THE UNIVERSITY OF
WARWICK

AUTHOR: **José Emilio Jiménez Roldán** DEGREE: **Ph.D.**

TITLE: **Rigidity analysis of protein structures and rapid simulations of protein motion**

DATE OF DEPOSIT: .....

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

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

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

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

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

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

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

DATE SIGNATURE ADDRESS

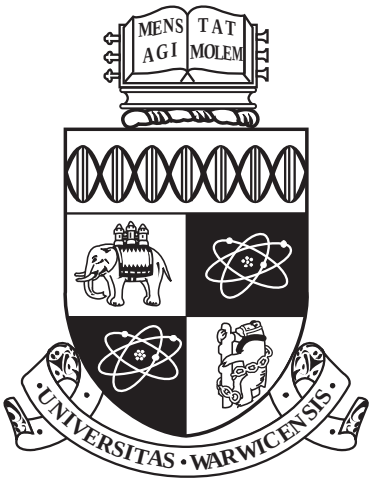
.....

.....

.....

.....

.....



Rigidity analysis of protein structures and rapid simulations of protein motion

by

José Emilio Jiménez Roldán

Thesis

Submitted to the University of Warwick

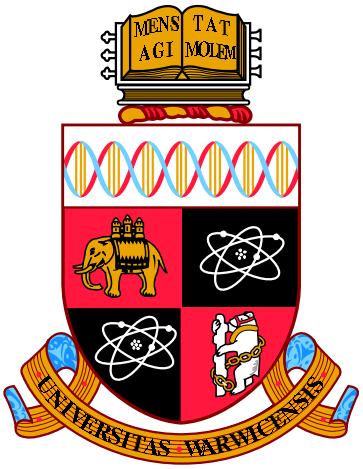
for the degree of

Doctor of Philosophy

Department of Physics and School of Life Sciences

July 2012

THE UNIVERSITY OF
WARWICK



Rigidity analysis of protein structures and rapid simulations of protein motion

by

José Emilio Jiménez Roldán

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Physics and School of Life Sciences

July 2012

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

Firstly, I would like to thank my supervisors Professor Rudolf A. Römer and Professor Robert B. Freedman for giving me the opportunity to carry out the work presented in this thesis, and also for their expert supervision, continued enthusiasm and guidance throughout. I am specially thankful for the shared passion to know more about how biological systems work by Professor Freedman. Also, for him inspiring my research by his constant enquire on how our simulations relate to in vitro protein behaviour. I am specially thankful to Professor Römer for his constant encouragement to strive for excellence during my PhD research, for his guidance in thinking strategically and planing my research. I am grateful to Dr. Stephen Wells for his guidance on the use of computational packages and continued support, to Dr. Katrina A. Wallis, John Blood and Kelly for their continued help and training in the lab, for their infinite patience in answering questions on biology to a physicist and for their guidance, training and teaching throughout the experiments. I would also like to express my thanks to our collaborators at the Indian Institute of Sciences (Bangalore-India), M. Bhattacharyya and Prof. S. Vishveshwara, for providing MD simulations data for this thesis and for many inspiring discussions. I am very thankful to our collaborators at the University of Warwick, H. Li and Professor P. B. O'Connor, for their hard work in our recent paper together. To our collaborators at the University of Kent, Professor M. Howard and Professor R. Williamson, I am thankful for their shared enthusiasm, teaching about NMR techniques and for our collaborative work. Further, I also wish to thank the rest of the members of the Freedman and Roemer groups, to the clerical and senior technical staff for their support and help and for making Warwick a very enjoyable place.

To all the selfless teachers, to my parents.

A los maestros del Ser, a mis padres.

Declarations

The work presented in this thesis is original work, conducted by myself under the supervision of Professor R. B. Freedman and Professor R. A. Römer. All sources of information have been acknowledged by means of references. None of this work has been used in any previous application for a degree. Some of the results presented in this thesis have been published in, or are in preparation to be submitted to a journal.

List of Publications

1.- 2012 Cross-linking experiments confirm computer simulations on yeast PDI mobility J. E. Jimenez-Roldan, H. Li, R.A. Römer, P.B. OConnor and R. B. Freedman. In preparation.

2.- 2012 Molecular dynamics and coarse graining simulations on yeast PDI mobility J. E. Jimenez-Roldan, M. Bhattacharyya, S. Vishveshwara, R.A. Römer, and R. B. Freedman. In preparation.

3.- 2012 Protein Flexibility is key to Cisplatin Cross-linking in Calmodulin H. Li, S.A. Wells, J. E. Jimenez-Roldan, R. A. Römer, Y Zhao, P.J. Sadler, P.B. OConnor. Mol. Cel. Proteomics, Submitted.

4.- 2012 Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses J. E. Jimenez-Roldan, R. B. Freedman, R. A. Römer and S. A. Wells. 2012 Phys. Biol. 9 016008.

5.- 2012 Inhibition of HIV-1 protease: the rigidity perspective J. W. Heal, J. E. Jimenez-Roldan, S. A. Wells, R. B. Freedman and R. A. Römer. Bioinformat-

ics (2012) 28 (3): 350-357.

6.- 2011 Rigidity analysis of HIV-1 protease J. W. Heal, S. A. Wells, J. E. Jimenez-Roldan, R. B. Freedman and R. A. Römer. 2011 J. Phys.: Conf. Ser. 286 012006.

7.- 2011 Characterisation of protein motion using a hybrid coarse graining method. J E Jimenez-Roldan, S A Wells and R A Römer. J. Phys.: Conf. Ser. Submitted

8.- 2010 Integration of FIRST, FRODA and NMM in a coarse grained method to study Protein Disulphide Isomerase conformational change J E Jimenez-Roldan, S A Wells and R A Rmer. 2011 J. Phys.: Conf. Ser. 286 012002

9.- 2009 Comparative analysis of rigidity across protein families S A Wells, J E Jimenez-Roldan and R A Römer. 2009 Phys. Biol. 6 046005

Abstract

It is a common goal in biophysics to understand protein structural properties and their relationship to protein function. I investigated protein structural properties using three coarse graining methods: a rigidity analysis method FIRST, a geometric simulation method FRODA and normal mode analysis as implemented in ELNEMO to identify the protein directions of motion. Furthermore, I also compared the results between the coarse graining methods with the results from molecular dynamics and from experiments that I carried out. The results from the rigidity analysis across a set of protein families presented in chapter 3 highlighted two different patterns of protein rigidity loss, i.e. "sudden" and "gradual". It was found that theses characteristic patterns were in line with the rigidity distribution of glassy networks. The simulations of protein motion by merging flexibility, rigidity and normal mode analyses presented in chapter 4 were able to identify large conformational changes of proteins using minimal computational resources. I investigated the use of RMSD as a measure to characterise protein motion and showed that, despite it is a good measure to identify structural differences when comparing the same protein, the use of *extensive* RMSD better captures the extend of motion of a protein structure. The in-depth investigation of yeast PDI mobility presented in chapter 5 confirmed former experimental results that predicted a large conformational change for this enzyme. Furthermore, the results predicted: a characteristic rigidity distribution for yeast PDI, a minimum and a maximum active site distance and a relationship between the energy cutoff, i.e. the number of hydrogen bonds part of the network of bonds, and protein mobility. The results obtained were tested against molecular dynamics simulations in chapter 6. The MD simulation also showed a large conformational change for yeast PDI but with a slightly different minimum and maximum inter-cysteine distance. Furthermore, MD was able to reveal new data, i.e. the most likely inter-cysteine distance. In order to test the accuracy of the coarse graining and MD simulations I carried out cross-linking experiments to test the minimum inter-cysteine distance predictions. The results presented in chapter 7 show that human PDI minimum distance is below 12Å whereas the yeast PDI minimum distance must be above 12Å as no cross-linking structures where found with the available (12Å long) cross-linkers.

Abbreviations

BM: Bismaleimide

BPTI: Bovine pancreatic trypsin inhibitor

DTT: Dithiothreitol

E.coli: Escherichia coli

E_{cut}: Cutoff energy

EDTA: Ethylenediamitetraacetic

FIRST: Floppy Inclusions and Rigidity Substructure Topography

FRET: Fluorescence resonance energy transfer

FRODA: Framework rigidity optimised dynamic algorithm

FRODAN: Framework rigidity optimised dynamic algorithm New

HCG: Hybrid coarse graining

IEC: Ion exchange chromatography

IMAC: Imomilised metal affinity chromatography

LB: Lysogeny broth

MD: Molecular dynamics

NEM: N-Ethylmaleimide

NMA: Normal mode analysis

NMR: Nuclear magnetic resonance

OD: Optical density

PDI: Protein disulphide isomerase

pLGIC: Pentameric ligand gated ion channel

RCD: Rigid cluster decomposition

RUM: Rigid unit motion

SDS-page: Sodium dodecyl sulfate polyacrylamide gel electrophoresis

List of Tables

List of Figures

Chapter 1

Introduction

Proteins are the main building blocks and functional molecules of the cell. Their function is determined by the polypeptide sequence and tertiary structure. These determine the proteins structural properties, i.e. rigidity, flexibility, mobility, reactive sites exposure, etc. Hence, understanding their structural and dynamic properties are crucial for understanding proteins biological function and cell function as a whole. There are many types of proteins and for some the relationship between mobility and function is very relevant. For example, some enzymes perform a multitude of biochemical and/or biomechanical reactions, such as altering, joining together or chopping up other molecules. These functions require the enzyme to move in space. Other proteins whose function requires structural motion are transmembrane proteins, which are key in maintaining a desirable cellular environment for the cell to function efficiently. These proteins regulate cell volume, ion transit across the cellular membrane, select molecules able to transit, etc. Motion is a key component of these protein structures but their dynamical behaviours may well span various long time scales and involve large numbers of residues. The wide range of protein architectures define and modulate the nature of the molecules' conformational dynamics in a complex way that it is still not completely understood. Therefore, the investigation of each protein structure on a case by case basis is essential.

Several experimental techniques are available to study protein structure and dynamics. The most commonly used to determine protein structure are X-ray or neutron crystallography, which provide a single snapshot of the spatial location of atoms. Nuclear magnetic resonance (NMR) [1] provides structural information but also dynamic information of the protein in solution. Other techniques to study protein motion are fluorescence resonance energy transfer (FRET) [2] or cross-liking experiments [3]. FRET involves attaching a fluorescent probe to different residues to

calculate the energy transfer between them and hence calculate the distance between the residues. Cross-linking involves attaching a polymer construct to two reactive sites in order to identify a distance between them via gel electrophoresis.

Computer simulations performance is defined in terms of the structural detail considered and the CPU-time employed to perform the simulations. Simulation techniques that require an all atom representation consider a great detail of the protein's structure so that all the atoms are accounted for. Molecular dynamics (MD) [4] is the gold standard for all atoms simulation method. It requires to solve Newton's equations of motion for the interacting atoms of the protein network where forces between the atoms and potential energy are defined by molecular mechanics force fields. Current detailed molecular dynamics methods typically require CPU-weeks or months to complete simulations of protein structures of the order of hundred residues.

MD has been one of the main simulation techniques, however, there is a need for techniques which are able to rapidly simulate large number of atoms and motions, e.g. hundreds or thousand residues. In this regard the emergence and increasing popularity of coarse graining models is due to their low computational expense and ability to provide quick responses. Coarse grained models like the model proposed by Go et al [5, 6], the Rosetta method [7] or the Elastic Network Model (ENM)[8, 9] use larger units than single atoms.

The difference in performance between models relies, among other factors, on how each model coarse grains the structure, i.e. on how the pseudo-units are defined in relation to the initial structure, its properties and the biological question to be addressed. For example, if blocks of atoms [10] or whole protein subunits [11] are considered as a unit, a further coarse graining step is achieved; however, this method does not distinguish the rigid from the flexible parts within the subunits and therefore essential structural information could be lost during the simplification process. As a consequence the results obtained with coarser models that do not take into account protein structural features could be far less accurate if they overlook essential structural features.

Although methods for fixing sub-units by using coarse grained models have come a long way since they started in 1976 [12], there is not a consensus yet or guidelines regarding how to choose the pseudounits. For example, the Rosetta method has been used in protein folding studies [13] by replacing a sequence of up to nine residues by a single body with six degrees of freedom. Whereas, other more sophisticated methods like the Elastic Network model (ENM) only focuses on the α-carbons, which are treated as point objects with three degrees of freedom [14].

1.1 Natural coarse graining

To address this issue, M. Thorpe and co-workers developed a new approach to determine the coarse graining pseudounits by identifying the rigid units of a crystallised biomolecule using topological and geometrical techniques. From this approach they developed the software package "FIRST" (Floppy inclusions and rigidity substructure topography) as a natural basis for coarse grain determination of rigid clusters [15] that integrates the "pebble game"[16] algorithm for rigidity analysis [17]. By matching degrees of freedom against constraints, it can rapidly divide a network into rigid regions and floppy "hinges" with excess degrees of freedom. The basic concept is that overconstrained regions will stay rigid during a mobility simulation and therefore can be treated as a single pseudounit.

In this approach, a protein is viewed as a network of different types of bonds holding the structure together. The strongest constraints, i.e. the covalent bonds, the locked and unlocked dihedral angles found along the polypeptide chain, are considered as strictly rigid. Whereas, the hydrogen bonds, salt bridges and hydrophobic tethers interactions that define the tertiary structure's constraint network, are considered to be breakable and affect protein rigidity and motion depending on their bond strength. Once the networks of constraints are obtained, FIRST is able to identify the rigid regions or clusters by assessing the number of constraints and degrees of freedom for each atom [17] and thus defining the rigid clusters as pseudounits that can be used to coarse grain a mobility simulation. Hence, using overconstrained or rigid regions as pseudounits is a method that varies from previous coarse graining methods in that the choice of the pseudounits is not arbitrarily but are defined according to the properties of the protein network, i.e. the overconstrained regions are considered as single rigid clusters and the underconstrained regions are considered as flexible.

1.2 Normal mode analysis

A well known coarse graining approach to simulate protein motion is the one implemented by normal mode analysis (NMA). This approach considers only the α-carbon atoms of a protein structure and models the interactions between them as Hookean springs based on a harmonic pairwise potential [18]. The α-carbon atom coordinates are derived from the crystal protein structures. The number interactions considered are limited by a cutoff value, typically between $9-14$Å, which is accounted for using a heavyside step function. According to the elastic network model [19] the elements

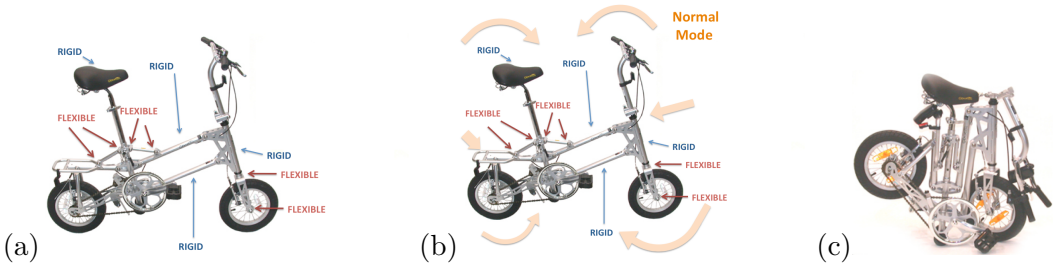


Figure 1.1: Pictorial metaphor comparing a folding bike with protein motion simulations using rigidity analysis and modes of motion. In panel (a) the rigid and flexible regions are identified, in (b) the directions of motion or normal modes, and (c) shows the bike in its folded state. The process using a hybrid coarse graining method to identify the rigid regions and NMA to identify the directions of motion before simulating protein motion is conceptually parallel.

of the Hessian matrix (H) are obtained from the second derivative of the potential (V) with respect to the Cartesian coordinates of the atoms. The normal modes are the eigenvectors obtained from the Hessian matrix, a $3Nx3N$ matrix composed of the second derivatives of the potential (V) with respect to residue fluctuations. Each normal mode define a network vibratory state which is characterised by a frequency and a mode, and independent of all the other modes. The calculated normal modes define an orthonormal mathematical basis set and provide information on all the possible directions that the protein structure can move. The dimensions of the orthonormal basis is the number of α-carbons contained in the protein structure. The vector defining the direction of motion for each α-carbon has three values which represent the directional component for each spatial axis. The normal modes do *not* suggest, at least right away, how the structure actually moves, which means that it is not possible to tell which modes are biologically relevant from the given set of modes calculated for a structure [8].

In recent years, NMA has emerged as a powerful computational method for studying large amplitude molecular motions. Due to the reduction in computational expense that its coarse graining procedure grants, it is one of the best suited methods for studying collective motions in proteins [8, 10, 14, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. The reasons underlying this success are not fully understood yet, especially since proteins are known to fold and function in a water environment, within a narrow rang of pH, temperature, ionic strength, etc., while NMA is performed in vacuum.

Intrinsic structural flexibility, as manifested in normal modes, facilitates the functionally important conformational changes. In simple terms, determining the

flexible and rigid regions is like examining a bicycle and finding out where the hinges and rigid bars are located as shown in Figure 1.1, whereas determining the normal modes of motions is like determining the possible configurations that the bicycle can adopt by exploring the range of motion of the movable parts.

There has been questions raised [27] with regards the limitations of NMA and the harmonic approximation. Studies of macromolecules are limited by the complex potentials used to describe the covalent and non-bonded interactions between atoms. There are high CPU-time requirements to compute such interactions limits the analysis of large proteins. The pioneering work of Tirion [18] demonstrated that the potential energy could be approximated by simpler pair-wise Hookean potential to sufficiently describe the low-frequency motion of large proteins. Although dismissing the an-harmonic terms, the simplified potential has proved to be an attractive alternative to model large conformational changes of large macromolecular assemblies.

1.3 Framework rigidity optimised dynamic algorithm

Protein motion along normal modes has been previously investigated using normal modes [28, 29]. However, the ability of the FRODA module [30] in FIRST to generate conformers is particularly useful to visualise conformers along the trajectory. A conformer is produced as the crystal structure is projected along the eigenvector and the structural constraints are meet. In a nutshell, a conformer is identical to the initial crystal structure but adopting a new conformation or distribution in space. The conceptual origins of FRODA are from studies on mineral crystal structures and the rigid-unit-mode (RUM) model [31, 32, 33, 34] which interprets the motion of a crystal network in terms of polyhedra moving as rigid units. Hence, the rigid parts are able to move as one block each in the directions that the flexible regions and structural constraints allow them to. The "Geometric Analysis of Structural Polyhedra" (GASP) is an implementation of the rigid-unit-mode model into a computer program written by S.A. Wells during his doctoral thesis to analyse mineral structures. The software performs real-space rigid unit analyses on framework structures with the primary aim of comparing two polyhedral framework structures to analyse their differences as a combination of rigid unit motion (displacement and rotation of polyhedra) and distortion of polyhedra [35].

FRODA's approach was therefore originally develop to simulate rigid-unit motion in framework mineral structures and then adapted to use protein rigid clusters as "ghost" templates instead of polyhedral rigid units. The "ghost" templates are

defined as the pseudounits that coarse grain the protein. They are defined as the rigid clusters defined in by FIRST rigidity analysis. Reducing the number of "units" in the simulation, i.e. atoms and/or pseudounits, allows for increased efficiency when simulating motion across the regions of the conformational space. Furthermore, the constraints associated with hard sphere steric repulsion effects are also accounted for so that atoms are not allowed to collide. Therefore, it is possible to interpret the protein motion through the conformational space as the movement of a dense packed assembly of rigid sphere clusters which can move while maintaining the covalent, hydrophobic, and hydrogen bond constraints between them. This approach is expected to yield good results especially for large biomolecules since the geometry will be largely determining the large scale motions. The directions of motion for a given protein can be explored in different ways using FRODA, either by: (a) a random bias, (b) a centrifugal motion from the centre of mass, or more elegantly (c) by using an eigenvector defined by an elastic network mode as defined by ELNEMO [25, 36].

1.3.1 Hybrid coarse graining methods

The use of FIRST as a coarse-graining method has been previously used as the basis for simulation methods exploring the large-amplitude flexible motion of proteins. The first mobility algorithm to be based on FIRST was the "ROCK" algorithm pioneered by Dr. Ming Lei [15, 37]. The ROCK method was developed to explore 3D conformations of a protein system constraint with the rigid clusters from FIRST. Its first application was done on a HIV protease structure and showed the extend that the flexible flaps could move [37].

Like ROCK, the concept behind the FRODA method [30] is to explore the allowed conformational space of a protein using the rigid clusters determined by FIRST to reduce the computational costs. FRODA improved significantly the sampling speed and prevented collision between atoms. These methods have been further developed by newer versions of FIRST and the new geometric simulation software FRODAN [38]. FRODAN resolves important outstanding issues with FRODA by integrating a new conceptual approach to handling rigid units during the simulation procedure and by enforcing constraints using a conjugate gradient minimization function.

The hybrid coarse graining (HCG) method I use in this thesis, integrates rigidity information from FIRST and NMA analyses as defined by ELNEMO with in the module FRODA to investigate conformational changes. Conceptually speaking, the multi-scale modelling approach merging rigidity information was first developed

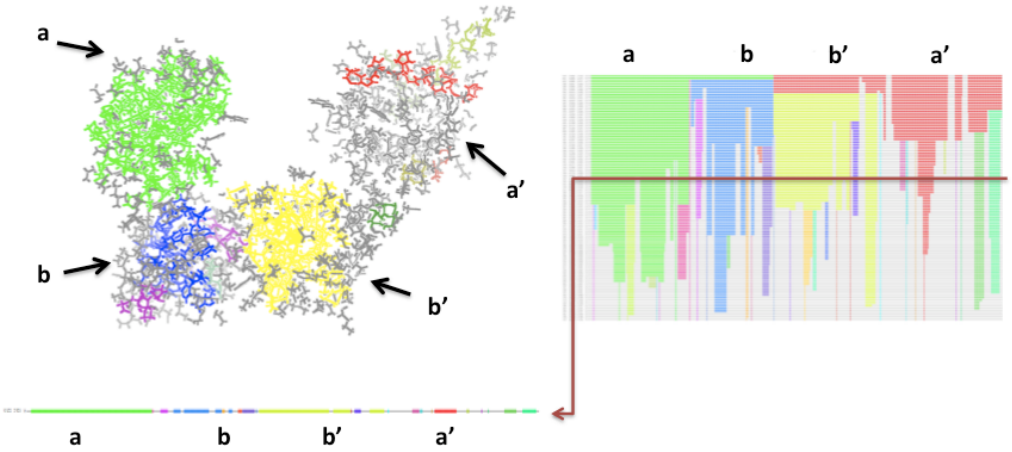


Figure 1.2: Rigidity distribution on the 3D structure and rigid cluster decomposition (RCD) graph of yeast protein disulphide isomerase. The right panel shows the rigidity dilution of yeast PDI and the selected polypeptide chain rigidity distribution chosen to coarse grain the protein structure. Each line represents the rigidity distribution along the polypeptide chain for a given energy cutoff (See Chapter 2 for a detailed explanation). The tertiary structure shown on the left hand side integrates the rigid clusters as defined in the RCD graph for the selected cutoff energy. The rigid clusters are coloured accordingly and the unconstrained regions are shown in grey. The protein domains of yeast protein disulphide isomerase are shown and labelled as ***a-b-b'-a'***.

by Ahmed et al. [21] and the integration of rigidity information and normal modes of motion to identify small harmonic displacements by Gholke et al. [39]. This outperforms the ENM technique on its own in terms of efficiency, allowing only translational and rotational degrees of freedom to the rigid clusters identified by FIRST but no relative motion within each cluster. Hence, the total number of degrees of freedom for the biomolecule is reduced to $\approx$ 30% compared with conventional ENM. Therefore, the memory requirements and computational times are reduced significantly by a factor of 9-125 [21].

These results support the hypothesis that identifying flexible and rigid regions to coarse grain the protein structure while performing a geometrical simulation as illustrated in Figure 1.2 provides an advantage to simulate protein motion that together with using normal modes of motion facilitates the ability to predict large-scale motions. The HCG method allows for a high level of versatility in modulating the motion of the protein structures for various directional modes, bonding condi-

tions and simulation parameters. A neat feature of FRODA is its ability to bias the motion of the initial structure using different mobility guides, e.g. random, centrifugal and using normal modes. These options make FRODA especially interesting to inspect large conformational changes of proteins at a very low computational cost and with a high degree of versatility in the simulation parameters. Furthermore, FRODA is able to provide intermediary conformers produced along the mobility simulations. This allows for visual inspection and comparison of the initial structure with respect to the conformers requested during the simulation.

Chapter 2

Materials and Methods

In this section I describe the overall computational methodology involved in integrating and performing the rigidity analysis, normal mode analysis and geometric simulations of flexible motion on a protein structure. A more detailed account of the methods used is included in each chapter. All the relevant computer codes (Reduce [40], PyMOL [41], First [16, 15], Froda [30] and Elnemo[25, 36]) are serial codes which run on the workstations of the Centre for Scientific Computing, either in interactive mode or in a scripted fashion using bash and pbs job submission scripts.

2.1 Select the relevant PDB files

The first step on each research project reported in this thesis starts by selecting a biological question, then the relevant X-ray crystal structures are selected from the protein data bank (PDB) [42] taking into account the experimental resolution and the crystallisation conditions. I choose to use X-ray crystal structures with the best resolution or at least better than 2.5Å when possible and with crystallising conditions that are relevant to the biological question investigated, e.g. a protein X-ray crystal structures with different ligands bound, as a dimer, monomer, etc.

2.2 Adding hydrogen bonds

X-ray protein crystal structures do not contain hydrogen atoms. Therefore in order to identify the hydrogen bonds, salt bridges and hydrophobic tethers that hold the tertiary structure in place they need to be added to the structure. Hydrogen bonds are formed when a charged residue of a protein that has a polar covalent bonds

Step	Action	Tool
1	Obtain crystal structure	PDB
2	Remove water and atoms added during crystallisation	PyMOL script
3	Add hydrogen bond to the structure	Reduce
4	Rank hydrogen bonds	First
5	Obtain RCD graphs	First
6	Obtain normal modes	Elnemo
7	Identify E_{cut} for mobility coarse grain	RCD graphs
8	Integrate rigidity distribution (at the chosen E_{cut}) and normal modes	Bash scripts and First
9	Simulate protein motion	Bash scripts and Froda
10	Obtain conformers	Bash scripts and Froda
11	Obtain structural information	Bash and fortran scripts

Table 2.1: Stepwise work-flow used to analyse protein structures. This table summarises the steps followed to obtain the results presented in this thesis. Starting from the choice of the PDB structure to obtain the respective structures and ending by obtaining structural data from the newly obtained structures.

forms an electrostatic interaction with a residue of opposite charge. Hydrophobic (non-polar) bonds are formed as hydrophobic tethers avoid contact with the polar water molecules. Whereas a salt bridge is actually a combination of hydrogen bonding and electrostatic interactions. I use the freely available software Reduce to "dress" the protein structure with its corresponding hydrogen atoms. This software takes into account the chemistry and geometry of the molecule to create the hydrogen bond network. The software aims to add hydrogen atoms to use contact dots to quantitatively analyse the network of bonds that pack proteins. It includes the correction of side-amide flips and avoids incorrect Histidine influence of Arsenie/Glutamine orientations. However, it does not include a compete analysis of Histidine protonation equilibria. This analysis could be included in future software versions but will require detailed knowledge of the pH and electrostatics [40].

The hydrogen bonds and salt bridges are defined using the geometry and energy of the interactions. Donor-acceptor distances (d ≤ 3.6Å), hydrogen-acceptor distances (r ≤ 2.6Å) and donor-hydrogen-acceptor angles ($90° \leq \theta \leq 180°$) define the sets of bonds included in the rigidity analysis [17]. Salt bridges interactions are considered as a special case of hydrogen bonds with a more significant ionic component, which is less geometrically sensitive. Salt bridges are identified by a maximum distance between donor and acceptor of ≤ 4.6Å and softening the angular dependence to ($80° \leq \theta \leq 180°$).

Since the strength of these bonds depends on the chemistry of the donor and acceptor atoms, and on their orientation[17], an energy function is used to rank hydrogen bonds as defined by equation 2.1. Where $F(\theta, \phi, \psi)$ depends on the geometrical constraints as defined in [17], $V_0 = 8kcal/mol$ and $d_0 = 2.8$Å.

$$E_{HB} = V_0 \left\{ 5 \left(\frac{d_0}{d} \right)^{12} - 6 \left(\frac{d_0}{d} \right)^{6} \right\} F(\theta, \phi, \psi) \tag{2.1}$$

2.3 Ranking of hydrogen bonds

Salt bridges or salt bonds are weak ionic bonds that contribute to the stability of the protein structure. They form between positively charged amino acids (arginine or lysine) and negatively charged amino acids (aspartic acid or glutamic acid). Once the hydrogen atoms have been added to the structure and their strength has been calculated based on their geometric properties, the hydrogen bonds and salt bridges are normalised and ranked according to their energetic strength from weakest to strongest. From now on we would refer to both hydrogen bonds and salt bridges as hydrogen bonds only. This ranking occurs within the program FIRST.

2.4 Floppy inclusions and rigidity substructure topography

FIRST implements the "pebble game", an integer algorithm for rigidity analysis which matches degrees of freedom against constraints to rapidly predict protein flexible regions. By using a protein crystallographic structure as an input file obtained for example from the PDB, FIRST is able to identify the network of bonds [17] between amino acids, describe them by their degrees of freedom and orientation and classify the bonds in order of bond strength. Once this process is completed, FIRST is able to identify rigid and flexible regions of a given protein structure by identifying the overconstrained and underconstrained regions using a flexibility index as defined in [17]. This approach is straightforward to implement and can be combined with other numerical simulation techniques. There have been a variety of studies applying rigidity analysis using FIRST to study phenomena such as virus capsid assembly [43] and "folding core" determination by simulating thermal denaturation [44] or structural properties of HIV-1 protease predictions [17].

The covalent bonding between atoms is of course included, as are hydrophobic interactions between adjacent hydrophobic side-chains. Hydrogen bonds are identified based on donor-hydrogen acceptor geometry; the "salt bridge" interaction between adjacent, oppositely charged ionic groups are also so identified. Non specific long range forces (such as general electrostatic and dispersion interactions) are not counted as constraints. This hierarchy and selection of constraints is discussed in detail in the literature on FIRST [17].

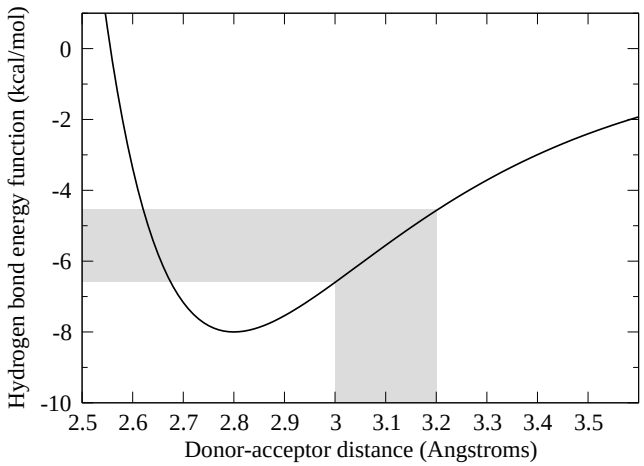


Figure 2.1: Dependence of hydrogen bond energy E in FIRST on the donor-acceptor distance. The shaded region indicates how an distance variation of ± 0.1Å can lead to a variation in the bond energy of more than 1 kcal/mol.

The energy of each potential hydrogen bond in the processed structure is calculated in FIRST using the Mayo potential [45]; the distance-dependent part of this potential is shown in Figure 2.1. For the dilution, FIRST performs an initial rigidity analysis including all the bonds with energies of 0 kcal/mol or lower; bonds are then removed in order of strength, gradually reducing, or "diluting", the rigidity of the structure.

2.5 Rigid cluster decomposition graphs

RCD graphs show the process of assessing rigidity distribution after removal of each single hydrogen bonds, which is done one by one in ascending order of bond energetic strength. In the RCD graphs the bond energy is defined in negative terms, so although the absolute bond energy strength increases as the bonds are removed from the weakest to the strongest, the energy scales in the RCD graphs go from zero to negative values to account for the attractive nature of the bonding force. The rigidity dilution pattern tell us about how rigidity is distributed across the three dimensional structure and how the hydrogen network strength evolves as we deepen into removing the stronger bonds. This gives a good idea of which areas are rigid but also of how rigidity loss falls as a function of energy. This capability of FIRST also allow us to compare rigidity of molecules at any given cutoff value.

Once the FIRST analysis is finished the results obtained can be plotted either in a RCD graph as shown in Figure 2.2, or into a three dimensional protein

structure representation as shown in Figure 2.3. The RCD graph is a bar graph (indicating the protein sequence) that allows us to visualise at a glance how rigidity is distributed across the protein sequence and how the rigid clusters evolve as more energetic hydrogen bonds are removed. The horizontal axis in Figure 2.2 represents the protein's linear primary structure. Flexible areas of the polypeptide sequence are shown as horizontal thin black lines while areas lying within a rigid cluster are shown as thicker coloured blocks. Colour is used to differentiate which residues belong to which rigid cluster. The three-dimensional protein fold makes it possible for residues that are widely separated along the backbone to be spatially adjacent and form a single rigid cluster. The vertical axis on the dilution plot represents the dilution of constraints by progressively lowering the cutoff energy for inclusion of hydrogen bonds in the constraint network. Each time the rigid cluster analysis of the mainchain α-carbon atoms changes as a result of the dilution, a new line is drawn on the plot, labelled with the energy cutoff and with the network mean coordination for the protein at that stage. It should be stressed that the RCD is always performed over the entire protein structure (mainchain and sidechain atoms) and a dilution is performed for every hydrogen bond removed from the set of constraints, typically several hundred bonds for a small globular protein.

The three dimensional representation of the rigidity distribution allows direct visualisation of the rigid clusters as they appear in the protein structure providing information on how neighbouring sites may work together by having similar rigidity. The comparison of the three dimensional structures allows for clear visualisation of how the changes in rigidity affect the mobility of the structure and it makes it easier to assess the biological significance of such changes.

As a natural place to investigate the rigidity and mobility behaviour of proteins has been suggested to be around room temperature by Jacobs et al.[17]. The correspondent equivalent to room temperature of 25° is equivalent to $-0.6kcal/mol$, which is equivalent to $1KT$ (where K is the Boltzmann constant and T is the temperature).

2.6 Structural comparison by RMSD

When dealing with slightly varying crystal structures of the same protein, the structural variation is quantified by aligning the α-carbon atoms of two structures and obtaining the root-mean-square deviation between α-carbon positions, where d_{ii} is

the distance between the α-carbon atoms of residue i in the aligned structures.

$$d = \sqrt{\frac{1}{N_{C\alpha}} \sum_{i=1}^{N_{C\alpha}} d_{ii}^2} \tag{2.2}$$

2.7 Normal modes of motion

There are several implementations of the elastic network models available. For the purpose of the work here presented I use the elastic network modelling implemented by the program ELNEMO.

