve as you increase the model order, however, the multiple components may have no physical basis. The AIC/BIC were determined using the residual sum of squares (RSS),

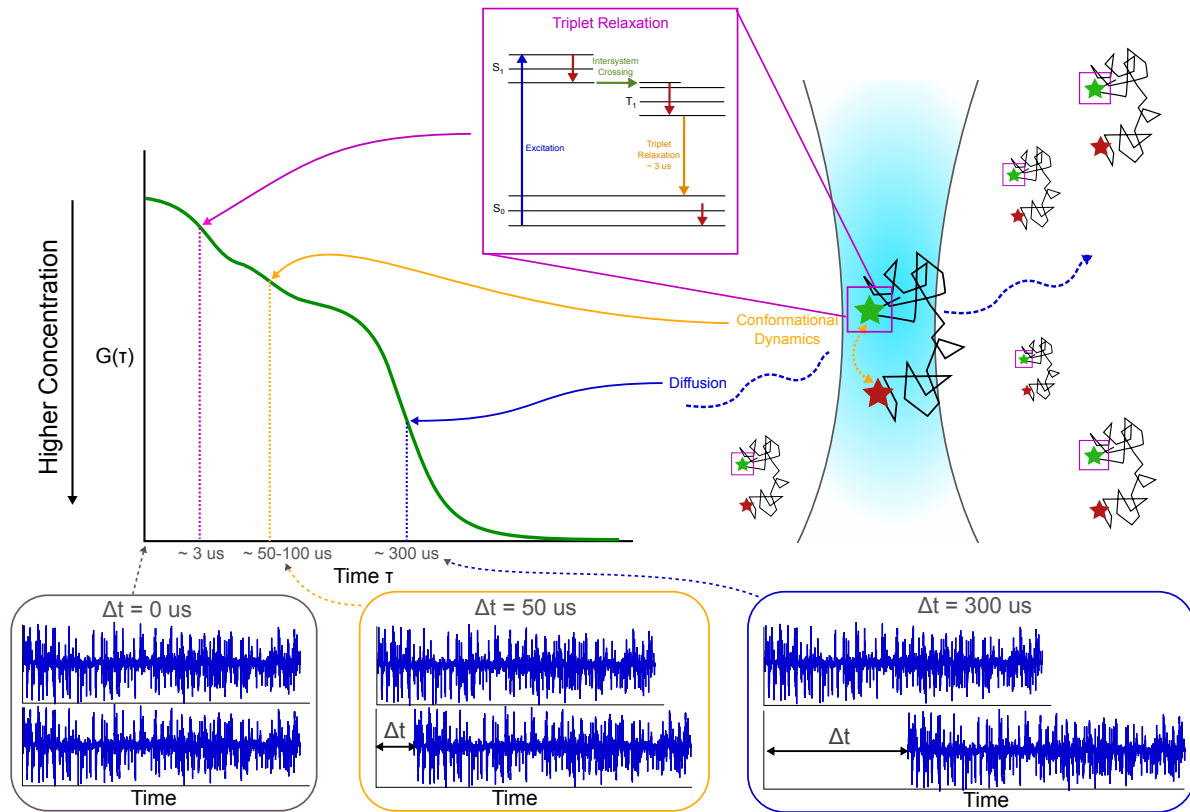


Figure 5.6: Schematic diagram of fluorescence correlation spectroscopy. The obtained fluctuations of fluorescence intensity are autocorrelated. Timescales of correlation appear that correspond to various physical kinetics such as diffusion, conformational dynamics and triplet relaxation.

$$AIC = N + N \log(2\pi) + N \log \left(\frac{RSS}{N} \right) + 2(K + 1) \quad (5.7)$$

$$BIC = N + N \log(2\pi) + N \log \left(\frac{RSS}{N} \right) + \log(N)(K + 1) \quad (5.8)$$

where N is the number of values in the data set and K is the number of model parameters.

Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) is a powerful method for studying kinetic processes across all temporal scales through the statistical analysis of equilibrium fluorescence intensity fluctuations. Since its invention by [Magde et al. \(1972\)](#), FCS has seen a wide range of applications including binding-unbinding reactions ([Icenogle and Elson, 1983](#)), translational diffusion in solutions ([Elson and Magde, 1974](#)) and membranes ([Fahey et al., 1977](#)) and molecular aggregation ([Thompson and Palmer, 2001](#)) to name but a few.

Accordingly, FCS is a robust method for detecting conformational dynamics in proteins. FCS is able to reveal fluctuations in fluorescence intensity of single-molecules as they pass through the confocal volume (Figure 5.6). These intensity fluctuations can be attributed to processes such as triplet kinetics, molecular diffusion and fluorescence quenching (Widengren et al., 1994). Previous studies have been able to detect microsecond dynamics using FCS techniques (Oikawa et al., 2013; Chung et al., 2012). In this research we explore the fluctuations of fluorescence intensity resulting from the kinetics of Förster resonance energy transfer as a result of conformational transitions using the confocal microscope setup.

The fluorescence fluctuation data appears to be a noisy mess, however, a statistical analysis using an autocorrelation function is able to extract coefficients that correspond to various kinetics (for a detailed review of the method and its applications see the review by Elson (2011)). Autocorrelation is a mathematical tool used to assess the similarity of a signals at times t and τ . The autocorrelation $G(\tau)$ can be represented as

$$G(\tau) = \frac{\langle \delta I(t) \delta(t + \tau) \rangle}{\langle I(t) \rangle^2} = \frac{\langle I(t) I(t + \tau) \rangle}{\langle I(t) \rangle^2} - 1, \quad (5.9)$$

where $I(t)$ and $I(t + \tau)$ are the photon counts at time t and $t + \tau$ respectively. The autocorrelation traces span all timescales (Figure 5.6) and features of the trace correspond to dynamical behaviours at particular timescales. The different timescales of the autocorrelation curve correspond to different dynamical behaviours in the system. In a scenario where only one APD is available, only the diffusion component could be measured (10^{-2} ms) due to the dead time of the detectors. When two APD's are available the fluorophore behaviour can be measured, such as triplet state relaxation (10^{-4} ms). In this setup, the signals between the two APD's are pseudo-autocorrelated (correlation of two similar signals). Without two APD's the dead time of the APD's would correlate and introduce a very steep component at 'triplet' timescales ($\sim 10^{-4}$ ms).

Due to the stochastic nature of diffusion, the autocorrelation signal from different molecules diffusing through the confocal volume may or may not 'line up' when comparing the signal with itself at long timescales. The rise in correlation observed at shorter timescales (Figure 5.6) is due to the signal from the same molecule aligning with itself. Each molecule will spend a differing length of time within the confocal volume. Some molecules will spend a long time in the confocal volume and increase $G(\tau)$ at high values of τ , whilst some molecules that spend less time within the confocal volume contribute to $G(\tau)$ at shorter timescales. At very high and very short values of τ , relative to the diffusion time, then no further molecules will contribute to $G(\tau)$. This produces the characteristic 2-shape where the position is dependent on the diffusion time of the molecules. Slow diffusion will shift the curve towards a longer τ and fast diffusion vice-versa. The amplitude of the 2-shape curve is inversely proportional to the concentration

of fluorescent molecules.

Autocorrelation can be calculated by way of Fourier transform $G(\tau) = I(\tau) \star I(\tau) = F^{-1}[F(I(\tau))^2]$ and, therefore, suffers from similar problems including the loss of temporal information.

Fitting the autocorrelation curves is not a simple task. The diffusion component can be fitted with,

$$G_{2D}(\tau) = \frac{\gamma}{N(1 + \frac{\tau}{\tau_D})} + C, \quad (5.10)$$

where γ is the shape factor, τ_D is the average time taken for a molecule to traverse the radial dimension of the confocal volume and N is the average number of molecules in the focus (Unruh, 2017). This diffusion assumes a 2-dimensional plane, whereas in reality a confocal volume is 3-dimensional. However, there is little signal fluctuation associated with movement along length of the confocal volume (approximately 1 μm) which is aligned with the beam axis.

The amplitude of the correlation function is defined at $\tau = 0$ where $G(0) = \gamma/N$. This is a useful estimate for the concentration of molecules (C),

$$C = \frac{N}{N_A \times V} \approx \frac{1}{G(0) \times N_A \times V} \quad (5.11)$$

where V is the focal volume (approximately 1 fl) and N_A is Avogadro's number. The diffusion time can be calculated as,

$$\tau_D = \frac{w_0^2}{4D}, \quad \text{where} \quad w_0 = \frac{\text{FWHM}}{\sqrt{2\ln(2)}}, \quad (5.12)$$

w_0 is the beam waist, D is diffusion coefficient and FWHM is the diameter of the confocal volume at half of the maximal intensity (full width at half maximum) (Unruh, 2017).

When there is triplet state behaviour a new component is introduced, and $G(\tau)$ is modified to,

$$G(\tau) = \left(\frac{G_1(0)}{(1 + \frac{\tau}{\tau_1})} \right) (1 - F + F e^{\frac{-\tau}{t_m}}) + C, \quad (5.13)$$

where t_m is the mean triplet state lifetime and F is the fraction of molecules in the dark state (Schwille and Haustein, 2009). Triplet state behaviour occurs typically in the 10^{-4} ms to 10^{-2} ms temporal scale. A higher laser intensity increases the probability of inter-system crossing by increasing the cycling rate of the fluorophore, this results in a larger contribution (amplitude)

from the triplet dynamics to the FCS curve. It must be noted that increasing the excitation intensity can also photobleach fluorophores and, therefore, decreases the apparent diffusion time of molecules.

In a similar fashion, we can introduce a third component which describes the conformational dynamics associated with FRET. Therefore, we can finally fit the autocorrelation curves with a global model that includes components for triplet excitation, conformational dynamics and diffusion, with the assumption that they differed by a factor of 1.6 to distinguish the components (Meseth et al., 1999),

$$G(\tau) = G(0) \left(\frac{1}{1 + \frac{\tau}{\tau_D}} \right) \left(1 - F + F e^{-\frac{\tau}{\tau_m}} \right) \left(1 - F_2 + F_2 e^{-\frac{\tau}{\tau_{conf}}} \right), \quad (5.14)$$

where τ_c , τ_m and τ_D are the dynamical timescales of the protein conformational dynamics, mean triplet relaxation and the protein diffusion respectively. F_1 is the fraction of molecules entering the triplet state and F_2 is the fraction of molecules conformationally fluctuating.

Samples were diluted to 200 pM in pH 7.5 FRET buffer (20 mM TRIS, 50mM NaCl) with 0.3 mg/ml BSA to prevent surface adsorption. Relative to the smFRET measurements, a slightly higher concentration increases the probability of two fluorescent molecules residing in the confocal volume at the same time. Such that a constant emission signal is present. Measurements were taken at thermal equilibrium such that all processes under analysis are statistical fluctuations around the equilibrium.

5.4.2. Melting temperatures curves from circular dichroism

There are two generally accepted methods for quantifying circular dichroism. The first is to calculate a differential absorbance between the right and left CP light:

$$\Delta A = A_l - A_r \quad (5.15)$$

The second method is to find the tangent of the ratio of the minor to major elliptical axes to express the results in degrees of ellipticity:

$$\theta = \tan \left(\frac{\text{minor}}{\text{major}} \right) \quad (5.16)$$

In fact, there is a simple linear relationship between the differential absorbance and ellipticity:

$$\Delta A = \frac{\theta}{32.982} \quad (5.17)$$

The protein secondary structure can be partially deconvoluted from the CD spectrum since β -sheets and α -helices each contribute their own ‘fingerprint’ CD spectrum to the protein characteristic CD spectrum (Figure 5.5). However, more recently it has become evident that the CD spectrum is not solely composed of secondary structures, e.g. residues close in space but not inside a secondary structure (Khrapunov, 2009).

Despite these limitations, CD is a useful method for analysing the secondary structure of proteins. In particular, calculating the ellipticity at a single wavelength across an increasing temperature range can provide a measure of melting temperature. Protein unfolding will invoke a large change in absorption of CP light. Since the α -helix composition is expected to unfold during denaturation we can follow the characteristic α -helix wavelength peak at 222 nm. Using a suitable mathematical model we can calculate thermodynamic parameters including the melting temperature (also known as the thermal denaturation midpoint). The melting temperature is the temperature at which 50% of the protein sample has denatured and is a widely used metric for molecular stability (Greenfield, 2007).

There are many available mathematical methods for determining the midpoint of unfolding T_M (Greenfield, 2009) as a function of temperature given the elliptical spectrum. Two particular thermodynamic models were used to analyse the CD melting curves in this thesis, they both assume a two-state transition of a monomer from a folded to unfolded form:

1. Assumes a change in heat capacity, ΔC_p , of the folded and unfolded states.

The dependence of Gibbs free energy ($\Delta G(t)$) on temperature can be established from the temperature dependence of entropy and enthalpy,

$$\Delta H(T) = \Delta H(T_m) + \int_{T_m}^T \Delta c_p dT = \Delta H(T_m) + \Delta c_p(T - T_m), \quad (5.18)$$

$$\Delta S(T) = \frac{\Delta H(T_m)}{T_m} + \int_{T_m}^T \Delta c_p d \ln(T) = \frac{\Delta H(T_m)}{T_m} + \Delta c_p \ln\left(\frac{T}{T_m}\right), \quad (5.19)$$

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) = \Delta H\left(1 - \frac{T}{T_m}\right) - \Delta c_p\left(T_m - T + T \ln \frac{T}{T_m}\right), \quad (5.20)$$

where T is temperature (in Kelvin), T_m is the melting temperature to optimise, $\Delta H(T)$ and $\Delta H(T_m)$ are the enthalpies at temperature T and at the melting temperature and $\Delta S(T)$ is the entropy. The most important inclusion is the change in heat capacity ΔC_p

between the folded and unfolded states. An equilibrium constant between the folded and unfolded states can be calculated as,

$$K = \exp\left(\frac{-\Delta G}{RT}\right), \quad (5.21)$$

such that the dependence of ellipticity on temperature can be calculated as,

$$[\theta](T) = \frac{K}{1+K}([\theta]_f - [\theta]_u) + [\theta]_u, \quad (5.22)$$

where $[\theta]_f$ and $[\theta]_u$ are the ellipticity of the unfolded and folded protein respectively. The melting temperature is then defined at $K = 1$.

2. **Introduces corrections for the pre- and post- transition linear changes to ellipticity.** A more simple method is to use the ellipticity calculated as a function of temperature,

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) = \Delta H\left(1 - \frac{T}{T_m}\right), \quad (5.23)$$

and, to correct for the assumed constant heat capacities, we introduce a pre- and post-transition linear component,

$$[\theta](T) = \frac{K}{1+K}([\theta]_f + C_f T - [\theta]_u + C_u T) + ([\theta]_u + C_u T), \quad (5.24)$$

where C_u and C_f are the linear corrections to ellipticity for the unfolded and folded states.

Measurements were performed using a Chirascan CD spectrometer (Photophysics, Surrey, UK) and its associated Chirascan software. The sample temperature was controlled using a MELCOR MTCA-9060 thermoelectric temperature controller combined with a Julabo AWC100 circulating water bath (JULABO GmbH, Seelbach, Germany).

Protein samples were prepared at 0.1mg/ml in 20mM phosphate buffer which was a reasonable trade-off between a higher CD signal strength, associated with higher protein concentrations, and the amount of light that can pass through the sample, which reduces with protein concentration. A total volume of 250ul protein sample is measured in a high quality quartz cuvette with a 1mm path length.

For melting temperature experiments, the monochromator wavelength was set at 222nm with a bandwidth of 1nm. The temperature was ramped from 20°C to 90°C in 2°C increments and the ellipticity was recorded 5 times at each temperature. Due to the high melting temperature of Aquifex ADK (109°C ([Henzler-Wildman et al., 2007b](#))) it is impossible to observe melting at

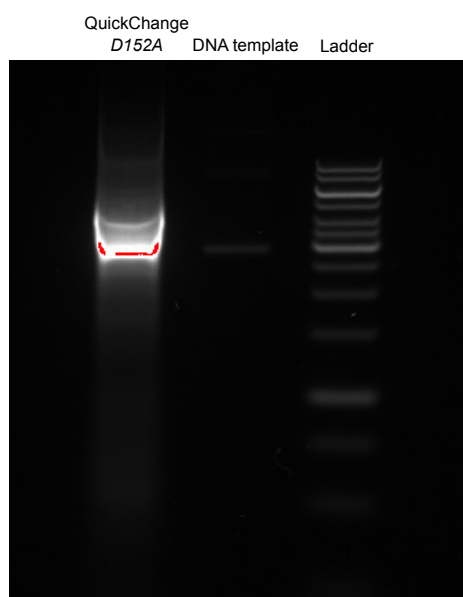


Figure 5.7: DNA gel electrophoresis of Quickchange mutagenesis. A comparison of the DNA produced directly after Quickchange mutagenesis with the DNA template (wildtype) and a DNA ladder after GelRed staining. If mutagenesis is successful there will be a bright band of similar length to the template DNA.

atmospheric pressures. Protein was eluted in 5M guanadine to lower the melting temperature into an instrumentally-accessible range.

5.5. Materials and Sample Preparation

5.5.1. *Aquifex* Adenylate Kinase construction

DNA Samples

Plasmids of *Aquifex* Adenylate Kinase (ID:18092 Plasmid:peT3a-AqAdk/MVGDH) were purchased from AddGene as deposited by ‘Dorothee Kern Lab Plasmids’¹. The plasmids were already encoded with two cysteine mutations for maleimide binding. The techniques described herein were applied to peT3a-AqAdk/MVGDH.

For mutant ADK species, DNA was mutated using Quickchange mutagenesis (Section 5.1.1) with 35-45 bp primers purchased from ThermoFisher Scientific. Successful PCR was checked with DNA gel electrophoresis (Section 5.1.1) to identify a DNA band as shown in Figure 5.7. Successful mutations were confirmed using DNA sequencing with Beckmann Genomics (now GeneWiz, NJ, US).

¹https://www.addgene.org/Dorothee_Kern/

Wildtype and mutant DNA was transformed into BL21 (DE3) cells using heat shock transformation (Section 5.1.1) and plated on LB-agar + 100 ug/ml Ampicillin plates overnight. A 50 ml overnight culture was grown from single colonies and the gene was re-sequenced to confirm that there were no unwanted mutations. Successful DNA mutations were labelled appropriately and stored at -20°C.

Protein extraction and expression

A volume of 10 ml of the 50 ml overnight cultures were used to inoculate a 1 litre culture of BL21 (DE3) cells containing the appropriate ADK gene. The cells were grown to 0.6 OD₆₀₀ after which ADK was expressed via a 1 mM IPTG (Isopropyl β -D-1-thiogalactopyranoside) induction. After 4 hours of expression the cells were harvested from the liquid culture by centrifugation at $10,000 \times g$ for 15 minutes. The solution is decanted and the pellet is drained to remove excess liquid. After measuring the weight of the cell pellet, 5ml of Bugbuster is added per gram of wet cell paste to re-suspend and lyse the cells.

1 mM TCEP (Tris(2-carboxyethyl)phosphine hydrochloride powder) and a protease inhibitor cocktail (for use in purification of Histidine-tagged proteins, DMSO solution) was added to the lysate to reduce oxidised cysteines and prevent peptide cleaving respectively. The cell suspension is incubated at room temperature on a shaking platform for 15 minutes. Insoluble cell debris is removed by centrifugation of the lysed cells at $16,000 \times g$ for 20 minutes at 4°C. The supernatant (10-15 ml) is subsequently transferred to a fresh tube for purification and further TCEP and protease inhibitor were added.

Purification

ADK was separated from unwanted proteins via a HIS-tag purification with a gravi-trap (GE-healthcare, Little Chalfont, USA). The protective tip of the gravi-trap column was removed and the excess liquid was poured away, after which the gravi-trap column was placed on a stand in a vertical manner. The column was equilibrated with 10 ml of 50 mM imidazole binding buffer (5.5.2), after no further solution was exiting the column the protein sample was added and allowed to run through to completion. The column was washed with 10 ml of 50 mM imidazole binding buffer to remove unwanted proteins and then a 300 mM imidazole elution buffer was used to elute ADK into a separate tube.

Imidazole acts like a salt and can be quite corrosive, therefore, buffer exchange is required. A Sephadex G-25 PD-10 desalting column (GE Healthcare) was used to remove excess imidazole and exchange into the required buffer for each particular experiment (Section 5.5.2). The cap and bottom tip are removed allowing the storage solution to be poured away. The column is

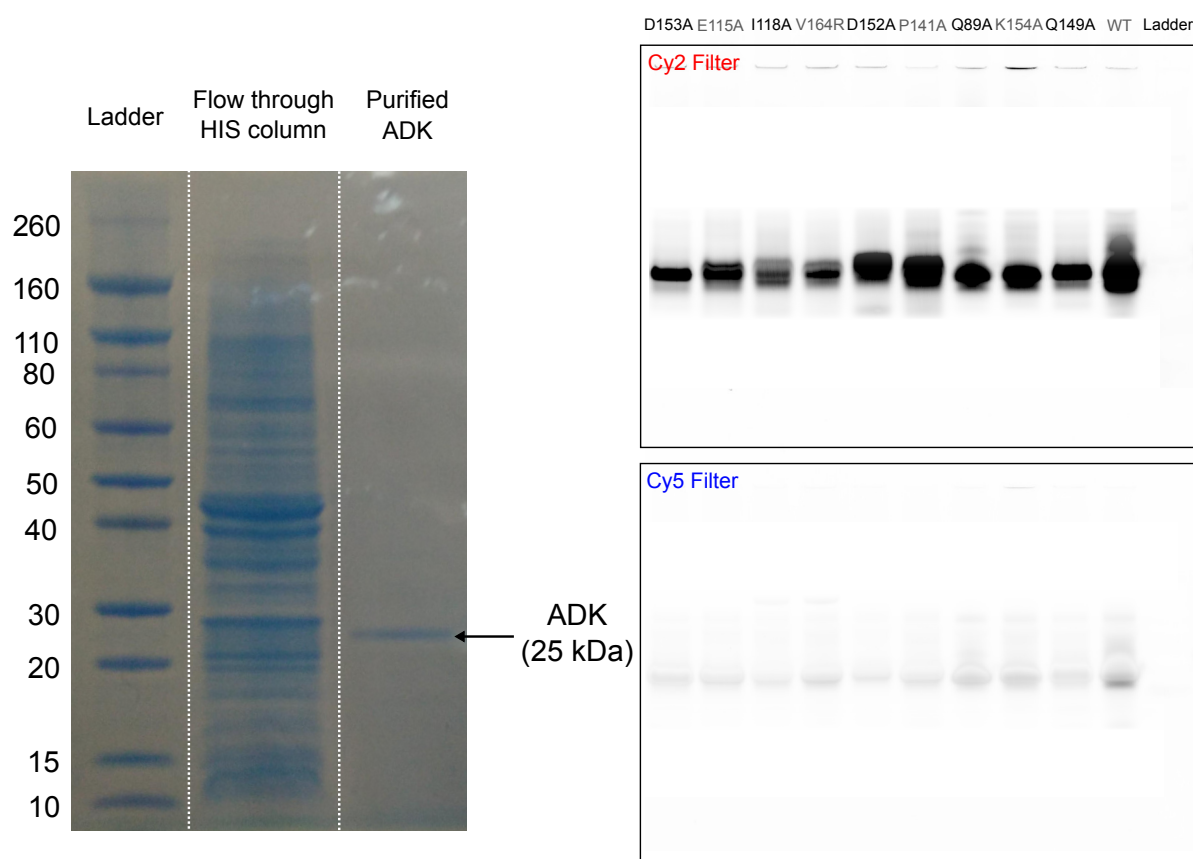


Figure 5.8: SDS page of purified and labelled proteins visualised using coomassie staining and typhoon fluorescence. SDS Page gel of the purification process of ADK, including the HIS column flow through and the purified product. Typhoon images of the labelled mutant proteins taken using 450 V PMT at a 50um resolution. Clear labelling is observed for all proteins, and no fluorescence is seen for the unlabelled protein ladder.

filled with 5 ml of the buffer chosen to suspend your final protein sample (equilibration buffer). The buffer is allowed to enter the packed gel bed completely and the process of equilibration is repeated another 4 times. The wash through is collected into a waste collection tube. A maximum of 2.5 ml of sample is added to the column and allowed to enter the gel bed completely. The waste collection tube is removed and a fresh tube is placed underneath the column, after which 3.5 ml of equilibration buffer is added to the column and the protein is eluted into the fresh tube.

After purification the concentration is measured using UV-VIS (Section 5.1.3) and a protein gel is performed (Section 5.1.2) to check the size of the protein (Figure 5.8).

Fluorophore Labelling

Alexa FluorTM 488 C₅ maleimide and Alexa FluorTM 633 C₅ maleimide are photostable and bright fluorophores with a maleimide derivative making them an excellent choice for thiol con-

jugation. They were dissolved in 1 ml of high-quality anhydrous dimethylsulfoxide (DMSO) respectively giving concentrations of 1.389 mM 0.769 mM for Alexa 488 and Alexa 633 respectively. The stocks solutions were protected from light and stored at -20°C .

Alexa 488/Alexa 633-labelled ADK was prepared using 20 μM protein with a molar ratio of 1:5:5 of protein, Alexa 488 and Alexa 633 respectively. The mixture was wrapped in aluminum foil in a tube overnight. Excess dye was removed using HIS-tag purification as described in the purification stage above. A PD-10 column was used as a further fluorophore removal step and to remove the imidazole associated with the HIS tag purification. A Typhoon FLA 9500 (GE-healthcare, Little Chalfont, USA) was used to examine the gel of the purified-labelled ADK product and showed no excess fluorophores (Figure 5.8).

5.5.2. Buffers

- (i) **FRET Buffer 1.** 20 mM TRIS, 50mM NaCl, pH 7.5 (used during purification of FRET labelled proteins).
- (ii) **FRET Buffer 2.** 20 mM TRIS, 50mM NaCl, 0.3 mg/ml BSA, pH 7.5 (used when imaging FRET labelled proteins).
- (iii) **CD Buffer 1.** 20 mM phosphate, pH 8.0 (single temperature full spectrum measurements).
- (iv) **CD Buffer 2.** 20 mM phosphate, 5M guanidine, pH 8.0 (melting temperature measurements).
- (v) **Enzymatic Assay** 20 mM TRIS, 50mM NaCl, pH 7.5.
- (vi) **Binding Buffer** 20 mM sodium phosphate, 500 mM NaCl, 50 mM imidazole, pH 7.4.
- (vii) **Elution Buffer** 20 mM sodium phosphate, 500 mM NaCl, 200 mM imidazole, pH 7.4.

6. Conformational dynamics of Aquifex ADK

The reality of proteins is far from the crystal structures that may be used to portray them, they are inherently mobile and dynamical to perform the necessary catalytic function that they evolved for. In fact, over 3/4 of eukaryotic proteomes are multi-domain proteins with large-scale inter-domain motions (Bashton and Chothia, 2007). These large-scale $ms - s$ motions are actually a manifestation of local motions that occur across the $fs - \mu s$ timescales, so how can we probe the dynamics across all these temporal scales? And how does the underlying protein structure enable these catalytically dependent dynamics? In this chapter we use Markov stability to identify the biologically relevant communities across all temporal and spatial scales in Adenylate Kinase (AdK). Adenylate Kinase provides an excellent testbed to probe protein dynamics, it boasts a core domain that is flanked by two seemingly independent dynamical subdomains and there is extensive literature available for comparison.

Markov stability calculations found that the AMP_{bind} domain appeared at earlier Markov times than the lid domain, suggesting that the AMP_{bind} would exhibit a higher rate of motion. Using FCS to measure the time-scale at which these subdomains fluctuated physically, we were able to show that the AMP_{bind} domain inter-converts between open and closed at a higher rate than the lid, revealing a relationship between Markov time and real time.

Literature studies have shown that introducing substrate results in a closed bound conformation (Henzler-Wildman et al., 2007b). We find that Markov stability calculations in the presence of substrate suggest that the communities corresponding to subdomains become less modular. This agreed with smFRET measurements that indicated a shift in the population equilibrium towards a closed conformation.

6.1. Using Markov stability to identify biologically relevant communities

As described in Chapter 4, Markov stability acts as a zooming lens that scans the temporal and spatial scales of an atomic resolution protein structure in search of robust community

structures. Such communities are deemed to possess biological relevance; groups of atoms that behave coherently at a particular timescale under the conditions of a dynamical diffusive process. Here we calculate Markov stability on *Aquifex Aeolicus* ADK in the open (PDB ID: 2RH5) and closed states (PDB ID: 2RGX). The results of a full Markov stability calculation over Markov time $t = 10^{-3} - 10^5$ are summarised in Figure 6.1. As Markov time increases the communities become larger and the partitions become coarser. This is consistent with the interpretation of Markov stability as a diffusive process acting on the protein structure.

Figure 6.1 is the first time we introduce a Markov Stability graph and since it forms an integral part of this thesis we will provide some intuition to understanding the graph. The top panel exhibits the number of communities on the left axis which corresponds to the discrete value trajectory which looks like a set of steps. The number of communities decreases with time. A plateau in the number of communities suggests that a particular community structure is robust in time and may be biophysically meaningful. On the right axis of the top panel we see the Stability metric which is a measure of the strength of a community structure, this is a continuous value which corresponds to the gradually decreasing curves. A higher value of Stability corresponds to a higher probability of the random-walker being trapped inside a community. The bottom panel exhibits the variation of information (VI) which tells us how often a particular community structure appears every time we attempt to find the optimal community structure at a single time point. If a particular community structure appears often at a single value of t then the variation of information is low and we can assume there is some importance to that community structure. Therefore, a dip in the variation of information is deemed to be biophysically relevant.

Plateaus in the number of communities and dips in the variation of information, indicative of robust partitions, are found at short Markov times $t = 2 \times 10^{-3}$ and $t = 10^{-1}$. These partitions correspond to chemical groups and amino acids respectively. Zooming out further we find a plateau in the number of communities at $t = 100$ which is accompanied by a drop in the variation of information. This partition corresponds to the secondary structures, such as α helices and β sheets, which form their own respective communities. The identification of these biologically relevant groups has been observed in every protein structure analysed thus far (Delmotte et al., 2011; Delmotte, 2014; Amor et al., 2014), confirming that we are exploring the expected structure at short and medium Markov times.

At short and intermediate timescales the partitions found in the open and closed structures are similar. This is a result of identical (covalent bonds) and similar (backbone inter-secondary structure) interactions at a local level. Beyond Markov time $t = 100$ the results of the separate structures begin to diverge as a result of the disparate weak interactions (Delmotte, 2014). Broadly, the value of Markov stability is higher for the open conformation, indicating a stronger community structure relative to the closed conformation. Moreover, the value of $\overline{\text{VI}}$ is generally higher for the closed conformation than the open conformation. Since the presence of a stronger

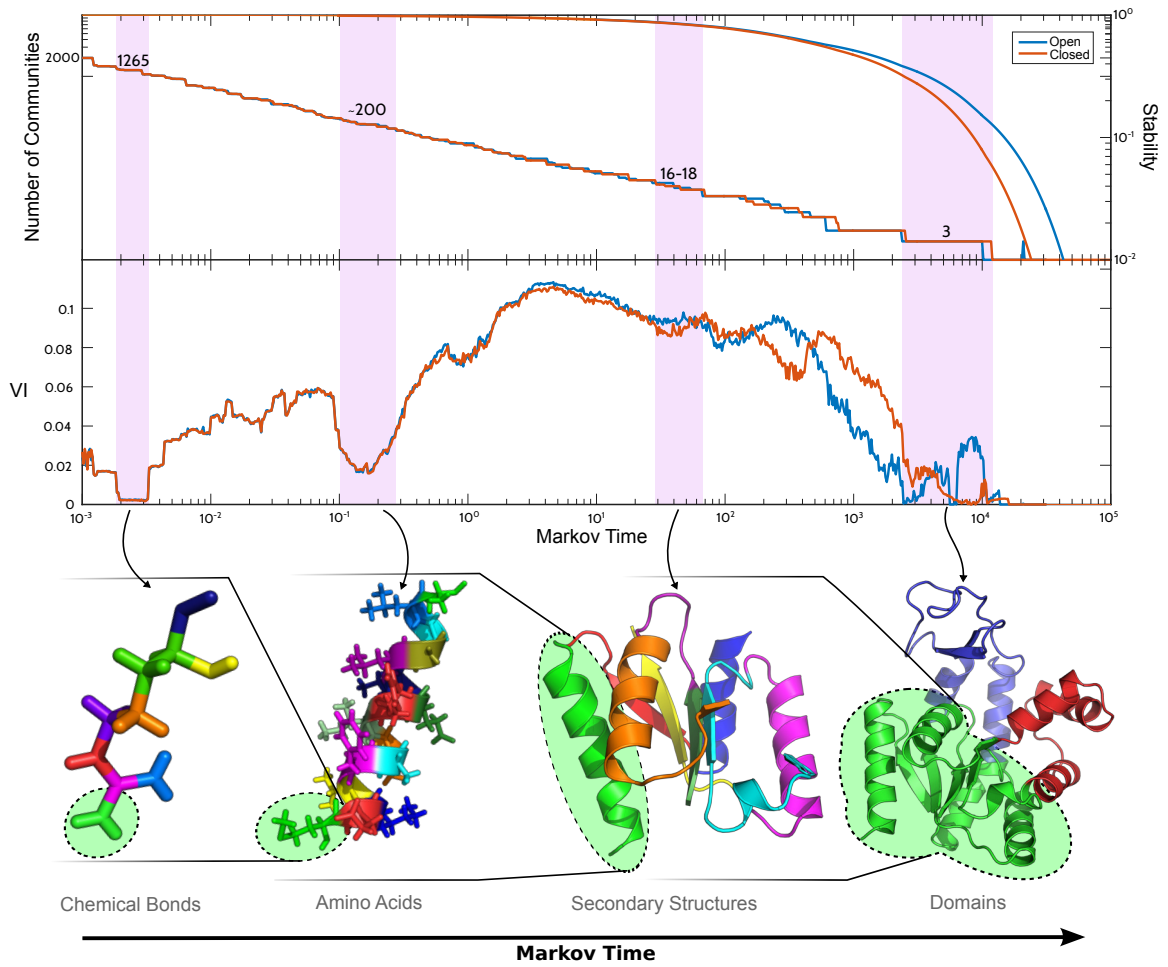


Figure 6.1: Markov stability results for the open (blue) and closed (orange) conformations of *Aquifex* ADK. Dips in the variation of information are found at Markov times 200-300 and $1 \times 10^{-1} - 2 \times 10^{-1}$ that correspond to chemical bonds and amino acids respectively. Another dip is found at 4×10^1 which corresponds to a partition that reveals the secondary structures such as alpha helices and beta sheets. Beyond 10^2 the results for the two conformations start to diverge in both their $\overline{VI}$ and their number of communities, reflecting the differences in their tertiary/quaternary structures. The general Markov stability of the open conformation is higher than the closed conformation, highlighting the robust and modular partition associated with the open conformation.

community structure is associated with a higher probability of trapping the random-walker, we can assume these regions are more disconnected than the closed conformation and therefore less dynamically constrained.

The first major difference between conformations is the large drop in $\overline{VI}$ between Markov times 200 to 400 in the closed conformation which corresponds to an 8-way partition. The community boundaries of this 8-way partition correspond closely to the dynamical hinges defined by [Henzler-Wildman et al. \(2007b\)](#). This suggests that Markov stability is able to identify areas of flexibility that link communities of atoms that move collectively at that particular timescale. There are two community boundaries that do not correspond to the 8 identified hinges. These are found at positions *Asp81 – Gly82* and *Asp105 – His106*. However, using HingeProt we were able to

show the presence of two hinges at these positions.

Both conformations have a robust long-lived 3-way partition during which their community structure is almost identical (Figure 6.2). In both conformations the partitions grouped residues 30-71 as a single community, residues 112-170 as a community and the remaining residues as one community. The theoretical subdomains are defined as (i) the lid which is inclusive of residues 124-154 + core-lid connector helices LIDC1 residues 114-123 and LIDC2 residues 155-170, (ii) AMP_{bind} domain which consists of residues 31-69 and (iii) the CORE domain which is the remaining residues. There is strong agreement between the community structure of both conformations and the theoretical positions of the subdomains.

Looking more closely at the Markov stability quality function (top-panel Figure 6.1) we observe a higher Markov stability indicating that, in contrast to its open counterpart, the closed conformation exhibits a more fluid structure where the Markov process can more easily escape the separate communities. This indicates that communities in the open conformation are relatively more independent of each other, suggesting that the dynamical rates in the open conformation are higher than from the closed conformation in the apo state. This qualitatively agrees with NMR experiments that found $k_{open} = 6500 \pm 500s^{-1}$ and $k_{closed} = 2000 \pm 200s^{-1}$ (Henzler-Wildman et al., 2007b).

Whilst the catalytic subdomains correspond to the community structure of the 3-way partition, the sub-domains actually appear as single communities separately at earlier Markov times. Figure 6.2b exhibits the hierarchy of communities and how they combine to create a coarser community structure as Markov time increases. We see that the AMP_{bind} domain first appears as a single community at Markov time $t = 300$ in the open conformation and Markov time $t = 400$ in the closed conformation. The lid domain doesn't appear as a single community till later Markov times, $t = 600$ in the open conformation and $t = 750$ in the closed. The random-walker becomes trapped within a subdomain indicating that the motion of atoms within that community are correlated at that particular timescale. This would suggest that the dynamical motion of the AMP_{bind} occurs at shorter timescales (a faster rate of motion) relative to the lid domain.

The lid subdomain is made up of two lid-connector helices, $\alpha 6$ (residues 114-122) and $\alpha 8$ (residues 155-170), and the floppy lid (residues 123-154). The floppy lid is a semi-floppy region with two beta-sheets and a 3/10 helix ($\alpha 7$). At shorter Markov times each lid connector helix and the floppy lid form their own respective community in both the open and closed conformations. As Markov time progresses, the lid connector helices join to form a single community. Interestingly, the 3/10 helix is grouped with the floppy lid in the closed conformation whilst in the open conformation it forms a community with the lid connector helices. This disparity indicates that 3/10 helix provides some structural importance in the motion of the floppy lid relative to the lid connector helices.

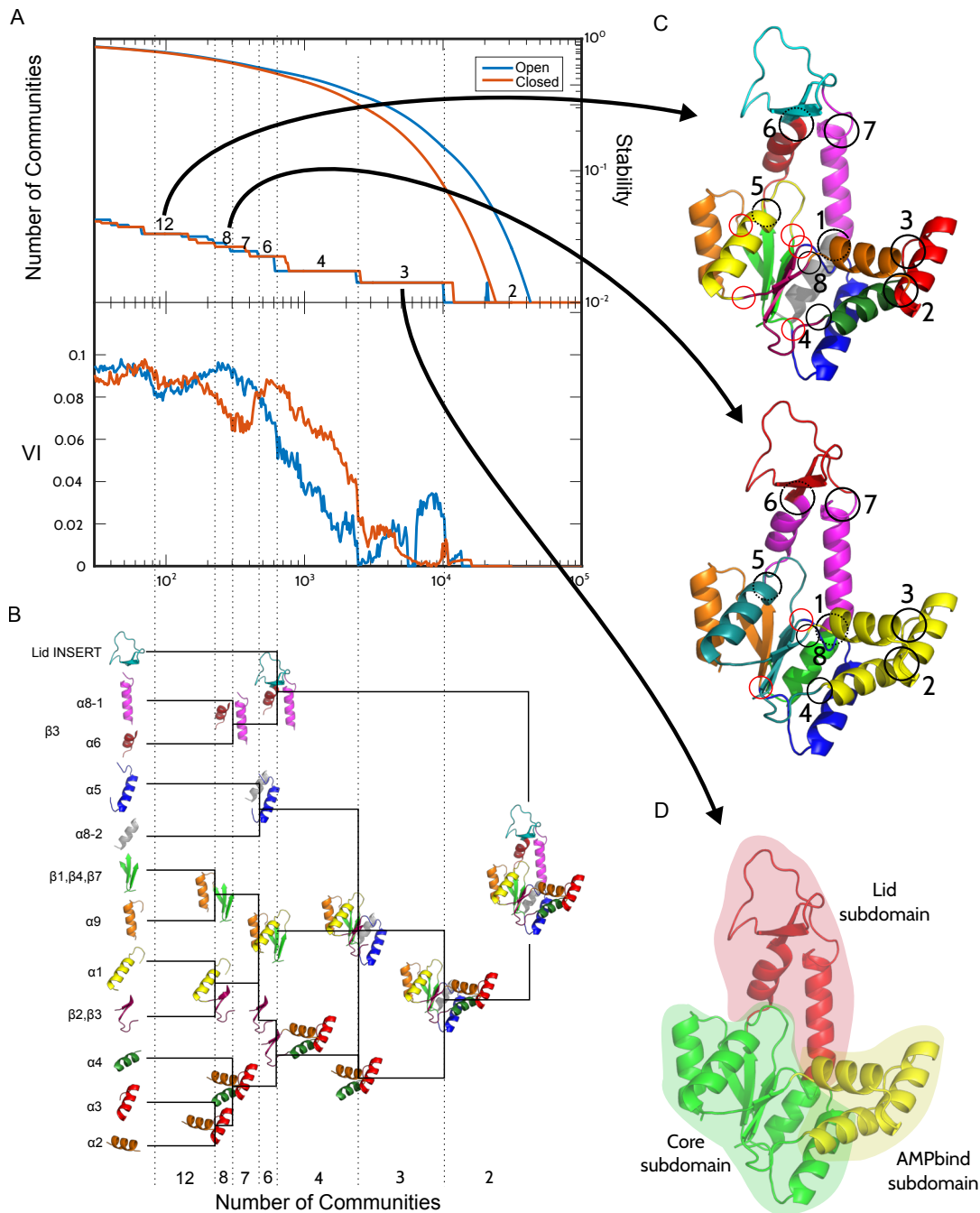


Figure 6.2: Multiscale partitioning of *Aquifex* ADK as a function of Markov time. (a)-(b) Various components in the secondary structures group into larger communities with Markov time. Markov stability does not preimpose any hierarchical grouping of communities, however, it is clear that the larger community structures are built from the smaller communities. Particular groups such as $\alpha 6$ and $\alpha 6 - 1$ last for long Markov times indicating their robustness. Drops in the variation of information are observed upon new clusterings. (c) The 8 community structures agree closely with the functional domains and hinges highlighted by Kern and coworkers (Henzler-Wildman et al., 2007b), however, the hinges found in the AMP_{bind} are actually found at an earlier Markov time during the 12-way partition. Community boundaries highlighted in red could be potential hinges that are yet unknown. (d) The 3 community partition agrees with the catalytic subdomains.

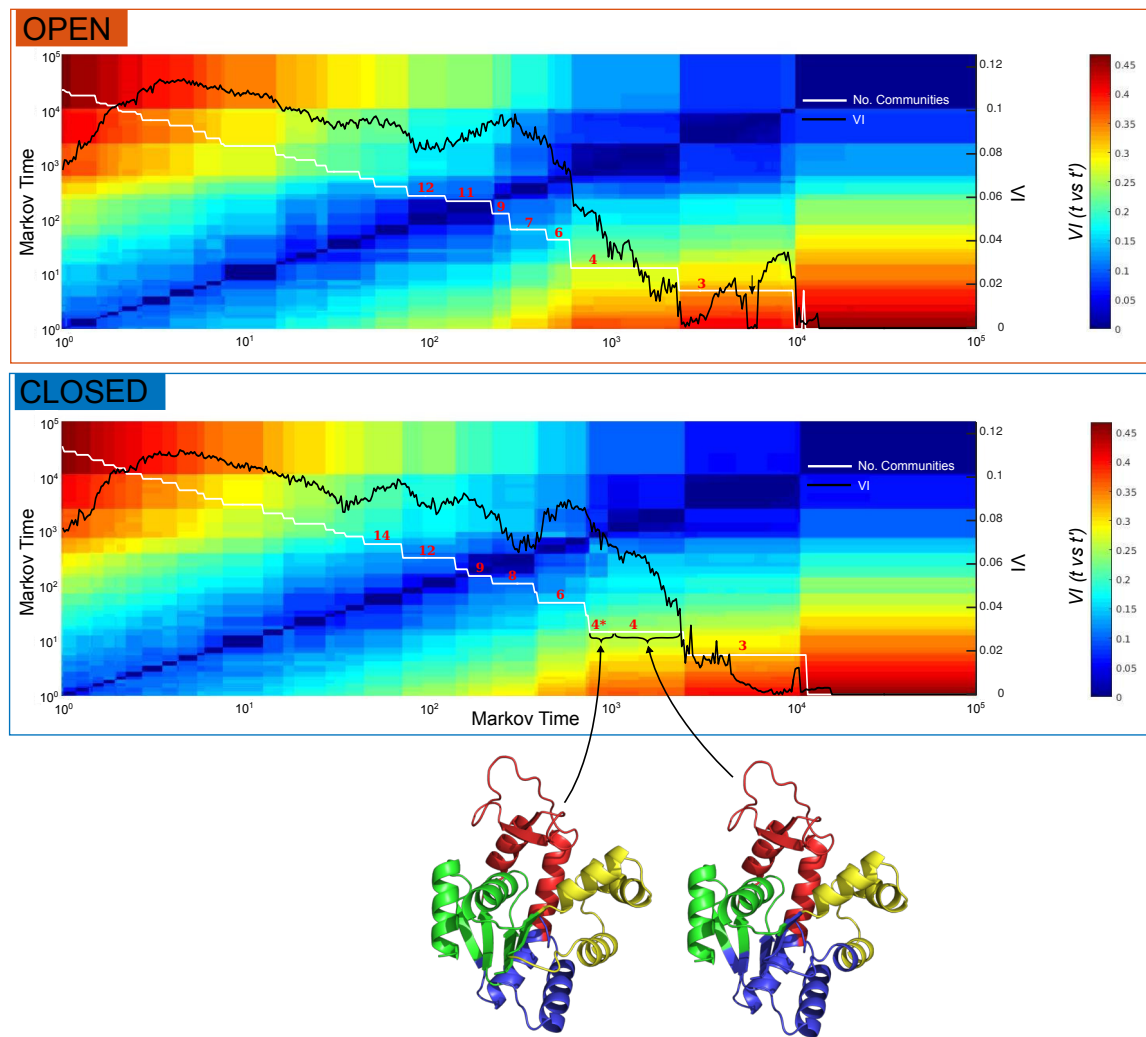


Figure 6.3: A heatmap showing the similarities between community structures optimised at different temporal scales. A $VI(P(t), P(t'))$ graph where the VI between partition $P(t)$ and $P(t')$ is calculated for every combination of t and t' . This provides a visual representation of the robustness of community structures against the perturbation of time. Dark blue indicates a high similarity whilst a dark red indicates very little similarity.

Figure 6.3 shows the $\overline{VI}$ and the number of communities superimposed on a $VI(P(t), P(t'))$ graph, where the VI between partition $P(t)$ and $P(t')$ is calculated for every combination of t and t' . The diagonal of such a graph is zero where the communities being compared are the same. Regions of dark blue are considered to be partitions that persist through time (a plateau in the number of communities could consist of two alternative partitions with an equal number of communities). Using this we can see that the 4-way partition of the closed conformation is not robust through time relative to the open conformation. Two separate 4-way partitions exist in the closed conformation with their differences associated with bend 2 (B2 - residues 72-77). In partition 4*, B2 groups itself with the AMP_{bind} subdomain but at longer Markov times it dissociates itself from the AMP_{bind} community and joins a community in the core domain. Interestingly, the same segment B2 appears as a single community in the open conformation at Markov time $t = 330$. This highlights the importance of this bend as dynamically independent relative to the rest of the protein.

6.2. Construction of three FRET pair mutants

The 3-way partition identified by Markov stability corresponded to the core, AMP_{bind} and lid subdomains. However, the AMP_{bind} first appeared at earlier Markov times than the lid, indicating that it possessed dynamics that occurred at a shorter timescale (a higher rate of motion). To test this hypothesis we designed three ADK structures (Figure 6.4), each with two surface exposed cysteine mutations:

- **1. Lid and AMP_{bind}.** Cysteine mutations were introduced at Y52 (AMP_{bind}) and C145 (lid). These cysteine positions were taken from [Henzler-Wildman et al. \(2007b\)](#).
- **2. Lid and core.** Positions C145 (Lid) and L199 (Core).
- **3. AMP_{bind} and core.** Positions Y52 (AMP_{bind}) and L199 (Core).

The resultant sequencing results from quickchange mutagenesis can be found in Appendix B. The cysteine mutations contain a thiol group which can be conjugated with a maleimide derivative of Alexa fluor 488 and Alexa fluor 633.

The three FRET pairs provide a 3-dimensional perspective of the large-scale dynamic motions. The labelling sites were selected to optimally detect interdomain distance changes between the open and closed conformations. Using FRET pair 1 we can measure the relative motion of both sub-domains which is essential for enzymatic catalysis. FRET pairs 2 and 3 allows us to distinguish the motion of the lid and AMP_{bind} subdomains separately such that we can highlight their independent contributions to enzymatic dynamics.

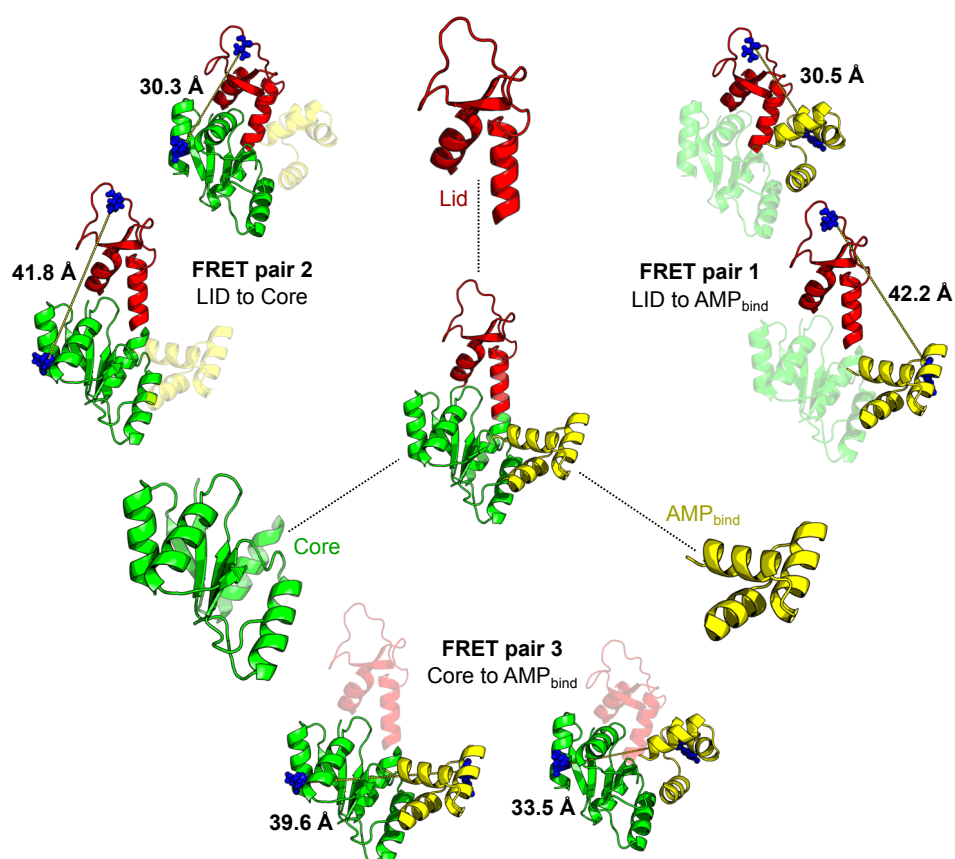


Figure 6.4: Crystallographic structure of the open and closed conformations with the different fluorophore positions labeled. Three ADK FRET pairs have been constructed that allow us to probe the relative dynamics between the lid & AMP_{bind} (FRET pair 1), lid & Core (FRET pair 2) and the AMP_{bind} & Core (FRET pair 3).

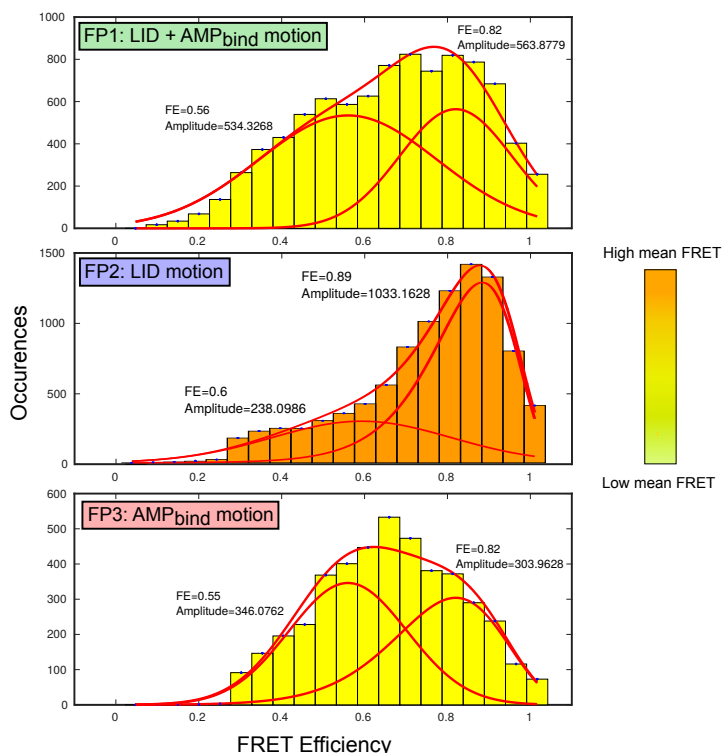


Figure 6.5: FRET efficiency histograms of ADK exhibiting the inter-domain dynamics. Each FRET pair shows that the subdomains sample the high and low FRET states. The lid to AMP_{bind} and AMP_{bind} to core FRET pairs show a distribution where the high and low FRET states are near equally sampled, whereas, the lid to core FRET pair prefers a high FRET state.

To confirm protein activity, an enzymatic assay (described in Section 5.1.2) was performed on the three FRET pairs. The enzymatic rate of the labelled species is $\sim 30\%$ of the wild type.

6.3. Single-molecule FRET details population equilibrium

To assess the independent dynamics of the separate subdomains we measured the FRET efficiency of the three FRET pairs. The three ADK FRET pairs were prepared and measured in conditions described in Section 6.3, in particular they were labeled using Alexa 488 and Alexa 633 as a donor and acceptor respectively. We measured the single-molecule FRET efficiency (E_t) of the three FRET pairs in the apo state freely diffusing in solution (Figure 6.5). FRET efficiencies were calculated from the total number of donor and acceptor photons counted during the dwell time of the laser focal volume. Each histogram was fitted with a two-state normal-distribution mixture model where the means were determined independently for each FRET pair. The substrate bound conformation was used as a reference to calibrate the relative FRET efficiencies with respect to distance for each FRET pair (Schuler et al., 2002).

Each FRET pair sampled a low and high FRET state in the absence of ligand (Figure 6.5), suggesting that the subdomains sample the open and closed conformations. This indicates the presence of a conformational selection model where ADK is continuously sampling the active and inactive states. From these FRET results we can draw two key conclusions. First, apo ADK samples both the open and closed conformations. Second, the timescale of interconversion must be of similar timescale to the residence time inside the confocal volume (~ 0.1 ms), otherwise complete averaging of the FRET populations would occur (Margittai et al., 2003). Interestingly, however, the ratios of the open and closed states differs between the three FRET pairs.

The core domain is perceived as static, whilst the lid and AMP_{bind} fluctuate relative to it. FP1 shows that the relative FRET efficiency of the lid to AMP_{bind} samples the open and closed states in a near equal manner, $E_t = 0.56$ and $E_t = 0.82$ at a ratio of $\sim 1:1$. This agrees with previous FRET measurements by Henzler-wildman and coworkers (Henzler-Wildman et al., 2007b) who used the same FRET pair.

FRET pair 2 (FP2) exhibited a majority high FRET state with $E_t = 0.60$ and $E_t = 0.89$ at a ratio of $\sim 1:4$ respectively in the absence of ligand. This tells us that the fluorophores on the lid and core subdomains are geometrically close, suggesting that the lid is in a closed state. Appropriately this agrees with Hanson et al. (2007) who found the closed conformation as the majority state when measuring FRET from lid to core. Moreover, a recent paper by Kovermann et al. (2017) found that the closed conformation of ADK does not prevent substrate binding. Suggesting that, geometrically, a lid closed conformation is a viable major transition state.

FP3 show a near equal open and closed conformation with $E_t = 0.55$ and $E_t = 0.81$ at a ratio of $\sim 1:1$. This suggests that the AMP_{bind} domain is continuously fluctuating back and forth with very little free energy difference between the conformations. To our knowledge this FRET pair (or similar) has not been measured before. FP1 and FP2 agreed with previous studies, suggesting that the our experimental methodology for FP3 was also reliable.

Whilst the FRET efficiency histogram of FP3 was fitted with a two state model (in accordance with the two-state paradigm), it seems equally likely that we could fit a broad single peak at $E_t = 0.65$. To explore this further we fitted the FP3 FRET histogram with a varying number of distribution components in the Gaussian mixture model (1 to 4 distributions). The AIC was calculated for each model as a function of its residual sum of squares, number of components and data set size (Section 6.3) and is presented in Figure 6.6. The BIC was also calculated and showed near identical results (not shown). Whilst the double component model exhibits the lowest AIC, the single component model is similarly ‘good’. Under the assumption of a two state system, the relatively low AIC for a single component model suggests that the two states are being averaged as a consequence of fast timescale inter-conversions of the AMP_{bind} domain (Margittai et al., 2003).

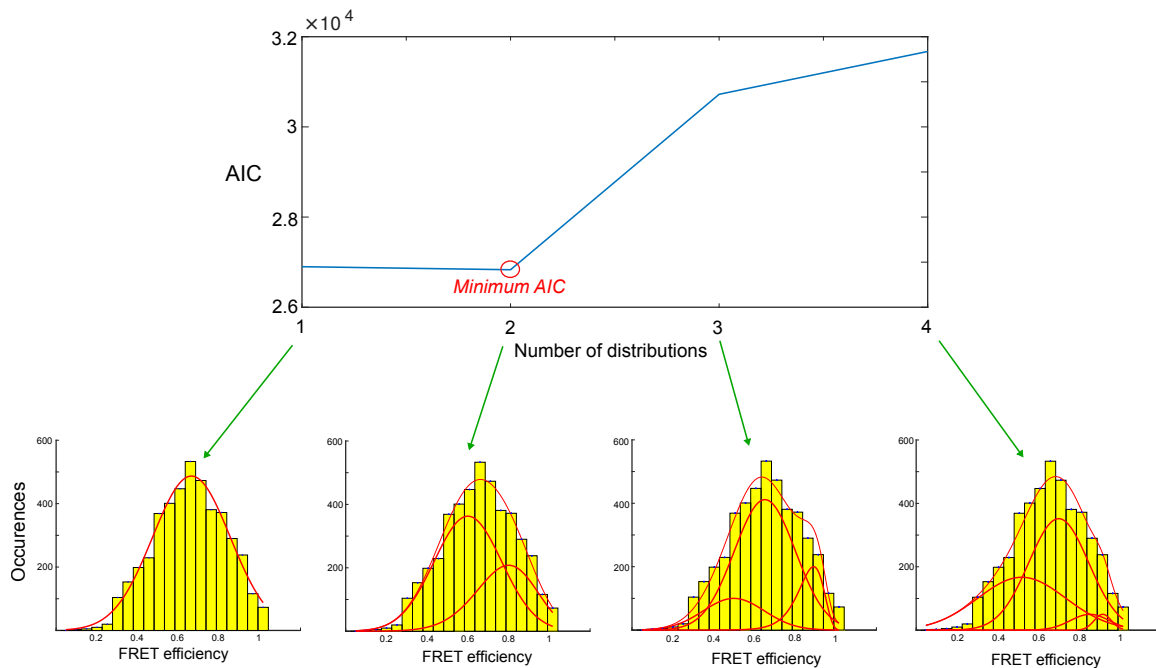


Figure 6.6: The AIC identifies the optimal number of distributions for the FP3 FRET efficiency histogram. The number of normal distributions was increased from 1 to 4 and the AIC was calculated for each one. The two-state mixture distribution exhibited the minimum AIC, however, the one-state mixture distribution also showed a relatively low AIC value. This suggests that the two states are being averaged as a consequence of the fast timescale inter-conversions.

To build a geometric model we need to consider all directions at once and acknowledge the orientational contributions to FRET efficiency. For now we will assume that R_0 is constant across the three FRET pairs. Using Eq. 5.2 we can determine the distances associated with the FRET efficiencies. These are shown in Table 6.1 and show good comparison with distances calculated from x-ray crystal structures. Both FP1 and FP3 exhibited a near equal population two-state model, where the common variable was that they both considered the motion of the AMP_{bind} domain. This suggests that the AMP_{bind} domain is continuously inter-converting between two similar free-energy states. Whereas, the lid domain, that exhibited a majority high FRET state, seems to prefer a closed conformation in the absence of ligand but still samples the open conformation intermittently. Biologically, this suggests that the lid domain is unavailable for binding to prevent unwanted binding events until substrate is bound at the AMP_{bind} domain. This would inherently support a model where the AMP_{bind} domain acts like a gate-keeper for ADK (Kubitzki and de Groot, 2008; Beckstein et al., 2009). Alternatively, the lid domain may already be partially unfolded in accordance with the cracking model discussed in Chapter 3 (Maragakis and Karplus, 2005; Whitford et al., 2007). This partially unfolded state would bring the donor and acceptor fluors geometrically closer together and produce the observed high FRET state. Another potential model is that the AMP_{bind} domain may be moving relative to the other two subdomains that are relatively static.

Table 6.1: A comparison of the FRET distances (calculated using the FRET efficiency of each state) with the geometric distances between the labeled residues in the x-ray crystallographic structure.

	FRET distances (Å)		X-ray Crystal distances (Å)	
	Open	Closed	Open	Closed
FP1	46.1	37.3	42.2	30.5
FP2	44.8	33.9	41.8	30.3
FP3	46.4	37.7	39.6	33.5

Markov stability predicted a robust 3-way partition which indicated a modular structure where the subdomains fluctuated independently. The FRET measurements confirm this, showing that the lid and AMP_{bind} both fluctuate around the core domain. However, whilst lid and AMP_{bind} motion is independent of the core domain, this doesn't necessarily imply an independent catalytic mechanism. The FRET efficiency of the AMP_{bind} to core motion indicated the presence of averaging (Figure 6.6) which is indicative of a fast interconversion between states relative to the lid subdomain which showed two separate peaks. In Section 6.1 we noted that the AMP_{bind} domain appeared at earlier Markov times which corroborates with a faster dynamical rate. To explore this further we must directly measure the rates of interconversion, to do this we will use FCS.

6.4. Comparison of FCS dynamical rates of sub-domains agrees with Markov time

The smFRET histograms detailed ADK's conformational equilibrium and showed the lid to assume a closed conformation whilst the AMP_{bind} sampled both open and closed states equally. Whilst the FRET efficiency measurements indirectly suggested that the AMP_{bind} fluctuated at a high rate of interconversion, the measurements don't calculate the rates and thus we are only exploring the relative free energy differences between conformations. Here, we use fluorescence correlation spectroscopy (FCS) to quantify the relative dynamical rates of the ADK subdomains by analysing the fluctuation of fluorescence intensity at equilibrium. FCS allows us to elucidate conformational protein dynamics down to $\sim 10 \mu\text{s}$ and thus provides a measure of the kinetic barriers to ADK dynamics.

The proteins were prepared and FCS measurements obtained as described in Section 5.4.1. Each autocorrelation measurement was calculated over 5 minutes of photon detection to improve the statistics. Additionally, a constant intensity of transient photons ($G_0 \approx 1.5$) across the 5 minutes was required to prevent the inclusion of new components of correlation e.g. before measurements it is important to allow for intensity drifts due to a temperature gradient between the protein solution and the confocal objective.

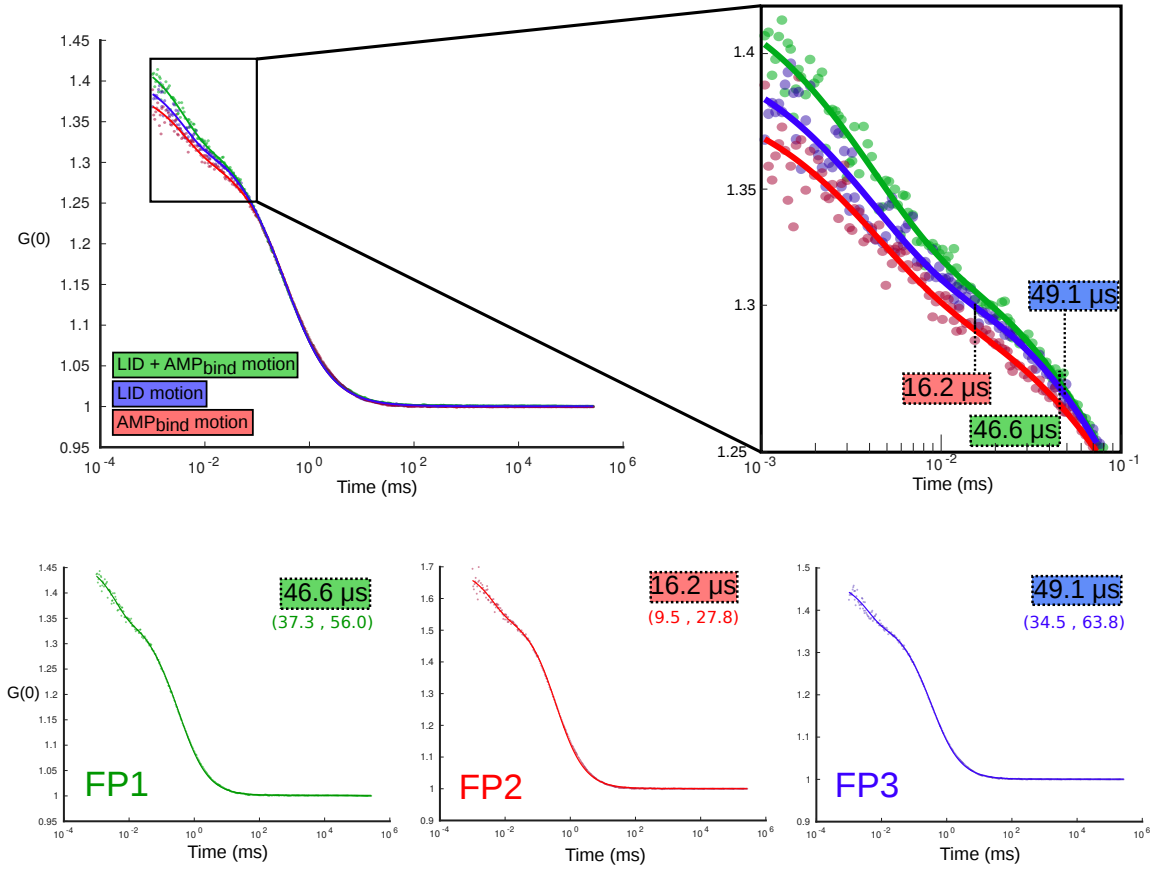


Figure 6.7: FCS autocorrelation measurements for the three ADK FRET pairs. FCS measurements of each FRET pair were normalised and overlaid onto a single graph. The zoom in details the differences between the dynamical timescales of the separate relative domain motions. The AMP_{bind} to core (FP3) exhibited a much faster dynamical rate, whilst the lid to core (FP2) and lid to AMP_{bind} (FP1) showed similar slower dynamical rates. The three separate FCS autocorrelation curves (un-normalised) are presented at the bottom.

The autocorrelation curves of the double-labeled proteins include components of (i) triplet relaxation, (ii) protein conformational dynamics and (iii) characteristic diffusion. To accurately calculate the protein conformational contribution to the autocorrelation curve we first need to determine the triplet relaxation rate and the characteristic diffusion rate of ADK across the confocal volume. A single-donor labeled species of ADK was constructed to measure the triplet and diffusion rates where conformational dynamics would be unseen without the acceptor fluorophore available for FRET. The triplet relaxation rate was determined using a high power laser, 480 μ W, to induce a higher probability of inter-system crossing from the fluorophore singlet-state to triplet-state. This increased the amplitude of the triplet relaxation component to $F_2 = 0.37$ (Eq. 5.14) which enabled us to accurately determine the mean triplet relaxation time $\tau_m = 3.6$ of the autocorrelation curve (Appendix C). Using the fixed triplet relaxation time, the diffusion rate could subsequently be determined at normal power conditions (180 μ W) $\tau_D = 0.36$ ms.

The photon streams of the double mutants were collected under normal power conditions (180

μW) and at a concentration of 1.5-2 molecules within the confocal volume at any time. Figure 6.7 presents the autocorrelation curves for the three FRET pairs, including a normalised overlaid comparison of the autocorrelation curves.

The transition dynamics between the lid and AMP_{bind} (FP1) occurred at a rate of $\tau = 46.4 \mu\text{s}$ which is not far from the $\sim 100\text{-}150 \mu\text{s}$ quoted by [Henzler-Wildman et al. \(2007b\)](#). Interestingly, the lid to core motion (FP2) was found to fluctuate at a similar timescale $\tau = 49.1 \mu\text{s}$ as the lid to AMP_{bind} (FP1), indeed it fell within experimental confidence levels, indicating that the lid is the rate-limiting motion. This was further evidenced by the short timescale fluctuations $\tau = 16.2 \mu\text{s}$ observed for the AMP_{bind} to core motion in the final FRET pair (FP3).

When dynamical motions of similar timescales occur, they will tend to average out to a single autocorrelation component ([Saffarian and Elson, 2003](#)). Therefore, the slow motion of the lid and fast motion of the AMP_{bind} domain would be expected to average out to a single timescale in the lid to AMP_{bind} FRET pair (FP1). Indeed, we observe a slight reduction in the dynamical rate ($\tau_{\text{FP1}} = 46.4 \mu\text{s}$), however, complete averaging would suggest an expected value of $\sim 32 \mu\text{s}$.

The rates of interconversion of the separate domains provide a good parallel with Markov stability; the faster dynamical rate of the AMP_{bind} agrees with the appearance of the AMP_{bind} as a single community at earlier Markov times (open conf: $t = 300$, closed conf: $t = 400$) relative to the lid domain that takes a longer time to inter-convert and appears at longer Markov times (open conf: $t = 600$, closed conf: $t = 750$). This highlights the ability to identify domains at the relevant timescales of their inherent dynamics.

6.5. Substrate bound ADK

6.5.1. Markov stability analysis of inhibitor bound ADK

The substrate-bound closed conformation of ADK (PDB ID: 2RGX) was used in Section 6.3 to calculate the distances between FRET pairs in the closed conformation. However, biophysically a protein/substrate complex induces a change in the dynamics which we can measure. To explore this we have calculated Markov stability across the substrate-bound complex and compared it with experimental rates.

Figure 6.8 presents a comparison of the Markov stability calculations for substrate (Ap5A) bound and unbound closed conformations (the sole structural difference is the presence of substrate at the active site). It is immediately obvious that the substrate reduces the Markov stability of the community structure at long Markov times. A high Markov stability indicates a loss of modular structure since partitioning is less viable i.e. the protein is beginning to act as a single unit when complexed with substrate. Additionally, the catalytically relevant 3-way partition of

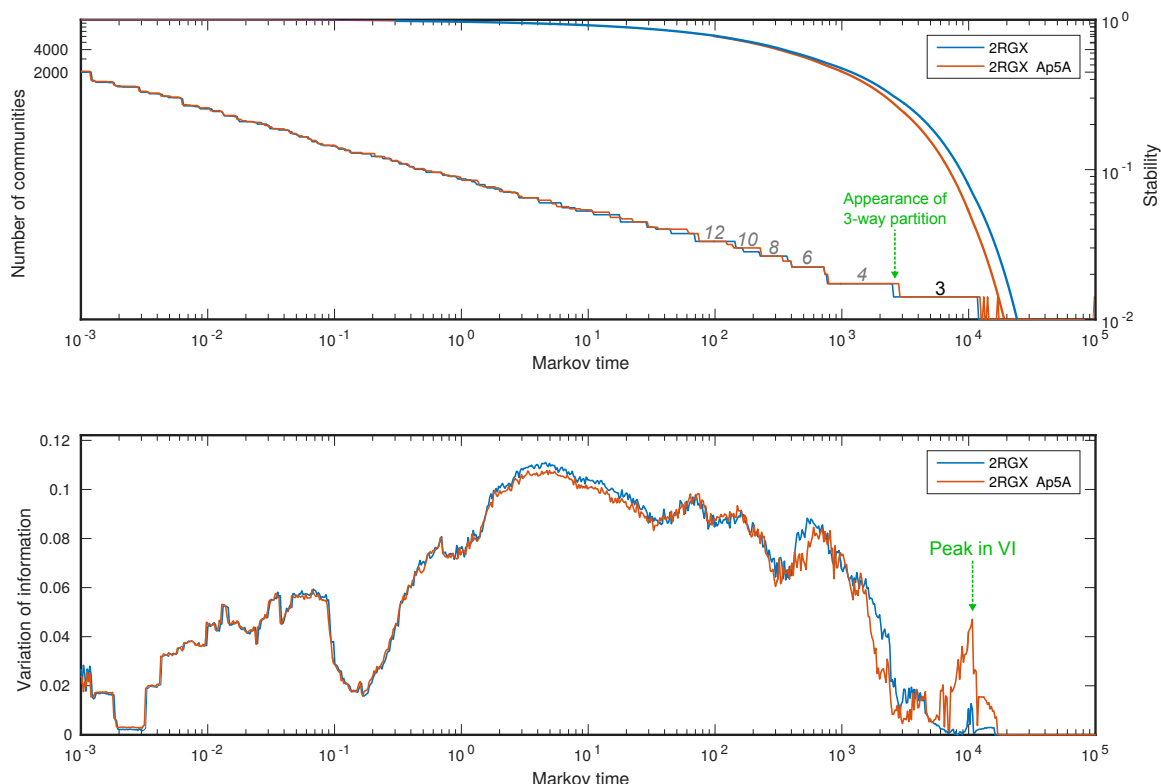


Figure 6.8: A Markov stability analysis of the inhibitor (Ap5a) bound conformation of Adenylate Kinase. The inhibitor bound structure (orange) exhibits a higher lower Markov stability value at long timescales, suggesting a less modular structure. A large peak in VI appears during the 3-way partition indicating a reduction in robustness corresponding to the biophysically relevant 3-way partition.

the substrate bound structure appears at a later Markov time. This suggests that the rate of interconversion in the bound state is lower than in the apo state. One particular major difference is the large VI peak for the substrate bound complex at $t = 10^4$, where the reduced robustness suggests that the community structure during the 3-way partition is weaker.

6.5.2. Relative FRET efficiencies of ADK in the presence of substrate/inhibitor

To investigate these differences we can look at the effect of substrate on the FRET efficiency histograms of the separate FRET pairs. Three separate substrates were used: (i) A substrate-inhibitor P1,P5-Di(adenosine-5') pentaphosphate pentasodium salt (Ap5A) was used to induce closure of both domains, (ii) ATP was used to bind at the lid domain and (iii) AMP was used to bind at the AMP_{bind} domain. For measurements of the molecules in the presence of substrate/inhibitors, 4mM substrate/inhibitor and 2mM MgCl₂ were added directly to the sample wells on the confocal laser.

Figure 6.9 exhibits the FRET histograms of the three FRET pairs in the absence and presence of substrate. FP1 showed a shift towards a majority high FRET state, suggesting that the lid

and AMP_{bind} domains are brought closer together. This result agreed with FRET measurements by [Henzler-Wildman et al. \(2007b\)](#) and with the bound crystal structures that appear to show the lid in closer proximity to the AMP_{bind} domain.

A rather non-intuitive result was the decrease in the mean FRET efficiency for FP3. Addition of any substrate/inhibitor saw a shift towards a low FRET state, implying an increase in population of the open conformation whilst still transiently occupying the closed conformation but to a lesser extent. At an atomistic level, these measurements suggest that the AMP_{bind} domain is moving away from the core domain in the presence of substrate. The crystal structures indicate a shortened distance between the backbone cysteines of FP3 when substrate bound, from 39.6 Å (apo crystal) to 33.5 Å (bound crystal). The reason could be geometrical, whereby the distance between the domains increases but the distance between the fluorophores could actually decrease, however, it is more likely that the rotational freedom of the fluorophores makes FRET less efficient in the substrate bound conformation. Additionally, it could be a consequence of a relative reorientation of the fluorophore polarisation that result in a different transfer efficiency.

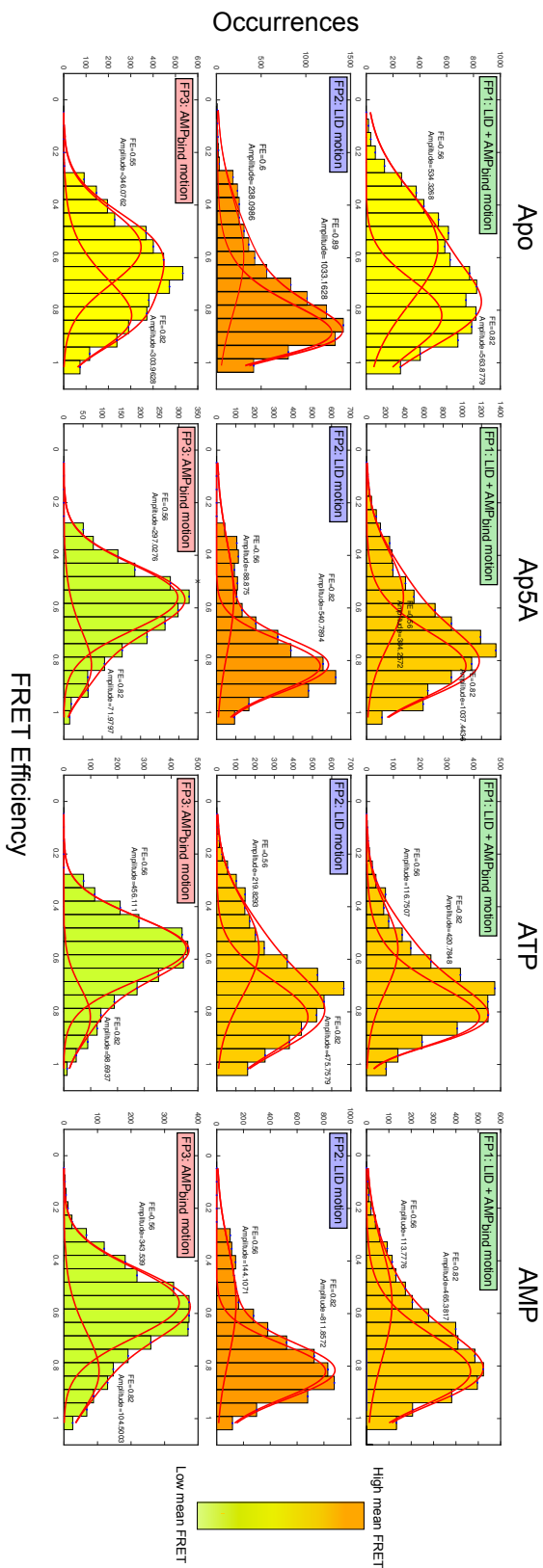


Figure 6.9: FRET histograms of the three ADK FRET pairs in the presence of various inhibitors and substrate. Substrates induced changes to the relative FRET efficiencies of the ADK FRET pairs. The zero peak has been corrected for and a two distribution mixture model is used to fit the histograms. The histograms are coloured according to their mean FRET efficiency.

What is clear from the substrate bound FRET measurements is the shift towards a single majority state, whilst in the apo conformation we find a near equal sampling of two states (the open and closed conformations) indicating the presence of a freely inter-converting system. A substrate thus stabilises the bound conformation, preventing the apo fluctuations previously observed. In the presence of substrate, Markov stability identified the same 3-way community structure found in the apo structure but at a reduced robustness, suggesting a structure that is less modular, and thus, less dynamic. Enzyme-substrate interactions induce a complexed structure where the fast time scale motions are quenched (Kovermann et al., 2017) which may explain the reduced robustness of the 3-way partition.

6.5.3. FCS indicates slower dynamical rates in the presence of substrate

In a similar manner to Section 6.4 we used FCS to explore the dynamical rates of the three FRET pairs, however, this time in the presence of inhibitor/substrate (Ap5A/ATP/AMP). FCS Experiments were performed in an identical manner to Section 6.4, with the inclusion of substrate/inhibitor at a concentration of 4 mM, accompanied by 2 mM MgCl₂.

The single-labeled ADK species was first measured in the presence of substrate/inhibitor at low (180 μ W) and high (480 μ W) laser power. Using this we determined the diffusion time $\tau_D = 0.320$ ms (confidence bounds: 0.314, 0.325) in the presence of ligands. Comparing this value with the apo diffusion rate ($\tau_D = 0.361$ ms (confidence bounds: 0.353, 0.365)) we find that ADK diffuses faster in the presence of ligands. According to Stokes' law, the drag force is proportional to the radius ($D = \frac{k_B T}{6\pi\eta r}$), therefore, a higher rate of diffusion is likely a consequence of a smaller radii. Comparison of the apo and bound x-ray crystal structures shows the radius of gyration to be $r_g = 16.52$ and $r_g = 18.99$ for the the substrate-bound and apo conformations respectively (measurements calculated using *rgyrate.py* written by Wassenaar (2011)). The geometric measurements suggest a drop in gyration radius of 13% which is agreeable with the 11% drop in diffusion rate. This suggests that substrate is effectively binding to ADK, reducing the conformational radii and resulting in higher rates of diffusion.

It became immediately clear that the conformational component of the autocorrelation model did not converge for all three FRET pairs in the presence of substrate, suggesting that the fluorescent fluctuations corresponding to the conformational component were not present or their contribution to the autocorrelation curve was too small to measure. Another plausible argument is that the timescale of conformational interconversion in the presence of substrate is on the order of the diffusion time.

In the presence of a single substrate (either ATP or AMP) or inhibitor (Ap5A), the rate of conformational fluctuations are severely reduced or completely stopped. Indeed, without a reaction to catalyse it is likely that the protein continues to stay bound to its single substrate,

patiently waiting in a transition state for the partnering substrate for reaction can occur. Whilst these measurements aren't definitive evidence for dynamics that occur at longer timescales, it does suggest this.

The Markov stability analysis of substrate-bound ADK in Figure 6.8 highlights the delayed drop into the 3-way partition, implying that dynamics associated with the 3-subdomains occur at longer Markov times. This would corroborate with the hypothesis that dynamics are retarded in the presence of single-substrates or inhibitors.

6.6. Discussion

In this chapter, we have showcased Markov stability as a powerful method to explore the dynamics of a protein graph across all spatial and temporal scales. By exploiting the transients of a Markov process on a protein graph, Markov stability offers an effective method for detection of biophysically relevant community structures. Markov stability reveals the similarities and differences in the hierarchical structures of the open and closed conformations of *Aquifex* Adenylate Kinase. Both structures exhibited a biophysically accurate 3-way partition that corresponded to the core, lid and AMP_{bind} subdomains. However, the open conformation exhibited a higher Markov stability at longer timescales, which reflects its highly modular structure where the subdomains are relatively independent of each other.

Using single-molecule FRET we measured three FRET pairs that modelled the relative dynamics of the three subdomains respectively. Analysis of the FRET histograms suggested independent motion of the subdomains relative to each other, agreeing with the Markov stability analysis that identified the three subdomains as distinct communities. Biophysically, the FRET histograms model the population equilibrium and suggest that ADK is transiently accessing two states (open/closed).

Whilst the 3-way partition agreed with the subdomain structure, the lid and AMP_{bind} subdomains actually appeared at earlier Markov times. In contrast to the lid community, the AMP_{bind} community appeared at an earlier Markov time suggesting that it possessed a faster dynamical rate. Using fluorescence correlation spectroscopy we were able to explore the conformational dynamics of the separate FRET pairs. We found the opening/closing rate of the AMP_{bind} subdomain to be $\sim 3\times$ faster than the lid subdomain. A comparison of the relative dynamics of lid to AMP_{bind} with the lid to core domain showed similar opening/closing rates.

A Markov stability analysis of the inhibitor bound closed conformation revealed a more fluid, connected protein graph, where the previously independent subdomains became less robust. Using smFRET we found that, in the presence of substrate/inhibitor, the population equilibrium shifted towards a single majority closed conformation (bound-state) that transiently sampled

a minor open population (unbound) to a low probability. Despite the component attributed to conformational dynamics not converging, the inhibitor/substrate bound FCS measurements suggested that the opening/closing motions were quenched. This agreed with the Markov stability analysis that showed the inhibitor bound conformation to drop into the 3-way partition at later Markov times.

This chapter has presented a good qualitative agreement between Markov stability and single-molecule fluorescence experiments. In the following chapter, we introduce a mutational scoring method that exploits a number of Markov stability metrics. This will allow us to directly probe the theory-reality relationship.

7. Computational mutagenesis of ADK reveals functionally important residues

We saw in chapter 6 that protein motion is a consequence of the peptide sequence and its associated protein structure. Weak interactions form the tertiary enzymatically-active structures, however, some residues are more critical than others. Substrate binding sites are often highly conserved across protein species making them easy to identify, however, residues that are essential for protein dynamics or molecular stability are not so easily predicted. In chapter 6, we used Markov stability to identify biologically relevant communities within ADK and the Markov times at which they appeared. It is likely that there are a number of residues/bonds that are critical to maintaining these important clusters and defining the dynamic character of ADK.

In this chapter, we used computational mutagenesis to identify residues that perturbed the biologically relevant community structures in ADK. We performed alanine mutagenesis across the entire ADK structure and analysed each mutant with Markov stability. A number of these mutations located within the lid region (to highlight the impact of mutations on the dynamics) were selected for experimental testing with single-molecule FRET to assess changes to the population equilibrium. The mutational analysis was able to separate mutations according to the Markov times at which they perturbed the community structure. We found that a number of the mutations resulted in a shift in the population equilibrium (smFRET), large changes in catalytic activity (kinase assays) and increased flexibility (MD simulations). Control mutations, which were not predicted to cause effect, showed no significant changes in the physical properties we experimentally measured.

7.1. Identification of important mutations using alanine scanning mutagenesis

Alanine scanning mutagenesis is a widely used experimental approach in which individual residues of a target protein are systematically substituted with alanine. Substitution with alanine residues eliminates the side-chain interactions while both preserving native protein structure and

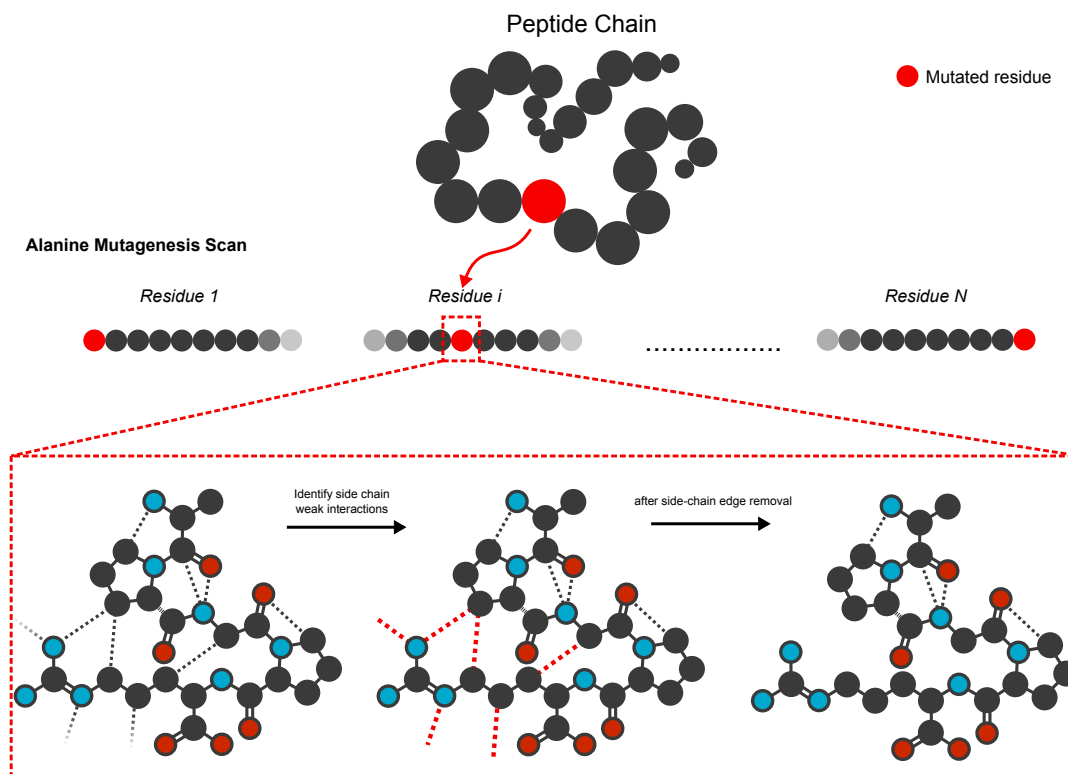


Figure 7.1: Illustration of mimicking alanine mutations *in silico*. Each residue in the peptide chain is sequentially mutated to alanine. The edges that correspond to side-chain weak interactions are removed directly from the graph structure.

not introducing any steric or electrostatic effects. Since the majority of side-chain interactions are removed upon alanine mutation we can mimic this *in silico* by removing all edges associated with the side-chain of the chosen residue (Figure 7.1). Due to the high computational efficiency of Markov stability we can evaluate the perturbation of every possible mutation in turn.

Once an alanine scan has been performed across a protein structure, and Markov stability has been calculated for each structure respectively, we must measure the perturbation or ‘effect’ of a mutation to the community structure. To measure the ‘effect’ of a mutation on the community structure we use two Markov stability metrics:

1. **The variation of information ($\overline{\text{VI}}$).** A measure of robustness of the community structure according to how many times the partition is found across multiple Louvain optimisations. Mutations could increase or decrease the robustness of a partition.
2. **The community variation (CV).** A direct measure of dissimilarity between the optimal community structures of the wildtype and mutant at time t . This is calculated using the variation of information between the wildtype and mutant structure.

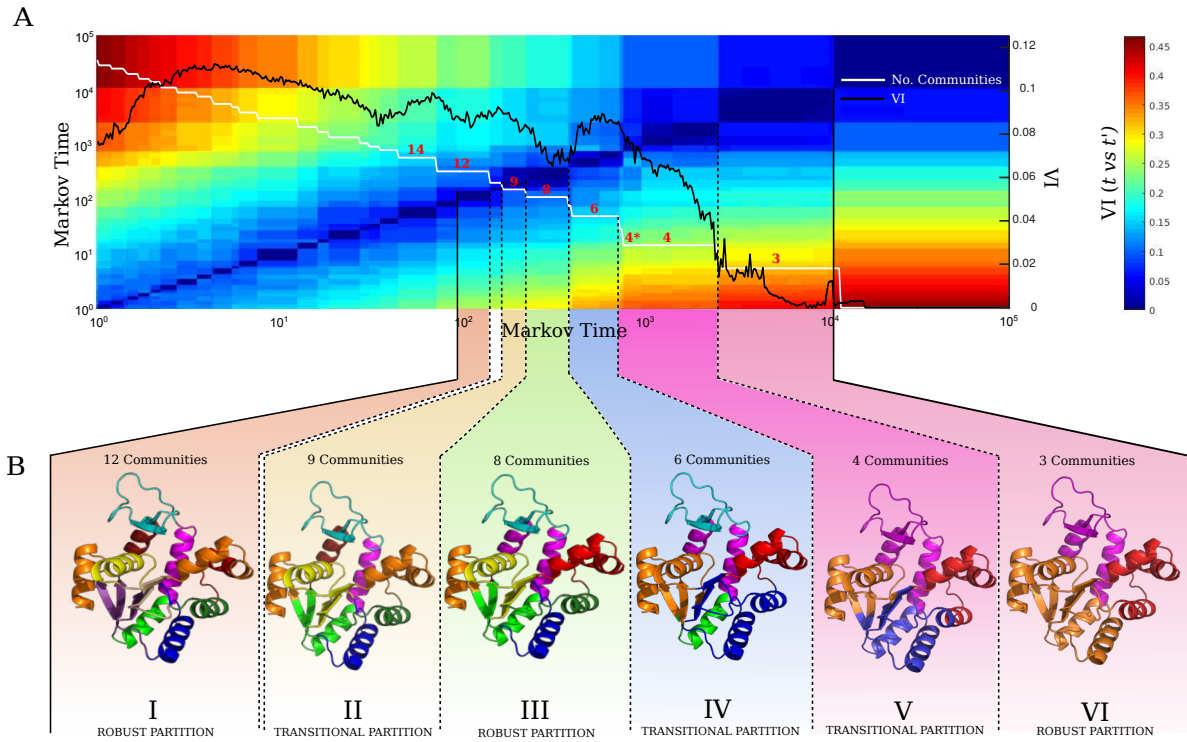


Figure 7.2: Markov stability results for the closed conformation of *Aquifex* ADK and the timescales of interest to look for perturbative mutations. (a) The Markov stability results from $t = 1$ to $t = 10^5$ for the closed conformation. The heatmap measures the community variation (CV) between two partitions, $P(t_i)$ vs $P(t_j)$. Large blocks of dark blue indicate partitions that are similar over a long Markov time. The 4-way partition lasts a long time, however, the exact structure of the partition changes significantly indicating it isn't robust. (b) For each Markov time range of interest the partitions are mapped onto the protein structures. Robust partitions correspond to dips in $\overline{VI}$ and transitional partitions correspond to peaks in $\overline{VI}$.

Using both metrics we are able to measure the isolated effect of a mutation on the community structure over a chosen timescale. Chapter 6 highlighted that there are various timescales at which we find biologically relevant community structures. In the closed conformation we find the 12, 8 and 3-way partitions and in the open conformation we find 12 and 3-way partitions that correspond to biologically relevant robust community structures. These partitions were found in accordance with troughs in the $\overline{VI}$ and plateaus in the number of communities. However, we are also interested in the transition from partition i to partition j where the community structures are susceptible to perturbations (such as mutation). Therefore, the intermediary partitions, in between the robust partitions, are also of interest. In contrast to the robust community structures, these intermediary transitional partitions are medium-lived and exhibit a peak in $\overline{VI}$. To explore this we have defined the 9, 6 and 4-way partitions in the closed conformation and the 9 and 6 -way partitions in the open conformation as transitional partitions. This gives separate ranges of Markov time over which we search for perturbative mutations. The timescales of interest and their associated community partitions are shown in Figure 7.2 for the closed conformation.

We carried out computational mutations of all residues on PDB ID: 2RGX and ran Markov stability calculations from Markov time $t = 10^2$ to $t = 10^5$ with 100 timesteps across every mutant structure. The following sections explore the effect of mutations on $\overline{VI}$ and CV separately, after which we combine both metrics to produce a single Markov stability mutant score.

7.1.1. Mutations that alter the robustness of partitions (VI)

Each mutant structure produces a $\overline{VI}$ trajectory. To find mutations that have a strong effect on the robustness of partitions at each timescale, we detect mutations that induce sustained changes in the robustness $\overline{VI}$. Changes to partition robustness $\overline{VI}$ occur in the continuous Markov time domain meaning that we can use a Gaussian Process regression (GPR) to calculate the mean $\overline{VI}$ trajectory across the ensemble of mutants. A GPR is a nonparametric kernel-based probabilistic model that uses a measure of similarity between points to predict the mean and standard deviation (Rasmussen and Williams, 2004). A further discussion of the GPR is available in Appendix D. The GPR is used to calculate a $\overline{VI}$ that represents the ensemble of mutant and wildtype $\overline{VI}$ trajectories. We can calculate the average distance away from the mean across the chosen timescale to attain a statistical $\overline{VI}$ score. Mutations with a high $\overline{VI}$ score appear as outliers with respect to a Gaussian Process fitted to the $\langle \overline{VI}(t) \rangle$ of the ensemble of all mutations.

$$\langle \overline{VI}(t) \rangle = \frac{1}{n} \sum_n^{i=1} \overline{VI}(i). \quad (7.1)$$

Figure 7.3 presents the $\overline{VI}$ trajectories for the ensemble of mutations (grey lines). A mean $\overline{VI}$ trajectory calculated using a GPR is plotted over the ensemble clearly showing that it accurately represents the data. On closer inspection, it is clear from the grayed ensemble of mutations that particular trajectories will deviate significantly from the mean. Below the ensemble of mutations in Figure 7.3 we have selected the mutation that results in the largest sustained deviation from the mean at each timescale. The mutations are plotted relative to a 2 standard deviation greyed region calculated from the GPR, where fluctuations outside of 2 standard deviations suggests a major disruption to the community robustness. These six mutations (D94A/K179A/I79A/I73A/R166A/R183A) all satisfy our criterion for a major disruption to their respective partitions. Despite causing a major disruption to their respective partitions, four of the mutations (D94A/K179A/I79A/ R166A) result in large deviations of $\overline{VI}$ during neighbouring partitions in addition. This suggests that a mutation can perturb dynamics across multiple timescales, although it may also be a consequence of similarities between partitions of different timescales.