The initial input required by ELNEMO is the spatial location of the α-carbons which I obtained from the pdb structure. ELNEMO has two main modules, the first one named PDBMAT generates the matrix of the network of bonds and the second module, DIAGSTD diagonalises the matrix and provides the eigenvalues and eigenvectors that define the normal modes of motion. Finally I split the eigenvectors into individual modes and chose the number of modes I want to investigate so that FRODA can use them as an input to bias the motion of the protein structure.

Since the lowest-frequency modes are expected to have the largest amplitudes and thus be most significant for large conformational changes I focus on the first fifth modes only. It is worth noting, however, that the very six lowest-frequency modes (modes 1 to 6) are trivial combinations of rigid-body translations and rotations of the entire protein. Hence, I refer to the lowest frequency mode as mode m_7 from hereof.

For the purpose of the simulations presented in this thesis I use the following FRODA options. The software FIRST requires of the FRODA option (-FRODA) to run the geometric simulation feature. For each simulation a list of hydrogen bonds (-hbin) and hydrophobic tethers (-phin) bonds are included as defined by the rigidity analysis in FIRST. It is also possible to define the number of conformers that we want to explore (-totconf), the frequency at which a conformer will be reported (-freq), e.g. for an option "-freq 100" the software will record every hundred conformer obtained from the total number of conformers calculated, define the number of iterations that the system will try to fit the rigid clusters at each simulation step (-maxfit), define that the directed motion is done for the direction of motion given by a normal mode (-modei), define the distance the atoms are projected at for a directed (-dstep) and random (-step) direction of motion.

2.8 Obtaining new conformers with FRODA

The FRODA method considers two types of mobile entities during the simulation process: atoms with three degrees of freedom and rigid "ghost templates" with six (rigid-body) degrees of freedom. The exploration of the conformational space is done following a series of two cyclical steps: (a) a random or guided perturbation is applied to all atoms and (b) the enforcement of constraints.

During the perturbation step each atom is displaced by a small perturbation, which can be random, guided (e.g. as defined by an eigenvector obtained from normal mode analysis) or a combination of both. The magnitude of the random motion is set to 0.1Å and the guided motion is set to 0.01Å during the simulations here presented. After the perturbation the atoms violate the network constraints as they no longer maintain the relationships with each other. Therefore, an iterative procedure is put into action to enforce these constraints. There are two steps in the procedure, first an iterative enforcement procedure and second a constraint fitting procedure to avoid steric overlap and maintain hydrophobic contact. First, the enforcement procedure is as follows: (a) Each ghost template is displaced to the location that minimises the sum of square distances between the physical atoms and the ghost atoms, (b) then the position of each physical atom is updated to the mean position of the ghost atoms they belong to. The iterative process of fitting or re-fitting ghost templates (again to minimise the sum of square distance between the ghost atoms) to the new positions of the atoms, and again the atoms are updated to the new positions of the ghost atoms. This iterative process continues until each single atom coincides with its corresponding ghost atom within some threshold, typically 0.125Å. When the minimum distance threshold between physical and ghost atoms is satisfied a new conformer is generated. Second, in order to prevent atom overlap and hydrophobic contact the iterative procedure is modified. The procedure to handle the minimum distance constraints that avoid overlap of non-bonded atoms starts by searching for any pairs of non-bonded atoms which relative positions are closer than the contact distance value determined by summing radii values for the atoms. This is done before the atoms are moved to the mean position of their ghost atoms. Likewise, the hydrophobic contact pairs are checked for any pairs that are farther apart than an allowed maximum distance. Hence, the distance violations, e.g. non-bonded atomic radii values and hydrophobic contact pairs, are addressed by displacing each atom a distance equal to the sum of the following vectors: (a) a vector that would displace the atom to the mean position of its ghost atom, (b) a vector that would displace the atom most directly away from an overlapping neighbour by half

of the overlapping inter-atomic distance. This applies for each overlapping atoms, i.e. if a given atom overlaps with multiple neighbouring atoms the displacement is added for each overlap, (c) a vector that would displace an atom directly towards a hydrophobic partner by half of the violated hydrophobic distance, again if an atom violates more than one hydrophobic contact there will be a displacement added for each case. The aim of these displacement movements is to avoid steric overlap and maintain hydrophobic contact by moving the 'conflictive' atoms towards and away from the atoms which are too close or too far and hence facilitate that the network converges to a conformer that is stereochemically acceptable.

The procedure to enforce constraints continues until three constraints are met; first, the 0.125Å tolerance distance between atoms and ghost atoms is respected; second, the distance between any two non-bonded atoms is greater than 85% of their Van der Walls radii; and third, the distance between hydrophobic atoms does not exceed the 0.125Å maximum distance.

2.9 Limitations and problems with FRODA

Despite providing a significant improvement compared to ROCK in terms of speed of exploring the conformational space and ensuring stereochemical constraints, several aspects of FRODA could be improved. Firstly, a fairly common occurrence during FRODA simulations is a sudden abort of the simulation after the fitting procedure repeatedly failed to satisfy constraints. Although it is difficult to exactly diagnose, the enforcement of constraints within the iterative procedure, which moves the constraint violating atoms half the distance of the violated space, seems the best candidate to account for the sudden jamming. The enforcement of constraints procedure does not ensure that the number of constraint violations are reduced at each step. For example, atoms in a crowded environment could face multiple overlaps at the same time or a group of overlapping atoms could simultaneously provoke alternate corrective distances that provoke recursively new violation of constraints. This could lead to the atoms being bounced back and forth from overlap to overlap, which could explain the limitations of the software in terms of how the jamming effects occur.

Another issue that is worth noting in FRODA is related to the rigid cluster templates that it incorporates from FIRST. The hydrogen bonds and hydrophobic contacts are considered as rigid within the rigid clusters but also are maintained rigid as the protein moves along the normal mode. Therefore, keeping the bonds distance and orientation fix, and unable to rotate so that the residues within the rigid

clusters are prevented from readjusting limits the motion of the protein artificially. Hence, a large scale motion may be inhibited or even blocked out if the geometries of the hydrogen and hydrophobic bonds are not allowed to re-arrange.

Besides the artificially imposed limitations in residue rearrangements, the use of rigid clusters to coarse grain protein motion could limit protein motion in other ways when used with FRODA. For example, higher cutoff energies implies bigger rigid clusters and therefore less atoms that are "free" to participate in the iterative process of fitting atoms to ghosts. Hence increasing the chances of FRODA being unable to generate an acceptable conformer. Likewise, when the structure has moved a given distance, the above mention effects would also manifest as the spatial disposition of the rigid clusters increasingly varies from the original one. Hence increasingly limiting the probabilities for FRODA finding new conformers. As the protein moves, the number of "free" atoms that reach near the constraint limits increases and therefore the difficulties for FRODA generating an acceptable conformer also increase.

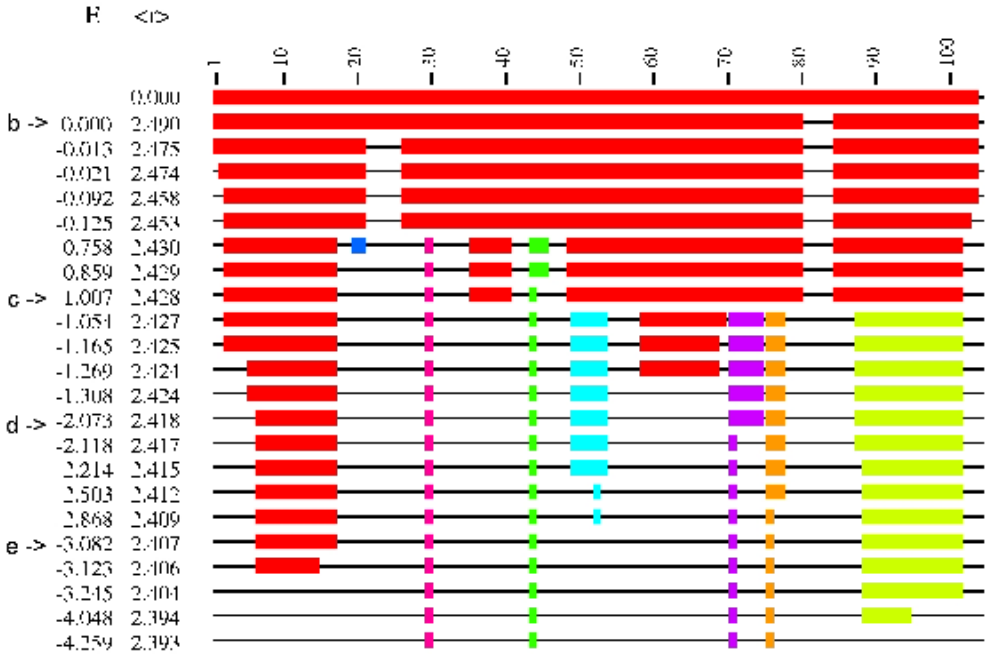


Figure 2.2: (a) Dilution plot for horse Cytochrome-c from the 1HRC structure. Flexible regions of the polypeptide chain appear as black thin lines, whereas rigid portions appear as coloured along the protein chain with α-carbon labelled from 1 to 105. The first column on the left indicates the energy (E) of the bond that is removed to generate the new rigidity distribution. A given bond energy is named as energy cut ($E_{\rm cut}$) to identify a cutoff which defines the protein rigidity distribution. The second column on the left indicates the mean number $\langle r \rangle$ of bonded neighbours per atom as the energy cutoff $E_{\rm cut}$ (kcal/mol) changes. When $E_{\rm cut}$ decreases (left-most column), rigid clusters break up and more of the chain becomes flexible. Colour coding shows which atoms belong to which rigid cluster.

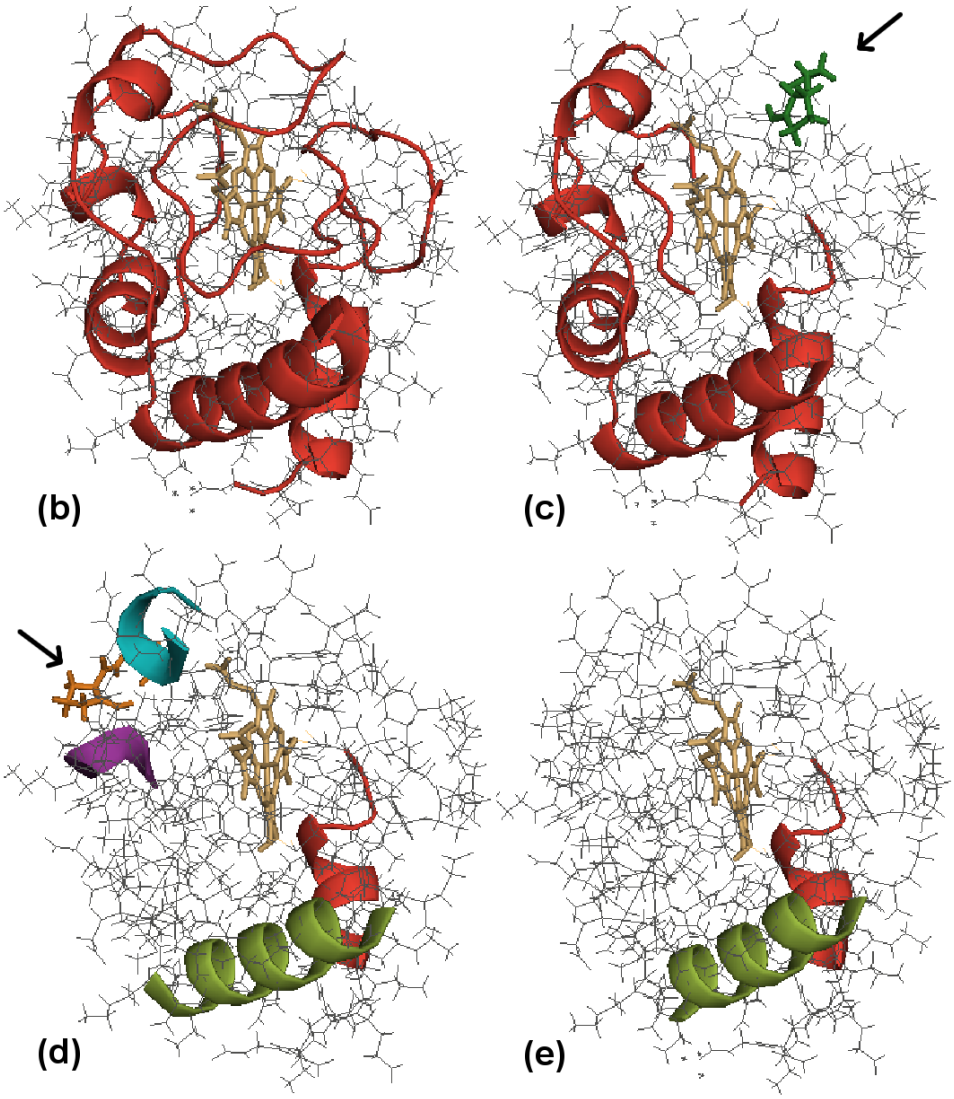


Figure 2.3: (b,c,d and e) Rigidity distribution for horse Cytochrome-c from the 1HRC structure in 3D. These figures represent in grey the flexible regions and in colour the largest rigid regions for the native state at energy cutoffs (b) $E_{\text{cut}} = 0.000$ kcal/mol, (c) $E_{\text{cut}} = 1.007$ kcal/mol, (d) $E_{\text{cut}} = 2.073$ kcal/mol and (e) $E_{\text{cut}} = 3.082$ kcal/mol, respectively. For each figure, the colour coding correlates with the colour coding given in (a). The arrows in (c) and (d) indicate two smaller rigid clusters shown in "stick" representation for clarity. The heme group is shown in "stick" representation (yellow).

Chapter 3

Rigidity analysis of protein families

This chapter presents a comparative study in which "pebble game" rigidity analysis is applied to multiple protein crystal structures, for each of six different protein families. The results show that the main chain rigidity of a protein structure at a given hydrogen-bond energy cutoff (E_{cut}) is quite sensitive to small structural variations, and conclude that the hydrogen bond constraints in rigidity analysis should be chosen so as to form and test specific hypotheses about the rigidity of a particular protein. Our comparative approach highlights two different characteristic patterns ("sudden" and "gradual") for protein rigidity loss as constraints are removed, in agreement with recent results on the rigidity transitions of glassy networks.

3.1 Introduction

The primary motivation for this chapter is to explicitly compare the results of rigidity analysis on groups of very similar crystal structures and particularly concentrates on six proteins (Cytochrome-c, Hemoglobin, Myoglobin, α-lactalbumin, Trypsin and HIV-1 protease). For each protein structure I observe the pattern of rigidity loss during the progressive removal of hydrogen bonds, or RCD plot [46, 44]. The *main-chain rigidity* is defined as a measure of the rigidity of the protein backbone in order to describe the rigidity loss during dilution. On the basis of this study I comment on the selection of E_{cut} values and the interpretation of rigidity analyses.

The second motivation for this chapter is to observe the pattern of rigidity loss during dilution. Previous studies on protein folding [46] have drawn comparisons between the folding transition of proteins and the rigidity transition of glassy

networks. A recent study [47] found that the rigidity transition in glasses could display either first-order or second-order behaviour depending on the character of the constraint network. In the first case, a small change in the constraints causes a sudden transition from an entirely floppy state to one in which the entire system becomes rigid. In the second, rigidity develops in a percolating rigid cluster which initially involves only a small proportion of the network and then gradually increases in size as more constraints are introduced. Our data on rigidity dilution shows that both types of transition are possible in proteins, with four of our proteins typically displaying "gradual" rigidity change and two (trypsin and HIV-1 protease) displaying "sudden" rigidity change [48].

3.2 Materials and Methods

3.2.1 Protein selection

The sets of proteins are chosen from the PDB [42] to obtain crystal structures for our comparison, as summarised in Table 3.1, and especially proteins that fall into two categories (i) examples of the same protein from different organisms, e.g. Cytochrome-c proteins from multiple different eukaryotic mitochondria, and (ii) protein structures obtained under different conditions of crystallisation, e.g. in complex with different ligands, proteins or substrates.

Rigidity analysis is best carried out on crystal structures with high resolution, therefore the X-ray crystal structures selected from the PDB have a resolution better than 2.5Å. A single protein chain was extracted from each PDB crystal structure and all water molecules were eliminated. Since the hydrogen atoms are absent from X-ray crystal structures they were added using the REDUCE software [40] which also performs necessary flipping of side chains. After the addition of hydrogens the atoms were renumbered using PyMOL to produce files usable as input to FIRST [17]. Only in the case of HIV protease I analysed the homo-dimer unit since it is the functional unit.

3.2.2 Mainchain rigidity loss during dilution

Dilution plots of very similar protein structures can be compared directly. This form of comparison, however, becomes unwieldy when comparing large numbers of structures, and can obscure differences in the hydrogen-bond energy scale. For glassy networks [47] the overall degree of rigidity of the structure was measured by the number of atoms in the largest spanning rigid cluster in a network with peri-

Table 3.1: List of proteins, organism of origin, PDB codes and figures they appear.

Protein	Organism	PDB ID	Figure	Comments
Cytochrome-c	Horse	1HRC	3.4	uncomplexed
		1WEJ		complexed with antibody E8
		1U75		complexed with peroxidase
		1CRC		at low ionic strength
Cytochrome-c	Tuna	5CYT	3.4	ferriCytochrome
		1I54		2FE:1ZN mixed-metal porphyrins
		1I55		2ZN:1FE mixed-metal porphyrins
		1LFM		Cobalt(III)-subsituted
Cytochrome-c	Rice	1CCR	3.5a	
	Bonito	1CYC		
	Bacteria	1A7V		
	Tuna	1I55		
	Yeast	1YCC		
		2YCC		
Myoglobin	Horse	1DWR	3.5b	
	Whale	1HJT		
	Turtle	1LHS		
α-lactalbumin	Baboon	1ALC	3.5c	
	Human	1HML		
	Goat	1HFY		
	Human	1A4V		
	Guinea pig	1HFX		
	Cattle	1F6R		
Hemoglobin	Human	1A3N	3.5d	deoxy
(α chain)		2DN1		oxy
		2DN2		deoxy
		2DN3		carbonmonoxy
	Goose	1A4F		
	Rice	1D8U		
	Bacteria	1DLW		
	Alga	1DLY		
	Cattle	1G09		
	Worm	1KR7		
	Clam	1MOH		
HIV-1 Protease	Virus	1HTG	3.5e	homodimers with inhibitors bound
		4HVP		
		7HVP		
		8HVP		
		9HVP		
Trypsin	Salmon	1A0J	3.5f	
	Cattle	1AQ7		
		1AUJ		
	Pig	1AVW		
	Pig	1AVX		
	Cattle	1AZ8		
	Rat	1BRA		
		1BRB		
		1BRC		
	Cattle	1BTH		
	Salmon	1BZX		
	Human	1H4W		
		1HPT		
	Cattle	1K1I		
		1K1J		
		1K1M		
		1K1N		
		1K10		
		1K1P		
	Pig	1LDT		
	Human	1TRN		
		2RA3		
	Rat	3TGI		

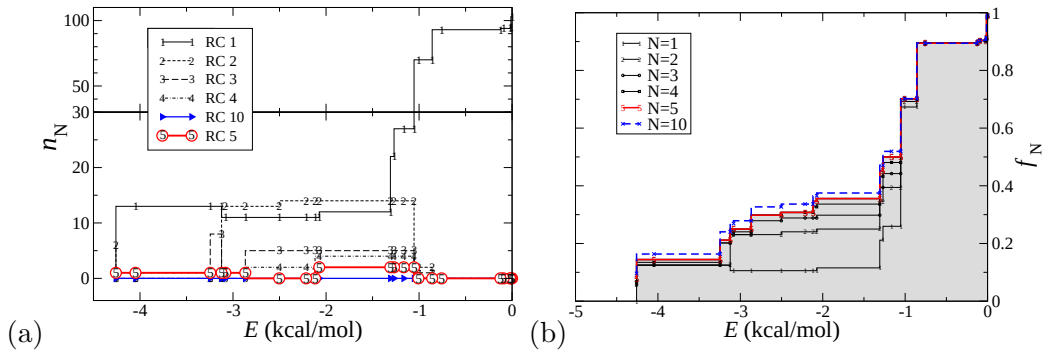


Figure 3.1: (a) The number n_N of α-carbon atoms contained within rigid clusters (RC) $N = 1, \ldots, 5$ and 10 of the 1HRC structure. Smaller, higher-numbered clusters do not contain more than one α-carbon. (b) The fraction f of the protein's α-carbon atoms contained within clusters 1 to N. The line corresponding to the $N = 5$ data has been shaded to show that the inclusion of rigid clusters 1 through 5 captures the large-scale rigidity of the protein.

odic boundary conditions. We therefore extract from the dilution plots a quantity measuring the overall degree of rigidity of the structure, and plot it as a function of the hydrogen-bond energy. Since the protein is not a periodic structure, its overall rigidity is measured by considering how many of its residues are included in large rigid clusters.

For protein structures, we first extract the number $n_N(E)$ of α-carbon belonging to each of the first N largest rigid clusters. Then $n_N(E)$ is normalized as a fraction of the total number $\mathcal{N}_{C_\alpha}$ of α-carbon. This allows us to compare the rigidity between different proteins.

So the measure of overall rigidity $f_N(E)$ defined in Eq. 3.1, is the fraction of atoms that are found in the N largest rigid clusters, essentially, those that appear within large blocks in the dilution plot.

$$f_N(E) = n_N(E)/\mathcal{N}_{C_\alpha} \tag{3.1}$$

Figure 3.1a shows the number n_N of α-carbon contained within the larger N rigid clusters of the horse Cytochrome-c structure 1HRC, for which the total number of α-carbon atoms equals $\mathcal{N}_{C_\alpha} = 105$. It is clear that only the first few rigid clusters (numbered 1–5) contain more than one α-carbon while higher-numbered clusters do not contain more than one C_α and do not represent two or more residues forming a single rigid unit. Typically single residues containing one α-carbon exist as rigid clusters when enough hydrogen bonds have been removed from the system. This

From\To:	1HRC	1CRC	1WEJ
1CRC	0.32	—	—
1WEJ	0.318	0.321	—
1U75	0.472	0.53	0.572

Table 3.2: RMSD variations for the α-carbon positions among four horse Cytochrome-c structures (Å) showing the similarity of the structures.

happens for example with proline due to its cyclic structure. Figure 3.1b shows the fraction f_N of α-carbon contained in the first N cluster, defined as in Eq. 3.2, for those α-carbon lying within rigid clusters $N = 1$ to 5 and also 10.

$$f_N(E) = \frac{1}{\mathcal{N}_{\mathrm{C}_\alpha}} \sum_{1}^{N} n_N(E) \tag{3.2}$$

The inclusion of the first five rigid clusters captures the large-scale rigidity of the protein; the difference between $N = 5$ and $N = 10$ is minimal. Therefore the use of the $N = 5$ measure, $f_5(E)$, to quantify protein rigidity is justified and will be referred to as *mainchain rigidity*.

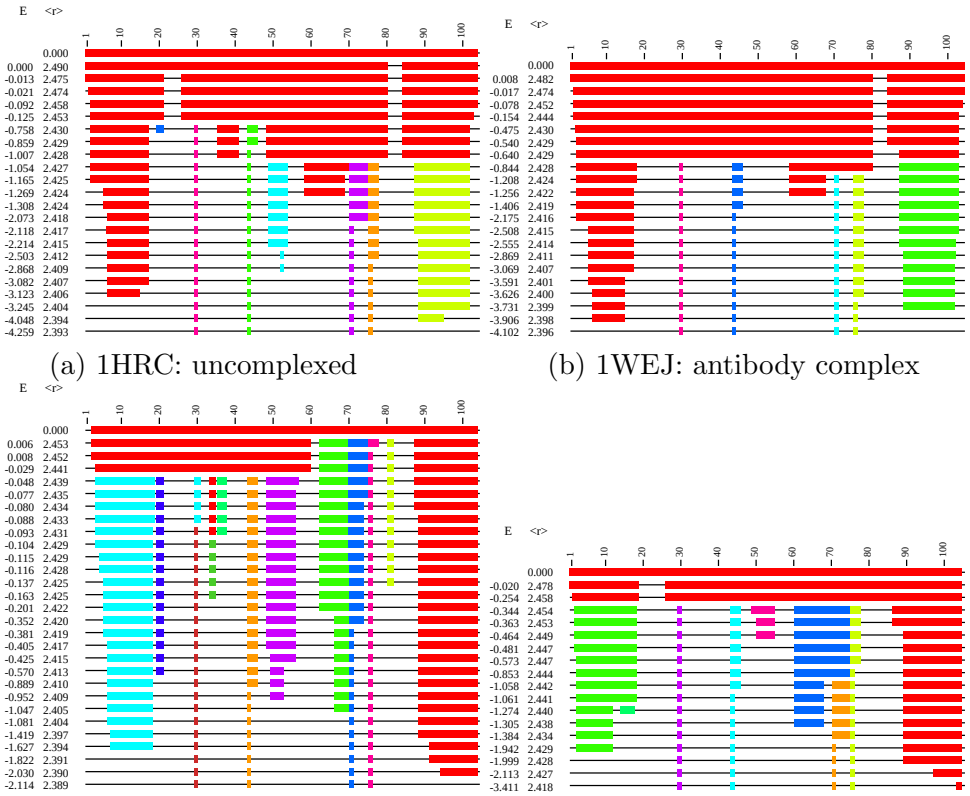
It is worth noting the "stepped" appearance of our graphs. This is because a given pattern of rigidity persists as the E_{cut} is lowered until at a specific value it changes and a certain amount of rigidity is lost.

3.3 Results

3.3.1 Rigidity variation of proteins crystalised under different conditions: Cytochrome-C

The dilution plots for four mitochondrial Cytochrome-c structures obtained from horse crystallised under different conditions are shown in Figure 3.2 and the crystallisation conditions are detailed in Table 3.1. The structural variations between these four structures are relatively small, the largest being 0.572Å between 1U75 and 1WEJ, see Table 3.2. The patterns of rigidity loss in Figure 3.2 appear quite similar on first inspection. The central portion of the protein sequence breaks up into smaller clusters and then becomes entirely flexible, while the rigidity of the two ends of the sequence, around residues 5–15 and 90–100, persists longer; due to this persistence, these portions (α-helical in secondary structure) were identified in [44] as being the folding core of Cytochrome-c, in agreement with experimental evidence.

On closer inspection, however, the differences between the rigidity distribu-



(a) 1HRC: uncomplexed

(b) 1WEJ: antibody complex

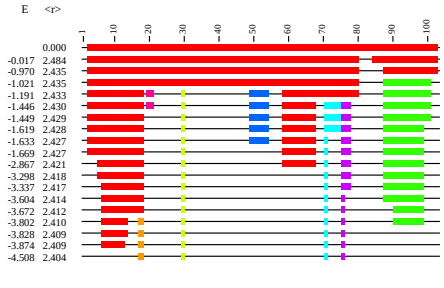
(c) 1U75: peroxidase complex

(d) 1CRC: low ionic strength

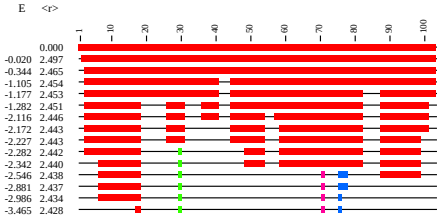
Figure 3.2: Dilution plots for four crystal structure of horse Cytochrome-c. The four structures are very similar to each other (see text) and display similar patterns of rigidity loss. The central portion of the protein sequence breaks up into smaller clusters (e.g. close to $E = -1$ kcal/mol for 1HRC and $E = -0.7$ kcal/mol for 1WEJ) and then becomes entirely flexible, while the rigidity of the two ends of the sequence, around residues $5 - 15$ and $90 - 105$, persists longer; these portions are α-helical in secondary structure.

tion of the four structures are clear. For example, in structures 1HRC and 1WEJ, the terminal α-helical sequences remain rigid down to $E_{\rm cut}$ values below -3 kcal/mol, while in 1CRC and 1U75 these sequences are already largely flexible at a $E_{\rm cut}$ value of -2 kcal/mol.

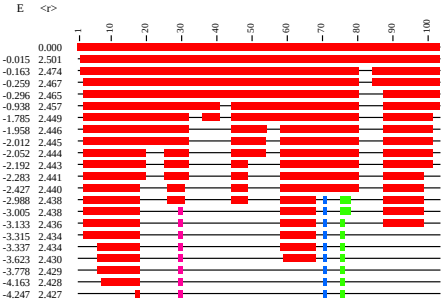
To illustrate this point I plot the main-chain rigidity for these four proteins as a function of $E_{\rm cut}$ during dilution in Figure 3.4a. The difference in energy scale of the rigidity loss is now clearly visible. It is worth noting that in the energy range around -0.1 to -0.6 kcal/mol, two of the structures retain mainchain rigidity ($f_5 > 0.9$) while the other two have already dropped to $f_5 < 0.5$. This means that there is a change in the number of α-carbons belonging to rigid clusters and in the



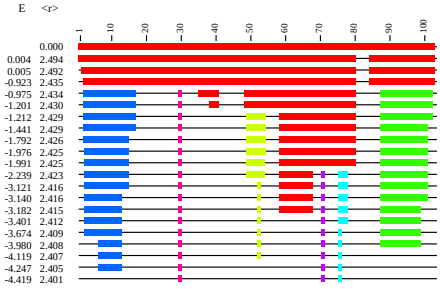
(a) 5CYT: normal ferriCytochrome



(b) 1I55: crystallised from 2Zn:1Fe mix



(c) 1I54: crystallised from 1Zn:2Fe mix



(d) 1LFM: with Co replacing Fe

Figure 3.3: Rigidity dilutions for four forms of tuna Cytochrome-c crystallised with different metal ion content in the heme groups. (a) normal Fe, (b) from a mixture with 2Zn:1Fe, (c) from a mixture with 2Fe:1Zn, (d) with Co.

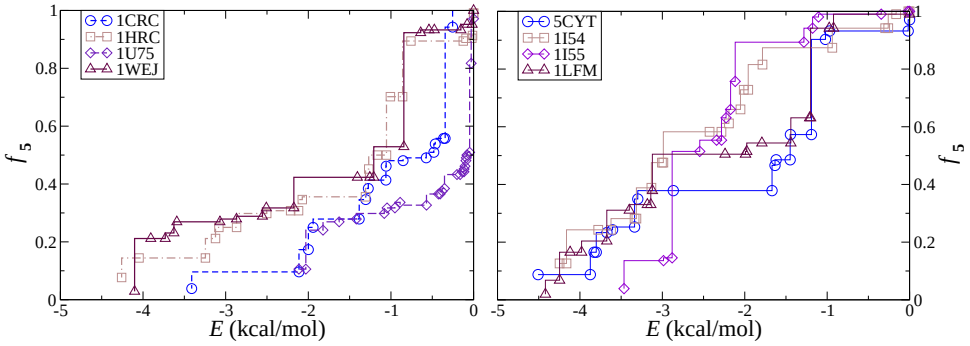
E_{cut}, which indicates that the network of bonds has also changed so that there is a new distribution of rigid and flexible regions that can limit or define new dynamical properties of the protein.

3.3.2 Effects of metal binding in protein rigidity

I now consider the rigid cluster decomposition graphs in Figure 3.4 of the mitochondrial Cytochrome-c structures (from tuna) which differ only in their heme-group metal content and are structurally very similar, see RMSD values in Table 3.3. The

From\To:	5CYT	1I55	1I54
1I55	0.27	—	—
1I54	0.2668	0.041	—
1LFM	0.286	0.116	0.087

Table 3.3: RMSD (Å) deviation for α-carbon positions among four tuna Cytochrome-c structures, showing the similarity of the structures.



(a) Horse cythochromes (b) Tuna cythochromes

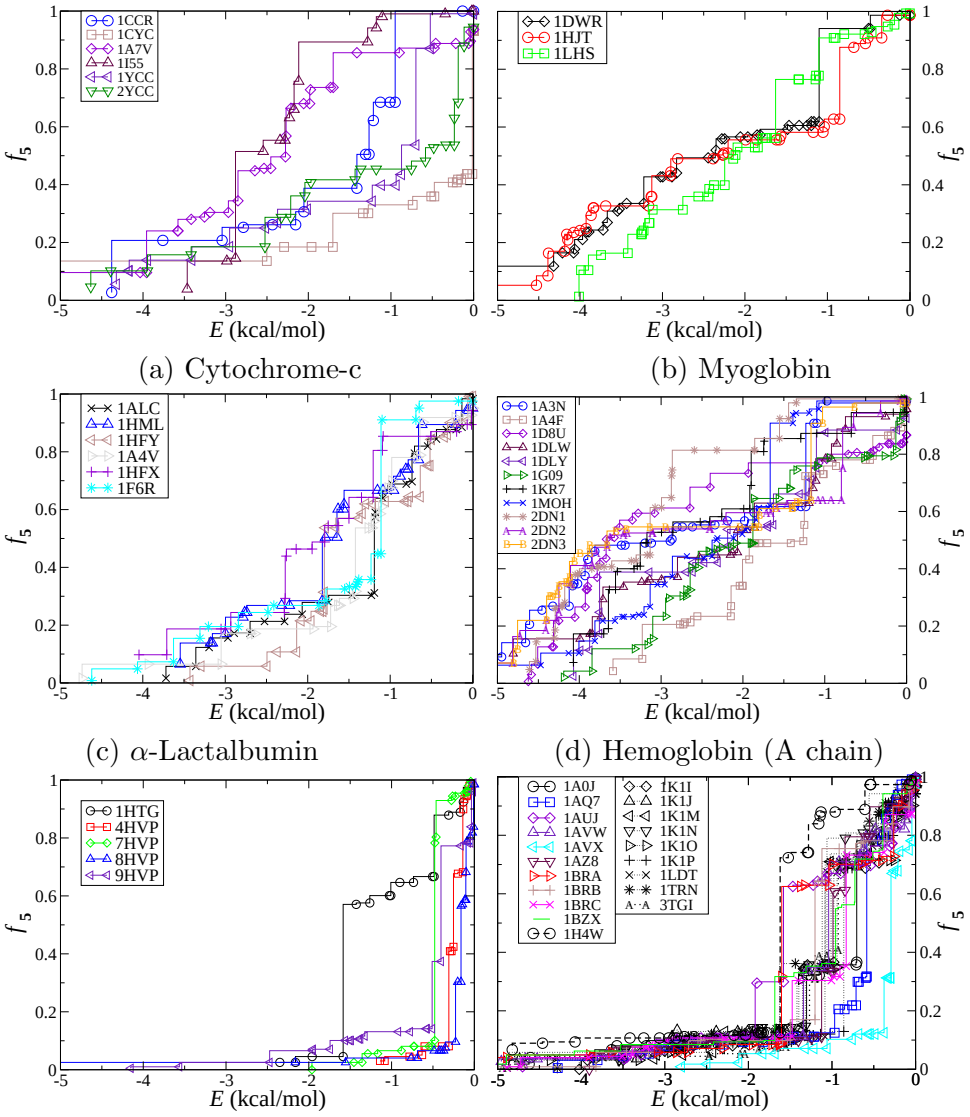
Figure 3.4: (a) Mainchain rigidity as a function of hydrogen bond E_{cut} during dilution for four horse mitochondrial Cytochrome-c structures. Note that for E_{cut} values in the region of -0.5 kcal/mol, structure 1HRC and 1WEJ are almost completely rigid while structures 1U75 and 1CRC are less than 50% rigid. (b) Mainchain rigidity for four tuna Cytochrome-c structures. Note the considerable differences in behaviour between, for example, 5CYT and 1I55 in the -1 to -2 kcal/mol energy range, even though the structures differ from each other only slightly.

tuna dilution plots have similar shapes for the structures crystalised with different metals bound into the heme group.