Table 7.1 details the top 10 most disruptive mutations to each partition across the entire ADK structure using the $\overline{VI}$ criterion. Mutation of D152 and R36 both decrease the robustness sig-

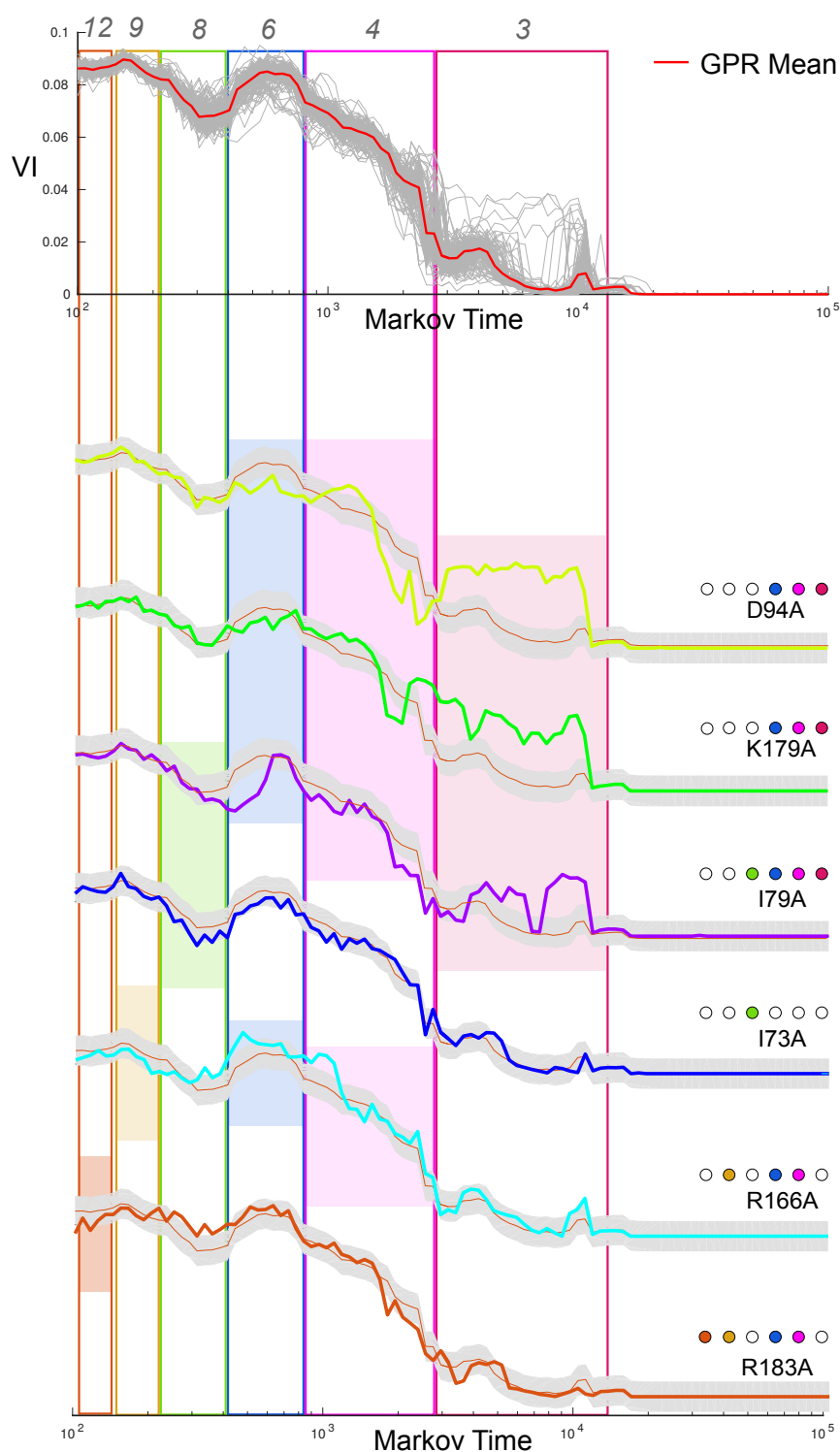


Figure 7.3: Identifying perturbative mutations to the community structure at every timescale. A mean VI (and its variance) is calculated using a GPR across the VI trajectories for the ensemble of mutations (in grey). At each timescale of interest we find mutants that cause large fluctuations according to a 2 standard deviation significance level.

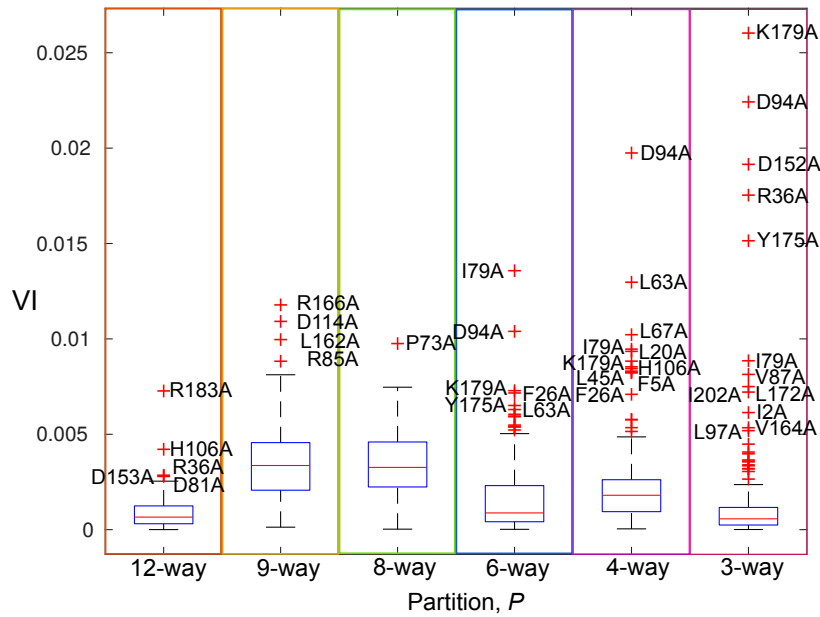


Figure 7.4: Mutations that cause the largest effects on VI. Boxplots of the extent to which mutation effect the VI. Outliers at each timescale have been highlighted.

nificantly during the 3-way partition. Interestingly, these two residues both interact via a salt bridge, stabilising the closed conformation. Another important observation is that during the long timescales partitions (3-way/4-way/6-way) the majority of mutations are localised deep within the core subdomain. This suggests that the robustness of communities is highly susceptible to changes within the core subdomain. Interestingly, none of the top ten residues during the 3-way, 4-way and 6-way partitions are binding residues. Intuitively this would make sense, residues that are structurally critical for long timescale dynamics are less likely to be conserved for binding.

7.1.2. Mutations that alter the make-up of the optimal partitions (CV)

To measure how much the makeup of a partition is affected by a mutation, we use the community variation CV,

$$CV_{[i]}(\hat{P}) = \min_{\tau} VI(\hat{P}, \hat{P}_{[i]}(\tau)), \quad (7.2)$$

which is the variation of information between the wildtype partition $\hat{P}$ and the mutant partition P_i at time τ . This produces a $CV(t)$ trajectory for each mutation and thus a total of N $CV(t)$ trajectories for an N residue protein. We then average the value of $CV(t)$ over a chosen timescale for each mutant,

Table 7.1: A ranking of the top 10 most disruptive alanine mutations during each relevant timescale according to the VI criterion.

Partitions (No. Communities)									
<i>9</i>		<i>8</i>		<i>6</i>		<i>4</i>		<i>3</i>	
R166A	0.0118	P73A	0.0098	I79A	0.0136	D94A	0.0197	K179A	0.0260
D114A	0.0109	A187A	0.0075	D94A	0.0104	L63A	0.0130	D94A	0.0224
L162A	0.0099	Q16A	0.0075	K179A	0.0072	L67A	0.0102	D152A	0.0191
R85A	0.0088	A11A	0.0074	F26A	0.0071	I79A	0.0094	R36A	0.0175
L121A	0.0081	V198A	0.0074	Y175A	0.0065	L20A	0.0093	Y175A	0.0151
I158A	0.0080	V27A	0.0073	L63A	0.0063	K179A	0.0088	I79A	0.0088
E57A	0.0078	I29A	0.0070	M96A	0.0060	I202A	0.0085	V87A	0.0081
K47A	0.0078	K154A	0.0067	L67A	0.0059	H106A	0.0084	L172A	0.0075
V117A	0.0070	F80A	0.0066	A90A	0.0059	L45A	0.0083	I202A	0.0072
Q168A	0.0068	I68A	0.0066	Y176A	0.0055	F5A	0.0082	I2A	0.0061

$$\overline{CV}(t) = \frac{1}{n} \sum_{i=1}^n CV(i), \quad (7.3)$$

providing each mutation with a single $\overline{CV}$ score. A high value of $\overline{CV}$ indicates a large disruption in the optimal partition $\hat{P}$ as a consequence of mutation.

Figure 7.5a shows the CV trajectories for the ensemble of mutations. During the 8-way/6-way/4-way partitions we find large relative differences between $\overline{CV}$ values for the different mutants as exhibited by the grey lines that fluctuates far from the ensemble. Figure 7.5b exhibits boxplots of the mutant $\overline{CV}$ values for each partition respectively. In a similar manner to Section 7.1.1 where we identified mutants which disrupted $\overline{VI}$, the same mutant pair, D152A and R36A, which are critical for stabilising the closed conformation, are found with a high $\overline{CV}$ value during the 3-way partition. Moreover, in a similar manner to using the $\overline{VI}$ criterion, a large number of the high scoring mutations are found deep within the core subdomain within close distance of each other.

The 12-way partition exhibits a much higher mean $\langle CV(t) \rangle \sim 0.45$ relative to the partitions at longer timescales, why is this? In fact, at very short timescales where each node is partitioned into its own separate community the difference between mutant community structures would be zero. At very long timescales, the community structure is partitioned into two communities where any differences between mutant community structure is negligibly small. However, at middling timescales, the potential for a large difference in community structure between the wildtype and mutant increases when the probability of assigning a node to a different community becomes higher. Therefore, we observe an increasing average $\overline{CV}$ as the number of communities within the partition increases (Figure 7.5b).

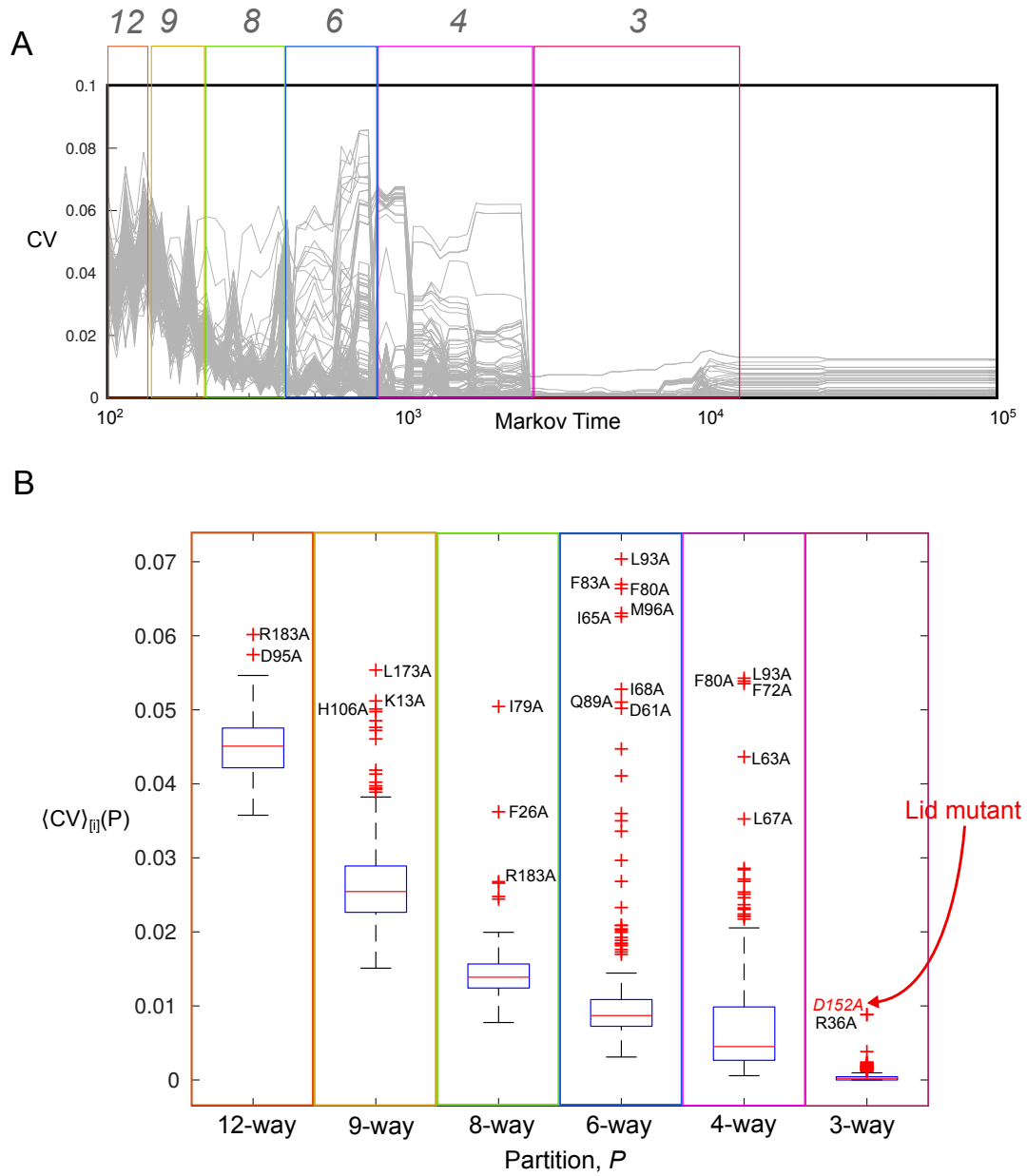


Figure 7.5: Identifying mutants that causes large changes to the community structure using CV. (a) CV was calculated for each mutant at each Markov time step, producing the grey ensemble of mutants. (b) The average CV for each mutant was calculated during the timescales of interest to produce a boxplot. Outliers that exhibited a large average CV across each timescales are highlighted on the boxplot.

7.1.3. Combining VI and CV to create a final mutant score

From the two previous sections it is clear that we can measure the ‘effect’ of a mutation according to its disruption to community robustness ($\overline{VI}$) and community structure (CV). Whilst there are some mutations that are found to perturb both $\overline{VI}$ and CV, others only impact one of the metrics. This highlights the importance of considering both metrics when scoring a mutation.

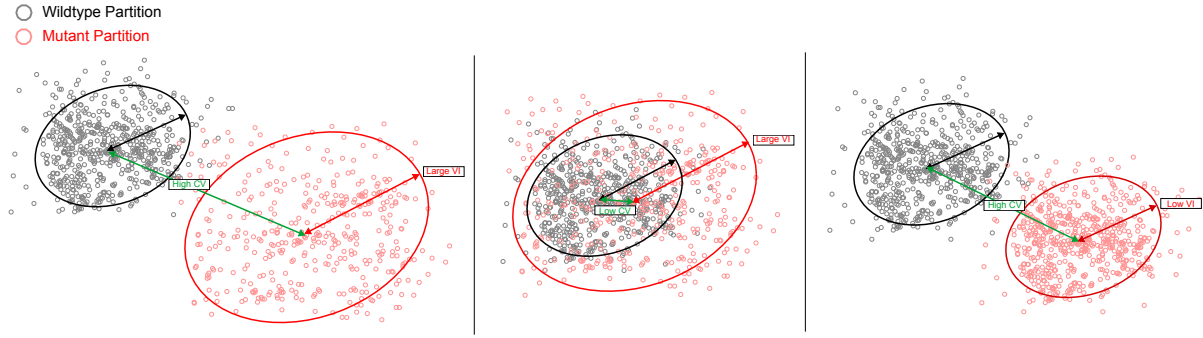


Figure 7.6: An illustration of the VI and CV metrics in partition space. Each panel illustrates a 2-dimensional representation of partition space. The grey cloud of points represent the wildtype which has a width of distribution given by VI, the red cloud represents a mutant where the width of the distribution can change relative to the wildtype (a fluctuation in VI). The mean distance of the mutant from the wildtype is measured by CV. This illustration showcases the importance of measuring both metrics; some mutations can perturb both metrics whilst other mutants will only perturb one.

Here we assign each mutation a score that combines each criteria from both CV and $\overline{VI}$, where both are equally weighted. A final mutation score, X , is calculated using the root mean square distance in respect to the CV and $\overline{VI}$.

$$X = \sqrt{\overline{VI}(t)^2 + CV(t)^2}. \quad (7.4)$$

It is difficult to provide an unbiased approach when combining two metrics into a single value, indeed, by not weighting the metrics we are still biasing one metric over the other. Therefore, throughout this report we will continue to consider each metric separately, whilst also providing the insight from a combined metric. Despite this, it becomes clear that the method of combining the two metrics is robust i.e. the high scoring mutations will continue to be high scoring regardless of the method chosen to combine them.

Both $\overline{VI}$ and CV are measures of distance in the highly dimensional partition space. Figure 7.6 exhibits a cartoon that illustrates the distances within partition space. The $\overline{VI}$ is a distance that represents the variance of the ensemble of Louvain optimisations, upon mutation this cloud of Louvain optimisations can shrink or expand with decreases or increases of robustness respectively. The CV is a distance between two optimal partitions, where a larger distance indicates a large difference between the wildtype and mutant partitions. To simplify the combination of both metrics we will use the root mean square distance to give a single score X_i for each mutant.

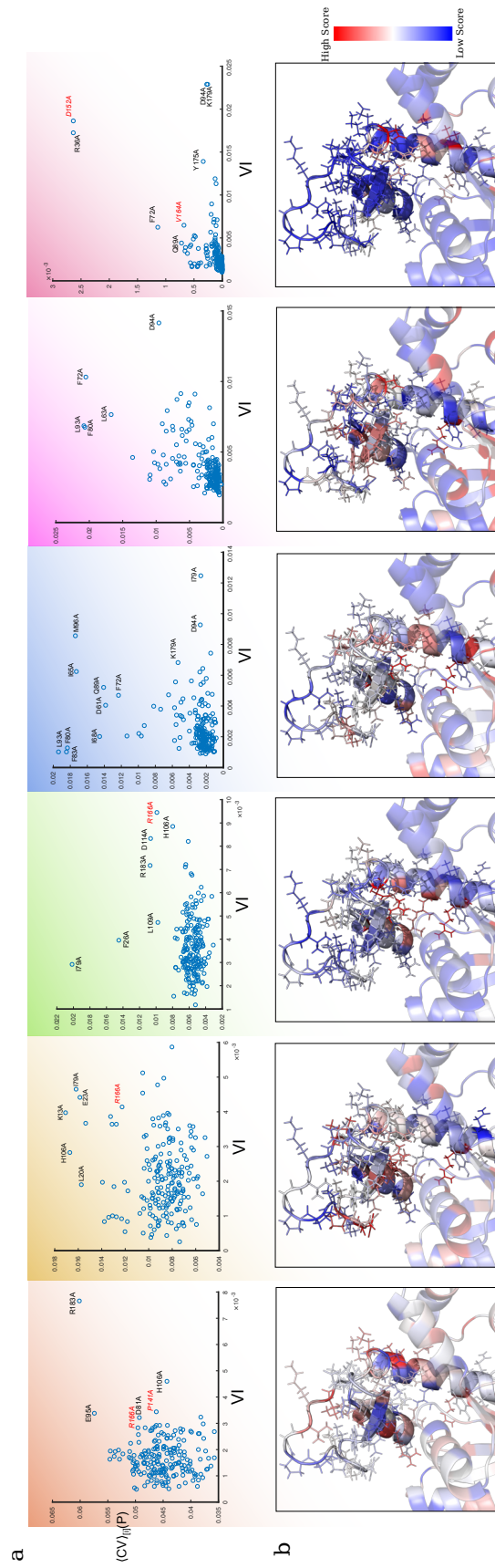


Figure 7.7: Markov stability predicts mutations that perturb different timescales of dynamics. (a) Markov stability calculates the effect of each mutation at each timescale. Each mutant is displayed on a scatter plot as a function of their $\overline{VI}$ and CV values (only lid mutations shown). Mutations that are found in the lid subdomain are italicized in red. (b) The scores of each mutant are mapped onto the protein structures where red is a high score and blue is low.

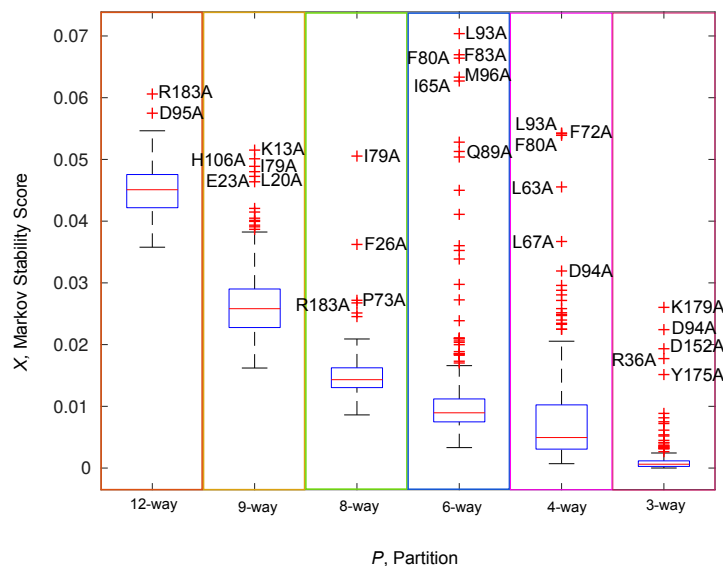


Figure 7.8: High scoring mutants are defined at each timescale relative to the ensemble of mutations. Box plots of the scores for each mutant at each relevant timescale, the outliers are highlighted in red.

Figure 7.7a presents a series of scatter plots that quantify the effect of *in silico* alanine mutations in respect to both $\overline{VI}$ and CV for each Markov time range of interest. The Markov timescales correspond to the community structures identified in Figure 7.2. The scores have been mapped onto the protein structures (Figure 7.7b) and we have focused on the lid domain to isolate mutations that will perturb lid dynamics (all calculations were made in respect to the full protein graph). It is immediately obvious why using both $\overline{VI}$ and CV is essential; the scatter plots present a number of mutants that only cause large changes in a single variable. Using a ‘scoring’ method that exclusively utilised either $\overline{VI}$ and CV would result in lost information and unreliability.

The boxplots (Figure 7.8) exhibit the root mean square distance of each mutation as a function of $\overline{VI}$ and CV, providing the raw relative score, X , of each residue at each timescale. The mean X becomes smaller at longer Markov timescales (lower number of communities), suggesting that the statistical ‘distance’ from the wildtype structure converges with increasing Markov time. This statistical distance is likely born of the stochasticity associated with a Markov process, at shorter timescales the stochastic behaviour manifests itself through small differences in community structure and the discrete number of Louvain optimisations.

As expected, the highest scoring mutations according to the combined score X were identified in the previous two sections using $\overline{VI}$ and CV as separate measures. To build a relationship between theoretical Markov stability scores and experiment, we focus on the dynamical aspect of Adenylate Kinase which we experimentally optimised in the previous chapter. The lid subdomain is well studied and undergoes large amplitude motions, therefore, we have isolated and ranked

Table 7.2: The top ten highest scoring lid subdomain mutants (as a combined metric of VI and CV) during each relevant timescale.

Highest scoring lid mutants											
<i>12</i>	<i>9</i>					<i>8</i>		<i>6</i>		<i>4</i>	<i>3</i>
V157A	0.0538	Y138A	0.0136	R166A	0.0137	R166A	0.0083	R166A	0.0083	D152A	0.0188
R120A	0.0529	R166A	0.0130	L162A	0.0102	R124A	0.0065	R120A	0.0061	V164A	0.0065
E151A	0.0511	I118A	0.0122	D153A	0.0096	D153A	0.0057	D152A	0.0052	R120A	0.0041
V147A	0.0509	P140A	0.0119	R124A	0.0087	R161A	0.0045	L162A	0.0049	R166A	0.0038
P141A	0.0505	I148A	0.0118	R120A	0.0081	E167A	0.0040	L121A	0.0047	Q168A	0.0037
G123A	0.0504	L121A	0.0117	I158A	0.0080	K137A	0.0039	E119A	0.0046	Y165A	0.0026
K160A	0.0503	D153A	0.0109	V117A	0.0080	P140A	0.0038	Q168A	0.0045	S122A	0.0024
G131A	0.0497	R124A	0.0108	R125A	0.0077	R150A	0.0038	D153A	0.0045	V157A	0.0023
R166A	0.0497	V133A	0.0107	H135A	0.0075	E119A	0.0038	R125A	0.0044	K160A	0.0022
K154A	0.0495	R125A	0.0106	E151A	0.0071	I118A	0.0038	V117A	0.0043	D153A	0.0022

Table 7.3: The high scoring arginines that cause a large effect on the community structure of the lid subdomain. Here we list their most important interactions with neighbouring residues.

Residue	Bond
Arg120	Arg120 HH22 : Ala187 O
	Arg120 HH12 : Ala187 O
	Arg120 HH11 : Gly10 O
Arg124	Arg124 HH21 : Asp153 OD1
	Arg124 HE : Asp153 OD2
Arg161	Arg161 HH22 : Asp153 OD1
	Arg161 HH21 : Asp152 O
Arg166	Arg166 HH22 : Val112 O
	Arg166 HH21 : Asp114 OD1

the mutations found inside the lid subdomain.

In the lid domain, there are a number of mutations that appear across multiple timescales as significant according to their mutant score, four of note are Arg120, Arg124, Arg161 and Arg166. Whilst Arg120, Arg124 and Arg161 are substrate binding residues located inside the catalytic site, Arg166 is found at the back of the lid connector helices complexed with a zinc ion in the crystal structure. Ignoring the substrate, each arginine residue makes a number of strong sidechain hydrogen bonds/salt bridges with neighboring residues as show in Table 7.3. Another two residues that appear across multiple timescales are D152A and D153A. Both of these aspartic acids are located on 3/10-helix-7 and hydrogen bond with a number of the arginines within the lid subdomain.

There are particular residues that appear at short timescales only, e.g. P141A and P140A score highly during the 12-way and 9-way partitions respectively. These mutants perturb the local community structure within the lid subdomain, suggesting that we would observe local changes to the dynamics. Indeed, P141A and P140A have been implicated in the reducing flexibility of the lid subdomain to attain the molecular stability required for thermophilic catalysis (Henzler-Wildman et al., 2007a).

During the 3-way partition the majority of top scoring mutants are located on 3/10-helix-7 and α -helix-8 which are in close proximity with the AMP_{bind} subdomain. This suggests that mutations during this timescale will perturb the communication and relationship that is essential for global dynamics. In particular, D152A and V164A, directly interact with the AMP_{bind} subdomain through hydrogen bonding and hydrophobic interactions.

7.1.4. Robustness of the scoring method

The $\overline{VI}$ provided some measure of width (or standard deviation) to the distribution of Louvain optimisations, whilst the CV provided a measure of distance (or mean) between the distribu-

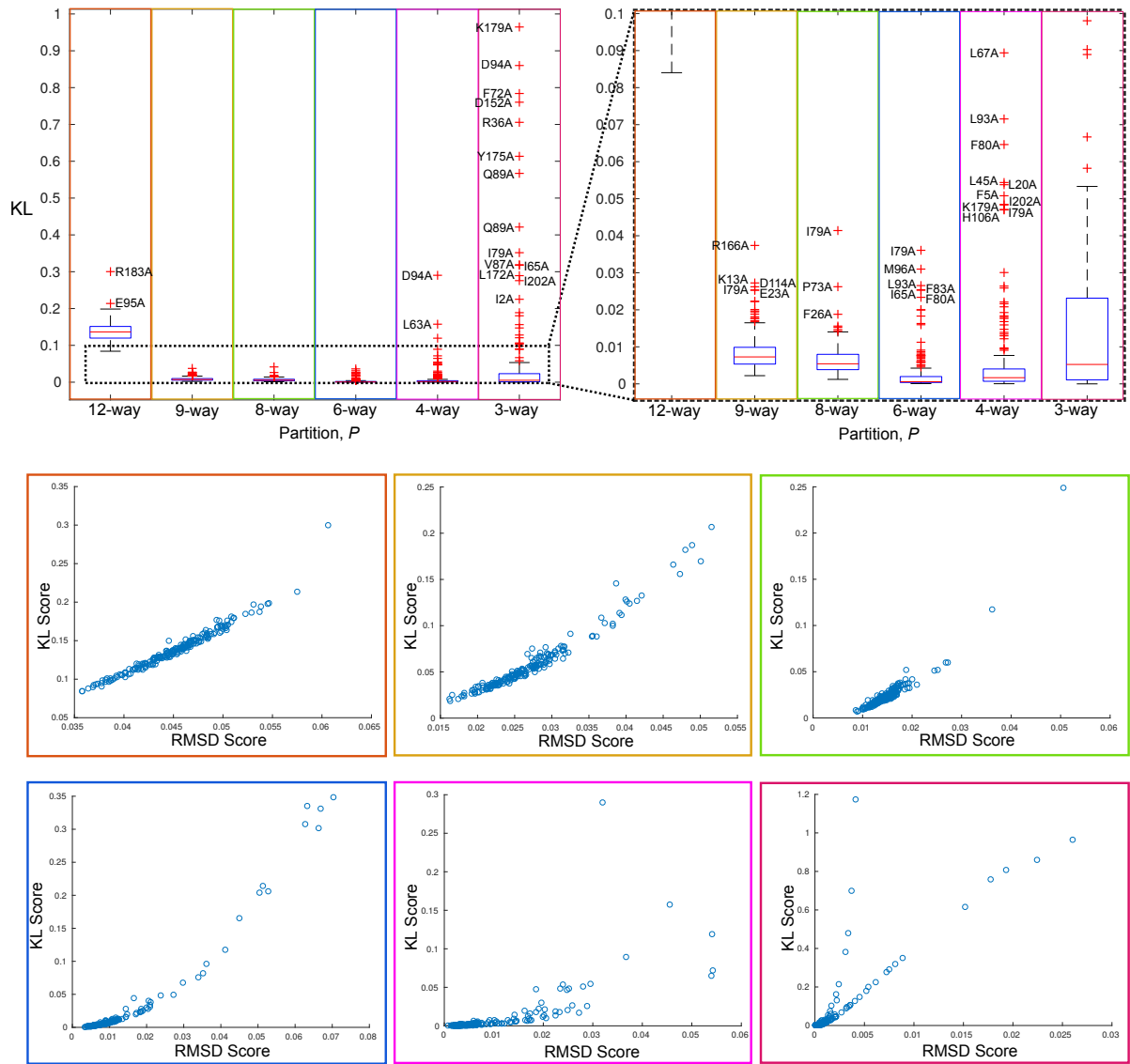


Figure 7.9: A comparison of the Kullback-Leibler method and RMSD method of combining VI and CV into a mutant score. (a) A boxplot of the Kullback-Leibler (KB) score attributed to each mutant as an alternative method of combining VI and CV. (b) Scatter plots of the KB method of combining VI and CV versus RMSD method during each timescale. The majority of timescales show a very strong correlation between the two mutant scoring methodologies, suggesting that the RMSD method chosen to combine VI and CV is robust.

tions. In the previous section we combined the $\overline{VI}$ and CV into a single mutant score using the un-normalised root mean square distance. However, combining a pseudo width ($\overline{VI}$) and pseudo mean difference (CV) as a single score using the RMSD could be considered arbitrary or statistically ambiguous. Indeed, this is why we also continue to use $\overline{VI}$ and CV as separate measures throughout this thesis. Considering this, we have used the Kullback-Leibler (KL) divergence as a measure of divergence between a mutant distribution and the wildtype distribution to showcase the robustness of combining the two metrics. The KL divergence is calculated pairwise, thus we

calculate the KL for each mutant relative to the wildtype,

$$KL(p||q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}, \quad (7.5)$$

where σ_1 and σ_2 are the VI values for the wildtype and mutant, and $\mu_1 - \mu_2$ is the value of $CV(P, P')$. A derivation of this formula is available in Appendix E.

Figure 7.9a exhibits the boxplots of KL values for each mutation during each Markov timescale. A first glance at the outliers during each timescale shows us that there is a strong agreement between the mutations identified by the original scoring method and those identified using the KL approach. To explore this further, we plotted the RMSD score versus the KL-score for each mutation during each timescale (Figure 7.9b). During each timescale, there is a strong correlation between the two scoring methods suggesting that the rmsd scoring method is robust. However, during the 3-way partition, there are two diverging populations that highlight a discrepancy between the scoring methods. A closer analysis showed that the population of steeper gradient is associated with mutations that obtained high CV values, whilst the population of shallower gradient is associated with mutations that obtained large changes in VI. This suggests that the KL scoring method provides a higher weighting to the differences in CV. Despite the difference in weightings, it is clear that the general trends are similar and we can highlight the most perturbative mutations. However, it is important to note that combining the two metrics is still statistically ambiguous and thus we will continue to consider each metric (VI and CV) separately and in conjunction with a combined score.

To explore the effects of these mutational predictions we primarily explore the dynamical motions using single-molecule FRET. The next section takes a number of mutations with varying scores at each timescale, constructs them individually and uses single-molecule FRET, activity assays and MD simulations to understand the effect of each mutation.

7.2. Experimental investigation of various alanine mutants

The previous section used Markov stability to identify relevant timescales, we then used Markov stability to score and rank mutations at each of these timescales. However, we need to understand what these ‘scores’ correspond to physically in a protein. In Chapter 6 we showed that Markov stability is able to identify regions of the protein that move independently of each other. In this section we use a similar single-molecule experimental approach to explore the physical effect of mutations and corroborate these findings with the Markov stability scores and rankings. Since we are investigating perturbations to dynamics, we have isolated and ranked mutations within the lid subdomain only.

7.2.1. Construction of ADK lid mutants according to prediction scores

We isolated FP1 (lid to AMP_{bind} FRET pair) from Chapter 6 and performed QuikChange mutagenesis to construct 7 mutant ADK structures. Each mutation is located in the lid subdomain to focus directly on perturbations to dynamics. In accordance with the computational *in silico* alanine scan, all chosen residues were substituted with alanine:

1. **Q149A**. Low scoring across all timescales in both conformations.
2. **E115A**. Low scoring across all timescales in both conformations.
3. **I118A**. High scoring at medium timescales in the open and closed conformations.
4. **P141A**. Appears in the open and closed conformation at short timescales.
5. **K154A**. High scoring at short timescales in the open conformation and closed conformation.
6. **D152A**. Very high scoring at long timescales in the closed conformation and short timescales in the open conformation.
7. **D153A**. High scoring at medium and long timescales in the open and closed conformation.

Two mutants, E115A and Q149A, were chosen as control mutations that represent low scores across all timescales in each conformation. The remaining residues had differing scores across the various timescales, offering us an opportunity to probe the timescale of perturbations.

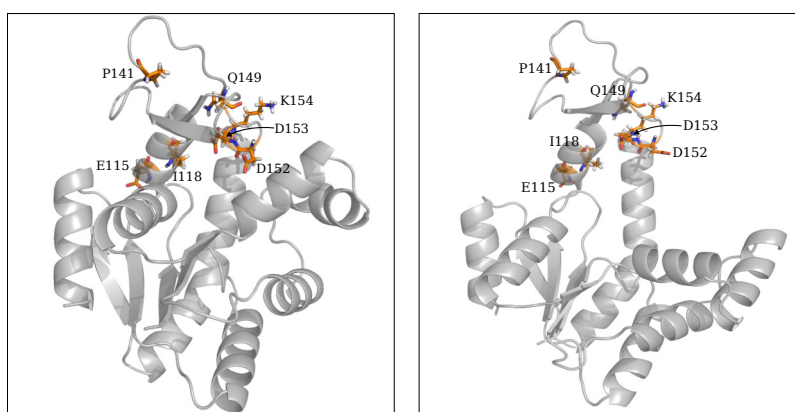


Figure 7.10: The mutations chosen for experimental measurement are highlighted on the protein structures. The positions of each mutation in the closed (left) and open (right) conformations.

7.2.2. Single-molecule FRET measurements of ADK mutants

To assess the effect of each mutation to global dynamics, we measured the relative FRET efficiency (E_t) between the lid and AMP_{bind} subdomains of each mutant ADK. The ADK mutants were prepared and measured in conditions described in Section 5.5, where the cysteines were randomly labelled with Alexa 488 and Alexa 633. Measurements were made in the absence of substrate whilst freely diffusing in solution. The mean FRET efficiencies of the two-state mixture model were fixed in accordance with Section 6.3 where the substrate bound conformation was used as a reference to calibrate the distributions.

Figure 7.11 presents the FRET efficiency histograms of the wildtype (apo and in the presence of osmolyte TMAO) and ADK mutants. The log-normal zero peak contribution to the E_t histograms has been removed such that only the conformational states are illustrated. The wildtype FRET efficiency histogram, measured previously in Chapter 6, was remeasured and appropriately showed a very similar distribution that suggested the open and closed conformations were sampled to near equal populations (ratio of wildtype = 1 : 1.34). As expected, in the presence of TMAO, the wildtype shifts to a closed conformation illustrating the propensity to shift the conformational equilibrium. The control mutations (E115A/Q149A), which showed low scores at all Markov stability timescales, exhibited similar histograms to the wildtype. The control mutations showed no significant change in the ratio of the open and closed conformations relative to the wildtype (ratio of E115A = 1 : 1.41 / ratio of Q149A = 1 : 1.53).

Mutation of I118A, P141A and K154A resulted in shifts towards a closed conformation where the ratios of open to closed states were $> 1 : 2$. These results highlight the fragility of the conformational equilibrium, where unsuspecting mutations can cause changes to the dynamical equilibrium. Interestingly, these three mutations all exhibited some perturbation to the Markov stability community structure at short to medium timescales as measured by their Markov stability score. Focusing on the two mutations that exhibited high scores at long timescales, D152A and D153A, we observe a significant change in the FRET efficiency histograms. Both neighbouring residues exhibited a shift towards a majority closed conformation (ratio of D152A = 1 : 4.56 / ratio of D153A = 1 : 8.3), highlighting their structural importance.

A molecular dynamics study by Brokaw and Chu (2010) found that the single point mutation of Pro177 in *E. Coli* ADK resulted in stabilisation of the closed conformation. The local unfolding of the hinge at Pro177 resulted in an increase in tertiary contacts that stabilised the closed state. Its likely that we are observing a similar mechanism for the mutations investigated here. We have observed changes in the conformational equilibrium as a result of mutation, but do they impact any other properties of ADK? And can we understand the atomistic mechanisms that have caused the change to the dynamics of ADK? The following sections measure the catalytic rates and atomistic flexibility using activity assays and molecular dynamics simulations respectively.

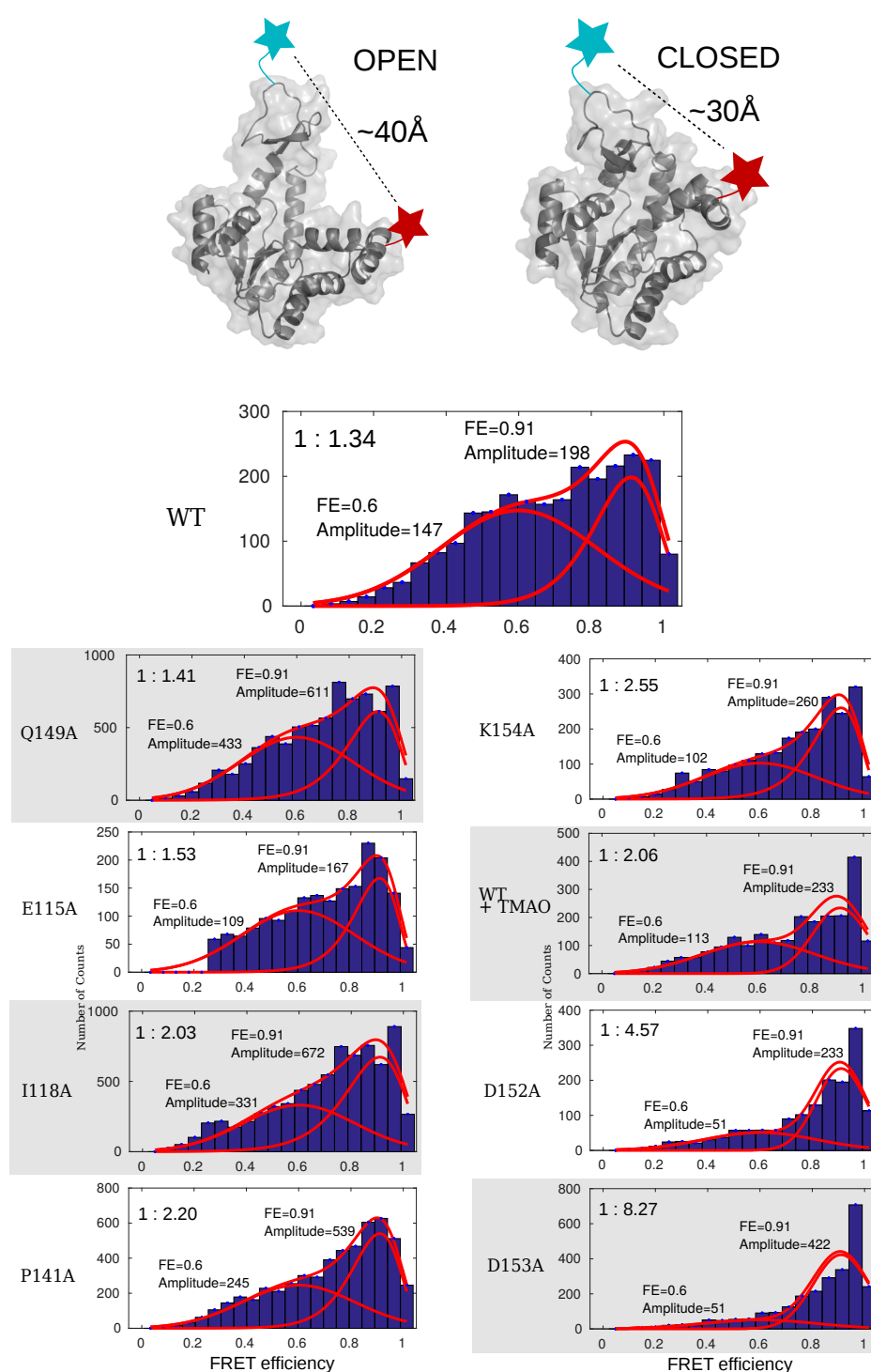


Figure 7.11: Single-molecule FRET measurements of ADK mutants that were predicted across various timescales. (a) The open (2RH5) and closed (2RGX) structures with Alexa Fluor 488 (cyan star) and Alexa Fluor 633 (red star) positions indicate the distances. (b) The FRET histograms for the wildtype and mutant AK_{aqu}s in the apo state. For the WT and each mutant a two Gaussian mixture function was used to fit the data. Large shifts in FRET distribution were observed upon mutation indicating a change in the energy landscape.

7.2.3. Catalytic rates measured using kinase assay

To measure the activity of the alanine ADK mutants, we employed luciferase assays to measure the output of ATP from an excess of ADP (full description in Section 5.1.2). The rates were relative measurements and do not provide absolute values, instead they provide a good comparison with the wildtype. A catalytic mutant control was constructed (Q89A) to observe complete inhibition by removal of a conserved binding residue.

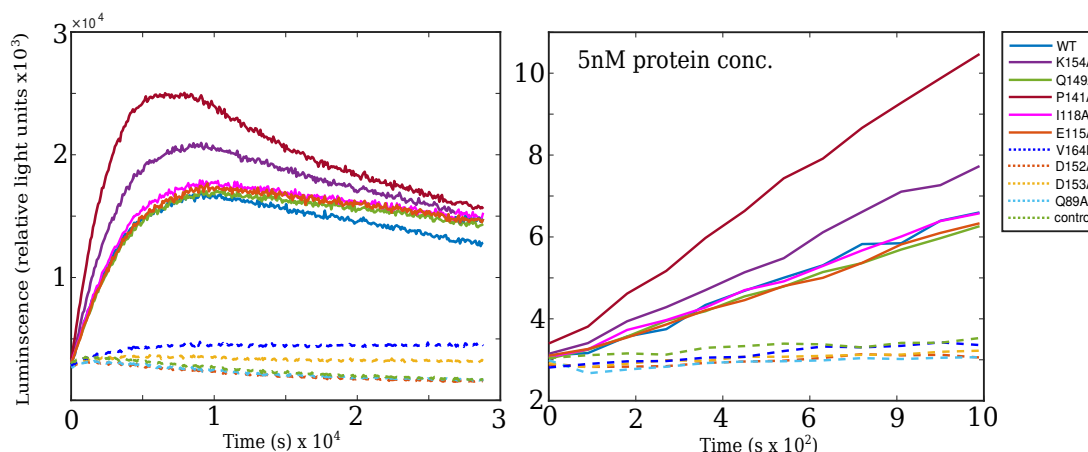


Figure 7.12: The activity of the ADK mutants as measured using a luciferase assay. The left panel exhibits the luminescence output over an 8 hour period. The right panel exhibits the initial rates (first 1000 seconds) before the rate limiting factor was no longer the production of ATP (ADK activity), but the luciferase reaction itself.

Figure 7.12 presents the catalytic rates of the ADK mutants. In agreement with the Markov stability scores they were assigned, the control mutations (E115A and Q149A) exhibited negligible differences to the wildtype activity. Similarly I118A also showed no change in the catalytic rate, suggesting that its high ‘score’ at medium timescales didn’t correspond to any functional aspect. Of particular interest are the K154A and P141A mutants which exhibited an increase in catalytic rate. In contrast to the long timescale global motions, both mutants scored highly at short Markov times which implicates them in more localised motions. Prolines have been previously suspected as an evolutionary mechanism to stabilise *Aquifex* ADK at high temperatures but consequently reducing the flexibility and catalytic rate at mesophilic temperatures (Henzler-Wildman et al., 2007a; Daily et al., 2011). Despite not binding substrate directly, inhibition was observed in D152A and D153A, suggesting that the large change in conformational dynamics may be responsible.

The addition of fluorophores to a protein introduces a large hydrophobic group that can reduce the rate of fluctuations. To allow direct comparison with FRET measurements, activity assays were repeated for the fluorophore labelled species and exhibited a 70% reduction in catalytic rate across all active mutants, agreeing with previous measurements by Henzler-Wildman et al. (2007b).

7.2.4. Molecular dynamics probes the flexibility of mutants

The smFRET and assay measurements provided crucial insight into the dynamics and function. In this section we employ short molecular dynamics (MD) simulations to attain a mechanistic understanding of the Markov stability predictions. MD is a routinely used principal tool for the theoretical study of biological molecules. Using an atomistic forcefield, the computational method calculates the time dependent behaviour of a biological molecule, allowing investigation of structure, dynamics and thermodynamics. Here, we use MD to explore the change in flexibility associated with the predicted alanine mutations.