There are also differences, in particular, in structure 1I54 the α-helical region at residues 60–70 remains rigid to lower E_{cut} values then that at residues 90–100, which would disagree with the "folding core" prediction of reference [44]. The mainchain rigidity as a function of E_{cut} graph shows the differences in the energy scales at which rigidity is lost, (Figure 3.4b). The greatest discrepancy appears in the energy range from -1 to -2 kcal/mol; here the 5CYT structure has $f_5 \simeq 0.4$ while 1I55 has $f_5 \simeq 0.9$, although the structures differ by less than $d = 0.3$Å in α-carbon RMSD.

3.3.3 Patterns of rigidity loss

Having established that the FIRST E_{cut} is effective in separating stronger from weaker hydrogen-bond constraints, it seems sensible to step back and consider what may be said about the pattern of rigidity loss during dilution. Figure 3.5 shows the patterns of rigidity loss for six different families of proteins as listed in Table 3.1. There are two classes, those proteins displaying a gradual pattern of rigidity loss (Figure 3.5a,b,c,d for proteins (a) Cytochrome-c, (b) myoglobin, (c) lactalbumin, and (d) hemoglobin); and those displaying a sudden loss of rigidity (Figures 3.5e,f,g for proteins (e) HIV-1 protease and (f) trypsin). For proteins in this second class,



(a) Cytochrome-c (b) Myoglobin

(c) α-Lactalbumin (d) Hemoglobin (A chain)

(e) HIV-1 protease (f) Trypsin

Figure 3.5: Rigidity dilutions for different families of proteins: Cytochrome-c, myoglobin, α-lactalbumin, hemoglobin, HIV-1 protease and trypsin. Hence, proteins can display either a "gradual" (a–d) or a "sudden" (e–f) pattern of rigidity loss.

all the 25 structures that are examined in this study display sudden loss of mainchain rigidity. As a given bond is removed the mainchain rigidity drops significantly and in most cases it becomes almost entirely flexible once the E_{cut} is reduced below -2 kcal/mol. This indicates that the rigidity of clusters in these proteins is due to weaker hydrogen bonds.

The rigidity distribution of proteins that display a gradual pattern of rigidity loss indicates that there is a continuum range of bond strength across the protein structure. The 'smooth' rigidity transition is well illustrated by the rigidity dilution of Myoglobin in Figure 3.5b, where the removal of a single bonds only provoke small losses of rigidity compared with proteins of the second class.

Comparison of these six protein families thus leads us to the conclusion that protein structures, like glassy networks, can display two distinct patterns of rigidity loss depending on the diversity of their constraint networks. There are two families of proteins, HIV protease and trypsin, whose members display rapid loss of rigidity as weaker hydrogen bonds are eliminated. In contrast, four other families of proteins display a gradual loss of rigidity indicating a gradual hierarchy of hydrogen-bond strengths that constraint and maintain protein rigidity.

Next, I review in more detail the rigidity dilution pattern of one protein for each of the above mention classes.

3.3.4 Cutoff values in previous studies using FIRST

Hespenheide et al. [44] identified the protein folding core with "the set of secondary structure that remain rigid the longest in the simulated denaturation", without regard to the exact values of the E_{cut} at which rigidity is lost. In considering the rigidity of virus capsid protein complexes, Hespenheide et al. [43] make use of a E_{cut} of -0.35 kcal/mol, a value chosen so that capsid protein dimers would be flexible while the inner ring of proteins in a pentamer of dimers would be rigid, and draw conclusions about the rigidity of other multimeric complexes. Meanwhile, Hemberg et al. [49] use a different E_{cut} of -0.7 kcal/mol in a study on the dynamics of capsid assembly.

Jacobs et al.[17] indicated that that E_{cut} should be at least -0.1 kcal/mol in order to eliminate a large number of very weak hydrogen bonds in the range $E_{cut} = 0.0$ to -0.1 kcal/mol and that a natural choice of the E_{cut} near the "room temperature" is located around $E_{cut} \simeq -0.6$ kcal/mol.

The FRODA geometric simulation algorithm [30] makes use of the RCD generated by FIRST as a coarse-graining. Simulations of protein mobility using FIRST/FRODA have tended to use E_{cut} values that are systematically lower than in

applications of FIRST alone; typically −1.0 kcal/mol or lower [30, 50, 51, 52, 53], as E_{cut} values closer to zero seem to include too many constraints to allow large-scale motion to occur. In a paper on the combination of rigidity analysis and elastic network modelling, Gohlke et al. [39] discuss RCDs of two protein crystal structures but do not specify a E_{cut} value, though the FRODA mobility simulations given in Figure 3a of that paper were performed using a E_{cut} of −1.5 kcal/mol and give an excellent match to experimental data from NMR ensembles.

In this chapter I have shown that structural variations can alter the ridigity dilution patterns and E_{cut} significantly and that therefore it is advisable that the choice of the E_{cut} is done on a case by case basis.

3.3.5 Secondary structure motifs and rigidity distribution

It was previously noted in section 3.3.1 that the portions of the crystal structure that retain rigidity longest during dilution are generally α-helical. The HIV-1 protease and trypsin structures consist almost entirely of β-sheet structure, in contrast to the other four families in our study all of which have mostly α structure. This difference of pattern in rigidity loss between mostly α-helical and mostly β-sheet structures has not, it seems, been explicitly remarked on in the FIRST literature.

A third possible pattern would be a sudden loss of rigidity mediated by stronger hydrogen bonds, i.e. persistence of near-complete mainchain rigidity down to much lower E_{cut} values, but this behaviour is not observed among the sample of proteins selected.

3.4 Conclusions

The first motivation in analysing and comparing the rigidity analysis of protein families crystallised under different conditions was to determine the robustness of the rigidity distribution and of the E_{cut} against relatively small structural variations. There is a considerable variation in the RCDs and E_{cut} values of structurally similar proteins, i.e. with similar RMSD values, see Tables 3.2 and 3.3. Figure 3.4, for example, shows that among a group of Cytochrome-c structures drawn from similar eukaryotic mitochondria, E_{cut} in the range from 0.0 to −2.0 kcal/mol (such as have typically been used for FIRST/FRODA simulations of flexible motion [30, 50, 51, 52]) can produce a wide range of degrees of main-chain flexibility. Different crystallisation conditions promote new bonds to the structure and/or change bond strength to form a network of bonds that can have unique characteristics. Therefore the RCD and energy E_{cut} values in each case could be very different. On this point it is possible to

conclude that the results of rigidity analysis on individual crystal structures should not be over-interpreted as being "the" RCD for a protein. Hence, while FIRST successfully divides weaker from stronger bonds, it is not possible to identify a unique value of the hydrogen bond E_{cut} which can be applied to all protein structures to give meaningful results. Rather, each protein structure should first be subjected to rigidity dilution to produce a dilution plot; a suitable value of the E_{cut} can then be chosen to test a specific hypothesis about the rigidity and flexibility of the protein.

The second motivation was to obtain an insight into the similarities between the patterns of rigidity loss during hydrogen-bond dilution of proteins. The results show that proteins can display either gradual (second-order-like) or sudden (first-order-like) patterns of rigidity loss during dilution. Sudden rigidity loss is found in two proteases, eukaryotic trypsin and viral HIV-1 protease. Both consist largely of β-sheet secondary structure with little α-helical content compared to the other proteins in our set, which may account for their different rigidity behaviour. The results reveal that the two distinct patterns of rigidity transition recently identified in glassy networks [47] are also seen in proteins.

Chapter 4

Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses

In this chapter I report the results of systematically investigating protein motion. A set of 6 proteins covering a range of sizes and structural characteristics are selected and using the HCG I determine their conformational motions and characterise their motion in terms of three measures, dot product, RMSD and extended RMSD. Hereby, I systematically explore the consistency of the trajectories, the amplitude and type of motion for a set of structurally very different proteins. The results show that it is possible to characterise protein structural mobility and that some measures are more suited to defined such motion.

4.1 Introduction

It is a common goal in biophysics to represent the flexibility of a protein and study its large-scale motion without incurring the full computational cost of MD simulations. Different levels of simplification can be combined in multi-scale methods [54, 55]. Here, we shall consider three such methods in particular: (a) The pebble-game rigidity analysis implemented in FIRST [17] which provides valuable information on the distribution of rigid and flexible regions in a structure [48]; (b) NMA [8, 22, 18, 10] of a coarse-grained elastic network model (ENM), implemented in ELNEMO [25, 36], generates eigenvectors for low-frequency motion which are po-

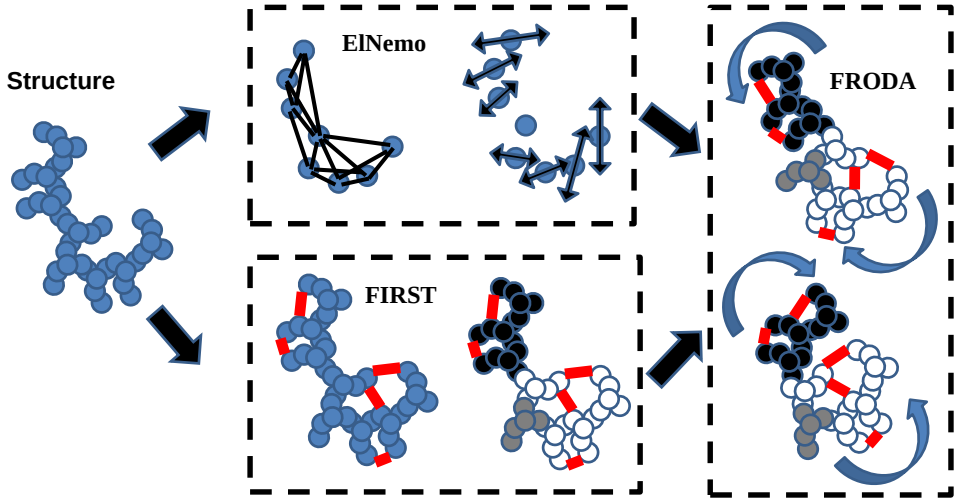


Figure 4.1: Schematic of the geometric simulation method. The input (at left) is an all-atom protein structure. Normal mode analysis (above) models the protein with a one-site-per-residue coarse graining and a simple spring model to produce an eigenvector for low-frequency motion. Rigidity analysis (below) identifies non-covalent interactions in an all-atom model of the protein and divides the protein into rigid clusters and flexible linkers. Geometric simulation (right) integrates normal-mode and rigidity information to explore the flexible motion of the protein.

tential sources of functional motion and conformational change [26, 27, 28, 29, 56]; (c) geometric simulation in the FRODA algorithm [30] uses rigidity information and explores flexible motion [50, 51].

The information from the rigidity analysis is used to coarse grain the structure and create pseudounits for the most rigid clusters. The eigenvectors from the NMA are used to bias the motion of the structure. These data is used by FRODA to generate new conformers that move along the chosen normal mode while maintaining rational bonding and sterics. The method is outlined schematically in Figure 4.1.

The method is tested with a set of six proteins of various sizes, from 58 to 1605 residues. The results show that it is possible to explore protein motion for large amplitudes in a few CPU-minutes.

Protein	PDB	Resolution	Residues	E_{cut} (kcal/mol)
BPTI	1BPI	1.1Å	58	-0.2, $\mathbf{-2.2}$
Cytochrome-c	1HRC	1.9Å	105	-0.7, $\mathbf{-1.2}$
Kinesin	1RY6	1.6Å	360	-0.4, $\mathbf{-1.1}$
α1-antitrypsin	1QLP	2.0Å	394	-0.1, $\mathbf{-1.1}$
PDI	2B5E	2.4Å	504	-0.015, $\mathbf{-0.522}$, -1.412
pLGIC	2VL0	3.3Å	1605	-0.4, $\mathbf{-0.5}$

Table 4.1: The proteins and specific structures selected for this study. For each protein, various E_{cut} values (kcal/mol) were chosen on the basis of rigidity analysis shown in Figure 4.3. This Table presents those values used in the simulations of motion that are presented in the main text. Bold values for E_{cut} have been used to compute the xRMSD.

4.2 Methods

4.2.1 Protein selection

Six proteins were selected for analysis that are diverse in function, structural characteristics and size, ranging from 58 to 1605 residues. Each selected protein is a representative high-resolution structure from the PDB [42]. The proteins and their PDB codes are listed in Table 4.1 and their structures are shown in Figure 4.2 with colour coding according to the results of rigidity analysis.

Bovine pancreatic trypsin inhibitor (BPTI) is a small well-studied protease inhibitor of 58 amino acids, comprising mainly random-coil structure plus two antiparallel β-strands and two short α-helices; the protein has only a small hydrophobic core, but is additionally stabilized by 3 intra-chain disulphide bonds [57, 58]. Mammalian mitochondrial cytochrome-c is a classic electron-transfer protein containing a redox-active haem group bound within a primarily α-helical protein fold. These two were selected as contrasting small proteins.

As medium size proteins we selected α1-antitrypsin and the core catalytic domain of the KinI motor protein kinesin [59]. The former is a protease inhibitor of the serpin family [60] which operates via a "bait" mechanism comparable to that of a mouse-trap, involving a very significant conformational change, whereas the latter is a mechanochemical device that transduces the chemical energy of ATP hydrolysis into mechanical work, specifically the depolymerisation of microtubules in the case of this particular kinesin. Both these proteins comprise an extensive β-sheet core flanked by several α-helices.

Protein disulphide-isomerase (PDI) is a large protein (with more than 500

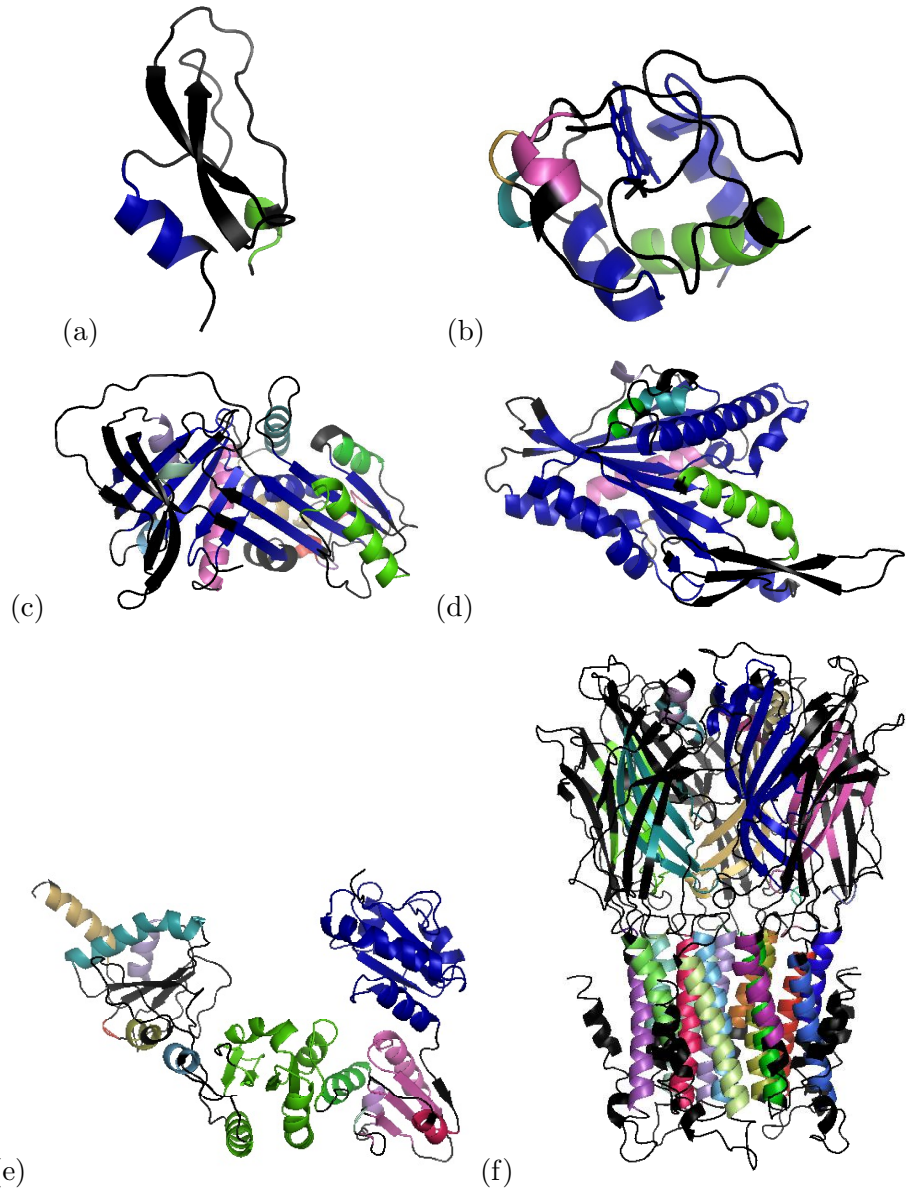


Figure 4.2: Tertiary structure of all six protein structures (a) BPTI (1BPI), (b) cytochrome-c (1HRC), (c) α1-antitrypsin (1QLP), (d) kinesin (1RY6,) (e) yeast PDI (2B5E) and (f) pLGIC (2VL0). The structures are given in standard PYMOL [41] format but broken into rigid clusters according to the rigidity analysis (see Figures 4.3 and 4.4) at the specific values of E_{cut} shown in Table 4.1. Each rigid cluster is represented in a different colour with the largest rigid cluster indicated in blue and flexible regions shown in black.

residues) comprising 4 distinct domains each with a thioredoxin-like fold, connected by two short and one longer linker [61]; the protein has both redox and molecular chaperone activity and intramolecular flexibility is essential for its action in facilitating oxidative folding of secretory proteins [62, 63]. The largest protein selected is an integral membrane protein (a bacterial protein of 1605 residues) that operates as a pentameric ligand-gated ion channel (pLGIC); it comprises an extracellular — mainly β-sheet — domain and a membrane-embedded domain, mainly comprising α-helices which form the lining of the ion-channel; the mechanisms of ion permeation and channel gating are not yet completely understood but it is clear that a conformational change is required for function [64].

4.2.2 Rigidity analysis and energy cutoff selection

Although the rigidity analysis procedure is explained in full detail in chapter 2.1.3, it is worth mentioning again the basic procedure to understand better how rigidity analysis links with protein mobility. First, the hydrogen atoms absent from the PDB X-ray crystal structures are added using the software REDUCE [40], the alternate conformations are removed and the hydrogen atoms are re-numbered in PYMOL [41]. This produces usable files for FIRST rigidity analysis. A RCD plot [17] is produced for each protein like the one presented in Figures 4.3 and 4.4. The plots show the dependence of the protein rigidity on E_{cut}, which determines the set of hydrogen bonds to be included in the rigidity analysis. An example of the tertiary structures with the residues coloured by the rigid clusters they belong is shown in Figure 4.2.

Previous studies [17] suggested that E_{cut} should be at least -0.1 kcal/mol in order to eliminate a large number of very weak hydrogen bonds, and that a natural choice is near the 'room temperature' energy of -0.6 kcal/mol. In a recent publication, we have further discussed the criteria for a robust selection of E_{cut}, showed that this criteria is not sufficient to avoid sensitivity to protein-specific structural variations. We argued that E_{cut} must be chosen on a case-by-case basis so as to test specific hypotheses about the rigidity of a particular structure [48].

Several E_{cut} listed in Table 4.1 are selected to explore flexible motion under different bond constraint conditions. A higher E_{cut} increases the number of constraints included in the simulation, and this is expected to restrict protein motion. We have used in each case at least one E_{cut} at which the protein is largely rigid (in the range -0.1 kcal/mol to -0.7 kcal/mol) and at least one lower E_{cut} at which the protein is largely flexible (in the range -0.5 kcal/mol to -2.2 kcal/mol).

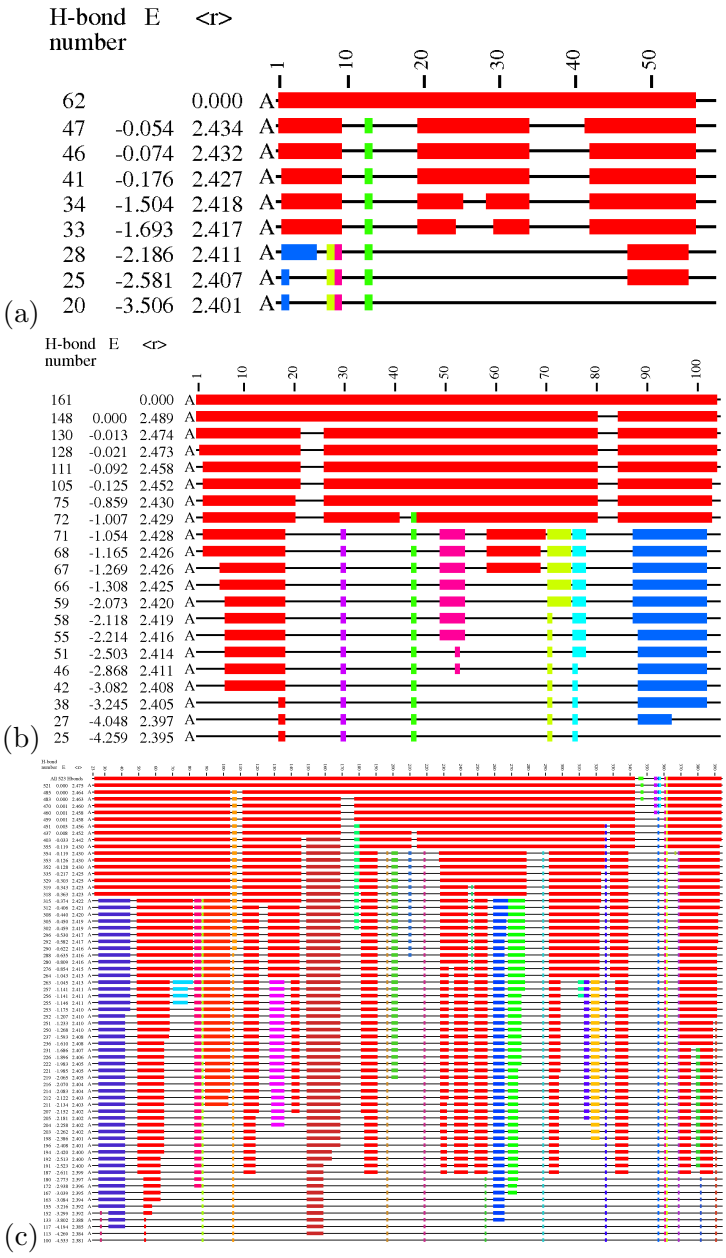


Figure 4.3: Rigid cluster decomposition graphs for: (a) BPTI (1BPI) (b) cytochrome-c (1HRC) and (c) α1-antitrypsin (1QLP). The x-axis represents the protein backbone and the y-axis the energy, E_{cut}, of the last hydrogen bond, which after being removed provokes a change in the rigidity distribution. Each line represents the new rigidity distribution of the polypeptide chain induced by removing a bond which alters the previous rigidity configuration. The residues belonging to rigid clusters are coloured — with the biggest rigid cluster coloured in red, whereas the flexible regions are shown as thin black lines. We choose the E_{cut}'s defining the number of rigidity constraints using the RCD plots.

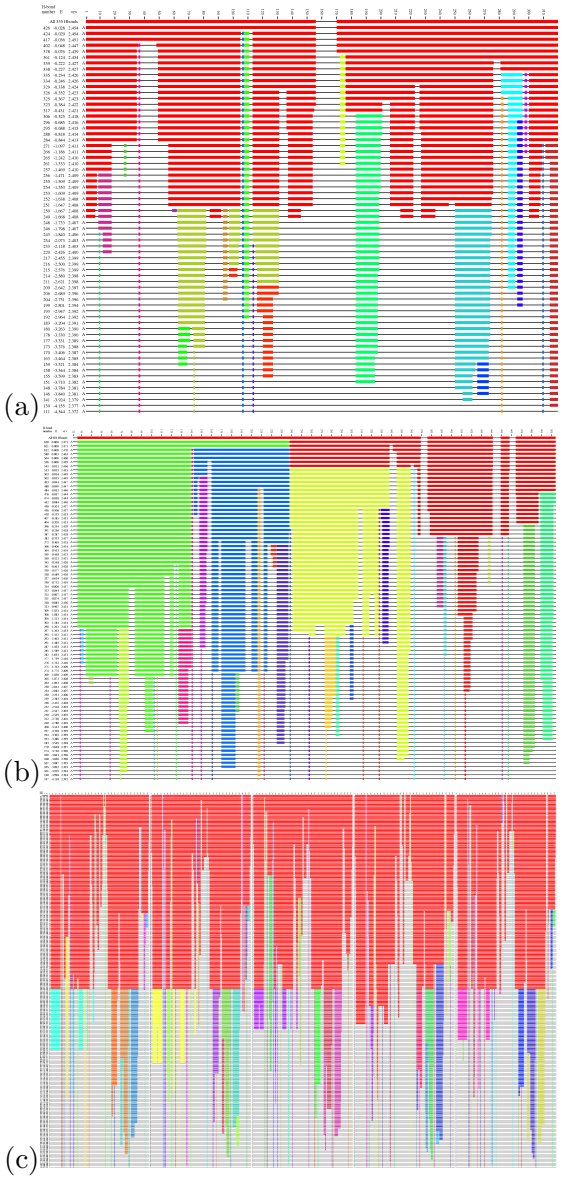


Figure 4.4: Rigid cluster decomposition graphs for: (a) internal kinesin motor domain (1RY6) (b) yeast PDI (2B5E) and (c) pLGIC (2VL0). The x-axis represents the protein backbone and the y-axis the energy, E_{cut}, of the last hydrogen bond, which after being removed provokes a change in the rigidity distribution. Each line represents the new rigidity distribution of the polypeptide chain induced by removing a bond which alters the previous rigidity configuration. The residues belonging to rigid clusters are coloured — with the biggest rigid cluster coloured in red, whereas the flexible regions are shown as thin black lines. We choose the energy E_{cut} defining the number of rigidity constraints using the RCD plots.

4.2.3 Normal modes of motion

The normal modes of motion are computed using the ENM [18] implemented in ElNemo software. This generates, for each protein, a set of eigenvectors and associated eigenvalues obtained from performing NMA analysis on the protein structure as explained in sections 1.2 and 2.7. Here we consider the five lowest-frequency non-trivial modes, that is modes 7 to 11 for each protein. We will denote these modes as m_7, m_8, ..., m_{11}. These eigenvectors are predicted on the basis of a single protein conformation and the amplitude to which a mode can be projected will be limited by bonding and/or steric constraints.

4.2.4 Mobility simulations

All the simulations were carried out using the normal modes defined by ElNemo and the First software capabilities. One of the First software command options, named Froda, allows to apply a mode eigenvector as a bias direction of motion [30, 38] and explore the flexible motion available to a protein. Another command option in the First software allows to introduce the protein rigidity distribution at a given cutoff energy as input in the simulation. Then, the residues part of a rigid cluster are treated as a single unit and projected along the chosen mode with minimal computer time and resources. New conformations are generated by applying a small random perturbation and the eigenvector directed biased motion simultaneously to all atomic positions; then First/Froda reapplies bonding and steric constraints to produce an acceptable new conformation. The capability to use a mode eigenvector as a bias has been reported previously [65, 21] and is illustrated schematically in Figure 4.1.

The α-carbons RMSD values between the aligned residues from the initial and current conformation are recorded for all the conformers generated. Furthermore, since Farrel et al. [38] reported several limitations in the Froda procedure we chose to carry out five parallel simulations for each structure, mode and direction of motion to have initial data to investigate such limitations.

The evolution of fitted RMSD during each run is illustrated in Figures 4.5, 4.6, 4.7 and 4.8.

The required input for all calculations is a `.pdb` format file containing an all-atom representation of the protein structure including hydrogen atoms, which we shall name `protein.pdb`. The procedure is to obtain a `.pdb` format file containing all-atom representation of the protein structure from the PDB; to remove alternate side chain conformations and nonbonded heteroatoms including water molecules; to

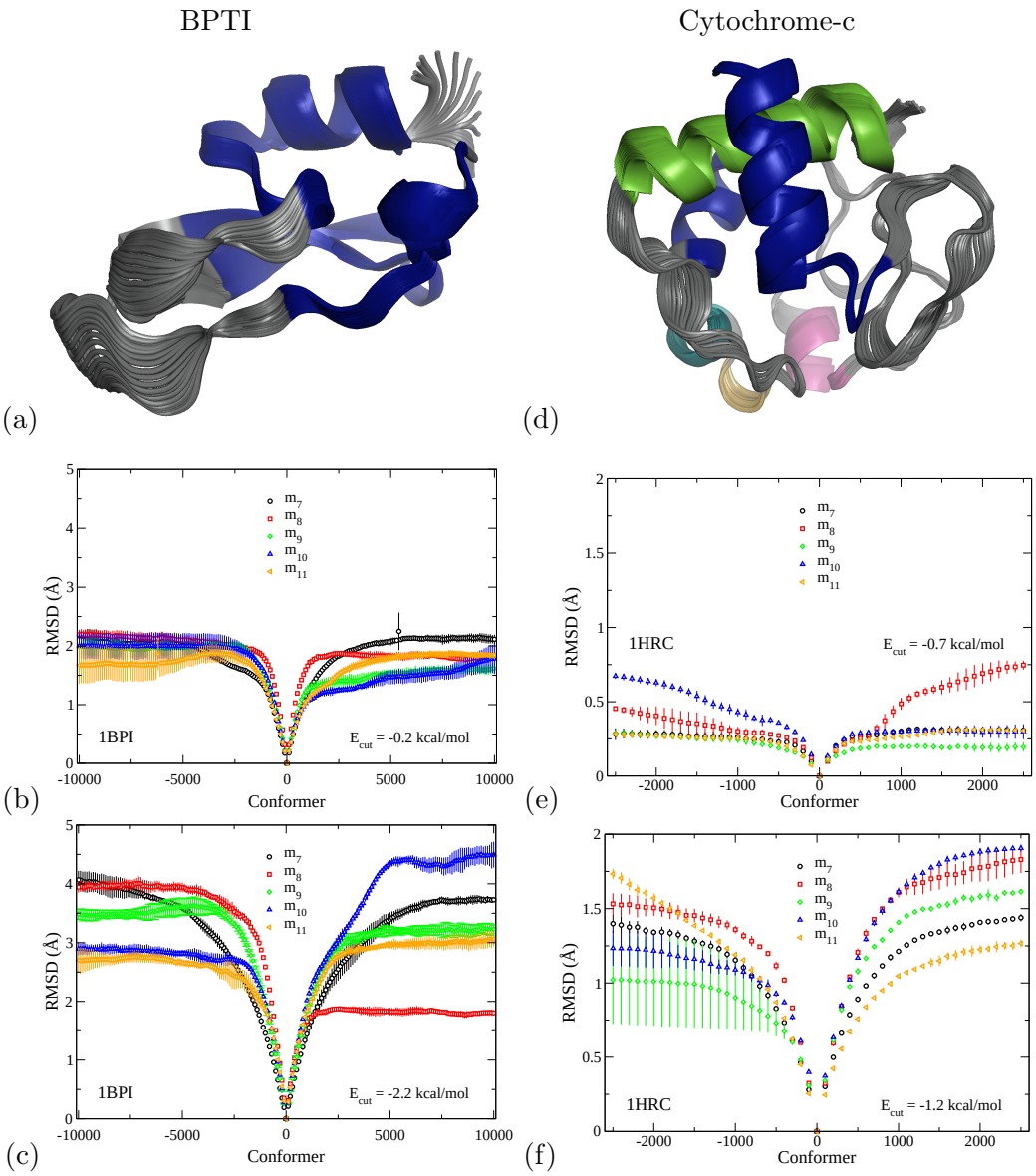


Figure 4.5: Superimposed structural variations and fitted RMSD for small loop motion as found in BPTI and cytochrome-c. Panels (a) and (d) indicate the range of projected tertiary structure for motion along mode m_7 at $E_{\text{cut}} = -2.2$ kcal/mol for BPTI and at $E_{\text{cut}} = -1.2$ kcal/mol for cytochrome-c, respectively. Panels (b,c) and (e,f) show the fitted RMSD as a function of FRODA conformations for BPTI (1BPI) and cytochrome-c (1HRC), respectively, for the non-trivial modes $m_7, \ldots, m_{11}$ at two values of E_{cut} as shown. Positive conformation values indicate motion along the direction of the corresponding ELNEMO mode, whereas negative conformation values indicate motion in the opposite direction. Points and error bars indicate mean and standard deviation obtained from five runs of the conformation generation for each mode.

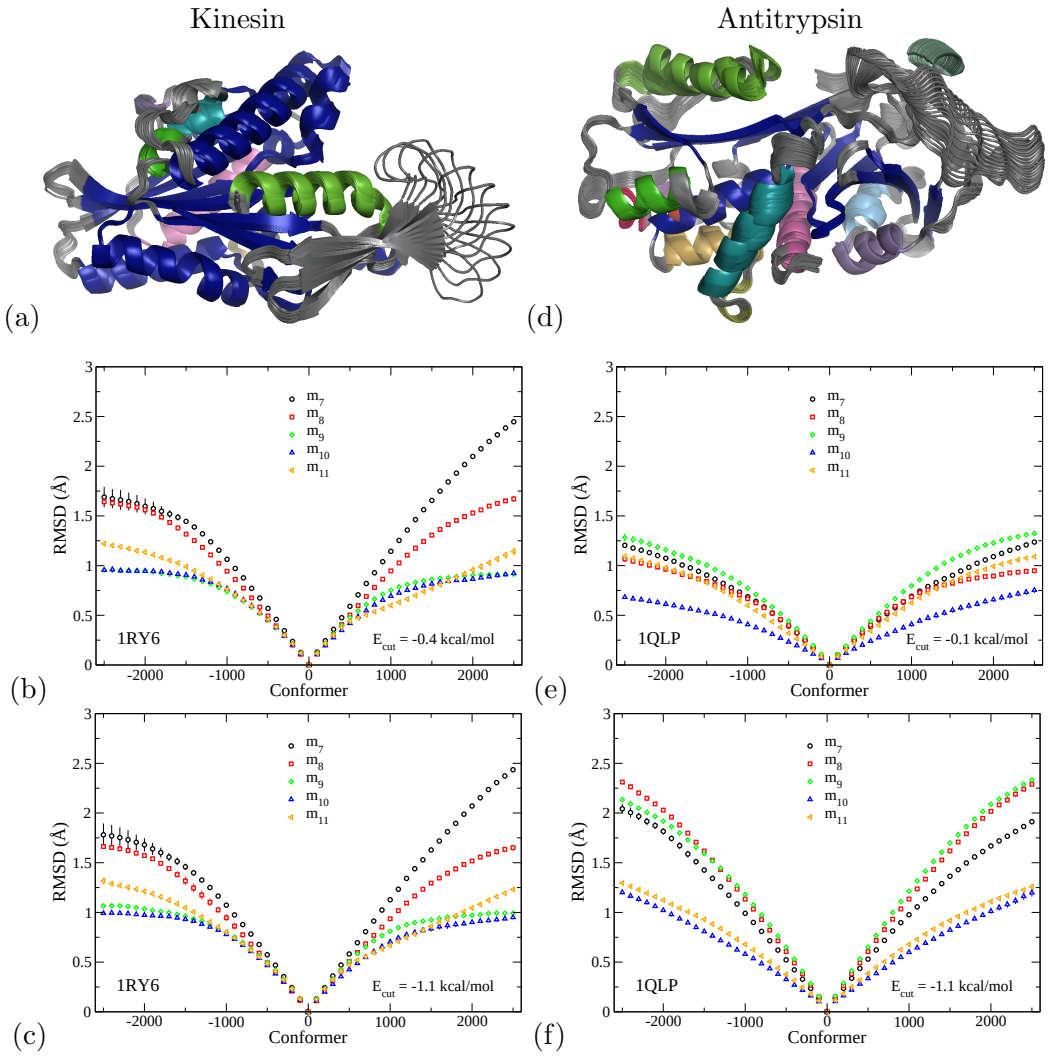


Figure 4.6: Superimposed structural variation and fitted RMSD for large loop motion as in (a) kinesin (1RY6) and (d) antitrypsin (1QLP) for $E_{\text{cut}} = -1.1$ kcal/mol. Panels (b,c) and (e,f) represent — as in figure 4.5 — the fitted RMSD at two values of E_{cut} for kinesin and antitrypsin, respectively. Points and error bars have been determined as in Figure 4.5.

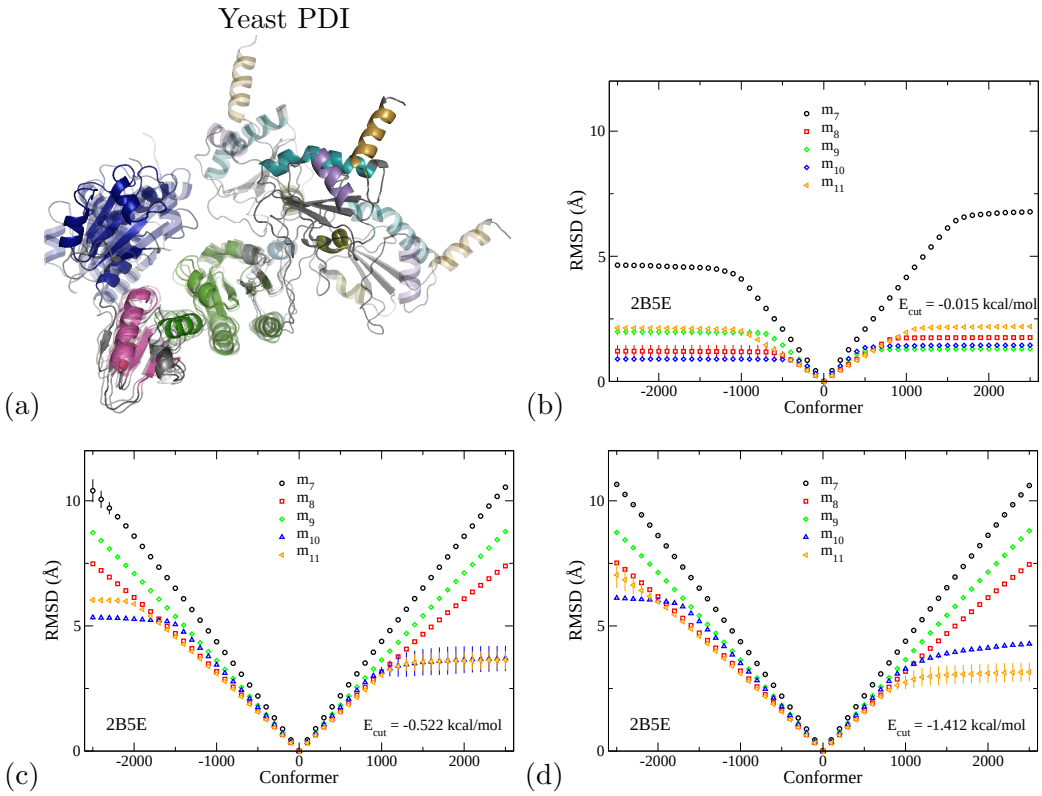


Figure 4.7: Superimposed structural variation of large domain motion and fitted RMSD for yeast PDI (2B5E). (a) We show the initial tertiary structure as opaque and the projected structures as partially transparent. All structures are aligned on the central two domains b–b' to highlight the motion of the a and a' domains. Motion represents large conformational change along m_7. Panels (b), (c) and (d) show the fitted RMSDs relative to the initial conformation for three values of E_{cut}. Points and error bars as in Figure 4.5.

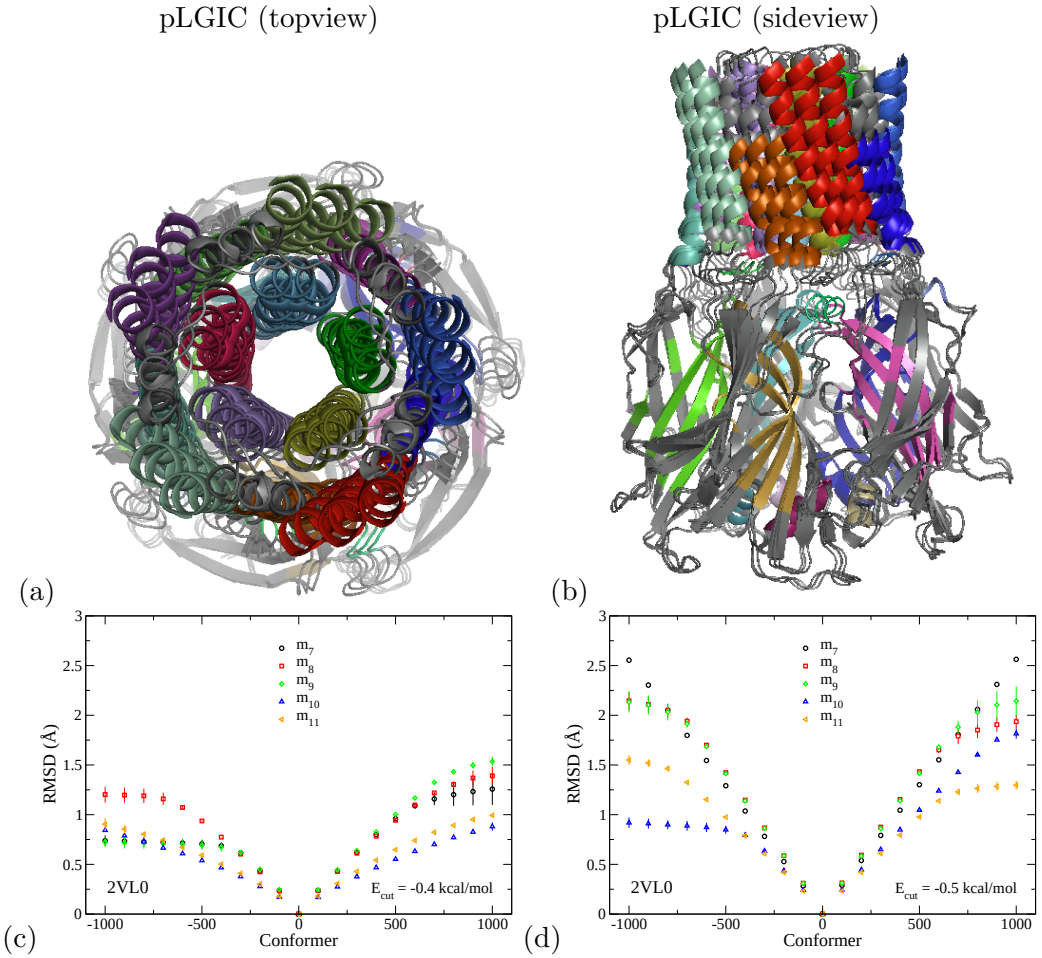


Figure 4.8: Large scale twist motion in a ligand gated ion channel (2VLO). (a) View down the transmembrane channel in its initial state and projected along m_7 in two directions. (b) Side view showing tilting of the helices during the motion. In both images the structures have been aligned on the extracellular β-sheet portion so as to highlight the relative motion of the domains, and residues from number 283 upwards in each chain are not shown to make the major helices visible. (c and d) Fitted RMSDs relative to initial conformation for low-frequency modes m_7, ..., m_{11} and two E_{cut}. Points and error bars as in Figure 4.5.

add hydrogens using the REDUCE software and to renumber the atoms sequentially using PYMOL.

Normal mode calculations were carried out using ELNEMO using the default setting of a 12Å E_{cut} in the spring network. The protein structure is given as consisting only of the α carbon from the all-atom structure. The output is a file that contains $3N$ mode eigenvectors for a protein of N residues. Each eigenvector is described with a mode number, a frequency, and N lines each giving a Cartesian vector; so that the ith line is the displacement to be applied to the ith residue. The vector is normalized so that the sum of the squares of all displacement vectors is unity.

Since the displacement from one conformation to the next is small, only the 100th conformation is recorded as the simulation runs continues for typically hundreds or thousand conformations. Each mode is projected to an amplitude of several Å in fitted RMSD (see section 4.2.5 for a detailed explanation on fitted RMSD) using a random component of 0.1Å and a directed component of 0.01Å. The run is considered complete when no further projection along the mode eigenvector is possible (due to steric clashes or bonding constraints) which manifests itself in slow generation of new conformations and poor reproducibility in the results of independent runs. The number of conformations selected to explore protein motion is 2500 for all proteins except pGLIC, for which 1000 conformations are sufficient to hit the slow conformer generation limit due to steric or bonding clashes. All the mobility simulations for each protein were performed at several selected values of E_{cut}, as shown in Table 4.1.

4.2.5 Raw vs fitted RMSD

The RMSDs reported in Table 4.2 are α carbon RMSDs from comparing the input structure to a generated conformation, obtained after least-squares fitting using the PYMOL `intra_fit` command. These values differ somewhat from the raw RMSD values reported by FRODA in its output files, which are calculated without any fitting being carried out. In particular, the fitted RMSD saturates once further motion along the mode direction is no longer possible due to steric clashes or limits imposed by covalent or noncovalent bonding constraints. The effects of fitting are illustrated in Figure 4.9a for mode m_7 of structure 1BPI. The raw RMSD values increase almost linearly during the generation of 10000 conformers, whereas the fitted RMSD values saturate for conformers from $\approx$ 5000 up to 10000. Conformers 5000 and 10000 differ by $\approx$ 3Å in raw RMSD but by only $\approx$ 0.8Å in fitted RMSD. Superposition of conformers 0, 5000 and 10000 with and without fitting, shown in

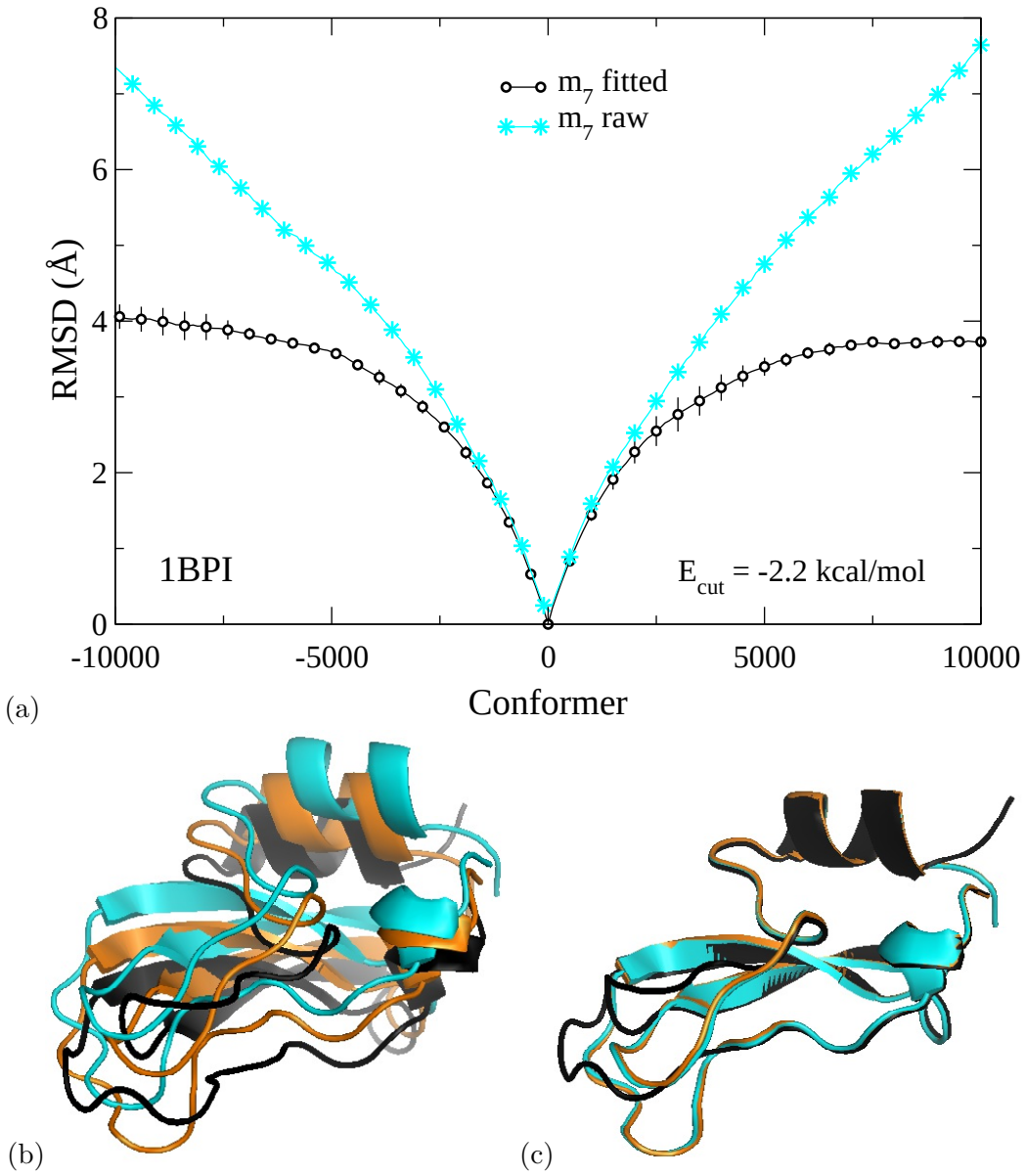


Figure 4.9: Raw vs fitted RMSD for BPTI. (a) Comparison between the RMSD values obtained before (raw RMSD) and after fitting the new conformers to the initial structure. We report the fitted and raw RMSD values of mode m_7 for every 100th conformer (only every 500th indicated by a symbol for clarity) and up to in total 10000 conformers at a E_{cut} of $E_{\text{cut}} = -2.2$ kcal/mol. The error bars denote the standard deviation obtained from including 5 different initial random perturbations to the guided motion, see section 4.2.4. The fitted RMSD values, also shown in Figure 4.5e for modes $m_7, \ldots, m_{11}$, saturate whereas the raw RMSD values increase linearly, due to FRODA weighting each residue by the number of atoms that it contains, thus producing a component of rigid body motion. Panels (b) and (c) show the superimposed structures for conformers 0 (black), 5000 (orange) and 10000 (green) of (b) raw and (c) fitted structures.

Protein	Residues	E_{cut} (kcal/mol)	RMSD pos	RMSD neg	xRMSD pos	xRMSD neg
BPTI	58	−2.2	2.62	2.66	152	154
Cytochrome-c	105	−1.2	1.44	1.40	151	146
Kinesin	360	−1.1	2.48	2.04	892	733
Antitrypsin	394	−1.1	1.91	2.04	753	804
PDI	504	−0.522	10.54	10.40	5314	5243
pLGIC	1605	−0.5	2.56	2.55	4113	4099

Table 4.2: Extensive RMSD (Å) values, maximum RMSD values and the E_{cut} chosen to calculate xRMSD for each protein based on the RCD graphs. For proteins that are expected to be rigid we have chosen a higher E_{cut} and for proteins with an expected conformational change we have chosen a more restrictive E_{cut}.

Figures 4.9b,c, show that conformers 5000 and 10000 are indeed very similar to each other. The raw RMSD is greater than the fitted RMSD and tends not to saturate, but rather to continue to increase slowly, once the motion is effectively jammed. The reason for this different behaviour is a small difference in the statistical weighting given to each residue by ElNemo and by First/Froda. In the ENM the normal modes are obtained for every α-carbon whereas in Froda, however, applies the bias to an all-atom representation of the structure; and thus the bias applied to a residue with many atoms affects the whole-body motion of the structure more than the bias applied to a residue with few atoms. The conformers generated by Froda therefore acquire a component of rigid-body translation and rotation, which increases the raw RMSD. Least-squares fitting to the input structure removes the rigid-body components.

Figure 4.9 shows the importance of using a fitted RMSD to identify and subtract correctly the rigid-body motion during a simulation. The raw RMSD values (in cyan) do not saturate, whereas the fitted RMSD values (in black) do. As presented in the Figure 4.9, the evolution of the fitted and raw RMSD values shows that raw RMSD values continue to increase linearly, hence accounting for network internal motion as well as spatial rigid-body translation as shown in Figure 4.9b. Whereas fitted RMSD values saturate as they account for network internal motion only, see Figure 4.9c. The good overlap of the conformers 5000 (orange) and 10000 (green) shown in Figure 4.9c correspond only to the minimal increase in fitted RMSD observed in Figure 4.9a beyond conformer 5000. That is, the conformer generation beyond conformer 5000 is slow and limited due to stereochemical constraints. The two structures overlap each other almost perfectly along the polypeptide chain, which indicates that there is little motion between the two conformers. However, the

flexible loop of BPTI have substantially moved for conformers 5000 and 10000 with respect to the initial structure, i.e. conformer 0 (shown in black). Hence, it is clear from this comparison that we account for the rigid-body motion effect by fitting all the conformers to the initial structure. Hence, fitting the protein structures before calculating RMSD values allows us to account and remove the rigid-body motion effect introduced by FRODA and identify real conformation change.

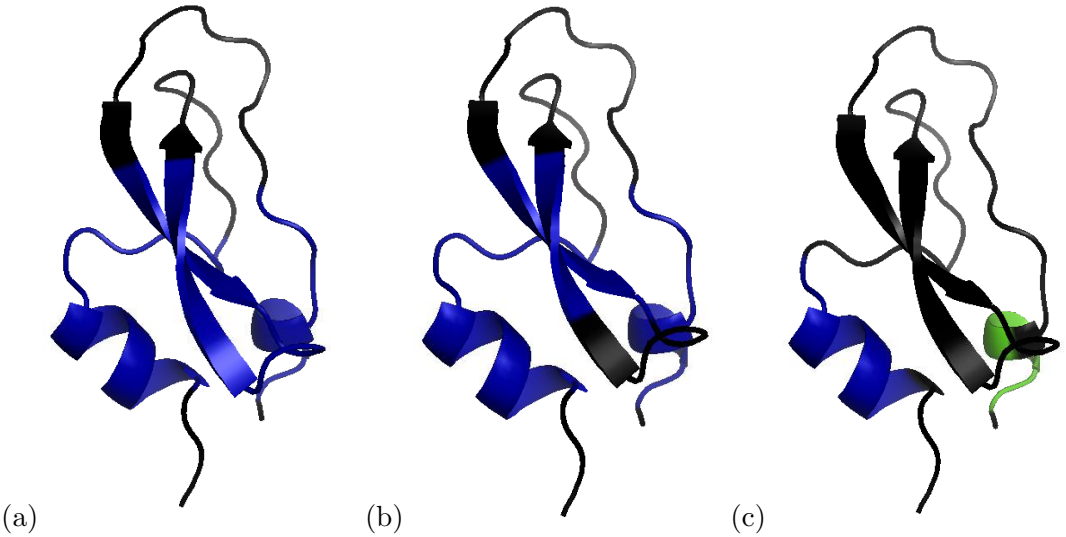
4.2.6 Monitoring the evolution of normal modes

It is implicit in normal-mode analysis of protein conformational change that a mode eigenvector should be a valid direction for motion over some non-zero amplitude. For example, Krebs et al. [20] have surveyed a large number of known conformational changes, using paired crystal structures, comparing the vector describing the observed conformational change to the low-frequency elastic network mode eigenvectors using a dot product. In many cases the observed change had a large dot product (> 0.5) with only one or two normal modes.

In each of our simulations we use an *initial* normal mode, $m_j^{(i)}$, as a bias throughout the simulation. We calculate a new set of *current* normal modes, $m_j^{(c)}$, for each newly generated conformation. We compute the dot product of the bias vector, $m_j^{(i)}$, with $m_j^{(c)}$, that is, the current normal mode with the same mode number j, as in $m_j^{(i)} \cdot m_j^{(c)}$. Graphs of these dot products are shown for all protein structures investigated here in the supplementary Figures 10.1–10.6.

NMA has been applied extensively to study protein motion [20, 28, 26, 29]. From the results of monitoring the evolution of normal modes presented in emerge three different motifs that characterise the evolution of the normal modes. The evolution of the dot product between the modes from the initial structure and the ones obtained from each conformer define the similarity between the modes. Although for some cases the dot product gives a rough indication of the type of motion the protein is undergoing, there is not a consistent description of protein motion in terms of dot product that correlates with the three types of motion investigated.

These results clarify that while the dot products provide useful additional information about the stability of the initial modes during the simulated motion, they are not simply correlated with loop or domain motion in contradistinction to the xRMSD.



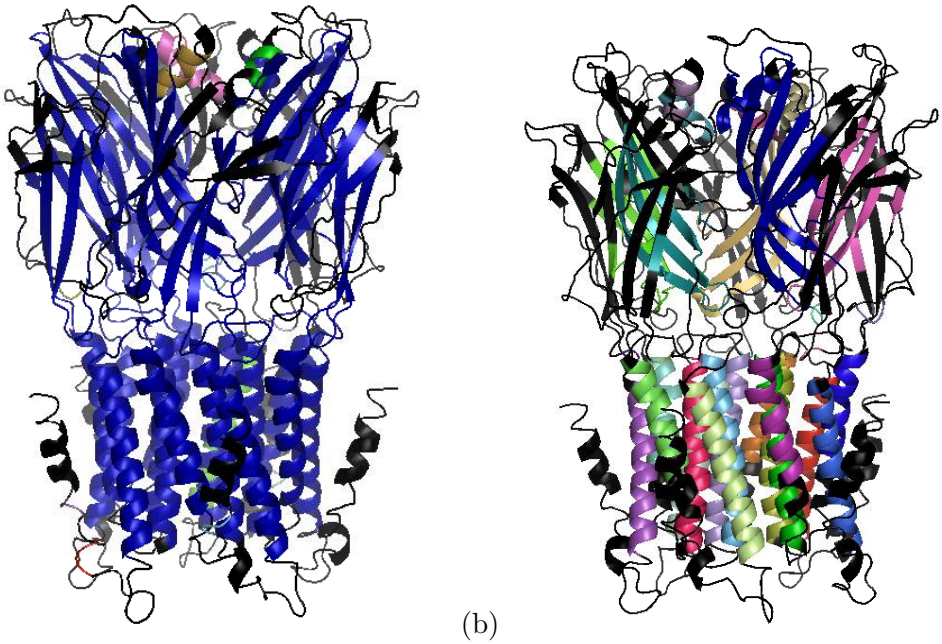
(a) (b) (c)

Figure 4.10: Tertiary structure of BPTI (1BPI). Colouring is defined using the rigidity analysis results shown in Figures 4.3 and 4.4. Flexible regions are illustrated in black whereas rigid residues are coloured as per the rigid cluster they belong to. The biggest rigid cluster is coloured in blue. The number and size of the rigid clusters vary depending on the chosen E_{cut} value, which for BPTI are (a) $E_{\text{cut}} = -0.2$ kcal/mol, (b) $E_{\text{cut}} = -1.7$ kcal/mol and (c) $E_{\text{cut}} = -2.2$ kcal/mol. Note that the colour code used to represent residues within the same rigid cluster is not the same in the RCD and in the tertiary structures.The biggest rigid cluster in the RCD graphs is noted in red and in the tertiary structures is noted in blue.

4.3 Results

The rigidity analysis of the protein structures shown in Figures 4.3 and 4.4 revealed some interesting features. The rigidity distribution and dilution is very different for each structure and highlights the structural difference among the selected proteins. For example,the rigidity dilution of BPTI shown in Figure 4.3a reveals that the rigidity of this protein is mainly represented by a single rigid cluster (in a red thick line) and by two flexible regions (in a black thin line). Figure 4.10 illustrates the rigidity distribution projected onto the tertiary structure of BPTI. The distribution of rigid and flexible regions corroborates previous experimental data about protein structures. For example, the RCD plot of BPTI structure correlates with biochemical studies that define the structure as a base with two disulphide bonds that rigidify the structure and two flexible loops [57, 58]. Besides yeast PDI rigidity distribution on which I will expand further in chapter 4.5, it is also interesting to mention that the pLGIC membrane protein shows a very unique rigidity behaviour. Its rigidity

dilution in Figure 4.4c shows a single rigid cluster (in red) which abruptly breaks up into several rigid clusters and flexible regions. This switch like rigidity transition corresponds to a first order transition as previously mentioned in chapter 3.4. The projection of the rigidity distribution on the tertiary structure for $E_{\text{cut}} > -0.4$ kcal/mol shows the rigidity distribution as a single rigid cluster (in blue) encompassing most of the protein structure, see Figure 4.11. Whereas at a lower E_{cut}, $E_{\text{cut}} < -0.5$ kcal/mol, the protein displays multiple rigid clusters linked by flexible regions. This suggests two potentially functional states divided by a switch like transition. In the first state, the structure's motion is constrained due to a single rigid cluster extending through most of the structure. Hence, the protein will not be able to achieve much internal motion beyond the mobility of the flexible regions. In the second state, however, the many rigid clusters are interlinked by flexible regions. This opens the door to a possible interdomain motion that accounts for protein functionality. Similarly, yeast PDI rigidity distribution shows that the rigid regions are located within the protein domains, except for very high E_{cut}, and that the intra-domain regions become flexible early in the rigidity dilution as shown in Figure 4.12.



(a) (b)

Figure 4.11: Tertiary structure of pLGIC (2VL0). Colouring of the tertiary structure is defined as in Figure 4.10 but with E_{cut} as (a) $E_{\text{cut}} = -0.4$ kcal/mol and (b) $E_{\text{cut}} = -0.5$ kcal/mol. Note that the protein appears to be rigid for $E_{\text{cut}} = -0.4$ kcal/mol and that there is a switch-like first order rigidity transition at $E_{\text{cut}} = -0.5$ kcal/mol which reveals the most flexible parts of the secondary structure which allow mobility.

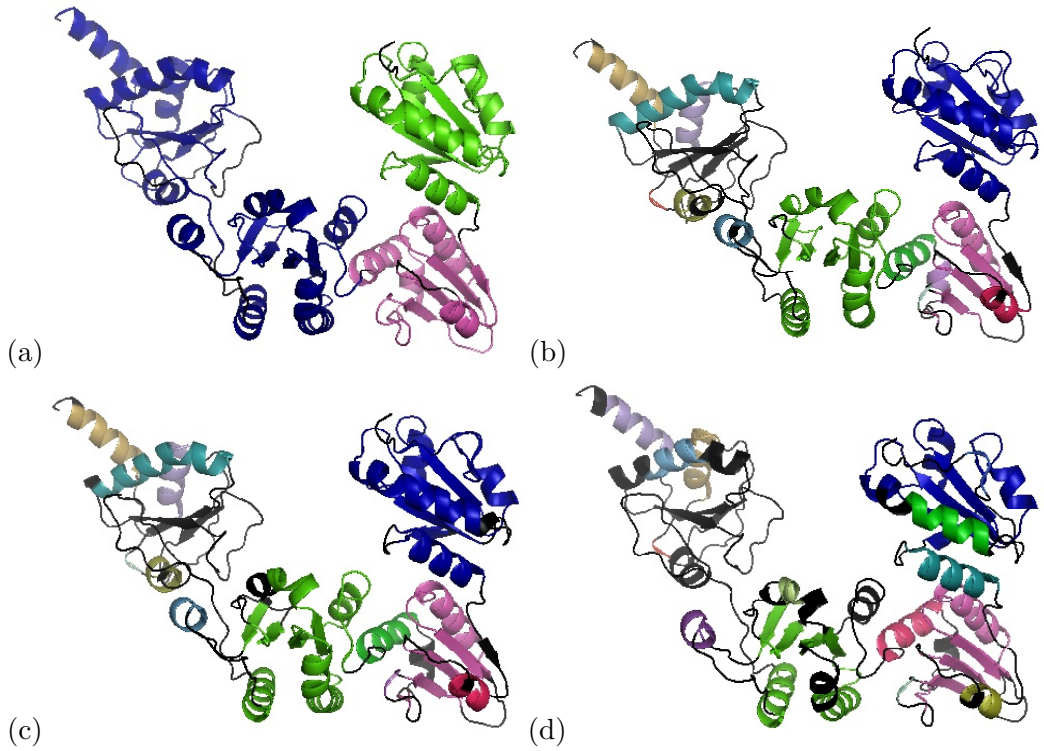


Figure 4.12: Tertiary structure of yeast PDI (2B5E). Colouring of the tertiary structure is defined as in Figure 4.10 but with E_{cut} as (a) $E_{\text{cut}} = -0.015$ kcal/mol, (b) $E_{\text{cut}} = -0.522$ kcal/mol, (c) $E_{\text{cut}} = -0.885$ kcal/mol and (d) $E_{\text{cut}} = -1.412$ kcal/mol. Note that colors are assigned according to cluster size which changes depending on the E_{cut}.

4.3.1 Tracking protein motion

The visual inspection of the superimposed tertiary structures (conformers) obtained from projecting the initial structure along a normal mode indicate that all the proteins move to some extend. In this section I present the results of tracking protein mobility using three different measures: scalar or dot product between initial set of normal modes and the subsequent conformers normal modes, RMSD and xRMSD between the initial structure and the conformers generated from the simulation.

4.3.2 Tracking protein motion: RMSD

RMSD is commonly used to compare structural similarity among protein structures. Hence I monitored the evolution of RMSD for each mode m_7, ...,m_{11}, and for at least two selected values of E_{cut}. The results are summarized in Figures 4.5-4.8c. In all cases, we observe an initial phase in which the RMSD increases almost linearly,

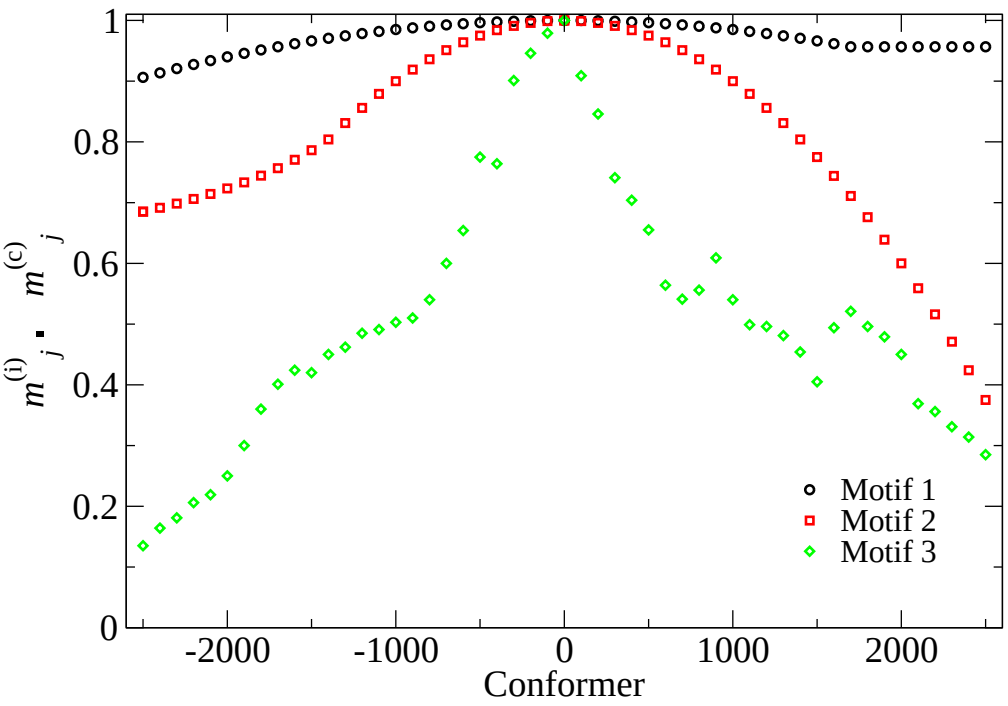


Figure 4.13: Dot product motifs. Schematic representation of the typical behaviours of the dot product of an initial mode with a current mode during projection along the initial mode eigenvector. Motif 1: gradual, nearly quadratic reduction in the dot product due to a progressive rotation of the current mode compared to the initial one. Eventual constant behaviour indicates that the motions has reached its amplitude limit. Motif 2: more rapid roughly quadratic reduction. Motif 3: sudden collapse of the dot product and the initial mode no longer resembles the current mode with the same mode number.

as the protein explores the mode direction without encountering significant steric or bonding constraints on the motion. During this phase, generation of new conformations is very rapid and the RMSDs from different runs are very similar to each other. The RMSD then displays an inflection, ceasing to rise linearly, and approaching an asymptote; this indicates that steric clashes and bonding constraints (such as hydrophobic tethers) are preventing further exploration along the mode direction. The asymptote is thus an amplitude limit on the mode. In this phase the generation of new conformations becomes slower as the fitting algorithm has increasing difficulty finding a valid conformation, and the RMSDs achieved by different runs differ. The mobility simulations for BPTI (1BPI) summarized in Figures 4.5a–c show the results of computing up to 10000 conformations and clearly illustrate the asymptotic behaviour.