Each chosen alanine mutation (Q149A/E115A/I118A/P141A/K154A/D152A/D153A) was introduced into the open and closed conformational structures using PyRosetta (Section 4.2.1). Before full simulation, the water molecules were energy minimised around the static protein structure for 100 ps. The wildtype and mutant structures were then temperature and pressure equilibrated for 100 ps at 300 K. Molecular dynamics simulations consisted of 10×2 ns simulations for the wildtype and each mutant structure, in each conformation respectively. The repeats are necessary due to the stochastic nature of an MD simulation; single trajectories could easily be trapped inside a local energy minima. The final structures of each 2 ns simulation were subsequently compared to the original equilibrated structure to calculate the root mean square distance (RMSD) and root mean square fluctuation (RMSF). The RMSD is a measure of distance between the two structures, whilst the RMSF is a measure of the distance that each residue fluctuated from its original position.

Figure 7.13a presents the RMSD of a single 2 ns simulation for the wildtype and D152A mutant. The RMSD reaches a plateau after 0.5 ns suggesting that the system has reached temperature equilibration, thus allowing us to assume that any transient dynamics we observe after this point are at equilibrium. To probe the changes in flexibility associated with the mutations predicted by Markov stability we must investigate the RMSF values. Figure 7.13b presents the average RMSF values for the D152A, I118A and E115A mutant structures relative to the wildtype. The grey area represents 2 standard deviations around the wildtype RMSF average, calculated from the $\times 10$ repeats of the wildtype MD simulation.

It is immediately clear that large root mean square fluctuations are located in the lid regions, appropriately agreeing with previous MD simulations (Henzler-Wildman et al., 2007b). In the open conformation (2RH5) an increase in RMSF in the lid region can be seen for D152A and I118A, however, there is very little difference for E115A. To illustrate the positions of these fluctuations, the RMSF values have been mapped onto the open conformation structures in Figure 7.13c.

Looking at the open conformation (2RH5) we saw interesting results; the control mutations predicted to have a lesser effect showed results as expected, E115A exhibited very little difference

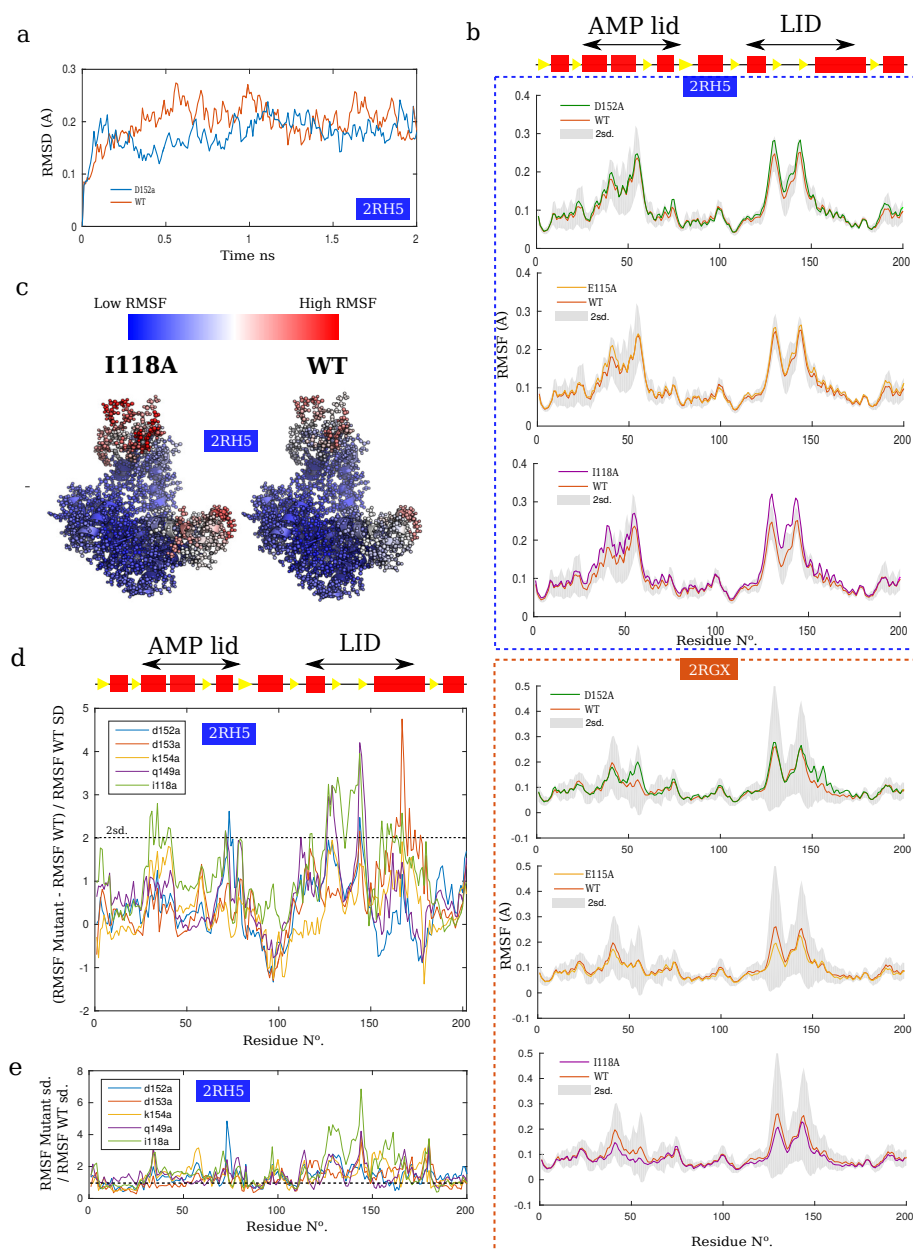


Figure 7.13: MD simulations initiated from the wildtype and alanine mutant structures. (a) Average RMSD values for the wildtype (orange) and D152A (blue) ADK structures over a 2 ns simulation. (b) The average backbone RMSF values as calculated relative to the original structures. The wildtype RMSF is shown as a guide for each mutant trajectory with the grey region representing 2 standard deviations either side of the mean. Increases in RMSF around the lid region are seen for D152A and I118A, whereas, only a very small change is seen for the E115A control mutation. (c) The RMSF values for the wildtype and I118A mutant are mapped onto the structures to indicate the regions of highest RMSF. As expected, the extremity of the lid and AMP bind domain exhibit the highest RMSF, however, the I118A mutant resulted in much larger RMSF values. (d) The difference between the mutant RMSF values and the wildtype RMSF values as a function of the wildtype standard deviation are plotted for each residue. This gives a measure of the change in RMSF. (e) The range of RMSF values as calculated using the ratio of mutant standard deviation to wildtype standard deviation tells us about any newly accessible local substates.

in RMSF values relative to the wildtype indicating that the mutation was not important in perturbing dynamics in the lid region (Figure 7.13d right). Interestingly, P141A exhibited very little change in the global flexibility of the lid region, despite having shown an increase in enzymatic rate. The mutation must be resulting in a faster rate of local conformational changes that allow a higher catalytic rate instead of affecting global dynamics. A very large deviation away from the WT trajectory was seen for I118A. Instead of just an increase in RMSF at the extremity of the lid, there is a large increase in RMSF of the whole lid domain. Increases in lid RMSF values were seen for D152A and Q149A, indicating that these mutations must be resulting in an increased flexibility to the lid region.

The MD results for the closed conformation (2RGX) showed less pronounced effects for the mutations. The wildtype exhibited a large standard deviation in RMSF for the lid region between the 10 repeats, however, very little difference in RMSF was seen for each mutation. In fact, the mean RMSF of the lid region was very slightly lower for all mutants except D152A. It is likely that the absence of the salt bridge between Arg36 and Asp152 resulted in a higher RMSF for D152A, reducing the free energy barrier to the open conformation. This indicates that stability of the closed conformation is defined by the interactions between the lid sub-domain and other neighbouring domains and not by internal changes.

7.3. A summary of the predicted mutations

In this section, we have performed an alanine mutagenesis scan across ADK and performed a Markov stability analysis of each mutant structure. Using the computational calculations we were able to score and rank each alanine mutation according to its perturbation to the community structure. Mutations were scored at each relevant Markov timescale and a selection of alanine mutations were physically constructed and their biophysical properties were measured. Here, we attempt to summarise and corroborate the theoretical and experimental results for each alanine mutation.

7.3.1. Control mutations (E115A/Q149A)

The control mutations, E115A and Q149A, were awarded relatively low scores by Markov stability across all timescales. Glu115 is solvent facing and located on α -helix-6, mutation to a hydrophobic alanine introduces the possibility of large solvent effects, however, there was no change in the FRET populations, catalytic activity or flexibility which corroborates with a highly variable residue position (low conservation). Gln149 was in close proximity with the aspartic acid pair, located between β -bridge-2 and 3/10-helix-7, however, it didn't exhibit a

large effect on the conformational equilibrium. This highlights the ability of Markov stability to discriminate between spatially close residues.

7.3.2. P141A

A particularly interesting mutation was P141A, located near the top of the lid subdomain. A comparison of Aquifex (thermophilic) and E.Coli (mesophilic) AK_{aqu} has previously shown that the inclusion of prolines in the thermophile acts to stabilise at higher temperatures but also reduce catalytic activity at lower temperatures (Daily et al., 2011; Henzler-Wildman et al., 2007a). P141A was predicted to have a perturbative effect at short timescales in the open and closed conformations by Markov stability, in fact, this was mimicked by Pro140 and Pro143 which were also predicted at short timescales in the open and closed conformations. The activity of P141A ADK increased which, from a biological perspective and based on literature studies, was expected (Daily et al., 2011; Henzler-Wildman et al., 2007a).

The molecular dynamics simulations showed no significant change in the flexibility of the lid domain after P141A mutation. However, this may be a result of a more local change in flexibility that didn't alter the flexibility of the entire lid domain. The prediction by Markov stability at short timescales also agrees with this hypothesis, indicating it is having a local effect within the floppy lid sub-domain. A consequence of the increased conformational entropy associated with proline removal is the increased catalytic rate, it is likely that the shift to a majority closed population (75%) may also be a result of increased conformational entropy in the open conformation.

7.3.3. I118A

I118A was ranked highly by Markov stability at medium Markov times in the open and closed conformations. It is located on alpha-helix-8 and its side-chain is located to the rear of the protein but slightly angled between the two lid-connector helices. In the closed conformation it connects to Asp114 and Glu115 through strong and weak hydrogen bonds respectively. The connection with Asp114 is maintained in the open conformation but there is a wealth of hydrophobic interactions with neighbouring residues. Ile118 is not a well conserved residue and upon mutation to alanine there was no change on the catalytic activity. However, there was a shift in the FRET populations to a majority closed state with a suitably sized minority open state. The molecular dynamics simulations indicated an impressive increase in flexibility of the open conformation but little change in the closed state. It is likely that the associated removal of hydrophobic interactions within alpha-helix-8 has resulted in an increase of independent movement between the two lid connector helices. Ile118 has been implicated in a hydrophobic pocket in previous studies (Rundqvist et al., 2009) that found mutation to glycine increased the lid

domains propensity to unfolding and, therefore, accessing the transition state to closing. In fact, further measurements by [Olsson and Wolf-Watz \(2010\)](#) found that mutation of residues in the same hydrophobic cluster to glycine resulted in a near 1 populated closed state. It is likely that our mutation of I118A is yielding similar results, shifting the equilibrium population towards closed. In contrast to [Olsson and Wolf-Watz \(2010\)](#) who observed an increase in binding affinity, we observed no change in catalytic rate. This disagreement is likely a result of our mutation to alanine instead of glycine.

Looking at the Markov stability community structures in detail we find that the lid connector helices only form a single community at later Markov times in the I118A mutant, indicating that there is reduced communication between the two helices which is resulting in the high Markov stability ‘score’. This disconnectivity between the two lid connector helices is reducing the lid molecular stability in the open conformation which would explain the shift in population towards a closed conformation and it also explains the relative free motion of the two alpha helices as observed by MD. However, the I118A mutated protein can still access the open and closed conformations, allowing it perform its catalytic function at a similar rate to that of the WT.

7.3.4. K154A

Lys154 is a positively charged solvent facing residue found between the 3/10-helix-7 and helix-8. It extends a strong hydrogen bond with Glu151 in the open conformation which is not present in the closed conformation. Markov stability calculations attributed K154A a high score at short timescale. In a similar regard to P141A, FRET measurements showed a slight shift towards a closed conformation. However, what was most striking was the 27% increase in catalytic activity, whilst not as significant as the increase observed for P141A, it was unexpected from a biophysical perspective. The increase in catalytic rate is accompanied with the slight increase in RMSF observed in the open conformation MD data, however, this doesn’t fully explain the mechanism of this mutation. It is likely that mutation of Lys154 is reducing a steric barrier to the rate-limiting opening of the lid, however, further experiments would be needed to validate this.

7.3.5. D152A & D153A

The final, and potentially most interesting, mutations are the aspartic acid pair: D152A and D153A. Mutation of either residue exhibited similar experimental results; both shifted towards a significantly closed conformation and both lost catalytic activity. This is likely a consequence of their similar spatial position and identical amino acid type. Asp153 and Asp152 are located

on a 3/10 helix at the top of alpha-helix-8 and the bottom of the floppy lid. In the open conformation, Asp153 binds to Arg124 and Arg150 through hydrogen bonds which are supplemented by a strong hydrogen bond with Arg161 upon transition to the closed conformation. On the otherhand, Asp152 extends multiple strong hydrogen bonds to Arg36 and Arg150 in the closed conformation which are broken during transition to the open conformation whereby a new hydrogen bond is formed with Arg161. The common factor is the hydrogen bonding with Arg150 in the closed conformation which has been recently implicated in significant reduction of activity after mutation (Kerns et al., 2015).

The Markov stability analysis showed high scores across the majority of timescales in both conformations for both D152A and D153A. A more in-depth analysis of the Markov stability community structures found that the boundary between the floppy lid community and the lid connector helices for the mutant structures had moved relative to the wildtype structure. The aspartic acids which originally grouped themselves with the lid connector helix community instead now form a community with the floppy lid region. These results suggest that Asp153 and Asp152 are acting as stabilisation residues in the lid domain and removal has introduced new, or softened existing, hinges between the lid connector helices and the lid. The de-stabilisation/hinge softening between the lid and lid connector helices for both the D152A and D153A mutants manifests itself through a highly closed state with a near 1.0 population.

An interesting observation of the Markov stability analysis for D152A was the large peak in $\overline{\text{VI}}$ that indicated a reduction in robustness of the 3-way partition. Comparing this peak with the substrate bound closed conformation in Chapter 6, we observe a similar peak at the same Markov times that was unseen in the closed apo conformation. This finding suggests the D152A ADK mutant acts in a similar manner to the substrate bound conformation. Indeed, we observe this experimentally - the wildtype in the presence of substrate shifted towards a closed conformation and, in a similar manner, the D152A mutant also shifted towards a closed conformation.

7.4. Discussion

In this chapter, we have introduced a mutagenesis scoring method that utilises the unique community structures found using Markov stability. This has allowed us to directly identify mutations in ADK that cause a large perturbation to biologically important community structures at different timescales. The scoring method is a combination of two Markov stability metrics. We first looked at the robustness of the community structure using the Variation of Information (VI). By calculating Markov stability across the ensemble of alanine mutations we are able to calculate the mean $\overline{\text{VI}}$, and its associated variance, using a Gaussian Process Regression. Comparing each mutant $\overline{\text{VI}}$ trajectory with the ensemble mean, we can identify mutations that fluctuate significantly away from the average. The second metric is the absolute change to the

optimised community partition at each timescale measured by the community variation (CV). We found that some mutations only changed one metric, whilst some mutations perturbed both, highlighting the importance of combining the two metrics into a single mutant score.

Using the two metrics we are able to combine them using the root mean square distance (RMSD) into a single mutant score X_i . This score tells us the extent to which a mutation effects the community structure during that timescale. We find that mutations which were originally only identified as perturbative using a single metric, were now scored highly using the combined methodology. We have demonstrated the robustness of the mutant scoring methodology using the Kullback-Liebler method. By assuming that the $\overline{VI}$ and CV modelled the standard deviation (σ_m) and the difference in means ($\mu_{mutant} - \mu_{wt}$) respectively, we could calculate a KB value for each mutant. There were clear correlations between the RMSD and KB scoring methods, suggesting that the original method of combining $\overline{VI}$ and CV was suitably robust.

To explore the effect of mutant scores on the physical attributes of ADK we constructed a set of lid mutants. Each mutant had been attributed different mutant scores at different timescales. Using single-molecule FRET we were able to measure changes to the conformational equilibrium and found that high scoring mutations at long timescales would result in large changes to protein dynamics. Looking at short timescales, we identified a couple of mutations that increased the catalytic rate of ADK, suggesting that local perturbations would increase the opening/closing rate of the lid. Using control mutations that didn't score highly we were able to show that no physical attributes were perturbed, highlighting the robustness of the mutant scoring method.

Our mutant scoring method is a computationally efficient method to study the many potential mutations in biomolecules using the complete atomistic physico-chemical structure. The Markov stability mutagenesis provides a unified understanding of the mutations that span different timescales, and therefore offering a robust way to look for candidate residues with important structural, dynamical and functional roles in proteins. In particular, this method can be used to identify residues that will perturb the global subdomain dynamics.

8. Validation of mutant predictions during the 3-way partition in ADK

In the previous chapter we used Markov stability to identify mutations that were deemed significant at each timescale. We saw a wide range of physical effects associated with these predictions, from changes in catalytic activity to shifts in the populations equilibrium. This highlighted the power of Markov stability to identify important mutations at all timescales. However, it was difficult to build a relationship between the physical effects and the computational predictions when the experimentally measured proteins were associated with different temporal scales. In the interest of building a fully validated relationship, we need to look at a single dynamical timescale, in particular a timescale that is accessible using single-molecule FRET.

The 3-way partition was the focal point of Chapter 6 because it models the 3 biophysical subdomains of ADK. We showed the single-molecule fluorescence measurements were effective at probing the motions associated with the 3 subdomains. In this section, we focus our effort on the 3-way community partition, whilst still focussing on the lid subdomain to probe conformational dynamics.

The most significant lid mutation observed in Chapter 7 was D152A, which saw a large shift towards a closed conformation as measured by smFRET. This section explores Asp152 further; through the introduction of different residue types at position 152 we can predict and experimentally validate changes to the conformational equilibrium. A mutant scan across all residue-types at position 152 gave a straight-line relationship between the Markov stability score attributed to each mutant and the ratio of the population equilibrium measured using smFRET.

To further validate our methodology and to showcase our ability to mutate to residues other than alanine, we perform an arginine scan across the ADK structure. We identify and experimentally explore V164R, a high scoring mutation during the 3-way partition. In agreement with the theory-experiment relationship defined by mutations at Asp152, we find a shift towards a closed conformation with smFRET.

8.1. Substitution of any residue type *in silico*

In reality an alanine will have side-chain interactions, albeit fewer, and the local side-chain packing will change dependent on the residue type. In the previous chapter we described a graph theoretical approach to alanine scanning where all side-chain interactions were removed to model a mutation to an alanine. In contrast, this section implements *in silico* substitution of residues directly into the PDB structure file, allowing us to introduce residues of any type at any position. A graph representation of the mutant protein structures can be constructed and subsequently analysed efficiently using Markov stability.

Computational mutations were introduced using PyRosetta, a python-based interface to the Rosetta molecular modelling suite (Chaudhury et al., 2010). The PDB was initially converted to a *pose* object and the PDB residue numbering was imported. The *mutants.py* function was used to replace residues at a chosen position and a PDB file was ‘dumped’ with the mutation present, further details are available in Section 4.2.1. Large side-chain residues can introduce steric clashes, to account for this we energy minimise each mutant PDB structure using GROMACS (Abraham et al., 2015), as described in Section 4.2.2. The water molecules, introduced during energy minimisation, are removed from the PDB file using PyMol (DeLano, 2002). Finally, a protein graph representation of the mutant structure can be constructed directly from the PDB file.

8.2. Building a relationship between theoretical scores and smFRET measurements

8.2.1. A Markov stability analysis of computational residue substitutions

Since we are interested in global dynamics, it is only natural that we choose the 3-way partition (corresponding to the three sub-domains) as our timescale. As discussed in Chapter 7, the D152A mutation was given a significantly high score during the 3-way partition of the closed conformation, which was reflected experimentally by a large shift in the conformational equilibrium, a high flexibility as measured by MD and a complete loss of catalytic activity. Therefore, in this section we have used Markov stability to explore the systematic introduction of *every* residue type at position Asp152 on the dynamical equilibrium. Markov stability calculations were performed from $t = 10^2$ to $t = 10^5$ with 400 timesteps for each mutant.

Figure 8.1 highlights the large differences in the size of the $\overline{\text{VI}}$ peak during the 3-way partition. It is immediately obvious that particular mutations, such as D152A, exhibit a large peak in $\overline{\text{VI}}$ relative to the wildtype indicating a reduced robustness of the community structure. In fact, a

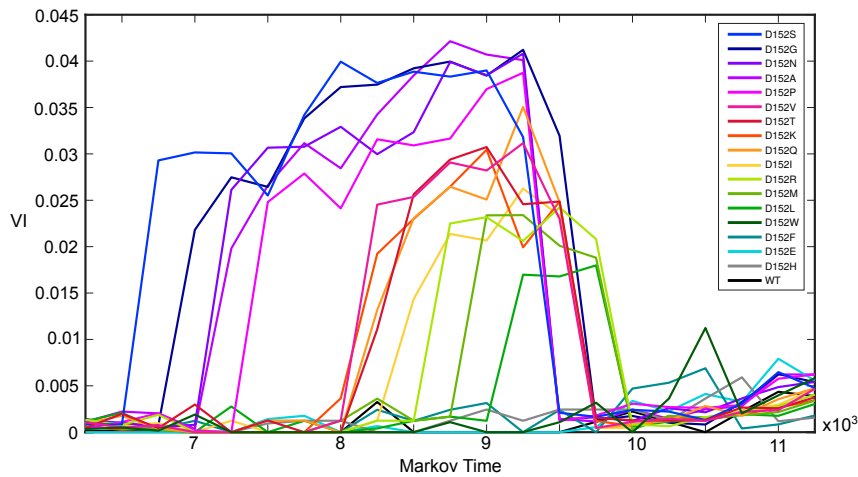


Figure 8.1: Peaks in $\overline{VI}$ exhibit the different effects of D152 mutants. A zoom-in of the $\overline{VI}$ peaks observed during the 3-way partition of the Markov stability analysis of each D152 mutant. The graphs are coloured in accordance with the average size of the peak, from blue (largest peak) through to black (smallest peak).

group of residues that include serine, glycine, asparagine, alanine and proline all exhibit a large change in $\overline{VI}$. On the other hand, some mutations such as D152E and D152W show no peak in $\overline{VI}$ indicating that there is no change in robustness relative to the WT. Importantly, there are varying reductions in community robustness made apparent from the wide range of $\overline{VI}$ peak sizes.

8.2.2. Measuring the significance of Asp152 mutations using mutant scoring

Here, we employ the same mutant scoring criterion as Chapter 7 i.e. scoring mutations according to their community variation (CV) and variation of information ($\overline{VI}$). We have calculated the Markov stability scores, X , and ranked each mutant accordingly in Figure 8.2. We observe a similar trend in the ranking of residues to the large peaks in $\overline{VI}$ that were observed in Figure 8.1.

The mutations that exhibit the highest $\overline{VI}(t)$, i.e. the largest peak in $\overline{VI}$, also tend to be smaller residues (serine, glycine, alanine etc.). Why do these smaller residues introduce a reduction in robustness of the community structure at long timescales? The original aspartic acid residue in the wildtype extended multiple salt-bridges/hydrogen bonds to Arg36 and Arg150, accompanied by hydrophobic interactions with Tyr52, Met53 and Glu54 located on the AMP_{bind} subdomain. However, mutation to smaller residues removed the majority of these strong side-chain hydrogen/salt-bridge interactions. Instead, the smaller residues substituted at position 152 only made low energy interactions with neighbouring residues: D152S no longer extends any inter-subdomain interactions, whilst D152P only extends weak hydrophobic interactions with Met53 and Arg36 on the AMP_{bind} subdomain.

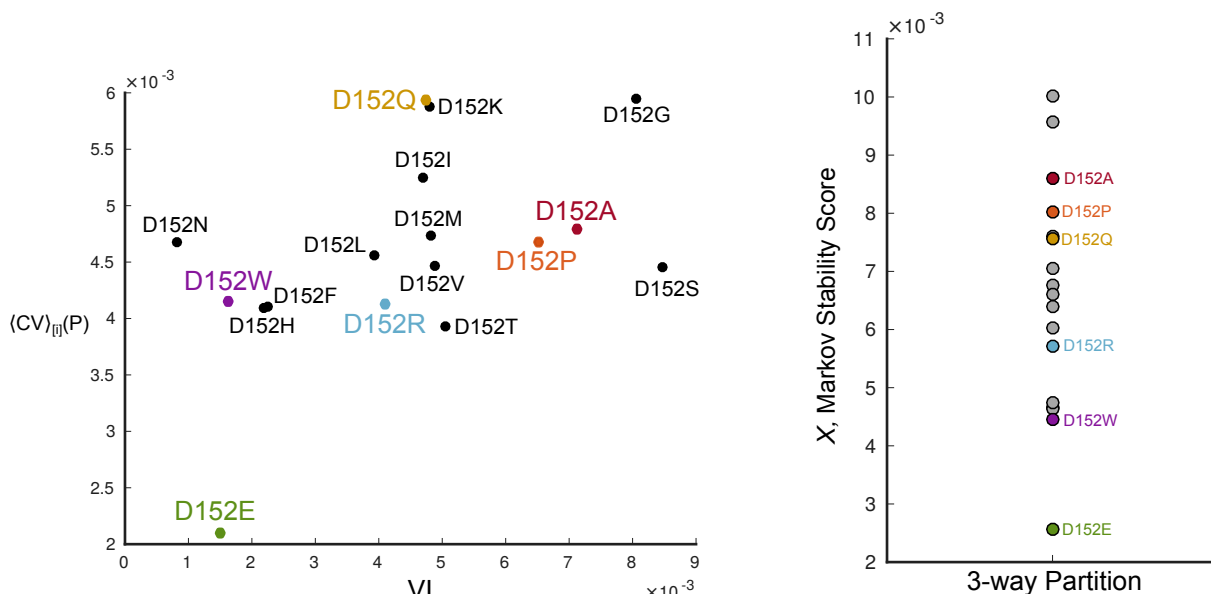


Figure 8.2: Scoring D152 mutations according to $\overline{VI}$ and CV. A scatter plot of the $\overline{VI}$ and CV for each D152 mutant which allows us to calculate the Markov stability score X .

Residues that presented a medium $\overline{VI}(t)$ relative to the ensemble, such as D152K, D152Q or D152L, showed a higher number of interactions with the AMP_{bind} subdomain relative to the ‘smaller’ residues, albeit low energy interactions. For example, D152I shared multiple hydrophobic interactions with the AMP_{bind} subdomain (Arg36, Glu52, Met53). Of notable interest is D152R, the introduction of a positively charged residue at position 152 would result in electrostatic repulsion with Arg150, Arg161 and Arg36. In contrast to the coordinated O atoms of the aspartic acid, the energy minimised D152R structure shows the arginine side chain to repel Arg36 such that their side-chains are anti-parallel. D152R coordinates with the AMP_{bind} subdomain through a hydrogen bond with Met53 and various hydrophobic interactions.

As expected, mutation of the aspartic acid to a glutamic acid saw no change in the community robustness $\overline{VI}$. Since they are both negatively charged residues we observed similar salt bridge interactions with the AMP_{bind} subdomain: Glu152 OE1 : Arg36 2HH1 (725 kcal/mol) / Glu152 OE1 : Arg36 2HH2 (72 kcal/mol). Despite the similarities between the two residues, glutamic acid is rarely substituted for the aspartic acid at position 152 according to sequencing results. This suggests that whilst the D152E mutation may not perturb long-timescale ADK dynamics, it may result in inhibitory effects. Another low $\overline{VI}$ mutation was D152F, the large bulky hydrophobic phenylalanine sidechain introduced a large number of hydrophobic interactions (to Leu35, Arg36, Val39, Lys50, Met53, Glu54). Despite not extending any strong salt bridge or hydrogen bonds, the sum of the many hydrophobic interactions of D152F accounts for the robust community structure as seen in the original aspartic acid.

8.2.3. Conformational dynamics of ADK D152 mutants measured by smFRET

The D152A mutation, which was scored highly by Markov stability, resulted in a large shift in the population equilibrium. Therefore, we can hypothesise that a higher Markov stability mutant score is linked to large changes in the conformational equilibrium. To investigate the relationship between the Markov stability score and the conformational equilibrium we have physically constructed six differing D152 mutant proteins using QuikChange mutagenesis. The 152-aspartic acid residue was substituted with:

- **1. Alanine.** This mutant was investigated in Chapter 7. The residue is small and hydrophobic. It exhibited a large change in $\overline{VI}$ during the 3-way partition.
- **2. Arginine.** Bulky and positively charge with a middling sized peak in $\overline{VI}$.
- **3. Glutamic Acid.** The control mutation with a negative charge and no change in $\overline{VI}$.
- **4. Glutamine.** A polar residue with a similar size to the aspartic acid but with a middling/large peak in $\overline{VI}$.
- **5. Proline.** A small hydrophobic residue which reduces the local entropy. It exhibited a large peak in $\overline{VI}$.
- **6. Tryptophan.** A bulky, hydrophobic residue with an aromatic group that exhibited a very small change in $\overline{VI}$.

DNA sequencing with Genewiz (NJ, US) was used to confirm that the substitutions were successful. The mutations were chosen to effectively sample high, medium and low Markov stability scores (i.e. some with larger and lesser fluctuations) and also provide diversity between the type of amino acids based on charge, bulkiness and polarity.

8.2.4. High resolution Markov stability analysis of chosen mutants

The Markov stability calculations for these chosen mutants (D152P, D152Q, D152W, D152E, D152R, D152A) were repeated over $t = 10^3$ to $t = 3.16^4$ with 400 timesteps and 1000 Louvain optimisations to improve the resolution and accuracy of the calculations over the 3-way partition. Figure 8.3 presents the high resolution Markov stability calculations for the six D152 mutants and wildtype. As expected, when compared to the lower resolution Markov stability calculations in Section 8.2.2 we see very similar results for every mutant. These results both highlight the reliability of Markov stability as a methodology and also provide more accurate results with experimental findings.

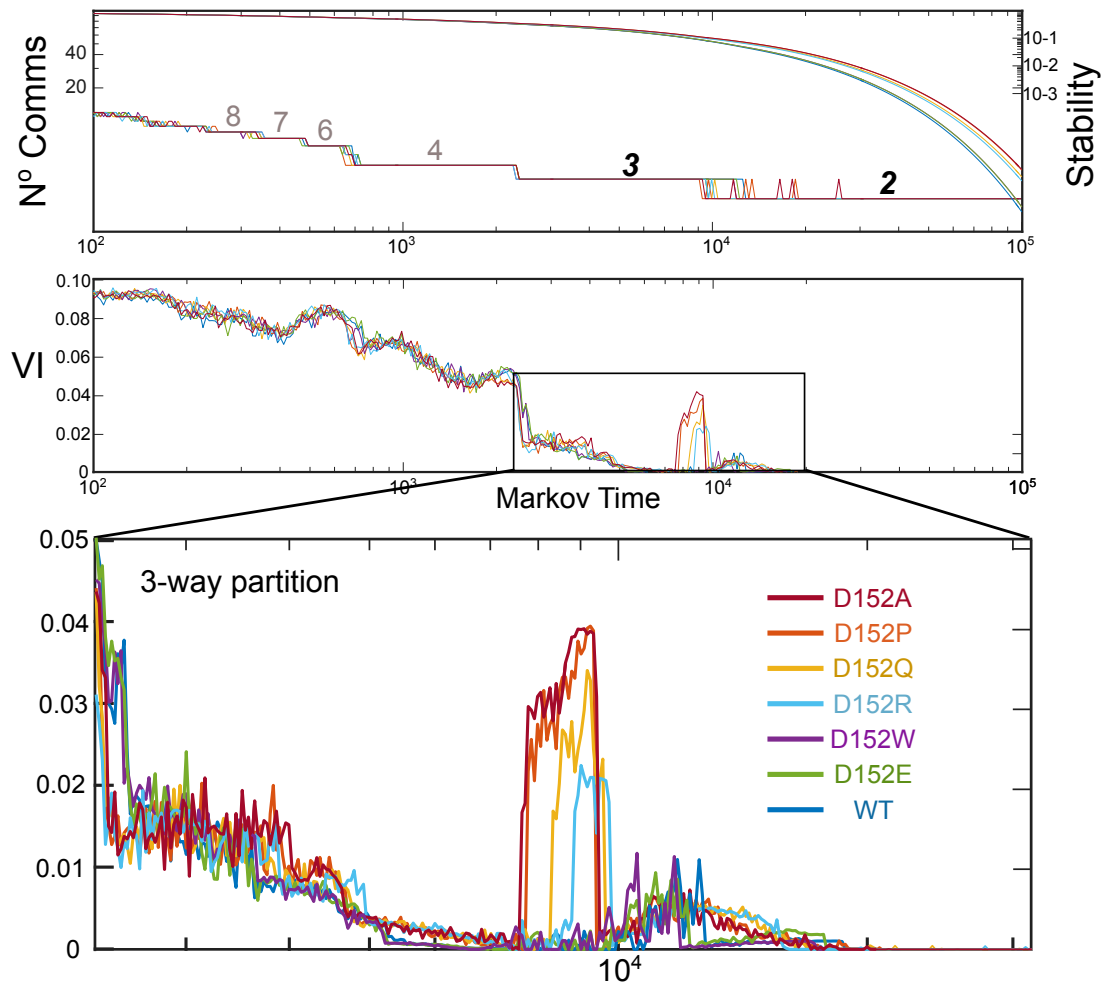


Figure 8.3: Markov stability analysis of D152 mutants. The high resolution Markov stability calculations for the D152 mutants. Magnified is the fluctuation of the $\overline{VI}$ across the 3-way partition over which the $\overline{VI}$ is averaged to obtain a ‘score’.

Whilst we observe large changes in the $\overline{VI}$ during the 3-way partition, there are also changes to the optimal number of communities shown in the upper panel of Figure 8.4. For example, D152A drops out of the 3-way partition at earlier Markov times relative to the wildtype, and then fluctuates between the 3-way and 2-way partitions over Markov time 1×10^4 and 3×10^4 . This suggests that the D152A mutant structure is unable to adopt a robust community structure. Similarly, D152P, D152Q and D152R all exhibit a reduction in robustness that leads to a less persistent 3-way partition and flip-flopping between the 3-way and 2-way partitions.

The Markov stability values (top panel Figure 8.3) provide a relative measure of the strength of a partition according to how well trapped the random-walker is within a community. The ‘strongest’ partition is D152A, followed by D152P, D152Q, D152R, D152W, D152E and finally the wildtype in decreasing order. The wildtype had the weakest partition, where the random-walker could easily escape a community, suggesting that there was communication between the

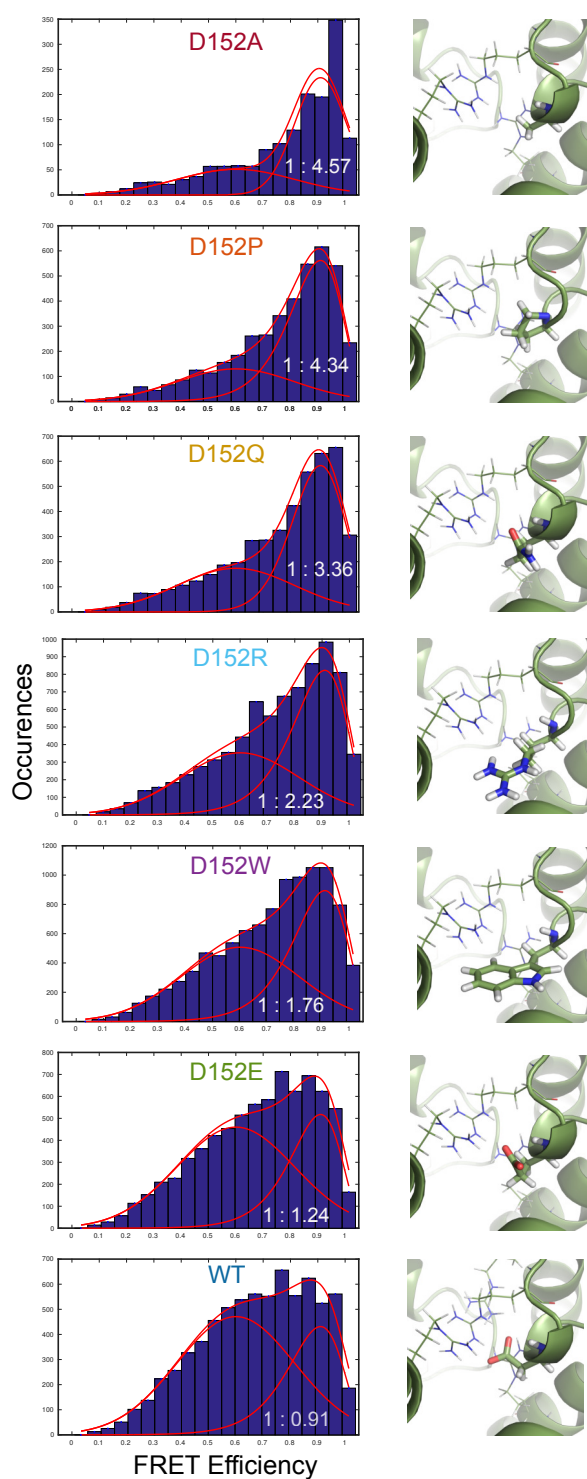


Figure 8.4: Single-molecule FRET efficiency histograms of the chosen set of D152 mutants. The FRET histograms exhibit a shift in ratio between the open and closed conformations as a result of mutation. The histograms are fitted with a two-state distribution mixture model. In-set is an illustration of each respective D152 mutation.

separate subdomains. However, increasing Stability suggested that this communication between the separate subdomains is weakened or removed. Li et al. (2015) have looked at the dynamic coupling of the subdomains in ADK and found that the lid can only reach the open state after the AMP_{bind} has opened as a result of strong salt bridges. Weakening or removing inter-subdomain interactions through mutation is reducing the communication between the subdomains and also reducing the dynamic cooperativity as measured by smFRET.

8.2.5. FRET efficiency histograms of D152 mutants

The FRET efficiency of each mutant was measured and fitted with the same parameters defined by the wildtype, Figure 8.4. The ratio between the open and closed conformations is reported on each histogram for comparison against the wildtype and the FRET histograms are ordered by their ratio from lowest to highest. The alanine mutation was previously measured in Chapter 7, we saw a large shift towards a closed conformation (open:closed 1:4.57) which mimicked the conformational equilibrium we observed using an osmolyte (TMAO). Looking at the remaining mutations, it is clear that there are differing shifts in the population equilibrium dependent on residue type. These results further highlight the fragility of the population equilibrium in ADK if single point mutations at D152 can cause such global dynamical changes.

The control mutation D152E, exhibited a near 1:1 population ratio of open to closed states in a similar manner to the wildtype structure. Suitably, this agreed with the Markov stability analysis that attributed D152E a low mutant score. Substitution of Asp152 with a tryptophan residue caused a slight shift towards a closed conformation; Markov stability found D152W to show slight fluctuations in $\overline{VI}$ (Figure 8.3), but overall the Markov stability score was relatively low. A larger shift in the conformational equilibrium was observed for D152R, followed by D152Q, D152P and finally D152A.

To present a visual relationship between fluctuations observed in Markov stability score and smFRET we plotted the ratio of the open and closed states for each mutant against the mutant score across the 3-way partition. Remarkably, we found a very strong correlation between the mutant score across the 3-way partition and the open to closed population ratio. A linear fit ($R^2 = 0.88$) is made for clarity in Figure 8.5, however, it is the presence of a monotonic relationship which is most spectacular.

The relationship in Figure 8.5 tells us that an increase in the mutant score during the 3-way partition is correlated with an increase in the population of the closed conformation. Whilst the correlation is strong evidence for a working model of the ADK system, it also provides us with an actionable relationship. We can use the correlation to predict the shift in conformation of the remaining residue types. For example, D152F was given a low score and we would expect a similar FRET histogram to the wildtype, whilst D152G scored highly and we would expect

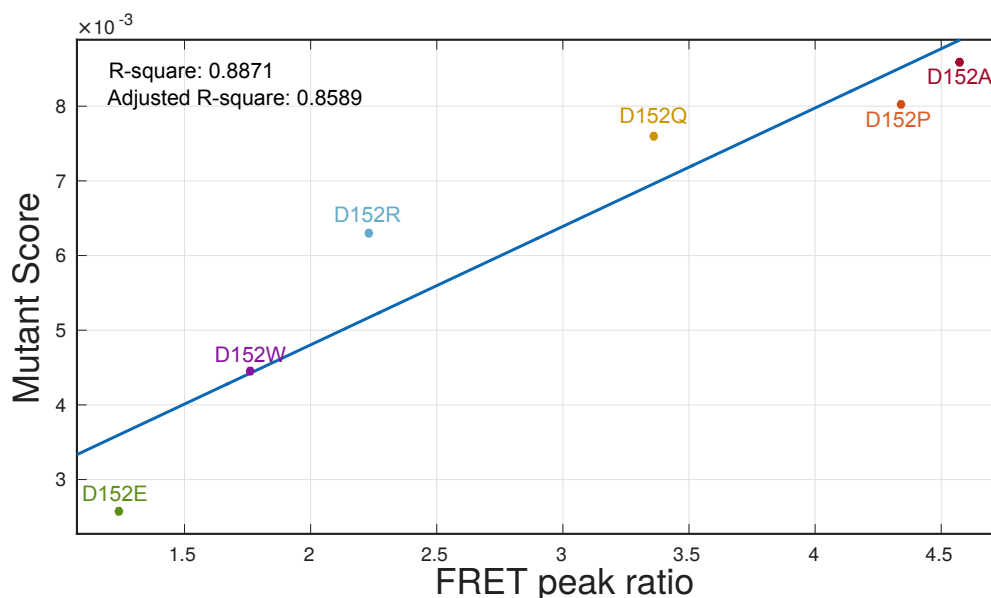


Figure 8.5: A correlation between the Markov stability mutant score across the 3-way partition and the shift in the conformational equilibrium as measured by smFRET. A linear regression with $R^2 = 0.88$ has been fitted to showcase the relationship between the mutant score and the ratio of the open and closed states.

a large shift towards the closed conformation. We are able to effectively control the global dynamics of ADK using a graph theoretical method.

Whilst we observe a correlation between the ratio of the conformational equilibrium and the Markov stability mutant score X . It is important to look at the separate metrics, $\overline{VI}$ and CV , that comprise the mutant score X . Figure 8.6 exhibits a scatter plot of FRET peak ratio versus $\overline{VI}$ and versus CV . It is clear that there is a distinct correlation with $\overline{VI}$ that is even stronger ($R^2 = 0.93$) than the correlation with the mutant score. However, the CV doesn't show a good correlation with the FRET peak ratio. This suggests that the relationship we observe for the mutant score is driven by the $\overline{VI}$. It is clear that the robustness of each mutant community structure ($\overline{VI}$) is a critical metric in defining the effect of a mutation on long-timescale dynamics, this contrasts with the KL scoring method.

8.2.6. Enthalpic changes drive the correlation between mutant score and the conformational equilibrium

But what is causing the varying shift in conformational equilibrium between different residue substitutions? Is it a consequence of varying levels of unfolding? In Chapter 7 we explored the mechanism for the D152A shift in conformational equilibrium. We found that Asp152 was located on a 3/10 helix and hypothesised that we were observing a local unfolding event. It is likely that the reduced number of inter- and intra- subdomain interactions associated with

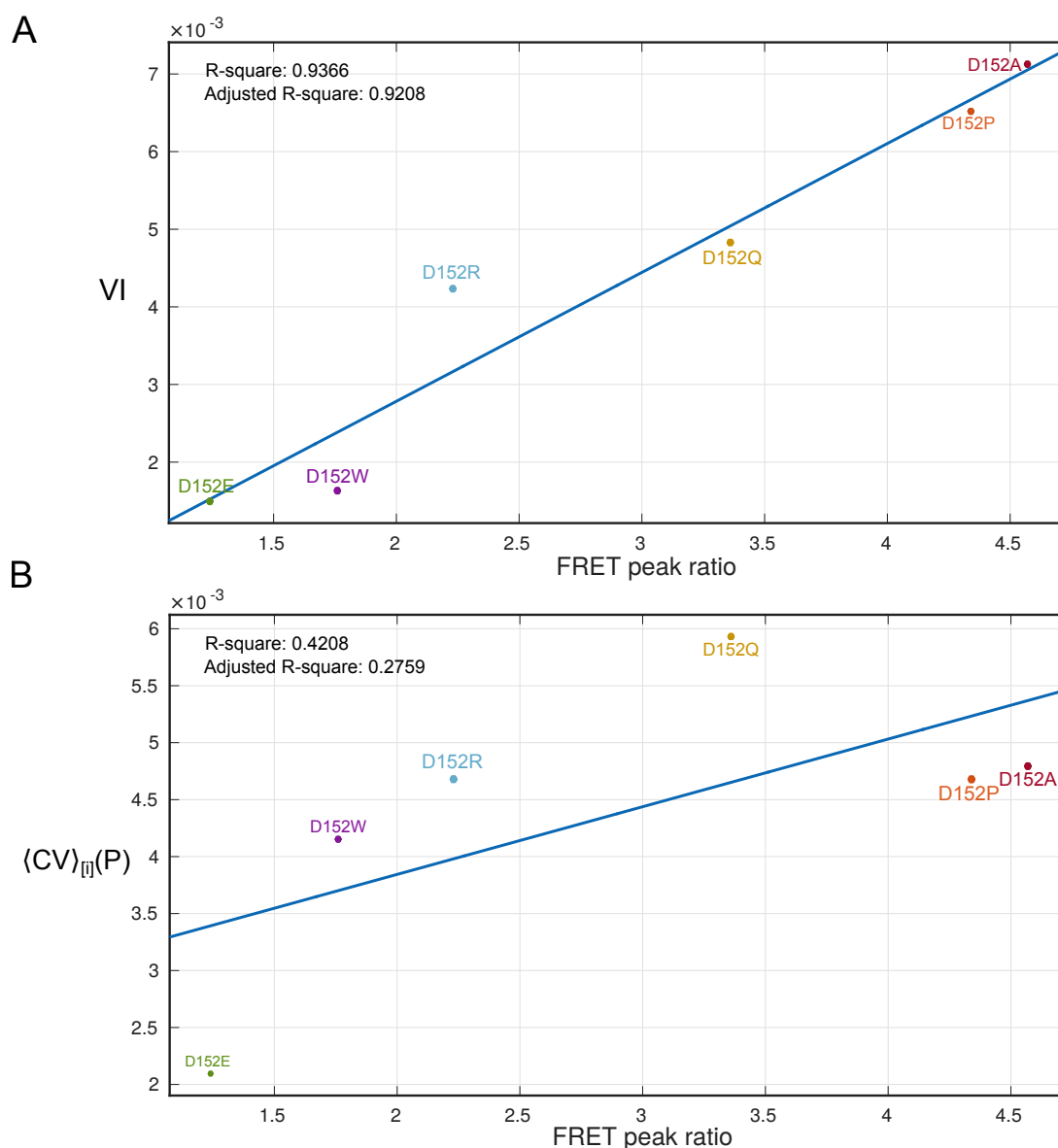


Figure 8.6: A correlation between the shift in the conformational equilibrium as measured by smFRET and $\overline{\text{VI}}$ and CV. (a) A scatter plot of $\overline{\text{VI}}$ versus FRET peak ratio finds a strong correlation whilst (b) a scatter plot of CV versus FRET peak ratio exhibits no useful correlation.

alanine and proline substitutions results in a destabilised lid region, whilst tryptophan and glutamic acid still form a large number of interactions.