In the regime of slow conformation generation, the mode bias is forcing the

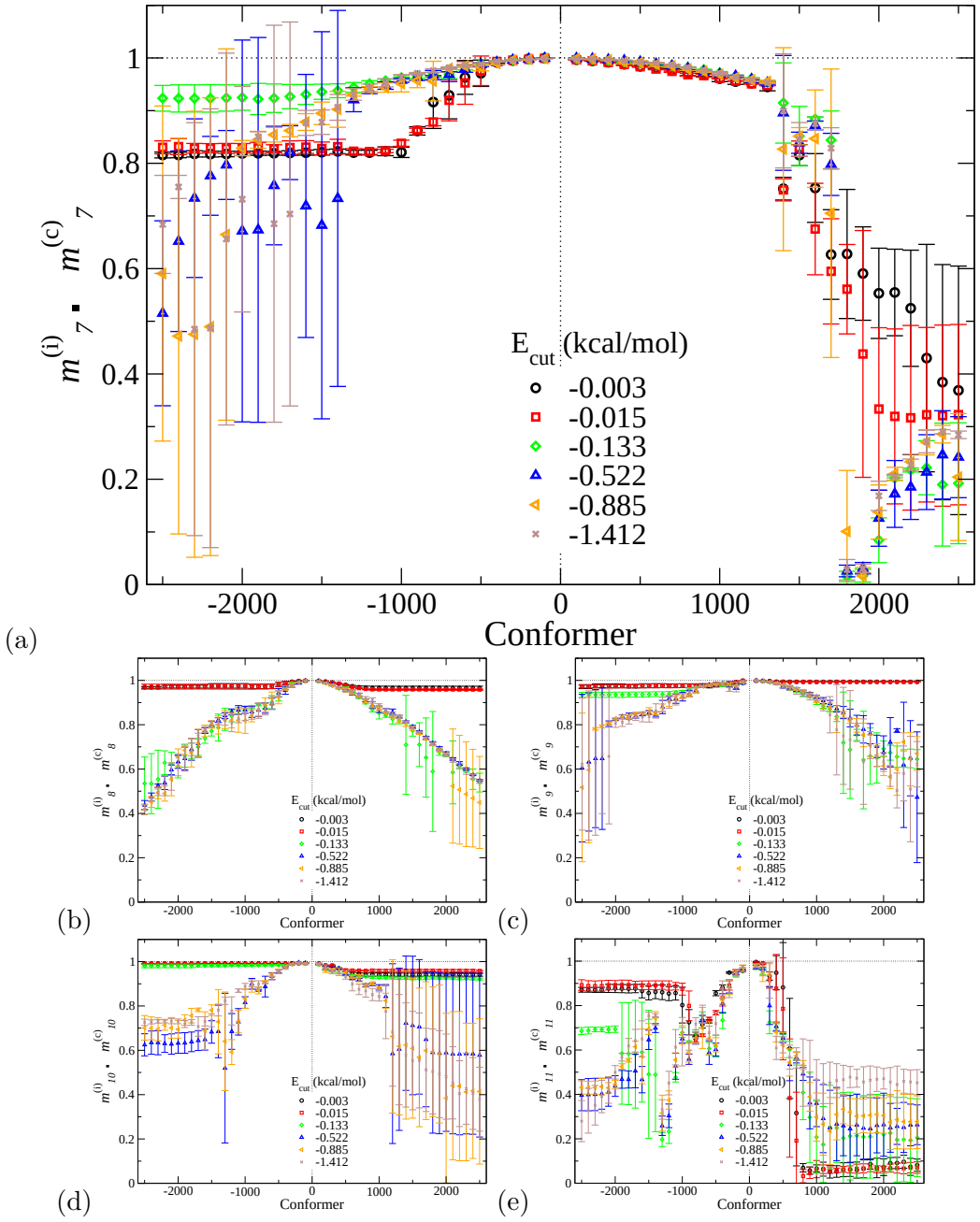


Figure 4.14: Dot product graph for yeast PDI (2B5E). The scalar product $m_j^{(i)} \cdot m_j^{(c)}$ between an initial starting mode $m_j^{(i)}$ and its current mode $m_j^{(c)}$, $j = 7, \ldots, 11$ as the initial structure is projected along the initial mode.The current modes, $m_j^{(c)}$, are obtained from performing normal mode analysis on the current conformations as the initial structure is projected along an initial mode $m_j^{(i)}$. For clarity, only the dot products for the first non trivial modes ($m_7^{(i)} \cdot m_{11}^{(c)}$), for a periodic sample of 25 conformations from the 2500 generated and for each direction of motion are shown. The evolution of the dot product along the conformations is reported for different E_{cut}.

structure into a regime of steric clashes and/or of bonding constraint limits, for example when residues connected by a hydrophobic tether are being pushed apart. The computing of RMSD data for larger proteins is therefore truncated once this "jamming" starts.

4.3.3 Tracking protein motion: Scalar product

The results of investigating and characterising protein mobility by computing the scalar or dot product for all the normal modes are summarised in Figures 4.14 and 10.1-10.6. These results reveal three main behaviours that I summarise schematically in three motifs, see Figure 4.13. In motif 1, the dot product remains close to 1 throughout the simulation, indicating that the initial $m_j^{(i)}$ and current $m_j^{(c)}$ modes remain very similar. In this case the motion of the protein is not introducing significant changes to the elastic network and the mode eigenvector of the new conformers remains almost unchanged. In motif 2, there is a gradual decline in the dot product; this suggests a gradual rotation of $m_j^{(c)}$ relative to $m_j^{(i)}$. However, motif 3 represents a rapid collapse of the dot product $m_j^{(i)} \cdot m_j^{(c)}$; this can occur at any point in the simulation, even if the RMSD between the initial and current conformations is small.

These motifs can be observed for the selected proteins in Figures 10.5 and 10.1-10.6 in the appendix. For example, a motif 1 behaviour is shown for mode m_7 of the yeast PDI as showin in Figure 10.5b,c,d, a motif 2 behaviour in shown for m_7 of kinesin in Figure 10.4 and a motif 3 character in shown by the m_7 of antitrypsin in Figure 10.3 as well as in Figure 10.1. Similar agreement with all motifs can be found for higher modes, although, as a general tendency, the smooth motifs 1 and 2 become gradually less visible and the more rapid changes exemplified by motif 3 more pronounced.

These sudden collapses do not indicate that the initial normal mode eigenvector has ceased to be a valid direction along which flexible motion is possible. Rather, the eigenvector has ceased to represent a "single pure mode" and may now have significant overlap with other modes.

Since yeast PDI reveals a large conformational change I investigated in more detail the evolution of the normal modes by grouping the results for PDI into a single panel for the dot product of each single mode at different E_{cut}. This reveals interesting features on the effects of varying the E_{cut}. The dot product results for each of the low-frequency modes at various E_{cut} presented show how E_{cut} affects protein mobility as shown in Figure 4.14a-e. Some modes display an asymptote for given E_{cut} which may suggest that the higher number of bonding constraints

restraint protein motion to the point that the structure stops moving and that the corresponding mode remains unchanged after that conformer. The evolution of the dot product data over the conformers varies depending on the mode and/or E_{cut}.

4.3.4 Tracking protein motion: RMSD, small loop motion

Figure 4.5a shows an ensemble of structures for BPTI generated by exploring the lowest-frequency non trivial mode, m_7, and Figures 4.5b,c show the RMSDs achieved for the five lowest-frequency non trivial modes. The amplitudes of flexible motion for BPTI at $E_{\text{cut}} = -0.2$ kcal/mol and $E_{\text{cut}} = -2.2$ kcal/mol reach an asymptote at RMSD values around 2 to 4Å. Considerable asymmetry, a factor of 2 difference in the achievable RMSD, is observable in some modes between the two possible directions of motion.

Cytochrome-c (1HRC) is a compact globular protein and although it is twice as large as BPTI in terms of residues, its capacity for flexible motion is visibly more limited, with amplitudes below 2Å for all modes. The RMSD results shown in Figure 4.5e,f support the validity of our method so that a larger protein does not necessarily displays larger mobility we now examine a compact globular protein, cytochrome-c (1HRC). RMSD results are shown in Figure 4.5e,f truncated once jamming had begun. Although this protein is twice as large as BPTI in terms of residues, its capacity for flexible motion is visibly more limited, with amplitudes below 2Å in all modes. This result is important in validating our method of projecting modes to large amplitude; if geometric simulation were capable of reaching unphysically large amplitudes, the method would lose its value.

4.3.5 Tracking protein motion: RMSD, Large loop motion

The mobility of the internal kinesin motor domain protein (1RY6) and α1-antitrypsin (1QLP) are of the same order, and both are much larger proteins than BPTI or cytochrome-c. Kinesin RMSDs values for the low-frequency modes shown in Figure 4.6b,c reveal that the amplitudes achievable for these modes differ little between the different E_{cut}; motion occurs principally in a very flexible loop region around residues 37–46. The flexible loop motion along mode m_7 is presented in Figure 4.6a for an $E_{\text{cut}} = -1.1$ kcal/mol, clearly illustrates this motion. Hence, the contribution to RMSD values is due, in this case, to the motion of this large loop. We find that the combination of the mode bias and the bonding constraints naturally causes the large flexible loop to follow a curved trajectory.

The character of α1-antitrypsin motion is very similar to kinesin; it displays

several low-frequency modes which easily explore amplitudes of up to 2–2.5Å and the contribution to those RMSD values is again due to the easy motion of a large flexible loops with respect to the relatively rigid β-sheet core of the protein, see Figures 4.6a-f. As shown in Figures 4.6e, f, α1-antitrypsin (1QLP) displays several low-frequency modes which easily explore amplitudes of up to 2–2.5Å RMSD depending on the rigidity cutoff. The motion shown in Figure 4.6d again involves the easy motion of large flexible loops with respect to the relatively rigid β-sheet core of the protein.

4.3.6 Tracking protein motion: RMSD, Domain motion

PDI (2B5E) is particularly interesting case for protein mobility since, unlike other proteins, PDI consists of four domains ***a-b-b'-a'*** connected by flexible linkers. It has been found that the structural flexibility is vital to its enzymatic function [61]. The rigidity distribution presented in section 4.3 already brings out the flexibility of the structure and its potential to undergo conformational changes. Rigidity analysis immediately brings out the flexibility of the molecule. Even at very high E_{cut} values ($E_{\text{cut}} = -0.015$ kcal/mol), the rigidity analysis reveals the domain organisation of the protein with each domain corresponding to a distinct rigid cluster flanked by flexible linkers.

The nature of the inter-domain flexibility and structural motion is portrayed in Figure 4.7a which shows the initial structure and the most extreme conformers, both at the amplitude limits of mode m_7 in the positive and negative directions respectively. The structures are aligned on the b–b' domains, bringing out the motion of the a domain and particularly of the a' domain. It is clear from Figures 4.7b-d that the RMSD achieved by the low-frequency modes does not depend significantly on the E_{cut}; motion is slightly limited at the very weakest E_{cut} (Figure 4.7b), but at other E_{cut} the achievable amplitudes are essentially the same (Figure 4.7c and d). Close examination of the conformation generation indicates that the amplitudes are limited eventually as further motion along the mode would over-extend covalent and hydrophobic-tether constraints.

The membrane protein pLGIC (2VL0) is the largest structure investigated; a pentameric structure which includes a transmembrane domain composed of α-helices and an extracellular domain consisting largely of β-sheets. The major rigidity transition in the protein identified by FIRST occurs between a E_{cut} of -0.4 kcal/mol, when almost the entire structure forms a single rigid cluster, and a E_{cut} of -0.5 kcal/mol, when the rigid linkers of the pentamer secondary structure units become flexible and the many intra-α-helices residues in the transmembrane are flexible. At the lower E_{cut} it is possible for the domains to move relative to each other, and the

transmembrane helices are also capable of relative motion. The increased RMSD possible at the lower E_{cut} is visible in Figure 4.8.

The flexible motion at the higher E_{cut}, with the protein largely rigid, involves only the motion of a few flexible loops. The motion at the lower E_{cut} is far more biologically interesting. The lowest-frequency non-trivial mode, m_7, involves a counter-rotation of the transmembrane and extracellular domains, including a change in the relative tilt of transmembrane helices lining the ion channel. This flexible motion is shown in Figure 4.8a,b.

4.3.7 Extensive RMSD as a characterisation of total flexible motion

The evolution of RMSD values is displayed in Figures 4.5–4.8 and the maximum values achieved by these motions, which range from 1.5Å to 10Å , are shown in Table 4.2. However, the character of the flexible motion does not seem well reflected by the RMSD values. For example, a small protein of 58 residues without a large conformational change such as BPTI shows RMSD values of up to 3.5Å in its small loop motion (Figure 4.5); whereas the channel protein pLGIC, a thirty times larger protein by residue count (1605 residues), shows a substantial domain motion, in which a large proportion of the atoms undergo relative motion as the transmembrane protein and extra cellular domains counter-rotate (see Figure 4.8a,b). Yet the channel protein pLGIC shows maximum RMSD values of around 2.5Å only. So, although RMSD is a good measure for comparing two similar structures, it does not necessarily capture the scale of motion in different structures.

The *extensive* RMSD measure characterizes the total protein motion without averaging. xRMSD is obtained by multiplying the RMSD (which describes the average displacement of atoms) by the number of residues in the protein. Thus, xRMSD is a measure of how much all the atoms, i.e. the structure, move rather than a measure of structural similarity like RMSD. The *extensive* RMSD is a measure of the total network coordinated motion. Figure 4.15 shows the xRMSD values for all the selected proteins moving along m_7. For the proteins with domain motion, I have chosen values of E_{cut} which correspond to lower flexibility. The observed large variation in xRMSD is therefore taking place despite a restrictive bond network. On the other hand, for the proteins with loop motion we selected E_{cut} which allow more structural flexibility. In this situation, although larger regions of the protein could become mobile, we still only observe localized loop motion. It is worth remembering that the apparent non convergence to saturated values in Figure 4.15 is due to the choice we make on the maximum number of cycles the simulation is allowed before exiting the fitting routine. We have already shown a long simulation in Figure

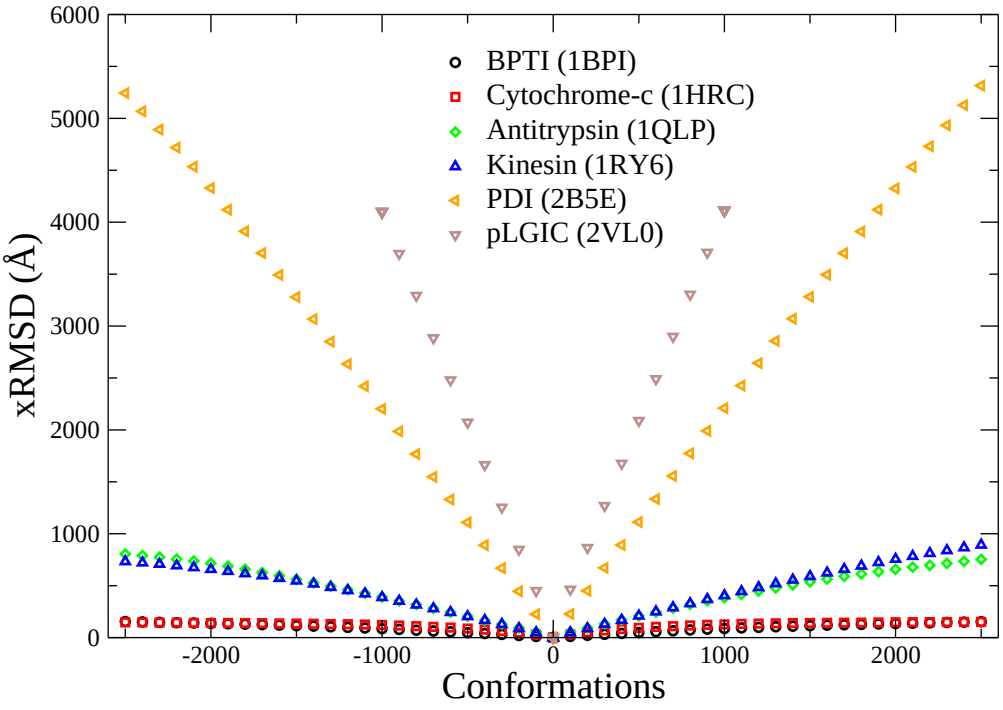


Figure 4.15: Extensive RMSD as a function of FRODA conformations for all six proteins moving along mode m_7. The maximum xRMSD values range from 150Å for BPTI to 5243Å for PDI. There are three clear categories of protein motion: large conformational changes achieved by domain motion (PDI and pLGIC), large loop motions (antitrypsin and kinesin) and small loop motions (BPTI and cytochrome-c). The selected E_{cut} for each protein are $E_{\text{cut}}^{\text{1BPI}} = -2.2$kcal/mol, $E_{\text{cut}}^{\text{1HRC}} = -1.2$ kcal/mol, $E_{\text{cut}}^{\text{1RY6}} = -1.1$ kcal/mol, $E_{\text{cut}}^{\text{1QLP}} = -1.1$ kcal/mol, $E_{\text{cut}}^{\text{2B5E}} = -0.522$ kcal/mol and $E_{\text{cut}}^{\text{2VL0}} = -0.5$ kcal/mol. The XRMSD values obtained for m_8, ..., m_{11} for the selected proteins are consistent with m_7 xRMSD.

4.5b,c presenting 10000 conformers, which show the corresponding RMSD saturation values.

The three categories of motion previously discussed — small loop motion, large loop motion and domain motion — become clearly visible in xRMSD. The xRMSD results for BPTI and cytochrome-c closely resemble each other even though cytochrome-c is almost double the size of BPTI. Similarly, the kinesin protein and the α1-antitrypsin display similar xRMSD behaviour to each other in their large loop motion. PDI and the pLGIC likewise have similar xRMSD behaviour reflecting their domain motion. Thus the character and extent of the flexible motions in proteins of various sizes, as shown in Figures 4.5a,d, 4.6a,d, 4.7a and 4.8a,b, is better reflected by the xRMSD than by the RMSD alone.

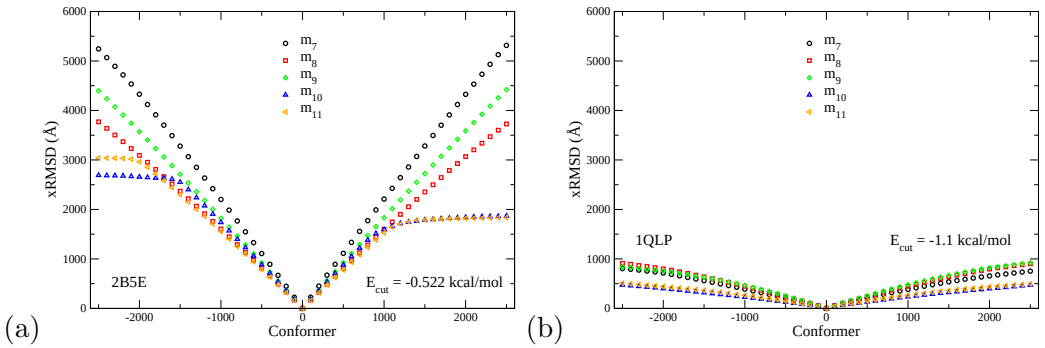


Figure 4.16: xRMSD graph for yeast PDI (2B5E) and antitrypsin (1QLP). The two proteins have a similar number of residues, 504 for PDI and 394 antitrypsin respectively. The evolution of xRMSD along the new conformations is reported for the lowest frequency modes $m_7 \cdot m_{11}$ for a E_{cut} of $E_{\text{cut}}^{\text{2B5E}} = -0.522$ kcal/mol, and $E_{\text{cut}}^{\text{1QLP}} = -1.1$ kcal/mol.

4.3.8 Extensive RMSD for all the modes

The total network coordinated motion captured by xRMSD reflects the character and extend of conformational changes not only for mode m_7, but also for all the lowest-frequency modes investigated modes $m_7 \dots m_{11}$. To illustrate this point, Figure 4.16 shows a comparison between two structures with very similar size, PDI and antitrypsin. The evolution of the xRMSD data and the order of magnitude of the xRMSD values are conserved for all the modes, hence confirming that PDI undergoes a much larger motion than antitrypsin.

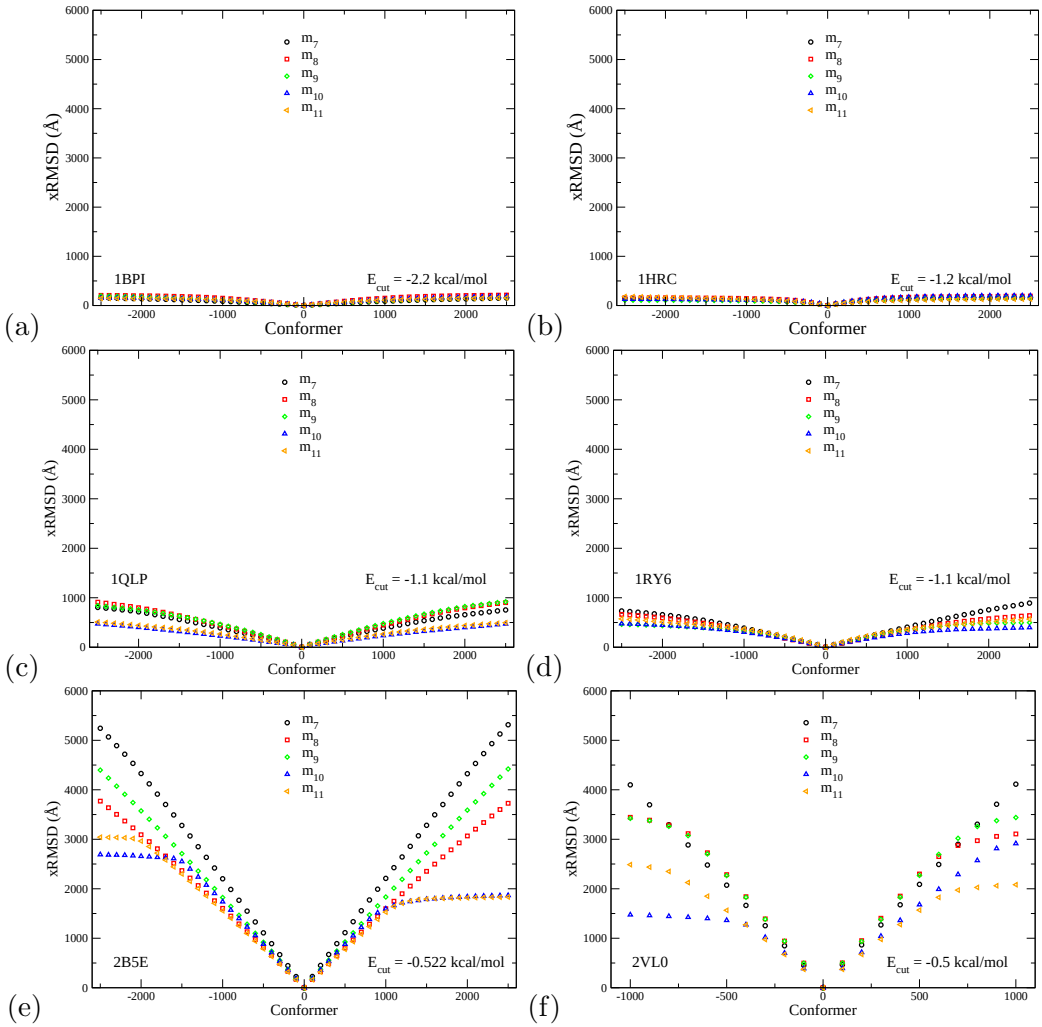


Figure 4.17: Extensive RMSD as a function of conformations for a selection of six proteins moving along modes $m_7^{(i)} \cdot m_{11}^{(c)}$ for a selected cutoff energy shown in table 4.1. The maximum xRMSD values for each protein range from 200Å for BPTI to 5243Å for PDI.

4.4 Discussion

4.4.1 Rigidity analysis

The results presented in this chapter on the rigidity and flexibility reveal that the rigidity dilution of the selected proteins is unique for each case. For example, the RCD distribution of the membrane protein displays a unique rigidity distribution. It shows a sharp change in the rigidity distribution. Hence, the selection of E_{cut} must be done based of the RCD graph. Further, the proteins displaying the highest mobility, in terms of xRMSD, have a very different pattern of rigidity dilution. This reinforces previous studies [48] which concluded that each protein structure needs to be examined individually and that the E_{cut} used for simulating protein motion must be chosen on a case by case. The choice of the E_{cut} to explore protein mobility depends on the RCD, on the E_{cut} ranges and on what question we are investigating. For all the proteins investigated in this chapter, the E_{cut} is chosen to identify at least one E_{cut} at which the protein is mostly rigid and one E_{cut} at which the protein is mostly flexible in order to quantify and compare protein mobility. In the next chapter however, we choose more E_{cut} from the RCD graph for extensively studying yeast PDI mobility under different rigid cluster constraints.

4.4.2 Significance of rigidity-analysis energy cutoff

In the case of small loop motion, it is clear that lowering the rigidity-analysis energy cutoff — thus making the structure more flexible as less hydrogen bonds are included in the rigid clusters— increases the amplitude of flexible motion, as one might expect. In the case of large domain motion, however, the most important criterion appears to be whether the domains are mutually rigid or not,. In general, the rigidity analysis of protein structures can thus add value to the simulation of flexible motion by identifying the constraints that must be eliminated in order for two residues to become independently mobile.

The proteins that undergo a large conformational change, i.e. PDI and pGLIC, display a different dependence with respect to the cutoff energy. In the case of PDI the amplitude of flexible motion for the lowest-frequency modes is almost unaffected by the choice of the E_{cut} provided it is set at a reasonable value of E_{cut} which represents each domain as a number of separate small rigid clusters. This conclusion can also be drawn in the case of the ligand-gated ion channel protein.

4.4.3 RMSD

RMSD is a standard measure of structural similarity between two proteins. However, the results presented in this chapter focus on using RMSD to characterise protein motion; this reveal its limitations to correctly describe protein large conformational changes. Visual inspection of the overlapped conformers revealed three types of motion which are not differentiated by using RMSD as a mobility measure. The most clear example is the comparison of RMSD values and conformer motion between BPTI and pLGIC; 58 and 1605 residue long respectively. The motion for BPTI is restricted to the motion of a small loop, whereas the membrane protein displays a domain motion where the intracellular domain moves with respect to the extracellular domain. The RMSD values for BPTI range $\approx$ 1.5–4.5Å whereas for pLGIC $\approx$ 0.7–2.5Å. The displacement of the overlapped conformers indicated a larger motion for the membrane protein than for BPTI, yet the RMSD measures suggest the contrary. This is due to the least-squares fitting routinely used to compare two very similar structures that averages and minimizes the structural variation; which in this case is performed by using PyMOL `intra_fit` command.

4.4.4 xRMSD

xRMSD characterises protein motion so that it correlates with the observed motion from the overlapping conformers. Furthermore, the results of comparing the xRMSD mobility of two proteins with similar number of residues (PDI and Antitrypsin) dismiss the possibility that the different types of motion here mentioned are due to the protein size. In addition, the results for the mobility characterisation are consistent not only for the lowest frequency mode but for the other modes too. Therefore, xRMSD properly identifies the character of protein motion and can be used as an useful measure to identify conformational change in proteins.

4.5 Conclusions

The reported hybrid method is able to explore protein motion by integrating both rigidity constraints from FIRST and low-frequency mode eigenvectors obtained using ELNEMO, into the geometric simulation FRODA. In order to illustrate the method, we have applied it here to a diverse selection of proteins whose flexible motion ranges from small loop motion (BPTI, cytochrome-c) and large loop motion (a kinesin and an antitrypsin) to large motions of entire domains (protein disulphide isomerase and a transmembrane pore protein). Detailed studies of dynamics in relation to func-

tion of particular proteins are currently in progress [66, 67]. The combined method can rapidly explore motion to large amplitudes in an all-atom model of the protein structure, maintaining steric exclusion and retaining the covalent and non-covalent bonding interactions present in the original structure. Significant amplitudes of motion are achieved with only CPU-minutes of computational effort even in a pentameric pore protein with more than 1600 residues. The amplitude of motion that can be achieved by flexible loops increases as the rigidity-analysis energy cutoff is lowered. For large-scale motion of domains, the most important criterion is that the energy cutoff should be low enough that different domains do not form a single rigid body. RMSD, a measure of structural similarity, does not properly reflect the scale of flexible motion between different proteins; this is better captured by an extensive measure, xRMSD, which reflects both the size of the protein and the amplitude of its motion. Examination of the behaviour of the elastic network eigenvectors during the motion shows many examples of mode mixing, so that a given vector of motion can change from being a pure mode to a mixed one after quite small displacements, without losing its character as an "easy" direction for flexible motion.

Chapter 5

Investigating PDI mobility with coarse graining methods

5.1 Introduction

5.1.1 Oxidation and isomerisation of disulphide bonds and the biological role of yeast PDI

A vital step in the correct folding of some proteins is the formation of disulphide bonds between two cysteine residues, which contain thiol groups. A thiol group (R-SH) contains a sulfhydryl group (S) linked to a carbon-containing group of atoms (R). The formation of disulphide bonds takes place within a specialised organelle of the eukaryotic cell, the endoplasmic reticulum (ER), and requires an oxidative environment. Disulphide bonds in proteins are formed as two cystein residues with thiol groups are oxidized and form a covalent bond. For the oxidative phase to take place two thiol groups(R-SH) must get close to each other and form a covalent bond, i.e. (R-SH) + (R-SH) = (R-S-S-R). Isomerisation requires that four thiol groups forming two disulphide bonds interchange their bonds. The main function of disulphide bonds is to stabilise a protein structure. A key component locked in the ER is protein disulphide isomerase (PDI), which catalyses the oxidation and isomerisation steps in the formation of disulphide bonds, a key component is stabilising protein structures as for example BPTI. PDI is a multifunctional protein and in addition to its role in catalysing disulphide bond formation acts as a molecular chaperone [68] and aids in the folding of other proteins [69].

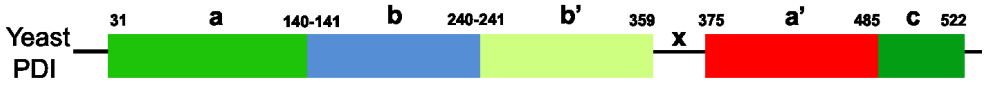


Figure 5.1: Domain organisation of yeast PDI deduced based on the crystal structure. There are four domains (***a-b-b'-a'***), a flexible loop ***x*** connecting domains ***b*** and ***b'*** and the c-terminal tail.

5.1.2 The PDI family

The elucidation of the sequence of cloned rat liver PDI in 1985 by Edman et al. [70] was a great advance at the time. They predicted a protein with two distinct regions homologous to other thioredoxin, the (***a*** and ***a'***) regions, plus two other regions (***b*** and ***b'***) in the order ***a-b-b'-b'***. This was the first indication that PDI family could also be catalysing disulphide bond formation via an internal disulphide sulphydryl interchange. Later on Farquhar et al. [71] isolated the gene encoding PDI for *Saccharomyces cerevisiae* (yeast). The amino acid sequence deduced from their analysis strongly suggested that this was a PDI encoding gene for various reasons: its predicted size is characteristic of full-length mammalian PDI translational products, the amino acid sequence shows a $30-32\%$ identity with mammalian and avian PDI sequences, and the amino acid sequence contained two copies of the 'thioredoxin-like' active sites i.e. two segments of approximately 100 amino acids, in the ***a*** and ***a'*** regions. Besides these similarities, the alignment of yeast and mammalian PDI sequence also revealed other regions that showed significant similarity. Later on these regions were identified as part of the same domains ***a*** and ***a'***. The thioredoxin fold is a secondary structure motif that consists of a four stranded β-sheet surrounded by three α-helices in a $\beta\alpha\beta\alpha\beta\beta\alpha$ configuration, and its function is associated with disulphide bond formation.

Freedman et al. [72] suggested the sequence of major structural features for PDI conserved in eukaryotes. Using sequence homologies and some limited proteolysis they proposed a domain sequence as follows, ***a-e-b-b'-a'-c***. This was one of the first attempts to identify the structural distribution of PDI family. A few years later Darby et al.[73], also using limited proteolysis on human PDI constructs, refined the identification of the domain limits to conclude that domain ***e*** was structurally part of domain ***a***, see Figure 5.1.

NMR studies provided the first detailed structural information about PDI families for the isolated domains ***a*** and ***b*** confirming the modular and multidomain structure [74, 75]. These investigations confirmed that the ***e*** domain was in fact part of the ***a*** domain and that the ***a*** domain resembles that of thierodoxin. Surprisingly,

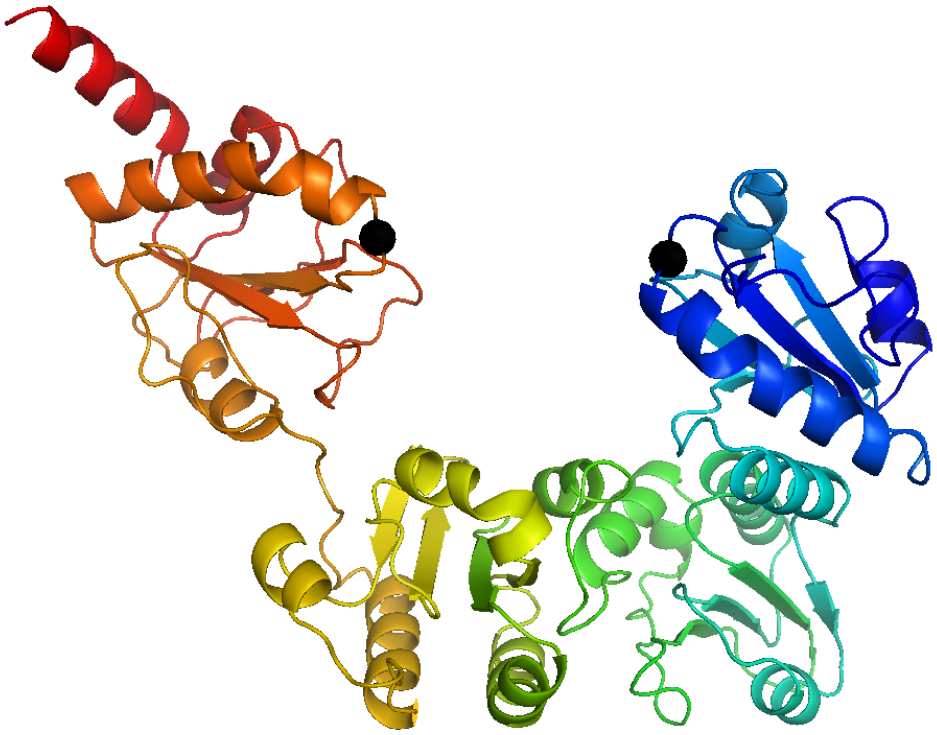


Figure 5.2: Tertiary structure of yeast PDI. Ribbon diagram of yeast PDI rainbow coloured from N-terminus to C-terminus with the effect that ***a*** domain is predominantly blue, ***b*** domain is predominantly green, ***b'*** domain is mostly yellow, ***a'*** domain is predominantly orange. The C-terminal (within the ***a'*** domain) is shown in red and the α-carbons of the active sites (cysteine residues) are shown as black spheres. The distance between these active sites from the crystal structure is $\simeq$ 27Å. The flexible ***x*** region is shown between domain ***a'*** and ***b'*** in light orange colour. Domains ***a***, ***b'*** and ***a'*** are situated in the same spatial plane but domain ***b*** (dark green) is displaced away from the reader. Note that some alpha helices appear to be placed closer to the reader than domain ***b*** indicating that such domain is indeed in a different spatial plane.

the ***b*** domain also showed a thioredoxin-like fold, but only the active sites on domains ***a*** and ***a'*** contribute to human PDI activity [76]. Both the isolated ***a*** and ***a'*** domains are folded and catalytically active during oxidative processes and their activities are similar to the full length molecule. However, neither domain by itself is able to display a significant catalysis of the isomerisation needed for refolding BPTI [77].

5.1.3 Structural properties and functions of yeast PDI

The availability of crystal structures in the study of PDI was a significant step ahead. The first available crystal structure for yeast PDI [63] shown in Figure 5.2, was published in 2006 with a resolution of 2.4Å and PDB code 2B5E. This structure allowed for the first time to observe the whole structural distribution of the yeast PDI protein. The spatial organization of the protein structure resembles the shape of a twisted "U" with the two active sites at domains ***a*** and ***a'***, located at the end of the "U" and domains ***b*** and ***b'*** defining the base. Domains ***a***, ***b'*** and ***a'*** are situated in the same spacial plane but domain ***b*** is displaced. The active sites at the inside surface of domains ***a*** and ***a'***, face each other and are $\simeq$ 27Å apart from each other.

The four domain distribution ***a-b-b'-a'*** shown by the crystal structure coincides with previous studies on mammalian PDI structures. The crystal structure [63] also confirmed that the ***a-b-b'-a'*** domains are linked by flexible linker regions of different lengths, which indicates different contact surfaces between adjacent domains. The linker regions between the ***a*** and ***b*** domains and between the ***b*** and ***b'*** domains are much shorter than the 19-residue linker region ***x*** situated between the ***b'*** and ***a'*** domains as previously predicted [72]. The contact surface between the ***b*** and ***b'*** domains is larger ($\simeq$ 700Å^2) than the area of contact between either the ***a*** and ***b*** or the ***a'*** and ***b'*** domains ($\simeq$ 200Å^2). In addition, at the end of domain ***a'***, the crystal structure reveals a highly acidic α-helix rich region, the ***c*** region. Hence the domain sequence ***a-b-b'-x-a'***, the flexible regions and the different contact surfaces between them suggests that the ***a*** and ***a'*** domains are flexible with respect to the 'base' formed by ***b*** and ***b'*** domains.

There is a strong presence of hydrophobic residues in the interior of the "U" surface, see Figure 5.2, which could facilitate interactions with misfolded proteins. This would suggest that misfolded proteins could dock at the base of the "U" shape i.e. domains ***b*** and ***b'***, and that ***a*** and ***a'*** domains would change their conformation so that the thiol groups can catalyse the rearrangement of the misfolded protein. Therefore, the biochemical information available reveals that the functional activities of yeast PDI require protein flexibility and domain conformational change.

The second yeast PDI structure [62] revealed a large scale conformational change compared with the initially reported structure [63]. This confirmed that PDI is a highly flexible molecule with its catalytic domains, ***a*** and ***a'***, representing two mobile arms connected to a more rigid core composed of the ***b*** and ***b'*** domains. Furthermore, limited proteolysis revealed that the linker in-between domain ***a'*** and the base ***b-b'*** was less susceptible to degrade that the one connecting the ***a*** and the base ***b-b'*** [62].

As previously said, yeast PDI has two active sites or active cysteine residues, one in domain ***a*** and one in domain ***a'***. Its main function is to catalyse three types of thio-disulfide exchange reactions: oxidation, reduction and isomerisation, all of which need the ***a*** and/or ***a'*** domain/s active sites involvement. The catalytic performance of PDI is quite remarkable, Freedman et al. [72] discovered that PDI can catalyse the oxidation and isomerisation of disulphide bonds up to 6000 times.

5.2 Methods

5.2.1 Rigidity distribution and mobility simulations of yeast PDI

A detailed explanation of the methods used to identify rigid regions and the creation of the rigidity dilutions patterns or RCD graphs can be found in the previous chapters. Figure 5.3 reports the complete rigidity distribution and the domain distribution for yeast PDI. The rigid regions in the RCD graph are defined with thick coloured lines to identify each rigid cluster and the flexible regions are shown with black thin lines. Six E_{cut} from the RCD graph are selected with the intention of investigating how varying the E_{cut} affects protein mobility. These energy E_{cut} represent either significant changes in the rigidity or changes in rigidity distribution of a cluster. The selected E_{cut} are $E_{\text{cut}} = -0.003$ kcal/mol, $E_{\text{cut}} = -0.015$ kcal/mol, $E_{\text{cut}} = -0.133$, $E_{\text{cut}} = -0.522$ kcal/mol, $E_{\text{cut}} = -0.885$ kcal/mol and $E_{\text{cut}} = -1.412$ kcal/mol. I have summarised the rigidity distribution for these energy E_{cut} in a 'mini' RCD plot, see Figure 5.4, for ease of comparison between them. In the 'mini' RCD plot it is easier to observe the correlation between the rigidity distribution along the polypeptide chain and the domains, but also it is easier to observe the changes in the rigidity distribution in each domain for various E_{cut}.

Likewise, the methodology to simulate protein motion has been extensively described in the previous chapters. Here, I also employ ElNemo to obtain the normal modes using a bond distance E_{cut} of 12Å and I focus on the five lowest frequency modes to investigate if there is a large conformational change.

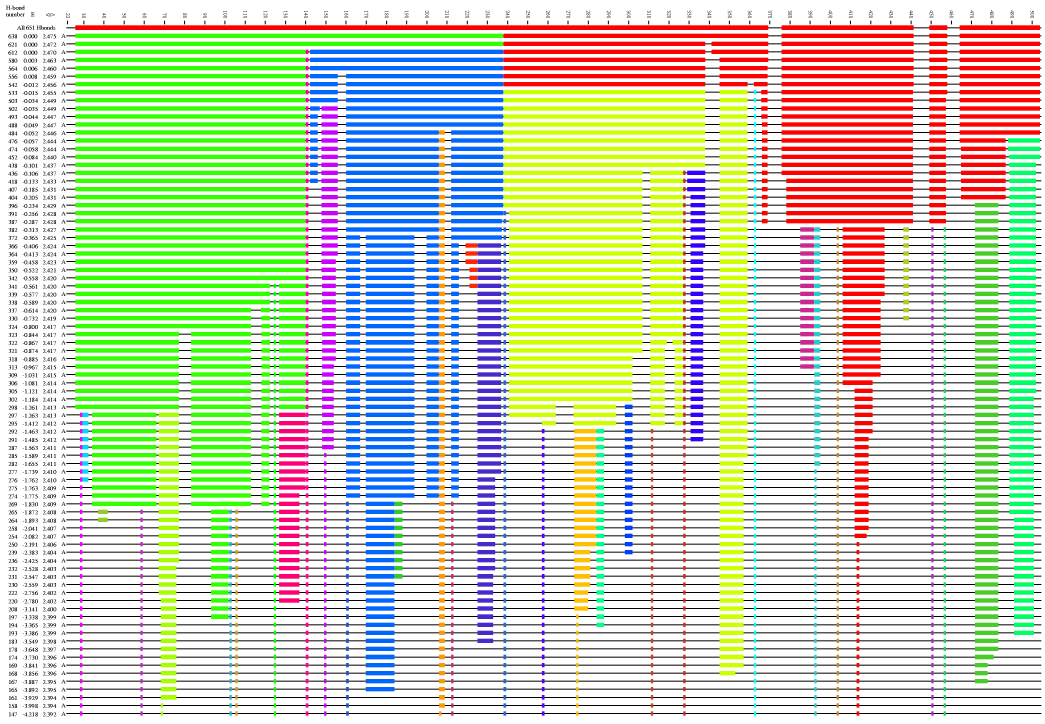


Figure 5.3: Rigid cluster decomposition graphs for yeast PDI (2B5E). The **x**-axis represents the protein backbone and the **y**-axis the energy, E_{cut}, of the last hydrogen bond, which after being removed provokes a change in the rigidity distribution. Each line represents the new rigidity distribution of the polypeptide chain induced by removing a bond which alters the previous rigidity configuration. The residues belonging to rigid clusters are coloured — with the biggest rigid cluster coloured in red, whereas the flexible regions are shown as thin black lines.

5.2.2 Computing the active sites distance

In order to quantify yeast PDI's conformational change I measure the distance between active sites $d_{cc}^{(m_j)}$; in particular the distance between the cysteine α-carbons in domains ***a*** and ***a'*** with atom id 599 and 5979 (residues 61 and 406) from the PDB file. The position of the active sites in the tertiary structure are shown as a black sphere in Figure 5.2 and the inter-cysteine distances are reported in Figures 5.7 and 5.8a-d. The inter-cysteine distance is computed using a self developed PYTHON script loaded in PYMOL and using the conformers recorded during the geometric simulation as an input. Although full details on how conformers are generated can be found in chapter 4, it is worth noting that a new conformer is recorded for every 100th generated structure and that each mode has a positive and negative direction of motion. The simulations start to get 'jammed' when reaching conformer 22 and up to 26, hence we report simulation runs that have reached 25 conformers for each

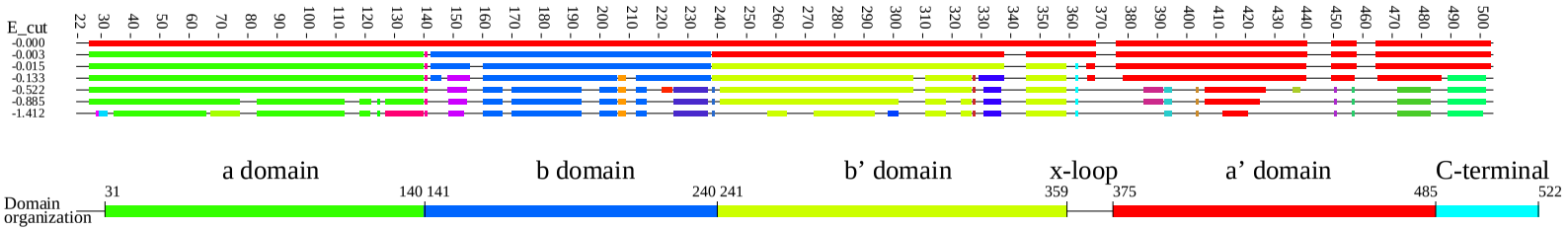


Figure 5.4: Rigidity distribution and domain organization for yeast PDI. (top) Rigidity distribution for selected E_{cut}, $E_{\text{cut}} = -0.003$ kcal/mol, $E_{\text{cut}} = -0.015$ kcal/mol, $E_{\text{cut}} = -0.133$, $E_{\text{cut}} = -0.522$ kcal/mol, $E_{\text{cut}} = -0.885$ kcal/mol and $E_{\text{cut}} = -1.412$ kcal/mol. The $\underline{x}$ axis represents the protein backbone and the $\underline{y}$ axis the E_{cut}, of the last hydrogen bond removed that provoked a change in the rigidity distribution. Each line represents the rigidity distribution of the polypeptide chain calculated by removing all the bonds with a E_{cut} below the selected E_{cut} threshold. Protein residues are coloured if they belong to a rigid cluster whereas the flexible regions are shown as thin black line.(bottom) Domain organization for yeast PDI as deducted from biochemical studies.

direction of motion and for each mode. This is one of the shortcomings of FRODA, due to the fact that it is purely a geometric projection of atoms across space and that it lacks a minimization function that stabilises the rigid clusters after projection.

We choose to use the ***b*** and ***b'*** domains as a structural base to align the different conformers and visualise their conformational change with respect to the initial structure. The suggestion of considering ***b*** and ***b'*** domains as a structural base appeared already in the literature [63] and it is backed up by the existence of bigger areas of contact between ***b*** and ***b'*** domains compared to the contact areas between either domains ***a'*** and ***b'*** or domains ***b*** and ***a***.

5.3 Results

5.3.1 Domain recognition

The rigidity analysis in Figures 5.3 and 5.4 show that yeast PDI has four main regions easily distinguishable across the RCD plot, the larger rigid clusters (residues 25-140, 140-240 and 240-360) and a region whose flexibility depends on the E_{cut} (residues 380-480). The larger rigid clusters approximately correspond to domains ***a-b-b'*** and the most flexible region corresponds to domain ***a'***. Whereas the three main rigid clusters corresponding to domains ***a-b-b'*** maintain their rigidity up to $E_{\text{cut}} < -1.412$ kcal/mol, the most flexible region, domain ***a'***, remains as a single rigid cluster for $E_{\text{cut}} > -0.313$ kcal/mol and breaks up into three-four similar sized clusters for $E_{\text{cut}} \leq -0.313$ kcal/mol.

The rigidity analysis also identifies several flexible regions, of which are of special interest the intra-domain regions ***a-b***, ***b-b'*** and ***b'-a'***. The intra-domain region ***b'-a'*** corresponds to the ***x*** linker region and it is shown to be completely flexible for $E_{\text{cut}} < -0.313$ kcal/mol in the RCD graph. The region that separates domain ***a*** and ***b*** is also shown to be flexible by the rigidity analysis, and although it contains a small rigid cluster the flexible linker is not bound neither to domain ***a*** nor ***b***. Hence these regions are good candidates to act as hinges that would allow domain motion, i.e. motion of domains ***a*** and ***a'*** with respect to the chosen structural base ***b-b'***. The intra-domain region separating domains ***b*** and ***b'*** is rigid for $E_{\text{cut}} \geq -1.412$ kcal/mol and only a few residues appear to be flexible for $E_{\text{cut}} < -1.412$ kcal/mol.

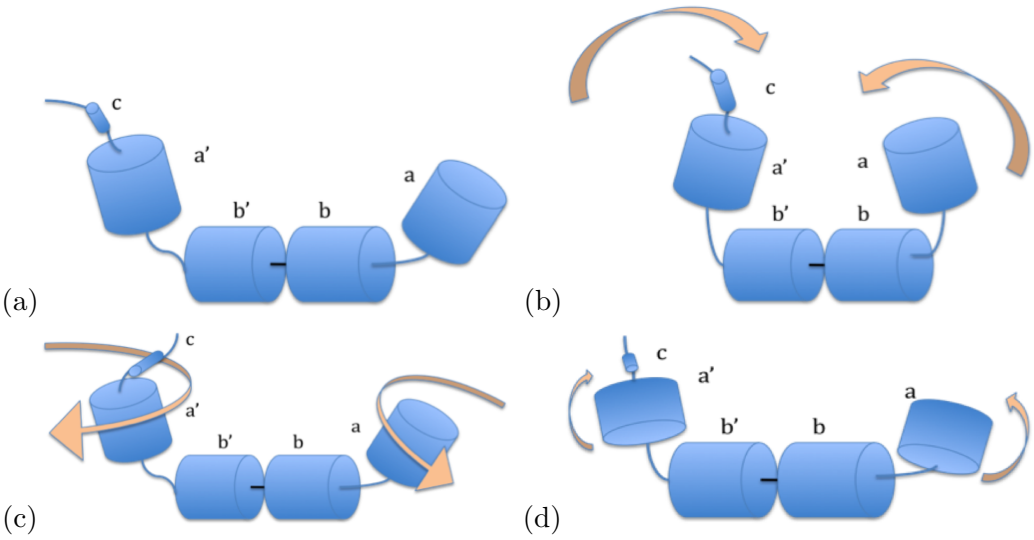


Figure 5.5: Cartoon representation of yeast PDI conformational motion along the lowest frequency modes. Panel (a) shows the cartoon representation of the crystal structure, panel (b) shows the double hinge motion by which the protein moves domains ***a*** and ***a'*** towards or away from each other, panel (c) illustrates domain rotation along a given axis, and panel (d) illustrates domains ***a*** and ***a'*** coordinated motion over the plane perpendicular to the axis defined by domains ***b-b'***.

5.3.2 Domain rigidity gradation

Another interesting feature that the complete RCD graph of Figure 5.3 reveals is the rigidity dilution pattern, i.e. how the rigidity distribution evolves for each of the domains as bonds are removed revealing new rigidity distributions. It is clear that the changes in rigidity distribution as we lower the E_{cut} is quite different for each domain. In particular domain ***a'*** becomes flexible and broken up in multiple clusters at E_{cut} in the region of $E_{\text{cut}} \simeq -0.250$ to $E_{\text{cut}} \simeq -0.300$ kcal/mol, much lower than other domains.

5.3.3 Yeast PDI modes of motion

The low-frequency modes of motion display the collective motion of protein residues potentially providing an insight into the functional mechanism of proteins [78]. In this study I observe typically a few different types of motion, which are summarised in Figure 5.5. These motions are based on a double-hinge structural mobility, i.e. domain motion of domains ***a*** and ***a'*** towards and away from each other, domains ***a*** and ***a'*** rotation (coordinated and anti-coordinated), domains ***a*** and ***a'*** rotations and domain ***a*** and ***a'*** sideways motions (coordinated and anti-coordinated). In line

with previous studies [63] we have chosen to have domains ***b-b'*** as the sequential basis to structurally align the generated conformers to the original structure.

During the mobility simulation a number of conformers are generated at each step, one in one hundred is reported as a new conformer, that is 25 in total. Five of these are selected to illustrate the evolution of the protein motion in Figures 5.6-5.12. The selected conformers are shown using two display formats: (a) the individual conformers one by one and (b) the conformers overlapped onto each other using the ***b-b'*** domains as a base for alignment. The single structures labelled (a) to (e) represent single conformers -2500 (a) and -1200 (b) for the negative direction of motion and conformers 1200 (d) and 2500 (e) for the positive direction of motion, whereas the initial crystal structure or conformer 0 is labelled as (c). I also present the structural alignment of these conformers aligned to each other using domains ***b-b'*** as a structural reference. For most modes I present two viewpoints of the aligned conformers which are label as panels (f) and (f').

5.3.4 Double hinge motion: mode m_7

The lowest frequency mode of the yeast PDI structure reveals a large conformational change. Figure 5.6a-e presents yeast PDI conformers from the most open conformation (a) up to the most closed (e) as the initial structure is projected along mode m_7. The overlapped structures in Figure 5.6f and 5.6f' indicate that domain ***a'*** and domain ***a*** move towards each other with domain ***a'*** moving the most. In order to quantify the motion I calculated the distance between active cysteines for all the conformers. The results shown in Figure 5.7 reveal that the active sites inter-cysteine distance ranges from a *minimum* distance of $d_{min}^{(m_7)} \simeq 15$Å up to a *maximum* distance of $d_{max}^{(m_7)} \simeq 55$Å.

Furthermore, the evolution of the inter-cysteine distance and its maximum and minimum are similar among the different E_{cut} with changes only appearing for the E_{cut} outside the biologically relevant E_{cut} range. It is worth noting that the minimum inter-cysteine distance is consistent among all the E_{cut} whereas the maximum inter-cysteine distance between active sites is restricted for the lowest E_{cut}. Mode m_7 defines a double hinge motion that is selectively restricted by the E_{cut} depending on the direction of motion but consistent among the biologically relevant E_{cut}.

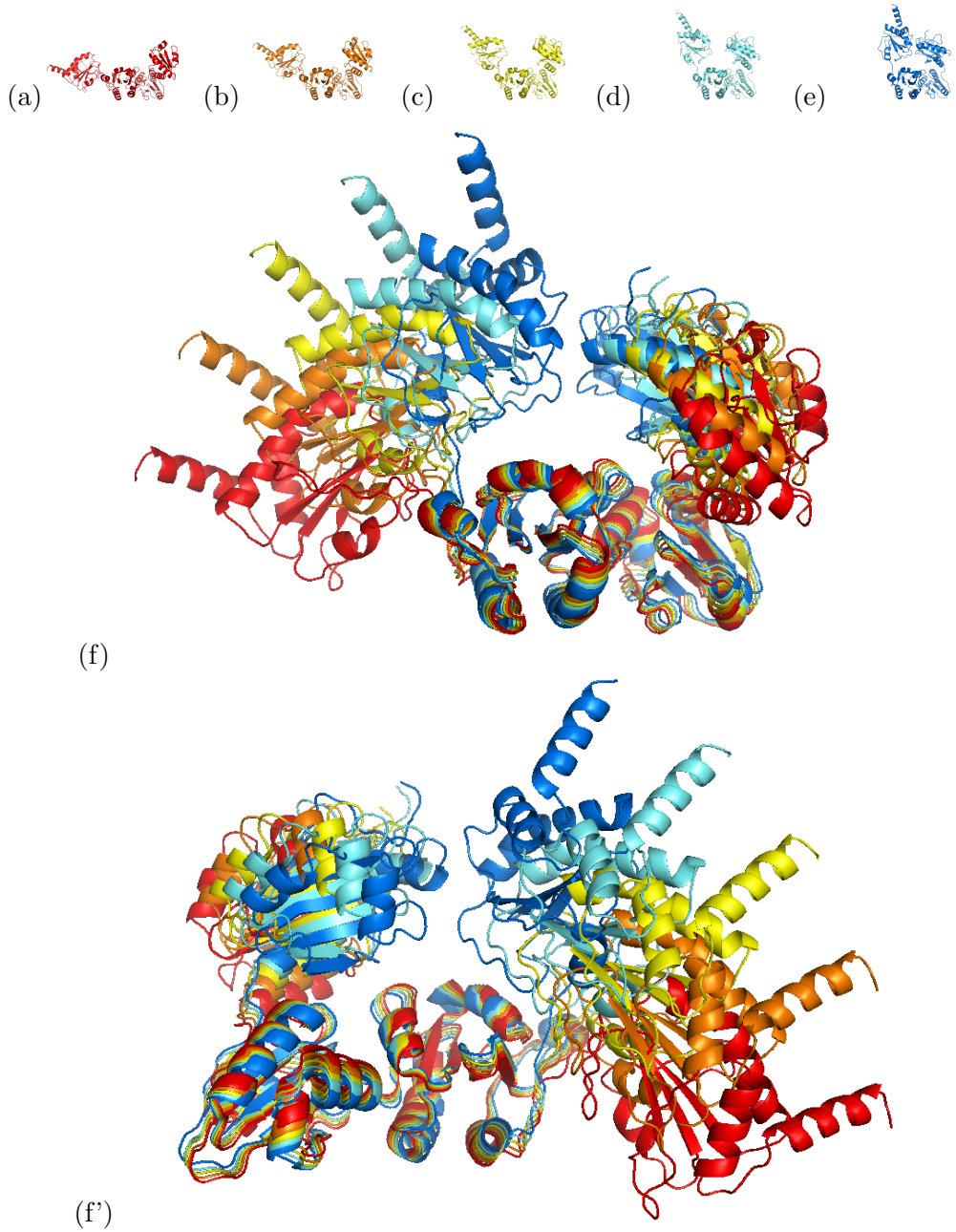


Figure 5.6: Conformational change for yeast PDI (2B5E) moving along mode m_7. Panels (a) to (e) correspond to conformers obtained from projecting the initial protein structure (c) along mode m_7 towards the opening or negative direction of motions i.e. conformers (a) and (b), and the closing or positive direction of motion i.e. conformers (d) and (e). The structures (a) and (e) correspond to conformers 2500 for the negative and positive direction of motion respectively as shown in Figures 5.7; whereas the structures (b) and (d) correspond to conformers 1200, also in the negative and positive direction of motion respectively. Figure (f) and (f') show the overlap of all the selected conformers aligned using domains ***b*** and ***b'*** as the structural basis.

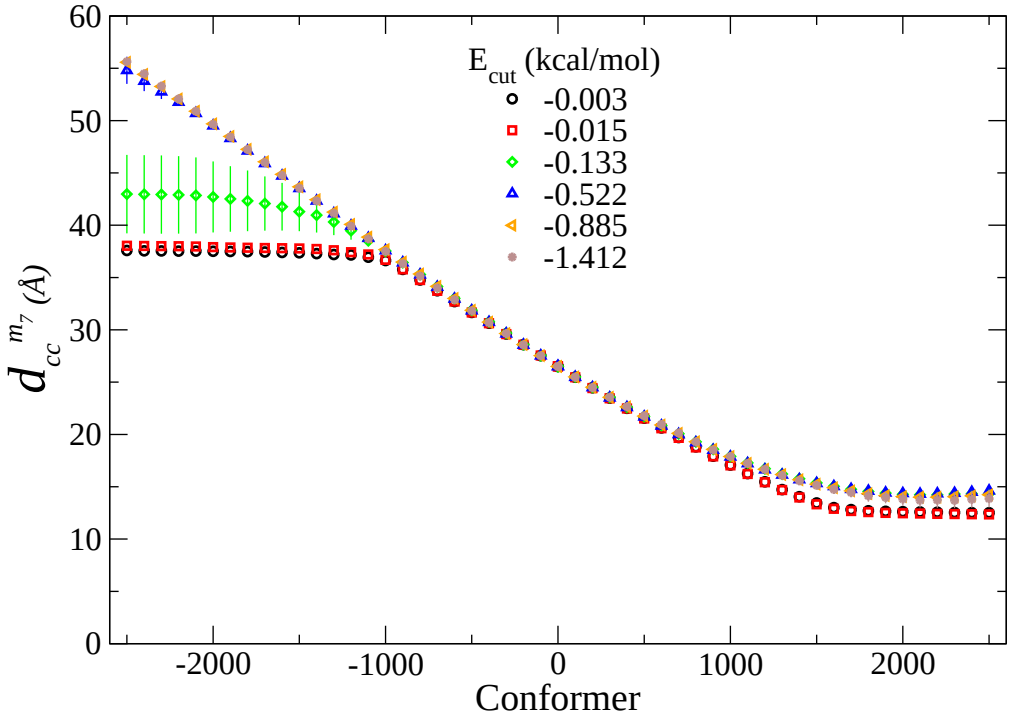


Figure 5.7: Distance between the cysteine active sites in the ***a*** and ***a'*** domains as the protein structure is projected along mode m_7. The active sites move with respect to each other as the initial structure is projected along each of the lowest frequency modes $(m_7^{(i)} \dots m_{11}^{(c)})$. Horizontal data points signal that the simulation has reach either its stereochemical boundary or software limitations. Higher E_{cut} include a higher number of hydrogen bonds.

5.3.5 Domain rotation: mode m_8

Mode m_8 mobility is shown in Figure 5.9. This mode directs the motion of domains ***a*** and ***a'*** to rotate almost perpendicular to the axis defined by the base ***b-b'***. The evolution of the inter-cysteine distance reported in Figure 5.8a also confirms the large conformational change for mode m_8 with maximum and minimum inter-cysteine distance $d_{max}^{(m_8)} \simeq 45$Å and $d_{min}^{(m_8)} \simeq 25$Å. Although a measure representing the rotational character of the motion would be better suited to quantify the domain rotations, the evolution of the inter-cysteine distance and the visual inspection of the overlapped conformers suffices to reveal the large conformational change that mode m_8 exhibits.

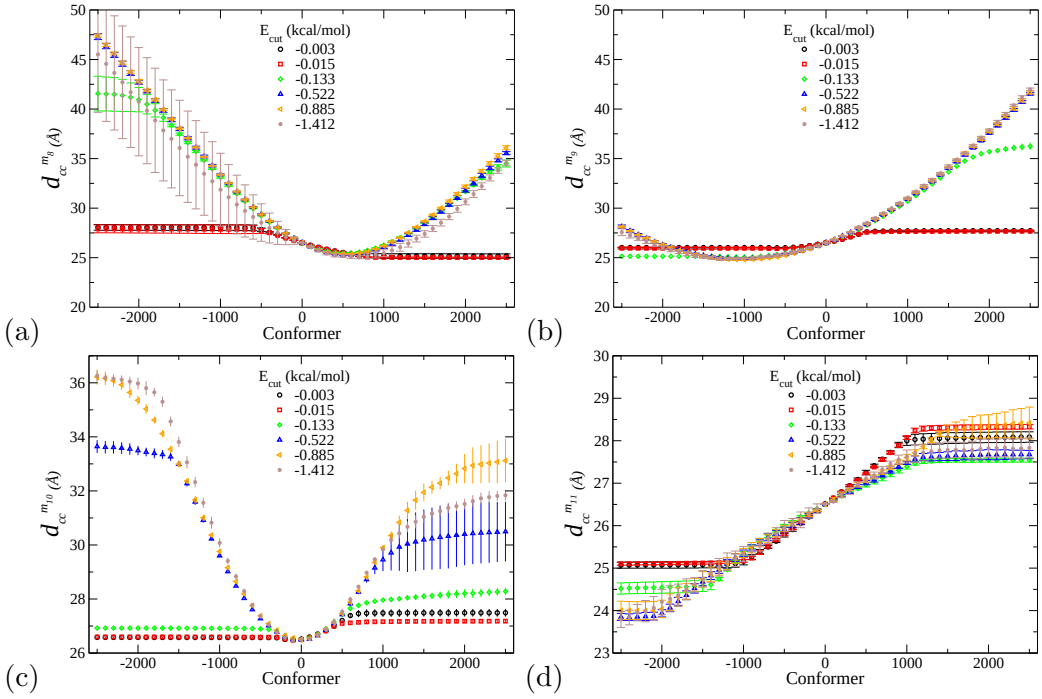


Figure 5.8: Distance between the cysteine active sites in the ***a*** and ***a'*** domains. The active sites move with respect to each other as the initial structure is projected along the initial modes $(m_8^{(i)} \ldots m_{11}^{(c)})$. The reported distances between active sites correspond to the intercysteine distance of the conformers obtained as the protein is projected along (a) mode m_8, (b) mode m_9, (c) mode m_{10} and (d) mode m_{11}.

5.3.6 Domain rotation and sideways motion: mode m_9

When it comes to the third lowest frequency mode yeast PDI reveals the rotation of the ***a'*** domain coupled with the sideways motion of domain ***a***. The set of conformers presented in Figure 5.10f show the rotation around the axis defined by the ***b-b'*** domains and the axial view of the conformers. The conformers shown in Figure 5.10f' reveal that the ***a'*** domain rotates along an axis located approximately at the end of the α-helix defined by residues 407-426. The maximum and minimum intercysteine distance for mode m_9 reported in Figure 5.8b is slightly smaller compared to previous modes, $d_{max}^{(m_9)} \simeq 43$Å and $d_{min}^{(m_9)} \simeq 25$Å, and the coordinated motion of the domains is different to the previous modes.

5.3.7 Domain rotation and sideways motion: mode m_{10}

The motion of mode m_{10} resembles mode m_9 but with the ***a*** and ***a'*** domains 'swapping their roles'. Figure 5.11f shows the sideways motion of domain ***a'*** and Figure 5.11f' the rotation of domain ***a'***. However, the range of domain rotation

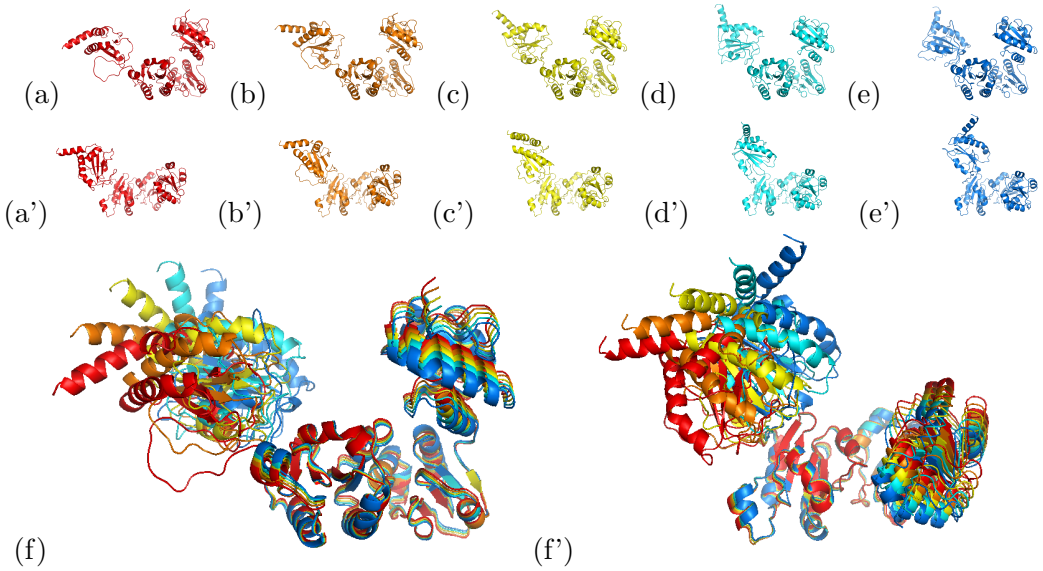


Figure 5.9: Conformational change for mode m_8 of the yeast PDI (2B5E) structure. The (a-e) series and the overlapped structures (f) correspond to a side view, and the (a'-e') series and the overlapped structures (f') correspond to a axial view. The selected conformers in Figures (f) and (f') are aligned using domains ***b*** and ***b'*** as the structural basis. These figures show an anti-coordinated sideways motion of domains ***a*** and ***a'*** with respect to ***b*** and ***b'***, and with domain ***a'*** moving to a greater extent than domain ***a***.

and sideways motion observed in the overlapped structures in Figure 5.11f and f', and in the maximum and minimum inter-cysteine distance in Figure 5.8c is much smaller than for mode m_9. The overlapping of domains ***a*** and ***a'*** between conformers -2500 (dark blue) and -1200 (light blue), and between conformers 2500 (red) and 1200 (orange) in Figure 5.11f and 5.11f' corresponds to the simulation stage where the structure is reaching the stereochemical or bond network limit. Therefore the geometric simulation struggles to find a conformation that respects stereochemical constraints and fit the ghost templates while moving along the normal mode of motion.

5.3.8 Coordinated sideways motion: mode m_{11}

The sequence of conformers obtained for mode m_{11} show a coordinated sideways motion of both domains ***a*** and ***a'*** with respect to the base ***b-b'***, see Figure 5.12f and 5.12f'. In this case the inter-cysteine distance does not reflect domain motion to its true extent since the domains and therefore the active sites move simultaneously to the same direction, see Figure 5.8d. This highlights the need to include different

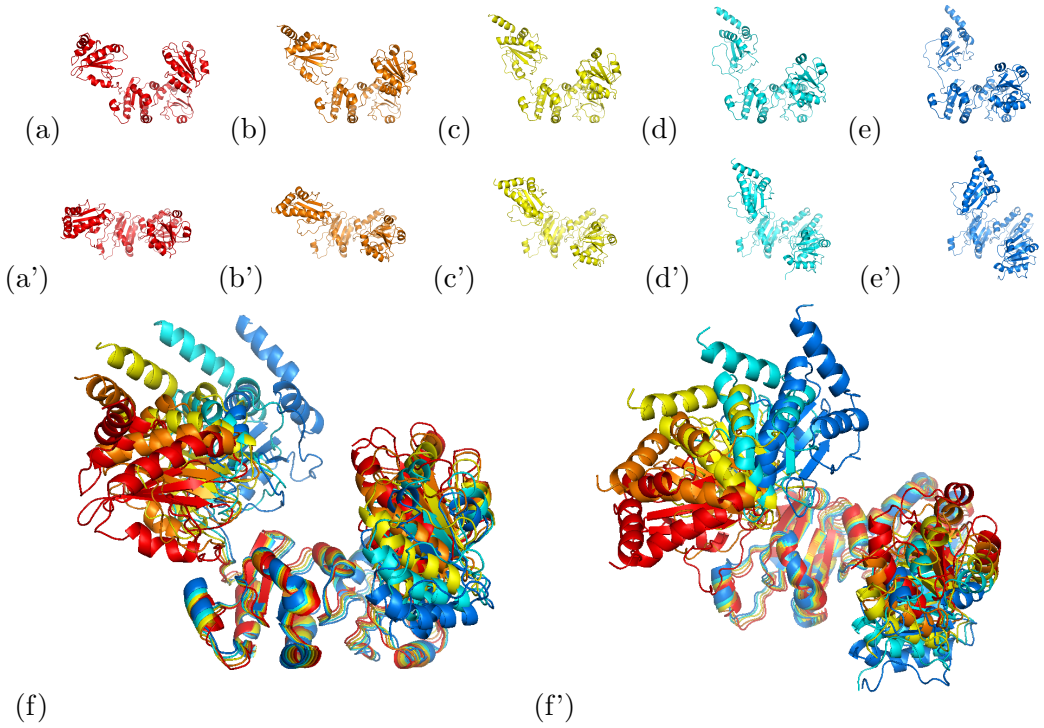


Figure 5.10: Conformational change for mode m_9 of the yeast PDI (2B5E) structure. The conformers in panels (a-e), (a'-e'), (f) and (f') correspond to the same conformer sequence as in Figure 5.6 and 5.9. A side view is shown in panels (a-e) and (f), and an axial view is shown in panels (a'-e') and (f'). The selected conformers in Figures (f) and (f') are aligned using domains ***b*** and ***b'*** as the structural basis. The sequence of conformers in panels (f) and (f') show an anti-coordinated rotation of domains ***a*** and ***a'*** with respect to ***b*** and ***b'***. Mode m_9 also shows that domain ***a'*** rotates to a greater extent than domain ***a***. However, the axis of rotation for domain ***a'*** is near the coinciding end of the α-helices defined by residues 407-426.

measures to track protein motion in the geometrical simulations software.