To analyse the change in interactions, and its associated effect on local unfolding, we can calculate the total enthalpy ΔH of each substitute residue as the sum of bond energies (in *kcal/mol*). Figure 8.7a shows a monotonic relationship between the number of interactions that the substituted residues extends with the shift in the conformational equilibrium (measured by smFRET). A decrease in the number of interactions causes a shift towards a closed conformation, suggesting that the further we destabilise the local structure the further we will observe a shift in the

population equilibrium. Figure 8.7b exhibits a 3D scatter plot of the Markov stability mutant score vs. ratio of the conformational equilibrium vs. the total enthalpy of the mutant residue. It is clear that local changes in enthalpy are the cause of mechanistic perturbation in ADK, and it is this spatial change in enthalpy that we can detect through Markov stability.

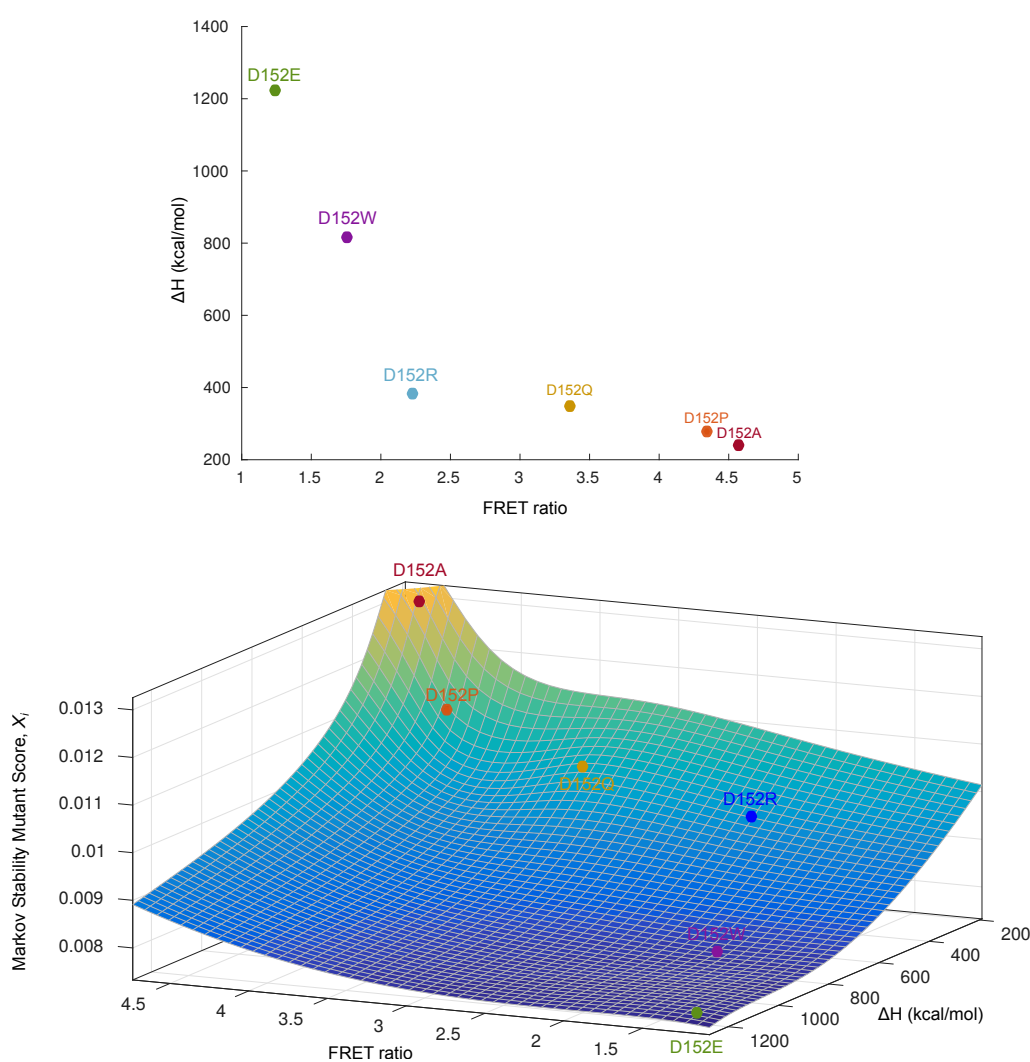


Figure 8.7: Exploring the relationship between the mutant score, the total bond enthalpy of mutants and the change to the conformational equilibrium. (a) A scatter plot of the total bond enthalpy of each mutant residue versus the ratio of open and closed states measured by smFRET. (b) A 3D scatter plot of FRET ratio vs. mutant score vs. change in enthalpy associated with the mutation. The scatter plot is fitted with a thin-plate spline computed from p where x is normalized by mean 2.917 and std 1.384 and where y is normalized by mean 547.7 and std 391.2.

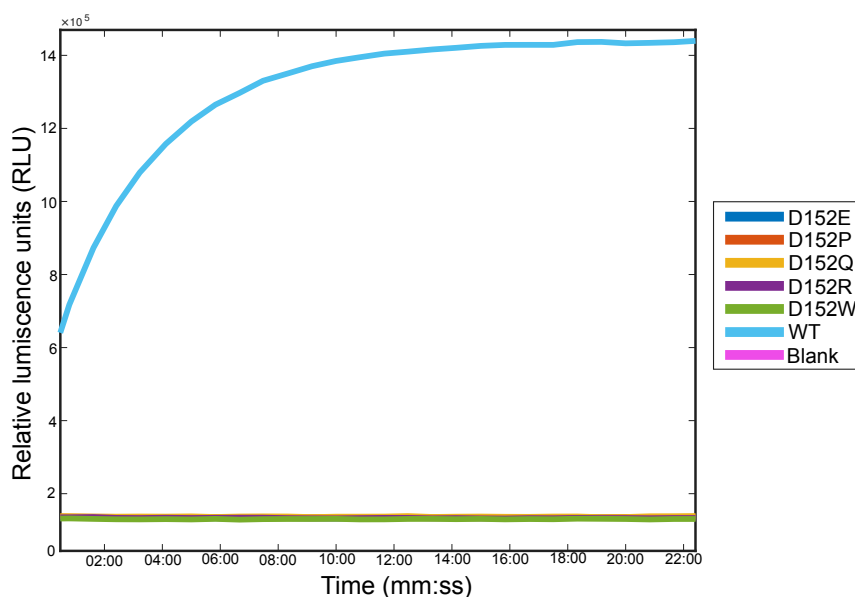


Figure 8.8: Activity assays for the D152 ADK variants. A 20 minute window of the ~ 8 hour luminescence activity assay. It is clear that all D152 mutants exhibit inhibition relative to the wildtype.

8.2.7. The role of the conformational equilibrium in catalysis.

Previous studies have shown that shifts in the conformational equilibrium result in large changes to catalytic activity (Nagarajan et al., 2011). A population shift towards the open conformation increases k_{cat} and a shift towards a closed conformation caused a significant reduction of k_{cat} (Ådén et al., 2012). These results demonstrate that the conformational equilibrium directly affects enzymatic catalysis.

To further explore the role of the conformational equilibrium in functional catalysis we performed an activity assay of the D152 mutants. Enzymatic activity of the D152 mutants was measured using Kinase-Glo luminescent kinase assay. The assay, Figure 8.8, shows the complete inhibition of all D152 mutants relative to the wildtype structure. Despite being a control mutation and holding similar properties, D152E also showed a complete lack of activity. These results suggests that changes to the conformational equilibrium are not the defining reason for inhibition, instead Asp152 is intrinsically important for catalysis as supported by its high conservation.

8.2.8. Comparing the $\overline{VI}$ ensemble of partitions uncovers communication between subdomains in D152 mutants

To understand the relationship in the previous section we must accurately analyse the Markov stability score that each mutant was attributed. In particular, we must understand why the higher score results in an increased propensity for the closed conformation and not the open

conformation. To do this we can look at the ‘hidden’ Louvain optimisations that are used to identify the optimal community partition.

The $\overline{VI}$ is calculated from a large number of Louvain optimisations at each Markov time (1000 partitions were Louvain optimised in these results) where the random-walker explores the protein structure and attempts to optimise the Stability function. Due to its stochastic nature and the use of a greedy algorithm, Markov stability will sometimes find a locally optimised partition in partition space. The $\overline{VI}$ across these 1000 partitions gives us a measure of partition robustness. A peak in $\overline{VI}$, as found for D152A and less so for D152R, indicates that the optimised partition is not robust and other partitions of the protein are also viable candidates. To understand the root cause of the $\overline{VI}$ fluctuations, Figure 8.9 shows the distribution of optimised partitions for the wildtype, D152R and D152A.

Looking at the 1000 partitions we actually find that D152A and D152R both find another set of locally optimal partitions that actually start to mix the core domain with the AMP_{bind} domain, these partitions appeared about 45% of the time in D152A and 15% of the time in D152R. Therefore, in this scenario the peaks of $\overline{VI}$ are a result of the AMP_{bind} domain merging with the core domain. This indicates an increased closeness between the subdomains that were originally separate in the wildtype structure. A similar analysis of the 2-way partition yields agreeable results (Appendix F); the lid community begins to encroach on the core domain.

The subdomain’s communities are becoming more closely connected after mutation which lends itself towards a more closed conformation. This is exactly what we observe in the FRET histograms, the lid and AMP_{bind} subdomains moved towards each other. This suggests that increased communication (by proxy of the random-walker) between the subdomains is a measure of dependence i.e. the subdomains will act to an extent as a single unit and won’t fluctuate relative to each other.

8.2.9. Multidimensional scaling reveals the differences between wildtype and mutant partitions

Whilst we have compared the 1000 Louvain partitions within each protein structure, we must also explore the relative differences between structures. We use the variation of information as a measure of pairwise distance between the 1000 Louvain optimised partitions. The $\overline{VI}$ provides a relative distance within partition space between the Louvain partitions. However, partition space is highly dimensional, making visualisations of the ‘distances’ difficult. Here, we use multidimensional scaling (MDS) to visualise the VI (similarity) between individual Louvain partitions across the wildtype and mutant structures. MDS is an ordination technique used to display information contained in a distance matrix (VI matrix) where each partition is assigned coordinates in a chosen number of dimensions N . For example, choosing $N = 2$ would

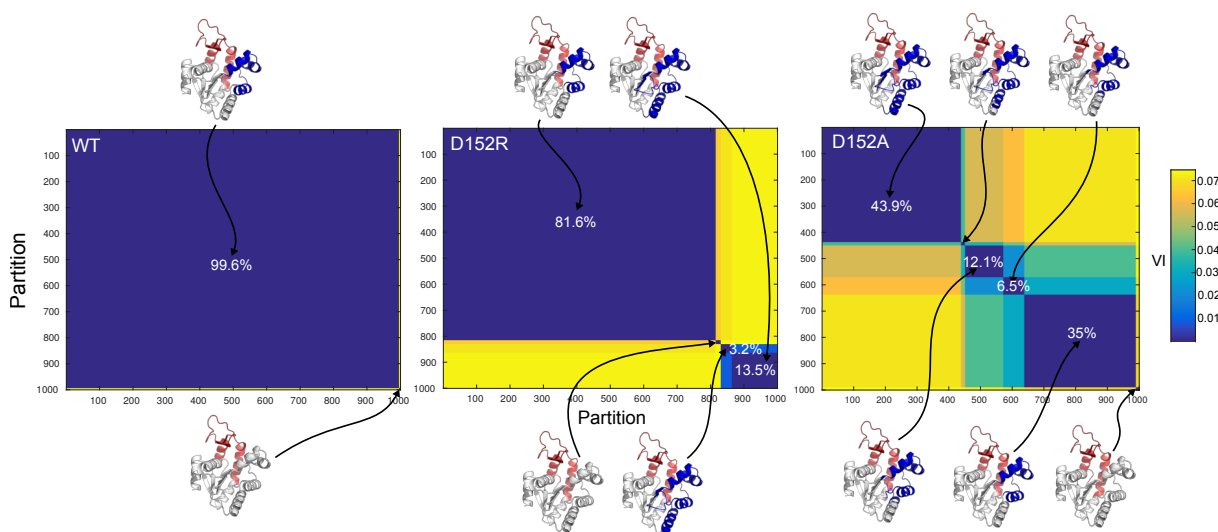


Figure 8.9: A Louvain optimisation partition matrix shows the varying partitions that appear upon mutation and reduce the robustness of the community structure. The wildtype finds very similar partitions across all the Louvain optimisations, suggesting that there is only one viable community structure. However, the mutants exhibit increasing differences between the community structures across the 1000 Louvain optimisations.

produce coordinates of partition location for a 2D scatter plot. Appendix G provides a thorough background to the MDS technique.

We produced a dissimilarity vector from every pairwise VI calculation of every Louvain optimised partition across the wildtype, D152A and D152R ADK structures (3000 total partitions). Using the ‘strain’ metric scaling criterion from the Matlab *mdscale.m* function we were able to recreate the data in 2-dimensions. Figure 8.10 exhibits the highly dimensional partition space transformed into a scatter plot. A lot of the partitions are exactly similar and lay on top of each other, therefore Figure 8.10 also exhibits a 2D bivariate histogram of the VI. In agreement with Figure 8.9, the wildtype clearly finds the same partition every time. Nearby to the wildtype is a slightly smaller population of D152R partitions, and a significantly smaller population of D152A partitions. It’s quite clear that D152A identifies a number of distinctly different partitions that are scattered around partition space as exhibited by the yellow histogram bars. Despite D152R sampling other partitions, the population of those other partitions is distinctly lower than D152A.

8.3. Arginine scan reveals reveals changes to long timescale dynamics

The previous section showed a relationship between the Markov stability mutant score and global dynamics and we also exhibited the ability to introduce different residues types and measure their effect on the community structure. However, we focussed on a single residue position

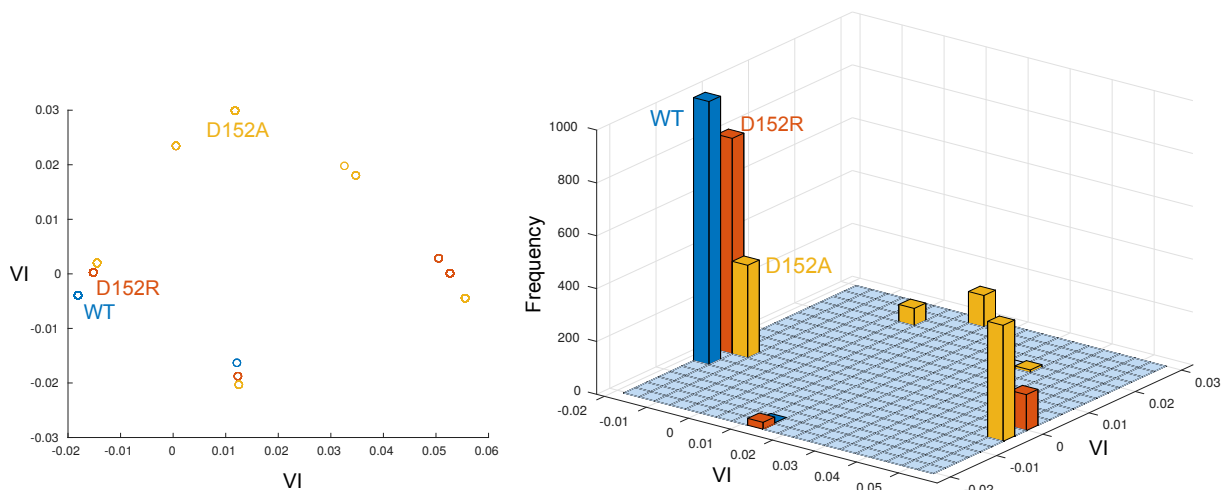


Figure 8.10: A multidimensional scaling (MDS) of ADK mutants in partition space. (a) Multidimensional scaling (MDS) is used to reduce the highly dimensional partition space into 2-dimensions such that we can calculate the ‘distance’ between optimal partitions. (b) A histogram of the MDS scatter plot highlights the changing populations of each partition for each mutant respectively. D152A exhibits a large shift away from the wildtype partition, explaining the underlying reason for the high mutant score.

(Asp152). In contrast to Chapter 7, in this section we explore the potential to perform complete mutagenesis scans of any residue type. In particular, we perform an arginine scan across ADK (mutating each residue to arginine separately), creating 203 mutant ADK structures.

8.3.1. Markov stability analysis of arginine mutant ensemble

To prevent sidechain clashes, energy minimisation of each mutant structures was essential. After which, a full Markov stability calculation was performed between Markov times 10^2 to 10^5 with 100 timesteps across each mutant structure. A workflow of the arginine scan methodology is presented in Figure 8.11A. In a similar manner to the previous section, we focus on mutations that perturb the community structure during the timescale associated with the 3-way partition so as to corroborate our findings with subdomain motions. A Gaussian process regression was used to calculate the mean and variance of the VI from the ensemble of mutations.

Figure 8.11 exhibits the scatter plot of $\overline{VI}$ versus CV for each mutant structure in the lid subdomain and the associated box plot to identify outliers. According to the interquartile ranges of the boxplot, we identify five outliers: V164R, V157R, D152R, K160R and P155R. In agreement with the previous section, D152R appears as a high scoring mutation, further highlighting the importance of the Asp152 residue. The highest scoring mutant, V164R, introduces an inter-subdomain salt bridge to E57 on the AMP_{bind} subdomain. P155R, V157R and K160R are located near the top of α -helix 8. V157R introduced strong hydrogen bonding with the backbone of Glu54, increasing the inter-subdomain communication. In a similar manner, mutation of K160 to arginine increase inter-subdomain communication through a salt bridge with Glu57 on the

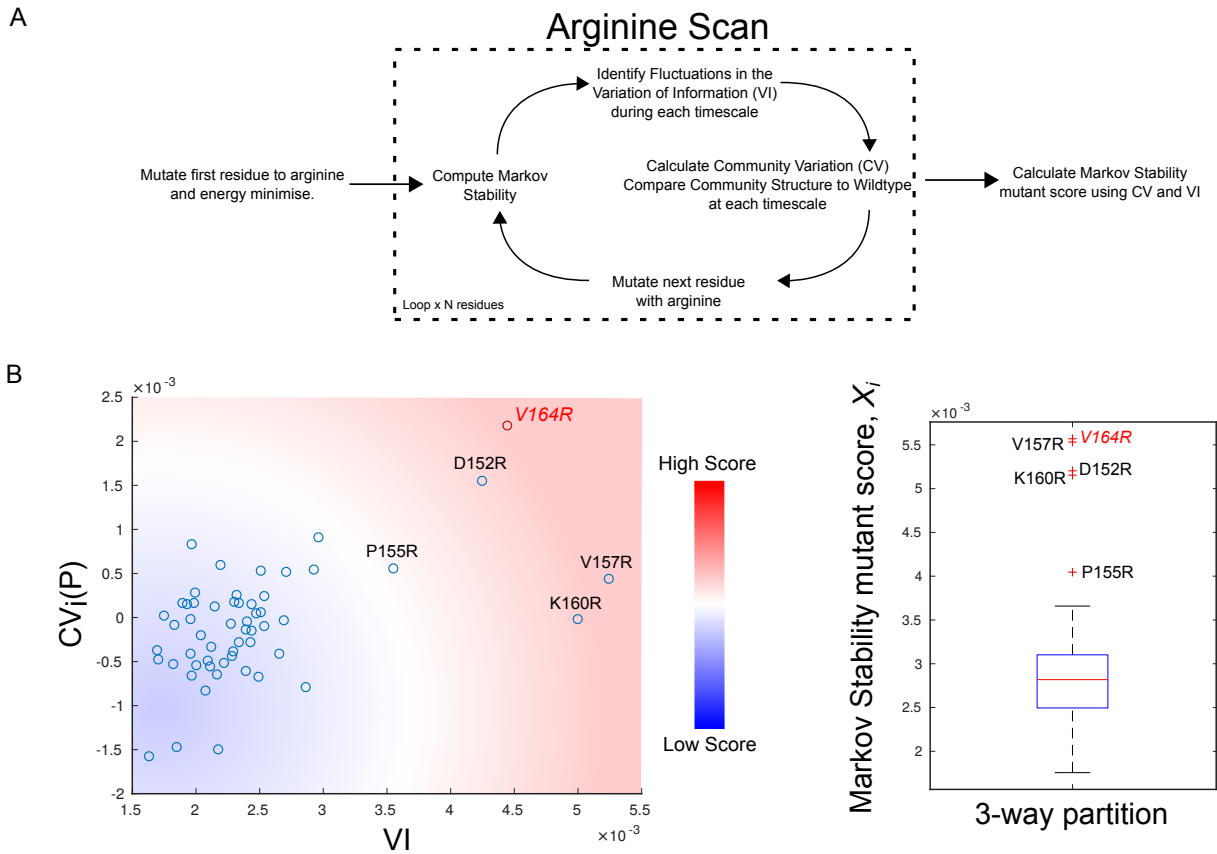


Figure 8.11: A Markov stability arginine scan to identify perturbative mutations. (a) A workflow of the arginine scan. (b) A scatter plot of VI versus CV to identify outlier arginine mutations. The boxplot of the mutant score calculated as a combined metric of VI and CV.

AMP_{bind} subdomain.

8.3.2. Single-molecule FRET of V164R

We have constructed and FRET labelled the V164R ADK mutant to investigate the Markov stability mutant scores attributed to the arginine scan. Single-molecule FRET measurements of the V164R mutant were performed and analysed in an identical manner to previous chapters. The FRET histogram in Figure 8.12b showed a significant shift towards a highly closed FRET state for V164R. The shift in conformational equilibrium is physically intuitive, the newly substituted arginine is forming a strong hydrogen bond with Glu57 on the AMP_{bind} domain and, therefore restricting the opening motion (Figure 8.12a).

In section 8.2 we found a correlation that told us that a higher Markov stability score during the 3-way partition would result in a shift towards a closed conformation. In accordance with this relationship, V164R exhibited a higher score ($X_{V164R} = 0.0056$) than D152R ($X_{D152R} = 0.0052$) and also exhibited a larger closed conformation population (V164R: $A_c = 80\%$, D152R: $A_c =$

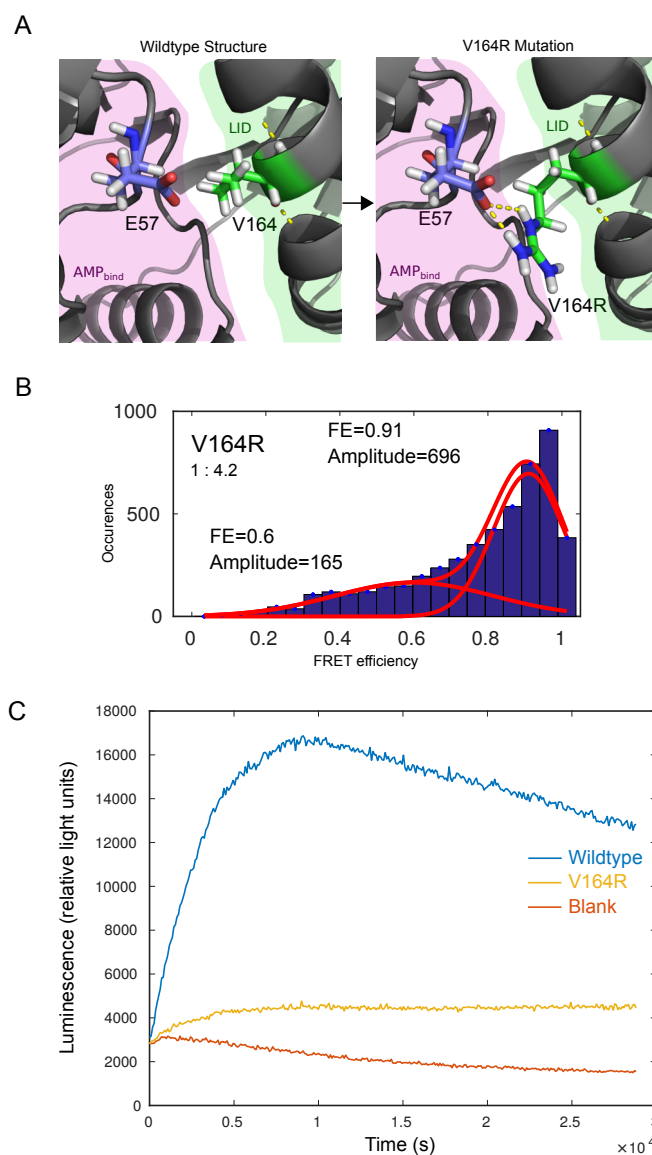


Figure 8.12: Experimental investigation of ADK V164R using smFRET and luciferase assays. (a) A side-by-side comparison of the sidechain packing of V164 before and after mutation to arginine. The hydrogen bonding of V164R to E57 has been highlighted. (b) The FRET efficiency histogram exhibits a large deviation towards the closed conformation as a result of the V164R mutation. (c) A luciferase activity assay suggests that, despite a large reduction, ADK V164R is still catalytically active relative to the wildtype.

68%). This showcases the robustness of the scoring methodology across the various residue types, providing further evidence for the relationship between Markov stability and global dynamics.

In the previous section we observed an increase in the Stability upon D152 mutations, however, V164R exhibits a decrease in the Stability of the 3-way partition. This is a result of the new salt bridge interactions with Glu57 on the AMP_{bind} subdomain, the random-walker is able to escape the lid community and explore the AMP_{bind}.

Despite the D152 mutations in section 8.2 exhibiting complete catalytic inhibition, V164R still exhibited catalytic activity albeit at a significantly decreased rate (Figure 8.12). Val164 is a fairly conserved residue, despite not binding substrate directly. Ådén et al. (2012) found that shifting the conformational equilibrium towards a closed conformation reduced the catalytic activity. We are observing a similar mechanism, the closed conformation induced by the V164R salt bridge is preventing the high catalytic turnover rate. In fact, another study by Kovermann et al. (2017) found that substrate could still bind to the active site whilst the protein was already in the closed conformation.

8.4. Discussion

Using the mutant scoring methodology in Chapter 7, we have exhibited a relationship between the graph-theoretic measure of Markov stability and the dynamical conformational equilibrium. We have used it to correctly identify mutations *without prior information* according to their perturbation to the conformational structure. Our mutant score reveals mutations that have *global* effects at long Markov timescales. We have shown that D152, an allosteric residue identified in Chapter 7, can be modified to modulate the conformational equilibrium of ADK. This finding suggests that the particular structural features embedded in the architecture of the protein are critical for functional motions. Our method further uncovers mutations that have not been previously identified as important, these high scoring mutations thus present potential targets for mutational analysis.

Our use of an atomistic graph has the advantage that we are able to calculate the total enthalpic contribution from single residues, allowing us to explore the relationship between the mutant score and the mechanistic perturbation that causes a change in the conformational equilibrium. Particular mutations significantly reduced the number of side-chain interactions, thus reducing the enthalpy associated with that position and also reducing the communication between regions of the protein structure.

In contrast to the alanine mutagenesis in Chapter 7, we have shown that we are able to effectively introduce new residues and perform mutagenesis scans across any residue type. Using energy minimisation we were able to effectively repack the side-chains of larger residues to allow the

introduction of large residues. The arginine scan of ADK highlighted residues at long timescales that introduced inter-subdomain interactions. We were able to produce a clamping mechanism that stabilised the closed conformation, whilst still allowing catalytic reactions to occur at a lower rate.

To what extent can we use the Louvain optimisations to understand the effect of mutations? At each Markov time the Louvain algorithm attempts to find the optimal partition through 100s-1000s of repeats. For the wildtype we found the same partition over 99% of the time, however, the high scoring mutants exhibited a large difference in the underlying Louvain partitions. The wildtype distinctly separated the subdomains into separate communities, whilst the mutants saw an increased mixing of the lid and AMP_{bind} with the core. These findings show that a stronger communication between the subdomains and the core corresponds with a higher shift towards a closed conformation.

The role of salt bridges in stabilising conformational structures and in catalytic activity has been widely studied (Anderson et al., 1990; Bosshard et al., 2004). Single mutations that remove a salt bridge can knock out kinase function (Dey et al., 2007). However, introduction of a salt bridge has so far received limited attention. Here, we have found that mutation of V164R introduces a salt bridge between two subdomains that clamps the protein closed, whilst still allowing a retarded catalytic rate. These results provide an excellent testbed for exploring the closed conformation of ADK, and suggest that the conformational equilibrium is correlated with catalytic activity (as long as the mutated residue was not a critical catalytic residue).

9. Thermostability analysis of ADK using Markov Stability and Circular Dichroism

In chapter 7, we found a correlation between the Markov Stability ‘score’ calculated for mutations at position D152 and the shift in the conformational population equilibrium as measured by smFRET. The physical reason for the shift to a closed conformation upon mutation is not fully understood. We discussed the potential for the lid subdomain to increase its propensity to ‘crack’ under these various mutations, indicating that we have perturbed the molecular stability of the lid subdomain. To explore this mechanism further we need to predict and validate changes in molecular stability.

In our graph representation we capture the physico-chemical details of the protein architecture. The bonds in a protein contribute to the free energy and a number of studies have found that the removal of a bond would result in a reduced molecular stability (Dunitz, 1995; Kristjansson and Kinsella, 1991; Pace et al., 2014b,a). However, even residues with similar enthalpic contributions will have differing entropic contributions based on their position (Dunitz, 1995). We would like to distinguish mutations that have a significant effect on molecular stability whilst inherently considering the spatial structure.

A chimeric analysis by Bae and Phillips (2006) showed that the core subdomain of ADK almost exclusively defines the thermal stability. Therefore, in this chapter we use Markov Stability to score mutations within the core domain according to their impact on the community structure. In a similar manner to the dynamics predictions, we hypothesise that high Markov Stability scores correspond to large changes in molecular stability.

To rigorously assess the predictive power of this method, we have expressed and purified 10 mutant proteins with ranging Markov Stability scores. By comparing the melting temperature with the mutant score, we find a good correlation between theory and experiment. Using the correlation as a training set we are able to identify other mutations that would cause a large change in molecular stability. Combining our results with multiple sequence alignment we are able to identify mutations inherent to ADK molecular stability and those that provide the thermophilic properties essential for *Aquifex* ADK.

9.1. Markov Stability and circular dichroism melting temperature study of ADK mutants

9.1.1. Computational mutagenesis identifies important residues in the core domain of ADK

In Chapter 8 we primarily looked at the 3-way partition to explore the global dynamics associated with the 3 subdomains. Here, we are interested in the local community structure of the core subdomain to explore molecular stability. We have continued to use the closed conformation (PDB ID: 2RGX) due to its low structural variability. Therefore, we have selected the transitional 6-way partition which is ‘fragile’ to perturbations such as mutation. Similar to Chapter 7, we have used alanine scanning combined with Markov Stability to identify interactions that perturb CV and $\overline{VI}$.

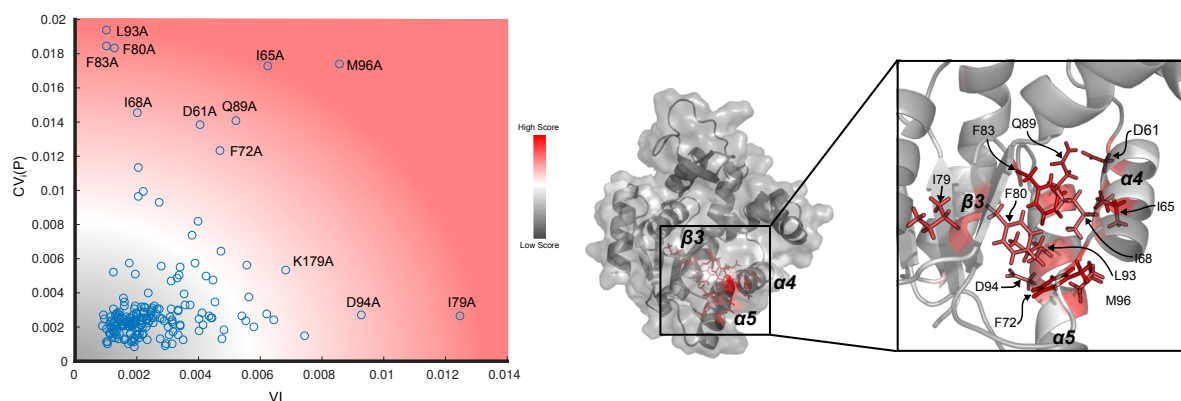


Figure 9.1: Mutations with the largest effect on 6-way partition of ADK. The normalised $\overline{VI}$ is plotted against $CV_i(P)$ to provide a measure of impact on the community structure. The higher scoring mutations are found in the red regions. The mutational score is mapped onto the protein structure.

Using the mutant structures and corresponding Markov Stability calculations from Chapter 6.1, we have calculated the $\overline{CV}$ and $\overline{VI}(t)$ during the timescale associated with the 6-way partition. Figure 9.1 presents the scatter plot of all mutations during this timescale. The scatter plot showcases the importance of using both CV and $\overline{VI}$, mutations such as D94 exhibit a large $\overline{VI}$ but low CV, whilst L93 exhibits a high CV and low $\overline{VI}$. In a similar manner to Chapters 7 and 8, the root mean square distance of each mutant defines its relevant score, this is illustrated by the grey-white-red blended background where grey indicates a low score and red indicates a high score.

The majority of high scoring mutations are located in a compact area of the core domain next to the AMP_{bind} domain (Figure 9.1). More exactly, they are located on $\beta 3$, $\alpha 4$ and $\alpha 5$. The most significant mutation, F72A, removed 36 hydrophobic interactions with surrounding residues. F72 interacts with M96, F80, I68, L93 which were all hydrophobic residues that ranked highly.

Table 9.1: Core mutations chosen for molecular stability experiments and their assigned Markov Stability score. The Markov Stability score was calculated during the 6-way partition.

Mutant	Markov Stability Score
WT	-
E95A	0.0027
K178A	0.0031
K179A	0.0087
D94A	0.0097
I79A	0.0127
Q89A	0.0150
I65A	0.0184
F83A	0.0185
Q89A + I65A	0.0150
F72A	0.0132

This indicates the presence of a hydrophobic core where F72 plays a key stabilising role. The presence of hydrophobic centres have emerged as essential components for stabilising folded proteins, however, hydrogen bonds also make a large contribution to protein stability (Pace et al., 2014b). We find that the community structure is also significantly affected by D94A, a negatively charged near-surface residue. D94 doesn't make many hydrophobic interactions but it does make significantly strong hydrogen bonds with K179 (676.1 kcal/mol) and Y175 (96.5 kcal/mol). The mechanisms of stabilisation between hydrophobic and polar residues are distinctly different, highlighting the importance of using an unbiased physico-chemical approach.

9.1.2. Construction of ADK core mutants according to prediction scores

A total of 10 ADK core-mutant structures were constructed using Quikchange mutagenesis and expressed as described in Section 5.5 (the protein was eluted into 20mM phosphate buffer since the Chloride of a salt buffer absorbs UV light used in CD experiments). The residue positions and their associated Markov Stability scores are shown in Table 9.1. Two mutations, E95A and K178A, were chosen for their low score to act as experimental controls. Additionally, their spatial closeness to D94A and K179A will help highlight the importance of using an atomistic description.

Due to the hyperthermophilic nature of Aquifex ADK ($T_m = 109^\circ\text{C}$), each protein was suspended in 5 M guanadine-HCL to destabilise the protein such that the wildtype melted at ($\sim 67^\circ\text{C}$). This allowed measurement of the full melting curve - from fully folded to fully unfolded. The 5 M gaunadine-HCL concentration was determined after progressive increases of guanadine-HCL to the wildtype (from 2 M in increments of 0.5 M up to 5 M).

9.1.3. Melting temperature measurements using circular dichroism

In this section we examined the wildtype and mutant ADK structures by circular dichroism (CD) spectroscopy to assess the effect each mutant made on the molecular stability. Circular Dichroism combined with temperature ramping were used to measure the melting temperature T_m of each ADK mutant. Proteins were measured in high transparency 1 mm path-length quartz cuvettes at a concentration of 0.1 mg/ml as described in Section 5.4.2.

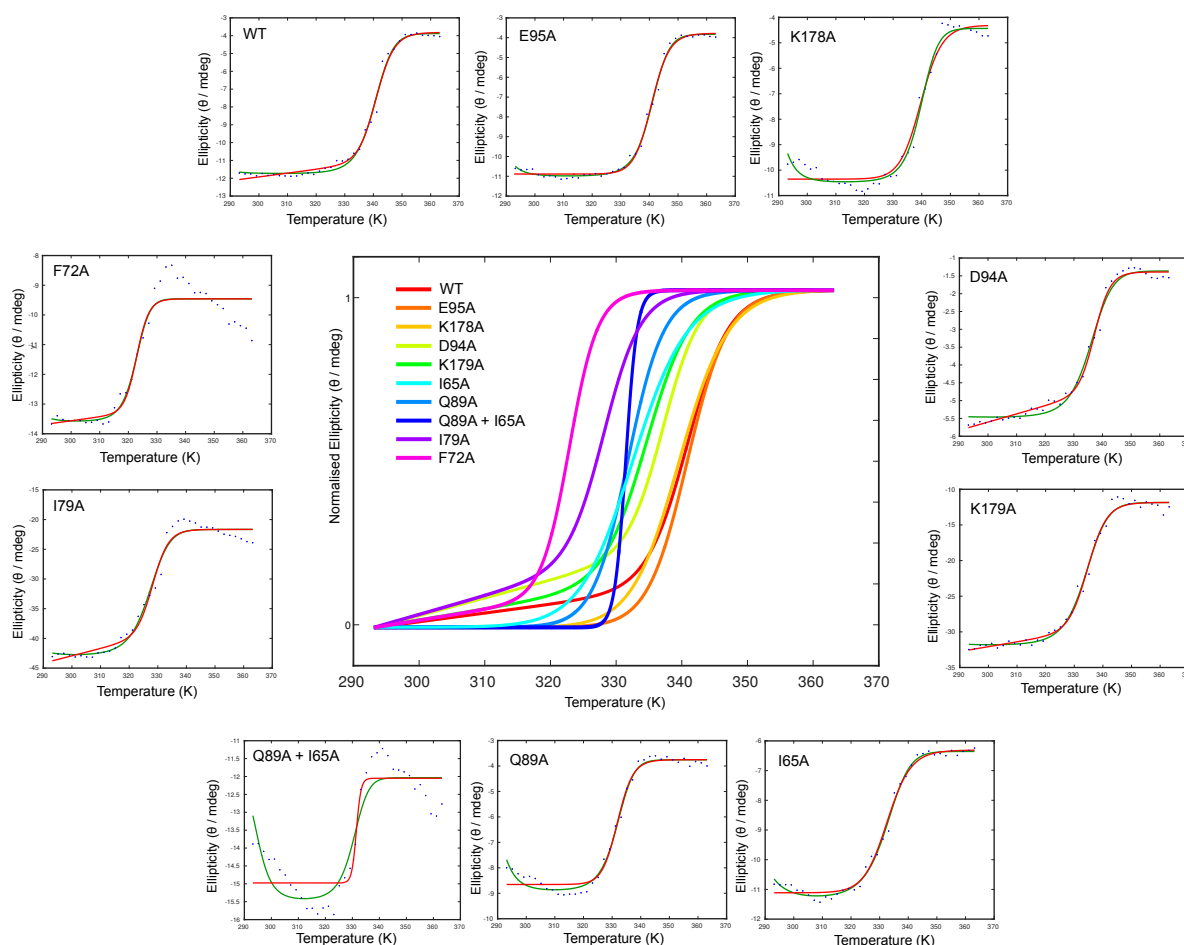


Figure 9.2: Melting curves of ADK mutants measured by circular dichroism. The blue dots represent the average ellipticity measured at 2°C intervals. The sub-graphs present the melting data and fits with model I (green) and II (red) (for fitting see Section 5.4.2) for each protein respectively. The central graph presents the normalised model I fit for each mutant, where the colour is representative of the melting temperature (red high and purple low).

CD melting trajectories of 10 ADK constructs (including the wildtype) are shown in Figure 9.2. All the mutants shown had a T_m between ~ 50 -70 °C. The majority of ADK constructs exhibited a near constant linear ellipticity in the fully folded and fully unfolded states. However, some large pre- and post-transitional irregularities in ellipticity were observed in F72A, I79A and Q89A+I65A. This is unlikely to be the buffer since all proteins were systematically diluted

Table 9.2: The melting temperature of the wildtype and mutant ADK structures are determined from the circular dichroism melting curves. Two thermodynamic models are used, the first (I) assumes a change in heat capacity between the folded and unfolded states, whilst the second model (II) introduces corrections for pre- and post- transition linear changes to ellipticity.

Mutant	$T_m(\text{I})$ / °C	$T_m(\text{II})$ / °C
WT	67.9 (67.2 - 68.7)	67.2 (66.6 - 67.8)
E95A	67.6 (67.1 - 68.1)	67.6 (67.2 - 68.0)
K178A	66.5 (63.6 - 69.3)	66.6 (65.8 - 67.3)
D94A	64.3 (63.5 - 65.0)	62.7 (61.8 - 63.7)
K179A	61.8 (60.8 - 62.7)	60.8 (59.9 - 61.7)
I65A	59.5 (58.4 - 60.6)	59.1 (58.4 - 59.8)
Q89A	58.7 (57.9 - 59.5)	58.5 (57.9 - 59.1)
Q89A + I65A	58.3 (56.9 - 59.8)	57.1 (54.3 - 59.9)
I79A	55.3 (53.9 - 56.7)	53.8 (52.4 - 55.1)
F72A	49.9 (47.6 - 52.1)	49.4 (47.4 - 51.4)

in the same concentration buffer and, therefore, it is likely a consequence of a highly destabilised protein that has already partially unfolded. Indeed, F83A (not shown in Figure 9.2) was already fully unfolded at room temperature and no T_m could be measured.

The results obtained from analysis of the melting trajectories are summarised in Table 9.2. The values for T_m given in the second column are considered reliable relative to their 95% confidence intervals shown in brackets. Additionally agreement between duplicate scans showed reliable results. The wildtype T_m was 67.9 °C. The two control mutations, E95A and K178A, both exhibited a T_m within the 95% confidence bounds of the wildtype indicating that these residues had a negligible contribution to molecular stability. Whilst polar residues are deemed important to coordinate hydration forces to stabilise protein structures (Levy and Onuchic, 2004), here we show that E95 and K178 are not essential.

The remaining 7 mutant structures have lower T_m values, outside of the wildtype 95% confidence bounds. The largest change was observed for F72A which melted at 49.4 °C and was assigned the highest Markov Stability ‘score’. The high score for F72A is a reflection of the large number of hydrophobic interactions removed upon mutation to alanine (discussed in Section 9.1). It is clear that these hydrophobic interactions played an essential role in the molecular stability of ADK. Moreover, F72 is a partially conserved amino acid, where it is often substituted for other hydrophobic residues, indicating that it is likely an important hydrophobic region for the stabilisation of the core.

The I65A and Q89A mutants both scored highly and exhibited T_m of 59.5 °C and 58.7 °C respectively (a drop of 8.4 °C and 9.2 °C). Interestingly, the double mutant (I65A+Q89A), which scored that same as Q89A, had a T_m of 58.3 °C. The combined mutation of I65A and Q89A clearly does not give an additive effect on the T_m , despite their separate significant perturbations. This is partially a consequence of the shared hydrophobic interactions between I65A and Q89A,

due to their close proximity in the tertiary structure, although these interactions are minor in comparison to the wide range of additional stronger interactions that were also broken. This highlights the importance of using a method that considers the spatial 3D structure.

Another notable mutation was D94A which saw a modest drop of 3.6 °C in T_m and neighbours the control mutation E95 which exhibited no change in T_m . The difference in T_m between D94A and E95A is reflected in their different computational predictions; D94A was assigned a mid-to-high score by Markov Stability whereas E95A scored very low. D94 extends hydrogen bonds with K179 (~700 kcal/mol) and Y175 (~100 kcal/mol) on α -helix 8 whilst E95 only interacts with residues inside the same secondary structure. This emphasises the importance of implementing an atomistic description of a protein structure and highlights the ability of Markov Stability to discern between similar residues that are spatially co-located.

Similarly, K179A saw a mid-high score and a drop of 6.1 °C whilst its neighbour K178A showed no change in T_m . Again, K178 only interacts with residues inside the same α -helix 8, whilst K179 extend inter-secondary interactions. The strong hydrogen bond between K179 and D94 was mentioned above, however, K179 extends a larger number of hydrophobic interactions relative to D94. This may explain the larger change in T_m observed for K179 relative to D94.

9.1.4. A detailed desconstruction of the mutant community structures

In the previous section we observed large changes in the melting temperature for mutations that scored highly, and we saw little change to the control mutations. Whilst we explored the underlying interactions of each mutated residue, we need to understand the changes to the community structure that each mutation introduced.

Figure 9.3a exhibits the VI trajectory for each mutant that we physically measured overlaid upon the 2 standard deviation region calculated from a Gaussian process regression of the ensemble of mutations. In justification of the low Markov Stability score they both attained, the control mutations E95A and K178A both show very little deviation from the ensemble mean. However, significant changes in the robustness of the partitions is observed for the remaining mutations which all fluctuate outside of two standard deviations multiple times.

To explore the evolution of the community structure, Figure 9.3b exhibits a colour map of the partitions over Markov time. The control mutations both show very similar community evolution maps to the wildtype, however, both of their neighboring residues D94A and K179A show large changes in the evolution of their community structure. Both K179A and D94A drop into the 3-way partition at an earlier Markov time. The 3-way partition forms a single community with a large section of the core subdomain, suggesting a large change to the structural make-up of the core. Interestingly, both K179A and D94A cause a large disruption to the 4-way partition,

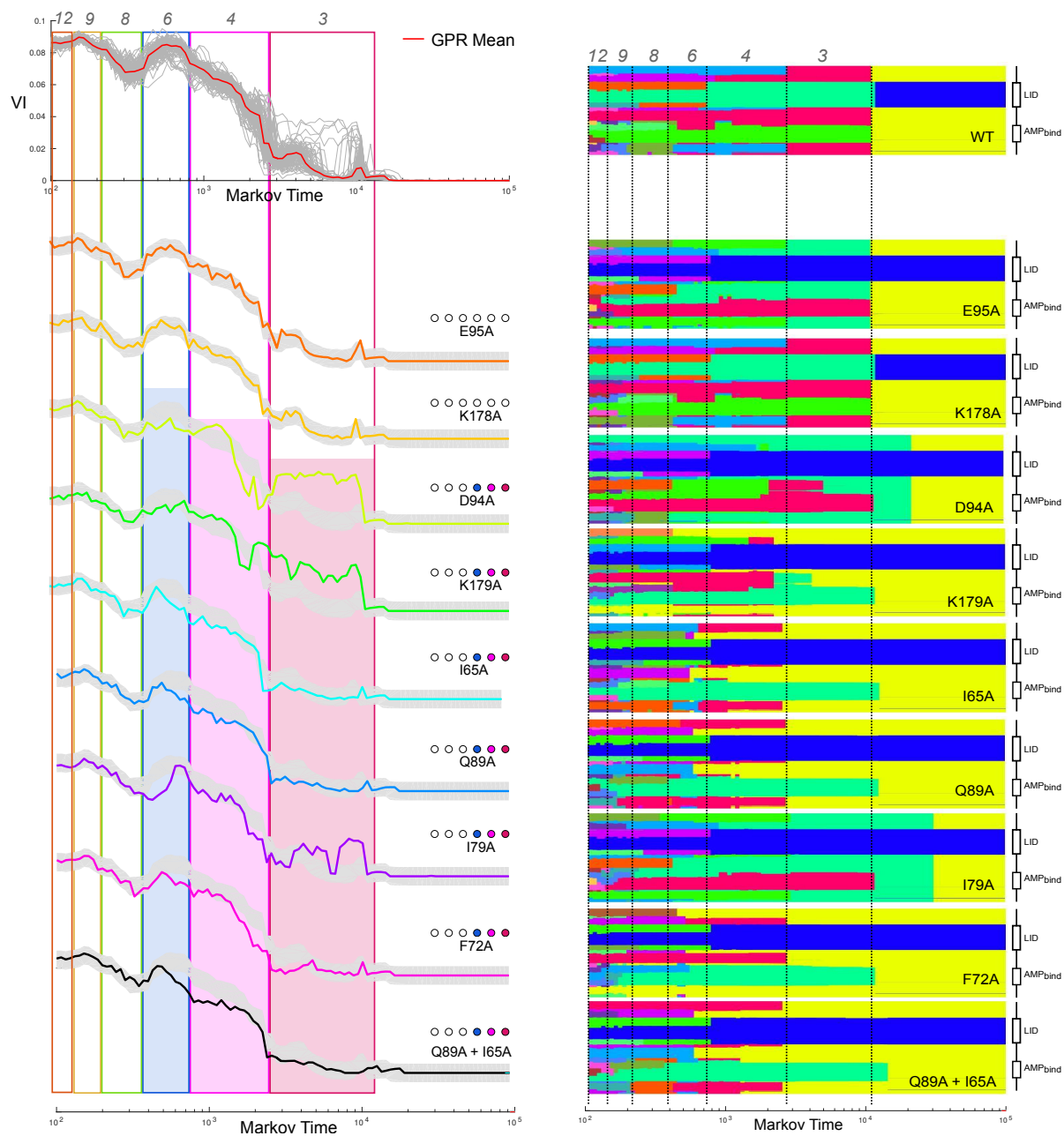


Figure 9.3: An analysis of the Markov Stability community structure of each core subdomain mutant. (a) The VI trajectories for the mutants that were experimentally investigated. The grey area is 2 standard deviations around the mean ensemble calculated by GPR. (b) The community structure calculated by Markov Stability for each mutant across Markov time. The black vertical lines denote the changes between the wildtype partitions.

where the mutant structures prefer to partition into 5 communities: the lower half of α -helix 8 has separated to form its own community. The similarity of perturbations to the community structure is a consequence of the hydrogen bond between D94 and K179. Since K179A is located on α -helix 8, the hydrogen bond between the two residues was essential for grouping the lower half of α -helix 8 with the core subdomain.