5.3.9 Effects of E_{cut} on protein mobility

Six different E_{cut} were selected to investigate how reducing the number of hydrogen bonds affects protein mobility.

The evolution of inter-cysteines active sites distance $d_{cc}^{(m_7)}$ shown in Figure 5.7 defines a double hinge-like motion where the active sites in domain ***a'*** and ***a*** move towards or away from each other for conformers 0 to 2500 and 0 to -2500 respectively. The simulation using the highest E_{cut} at $E_{\text{cut}} = -0.003$ kcal/mol and $E_{\text{cut}} = -0.015$ kcal/mol show that yeast PDI motion is restricted compared with lower E_{cut}. Similarly, the distance between active sites for $E_{\text{cut}} = -0.133$

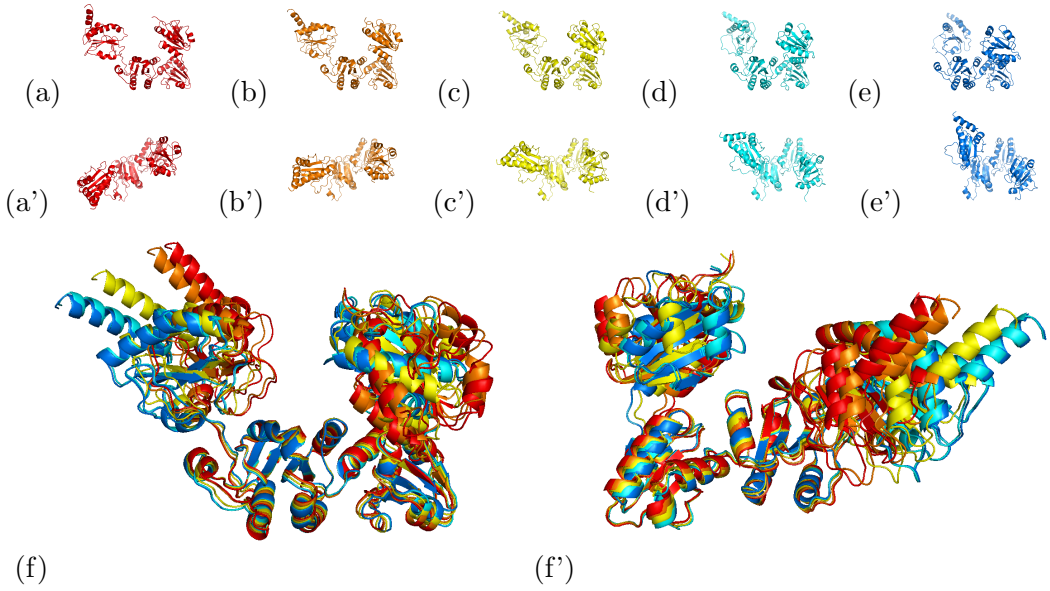
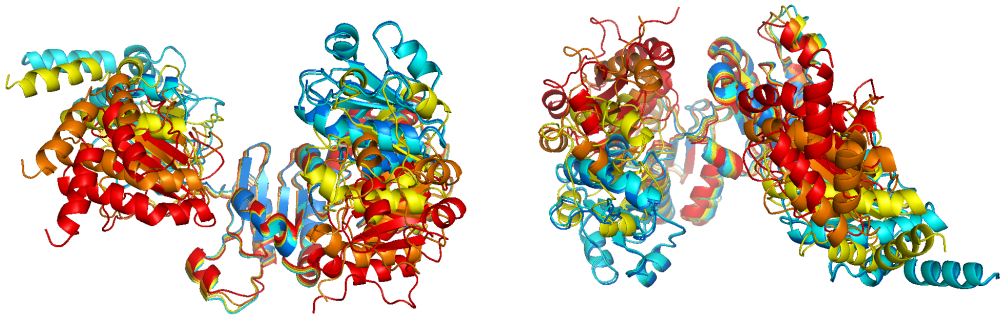


Figure 5.11: Conformational change for mode m_{10} of the yeast PDI (2B5E) structure. Conformers in panels (f) and (f') show the coordinated motion of domains ***a*** and ***a'*** with respect to ***b*** and ***b'***. In this case, domain ***a*** rotates over itself, whereas domain ***a'*** moves sideways. Despite mode m_{10} showing similar types of motions to other modes, the direction of motion ***a'*** and the axis of rotation for domain ***a*** are different compared to other modes.

kcal/mol also appears to be restricted but to a lesser extent, with the inter-cysteine distance reaching a constant value before reaching conformers ±2500. However, the subsequent simulations at lower E_{cut}, $E_{\text{cut}} = -0.522$ kcal/mol, $E_{\text{cut}} = -0.885$ kcal/mol and $E_{\text{cut}} = -1.412$ kcal/mol show increased and consistent protein mobility so that the maximum and minimum distance between active sites $d_{cc}^{(m_j)}$ increases notably. This behaviour is consistent for $E_{\text{cut}} < -0.522$ kcal/mol. The lowering of the E_{cut} affects protein motion differently when the protein structure is opening, i.e. from conformer 0 to −2500 or negative direction of motion, than when the structure is closing, i.e. from conformer 0 to 2500 or positive direction of motion. It appears that the high density of weak bonds at low E_{cut}, i.e. $E_{\text{cut}} > -0.522$ constrains the opening (Figure 5.7) of yeast PDI while only steric effects limit the closing.

The effects of varying the E_{cut} on the rotations and side translations shown in modes $m_8, \ldots, m_{11}$ also implies that the high density of bonds present at higher E_{cut} restrain domain rotation and translation. The motion of the initial structure always saturates after just a few conformers when the protein network is constrained by the lowest E_{cut}, i.e. $E_{\text{cut}} = -0.003$ kcal/mol and $E_{\text{cut}} = -0.015$ kcal/mol.



(a) (a')

Figure 5.12: Conformational change for mode m_{11} of the yeast PDI (2B5E) structure. The conformers in panels (a) and (a') show two axial views of the coordinated side motion of domains ***a*** and ***a'*** with respect to domains ***b*** and ***b'***. Mode m_{11} shows a larger motion of domain ***a'*** than domain ***a***. As for previous modes, the axis of motion for domain ***a'*** is quite different compared to modes $m_8 \ldots m_{10}$.

5.4 Discussion

5.4.1 Rigidity analysis: Domain recognition

The rigidity analysis computed on yeast PDI structure reveals very similar structural features to the ones described previously by biochemical and structural studies [63, 61]. Biochemical previously identified yeast PDI domain organization and flexible regions [61]. The complete crystal structure of Yeast PDI [63] revealed four well differentiated domains ***a-b-b'-a'*** and confirmed the existence of the linker region ***x*** between domains ***b'*** and ***a'*** that had previously been identified by experimental proteolysis [61]. Such distribution of domains and flexible regions corresponds with the rigidity and flexibility distribution shown by the complete and 'mini' RCD graphs 5.3 and 5.4, which show four rigid clusters that correspond to domains ***a-b-b'-a'*** and two main inter-domain flexible regions that link domain ***b'*** with ***a'***, and domain ***b*** with ***a***. The examination of the inter-domain contacts of the crystallised protein structure revealed that the interaction between domains ***b*** and ***b'*** is quite tight with a buried area of ($\simeq 700$Å^2). The inter-domain contacts between the base ***b-b'*** and domains ***a*** and ***a'*** are loosely connected with a negligibly small contact area of ($\simeq 200$Å^2) in each case [63]. Therefore, the rigid analysis captures not only the overall structural distribution of the domains but also more detailed features that have been previously identified experimentally.

5.4.2 Domain motion

It has been shown experimentally that the catalytic activity of yeast PDI requires a conformationally flexible molecule in order to perform its function [62]. This implies that the domains containing the active sites must move with respect to each other [72, 63, 62]. This prompted our choice of the inter-cysteine distance $d_{cc}^{(m_j)}$ as an initial measure to investigate more detailed mobility features. The monitoring of the inter-cysteine distance for the lowest frequency modes clearly indicates that yeast PDI undergoes a large conformational change for several modes. Hence, the span of conformational space that the protein explores appears to be wide (≈ 40Å). In addition to these large conformational changes, the methodology reveals more details about the protein motion. It identifies the ranges of motion for each normal mode and the type of motion for each of them while taking into account the constraints introduced by the network of bonds and by the stereochemical constraints.

The use of the inter-cysteine distance as a measure to investigate yeast PDI motion shows that the structure undergoes a large conformational change for the lowest frequency modes investigated. This gives an indication of how much the domains can move with respect to each other. This is particularly important since it makes possible to have an indication of the evolution of interatomic distances during protein motion. This implies that it is possible to answer specific questions for a given protein structure or guide experiments that would otherwise be impossible to tackle. For example, cross-linking experiments require the cross-linker ends to bind to the cystine groups on both domains to verify the distance between them. Hence, having an indication of the maximum and minimum distances between active sites makes it possible to narrow down the number of cross-linker lengths to choose from and therefore to reduce costs and experimental time. Likewise, the information provided by this methodology is very valuable to assist in deciding which fluorescent pairs are more suitable to undertake FRET experiments since the distance range of fluorescent efficiency for a given set of fluorescent pairs varies depending on the pairs available.

5.4.3 Comparison with experimental data

Several structural studies pointed out at the ***a*** and ***a'*** domain linkers flexibility [62, 63] without bringing consensus of which one is the most flexible. Proteolysis experiments [61] indicate that the ***a*** domain arm was cleaved off much easier than the other. On the basis of this result, the authors suggested that the ***a*** domain arm must be the most flexible and therefore that domain ***a*** will be moving the most.

The results obtained from the rigidity analysis indicate that both domain linkers have similar rigidity properties. Both regions become flexible so that the linkers can move freely for $E_{\text{cut}} < -0.522$ kcal/mol. However, the ***a*** domain arm still retains a small rigid α-helix within the flexible arm up to very low E_{cut}, i.e. $E_{\text{cut}} < -1.412$ kcal/mol. Meanwhile, the ***a'*** domain arm, i.e. the ***x*** region, appears to be fully flexible for $E_{\text{cut}} < -0.522$ kcal/mol. The geometric simulation results indicate that domain ***a'*** is moving the most with respect to the base defined by domains ***b-b'***.

It is not possible to make definitive conclusions on which linker is the most flexible one and which domain achieves the largest range of motion with the data available. However, a suggestion arising from the rigidity analysis and which is not in contradiction with previous experimental data [61] is that, although the ***x*** region is longer and both linkers are similarly flexible, the enzyme could find it easier to digest the ***a*** domain arm due to other structural or biochemical features rather than how flexible the region might be. For example, if the structure spends most time in a conformation that leaves the ***a*** domain arm more exposed.

A recent study [29] applying NMA to yeast PDI reported the motion using the same yeast PDI crystal structure (code 2B5E). However, the authors reported only the lowest frequency mode m_7 and gave an indication of the protein's motion without exploring the limits of motion. It is reassuring that their results correlate with the negative direction of motion for mode m_7 as reported here. The difference being that I choose the ***b-b'*** domains as a base and they use the center of domains ***a-b-b'*** as an origin to define the motion.

5.4.4 Cutoff energies and protein mobility

I reviewed the effects of generating geometric simulations using different E_{cut} from the rigidity analysis to investigate the mobility under different gradations of coarse graining, i.e. the number of residues considered to be rigid is different for each of the E_{cut} selected from the rigidity analysis as reported in Figure 5.4. By decreasing the E_{cut}, the number of hydrogen bonds participating in the hydrogen bond network decreases, i.e. the weaker bonds up to the chosen E_{cut} are removed. Therefore, by decreasing the number of bonds the protein network can potentially acquire improved mobility. The analysis of yeast PDI mobility over a set of normal modes and over a range of six E_{cut} shows restricted mobility for higher E_{cut}.

It is worth noting that inter-cysteine distance is not consistently increasing as a the E_{cut} decreases. Lower E_{cut} reduce the number of bonds that constraint protein mobility and therefore it is expected that this will increase protein mobility. However, the evolution of the inter-cysteine distance shown in Figures 5.7 and 5.8

is, for some modes, different of what it would be expected. For example, the inter-cysteine distance for mode m_{10} does not show a homogeneous increase in mobility as the E_{cut} decreases. The maximum inter-cysteine distance at $E_{\text{cut}} = -0.885$ kcal/mol as the motion is directed along the positive direction is $d^{(m_{10})}_{-0.885} = 33$Å, which is above $d^{(m_{10})}_{-1.412}$. Also, some modes show higher standard deviation values for a given E_{cut} and not for others without a clear pattern that could be extracted between them. It is unclear what could explain this behaviour. However, the limitations of fitting atoms to ghost templates and the lack of an energy minimisation procedure in FRODA have been previously highlighter and could be the responsible of these abnormalities.

5.5 Conclusions

It is very reassuring that the results from the rigidity and mobility analyses are in good agreement with previous structural and biochemical studies. Nevertheless, this also brought up new questions to be addressed. Three or four questions are the most urgent. First, would an all atoms simulation method like M.D. also reveal similar mobility features? Second, would the distribution of rigid and flexible regions hold during such simulation? In other words, would the intradomain structural distribution change during protein motion? And third, would the distance between active sites be comparable across different simulation methods? To address these questions I started up a collaboration with M. Bhattacharyya and Prof. S. Vishveshwara at the Indian Institute of Sciences (Bangalore-India). They carried out a 30ns and a 10ns MD simulations and extracted data to monitor the mobility of yeast PDI (2B5E). The results from these simulations are reported in chapter 6.

The range of motion defined in terms of active sites distance provides quantitative data that could be compared with experimental data but also guide experiments. In this regard, I performed cross-linking experiments to provide a rough measure of the inter-cysteine distance, see chapter 7. The inter-cysteine maximum and minimum distances $d^{(m_j)}_{max} \simeq 55$Å and $d^{(m_j)}_{min} \simeq 15$Å were a valuable guide to narrow down the choice of cross-linkers. Likewise, we also found the range of inter-cysteine motion provides useful information to narrow down the choice of FRET pairs that would be functional or have a better resolution within that distance range.

In summary, the investigation of yeast PDI mobility guided by normal modes has proven to identify the large conformational changes previously revealed by experimental data, provided new insights that had not previously been revealed and served as basis to guide experiments.

Chapter 6

MD simulations of yeast PDI

6.1 Introduction

Although protein mobility simulation methods have been evolving dramatically over the last decades, the gold standard during these years has been MD. Several coarse graining techniques have emerged with different capabilities and limitations, but usually faster time performances. Although the accuracy of coarse graining techniques has been held under certain scepticism by biologists and biochemists [27], there is a growing interest in using coarse graining models. However, there are still many questions to be answered in regard to the validation and limitations of coarse graining models.

This chapter reports the results from the 30ns and 10ns MD simulations carried out by our collaborators M. Bhattacharyya and Prof. S. Vishveshwara at the Indian Institute of Sciences (Bangalore-India) on yeast PDI (2B5E). The purpose is two fold: first, to present and compare the results and limitations of two protein simulation methods to identify the intra-domain variability, i.e. to compare structural similarity between each individual domain during an MD simulation and the rigidity distribution; and second, to compare the conformational space explored by MD and the hybrid coarse graining method (HCG) as presented in chapter 4.

6.2 Methods

6.2.1 Protein preparation

The structural properties of yeast PDI were examined by performing all atom MD simulations using the standard AMBER9 force field package and the same PDB structure (2B5E) used during the HCG simulations in chapter 5. Aqueous solvation

was modelled using the explicit solvent TIP3P water model and temperature maintained constant at 300K. The initial structure was stabilised by doing an energy minimization. Thereafter two MD simulation runs were performed. The first simulation was initiated from crystallographic coordinates to explore yeast PDI conformational space during a long-range 30ns simulation which took up to 94 days using 16 processors on an Intel Xeon Sun Microsystems cluster. The purpose of this simulation was to identify the conformational space that yeast PDI is able to explore, extract structural and mobility data to compare with the geometric simulations presented in the previous chapter. The second simulation was initiated using the most closed conformational structure defined by the geometric simulation over mode m_7 presented in chapter 5 and lasted 10ns. The purpose of this simulation was to clarify if the most closed conformational state obtained by the coarse graining method is a stable structure. Note that the initial results on the MD 30ns simulations identified a conformation with a minimum inter-cysteine distance of $d_{min}^{(MD)} \simeq 22$Å.

The approach of coarse graining rigid regions as pseudo units to minimise the computational expense assumes that the thermal variations of the residues within the selected rigid clusters is small enough to consider that the residues in such regions are moving together with no intra-cluster relative motion. Obviously the rigid clusters are considered to be strictly rigid during the coarse grained simulation and the domains move along the lowest frequency modes as single units so the intra-domain RMSD values are the same during these simulations. The hypothesis that a rigid cluster stays constrained throughout moving along a normal mode has not been tested to our knowledge. Therefore, comparing with an all atom simulation technique to corroborate this fact seems the most straightforward available path. Hence, during the 30ns MD simulation on the same yeast PDI structure the structural variation was monitored in terms of RMSD values for the whole protein and for each domain.

Although I highlighted the limitations of using RMSD as a measure for characterising conformational change in chapter 4, specially when comparing between different proteins, RMSD is generally a useful measure to compare structural variations of the same protein.

6.2.2 Inter-cysteine distance

In the previous chapter I introduced the inter-cysteine active sites distance as a measure to identify if yeast PDI undergoes a large conformational change. Although using the inter-cysteine distance is a simple approach to identify protein motion, it provides interesting and useful information in this case. The results in chapter 5

Domain	Residue	Atom ID
a	61	599
a	90	1039
a'	406	5979

Table 6.1: Yeast PDI domains containing the residues and α-carbons ID numbers for the cysteine residues of interest. This table summarises the domains where the active site cysteine residues are located in the tertiary structure and the atom identification numbers of the α-carbons that are used to track the active sites relative motion in the ***a*** and ***a'*** domains of yeast PDI. I use residue number and atom ID interchangeably but the inter-cysteine distance is always accounted using the α-carbons as a reference.

show that the largest conformational change that yeast PDI undergoes along the lowest frequency modes, is a double-hinge motion where the active site cysteine residues move away and towards each other. Thus, during the MD simulation the inter-cysteine distances is monitored. The atoms ID used as reference points are the α-carbon's atom ID chosen to represent the cysteine groups of interest, see Table 6.1. There are two active sites in the protein structure that contribute to the catalytic activity of yeast PDI, residues 61 and 64 in domain ***a*** and residues 406 and 409 in domain ***a'***. The other cysteine residue of interest is buried within the domain ***a*** structure (residue 90). For the purposes of tracking protein motions, the α-carbons of interest are reported in Table 6.1 and shown in Figure 6.1. These α-carbons will be used as end points to measure the "active sites" distance $d_{cc}^{(MD)}$.

6.3 Results

6.3.1 RMSD: structural variation

Initially, when comparing the motion of the whole structure, the RMSD values increase up to $\approx$ 8Å in just $\approx$ 0.25ns of MD simulation, which indicates that the structure is very inclined to move away from the initial crystal coordinates. Thereafter, the protein structure explores the conformational space, as shows Figure 6.3a, with RMSD values oscillating around $\approx$ 7Å for most of the remaining simulation time and reach a maximum values of $\approx 10 - 12$Å but only very briefly.

The evolution of the RMSD values from the HCG method and for the lowest frequency mode m_7 at a $E_{cut} = -0.522$ kcal/mol, show a smooth increase up to RMSD $\approx$ 10Å. At this point the increase in the standard deviation for the negative direction of motion suggested that the protein is hitting stereochemical constraints,

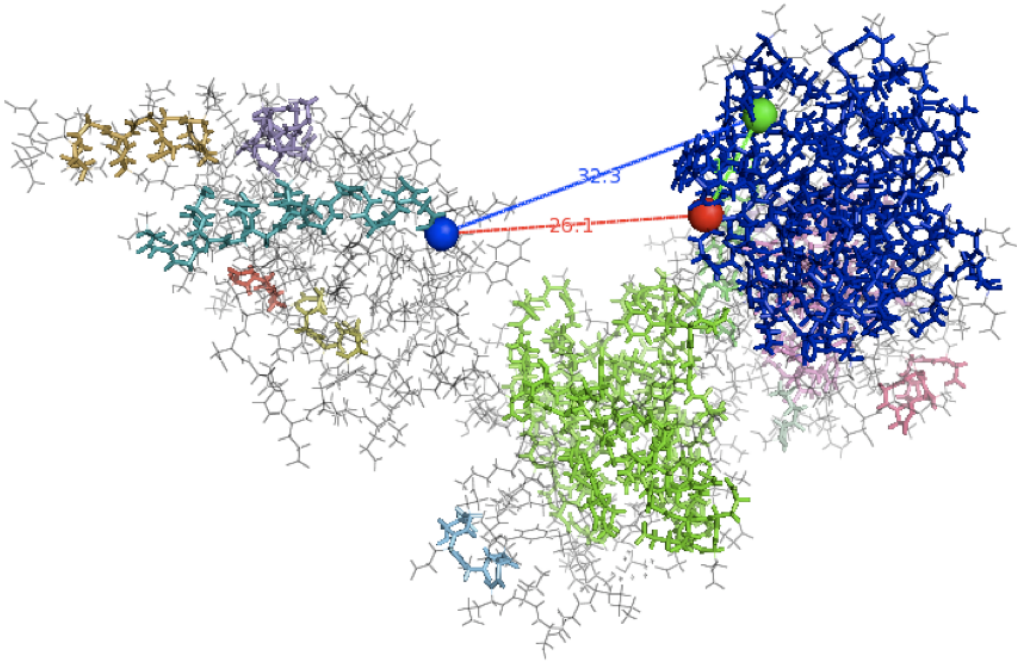


Figure 6.1: Yeast PDI tertiary structure from HCG simulations. Each coloured area identifies the residues belonging to a rigid cluster as defined by FIRST (see chapter 5.5) and the coloured spheres indicate the cysteine α-carbons of interest. The blue sphere corresponds to the α-carbon atom of the cysteine active site within the ***a'*** domain (ID 5979), the red sphere to the α-carbon atom also of a cysteine active site but within the ***a*** domain (ID 599) and the green sphere corresponds to the α-carbon atom of the buried cysteine in the ***a*** domain (ID 1079). The coloured lines represent the inter-cysteine distance between the cysteine's α-carbon atoms (599-5979) and between the (5979-1039) atoms. The red dashed line corresponds to the distance between the two active sites in domains ***a'*** and ***a***; and the blue line between the active site in domain ***a'*** and the α-carbon atom of the buried cysteine in domain ***a***. The green dashed line between the two cysteine's α-carbon atoms in domain ***a*** goes through the rigid cluster since the (ID 1039) α-carbon atom is buried within the rigid cluster. The inter-cysteine distances shown for this conformer are $d^{HCG}_{599-5979} = 26.2$Å and $d^{HCG}_{1079-5979} \simeq 32.3$Å.

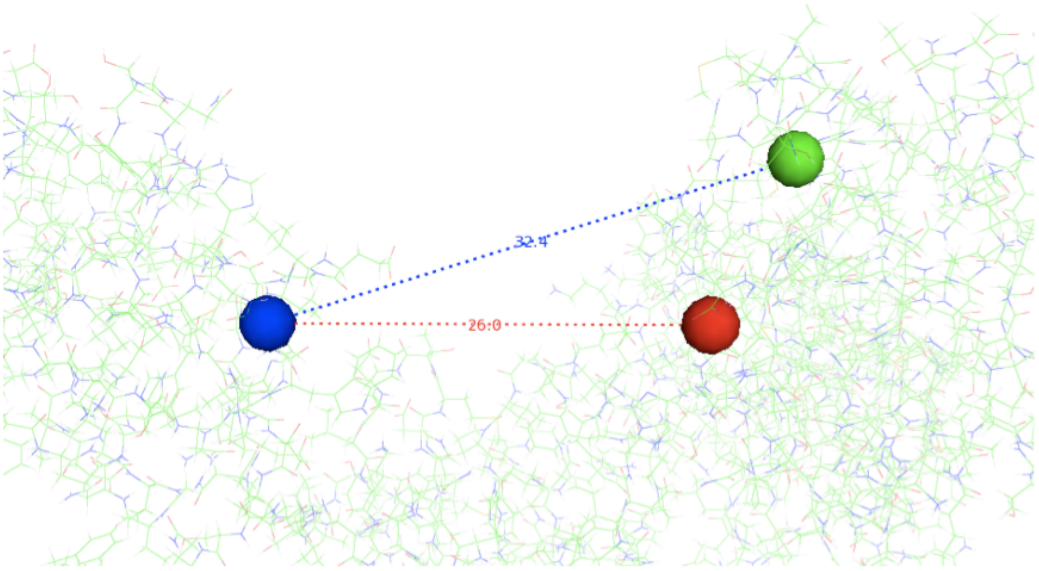


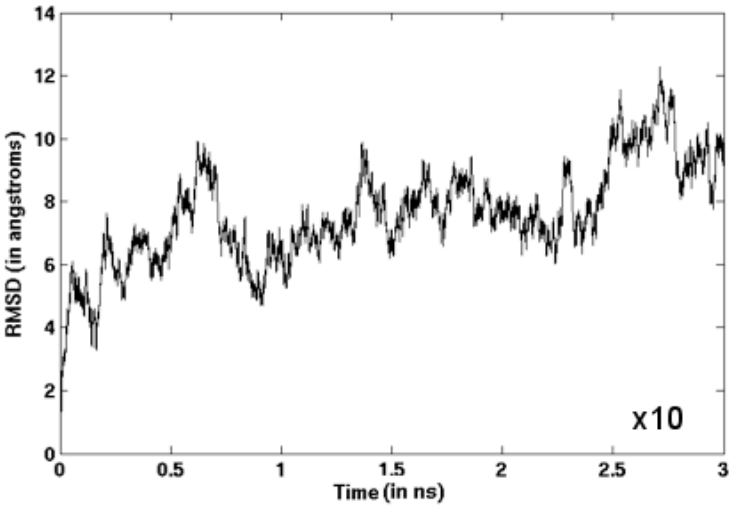
Figure 6.2: Close up view of the yeast PDI tertiary structure from MD simulations. The coloured spheres and coloured lines identify the same cysteine α-carbon atoms and distances between them shown in Figure 6.1. The inter-cysteine distances shown for this particular MD conformer are $d^{MD}_{599-5979} = 25.8$Å and $d^{MD}_{1079-5979} \simeq 33.9$Å.

see Figure 6.3b.

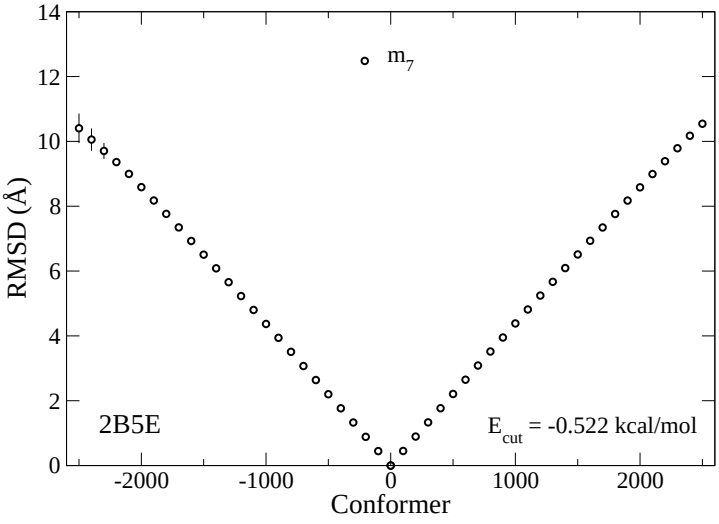
Although the methodologies to investigate protein motion are different, the RMSD maximum values are similar in both cases. Each method explores the protein's conformational space differently. The hybrid coarse graining method explores protein mobility using a bias motion along a normal mode, whereas MD explores all the available conformational space during each simulation using force fields.

6.3.2 Intra-domain RMSD

The intra-domain MD RMSD values in Figure 6.4 show that there is a rigidity gradation across the domains. Domain ***a*** shows the lowest structural variation and ***a'*** is the domain showing the highest. Lower RMSD values indicate that the domain structure is more stable and with less flexible regions or residues that contribute to intra-domain mobility. Therefore, domain ***a*** is the most tightly bound domain according to the MD simulation. This correlates with the rigidity analysis previously shown in chapter 4 where it was shown that there is a rigidity gradation across domains, and that domain ***a*** is the most rigid and domain ***a'*** the most flexible one. Therefore the RMSD intradomain variations throughout the MD simulation show similar rigidity gradation across the domains to the rigidity distribution identified



(a)



(b)

Figure 6.3: RMSD as a function of time for MD simulation and versus conformer generated during the HCG simulation for yeast PDI. MD data courtesy of M. Bhattacharyya and Prof. S. Vishveshwara. (a) The values are obtained by comparing the initial structure from crystal coordinates to the conformers generated during a 30ns MD simulation. After starting the simulation, at time $\approx$ 0ns, the protein quickly moves to reach RMSD values of 6Å and then up to 10Å in a relatively short period of time. Thereafter, the conformational RMSD values oscillate around $\approx$ 6Å for most of the simulation, with the exception of a few higher values between $8 - 12$Å. Panel (b) shows the RMSD values relative to the initial structure against the conformers generated using the hybrid coarse graining method presented in chapter 4.5. The data presented corresponds to the projection of the initial structure along the lowest frequency mode m_7, for the positive and negative direction of motion and at a $E_{cut} = -0.522$ kcal/mol.

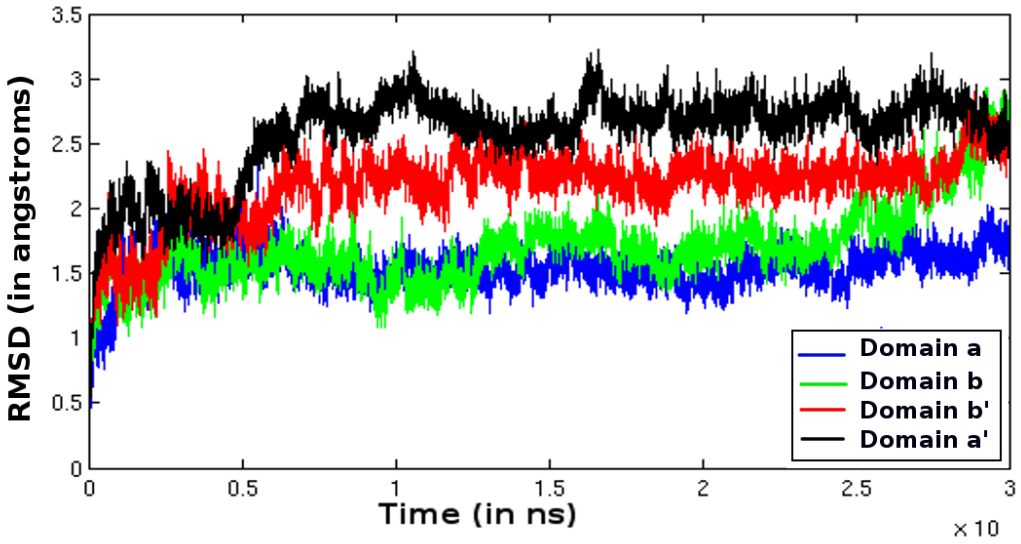


Figure 6.4: RMSD as a function of simulation time for yeast PDI domains. MD data courtesy of M. Bhattacharyya and Prof. S. Vishveshwara. The values are obtained by overlapping each domain from the initial crystal structure with itself from the conformers generated during a 30ns MD simulation. Domain ***a'*** exhibits the highest values and domain ***a*** the lowest.

using FIRST.

6.3.3 Monitoring the inter-cysteine distances

During the MD simulation the inter-cysteine distance was monitored for the three cysteine α-carbon pairs of interest $(61-406)$, $(90-406)$ and $(61-90)$, see Figure 6.5. Residues $(61-90)$ belong to domain ***a***, the most rigid of the four domains. Therefore, the inter-cysteine distance $d_{599-1079}$ between the α-carbons $(599-1079)$ of the residues $(61-90)$ are expected to remain constant throughout the 30ns simulation. Indeed the results in Figure 6.5 show that the inter-cysteine distance remains constant $d_{599-1079} \simeq 11$Å during the MD simulation.