If we look closely at I65A during the 6-way partition we can see a large amount of flip-flop between partitions. In particular, residues located on β -helix 3 and α -helix 5 continuously swap between neighbouring communities. Q89A shows similar changes in the evolution of the community structure, where β -helix 3 begins to break away from the core subdomain. The double mutant (I65A+Q89A) actually exhibits a very similar community evolution to Q89A, suggesting that the structural changes introduced by Q89A actually dampen or exceed the perturbations introduced by I65A. This helps clarify the similar Markov Stability score that was assigned to the double mutant relative to its independent mutations.

Mutation of I79A showed slight changes to the core subdomain, α -helix 4 begins to flip-flop between the core and AMP_{bind} communities. Similarly, F72A showed large changes to α -helix 4 which temporarily formed its own independent community. Suggesting that mutation to I79 or F72 increases the flexibility of α -helix 4 relative to the core. Curiously, I79A is located at a distal site to α -helix 4 and makes no direct interactions, highlighting the allosteric nature of mutations. On the other hand, F72A is located at the end of α -helix 4 in an essential hinge region identified in Chapter 6 and by [Henzler-Wildman and Kern \(2007\)](#). It is likely that perturbation to this hinge is separating α -helix 4 from the core subdomain.

9.1.5. Combining Markov Stability predictions and melting temperatures

We suggested that the relationship between Markov Stability scores and the conformational equilibrium in Chapter 8 was caused by changes in molecular stability. Here we look for a direct relationship between the computational predictions and melting temperature. Figure 9.4 presents the plot of Markov Stability ‘score’ against T_m . The linear fit is to represent the observed correlation and not chosen to present any physical meaning. Despite an R-square of 0.58, the regression shows a clear correlation between an increased Markov Stability score and a reduction in T_m .

The results of this correlation imply that the melting temperature is decreasing with an increasing Markov Stability score. The high Markov Stability scores were a result of a less robust community structure ($\overline{VI}$) or a significant change to the community structure (CV). The changes to $\overline{VI}$ and CV were a result of removing key interactions, identified by Markov Stability as spatially crucial at the chosen timescale. In fact, the clustering of high scoring mutations suggested that there is a ‘hot spot’ of residues that are essential for stabilising ADK.

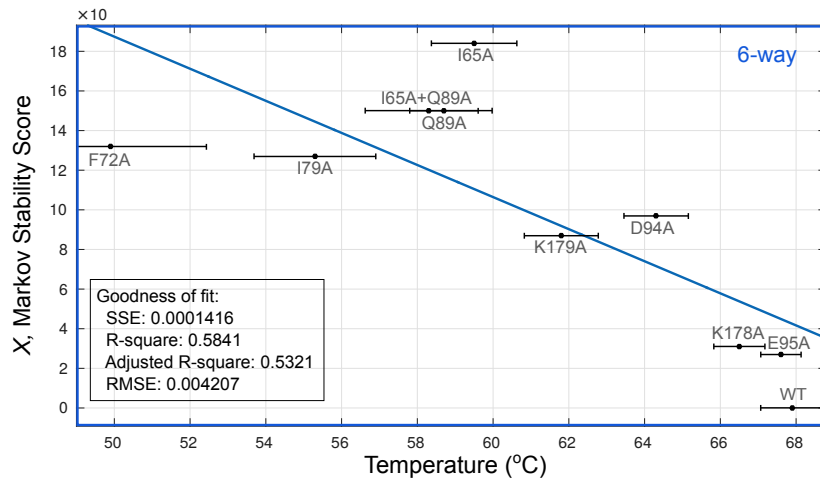


Figure 9.4: A plot of computational scores against melting temperatures for the ADK mutants. The scores The Markov Stability scores of each ADK core mutation were plotted against their respective melting temperature as measured by circular dichroism. The error bars are the 95% confidence intervals of the sigmoidal fit to the CD curves. The blue line represents a linear fit with an R-square of 0.58, suggesting the presence of a correlation between the theoretical and experimental parameters.

Using an approach to ‘score’ mutations according to their interaction energies would have led one to believe that removing two residues would result in the direct addition of their respective bond energies. However, a Markov Stability analysis of the I65A+Q89A mutant attributed a score only slightly higher than the separate mutants. Indeed, the double mutant exhibited a decrease in melting temperature that wasn’t a combination of the difference in melting temperature of the separate mutants $\Delta T_m(12) \neq \Delta T_m(1) + \Delta T_m(2)$. Appropriately, this also agreed with the correlation found between the Markov Stability scores and the melting temperatures. This correlation allows us to benchmark and predict the position of potential future mutations to the core subdomain of ADK.

9.1.6. Robustness of scoring across timescales

In the previous analysis we selected the 6-way partition to explore mutations during the a transitional partition where the community structure inside the core domain was undergoing change. This was chosen because literature studies have shown the core domain to be the main region of molecular stability in ADK. Additionally, a transitional partition is significantly susceptible to mutation allowing us to observe large effects on the community structure. However, to remove doubts associated with ‘cherry picking’ we must additionally explore the neighbouring timescales and their relationship with the melting temperature studies.

Figure 9.5a exhibits a CV versus VI plot for the mutants during the 4-way partition. We see that F72A is attributed a much higher score than during the 6-way partition. The Markov Stability mutant scores X_i associated with the scatter plot are calculated and plotted against the melting

temperature in Figure 9.5b. We observe a correlation of similar strength (R -Squared = 0.56) to the scores from the 4-way partition. The control mutations show a low score regardless of which partition we choose, whilst the scores of the remaining mutants is still higher but have changed. In particular, the double mutant now scored more highly (Table 9.3) than its constituent separate mutants ($X_{Q89A+I65A} = 0.0125$, $X_{Q89A} = 0.0099$, $X_{I65A} = 0.0086$), agreeing with the slightly increased melting temperature measured for the double mutant.

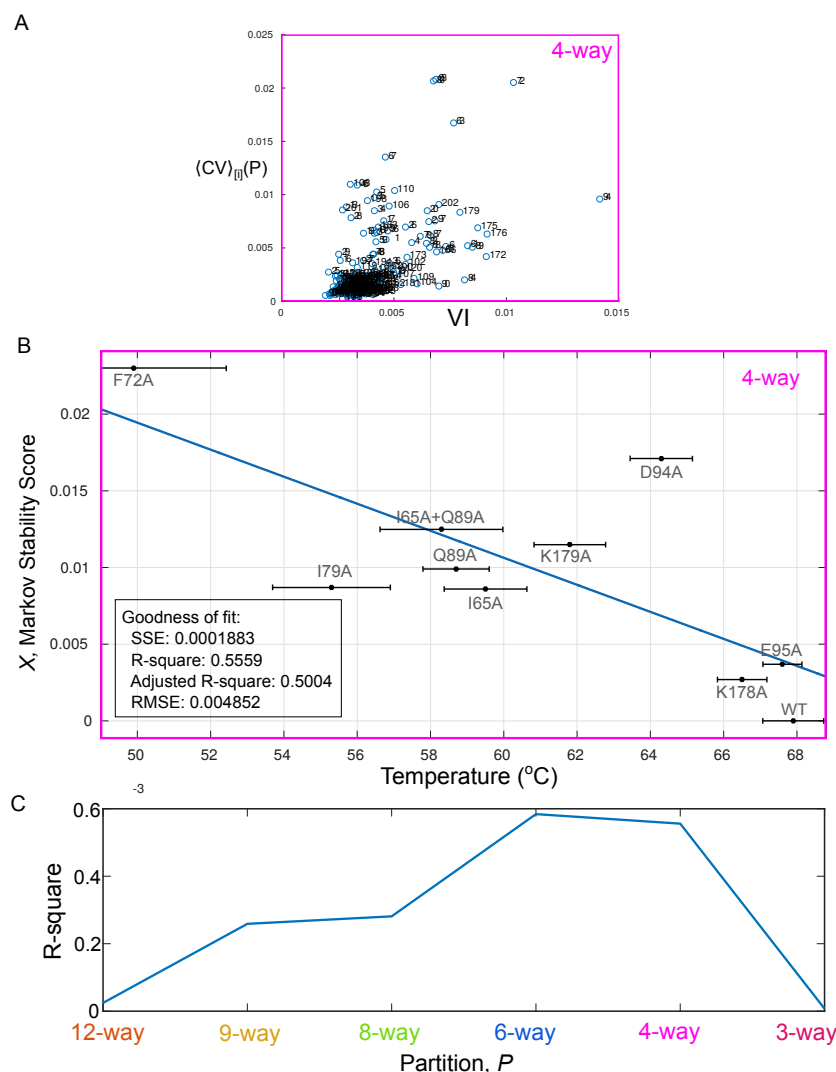


Figure 9.5: Comparing the melting temperature of each mutant with the scores across all Markov timescales. (a) A scatter graph of CV versus VI during the 4-way partition to identify mutant outliers. (b) A plot of mutant score during the 4-way partition versus melting temperature. A linear regression shows an $R^2 = 0.55$. (c) A plot showing the R-square values of correlations for the mutant score versus melting temperature for each relevant timescale. The 4-way and 6-way partition mutant scores show the strongest correlation with experimental data.

To explore the correlation across all Markov timescales, we plotted the mutant score against melting temperature for the 3-way/4-way/6-way/8-way/9-way/12-way partitions and calculated

Table 9.3: The Markov Stability mutant score for chosen mutations during the 4-way partition. The double mutant exhibits a higher score than its separate constituent mutations.

Mutant	Markov Stability Score
WT	-
E95A	0.0037
K178A	0.0027
K179A	0.0115
D94A	0.0171
I79A	0.0087
Q89A	0.0099
I65A	0.0086
F83A	0.0084
Q89A + I65A	0.0125
F72A	0.0230

the R-squared value for a linear regression to each (Figure 9.5c). It is clear that the 4-way and 6-way partitions offer the highest correlation, whilst the correlation significantly decreases at shorter and longer timescales. Despite changes to the number of communities, the 4-way and 6-way partitions don't exhibit changes to the lid and AMP_{bind} subdomains community structure. Instead, we observe large changes to the community structure within the core subdomain. This suggests that the melting temperature is dictated by interactions between large communities within the core subdomain. This would agree with the experimental observations by Bae and Phillips (2006) that used chimeric analysis of mesophilic and thermophilic ADK subdomains to show that the core subdomain was responsible for the molecular stability whilst the lid subdomains were responsible for the catalytic rate.

9.2. Revealing residues essential for thermophilic properties

Using Markov stability as a mutant scoring method we can start to uncover residues that are essential for molecular stability. However, its difficult to differentiate between residues that are essential for inherent mesophilic molecular stability and residues that have been evolutionarily introduced to provide the thermophilic properties of *Aquifex* ADK. To explore this further we need to combine the conservation scores from a multiple sequence alignment with the Markov stability scores.

Table 9.4 presents the melting temperatures versus conservation score for the mutants measured in Section 9.1.3 (the double mutant was removed since we are focussed on single residue conservation scores). The conservation scores were calculated using the ConSurf server database (Ashkenazy et al., 2010)¹ with the input of PDB ID: 2RGX and are available in Appendix H.

¹http://bental.tau.ac.il/new_ConSurfDB/

Table 9.4: The conservation score (Consurf (Ashkenazy et al., 2010)) for each mutant and their respective melting temperature.

Residue / Mutant	T_m / °C	Conservation Score
E95	67.6	2.615
K178	66.5	1.969
D94	64.3	-0.261
K179	61.8	0.694
I65	59.5	-0.385
Q89	58.7	-1.082
I79	55.3	-0.616
F72	49.9	-0.243

A more negative conservation score indicates a high evolutionary conservation, whilst a positive conservation score suggests no evolutionary conservation. It is important to note that the evolutionary conservation of a residue at a particular position isn't necessarily for molecular stability. Indeed, since nature tends to reuse the active site in enzymes of related species, the majority of conserved residues are usually located around the active site for substrate binding. According to the crystal structure only Q89 binds directly to Ap5A, whilst E95 coordinates a Zinc ion (likely a consequence of crystallisation rather than functional).

The two control mutations, E95A and K178A, both showed a low conservation, low Markov stability score and a negligible change in melting temperature relative to the wildtype. Both residues are solvent facing charged residues capable of interacting with water molecules. Hydration forces play a critical role in protein stability (Sorenson et al., 1999), however, we don't implicitly include water molecules inside the protein graph. Instead our atomistic graph representation indirectly considers hydration forces through the introduction of hydrophobic interactions. These two mutations highlight the robustness of our graph construction methodology, the indirect inclusion of hydration forces is sufficient to model the protein system whilst also simplifying it for computational efficiency. E95 hydrogen bonds with K99, however, its low conservation indicates that the interaction, whilst strong, isn't necessary for molecular stability.

The substrate binding residue, Q89, showed the highest conservation, supporting its critical role in enzymatic catalysis. However, the drop in melting temperature $\Delta T_m = 9.2^\circ\text{C}$ suggests that Q89 also played a critical role in the molecular stability of ADK. This highlights the multiple roles that single residues play in protein systems. In fact, Kerns et al. (2015) previously showed the dual importance of R150 in the lid subdomain of ADK.

The remaining residues, except K179, all showed relatively high conservation which correlated with the high Markov stability score and their associated drops in T_m . In section 9.1 we mentioned the presence of a molecular stability 'hot spot' in the core subdomain. The high conservation of these residues provides further evidence for their importance to molecular stability. However, since they are conserved across the majority of ADK species we can assume that they

aren't providing the thermophilic properties that have optimised Aquifex ADK. The large variability of K179, which is a partially outward-facing residue located at the bottom of α helix 8, is of particular interest. Beyond interactions with residues within its own secondary structure, K179 extends a strong hydrogen bond to D94 on α helix 5. This analysis suggests that the introduction of K179 provides a higher molecular stability allowing Aquifex ADK to adapt to its thermophilic conditions.

Using the correlation in Section 9.1.5 we can identify mutations that have a high Markov stability score (highly destabilising) but a low variability (non-conserved across ADK species), providing us with a pool of residues that are potential thermophilic adaptations. Figure 9.6 shows a plot of Markov Score against conservation score for each residue in Aquifex ADK. Inhibitor (Ap5A) binding residues from the crystal structure have been coloured mauve. The general trend exhibits a clustering of high conservation and low scoring residues in the bottom right, suggesting that residues are conserved for more than just molecular stability. No residues exhibited a significantly high Markov stability score and also a low conservation (top right of the graph), this indicates that all residues essential for molecular stability tend to be fully or partially conserved. Highlighted by red lines towards the middle of the plot, we find residues that are relatively variable between amino acid types but exhibit a high Markov stability score. Using the correlation identified in Figure 9.4 we can posit that these residues are destabilising and, due to their variability across ADK species, play a role in the thermophilic adaption of Aquifex ADK.

The lysine mutation K179A, which falls into the 'thermophilic adaption zone' on Figure 9.6, was identified and measured to show a large change, $\Delta T = 6.1^\circ\text{C}$, to the melting temperature. The key salt bridge interaction between K179 and D94 is not present in E.Coli ADK, suggesting that it was introduced to provide greater molecular stability. Whilst D94 is present in E.Coli ADK (PDB ID: 1AKE, D93), it doesn't make any alternative salt bridge interactions.

Aquifex ADK (thermophilic) has 7 unique prolines, whilst E.Coli ADK (mesophilic) has only 2. [Henzler-Wildman et al. \(2007a\)](#) simulated four proline mutations (P44E, P60T, P73A, P155A) to alanine and found that they were the physical origin for the observed differences between meso and thermo ADK. Our analysis here identifies P73A has a potential thermophilic residue, attaining both a high Markov stability score and showing low conservation across the ADK species. Both P155A and P44A sit on the edge of the region of thermophilic adaption, their relatively low score may be a consequence of our network based method not modelling entropic contributions effectively. However, whilst proline substitutions introduce local rigidity to the structure, [Daily et al. \(2011\)](#) found that hinge flexibility and global functional motions are correlated but not exclusively determined by hinge residues. This suggests that the mutations we highlight as potential thermophilic stabilising residues may still be crucial for thermophilic adaption.

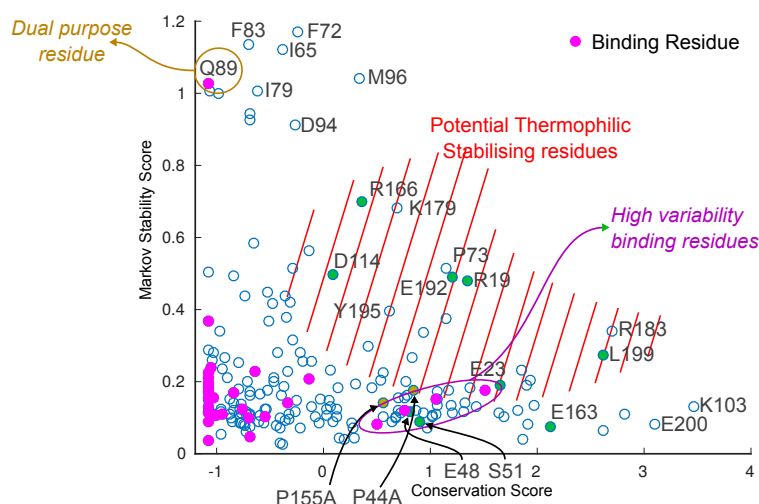


Figure 9.6: Using the mutant score and conservation scores to identify thermophilic adaption residues. A scatter plot of Markov stability mutant score versus conservation score for each alanine mutant in ADK. The red zone are mutations that have low conservation but also a high mutant score, these are residues that were potentially introduced to increase thermostability. Mutants highlighted in pink are binding residues and the majority have a low mutant score, suggesting they aren't important for the structural integrity of ADK. Residues highlighted in green are those identified by [Nguyen et al. \(2017\)](#) as thermophilic adaption residues in ADK, of these 6/9 fall into the red prediction zone.

A more recent study by [Nguyen et al. \(2017\)](#) identified, by ancestral sequence reconstruction, a number of residues that were responsible for the thermostability of ADK enzymes. In particular they identify R19, E23 (D23), E192 (Q202), L199 (K209), D114 (K117), R166 (M176), K48 (E48), E51 (S51) where the residues in brackets are the equivalent positioned residues. Of these residues, only K48 and E51 don't appear inside the thermophilic zone in Figure 9.6, suggesting that our prediction of thermophilic adaption residues is accurate. The reason that K48 and E51 aren't also predicted by our method is that they don't produce a salt bridge in Aquifex ADK. Combining this with the prediction of proline mutations within the thermophilic adaption zone (i.e, P73), we can hypothesise that we are able to predict residues that are not only crucial for molecular stability but residues have been introduced to adapt the catalytic rate according to the environmental temperature.

9.3. Fuzzy community structures identify molecular stability hotspots

We saw in the previous section that mutations with a large change in melting temperature also exhibited large changes to their community structure within the core subdomain. To explore these structural changes further we need to look at the 'cross-talk' between communities, how well defined are the communities within the core subdomain? To do this we can look at the *fuzziness* of the communities.

Markov stability optimises the communities within a graph using a transient random walk. These communities are defined on a discrete basis; a node is either in or out of a particular community and a single node can only be found in a single community. The ‘fuzziness’ is when a node has a high probability of joining a neighbouring community and can help us identify rigid and flexible regions. A rigorous approach to identify fuzzy communities would be to optimise a probability partition in a similar manner to how Markov stability optimises a discrete partition. However, this isn’t easy and would be computationally inefficient. Here, we propose a ‘cheat’ to be able to observe the fuzzy communities simply using information already derived from Markov stability. The communities are pre-defined using Markov stability to find the optimised partition. At a particular timescale if we sum the probabilities of the random-walker leaving node i and ending on any node in community j we can define the probability of a node being part of that community. Mathematically this requires multiplying the matrix exponential by the community node matrix H ,

$$Z = e^{-tL}H, \quad (9.1)$$

where Z is an $N \times c$ matrix that gives the probability of a node being in each community and each row should therefore sum to 1. We can then look at the entropy of distribution of this community probability matrix by calculating the entropy of each node across the communities,

$$S_i = -p_i \log(p_i), \quad (9.2)$$

at every time point. This can be used as a measure of the robustness of a community structure, where a lower entropy would suggest a more robust community. We can also use it to identify regions that undergo large changes in the structure of a particular Markov timescale.

Figure 9.7 exhibits the entropy of each node mapped onto ADK for each relevant timescale. It is clear that during the 12-way through to 4-way partitions there is a central red region inside the core subdomain that reflects the high probability of these atoms joining a nearby community instead of its assigned community. This suggests that the community structure within the core is not strong, any small changes, i.e. mutations, can cause large changes to the community structure within the core. The 4-way and 6-way partitions seem to show the highest node entropy within the core subdomain, the region that is responsible for the molecular stability (Bae and Phillips, 2006). This may explain why we observe a correlation between the mutant score (the effect of a mutation on the community structure) and the melting temperature during these timescales.

In contrast to the other timescales, the 3-way partition exhibited a higher node promiscuity (i.e. a node had a high probability of being in multiple communities) between the lid and AMP_{bind}

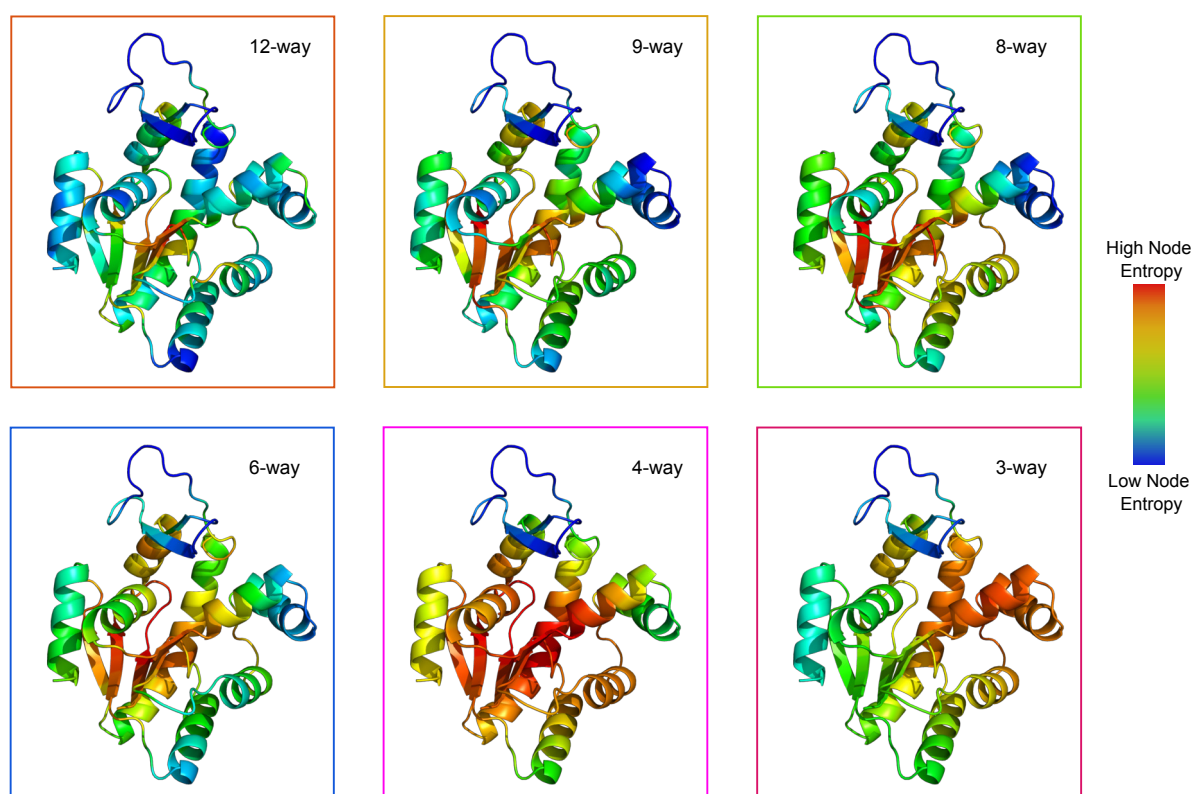


Figure 9.7: Using fuzzy communities to identify residues/regions that have no stable community structure. The entropy of distribution (the inverse probability of staying within its optimal community) for each node is mapped onto the ADK structure from high (red) to low (blue) for each timescale of interest.

communities, whilst atoms in the core subdomain were suitably stable within their community. Since the previous chapter found a correlation between the smFRET dynamics and the mutant score, we can hypothesise that the mutant score will identify mutations that perturb dynamics when there is a large node entropy between the dynamical subdomains. Vice-versa, we will identify mutations that perturb melting temperature when there is a large node entropy within the core subdomain.

9.4. Comparison of ADK temperature homologues

In Section 9.2 we used the relationship between Markov stability scores and melting temperature to identify residues that had been evolutionarily introduced into Aquifex ADK to adapt to higher temperatures. To explore the evolutionary adaption of ADK we performed a Markov stability analysis of three temperature divergent ADK homologues: a psychrophile (PDB ID: 1S3G), mesophile (PDB ID: 1P3J), and a thermophile (PDB ID: 1ZIN). These ADK structures were used preferentially over similar ADK structures because they were derived from the Bacillus family and, therefore, shared high homology. The differences in their peptide sequences are

solely derived from temperature adaption which makes them a suitable system to explore (Bae and Phillips, 2004).

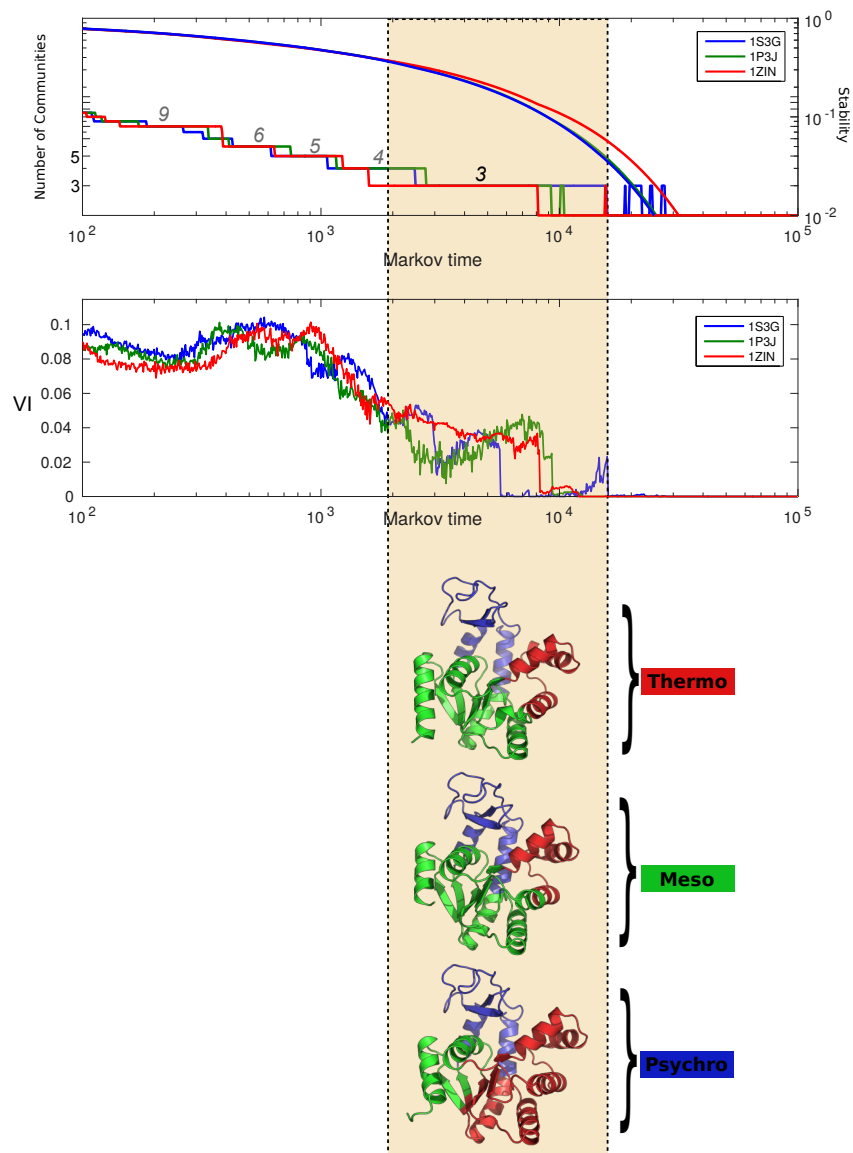


Figure 9.8: Markov stability calculations of wildtype ADK species. *Bacillus globisporus* AKpsychro (PDB ID:1S3G), (2) *Bacillus subtilis* AKmeso (PDB ID:1P3J) and (3) *Bacillus stearothermophilus* AKthermo (PDB ID:1ZIN). The MS value (top curve that decays) can be difficult to observe differences in without zooming, however, it can already be observed that there is a significant difference between the thermophile MS value and that of the mesophile and psychrophile. Below the graphs are the community partitions mapped onto the respective proteins..

A Markov stability analysis, as exhibited in Figure 9.8 indicates a variety of differences in the community structure, Markov stability and $\overline{VI}$ between the three wildtype ADK homologues. The literature defined domains are residues 1:29 74:113 180:217 for the core, residues 30:73 for the AMP_{bind}, and residues 114:179 for the lid domain (Müller et al., 1996) (the numbering

was shifted using a structure based comparison). During the 3-way partition the thermophile and mesophile partition have two similar partitions. The first is very similar to the literature defined domains where the communities are residues 1:29 76-115 179-217 for the core, residues 30:75 for the AMP_{bind}, residues 116-178 for the lid. The second partition is very similar again but α -helix 6 now joins the AMP_{bind} domain community after having originally been part of the core domain community. This is likely indicating that α -helix 6 does shift its position during ADK opening and closing. The attachment of α -helix 6 to the AMP_{bind} domain was not seen when Markov stability was used to analyse Aquifex ADK (PDB ID:2RGX) in Chapter 6. In contrast to the thermophile and mesophile, the psychrophile core subdomain is split into two during the 3-way partition, likely a consequence of weaker/fewer cross core bonds. This provides early indication that the core domain of the psychrophile has a higher flexibility due to a lower enthalpic contribution. The increased flexibility would explain the lower thermal stability exhibited by the psychrophile.

The Markov stability trajectories in the upper panel of Figure 9.8 show that thermophile to have a much higher community stability i.e. the community structure is much better defined. Despite the mesophile and psychrophile show similar Markov stability trajectories to each other, the mesophile shows a slightly higher community textitstability. From the perspective of the random-walker, there is a higher probability that it becomes trapped within one of the subdomains of the thermophile relative to its ADK homologues. This could be a result of a higher density of intra-subdomain interactions in the thermophile which provide the higher enthalpic penalty for unfolding.

A study by [Nguyen et al. \(2017\)](#) structurally compared these three structures (1P3J, 1S3G, 1ZIN) with ancestor structures and found that salt bridges were the primary source for differential molecular stabilities. In particular, they found several unique salt bridges were the principal contributors to thermostability. This was confirmed by interchanging the salt bridge residues into a mesophile and measuring the T_m .

9.5. Discussion

In Chapter 7, we introduced a Markov stability mutant score that measures the perturbation that a single mutation causes to the community structure. We then used this methodology to identify mutations in the lid subdomain that altered the conformational equilibrium of ADK. In this chapter, we have extended this comparison of theory and experiment to the core subdomain to identify mutations that result in a global reduction in molecular stability.

To validate the Markov mutant score in the core subdomain, we constructed 10 ADK mutants with high mutant scores. Using circular dichroism, we were able to measure the melting tem-

perature of each mutant ADK structure. We observed a correlation between Markov stability mutant score and a reduction in melting temperature. This relationship suggests that removing interactions that are essential for the community structure of the core subdomain causes a large reduction in molecular stability. Two control mutations that were attributed low scores by the mutant scoring methodology were used as references. Reassuringly, both controls saw very small changes in melting temperature.

A graph theoretical approach by [Brinda and Vishveshwara \(2005\)](#) found that residues with a high degree of edges acted as thermostability hubs. These hubs play a key role in bringing together secondary structural elements in the tertiary structure. This agrees with the hub-like residues we identify within the core subdomain of ADK, their interactions span across secondary structures that are split into separate communities. Another study by [Alves and Martinez \(2007\)](#) built a network based on inter-residue interactions that found the average degree of a node to be related to the hydrophobic character of each residue, suggesting the importance of hydrophobic hubs for protein stability. Further network based studies by [Aftabuddin and Kundu \(2007\)](#) found that the degree of hydrophobic networks was slightly higher than hydrophilic and charged networks, whilst their average strength was similar. They found that hydrophobic networks were highly hierarchical and would explain the reason we only identify these key residues during the 6-way and 4-way partition ([Aftabuddin and Kundu, 2007](#)).

A bioinformatic approach to measuring changes in enthalpy often ignore the 3-Dimensional contacts and thus the change in enthalpy for a double mutant would be additive. However, the double mutation doesn't attain a Markov stability mutant score that is the addition of its constituent mutants which is reflected in the non-additive change in melting temperature. These results highlight the importance of using a 3-dimensional approach.

To validate the mutant scoring methodology, we calculated the correlation of Markov mutant scores with melting temperature across each relevant timescale. Using a linear regression to fit the data, we found a similar strength correlation between the theoretical scores and experimental melting temperatures in the 4-way and 6-way partitions. However, at shorter timescales (12-way to 8-way) the correlation is weak or non-existent. These results suggest that we can predict changes to global molecular stability using long timescale partitions. Despite this, the 3-way partition was less successful at producing a correlation which highlighted the importance of the 3-way partition for modeling the dynamics of ADK.

If we assume that a higher mutant score indicates a large reduction in melting temperature, as indicated by the correlations, we can combine the mutant score with the conservation score for each mutant. Using this we were able to identify residues that were conserved across ADK species and high scoring, but also identify residues that were unique to *Aquifex* ADK whilst still attaining a high mutant score. The high scoring conserved residues are likely essential for molecular stability, despite not directly binding substrate. The high scoring unique residues are

likely evolutionary adaptations to the high temperatures that *Aquifex* *Aeolicus* is subject to; they provide additional molecular stability to the basic ADK structure. In fact, our predictions of residues essential for thermophilic adaptation agreed with a number of studies that identified the key thermophilic residues in ADK (Henzler-Wildman et al., 2007a; Nguyen et al., 2017). Only 1/28 binding residues is attributed a high Markov stability mutant score. One interpretation of this result is that binding residues need to be flexible for substrate binding and thus don't provide the essential molecular stability.

We introduced a measure of fuzziness or promiscuity which measured the probability of a node/atom leaving its optimal community and joining a neighboring community. Using this measure we were able to show that the 4-way and 6-way partitions showed a high fuzziness within the core subdomain, suggesting that atoms and residues were unsure of their optimal community. Any large changes during this timescale, such as mutation, would result in large effects on the community structure of the core subdomain. Moreover, the 3-way partition showed a high entropy between the lid and ampbind subdomains, suggesting that any mutations within the lid would cause large effects.

10. Conclusions

The principle contribution of this thesis has been to understand the relationship between Markov stability and the physical properties of proteins, and to apply this understanding to gain insight into the dynamical and functional behaviour of *Aquifex* Adenylate Kinase. Building on previous work by [Delvenne et al. \(2010\)](#), [Meliga \(2009\)](#), [Delmotte \(2014\)](#) and [Amor \(2015\)](#) we have used Markov stability to measure the effect of mutations within a protein system and then used single-molecule FRET to experimentally validate these predictions. Our graph theoretical approach offers two significant advantages over previous work. Firstly, we are able to probe long timescale dynamics that are out of reach for molecular dynamics simulations. Secondly, our graphs have atomistic detail, we consider covalent and non-covalent interactions, and we weight nodes by the bond strengths between atoms. In this manner we are able to probe the key interactions and mechanisms underlying dynamical and functional behaviour in proteins.

In Chapter 7, we introduced a novel graph-theoretic measure that we used to identify, score and rank mutations that caused a large perturbation to the community structure. We used two metrics in particular to measure the ‘effect’ of a mutation on a protein structure. The first was the $\overline{VI}$, where increases suggested a less robust community structure. The second was the CV, which quantified the change in the community structure of a mutant relative to the wildtype. We considered each metric separately and also combined both metrics into a single mutant score X . In Chapter 8 we introduced a developed and implemented a methodology to perform mutational scans to any residue type, expanding the potential for Markov stability. Using energy minimisation we were able to effectively pack the side-chains of substituted residues before converting the 3-dimensional structure into a graph.

We have constructed a set of mutant structures and experimentally measured a number of properties to test our graph theoretic results: single-molecule FRET explores the conformational equilibrium of the subdomain motions (Section 6.3); fluorescence correlation spectroscopy to measure the dynamical rates of the subdomains (Section 6.4); activity assays are used to measure the functional rates of the mutants (Section 7.2.3), and molecular dynamics simulations to probe changes in the flexibility after mutations (Section 7.2.4). Finally, we have combined the experimental measurements and graph theoretical predictions to produce a monotonic relationship between the mutant scores assigned by Markov stability and the conformational equilibrium

of the open and closed states.

We have applied the computational and experimental methods to *Aquifex* Adenylate Kinase. Our investigation identified important residues found in literature and also previously unseen residues that are critical for functional dynamical behaviour. We have provided a novel insight into Adenylate Kinases dynamical and allosteric behaviour. In the following section, we summarise the main biological results from the computational and experimental analyses. After that, we provide a short discussion of our principal contribution to the current understanding of protein dynamics. In the final section, we discuss some open questions that our work has raised, and we speculate on conducive future experimental and computational analyses.

10.1. Summary of main biological results

In Chapter 6, we applied Markov stability to *Aquifex* Adenylate Kinase, a well studied and understood dynamical protein system. We found three robust timescales that were biologically relevant and also unique to ADK. The first two community structures (12-way and 8-way partitions) highlighted the dynamical hinges known in literature whilst the third community structure (3-way partition) corresponded to the known subdomains in Adenylate Kinase. We found that the AMP_{bind} subdomain appeared at earlier Markov times to the lid subdomain which was reflected in the slower dynamical motions that we measured using FCS. Through a Markov stability analysis of the inhibitor bound conformation, we observed a more fluid, less compartmentalised community structure. Single-molecule FRET analysis of ADK in the absence of substrate found the open and closed states to be sampled equally, whilst introduction of substrate saw a shift towards a majority closed conformation.

In Chapter 7, we performed alanine mutagenesis across the entire ADK structure, analysed each mutant with Markov stability and scored mutations according to their perturbation to the community structure. To isolate mutations that would affect dynamics we looked at the lid subdomain. We found that particular mutations such as R166A, R120A, R124A, D152A and D153A would rank highly across multiple timescales. These residues form a hydrogen bonding network that when perturbed has a significant functional effect. At medium timescales we identified I118A as important, this residue has previously been implicated in a hydrophobic core within the lid subdomain that, when mutated, increases the propensity for the lid to ‘crack’. We observed that I118A shifted the conformational equilibrium whilst still catalytically active. At short timescales we identified P141A and K154A which were predicted to have local dynamical effects, we observed an increase in catalytic rate after mutation. Importantly, low scoring control mutations showed no significant changes to function or dynamics.

In Chapter 8, we further explored the D152A mutation that saw both inhibition and a large

change in the conformational equilibrium. We computationally substituted D152 for each residue type and analysed each mutant structure with Markov stability. Depending on the ‘effect’ of each substitution on the community structure, we saw a correlated shift in the population equilibrium. These results suggest that particular structural features embedded in the architecture of the protein are critical for functional motions.

The importance of D152 to ADK function is not well-known, despite not directly binding substrate it causes catastrophic failure of catalytic function. Allostery is a widely studied field however there is no mechanistic consensus, here we highlight a key allosteric residue, whereby we can modulate the dynamical equilibrium through a single mutation.

We extended our analysis to a fully encompassing arginine scan of ADK. This revealed that V164R causes a large change to the global community structure through the introduction of a salt bridge between the lid and AMP_{bind} subdomains. Single-molecule FRET showed that we were clamping the two subdomains together and had the effect of retarding the catalytic rate.

In Chapter 9, we extended our analysis of mutations on the community structure to the core subdomain. In Chapter 8 we hypothesised that we were identifying mutants that removed key interactions that would cause a global reduction in molecular stability. High scoring mutations were experimentally constructed and their melting temperatures were measured. We showed that, in general, the higher the ‘effect’ on community structure the lower melting temperature. Mutation of residues located within a hydrophobic pocket on $\beta 3$, $\alpha 4$ and $\alpha 5$ resulted in significant drops in melting temperature. We highlight the importance of an atomistic 3-dimensional approach, K179A ranked highly and also resulted in a large changing in melting temperature whilst its neighbour and twin, K178A, scored low and saw very little changing in melting temperature. This was confirmed further with D94A and D95A. We also explore the effect of double mutations and find that the effect on melting temperature is not additive of the separate mutations which agreed with the non-additive mutant score.

By combining the mutant score with the conservation score of each residue we were able to identify residues that were unique to *Aquifex* ADK but would cause a large change to the molecular stability. We find that P73A is a potential thermophilic adaption residue which agreed with previous studies that showed P73A to be an evolutionary method to increase molecular stability (Henzler-Wildman et al., 2007a). We also find previously unseen non-conserved residues such as R166A and K179A that may be the source of thermophilic adaption for *Aquifex* ADK.

10.2. A philosophical outlook

Thus far, we have provided a discussion of the observations and results seen throughout this thesis. Here, we would like to put this work into context of the bigger picture.

A review of ADK (Chapter 3) detailed its dynamical mechanisms. Firstly, our experimental results provided further evidence for the equilibrium between the open and closed conformations. Secondly, in regards to the ‘cracking’ mechanism that explains the transition from the open to closed conformations, we found that certain mutations within the lid subdomain are destabilising the lid which potentially provides further evidence for this theory. Thirdly, our analysis of ADK has led us to hypothesise that the lid subdomain may be partially disordered in the open conformation. In contrast to a ‘cracking’, we propose the possibility that the lid may be partially unfolded in the unbound open and closed conformations and simply folds into a closed conformation when induced by substrate. The shift to high FRET efficiency seems most likely to be a consequence of the lid partially unfolding, allowing the fluorophores to reorientate or become geometrically closer as a consequence of the less structured lid. Indeed, the folded crystal structures of the unbound open conformation may just be a consequence of crystal packing or the chosen buffer. This provides an important point of discussion, a large number of the studies in Chapter 3 use the crystal structures of the unbound open conformation as the initial structures for their MD and coarse-grained simulations. There is a distinct possibility that a large number of these studies have been biased by their initial starting structure.

Looking more broadly, the results within this thesis provide a larger context for protein dynamics. We have shown that graph theoretical methods can provide a novel insight into the medium- to long- timescale motions of protein structures. A one-dimensional approach to protein dynamics does lose information regarding the rotational motions, however, we have shown here that we can still extract important information from this reduced approach. Indeed, the rates at which we are able to perform complete mutagenesis scans is currently incomparable. We also emphasise the importance of taking an atomistic approach to capture the key interactions that can often govern functional dynamics.

10.3. Open questions and future work

The work in this thesis has opened a myriad of potential avenues for future research. In this section, we explore the most promising areas.

10.3.1. Derivation of combining two metrics into a single score.

During chapter 7 we introduced a combined mutant score that included both $\overline{VI}$ and CV. Whilst we found interesting relationships between the combined score and experimental properties (such as the conformational equilibrium and melting temperature), the derivation of the combined metric isn’t entirely robust. As such we continued to use $\overline{VI}$ and CV as separate measures alongside the combined score throughout the thesis.

However, there is potential for a mathematically robust statistical combination of $\overline{\text{VI}}$ and CV into a single metric to produce a true mutant score. The $\overline{\text{VI}}$ is a measure of the number of wells within partition space, if the $\overline{\text{VI}}$ is high then there are multiple wells that offer a high Markov stability. The CV measures the change of the optimal partition well of a mutant relative to the wildtype. Both measures are calculated using the Variation of Information and therefore both sit in the same ‘space’. A potential avenue is to use the Mann-Whitney U test which is a non-parametric test without the requirement of a normal distribution.

10.3.2. Mutating V164 to all residue types

Mutation of V164R saw a large shift in the conformational equilibrium towards a closed conformation, however, unlike the ensemble of D152 mutations we only saw a reduction in catalytic rates. This suggests that we are able to potentially modulate the catalytic rate using the conformational equilibrium. Modulation of the open and closed conformations has been shown to correlate with dissociation constants (Seo et al., 2014) in the maltose-binding protein and the overall turnover rate in Adenylate Kinase (Ådén et al., 2012).

Performing a similar mutagenesis study on V164 (changing to every potential amino acid) we can identify a ranging effect of mutations. In a similar manner we can measure the conformational equilibrium using smFRET and then measure the catalytic turnover of each mutant, providing a direct linkage between theory, dynamics and function.

10.3.3. Amide hydrogen-deuterium exchange to identify local unfolding events

Local unfolding events as a consequence of mutation are difficult to probe, however, amide hydrogen-deuterium (H/D) exchange is a powerful tool for investigating protein folding and stability (Englander and Kallenbach, 1983). Hydrogen-deuterium exchange is a chemical reaction in which the amide proton of the protein backbone is replaced by a deuterium atom (or vice-versa). The methodology gives localised information regarding the solvent accessibility of various parts of the protein. H/D exchange is usually monitored using NMR spectroscopy or mass spectrometry.

The results in this thesis suggest that the lid subdomain may be unfolding or increasing its propensity to unfold after mutation. Using H/D exchange we could both monitor the effect of structural mutations in more detail and also probe the ‘cracking’ mechanism that is hypothesised for the lid subdomain of ADK.

10.3.4. Prediction and validation in multimer systems

Protein oligomers are abundant in nature; over 60% of all proteins in any organism are assumed to form homo- or hetero- oligomers in various stoichiometries. Many questions concerning oligomer structure remain unexplored; what is the distribution of conformations sampled by individual monomers within oligomers? Are there differences between the conformational ensembles of different order oligomers? Do oligomers and monomers exhibit different conformational ensembles?

The methods used and developed in this thesis offer promising ways to explore oligomer dynamics. Our limited understanding of oligomer dynamics and allostery is a consequence of standard computational techniques being unable to handle their large size. Markov stability is considerably less demanding than approaches that include molecular dynamics, whilst still retaining the atomistic detail that is lost in elastic network models. Furthermore, the inherent multi-resolution properties of Markov stability make it an ideal method for identifying community structures related to domains. The quaternary structure of oligomers offers an additional spatial and temporal scale which is unseen in monomers but would be identified in an unsupervised manner by Markov stability.

10.3.5. Edge based Markov stability

We have used Markov stability to cluster atoms together at all timescales, from this we were able to identify regions that moved independently of other regions at a particular Markov time. Another actionable use of Markov stability is to look at the edge space rather than the node space. Protein motions are a consequence of the extension, rotation and vibration of interactions (edges), suggesting that an edge-centric analysis is more applicable. We can define a flow-redistribution matrix K (or edge-to-edge matrix) which quantifies the likelihood that each edge is critical to flow redistribution in the network. The flow redistribution matrix can be interpreted as an adjacency matrix of the effective edge-to-edge network. Application of Markov stability will cluster edges according to their influence on each other (Schaub et al., 2014).

The analysis will potentially allow us to cluster (long distance) regions of the protein at different timescales. For example, we would assume that hinges regions around the protein would cluster together given that they corresponded to motions of a similar timescale. Additionally, we may be able to cluster interactions that after mutation would have a significant effect on structure or function respectively (edges).

10.3.6. Markov State Models using Markov Stability

Markov State Models (MSMs) are constructed from detailed molecular dynamics simulations and provide a kinetic model of a random process e.g. a protein folding pathway, or transitions between conformational states. These simplified models are coarse-grained representations of the more detailed models. A network of the MD conformational states (micro-states) is produced using some measure of similarity (k-means), the nodes correspond to conformational snapshots of the protein and the edges correspond to the similarity between two conformational states. A transition matrix is produced that is based on the MD probability of going from state i to state j . Current clustering of micro-states and macro-states in MSMs are performed using methods such as k-means and PCCA.

Using Markov Stability, we would be able to provide an unbiased method of clustering the conformational states of an MD simulation to produce a Markov State Model. Using this, we would be able to identify the minor-states at short Markov timescales and the major-states at long Markov timescales. Combining this methodology with free energy calculations of each state, we would be able to map the energy landscape of a protein structure, including the inherent coarseness and width of the energy wells, and kinetic barriers that separate them. Since we are able to access the hierarchy of timescales, we would be able to access the hierarchy of conformational motions across the energy landscape. Of course, this assumes high-quality MD data that samples the complete ensemble of states.

Bibliography

- Abraham, M. J., Murtola, T., Schulz, R., Pall, S., Smith, J. C., Hess, B., and Lindah, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25.
- Ådén, J., Verma, A., Schug, A., and Wolf-Watz, M. (2012). Modulation of a pre-existing conformational equilibrium tunes adenylate kinase activity. *Journal of the American Chemical Society*, 134(40):16562–16570.
- Ådén, J., Weise, C. F., Brännström, K., Olofsson, A., and Wolf-Watz, M. (2013). Structural topology and activation of an initial adenylate kinase-substrate complex. *Biochemistry*, 52(6):1055–1061.
- Ådén, J. and Wolf-Watz, M. (2007). NMR identification of transient complexes critical to adenylate kinase catalysis. *Journal of the American Chemical Society*, 129(45):14003–14012.
- Aftabuddin, M. and Kundu, S. (2007). Hydrophobic, Hydrophilic, and Charged Amino Acid Networks within Protein. *Biophysical Journal*, 93(1):225–231.
- Ahmed, A., Rippmann, F., Barnickel, G., and Gohlke, H. (2011). A Normal Mode-Based Geometric Simulation Approach for Exploring Biologically Relevant Conformational Transitions in Proteins. pages 1604–1622.
- Alder, B. J. and Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *J. Chem. Phys.*, 27(5):1208–1209.
- Alexandrov, V. (2005). Normal modes for predicting protein motions: A comprehensive database assessment and associated Web tool. *Protein Science*, 14(3):633–643.
- Alexiev, U., Rimke, I., and Pohlmann, T. (2003). Elucidation of the nature of the conformational changes of the EF-interhelical loop in bacteriorhodopsin and of the helix VIII on the cytoplasmic surface of bovine rhodopsin: A time-resolved fluorescence depolarization study. *Journal of Molecular Biology*, 328(3):705–719.

- Alves, N. A. and Martinez, A. S. (2007). Inferring topological features of proteins from amino acid residue networks. *Physica A: Statistical Mechanics and its Applications*, 375(1):336–344.
- Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425.
- Amor, B. (2015). Exploring allostery in proteins with graph theory. *PhD Thesis Imperial College*.
- Amor, B., Yaliraki, S. N., Woscholski, R., and Barahona, M. (2014). Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Mol. BioSyst.*, 10(8):2247–2258.
- Anderson, D. E., Becktel, W. J., and Dahlquist, F. W. (1990). pH-Induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, 29(9):2403–2408.
- Arenas, A., Diaz-Guilera, A., and Perez-Vicente, C. J. (2006). Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 96(11).
- Arkhipov, A., Shan, Y., Das, R., Endres, N. F., Eastwood, M. P., Wemmer, D. E., Kuriyan, J., and Shaw, D. E. (2013). Architecture and membrane interactions of the EGF receptor. *Cell*, 152(3):557–569.
- Arora, K. and Brooks, C. L. (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18496–18501.
- Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 38.
- Atilgan, A. R., Akan, P., and Baysal, C. (2004). Small-World Communication of Residues and Significance for Protein Dynamics. *Biophysical Journal*, 86(1):85–91.
- Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–73.
- Bae, E. and Phillips, G. N. (2004). Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *Journal of Biological Chemistry*, 279(27):28202–28208.
- Bae, E. and Phillips, G. N. (2005). Identifying and engineering ion pairs in adenylate kinases: Insights from molecular dynamics simulations of thermophilic and mesophilic homologues. *Journal of Biological Chemistry*, 280(35):30943–30948.

- Bae, E. and Phillips, G. N. (2006). Roles of static and dynamic domains in stability and catalysis of adenylate kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2132–2137.
- Bahar, I. and Jernigan, R. L. (1998). Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. *Journal of Molecular Biology*, 281(5):871–884.
- Balakrishnan, G., Case, M. A., Pevsner, A., Zhao, X., Tengroth, C., McLendon, G. L., and Spiro, T. G. (2004). Time-resolved absorption and UV resonance raman spectra reveal stepwise formation of T quaternary contacts in the allosteric pathway of hemoglobin. *Journal of Molecular Biology*, 340(4):843–856.
- Barahona, M. and Pecora, L. M. (2002). Synchronization in Small-World System. *Phys. Rev. Lett.*, 89(5):054101.
- Bashton, M. and Chothia, C. (2007). The Generation of New Protein Functions by the Combination of Domains. *Structure*, 15(1):85–99.
- Beckstein, O., Denning, E. J., Perilla, J. R., and Woolf, T. B. (2009). Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open to Closed Transitions. *Journal of Molecular Biology*, 394(1):160–176.
- Bellinzoni, M., Haouz, A., and Gran, M. (2006). The crystal structure of Mycobacterium tuberculosis adenylate kinase in complex with two molecules of ADP and Mg²⁺ supports an associative mechanism for phosphoryl transfer. pages 1489–1493.
- Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing.*, volume 17, pages 6–17.
- Benson, N. C. and Daggett, V. (2012). A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *Journal of bioinformatics and computational biology*, 10(4):1250008.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.
- Berry, M. B., Meador, B., Bilderback, T., Liang, P., Glaser, M., and Phillips, G. N. (1994). The closed conformation of a highly flexible protein: The structure of E. coli adenylate kinase with bound AMP and AMPPNP. *Proteins: Structure, Function and Genetics*, 19(3):183–198.
- Biarnés, X., Ardèvol, A., Iglesias-Fernández, J., Planas, A., and Rovira, C. (2011). Catalytic itinerary in 1,3-1,4- β -glucanase unraveled by QM/MM metadynamics. Charge is not yet

- fully developed at the oxocarbenium ion-like transition state. *Journal of the American Chemical Society*, 133(50):20301–20309.
- Bilderback, T., Fulmer, T., Mantulin, W. W., and Glaser, M. (1996). Substrate binding causes movement in the ATP binding domain of Escherichia coli adenylate kinase. *Biochemistry*, 35(19):6100–6106.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *Pattern Recognition*, 4(4):738.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10008(10):6.
- Boccaletti, S., Ivanchenko, M., Latora, V., Pluchino, A., and Rapisarda, A. (2007). Detecting complex network modularity by dynamical clustering. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 75(4).
- Boehr, D., McElheny, D., Dyson, H., and Wright, P. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. *Science (New York, NY)*, 313(5793):1638–1642.
- Borgia, A., Williams, P. M., and Clarke, J. (2008). Single-Molecule Studies of Protein Folding. *Annual Review of Biochemistry*, 77(1):101–125.
- Bosshard, H. R., Marti, D. N., and Jelesarov, I. (2004). Protein stabilization by salt bridges: Concepts, experimental approaches and clarification of some misunderstandings. *Journal of Molecular Recognition*, 17(1):1–16.
- Brandes, U., Dellinger, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.
- Brinda, K. V. and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophysical journal*, 89(6):4159–70.
- Brokaw, J. B. and Chu, J. W. (2010). On the roles of substrate binding and hinge unfolding in conformational changes of adenylate kinase. *Biophysical Journal*, 99(10):3420–3429.
- Cammarata, M., Levantino, M., Schotte, F., Anfinrud, P. A., Ewald, F., Choi, J., Cupane, A., Wulff, M., and Ihee, H. (2008). Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat Methods*, 5(10):881–886.
- Celej, M. S., Montich, G. G., and Fidelio, G. D. (2003). Protein stability induced by ligand binding correlates with changes in protein flexibility. *Protein science : a publication of the Protein Society*, 12(7):1496–1506.

- Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26:689–691.
- Chennubhotla, C. and Bahar, I. (2007). Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS computational biology*, 3(9):1716–26.
- Chu, J.-W. and Voth, G. A. (2007). Coarse-Grained Free Energy Functions for Studying Protein Conformational Changes: A Double-Well Network Model. *Biophysical Journal*, 93(11):3860–3871.
- Chung, H. S., Cellmer, T., Louis, J. M., and Eaton, W. A. (2013). Measuring ultrafast protein folding rates from photon-by-photon analysis of single molecule fluorescence trajectories. *Chemical Physics*, 422:229–237.
- Chung, H. S., McHale, K., Louis, J. M., and Eaton, W. A. (2012). Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science*, 335(6071):981–984.
- Cieplak, M. and Banavar, J. R. (2013). Energy landscape and dynamics of proteins: An exact analysis of a simplified lattice model. *Physical Review E*, 88(4):040702.
- Clamme, J. (2005). Three Color Single Molecule Fluorescence Resonance Energy Transfer. *ChemPhysChem*, 6(1):74–77.
- Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.
- Combs, C. A. (2010). Fluorescence microscopy: A concise guide to current imaging methods. *Current Protocols in Neuroscience*, (SUPPL.50).
- Cukier, R. I. (2009). Apo adenylate kinase encodes its holo form: A principal component and varimax analysis. *Journal of Physical Chemistry B*, 113(6):1662–1672.
- Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Science*, 6(6):1333–1337.
- Daily, M. D., Phillips, G. N., and Cui, Q. (2010). Many Local Motions Cooperate to Produce the Adenylate Kinase Conformational Transition. *Journal of Molecular Biology*, 400(3):618–631.
- Daily, M. D., Phillips, G. N., and Cui, Q. (2011). Interconversion of functional motions between mesophilic and thermophilic adenylate kinases. *PLoS Computational Biology*, 7(7):e1002103.
- del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology*, 2(2006.0019).

- del Sol, A., Fujihashi, H., and O'Meara, P. (2005). Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315.
- DeLano, W. L. (2002). The PyMOL Molecular Graphics System. *Schrödinger LLC* www.pymol.org, Version 1.:<http://www.pymol.org>.
- Delmotte, A. (2014). All-scale structural analysis of biomolecules through dynamical graph partitioning. *PhD Thesis Imperial College*.
- Delmotte, A., Tate, E. W., Yaliraki, S. N., and Barahona, M. (2011). Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin myosin light chain interaction. *Physical Biology*, 8(5):pp. 55010–55021(12).
- Delvenne, J.-C., Schaub, M., Yaliraki, S., and Barahona, M. (2013). The stability of a graph partition: A dynamics-based framework for community detection. *Dynamics On and Of Complex Networks, Springer 2013*, Volume 2:pp 221–242,.
- Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12755–12760.
- Dey, M., Cao, C., Sicheri, F., and Dever, T. E. (2007). Conserved intermolecular salt bridge required for activation of protein kinases PKR, GCN2, and PERK. *Journal of Biological Chemistry*, 282(9):6653–6660.
- Dill, K. A., Bromberg, S., Yue, K., Chan, H. S., Ftebig, K. M., Yee, D. P., and Thomas, P. D. (1995). Principles of protein folding - A perspective from simple exact models.
- Dokholyan, N. V., Li, L., Ding, F., and Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proceedings of the National Academy of Sciences*, 99(13):8637–8641.
- Dongen, S. V. (2000). A cluster algorithm for graphs. *Information Systems [INS]*, (R 0010):1–40.
- Dorogovtsev, S. N., Goltsev, a. V., and Mendes, J. F. F. (2004). Potts model on complex networks. *The European Physical Journal B - Condensed Matter*, 38(2):177–182.
- Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104.
- Dunitz, J. D. (1995). Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chemistry and Biology*, 2(11):709–712.
- Elber, R. and Karplus, M. (1987). Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science (New York, N.Y.)*, 235(4786):318–21.

- Elson, E. L. (2011). Fluorescence correlation spectroscopy: Past, present, future. *Biophysical Journal*, 101(12):2855–2870.
- Elson, E. L. and Magde, D. (1974). Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers*, 13:1–27.
- Enderlein, È. and Keller, R. a. (2002). Single molecule detection in solution methods and applications. *Molecules*, 8:371.
- Englander, S. W. and Kallenbach, N. R. (1983). Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Quart. Rev. Biophys.*, 16(4):521–655.
- Esteban-Martín, S., Fenwick, R. B., Ådén, J., Cossins, B., Bertoncini, C. W., Guallar, V., Wolf-Watz, M., and Salvatella, X. (2014). Correlated Inter-Domain Motions in Adenylate Kinase. *PLoS Computational Biology*, 10(7):1–7.
- Fahey, P. F., Koppel, D. E., Barak, L. S., Wolf, D. E., Elson, E. L., and Webb, W. W. (1977). Lateral diffusion in planar lipid bilayers. *Science*, 195(4275):305–306.
- Felli, I. C. and Brutscher, B. (2009). Recent advances in solution NMR: Fast methods and heteronuclear direct detection. *ChemPhysChem*, 10(9-10):1356–1368.
- Feng, Y., Yang, L., Kloczkowski, A., and Jernigan, R. L. (2009). The energy profiles of atomic conformational transition intermediates of adenylate kinase. *Proteins: Structure, Function and Bioinformatics*, 77(3):551–558.
- Fenwick, R. B., Esteban-Martin, S., Richter, B., Lee, D., Walter, K. F. A., Milovanovic, D., Becker, S., Lakomek, N. A., Griesinger, C., and Salvatella, X. (2011a). Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *Journal of the American Chemical Society*, 133(27):10336–10339.
- Fenwick, R. B., Esteban-Martin, S., and Salvatella, X. (2011b). Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *European Biophysics Journal*, 40(12):1339–1355.
- Fernandez-Leiro, R., Conrad, J., Scheres, S. H. W., and Lamers, M. H. (2015). Cryo-EM structures of the E. Coli replicative DNA polymerase reveal its dynamic interactions with the DNA sliding clamp, exonuclease and τ . *eLife*, 4(e11134).
- Fernandez-Leiro, R. and Scheres, S. H. W. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537(7620):339–346.
- Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305.

- Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633.
- Formoso, E., Limongelli, V., and Parrinello, M. (2015). Energetics and structural characterization of the large-scale functional motion of adenylate kinase. *Scientific reports*, 5:8425.
- Förster, T., Energiewanderung, Z., and Von, F. (1939). Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik*, 248(1938):55–75.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.
- Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41.
- Fortunato, S. and Castellano, C. (2012). Community structure in graphs. In *Computational Complexity: Theory, Techniques, and Applications*, volume 9781461418, pages 490–512.
- Frank, J., Verschoor, A., and Boublik, M. (1981). Computer averaging of electron micrographs of 40S ribosomal subunits. *Science (New York, N.Y.)*, 214(4527):1353–1355.
- Gagné, D., Charest, L. A., Morin, S., Kovrigin, E. L., and Doucet, N. (2012). Conservation of flexible residue clusters among structural and functional enzyme homologues. *Journal of Biological Chemistry*, 287(53):44289–44300.
- Gfeller, D. and De Los Rios, P. (2007). Spectral coarse graining of complex networks. *Physical Review Letters*, 99(3).
- Gilson, M. K. and Honig, B. H. (1986). The dielectric constant of a folded protein. *Biopolymers*, 25(11):2097–2119.
- Giri Rao, V. V. H. and Gosavi, S. (2014). In the Multi-domain Protein Adenylate Kinase, Domain Insertion Facilitates Cooperative Folding while Accommodating Function at Domain Interfaces. *PLoS Computational Biology*, 10(11).
- Gómez, H., Polyak, I., Thiel, W., Lluch, J. M., and Masgrau, L. (2012). Retaining glycosyltransferase mechanism studied by QM/MM methods: Lipopolysaccharyl- α -1,4-galactosyltransferase C transfers α -galactose via an oxocarbenium ion-like transition state. *Journal of the American Chemical Society*, 134(10):4743–4752.
- Gopich, I. V. and Szabo, A. (2007). Single-molecule FRET with diffusion and conformational dynamics. *Journal of Physical Chemistry B*, 111(44):12925–12932.
- Gramatikov, B. I. (2014). Modern technologies for retinal scanning and imaging: An introduction for the biomedical engineer.

- Greenfield, N. and Fasman, G. D. (1969). Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, 8(10):4108–4116.
- Greenfield, N. J. (2007). Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nature Protocols*, 1(6):2527–2535.
- Greenfield, N. J. (2009). Analysis of the kinetics of folding of proteins and peptides using circular dichroism. *Nature protocols*, 1(6):2891–2899.
- Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 70(2 2).
- Gur, M., Madura, J. D., and Bahar, I. (2013). Global transitions of proteins explored by a multiscale hybrid methodology: Application to adenylate kinase. *Biophysical Journal*, 105(7):1643–1652.
- Ha, T., Enderle, T., Ogletree, D. F., Chemla, D. S., Selvin, P. R., and Weiss, S. (1996). Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences*, 93(13):6264–6268.
- Hammes, G. G., Chang, Y.-C., and Oas, T. G. (2009). Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13737–41.
- Hanson, J. a., Duderstadt, K., Watkins, L. P., Bhattacharyya, S., Brokaw, J., Chu, J.-W., and Yang, H. (2007). Illuminating the mechanistic roles of enzyme conformational dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):18055–18060.
- Hayward, S., Kitao, a., and Berendsen, H. J. (1997). Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins*, 27(3):425–437.
- Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–972.
- Henzler-Wildman, K. a., Lei, M., Thai, V., Kerns, S. J., Karplus, M., and Kern, D. (2007a). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916.

- Henzler-Wildman, K. a., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. a., Petsko, G. a., Karplus, M., Hübner, C. G., and Kern, D. (2007b). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844.
- Hess, S. T. and Webb, W. W. (2002). Focal Volume Optics and Experimental Artifacts in Confocal Fluorescence Correlation Spectroscopy. *Biophysical Journal*, 83(4):2300–2317.
- Hobza, P. (2008). Stacking interactions. *Physical chemistry chemical physics : PCCP*, 10(19):2581–2583.
- Hoffmann, N. A., Jakobi, A. J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W. J. H., Sachse, C., and Müller, C. W. (2015). Molecular structures of unbound and transcribing RNA polymerase III. *Nature*, 528(7581):231–236.
- Hopf, T. a., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621.
- Huang, C. Y., Getahun, Z., Wang, T., DeGrado, W. F., and Gai, F. (2001). Time-resolved infrared study of the helix-coil transition using C-13-labeled helical peptides. *Journal of the American Chemical Society*, 123(48):12111–12112.
- Huheey, J., Keiter, E., and Keiter, R. (1993). *Inorganic Chemistry, Principles of Structure and Bonding*. Harper Collins College Publishers, New York.
- Icenogle, R. D. and Elson, E. L. (1983). Fluorescence correlation spectroscopy and photobleaching recovery of multiple binding reactions. I. Theory and FCS measurements. *Biopolymers*, 22(8):1919–1948.
- Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–65.
- Jana, B., Adkar, B. V., Biswas, R., and Bagchi, B. (2011). Dynamic coupling between the LID and NMP domain motions in the catalytic conversion of ATP and AMP to ADP by adenylate kinase. *Journal of Chemical Physics*, 134(3).
- Jensen, M. O., Jogini, V., Borhani, D. W., Leffler, A. E., Dror, R. O., and Shaw, D. E. (2012). Mechanism of Voltage Gating in Potassium Channels. *Science*, 336(6078):229–233.
- Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., and Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, 351(6275):867–871.
- Jorgensen, W. L. and Tirado-rives, J. (1988). The OPLS Potential Functions for Proteins. Energy Minimisations for Crystals of Cyclic Peptides and Crambia. *Journal of the American Chemical Society*, 110(4):1657–1666.

- Jorgensen, W. L. and Tirado-Rives, J. (2005). Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J Comput Chem*, 26(16):1689–1700.
- Jung, J., Lee, J., and Moon, H. T. (2005). Topological determinants of protein unfolding rates. *Proteins: Structure, Function and Genetics*, 58(2):389–395.
- Kamerlin, S. C. L. and Warshel, A. (2010). Reply to Karplus: Conformational dynamics have no role in the chemical step. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):E72.
- Kannan, N. and Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of molecular biology*, 292(2):441–64.
- Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Folding & design*, 2(4):S69–S75.
- Karrer, B., Levina, E., and Newman, M. E. J. (2008). Robustness of community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 77(4).
- Kern, D., Eisenmesser, E. Z., and Wolf-Watz, M. (2005). Enzyme dynamics during catalysis measured by NMR spectroscopy. *Methods in Enzymology*, 394:507–524.
- Kern, D. and Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, 13(6):748–757.
- Kerns, S. J., Agafonov, R. V., Cho, Y.-J., Pontiggia, F., Otten, R., Pachov, D. V., Kutter, S., Phung, L. A., Murphy, P. N., Thai, V., Alber, T., Hagan, M. F., and Kern, D. (2015). The energy landscape of adenylate kinase during catalysis. *Nature Structural & Molecular Biology*, 22(2):124–131.
- Khrapunov, S. (2009). Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis. *Analytical Biochemistry*, 389(2):174–176.
- Kim, B. and Hirata, F. (2013). Structural fluctuation of protein in water around its native state: A new statistical mechanics formulation. *Journal of Chemical Physics*, 138(5).
- Kim, E., Lee, S., Jeon, A., Choi, J. M., Lee, H.-S., Hohng, S., and Kim, H.-S. (2013). A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. *Nature Chemical Biology*, 9(5):313–8.
- Kim, H. S., Wilce, M. C. J., Yoga, Y. M. K., Pendini, N. R., Gunzburg, M. J., Cowieson, N. P., Wilson, G. M., Williams, B. R. G., Gorospe, M., and Wilce, J. A. (2011). Different modes of interaction by TIAR and HuR with target RNA and DNA. *Nucleic Acids Research*, 39(3):1117–1130.

- Kim, Y. S., Liu, L., Axelsen, P. H., and Hochstrasser, R. M. (2009). 2D IR provides evidence for mobile water molecules in beta-amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 106(42):17751–17756.
- König, I., Zarrine-Afsar, A., Aznauryan, M., Soranno, A., Wunderlich, B., Dingfelder, F., Stüber, J. C., Plückthun, A., Nettels, D., and Schuler, B. (2015). Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nature Methods*, 12(8):773–779.
- Kovermann, M., Grundström, C., Sauer-Eriksson, A. E., Sauer, U. H., and Wolf-Watz, M. (2017). Structural basis for ligand binding to an enzyme by a conformational selection pathway. *Proceedings of the National Academy of Sciences*, 114(24):201700919.
- Krishnamurthy, H., Lou, H., Kimple, A., Vieille, C., and Cukier, R. I. (2005). Associative mechanism for phosphoryl transfer: A molecular dynamics simulation of Escherichia coli adenylate kinase complexed with its substrates. *Proteins: Structure, Function and Genetics*, 58(1):88–100.
- Kristjansson, M. M. and Kinsella, J. E. (1991). Protein and Enzyme Stability: Structural, Thermodynamic, and Experimental Aspects. *Advances in Food and Nutrition Research*, 35(C):237–316.
- Kubitzki, M. B. and de Groot, B. L. (2008). The Atomistic Mechanism of Conformational Transition in Adenylate Kinase: A TEE-REX Molecular Dynamics Study. *Structure*, 16(8):1175–1182.
- Kuzmanic, A., Pannu, N. S., and Zagrovic, B. (2014). X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature Communications*, 5.
- Lafferty, J. and Kondor, R. I. (2002). Diffusion Kernels on Graphs and Other Discrete Input Spaces. *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315—322.
- Lakowicz, J. (1999). *Principles of Fluorescence Spectroscopy*. Plenum Press, New York, 2nd edition.
- Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2014). Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76 – 90.
- Lee, J., Joo, K., Brooks, B. R., and Lee, J. (2015). The Atomistic Mechanism of Conformational Transition of Adenylate Kinase Investigated by Lorentzian Structure-Based Potential. *Journal of Chemical Theory and Computation*, 11(7):3211–3224.

- Lee, J., Lee, S., Ragunathan, K., Joo, C., Ha, T., and Hohng, S. (2010). Single-molecule four-color FRET. *Angewandte Chemie - International Edition*, 49(51):9922–9925.
- Lee, W. G., Gallardo-Macias, R., Frey, K. M., Spasov, K. A., Bollini, M., Anderson, K. S., and Jorgensen, W. L. (2013). Picomolar inhibitors of HIV reverse transcriptase featuring bi-cyclic replacement of a cyanovinylphenyl group. *Journal of the American Chemical Society*, 135(44):16705–16713.
- Lenn, T., Gkekas, C. N., Bernard, L., Engl, C., Jovanovic, G., Buck, M., and Ying, L. (2011). Measuring the stoichiometry of functional PspA complexes in living bacterial cells by single molecule photobleaching. *Chem. Commun.*, 47(1):400–402.
- Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. (2005). An Analysis of Core Deformations in Protein Superfamilies. *Biophysical Journal*, 88(2):1291–1299.
- Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253(5494):694–698.
- Levy, Y. and Onuchic, J. N. (2004). Water and proteins: A love-hate relationship. *Proceedings of the National Academy of Sciences*, 101(10):3325–3326.
- Li, D., Liu, M. S., and Ji, B. (2015). Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. *Biophysical Journal*, 109(3):647–660.
- Lichtman, J. W. and Conchello, J.-A. (2005). Fluorescence microscopy. *Nature Methods*, 2(12):910–919.
- Lin, M. S., Fawzi, N. L., and Head-Gordon, T. (2007). Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure (London, England : 1993)*, 15(6):727–40.
- Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132.
- Lindorff-Larsen, K., Maragakis, P., Piana, S., and Shaw, D. E. (2016). Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *Journal of Physical Chemistry B*, 120(33):8313–8320.
- Lou, H. and Cukier, R. I. (2006). Molecular dynamics of apo-adenylate kinase: A distance replica exchange method for the free energy of conformational fluctuations. *Journal of Physical Chemistry B*, 110(47):24121–24137.
- Lu, Q. and Wang, J. (2008). Single molecule conformational dynamics of adenylate kinase: Energy landscape, structural correlations, and transition state ensembles. *Journal of the American Chemical Society*, 130(14):4772–4783.

- Magde, D., Elson, E., and Webb, W. W. (1972). Thermodynamic fluctuations in a reacting system measurement by fluorescence correlation spectroscopy. *Physical Review Letters*, 29(11):705–708.
- Maragakis, P. and Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *Journal of Molecular Biology*, 352(4):807–822.
- Margittai, M., Widengren, J., Schweinberger, E., Schroder, G. F., Felekyan, S., Haustein, E., Konig, M., Fasshauer, D., Grubmuller, H., Jahn, R., and Seidel, C. A. (2003). Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proc Natl Acad Sci U S A*, 100(26):15516–15521.
- McCammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267(5612):585–590.
- Meila, M. (2007). Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.
- Meliga, S. (2009). *Master’s Thesis, Imperial College London*. PhD thesis.
- Menk, R. H., Sarvestani, A., Amenitsch, H., Bernstorff, S., Besch, H. J., Orthen, A., Pavel, N., Rappolt, M., Sauer, N., and Walenta, A. H. (2000). Novel detector systems for time resolved SAXS experiments. In *Journal of Applied Crystallography*, volume 33, pages 778–781.
- Meseth, U., Wohland, T., Rigler, R., and Vogel, H. (1999). Resolution of fluorescence correlation measurements. *Biophysical journal*, 76(3):1619–1631.
- Milles, S., Koehler, C., Gambin, Y., Deniz, A. A., and Lemke, E. A. (2012). Intramolecular three-colour single pair FRET of intrinsically disordered proteins with increased dynamic range. *Molecular BioSystems*, 8(10):2531.
- Minsky, M. (1988). Memoir on inventing the confocal scanning microscope. *Scanning*, 10(4):128–138.
- Mittermaier, A. and Kay, L. E. (2006). New tools provide new insights in NMR studies of protein dynamics. *Science (New York, N.Y.)*, 312(5771):224–8.
- Miyashita, O., Onuchic, J. N., and Wolynes, P. G. (2003). Nonlinear elasticity, protein quakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences*, 100(22):12570–12575.
- Miyashita, O., Wolynes, P. G., and Onuchic, J. N. (2005). Simple energy landscape model for the kinetics of functional transitions in proteins. *Journal of Physical Chemistry B*, 109(5):1959–1969.

- Moerner, W. E. and Fromm, D. P. (2003). Methods of single-molecule fluorescence spectroscopy and microscopy. *Review of Scientific Instruments*, 74(8):3597–3619.
- Møller, M., Nielsen, S. S., Ramachandran, S., Li, Y., Tria, G., Streicher, W., Petoukhov, M. V., Cerione, R. A., Gillilan, R. E., and Vestergaard, B. (2013). Small Angle X-Ray Scattering Studies of Mitochondrial Glutaminase C Reveal Extended Flexible Regions, and Link Oligomeric State with Enzyme Activity. *PLoS ONE*, 8(9).
- Moore, C. (2017). The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. *arXiv:1702.00467*.
- Morcos, F., Jana, B., Hwa, T., and Onuchic, J. N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538.
- Müller, C. W., Schlauderer, G. J., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure (London, England : 1993)*, 4(2):147–156.
- Müller, C. W. and Schulz, G. E. (1992). Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state. *Journal of molecular biology*, 224:159–177.
- Munro, I., Pecht, I., and Stryer, L. (1979). Subnanosecond motions of tryptophan residues in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 76(1):56–60.
- Nagarajan, S., Amir, D., Grupi, A., Goldenberg, D. P., Minton, A. P., and Haas, E. (2011). Modulation of functionally significant conformational equilibria in adenylate kinase by high concentrations of trimethylamine oxide attributed to volume exclusion. *Biophysical journal*, 100(12):2991–2999.
- Nanodrop (2009). <http://www.nanodrop.com/Library/NanoDrop%202000%20User%20Manual.pdf>.
- Neudecker, P., Lundstrom, P., and Kay, L. E. (2009). Relaxation dispersion NMR spectroscopy as a tool for detailed studies of protein folding. *Biophysical Journal*, 96(6):2045–2054.
- Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):1–16.
- Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69(6 2).
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582.

- Nguyen, V., Wilson, C., Hoemberger, M., Stiller, J. B., Agafonov, R. V., Kutter, S., English, J., Theobald, D. L., and Kern, D. (2017). Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science*, 355(6322):289–294.
- Ohki, S. y. and Kainosho, M. (2008). Stable isotope labeling methods for protein NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 53(4):208–226.
- Oikawa, H., Suzuki, Y., Saito, M., Kamagata, K., Arai, M., and Takahashi, S. (2013). Microsecond dynamics of an unfolded protein by a line confocal tracking of single molecule fluorescence. *Scientific Reports*, 3.
- Olsson, U. and Wolf-Watz, M. (2010). Overlap between folding and functional energy landscapes for adenylate kinase conformational change. *Nature communications*, 1(8):111.
- Omega (2017). <http://www.omegafilters.com/capabilities/neutral-density>.
- Ono, J., Takada, S., and Saito, S. (2015). Couplings between hierarchical conformational dynamics from multi-time correlation functions and two-dimensional lifetime spectra: Application to adenylate kinase. *Journal of Chemical Physics*, 142(21).
- Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences*, 92(8):3626–3630.
- Optoelectronics, P.-E. (2002). SPCM-AQRH Single Photon Counting Module. *Perkin-Elmer Optoelectronics Datasheet*, pages 1–10.
- OptoSigma (2017). <http://www.optosigma.com/en.eu/Catalogs/>.
- Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chacón, P., and Orozco, M. (2010). Approaching elastic network models to molecular dynamics flexibility. *Journal of Chemical Theory and Computation*, 6(9):2910–2923.
- Orellana, L., Yoluk, O., Carrillo, O., Orozco, M., and Lindahl, E. (2016). Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. *Nature communications*, 7:12575.
- Ostmeyer, J., Chakrapani, S., Pan, A. C., Perozo, E., and Roux, B. (2013). Recovery from slow inactivation in K⁺ channels is controlled by water molecules. *Nature*, 501(7465):121–124.
- Otosu, T., Ishii, K., and Tahara, T. (2015). Microsecond protein dynamics observed at the single-molecule level. *Nature communications*, 6:7685.
- Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Schell, D., Thurlkill, R. L., Imura, S., Scholtz, J. M., Gajiwala, K., Sevcik, J., Urbanikova, L., Myers, J. K., Takano, K.,

- Hebert, E. J., Shirley, B. A., and Grimsley, G. R. (2014a). Contribution of hydrogen bonds to protein stability. *Protein Science*, 23(5):652–661.
- Pace, N. C., Scholtz, M. J., and Grimsley, G. R. (2014b). Forces stabilizing proteins. *FEBS Letters*, 588(14):2177–2184.
- Pan, A. C., Borhani, D. W., Dror, R. O., and Shaw, D. E. (2013). Molecular determinants of drug-receptor binding kinetics. *Drug Discovery Today*, 18(13-14):667–673.
- Peng, C., Zhang, L., and Head-Gordon, T. (2010). Instantaneous normal modes as an unforced reaction coordinate for protein conformational transitions. *Biophysical journal*, 98(10):2356–2364.
- Pesonen, J., Henriksson, K. O. E., López-Blanco, J. R., and Chacón, P. (2012). Normal mode analysis of molecular motions in curvilinear coordinates on a non-Eckart body-frame: An application to protein torsion dynamics. *Journal of Mathematical Chemistry*, 50(6):1521–1549.
- Piana, S., Bucher, D., Carloni, P., and Rothlisberger, U. (2004). Reaction mechanism of HIV-1 protease by hybrid car-parrinello/classical MD simulations. *Journal of Physical Chemistry B*, 108(30):11139–11149.
- Pierdominici-Sottile, G., Horenstein, N. A., and Roitberg, A. E. (2011). Free Energy Study of the Catalytic Mechanism of Trypanosoma cruzi trans-Sialidase. From the Michaelis Complex to the Covalent Intermediate. *Biochemistry*, 50(46):10150–10158.
- Ping, J., Hao, P., Li, Y. X., and Wang, J. F. (2013). Molecular dynamics studies on the conformational transitions of adenylate kinase: A computational evidence for the conformational selection mechanism. *BioMed Research International*, 2013(628536):7.
- Pirchi, M., Ziv, G., Riven, I., Cohen, S. S., Zohar, N., Barak, Y., and Haran, G. (2011). Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nature Communications*, 2:493.
- Pisliakov, A. V., Cao, J., Kamerlin, S. C. L., and Warshel, A. (2009). Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17359–17364.
- Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks (long version). *Computer and Information Sciences - ISCIS 2005*, pages 284–293.
- Pontiggia, F., Zen, A., and Micheletti, C. (2008). Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophysical journal*, 95(12):5901–12.

- Potoyan, D. A., Zhuravlev, P. I., and Papoian, G. A. (2012). Computing free energy of a large-scale allosteric transition in adenylate kinase using all atom explicit solvent simulations. *Journal of Physical Chemistry B*, 116(5):1709–1715.
- Rader, a. J., Hespenheide, B. M., Kuhn, L. a., and Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Track II):3540–3545.
- Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.
- Rasmussen, C. E. and Williams, C. K. I. (2004). Gaussian processes for machine learning. *International journal of neural systems*, 14(2):69–106.
- Reichardt, J. and Bornholdt, S. (2006). Statistical Mechanics of Community Detection. *Physical Review E*, 74(1):1–16.
- Reuveni, S., Klafter, J., and Granek, R. (2012). Dynamic structure factor of vibrating fractals. *Physical Review Letters*, 108(6).
- Rhoades, E., Gussakovsky, E., and Haran, G. (2003). Watching proteins fold one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3197–3202.
- Ribeiro, A. A. S. T. and Ortiz, V. (2014). Determination of signaling pathways in proteins through network theory: Importance of the topology. *Journal of Chemical Theory and Computation*, 10(4):1762–1769.
- Ronhovde, P. and Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(1).
- Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks Martin Rosvall, and Carl T. Bergstrom. *PNAS*, 104:7327.
- Roy, R., Hohng, S., and Ha, T. (2008). A practical guide to single-molecule FRET. *Nature methods*, 5(6):507–516.
- Rule, G. S. and Hitchens, T. K. (2006). Fundamentals of Protein NMR Spectroscopy. *Focus on structural biology*, 5:532.
- Rundqvist, L., Ådén, J., Sparrman, T., Wallgren, M., Olsson, U., and Wolf-Watz, M. (2009). Noncooperative folding of subdomains in adenylate kinase. *Biochemistry*, 48(9):1911–1927.

- Ruschak, A. M. and Kay, L. E. (2010). Methyl groups as probes of supra-molecular structure, dynamics and function. *Journal of Biomolecular NMR*, 46(1):75–87.
- Saffarian, S. and Elson, E. L. (2003). Statistical analysis of fluorescence correlation spectroscopy: the standard deviation and bias. *Biophysical journal*, 84(March):2030–2042.
- Sanchez, K. M., Kang, G., Wu, B., and Kim, J. E. (2011). Tryptophan-lipid interactions in membrane protein folding probed by ultraviolet resonance Raman and fluorescence spectroscopy. *Biophysical Journal*, 100(9):2121–2130.
- Schaub, M., Lehmann, J., Yaliraki, S., and Barahona, M. (2014). Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution. *Network Science*, 2(2014):1–24.
- Schaub, M. T., Delvenne, J. C., Yaliraki, S. N., and Barahona, M. (2012). Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. *PLoS ONE*, 7(2).
- Scheres, S. H. W., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P. B., Frank, J., and Carazo, J.-M. (2007). Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1):27–29.
- Schlauderer, G. J. and Schulz, G. E. (1996). The structure of bovine mitochondrial adenylate kinase: comparison with isoenzymes in other compartments. *Protein science : a publication of the Protein Society*, 5(3):434–41.
- Schuler, B. (2013). Single-molecule FRET of protein structure and dynamics - a primer. *Journal of nanobiotechnology*, 11 Suppl 1(Suppl 1):S2.
- Schuler, B., Lipman, E. A., and Eaton, W. A. (2002). Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747.
- Schüttelkopf, A. W. and Van Aalten, D. M. F. (2004). PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*, 60(8):1355–1363.
- Schwille, P. and Haustein, E. (2009). Fluorescence Correlation Spectroscopy: An Introduction to its Concepts and Applications. *Spectroscopy*, 94:1–33.
- Seo, M.-H., Park, J., Kim, E., Hohng, S., and Kim, H.-S. (2014). Protein conformational dynamics dictate the binding affinity for a ligand. *Nature Communications*, 5.
- Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6620–5.

- Shapiro, Y. E. and Meirovitch, E. (2006). Activation energy of catalysis-related domain motion in E. coli adenylate kinase. *Journal of Physical Chemistry B*, 110(23):11519–11524.
- Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. a., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science (New York, N.Y.)*, 330(6002):341–346.
- Sinev, M. A., Sineva, E. V., Ittah, V., and Haas, E. (1996). Domain closure in adenylate kinase. *Biochemistry*, 35(20):6425–6437.
- Smyth, S. and White, S. (2005). A spectral clustering approach to finding communities in graphs. *Proceedings of the 5th SIAM International Conference on Data Mining*, pages 76–84.
- Song, H. D. and Zhu, F. (2013). Conformational Dynamics of a Ligand-Free Adenylate Kinase. *PLoS ONE*, 8(7).
- Sorenson, J. M., Hura, G., Soper, a. K., Pertsemlidis, a., and Head-Gordon, T. (1999). Determining the role of hydration forces in protein folding. *Journal of Physical Chemistry B*, 103(26):5413–5426.
- Stryer, L. (1978). Fluorescence Energy Transfer as a Spectroscopic Ruler. *Annual Review of Biochemistry*, 47(1):819–846.
- Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69.
- Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151.
- Sullivan, S. M. and Holyoak, T. (2008). Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):13829–34.
- Swint-Kruse, L. (2004). Using networks to identify fine structural differences between functionally distinct protein states. *Biochemistry*, 43(34):10886–10895.
- Tawfik, S. D. and Khersonsky, O. (2010). Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1):471–505.
- Thompson, N. L. and Palmer, A. G. (2001). Molecular aggregation characterized by high order autocorrelation in fluorescence correlation spectroscopy. *Springer Series in Chemical Physics*, 65:438–458.

- Thorpe, M. (2009). *FIRST Users Guide*. Arizona State University.
- Unruh, J. (2017). Fluorescence Correlation Spectroscopy (FCS). <http://research.stowers-institute.org/microscopy/external/Technology/FCS/index.htm>.
- Uversky, V. N., Gillespie, J. R., Millett, I. S., Khodyakova, A. V., Vasiliev, A. M., Chernovskaya, T. V., Vasilenko, R. N., Kozlovskaya, G. D., Dolgikh, D. A., Fink, A. L., Doniach, S., and Abramov, V. M. (1999). Natively unfolded human prothymosin α adopts partially folded collapsed conformation at acidic pH. *Biochemistry*, 38(45):15009–15016.
- Valadié, H., Lacapčre, J., Sanejouand, Y.-H., and Etchebest, C. (2003). Dynamical Properties of the MscL of Escherichia coli: A Normal Mode Analysis. *J Mol Biol*, 332(3):657–674.
- Valeur, B. (2002). *Molecular Fluorescence: Principles and Applications*. Wiley-VCH Verlag GmbH & Co. KGaA.
- Velázquez-Muriel, J. A., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., and Carazo, J.-M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC structural biology*, 9:6.
- Vendruscolo, M., Dokholyan, N. V., Paci, E., and Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 65(6):4.
- Vijayabaskar, M. S. and Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophysical Journal*, 99(11):3704–3715.
- Wang, Y., Gan, L., Wang, E., and Wang, J. (2013). Exploring the dynamic functional landscape of adenylate kinase modulated by substrates. *Journal of Chemical Theory and Computation*, 9(1):84–95.
- Wassenaar, T. (2011). r.gyrate. https://pymolwiki.org/index.php/Radius_of_gyration.
- Whitford, P. C., Miyashita, O., Levy, Y., and Onuchic, J. N. (2007). Conformational Transitions of Adenylate Kinase: Switching by Cracking. *Journal of Molecular Biology*, 366(5):1661–1671.
- Whitford, P. C., Onuchic, J. N., and Wolynes, P. G. (2008). Energy landscape along an enzymatic reaction trajectory: hinges or cracks? *HFSP journal*, 2(2):61–64.
- Widengren, J., Rigler, R., and Mets, Ü. (1994). Triplet-state monitoring by fluorescence correlation spectroscopy. *Journal of Fluorescence*, 4(3):255–258.
- Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z., and Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature structural & molecular biology*, 11(10):945–9.

Word, J. and Richardson, D. (2012). Reduce Version 2.

Wu, F. and Huberman, B. A. (2004). Finding communities in linear time: A physics approach. In *European Physical Journal B*, volume 38, pages 331–338.

Zeller, F. and Zacharias, M. (2015). Substrate Binding Specifically Modulates Domain Arrangements in Adenylate Kinase. *Biophysical Journal*, 109(9):1978–1985.

Appendices

A. Graph construction and parameters

In contrast to the majority of network approaches to protein dynamics, our network is constructed by assigning edges that correspond to covalent and non-covalent interactions. The weight of each edge is determined by the strength of the physicochemical interaction between two atoms. Covalent bond strengths are obtained from tables that assuming standard bond lengths, whilst non-covalent interaction energies are obtained using a variety of approaches outlined here. We include: hydrophobic interactions, hydrogen bonds, electrostatic interactions and π -stacking.

A.1. Covalent bonds

Covalent bonds are identified using FIRST (Thorpe, 2009) using the distances between atoms. The weights are assigned according to a bond energy table¹ (Table A.1).

Table A.1: The energies used to weight covalent interactions in the graph representation of a protein. These values were taken from Huheey et al. (1993)

Bond	Energy (kJ/mol)	Bond	Energy (kJ/mol)	Bond	Energy (kJ/mol)
C — C	346	C = O	799	H — O	459
C = C	602	C — S	272	H — S	363
C — N	305	H — H	432	H — Se	276
C = N	615	H — C	411	P — O	335
C — P	264	H — N	386	P = O	544
C — O	358	H — P	322		

A.2. Hydrogen bonds

The hydrogen bonds strengths are calculated using the same formula implemented by FIRST (Jacobs et al., 2001):

¹http://www.wiredchemist.com/chemistry/data/bond_energies_lengths.html (Huheey et al., 1993)

$$E_{HB} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{10} - 6 \left(\frac{R_0}{R} \right)^{12} \right\} F(\theta, \phi, \varphi) \quad (\text{A.1})$$

It employs a potential proposed by Mayo *et al* (Dahiyat *et al.*, 1997), where $V_0 = 8\text{kcal/mol}$, $R_0 = 2.80\text{\AA}$ is the donor acceptor equilibrium distance, and R is their actual distance. See Figure A.1 for an intuitive schematic. $F(\theta, \phi, \varphi)$ is a function of the bond angles between the donor, acceptor and hydrogen atoms. It is dependent on the bond type:

- sp^3 donor - sp^3 acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\phi - 109.5)$
- sp^3 donor - sp^3 acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\phi)$
- sp^2 donor - sp^3 acceptor: $F = \cos^4\theta e^{-2(\pi-\theta)^6}$
- sp^2 donor - sp^2 acceptor: $F = \cos^2\theta e^{-(\pi-\theta)^6} \cos^2(\max[\phi, \psi])$

where ϕ is the hydrogen-acceptor-base angle, ψ is the angle between the normal of the planes that are given by the atoms at the sp^2 centres, and θ is the donor-hydrogen-acceptor angle.

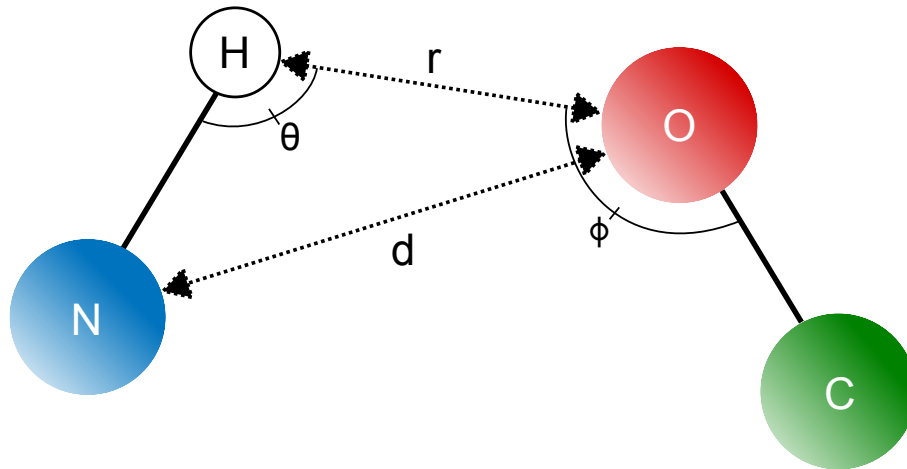


Figure A.1: Variables used to calculate hydrogen bond strength.

A.3. Electrostatic interactions

The electrostatic interactions between ligands and proteins are identified using the LINK entries in the proteins PDB file. Using the relative distance, r , and the charge of each atom (q_1, q_2) we can determine the Coulomb potential

$$E_c = \frac{332}{\epsilon} \frac{q_1 q_2}{r} \quad (\text{A.2})$$

where ϵ is the dielectric constant and set to $\epsilon = 4$ according to [Gilson and Honig \(1986\)](#). The atoms charges for standard residues are calculated from the OPLS-AA forcefield [Jorgensen and Tirado-rives \(1988\)](#), whilst the charges for ligands and non-standard residues (i.e. phosphorylated residues) are found using the PRODRG web-server ([Schüttelkopf and Van Aalten, 2004](#)).

A.4. Hydrophobic interactions

Hydrophobic tethers are also identified using FIRST ([Thorpe, 2009](#)), however, the energy is not assigned. Hydrophobic interactions are assigned between C-C or C-S pairs based on their relative proximity to each other: two atoms will be tethered if their Van der Waal's radii are within $2\text{\AA}$. The interaction energy is then determined based on a double well potential of mean force ([Lin et al., 2007](#)) which gives an energy of -0.8 kcal/mol for atoms within $2\text{\AA}$.

B. DNA Sequencing results

B.1. Wildtype (FP1)

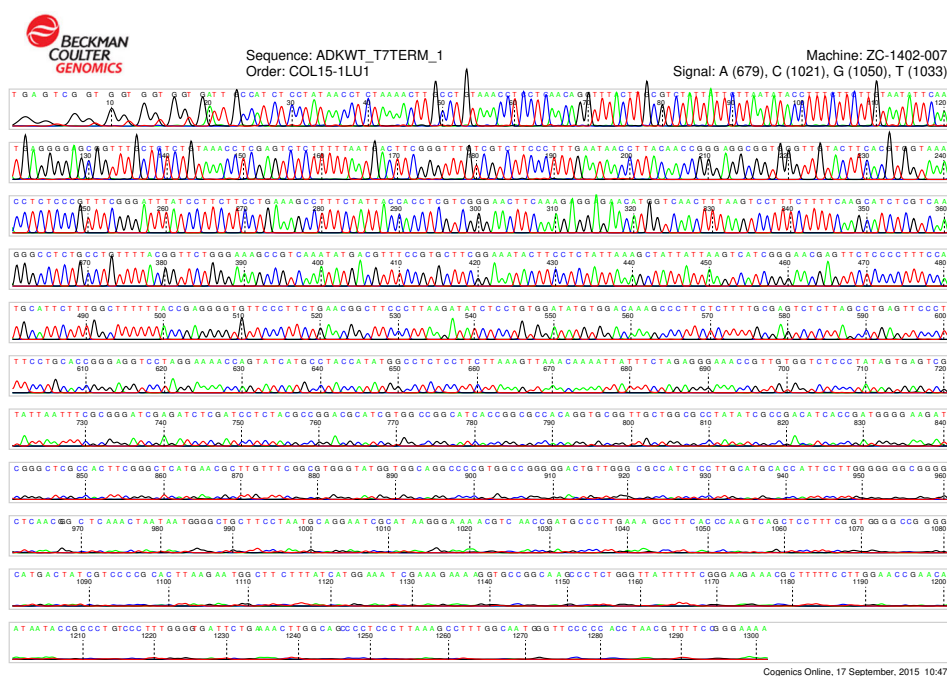


Figure B.1: DNA sequencing results of FP1 from GENEWIZ.

```

Met V G Met I L V F L G P P G A G K G T Q A K R L A K E K G F V H I S T G D I L R E A
V Q K G T P L G K K A K E C Met E R G E L V P D D L I I A L I E E V F P K H G N V
I F D G F P R T V K Q A E A L D E Met L E K K G L K V D H V L L F E V P D E V V I
E R L S G R R I N P E T G E V Y H V K Y N P P P G C K V I Q R E D D K P E V I K
K R L E V Y R E Q T A P L I E Y Y K K K G I L R I I D A S K P V E E V Y R Q V L E
V I G D G N H H
  
```

Met V G Met I L V F L G P P G A G K G T Q A K R L A K E K G F V H I S T G D I L R E A
V Q K G T P L G K K A K E Y Met E R G E L V P D D L I I A L I E E V F P K H G N V
I F D G F P R T V K Q A E A L D E Met L E K K G L K V D H V L L F E V P D E V V I
E R L S G R R I N P E T G E V Y H V K Y N P P P P G C K V I Q R E D D K P E V I K
K R L E V Y R E Q T A P L I E Y Y K K K G I L R I I D A S K P V E E V Y R Q V C E
V I G D G N H H H H H H

B.3. L199C & C145V mutant (FP3)

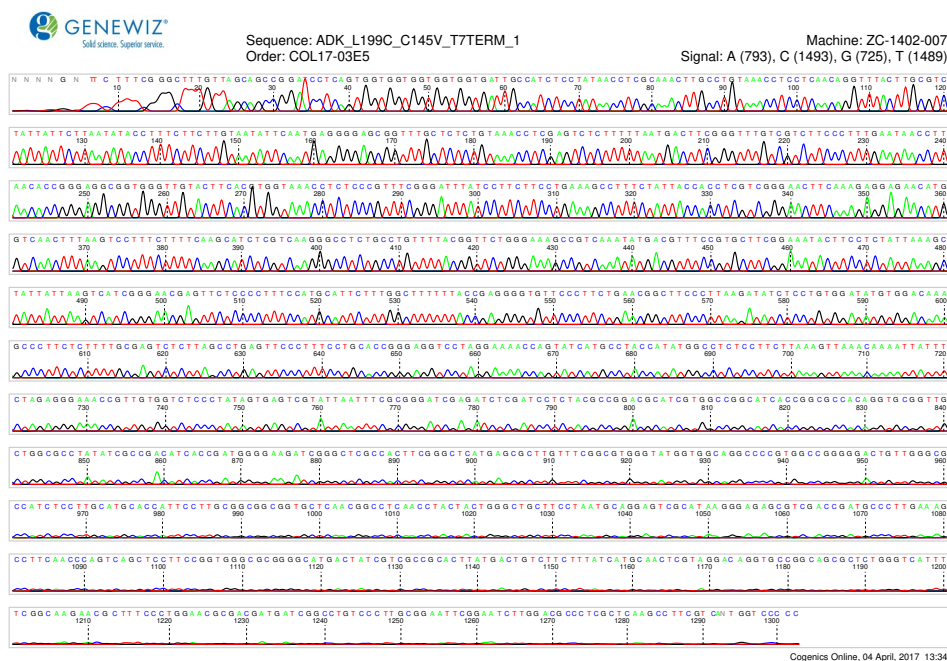


Figure B.3: DNA sequencing results of FP3 from GENEWIZ.

Met V G Met I L V F L G P P G A G K G T Q A K R L A K E K G F V H I S T G D I L R E A
V Q K G T P L G K K A K E C Met E R G E L V P D D L I I A L I E E V F P K H G N V
I F D G F P R T V K Q A E A L D E Met L E K K G L K V D H V L L F E V P D E V V I
E R L S G R R I N P E T G E V Y H V K Y N P P P P G V K V I Q R E D D K P E V I K
K R L E V Y R E Q T A P L I E Y Y K K K G I L R I I D A S K P V E E V Y R Q V C E
V I G D G N H H H H H H H

C. FCS measurements

To determine the rate that corresponded to motion of the lid subdomains in ADK we had to accurately define the triplet dynamics and diffusion rate respectively. This required using single-donor labelled ADK species such that no FRET contributed to the FCS correlation curve.

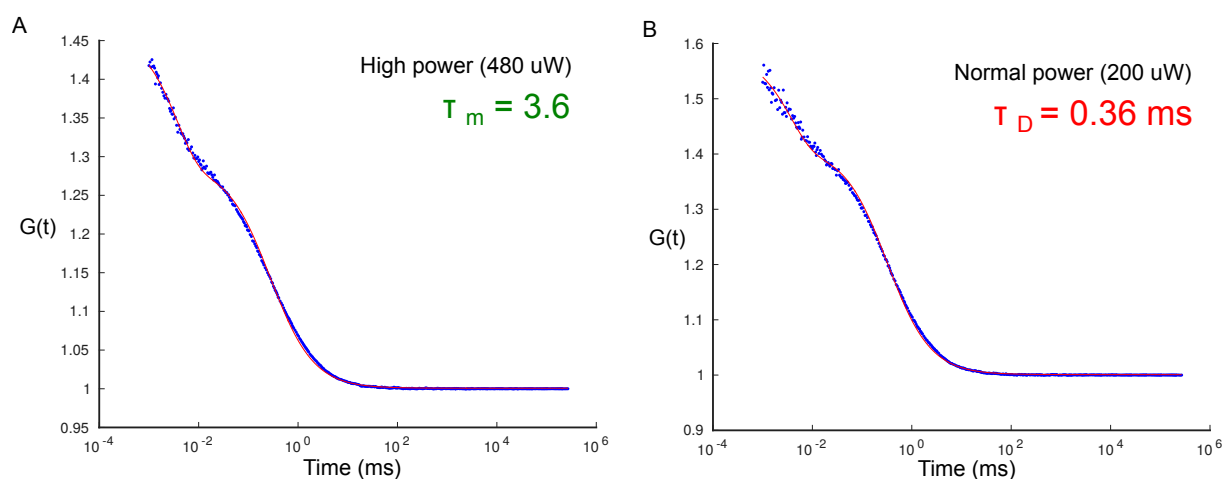


Figure C.1: Determining the triplet relaxation rate and the diffusion rates for ADK. (a) A high power measurement increases the amplitude of the triplet relaxation component of the autocorrelation curve, as such we can accurately measure the triplet rate as $3.6 \mu\text{s}$. (b) Using the fixed triplet rate we can now measure the diffusion rate accurately at low power as 0.36 ms.

D. Gaussian Process Regression

To measure the extent to which a mutation perturbed the community structure of a protein we performed a full computational mutagenesis. Computational mutations altered the network structure associated with the protein through removal of side-chain weak interactions. Markov Stability was computed for each mutant structure and it was clear that different mutations affected different Markov timescales. We can use Gaussian process regression to identify important mutations at each relevant scale.

Each Markov Stability computation produces a $\overline{\text{VI}}$ ‘trajectory’ that describes the robustness of the community structure at each Markov time. Using Gaussian process regression (GPR) we can produce a mean $\overline{\text{VI}}$ trajectory that is representative of (and constrained by) the ensemble of mutant trajectories. The associated variance of the representative trajectory is used as a statistical measure to identify mutant trajectories that deviate significantly from the ensemble at that timescale.

In the following section we describe GPR with a simple test set.

D.1. Using GPR on a simple test set

Gaussian process regression (GPR) models are non-parametric kernel-based probabilistic models. If we consider the data in figure D.1 we could attempt to fit a sinusoidal function of the form $f(x) = x * \sin(x)$, however, there is no obvious relationship between the dependent and independent variables given the low number of data points. GPR avoids making assumptions regarding the relationship between the variables making it a highly flexible form of regression.

A Gaussian process (GP) is an infinite dimensional generalisation of a univariate (standard 1-dimensional) Gaussian distribution. Suppose we generate data using a GP over a certain domain. Any subset of points within the domain will also follow a multivariate Gaussian distribution. Therefore, we can consider a data set of 25 points $\{y_1, \dots, y_{25}\}$ as a single subset of a 25-dimensional Gaussian distribution. How do we use this approach to estimate y^* for a new test point x^* ?

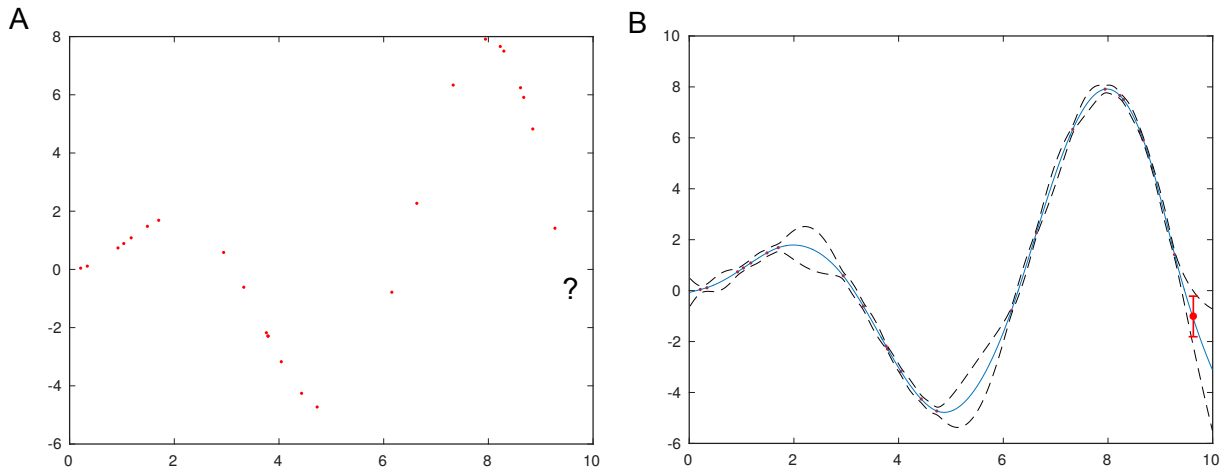


Figure D.1: A Gaussian process regression example with 25 training-set points. (a) The training set points are plotted in red. We want to estimate the value of a new output y^* given a new test-set point x^* . (b) The expectation of $\bar{y}^*$ at test point x^* is given by the red vertical error bars. The blue line is the mean of 1000 test set data points randomly selected between 0 and 10. The dashed lines correspond to a 2 standard deviation interval of the predicted test set using the GPR model trained on the training set.

A multivariate Gaussian (specifically a GP in this case) is specified by a mean (vector) and covariance function (diagonal matrix). The mean function is generally taken to be 0, and the covariance matrix is often chosen as the *squared exponential function*

$$k(x, x') = \sigma_f^2 \exp\left(\frac{-(x - x')^2}{2l^2}\right), \quad (\text{D.1})$$

where l and σ_f are hyperparameters that control the length scale and the maximum covariance respectively. The value of $k(x, x')$ will be higher for points that are closer together. The hyperparameters must be estimated, for this test set we choose $\sigma_f = 1$ and $l = 1.5$.

We can combine the 25 ‘training-set’ points into a single vector y with an unknown ‘test-set’ point y^* to produce a 26-dimensional vector. Under a GP we assume that this new vector has a multivariate Gaussian distribution

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right). \quad (\text{D.2})$$

The covariance matrices K and K_* are obtained through evaluation of the covariance function $k(x, x')$ (Equation D.1) for all possible pairs of x values

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_{25}) \\ & \ddots & \vdots \\ k(x_{25}, x_1) & \cdots & k(x_{25}, x_{25}) \end{bmatrix}, \quad (\text{D.3})$$

and

$$K_* = [k(x^*, x_1) \cdots k(x^*, x_{25})], \quad (\text{D.4})$$

and

$$K_{**} = k(x^*, x^*). \quad (\text{D.5})$$

We are interested in the probability of the new test output given the test data i.e. $p(y^*, y)$. This is given by

$$y^* | y = \mathcal{N}(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*^T) \quad (\text{D.6})$$

for a multivariate Gaussian distribution. Therefore, the expected value and variance of the new test output y^* is

$$\bar{y}^* = K_* K^{-1} y \quad (\text{D.7})$$

$$\text{Var}(y^*) = K_{xx} = K_* K^{-1} K_*^T. \quad (\text{D.8})$$

Figure D.1b exhibits the new point $\bar{y}^*$ and its corresponding variance exhibited as error bars. Carrying out the same procedure with a large test-set of points (1000 points) distributed across the x-axis we are able to plot a 2 standard deviation interval corresponding to the probability distribution of these points.

Until this point we assumed that the hyperparameters $\theta = \{f, l\}$ were known *a priori*. In reality, this isn't the case and we must estimate them from the data. We want hyperparameters θ that maximise $p(\theta | x, y)$ which is equivalent to maximising $\log p(y | x, \theta)$ according to Baye's theorem. Continuing the assume a Gaussian distribution with 0 mean and covariance matrix K , we can calculate this by

$$\log p(y | x, \theta) = -\frac{1}{2}y^T K^{-1}y - \frac{1}{2}\log |K| - \frac{n}{2}\log(2\pi) \quad (\text{D.9})$$

and use standard optimisation techniques such as gradient descent to maximise this value. Once the hyperparameters are optimised we can use them to calculate the mean output and its associated variances for any new set of test data.

D.2. Using GPR on the ensemble of mutated VI trajectories

Each Markov Stability calculation produces a $\overline{\text{VI}}$ trajectory $\mathbf{v}_i = [v_1, v_2, \dots, v_u]$ at Markov times $\mathbf{t}_i = [t_1, t_2, \dots, t_u]$. A complete mutagenesis analysis of an N residue protein gives $N + 1$ $\overline{\text{VI}}$ trajectories (including the wildtype structure). The training-set $(\mathbf{v}, \mathbf{t})$ is produced by appending the trajectories, $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{N+1}]$ and $\mathbf{t} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{N+1}]$. Equation D.9 can be used to infer and optimise the hyperparameters, this is completed using the GPML Matlab toolbox¹. Afterwhich we can evaluate the mean and variance, using Equations D.7 and D.8, that represents the ensemble of $\overline{\text{VI}}$ trajectories.

¹Available from <http://www.Gaussianprocess.org/gpml/code/> (Rasmussen and Williams, 2004)

E. Kullback-Leibler divergence

To exhibit the robustness of the combined scoring methodology of VI and CV we used Kullback-Leibler (KL) to combine the difference between two normal distributions. Here we exhibit workigs required to determine the KL-divergence between two Gaussians given the generalised form.

Letting $p(x) = \mathcal{N}(\mu_1, \sigma_1)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2)$ we can use the definition of KL by [Bishop \(2006\)](#),

$$KL(p, q) = - \int p(x) \log q(x) dx + \int p(x) \log p(x) dx. \quad (\text{E.1})$$

where

$$\int p(x) \log p(x) dx = \frac{1}{2} (1 + \log 2\pi\sigma_1^2). \quad (\text{E.2})$$

Just restricting myself to $\int p(x) \log q(x) dx$, we can rewrite this as

$$- \int p(x) \log \frac{1}{(2\pi\sigma_2^2)^{1/2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx. \quad (\text{E.3})$$

This can be separated into

$$\frac{1}{2} \log(2\pi\sigma_2^2) - \int p(x) \log e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx \quad (\text{E.4})$$

Taking the log, separating the sums and removing σ_2^2 from the integral we get

$$\frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\int p(x)x^2 dx - \int p(x)2x\mu dx + \int p(x)\mu^2 dx}{2\sigma_2^2} \quad (\text{E.5})$$

If we denote the expectation operator under p as $\langle \rangle$ then we can rewrite equation E.5 as

$$\frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\langle x^2 \rangle - 2\langle x \rangle \mu_2 + \mu_2^2}{2\sigma_2^2} \quad (\text{E.6})$$

Knowing that the variance of x is $\text{var}(x) = \langle x^2 \rangle - \langle x \rangle^2$, we can define

$$\langle x^2 \rangle = \sigma_1^2 + \mu_1^2. \quad (\text{E.7})$$

Inputting equation E.7 into equation E.6 and rearranging we find

$$\frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}. \quad (\text{E.8})$$

This can finally be added to equation E.1 to produce

$$\begin{aligned} KL(p, q) &= \frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} (1 + \log 2\pi\sigma_1^2) \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \end{aligned} \quad (\text{E.9})$$

This now provides us with the KL-divergence between two Gaussian distributions. Appropriately, the divergence for $KL(p, p)$ equals zero in this case since there is no difference between two equal distributions.

F. 2-way community structure of D152 mutations

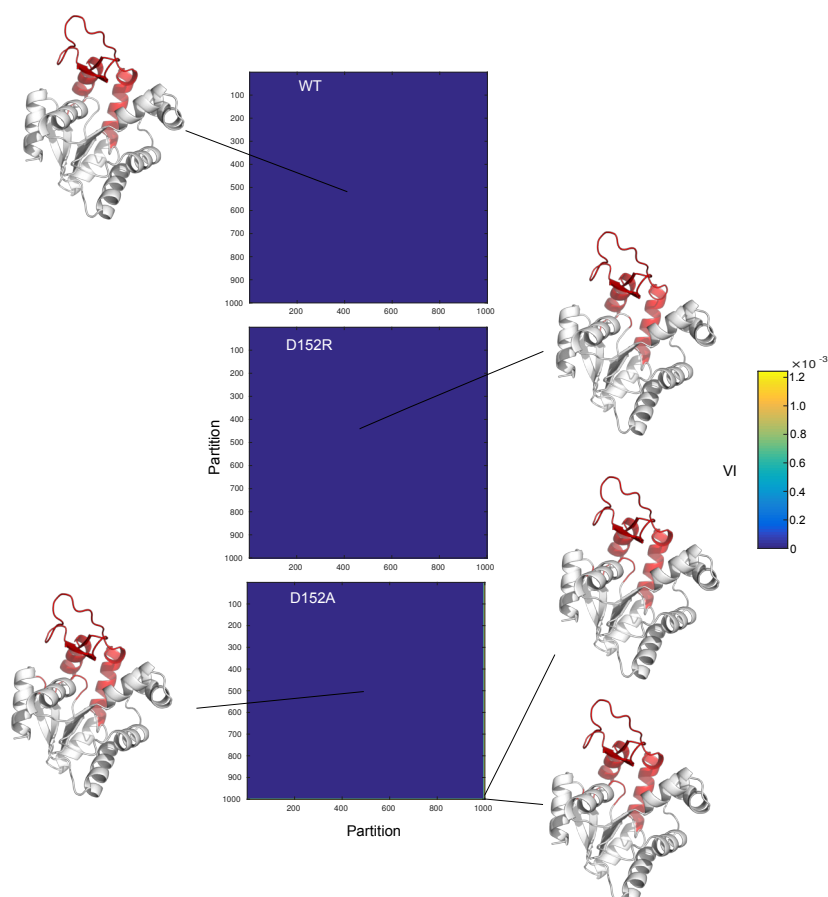


Figure F.1: The variation between the optimal partitions of the 2-way community structures upon mutation. A comparison of the 1000 Louvain partitions for suggests that the same partition is found every time. Interestingly, the optimal partitions for the wildtype, D152R and D152A are different. The binding loop located in the core subdomain of D152R and D152A begins to join the community of the lid subdomain. Mutation has produced an increased closeness between the lid and core subdomains, owing itself towards a closed conformation.

G. Multidimensional scaling

One of the most important goals in visualising data is to provide a sense of distance (how near or far) between points. When the data lies in a 2-dimensional plane with a set of coordinates (x,y) it is easy to visualise using a 2D scatter plot, however, not all data is this simple. Instead, the data may be in the form of pairwise similarities or dissimilarities between observations or cases, where there are no points to plot. The easiest way to visualise the ‘distance’ between to points might not be their Euclidean distance. For example within a city block setting the absolute distance (as the crow flies) is shorter than the actual distance travelled to get from one block to another in a diagonal manner. Moreover, with a large number of variables (a multidimensional data set) it is very difficult to visualise the data unless it can be represented in a smaller number of dimensions (dimension reduction).

G.1. Details of MDS

Multidimensional scaling (MDS) is a set of methods that address these problems, allowing us to visualise the distances between multidimensional data. Here we will address non-metric multidimensional scaling (a subset of classical MDS), since classical MDS is not applicable for dissimilarity ratings ($\overline{VI}$). Given a dissimilarity matrix

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix} \quad (\text{G.1})$$

we want to find I vectors $\mathbf{x}_1, \dots, \mathbf{x}_I \in \mathbb{R}^2$ such that

$$\|x_i - x_j\| \approx \delta_{i,j} \quad (\text{G.2})$$

which gives the vector norm to provide a distance. This means that MDS attempts to find an

embedding from I objects (or points) into 2 dimensional space such that distances are preserved. The vectors $\mathbf{x}_i$ are not unique and recalculation can produce arbitrary transformations, rotations and reflections since they do not affect the pairwise distance $\|x_i - x_j\|$.

To determine the vectors $\mathbf{x}_i$ there are a variety of approaches, the most common is to use an optimisation formulate and the vectors are found as a minimiser of a cost function i.e. let x be the vector of proximities, $f(x)$ denotes a monotonic transformation of x and d the point distances. The coordinates are found that minimise the stress,

$$stress = \sqrt{\frac{\sum (f(x) - d)^2}{\sum d^2}}. \quad (\text{G.3})$$

Two optimisation processes must be applied for non-metric MDS. Firstly we must optimise the monotonic transformation of proximities, then the points of the configuration must be optimally arranged. Such that the distances between points match the scaled proximities as accurately as possible.

G.2. Using MDS on the ensemble of Louvain partitions

At each Markov time point we calculate N Louvain optimisations that attempt to optimise the community structure. The dissimilarity between the community structures is measured using the VI between each Louvain optimisation, producing an $N \times N$ dissimilarity matrix. The stress function (Equation G.3) is implemented and minimised using the *mdscale* code available in Matlab. Afterwhich, we can plot a 2-dimensional representation of the multidimensional dissimilarity matrix.

H. Conservation scores of Adenylate Kinase

Our analysis of Adenylate Kinase also used the evolutionary conservation scores of each residue. This helped us determine critical catalytic residues but also allowed to identify potential thermophilic residues. Conservation scores were calculated for *Aquifex* Adenylate Kinase using PDB ID: 2RGX using the ConSurf server¹.

- **POS:** The position of the AA in the SEQRES derived sequence.
- **SEQ:** The SEQRES derived sequence in one letter code.
- **3LATOM:** The ATOM derived sequence in three letter code, including the AA's positions as they appear in the PDB file and the chain identifier.
- **SCORE:** The normalized conservation scores.
- **COLOR:** The color scale representing the conservation scores (9 - conserved, 1 - variable).
- **CONFIDENCE INTERVAL:** When using the bayesian method for calculating rates, a confidence interval is assigned to each of the inferred evolutionary conservation scores.
- **CONFIDENCE INTERVAL COLORS:** When using the bayesian method for calculating rates. The color scale representing the lower and upper bounds of the confidence interval.
- **MSA DATA:** The number of aligned sequences having an amino acid (non-gapped) from the overall number of sequences at each position.
- **RESIDUE VARIETY:** The residues variety at each position of the multiple sequence alignment.

¹Available from http://bental.tau.ac.il/new_ConSurfDB/ (Ashkenazy et al., 2010)

POS	SEQ	3LATOM	SCORE (normalized)	COLOR	CONFIDENCE	INTERVAL	CONFIDENCE	INTERVAL	COLORS	MSA DATA	RESIDUE VARIETY
1	M	MET1:A	-1.039	9	-1.084	-1.027	9,9		84/276	M, I	
2	I	ILE2:A	-0.839	8	-0.928	-0.783	9,8		85/276	F, H, Q, N, I, V	
3	L	LEU3:A	0.127	4	-0.069	0.287	5,4		273/276	F, A, M, I, Y, L, V	
4	V	VAL4:A	-0.460	7	-0.578	-0.364	7,7		276/276	M, I, L, V	
5	F	PHE5:A	-0.510	7	-0.636	-0.443	8,7		276/276	F, M, I, L	
6	L	LEU6:A	-0.473	7	-0.636	-0.364	8,7		276/276	F, M, I, L, V	
7	G	GLY7:A	-1.077	9	-1.092	-1.074	9,9		276/276	G	
8	P	PRO8:A	-0.746	8	-0.825	-0.690	8,8		276/276	S, A, P, G, L	
9	P	PRO9:A	-1.079	9	-1.092	-1.074	9,9		276/276	P	
10	G	GLY10:A	-1.077	9	-1.092	-1.074	9,9		276/276	G	
11	A	ALA11:A	-0.839	8	-0.897	-0.783	9,8		276/276	S, A, T, C, I, G, V	
12	G	GLY12:A	-1.077	9	-1.092	-1.074	9,9		276/276	G	
13	K	LYS13:A	-1.081	9	-1.092	-1.074	9,9		276/276	K	
14	G	GLY14:A	-1.055	9	-1.084	-1.046	9,9		276/276	S, G	
15	T	THR15:A	-1.073	9	-1.090	-1.061	9,9		276/276	S, T	
16	Q	GLN16:A	-0.975	9	-1.027	-0.957	9,9		276/276	H, Q, I, L, Y	
17	A	ALA17:A	-0.905	9	-0.957	-0.862	9,9		276/276	S, A, C, G	
18	K	LYS18:A	0.449	3	0.207	0.597	4,3		276/276	F, S, A, T, N, P, K, E, V, H, Q, D, R, G, L	
19	R	ARG19:A	1.343	1	0.872	1.250	1,1		276/276	F, S, T, N, K, Y, E, V, Q, M, C, L, A, H, D, I, R, G	
20	L	LEU20:A	-0.371	7	-0.513	-0.277	7,6		276/276	F, M, I, L, V	
21	A	ALA21:A	0.647	2	0.382	0.872	3,1		276/276	S, A, T, N, K, E, V, Q, M, D, C, R, I, L	
22	K	LYS22:A	1.310	1	0.872	1.250	1,1		276/276	A, S, T, N, K, Y, E, H, Q, M, D, I, R, G	
23	E	GLU23:A	1.649	1	0.872	1.841	1,1		276/276	S, F, A, T, N, K, E, H, Q, M, D, I, R, L	
24	K	LYS24:A	1.145	1	0.597	1.250	3,1		276/276	F, T, N, K, E, Y, H, Q, R, G	
25	G	GLY25:A	1.204	1	-0.513	-0.277	1,1		276/276	S, A, S, N, P, K, E, V, H, Q, M, D, C, I, R, G	
26	F	PHE26:A	-0.526	7	-0.636	-0.443	8,7		276/276	S, A, F, T, V, M, C, I, L	
27	V	VAL27:A	-0.110	5	-0.277	-0.069	6,5		276/276	F, S, A, T, N, K, P, E, V, Q, C, R, I, L	
28	H	HIS28:A	-0.874	9	-0.957	-0.825	9,8		276/276	H, Q, R, K, Y	
29	I	ILE29:A	-0.920	9	-1.084	-0.862	9,9		276/276	M, I, L, V	
30	S	SER30:A	-0.964	9	-1.066	-0.928	9,9		276/276	A, S	
31	T	THR31:A	-1.050	9	-1.074	-1.046	9,9		276/276	S, T	
32	G	GLY32:A	-1.077	9	-1.092	-1.074	9,9		276/276	G	
33	D	ASP33:A	-0.920	9	-1.074	-0.862	9,9		276/276	N, D, E	
34	I	ILE34:A	-0.715	8	-0.825	-0.690	8,8		276/276	M, I, L	
35	L	LEU35:A	-0.756	8	-0.862	-0.690	9,8		276/276	F, M, I, L	
36	R	ARG36:A	-1.082	9	-1.092	-1.074	9,9		276/276	R	
37	E	GLU37:A	-0.261	6	-0.443	-0.179	7,6		276/276	F, S, A, T, N, K, Y, E, V, Q, D, R, G, L	
38	A	ALA38:A	-0.721	8	-0.825	-0.690	8,8		276/276	A, T, N, Y, E, H, Q, M, I, L	
39	V	VAL39:A	-0.437	7	-0.578	-0.364	7,7		276/276	F, A, K, V, M, I, R, L	
40	Q	GLN40:A	0.085	5	-0.179	0.287	6,4		276/276	A, S, Q, T, N, R, K, G, E	
41	K	LYS41:A	-0.720	2	0.382	0.872	3,1		276/276	A, A, N, K, V, Q, M, C, D, R, G	
42	G	GLY42:A	0.673	2	0.382	0.872	3,1		276/276	S, A, N, K, E, Q, M, C, D, R, G, L	
43	T	THR43:A	-0.719	8	-0.825	-0.690	8,8		276/276	S, T, D, N, K	
44	P	PRO44:A	1.019	1	0.597	1.250	3,1		275/276	S, A, T, K, P, E, V, Q, D, G	
45	L	LEU45:A	-0.683	8	-0.783	-0.636	8,8		276/276	F, T, M, I, L, V	
46	G	GLY46:A	-1.030	9	-1.074	-1.006	9,9		276/276	A, G, E	
47	K	LYS47:A	1.816	1	1.250	1.841	1,1		276/276	S, A, T, N, K, E, V, Q, M, D, I, R, G, L	
48	K	LYS48:A	0.810	2	0.382	0.872	3,1		276/276	S, A, T, N, K, E, V, Q, M, D, I, R, L	
49	A	ALA49:A	-0.780	8	-0.862	-0.739	9,8		276/276	S, A, N, V, M, C, I, L	
50	K	LYS50:A	-0.554	7	-0.690	-0.513	8,7		276/276	F, A, T, N, P, K, E, V, H, Q, D, I, R, G	
51	E	GLU51:A	0.901	1	0.597	0.872	3,1		276/276	A, S, T, N, P, K, E, V, Q, D, I, R, G	
52	Y	TYR52:A	0.130	4	-0.069	0.287	5,4		276/276	A, F, T, K, Y, V, H, I, L	
53	M	MET53:A	-0.689	8	-0.783	-0.636	8,8		276/276	T, M, I, L, V	
54	E	GLU54:A	-0.411	7	-0.578	-0.364	7,7		276/276	T, M, I, L, V	
55	R	ARG55:A	0.680	2	0.382	0.872	3,1		276/276	S, A, T, N, K, E, V, Q, M, D, R, G	
56	G	GLY56:A	-1.077	9	-1.092	-1.074	9,9		276/276	G	
57	E	GLU57:A	1.051	1	0.597	1.250	3,1		276/276	S, A, N, K, E, H, Q, D, R, G, L	
58	L	LEU58:A	-0.620	9	-1.074	-1.006	9,9		276/276	Y, L	
59	V	VAL59:A	-1.069	9	-1.046	-0.983	9,9		276/276	I, L, V	
60	P	PRO60:A	-0.723	8	-0.825	-0.690	8,8		276/276	S, T, N, R, P	
61	D	ASP61:A	-1.068	9	-1.090	-1.061	9,9		276/276	N, D	
62	D	ASP62:A	-0.381	6	0.059	0.382	5,3		276/276	A, S, T, N, K, E, Q, D, I, G	
63	L	LEU63:A	-0.141	6	-0.277	-0.069	6,5		276/276	F, A, T, M, I, L, V	
64	I	ILE64:A	-0.640	8	-0.739	-0.578	8,7		276/276	M, T, I, L, V	
65	I	ILE65:A	-0.385	7	-0.513	-0.277	7,6		276/276	S, M, T, N, V, M, C, I, R, L	
66	A	ALA66:A	-0.553	7	0.207	0.597	4,3		276/276	S, A, T, N, K, E, V, H, Q, M, D, I, R, G	
67	L	LEU67:A	-0.361	7	-0.513	-0.277	7,6		276/276	T, M, I, L, V	
68	I	ILE68:A	-0.687	8	-0.783	-0.636	8,8		276/276	A, M, D, I, L, V	
69	E	GLU69:A	0.288	4	0.059	0.382	5,3		276/276	S, A, M, T, N, K, E, Q, M, D, I, R, G, L	
70	G	GLY70:A	0.025	5	-0.179	0.059	6,5		276/276	A, S, T, N, K, E, V, H, Q, M, D, I, R, G	
71	V	VAL71:A	-0.683	8	-0.783	-0.636	8,8		276/276	F, A, T, N, K, E, Y, V, H, Q, C, R, I	
72	F	PHE72:A	-0.243	6	-0.443	-0.179	7,6		276/276	F, M, T, Y, V, M, I, L	
73	P	PRO73:A	1.152	1	0.597	1.250	3,1		276/276	A, S, T, N, K, P, E, V, H, Q, M, D, R, G, L	
74	K	LYS74:A	-1.190	1	0.872	1.250	1,1		275/276	S, A, T, N, K, E, V, H, Q, M, D, C, R, G, L	
75	H	HIS75:A	-0.541	7	-0.690	-0.513	8,7		266/276	A, S, T, N, P, K, E, V, H, Q, D, I, G, L	
76	G	GLY76:A	1.170	1	0.597	1.250	3,1		273/276	A, S, T, N, P, K, Y, E, V, H, Q, C, D, R, G	
77	N	ASN77:A	-0.766	8	-0.862	-0.690	9,8		276/276	A, S, N, K, Q, D, R, G	
78	V	VAL78:A	-0.354	7	-0.513	-0.277	7,6		276/276	A, F, M, C, I, L, Y, V	
79	I	ILE79:A	-0.616	8	-0.739	-0.578	8,7		276/276	M, I, L, V	
80	F	PHE80:A	-0.689	8	-0.783	-0.636	8,8		276/276	F, M, I, L	
81	D	ASP81:A	-1.082	9	-1.092	-1.084	9,9		276/276	D	
82	G	GLY82:A	-1.077	9	-1.092	-1.074	9,9		276/276	F, I, Y, V	
83	F	PHE83:A	-0.783	7	-0.825	-0.636	8,8		276/276	P	
84	P	PRO84:A	-1.079	9	-1.092	-1.074	9,9		276/276	R	
85	R	ARG85:A	-1.082	9	-1.092	-1.074	9,9		276/276	S, T, N	
86	T	THR86:A	-0.942	9	-1.006	-0.928	9,9		276/276	F, A, S, T, K, P, E, V, Q, D, C, I, L	
87	V	VAL87:A	-0.219	7	-0.364	-0.179	7,6		276/276	F, S, A, T, N, P, K, Y, E, V, H, Q, D, R, G, L	
88	K	LYS88:A	0.830	2	0.382	0.872	3,1		276/276	Q	
89	Q	GLN89:A	-1.082	9	-1.092	-1.084	9,9		276/276	A, C, G, V	
90	A	ALA90:A	-1.034	9	-1.061	-1.027	9,9		276/276	A, C, V	
91	E	GLU91:A	-0.146	7	-0.069	0.287	5,4		276/276	F, A, S, T, N, K, E, V, H, Q, D, I, R, G, L	
92	A	ALA92:A	-0.527	7	-0.636	-0.443	8,7		276/276	S, F, A, W, T, K, Y, E, V, H, Q, M, I, R, G, L	
93	L	LEU93:A	-0.985	9	-1.046	-0.957	9,9		180/276	F, T, I, L	
94	D	ASP94:A	-0.261	6	-0.443	-0.179	7,6		180/276	A, S, T, N, K, E, H, Q, D, I, G	
95	E	GLU95:A	2.615	1	1.250	2.467	1,1		180/276	A, S, T, N, K, E, V, H, Q, M, D, I, R, G	
96	M	MET96:A	0.331	4	0.059	0.382	5,3		180/276	S, A, F, T, W, N, K, Y, V, M, I, G, L	
97	L	LEU97:A	-0.652	8	-0.783	-0.578	8,7		276/276	F, M, T, I, L, V	
98	E	GLU98:A	2.815	1	1.250	3.467	1,1		276/276	F, S, T, N, K, Y, E, V, Q, M, L, A, P, H, D, R, I, G	
99	K	LYS99:A	1.862	1	1.250	1.841	1,1		276/276	F, S, A, T, N, K, E, V, H, Q, M, D, I, R, G	
100	K	LYS100:A	1.064	1	0.597	1.250	3,1		274/276	S, A, T, N, K, E, Y, V, H, Q, M, D, R, I, G, L	
101	G	GLY101:A	1.016	1	0.597	1.250	3,1		273/276	A, S, T, N, K, E, H, Q, C, D, R, G	
102	L	LEU102:A	0.762	2	0.382	0.872	3,1		273/276	A, S, T, N, K, E, V, H, Q, M, I, R, L	
103	K	LYS103:A	2.467	1	1.841	2.467	1,1		274/276	F, S, T, N, K, E, V, Q, L, A, P, H, D, R, I, G	
104	V	VAL104:A	-0.030	5	-0.277	0.059	6,5		275/276	T, M, D, I, L, V	
105	D	ASP105:A	-0.469	7	-0.578	-0.364	7,7		275/276	S, T, N, K, Y, E, Q, D, G	
106	H	HIS106:A	0.930	1	0.597	1.250	3,1		276/276	F, S, A, T, N, K, Y, E, V, H, M, C, R, G, L	
107	V	VAL107:A	-0.799	9	-0.862	-0.739	9,9		276/276	A, T, C, I, L, V	
108	L	LEU108:A	0.063	5	-0.179	0.287	6,4		276/276	F, M, P, I, L, V	
109	L	LEU109:A	-0.336	6	-0.513	-0.277	7,6		276/276	A, F, S, N, E, Y, H, Q, C, D, L	
110	F	PHE110:A	-0.424	7	-0.578	-0.364	7,7		276/276	F, M, I, Y, L, V	
111	I	ILE111:A	-0.065	7	-0.597	-0.250	3,1		276/276	F, S, T, N, K, Y, E, V, Q, C, L, A, H, D, R, G	
112	V	VAL112:A	-0.847	9	-0.928	-0.825	9,8		276/276	S, A, F, T, L, V	
113	P	PRO113:A	0.702	2	0.382	0.872	3,1		276/276	F, A, S, T, N, K, P, E, Q, D, R, G	
114	D	ASP114:A	0.094	5	-0.179	0.287	6,4		276/276	S, A, F, T, N, P, K, Y, E, Q, M, D, R, G, L	
115	E	GLU115:A	0.938	1	0.597	1.250	3,1		276/276	A, S, N, P, K, E, V, Q, D, R, G	
116	V	VAL116:A	0.716	2	0.382	0.872	3,1				

157	V	VAL157:A	-0.591	7	-0.690, -0.513	8,7	276/276	A, S, T, N, I, K, V
158	I	ILE158:A	-0.731	8	-0.825, -0.690	8,8	276/276	F, A, M, I, L, V
159	K	LYS159:A	0.787	2	0.382, 0.872	3,1	276/276	A, S, T, N, K, E, V, H, Q, M, C, I, R, G, L
160	K	LYS160:A	0.539	3	0.207, 0.597	4,3	276/276	A, S, T, N, K, E, V, H, Q, D, R, I, L
161	R	ARG161:A	-1.082	9	-1.092, -1.074	9,9	276/276	R
162	L	LEU162:A	-0.844	9	-0.928, -0.783	9,8	276/276	Q, M, I, Y, L
163	E	GLU163:A	2.124	1	1.250, 1.841	1,1	276/276	A, S, T, N, K, E, V, Q, M, D, R, I, G, L
164	V	VAL164:A	-0.405	7	-0.513, -0.364	7,7	276/276	S, A, Q, T, I, E, L, V
165	Y	TYR165:A	-0.962	9	-1.027, -0.928	9,9	276/276	F, H, Y
166	R	ARG166:A	0.357	4	0.059, 0.382	5,3	276/276	F, S, T, N, K, Y, E, V, Q, M, L, A, W, H, D, R, I, G
167	E	GLU167:A	1.940	1	1.250, 1.841	1,1	276/276	S, A, T, N, K, E, V, Q, D, R, G
168	Q	GLN168:A	-0.398	7	-0.513, -0.364	7,7	276/276	A, S, F, T, N, K, E, Q, M, L
169	T	THR169:A	-0.955	9	-1.006, -0.928	9,9	276/276	A, S, M, T, N, G
170	A	ALA170:A	0.845	1	0.382, 0.872	3,1	276/276	S, F, A, T, K, Y, E, V, H, Q, M, R, G, L
171	P	PRO171:A	-0.865	9	-0.957, -0.825	9,8	276/276	S, A, K, P, V, H, Q, L
172	L	LEU172:A	-0.735	8	-0.825, -0.690	8,8	276/276	I, L, V
173	I	ILE173:A	-0.311	6	-0.443, -0.277	7,6	276/276	A, T, C, K, I, L, V
174	E	GLU174:A	1.431	1	0.872, 1.841	1,1	276/276	S, A, T, N, K, P, E, V, H, Q, D, R, G
175	Y	TYR175:A	-0.283	6	-0.443, -0.179	7,6	276/276	H, F, I, Y, V
176	Y	TYR176:A	-1.076	9	-1.092, -1.074	9,9	276/276	Y
177	K	LYS177:A	1.908	1	1.250, 1.841	1,1	262/276	S, A, T, N, K, Y, E, H, Q, C, D, I, R, G, L
178	K	LYS178:A	1.969	1	1.250, 1.841	1,1	255/276	A, S, T, N, K, E, Y, V, H, Q, M, D, R, G, L
179	K	LYS179:A	0.694	2	0.382, 0.872	3,1	241/276	S, A, W, T, N, K, E, V, H, Q, D, R, L
180	G	GLY180:A	-0.640	5	-0.277, 0.059	6,5	238/276	S, A, N, K, E, V, Q, D, R, G
181	I	ILE181:A	1.692	1	1.250, 1.841	1,1	238/276	F, A, S, T, N, K, E, V, V, H, Q, D, I, R, G, L
182	L	LEU182:A	-0.328	6	-0.513, -0.277	7,6	237/276	F, A, W, K, P, Y, V, H, M, I, L
183	R	ARG183:A	2.702	1	1.250, 3.467	1,1	236/276	A, F, S, T, N, K, E, Y, V, H, Q, M, D, C, R, I, L
184	I	ILE184:A	1.365	1	0.872, 1.841	1,1	236/276	S, T, N, K, Y, E, V, Q, M, C, A, P, H, D, I, R, G
185	I	ILE185:A	-0.194	6	-0.364, -0.069	7,5	236/276	H, F, A, I, L, V
186	D	ASP186:A	0.060	5	-0.179, 0.207	6,4	235/276	A, S, N, K, P, E, Y, V, H, Q, D, I, R
187	A	ALA187:A	-0.794	8	-0.897, -0.739	9,8	235/276	S, A, N, P, G
188	S	SER188:A	0.554	3	0.207, 0.597	4,3	235/276	A, S, T, N, K, E, V, H, Q, M, D, I, R, G, L
189	K	LYS189:A	0.753	2	0.382, 0.872	3,1	235/276	A, T, N, P, K, E, V, Q, M, D, R, I, G, L
190	P	PRO190:A	1.517	1	0.872, 1.841	1,1	235/276	A, S, T, N, K, P, E, H, Q, D, G, L
191	V	VAL191:A	-0.136	6	-0.364, -0.069	7,5	233/276	F, S, A, T, N, K, P, V, Q, M, I, L
192	E	GLU192:A	1.201	1	0.872, 1.250	1,1	233/276	S, A, T, N, K, P, E, V, Q, D, G, L
193	E	GLU193:A	1.640	1	0.872, 1.841	1,1	233/276	S, A, T, N, K, E, V, Q, D, R, G, L
194	V	VAL194:A	-0.775	8	-0.862, -0.739	9,8	232/276	T, I, E, L, V
195	Y	TYR195:A	0.616	2	0.207, 0.872	4,1	231/276	A, S, F, W, T, N, K, E, Y, V, Q, M, R, I, L
196	R	ARG196:A	1.897	1	1.250, 1.841	1,1	230/276	S, A, T, N, K, E, V, H, Q, D, R, G, L
197	Q	GLN197:A	0.970	1	0.597, 1.250	3,1	230/276	A, S, T, N, K, E, V, H, Q, M, D, I, R, G, L
198	V	VAL198:A	-0.349	6	-0.513, -0.277	7,6	230/276	A, M, T, K, I, L, V
199	L	LEU199:A	2.616	1	1.250, 3.467	1,1	226/276	F, S, T, N, K, Y, E, V, Q, M, C, L, A, D, R, I, G
200	E	GLU200:A	3.099	1	1.841, 3.467	1,1	223/276	A, S, T, N, P, K, E, V, Q, D, R, G, L
201	V	VAL201:A	0.415	3	0.059, 0.597	5,3	222/276	S, A, T, K, E, V, H, M, C, R, I, L
202	I	ILE202:A	-0.286	6	-0.443, -0.179	7,6	220/276	A, F, N, I, L, V
203	G	GLY203:A	-0.076	5	-0.364, 0.059	7,5	101/276	A, S, N, K, E, Q, D, G, L
204	D	-	1.928	1	0.872, 1.841	1,1	54/276	S, N, K, E, H, Q, D, C, R, G
205	G	-	0.474	3*	-0.179, 0.872	6,1	11/276	D, N, K, G
206	N	-	-0.329	6*	-0.825, -0.069	8,5	4/276	N, K