The inter-domain distances between the α-carbons atoms $(599-5979)$ and $(1039-5979)$ of the cysteine groups $(61-406)$ and $(90-406)$ vary approximately between a *minimum* of $d_{min}^{(MD)} \simeq 22$Å and a *maximum* of $d_{max}^{(MD)} \simeq 70$Å. Hence the range of motion for the cysteine groups with respect to each other is ≈ 50Å. It is worth noting that $d_{1079-599}$ and $d_{599-5979}$ vary with respect to each other during the simulation. For example, for $t = 0.5$ns the $d_{1079-599}$ and $d_{599-5979}$ are almost identical whereas at $t = 2$ns the difference is ≈ 11Å. This indicates that there must

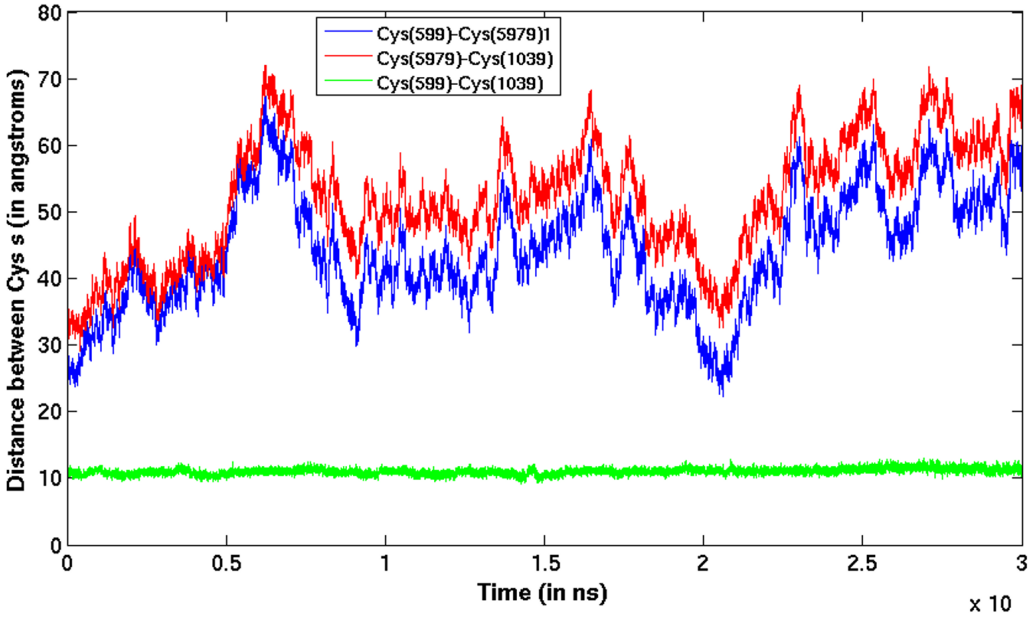


Figure 6.5: Evolution of inter-cysteine distances between cysteine pairs for the 30ns simulation. MD data courtesy of M. Bhattacharyya and Prof. S. Vishveshwara. The three cysteine residues are (61 and 406) for the active sites and (90) for the buried cysteine group. The α-carbon atoms ID numbers corresponding to these residues are $(599 - 5979)$ for the active sites at the ***a*** and ***a'*** domains respectively, and (1039) for the buried cysteine group in the ***a*** domain. The intra-domain distance between cysteine α-carbon $(599 - 1039)$ is measured and shown to remain constant during the simulation, whereas the inter-cysteine distance between α-carbon atoms $(599 - 5979)$ and $(5979 - 1039)$ changes according to the conformational change adopted by the protein structure.

be at least two types of motion occurring during the MD simulation, a double-hinge motion to account for the large inter-cysteine distance range of ≈ 50Å and a domain rotation or translation to account for the changes in relative values between $d_{599-5979}$ and $d_{1079-599}$ as illustrated in Figure 6.2. Hence, the comparison of results between the MD and the HCG simulations reveal that the minimum and maximum inter-cysteine distances are: $d_{min}^{(HCG)} \simeq 15$Å and $d_{max}^{(HCG)} \simeq 55$Å , i.e. a maximum range of ≈ 40Å, and $d_{min}^{(MD)} \simeq 22$Å and a maximum of $d_{max}^{(MD)} \simeq 70$Å. This suggested that perhaps that the most closed conformer (obtained from the HCG method) is not a stable structure. Since MD simulations showed no conformers with such low inter-cysteine distance. This would imply that the hybrid coarse graining method is able to reach a conformation that goes unnoticed for the 30ns MD simulation due to the different simulation approach.

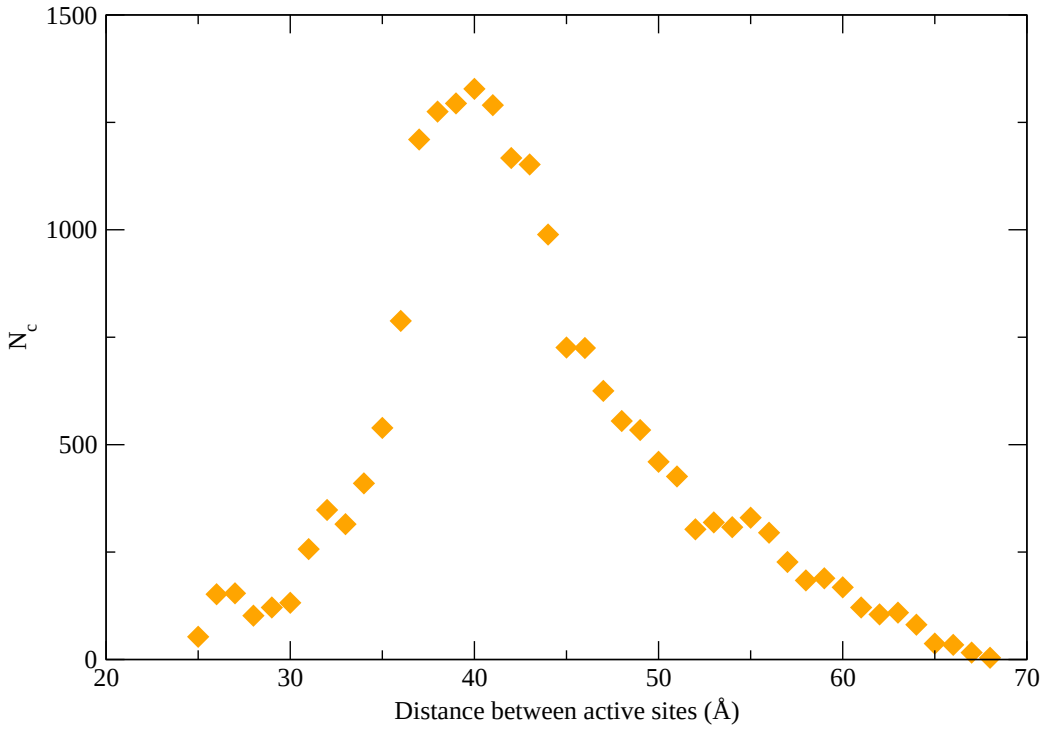


Figure 6.6: Conformers with same active sites distance during the MD 30ns simulation. MD data courtesy of M. Bhattacharyya and Prof. S. Vishveshwara. The number of conformers (N_c) that display the same distances between cysteine active sites in domains ***a*** and ***a'*** reveal that the preferred inter-cysteine distance is ≈ 40Å.

6.3.4 Stability of the closest conformer

The discrepancy between MD and HCG minimum distance prompted a 10ns MD simulation using the closest conformer from the HCG simulation as an initial input to determine if the most closed conformer was a stable structure. The results of the 10ns simulation as shown in Figure 6.7 indicate that the inter-cysteine distances remains stable throughout the 10ns simulation, which implies that the protein structure is stable.

6.3.5 Preferred inter-cysteine distance

The inter-cysteine distance histogram for the MD simulation shown in Figure 6.6 reveals that yeast PDI spends most of the 30ns simulation time displaying conformations that have an inter-cysteine distance between $\simeq 35$Å and $\simeq 50-55$Å. And it spends very little time in conformations with an inter-cysteine distance close to the initial inter-cysteine distance $d^0_{599-5979} \simeq 27$Å from the initial crystal structure.

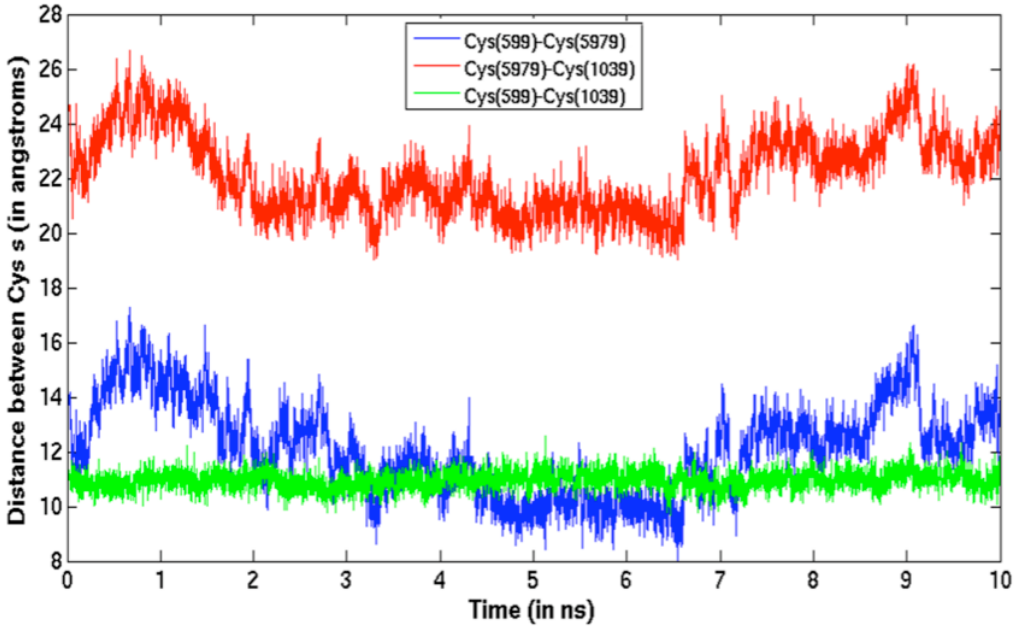


Figure 6.7: Evolution of inter-cysteine distances between cysteine pairs for the 10ns simulation. MD data courtesy of M. Bhattacharyya and Prof. S. Vishveshwara. The three cysteines residues and α-carbon ID numbers are the same as before, i.e. atom ID (599-5979) for the active sites at the ***a*** and ***a'*** domains respectively, and (1039) for the buried cysteine α-carbon in the ***a*** domain. The intra-domain distance between cysteines (599-1039) again remains constant during the simulation. In this case, the inter-cysteine distance between (599-5979) and between (5979-1039) change but the inter-cysteine distances vary within a range of $\approx 6-7$Å, i.e. between $\approx 20-36$Å for the (5979-1039) pair and between $\approx 9-16$Å for the (599-5979) pair. However, the two inter-cysteine distances variations appear to be coordinated or symmetrical.

6.4 Discussion

One of the main differences between the MD and the NMA based HCG method is that NMA identifies all the possible directions or modes of motion and ranks them according to their amplitude. A given direction of motion (or mode) can be used to guide protein motion individually or as a combination of various modes. By contrast, an MD simulation explores all the conformational space accessible to the protein according to the force fields used and produces a single trajectory of the protein structure along its conformational space. Despite these methodological differences between modelling using an all atom force field method and a coarse

graining approach to model protein motion, both methods identify yeast PDI's ability to undergo a large conformational change, identify the most rigid and most flexible domains, and give a good indication of the types of motion of yeast PDI.

The two approaches reveal results that are only possible for one of the two methods. On one hand, the 30ns MD simulation data reveals in Figure 6.6 yeast PDI's preferred conformational state in terms of inter-cysteine distances. On the other hand, the HCG method is able to explore protein motion at a low computational cost for directions of motion defined by normal modes of motion. Hence, computational resources to run the simulations are much less for the HCG method than for MD simulations; that is CPU time of hours or days for the HCG method and weeks or months for MD simulations. Second, the HCG method is able to identify a conformational state that may be difficult for MD simulations to reach.

6.5 Conclusions

The comparison of the protein simulations over yeast PDI clearly reveals several points. First, there is a consensus in identifying that yeast PDI undergoes a large conformational change and that there is an double-hinge and a domain rotation type of motion. Second, the rigidity distribution of each domain is well captured by both methods. Third, the short MD 10ns simulation confirms that the most closed structure identified by NMA is stable and suggests that there part of the conformational space was not explored during the MD 30ns simulation.

Chapter 7

Crosslinking experiments with yeast and human PDI

7.1 Introduction

In the previous chapters I have reported the results of investigating yeast PDI using a HCG method and MD simulations. Both method agreed that yeast PDI undergoes a large conformational change but there are differences in identifying the maximum and minimum inter-cystein distances. There are at least two experimental techniques that can be used to test these results, protein cross-linking and FRET. It is possible to crosslink the active sites of yeast PDI in the ***a*** and ***a'*** domains with the aid of Bismaleimide constructs which bind to the exposed active cysteine residues. By using constructs of different lengths it should be possible to identify a minimum inter-cysteine distance. Alternatively, a more sophisticated experimental technique that will allow to identify inter-cysteine distances but also to monitor the dynamics of yeast PDI is FRET. By attaching two fluorescent markers, one on each active site, it is possible to monitor the distance between the two markers and therefore monitor the relative distance between active sites[1].

Here I report the crosslinking experiments we carried out on yeast and human PDI using Bismaleimide constructs containing six (BM-6) and two (BM-2) α-carbon groups respectively and space arm distance separating the maleamide groups of approximately 12Å and 6Å respectively. Since the maleamide groups bind to PDI's exposed active sites and if the inter-cysteine is close enough they will crosslink to

[1]The experimental work presented in this chapter was carried out in partnership with John Blood, a fellow PhD student, at the Structural Biology laboratory, School of Life Sciences, Warwick. Hereby, I acknowledge John Blood's guidance and the joint work in carrying out during the experimental work here presented.

both sites.

7.2 Methods

7.2.1 Sample preparation: Cell inoculation

To express yeast PDI with a six histidine tag (his-tag), the plasmid 'LR370' (1μl) was used to transform 1μl of competent cells containing the E. coli strain BL21 (pLysS). Then the mixture was left to rest for 15 minutes. In the meantime, Chloroamphenicol antibiotic ($\approx$ 25μl) was spread onto an agar Lysogeny broth (LB) plate (a nutritionally rich medium primarily used for the growth of bacteria) that already contained the antibiotic Ampicillin. After sterilising the bench with alcohol and using a bunsen burner to create a sterile space, the sample of competent cells transformed with the plasmid was added to the agar plate and left overnight at 37^{o}C.

7.2.2 Sample preparation: Cell growth

On the next day, a single colony was picked from two agar plates and inoculated with 50ml of LB medium, 50μ l of Ampicillin and 50μl of Chloroamphenicol in a 250ml flask. The flax was placed in the incubator overnight stirring at 200 r.p.m. and 37^{o}C for the culture to grow.

The optical density (OD) of the cell culture was measured at 600nm to identify the optimal culture density necessary to inoculate a larger volume of culture. The spectrometer was calibrated using sterile water. A sample from a 100μl cell culture diluted in 900μl of water was measured to obtain the OD and identify the concentration needed to inoculate a larger culture volume. This step is required to obtain cells replicating at a favourable life-cycle point so that the PDI yield is maximised. The new culture was grown in two 2l flasks containing 400ml of LB, 400μl of Chloroamphenicol and 400μl Ampicillin and 4.4ml sample of the previous culture. Then, the culture was placed in the incubator at 37^{o}C and the OD was monitored regularly after one hour to identify the start of the exponential phase of cell growth. When cell growth reached the optimal rate, i.e. OD $= 0.45 - 0.5$, the culture was induced with Isopropyl-β-D-thiogalactopyranoside (IPTG), which is a lactose like compound that allows control of gene expression for PDI production. The culture was incubated at 37^{o}C and 180 r.p.m. for approximately four hours. The culture was then centrifuged at 5000 r.p.m. for 20 minutes to harvest the cells which form a pellet. Thereafter the media was poured out and the pellet resuspended in 1/10 of

the original volume. Phosphate solution A buffer (20mM) was added to this mixture to stabilise pH.

The re-suspended cell pellet was stored in the freezer over night and then brought to room temperature by placing it in a water bath at room temperature. The overnight freezing and 30 seconds of sonication lyse (breaks down) the cells. The lysed cells were centrifuged again to concentrate all the insoluble cell debris down to a pellet and leave the soluble protein in the media.

The purification of PDI from the soluble part of the cells was achieved using an immobilised metal affinity chromatography (IMAC) column. The IMAC column was constructed using a syringe as a container and a piece of glass wool acting as a filter. Then 6ml of Sepharose gel was added onto the glass wool and rinsed with 25ml of pure water after the gel had sank for 10 minutes to remove any dirt that would be deposited. Subsequently 2ml of 0.2M Nickel Cloride ($NiCl_2$) was added to the gel. The Ni binds to the Sepharose gel and will also bind to the hexa-histidine groups that are expressed at the beginning of the PDI sequence. Finally, 25ml of Acetate (0.5M NaCl at pH 3.0) was added into the syringe to remove the weakly bound Ni and then 25ml of buffer A to stabilise the pH and leave the IMAC ready for use before adding the solution containing the cell soluble protein media.

The total protein media from the soluble part of the broken up cells was then loaded into the IMAC column so that it will flow through the gel and the histidine tag binds to the gel. 10ml of 25mM Imidazole, 0.5M Sodium Cloride and 20mM Sodium phosphate (pH 7.3) were added to the column followed by 25ml of Buffer A with a low salt solution (20mM Sodium Phosphate pH 7.3) to clear up the proteins and cell debris that are not attached to the column. The PDI bound to the Ni^{2+}-sepharose gel was then eluted using 25ml 50mM of Ethylenediaminetetraacetic acid (EDTA), 20mM sodium phosphate (pH 7.3). The OD of the elute sample was monitored at an amplitude of 280nm (A_{280}) to identify PDI in the sample and a SDS-PAGE gel experiment was carried out to corroborate that the sample contains PDI. In order to remove the nickel and EDTA, the eluted sample was placed in a dialysis tube-membrane.

7.2.3 Ion exchange chromatography

The next purification step was performed using ion exchange chromatography (IEC). This technique allows to separate proteins with different binding charge or strength. The proteins flow down the IEC column and their charge determine how strongly bound they are to the column's electrically charged resin. The IEC column was connected to a pumping system to circulate the solutions through the column and

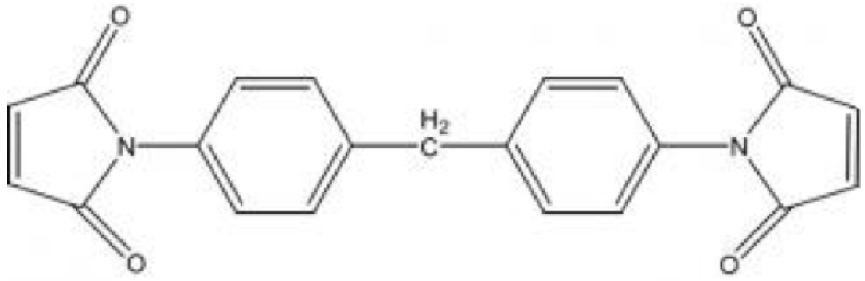


Figure 7.1: Bismaleimide construct. The Bismaleimide constructs bind to the active sites residues represented by the group at the chain. The maleimide groups bind to the active residues and to the $(CH_2)^N$ groups. The number of $(CH_2)^N$ groups determines the length of the space arm distance separating the maleimide groups.

a system to collect fractions of samples according to the protein's charges. The column was first cleaned and equilibrated by filtering a low salt buffer (using buffer A) to get rid of the lightly bound proteins. Then a salt gradient was applied over time using a mix of high and low salt concentration buffer (buffer B). The process starts with a flow through of 100% buffer A and 0% buffer B and slowly increasing the percentage of buffer B meanwhile reducing the % of buffer A until reaching 0% buffer A and 100% buffer B. During this process the proteins bound to the resin are progressively being collected into fraction tubes according to their ionic binding strength.

7.2.4 Calculating protein concentration

Protein concentration was determined using Beer-Lambert law:

$$P_c = (A_{280} \div (\epsilon d))D \tag{7.1}$$

Where P_c denotes protein concentration [mol $\cdot$ l^{-1}], (A_{280}) is the sample absorbance at 280nm (no units), ϵ the extinction coefficient [$M^{-1} \cdot cm^{-1}$], 'd' the path length [cm] and D the sample dilution. Then the yield was calculated using:

$$Y = P_c \cdot V_o \tag{7.2}$$

Where V_o denotes the total volume collected and Y the yield. Finally the total mass (M) of the protein sample was calculated using the yield and the protein's molecular weight (m) as follows:

$$M = Y \cdot m \tag{7.3}$$

7.2.5 Crosslinking experiment and SDS page gel

Sodium Dodecyl Sulphate Polyacrylamide gel electrophoresis (SDS-PAGE) was used to differentiate between proteins with different cross-sections. The gels are discontinuous and consisted of a resolving gel (pH 8.8) and a stacking gel (pH 6.8). The human and yeast PDI samples are crosslinked using two different Bismaleimide constructs containing six (BM-2) and two (BM-6) α-carbon groups respectively as shown in Figure 7.1. The "space-arm" distance separating the maleamide groups is approximately 6Å and 12Å respectively. The maleamide groups bind to PDI's exposed active sites and if the inter-cysteine is close enough they will crosslink to both sites. When the samples are reduced and unfolded after adding DTT (a strong reducing agent) and SDS (a denaturing detergent), the crosslinked constructs display a smaller cross section than the reduced structures. When crosslinked and not crosslinked samples travel through the SDS-gel they travel different distances due to their different cross sections. The speed of the samples across the gel depends on the voltage applied and on the proteins cross sections. Hence, different proteins will travel through the SDS-gel at different speeds depending on their cross section. If comparing a crosslinked and not crosslinked protein, the cross sections will be different as shown in Figure 7.2, so that they will travel at different speeds. Thus, it is possible to identify if the two constructs are present in the sample with the aid of appropriate molecular weight markers.

Each of three different compounds were mixed with a yeast and human PDI protein volume, i.e. NEM (N-Ethylmaleimide, containing an Imide functional group), BM-2 and BM-6. The three sets of compounds for each PDI species are placed into a 30°C bath and 25μl samples are collected at intervals of 30, 40, 60 and 80 minutes for the yeast and human PDI containing either BM-2 or BM-6, and then at 80 minutes for the NEM containing yeast and human PDI containing compounds. Right after sample collection, SDS loading buffer (10% Mercaptoethanol) and of DTT (0.5M) were added. Thereafter, the samples are heated to 95°C for 10 minutes before placing 10μl of each mixture in an gel lane. Hence, gel lanes contain the following selected samples for yeast and human PDI: a marker (M), a sample of pure PDI as control (PDI), a sample of PDI with NEM and a series of four samples collected from the bath at the above-mention times for PDI with BM-2 and BM-6 respectively.

The samples run across the gel driven by means of applying a voltage of 60V for 20 minutes for the sample to run through the stacking gel and then a voltage of 180V for approximately 50 minutes until the dye marker runs out of the gel. Thus, the protein sample reaches closest to the end of the gel to allow for a better

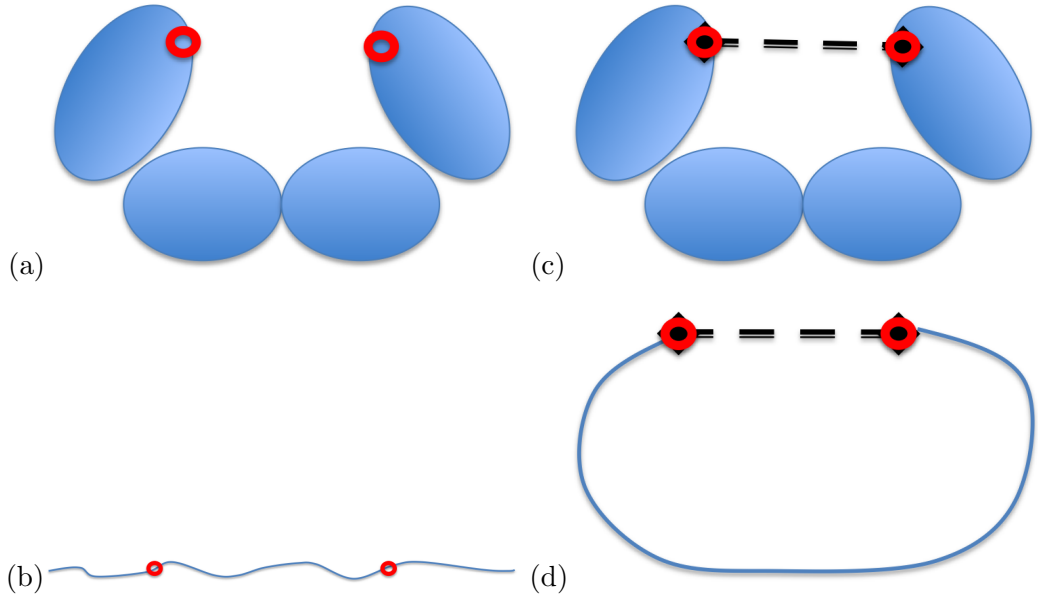


Figure 7.2: Cartoon representation of crosslinked PDI. The domains are represented by blue ellipses when folded and a thin line when unfolded, the active sites are shown as red rings and the crosslinker as a dashed black line.(a) Shows the folded PDI in its native state, in red the active sites. (b) Shows the denatured PDI sequence. (c) Shows a crosslinked PDI structure in its folded state. (d) Shows a denatured crosslinked PDI sequence. The samples that run through the gel will move faster when crosslinked than when unfolded as the protein cross section is smaller. Hence the crosslinked polypeptide chain offers less resistance to moving through the gel than the unfolded-not-crosslinked chain. Note: drawings not to scale.

resolution between the bands, i.e. between the crosslinked and not crosslined PDI structures. Finally, the gels were stained using Coomassie Blue dye.

7.3 Results

The two SDS-PAGE gels were produced one for yeast and one for human PDI are shown in Figure 7.3. The gels show a different results for the two PDI species. Human PDI appears to crosslink with BM-6 as the double blue bands below the 97kDa marker height of human PDI gel and BM-6 bands indicate. This is true for all the samples collected at different time intervals, although it is less clear for the sample collected at 30 minutes. This indicate that the reactive cysteine groups of the human structure can get closer than 12Å.

On the contrary there appears to be no crosslinking for yeast PDI for neither the Bismaleimide constructs BM-2 nor BM-6. Therefore, the minimum yeast PDI

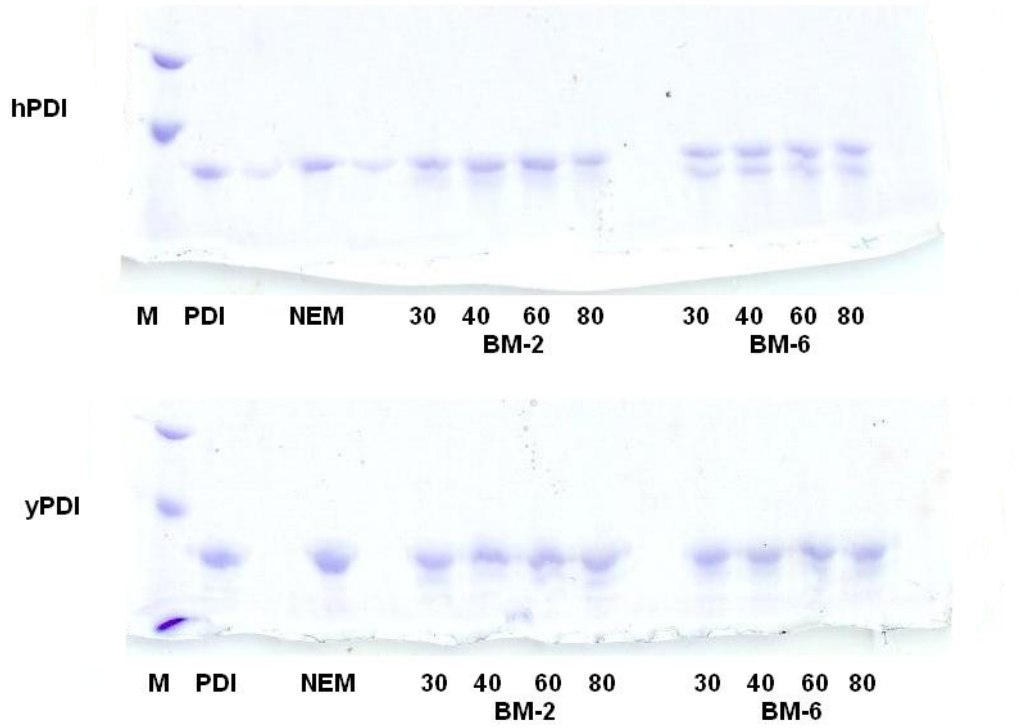


Figure 7.3: SDS-Page gel. The top gel reports the bands for human PDI and the gel at the bottom the bands for yeast PDI. The letter (M) denotes the lane for the marker and the blue bands within this lane denote markers at 97kDa for the top one and 66kDa for the bottom one. The direction that the samples move across the gel is from top to bottom and the typical band for PDIs is $\approx 56kDa$. The next lane contains a pure sample of human or yeast PDI as a control and without any modifications, (NEM) denotes the gel lane containing human or yeast PDI NEM treated and the remaining lanes contain samples of human or yeast PDI with the crosslinker BM-2 or BM-6 collected at the given time intervals from the bath. The gels initially suggest that human PDI crosslinks with BM-6 but not for yeast PDI.

inter-cysteine distance must be bigger than 12Å to the BM-6 distance of 12Å. These results agree so far with the simulation results presented in chapters 5 and 6. The HCG and the MD simulation showed that the minimum yeast PDI inter-cysteine distance was found to be $d_{min}^{(HCG)} \simeq 15$Å and $d_{min}^{(MD)} \simeq 22$Å.

7.4 Conclusions

Despite the simplicity of the experiment it is clear that the two protein structures have a different minimum inter-cysteine distance. Furthermore, despite the similarity of yeast and human PDI protein sequence, there is a functional difference between the two structures which could be explained by some different tertiary structure dispositions. However, since there is no crystal structure for human PDI available yet, it is not possible to compare the two structural features, i.e. the atomic spatial disposition, the rigidity distribution nor mobility between, using either HCG nor MD simulations. As soon as the structure is available I will be corroborating the experimental result here presented with mobility simulations on the human PDI crystal structure when it becomes available.

The use of the available crosslinkers reveal that the minimum inter-cysteine distance for yeast PDI is above 12Å. Although this results do not confirm the predicted HCG and MD minimum inter-cysteine distance for yeast PDI, i.e. 15Å and 22Å respectively, they do not contradict the predictions. In order to test the accuracy of the predictions it will be necessary to employ a range of longer crosslinkers, e.g. using a crosslinkers of lengths 15Å 17Å ... 27Å.

Chapter 8

Conclusions

8.1 Rigidity analysis

The analysis and comparison of rigidity distribution for different proteins in chapter 3 revealed that proteins during a hydrogen-bond dilution can show a pattern of rigidity loss that resembles the patterns of rigidity loss of glassy networks. Furthermore, the results revealed how small structural variations could lead to network variations that alter the rigidity distribution of proteins. A natural follow up of the results presented in chapter 3 has been to investigate the rigidity distribution of HIV-1 protease when bound to different ligands [79, 80]. These studies reveal the effects on rigidity that different drugs have on the rigidity of the protein flaps which cover the active site. These results add to the contribution made by other studies [17, 43, 44] that successfully applied FIRST rigidity analysis to identify structural properties of proteins, e.g. fitting of low-resolution cryo-EM maps of proteins [51], the study of viral capsid assembly [49], the use the rigidity distribution as a basis for coarse graining protein motion to simulate large biomolecular motions using and elastic network models [21, 39] and by ourselves as presented in chapter 4 and the identification of the folding core of proteins [44, 17].

Recently, FIRST has been made available as part of the flexweb.asu.edu website portal [81] where it is possible to upload a structure of choice and obtain its structural information analysis. The rigidity analysis using FIRST to obtains structural information and the wide range of applications to other closely related fields highlights the usefulness of the method. However, there is a wide range of improvements and new applications possible, from the most simple ones related to further enquires on protein rigidity to more methodological ones.

8.2 Geometric simulations

In this thesis I have reported the use of FIRST/FRODA and normal modes of motion from the elastic network model to geometrically explore the conformational space of proteins and obtain several insights about protein motion. This hybrid method is able to explore protein motion by integrating both rigidity constraints from FIRST and low-frequency mode eigenvectors obtained using ELNEMO, into the geometric simulation FRODA to simulate large biomolecular motions of proteins. Although a similar method has been previously reported [21, 39], it is the first time that the methodology has been applied to a large set of proteins to investigate their full conformational space. A great advantage of the method is that it makes it possible to achieve significant amplitudes of motion with only a few CPU-hours of computational effort even for a pentameric pore protein with more than 1600 residues. The geometrical simulations permitted to investigate protein motion and qualitatively differentiate three different types of protein motion and propose a new measure to quantify protein motion that takes into account how much the protein as a whole changes position rather than an minimized average like the RMSD, which can be susceptible to protein shape. The xRMSD measure accounts for the total deviation of the protein from the initial structure. Therefore, it is not a measure of structural similarity like RMSD, but a measure of protein mobility. A natural follow up of this research would be to investigate a wider set of proteins and identify whether the new measure quantifies the mobility of proteins.

8.3 Large conformational changes of yeast PDI

The hybrid coarse graining method was put to the test to investigate the conformational space of yeast PDI protein, see chapter 5. Each normal mode defined a direction of motion that was explored to identify the motion limits and the effects of using different networks of hydrogen bonds by using different E_{cut} from the rigidity analysis. The rigidity analysis identifies the modular nature of the protein and the geometric simulations identifies the large conformational change suggested by previous experimental data [63]. The distance between the active sites of yeast PDI provided a rough measure to quantify PDI motion and confirms that the protein undergoes a large conformational change. Further, the use of different E_{cut} showed that the protein mobility is impaired at high E_{cut}. Although initially it was thought that such limitation was due to the high density of hydrogen bonds constraining protein motion, recent work [38] highlighted several limitations of FRODA. For ex-

ample, the enforcement of constraints procedure in FRODA does not guaranty that the number of constraint violations are reduced at each step. Therefore, although projecting the initial structure along the normal modes has confirmed yeast PDI large conformational change, further investigation will be advisable to prove the robustness and limitations of the enforcement of constraints procedure by FRODA.

Furthermore, in chapter 6 the MD simulations showed that the active sites distance between α-carbons atoms $(599-5979)$ vary approximately between a *minimum* of $d_{min}^{(MD)} \simeq 22$Å and a *maximum* of $d_{max}^{(MD)} \simeq 70$Å . Since the motion along normal modes using FRODA is not modulated by force fields but limited by the stereochemical constraints it is expected that protein motion will reach larger conformational changes. However, in the series of simulations here presented the maximum inter-cysteine distance appears to be bigger during the MD simulations. On the contrary, the minimum intra-cysteine distance is smaller for the HCG than for the MD simulations, which is to be expected. MD will seldom guide the protein to energetically not favourable conformations whereas HCG method will explore the stereochemically accessible space as is. The experiments presented in chapter 7 using cross-linkers to identify the minimum distance between active sites revealed that the minimum intra-cysteine distance for yeast PDI is longer that the largest cross-linker used, i.e. a distance of 12Å between the two maleimide groups. Longer cross-linkers are needed to identify such minimum intra-cysteine distance.

The HCG method has proven to reveal interesting and useful results. It allows to project a protein along the pathway defined by a normal mode and generate the corresponding conformers along the pathway by using a minimum of computer resources, typically hours of CPU time. However, it will be revealing to investigate further whether the more restrictive motions displayed by the HCG method are due to FRODA limitations and to which extend do they affect protein mobility for simulations at different E_{cut}. Hence, new simulations to test the robustness of the mobility limits will be advisable. In regard to yeast PDI mobility predictions, it will be interesting to perform mobility simulations with other methods which can circumvent FRODA's issues and compare with new experiments using longer cross-linkers or FRET experiments to determine the accuracy of yeast PDI mobility predictions.

The minimum and maximum inter-cysteine distance predicted by the HCG and MD, and the predicted most likely inter-cysteine distance by the MD simulation are very useful results to provide a good indication for the type of fluorophore pairs that are functional within the distance ranges that we wish to explore. The distance range that a given pair of fluorophores can explore varies depending on

the fluorophore pairs, e.g. a given set of pairs will be able to explore motion for distances between 45 to 55Å. Hence, using the inter-cysteine distance ranges will be very useful to narrow down the number of useful fluorophore pairs. Single-molecule FRET experiments with alternating laser excitations (ALEX) is a new and suitable technique to investigate the relative motion between yeast PDI cysteine groups. Currently, members of my co-supervisor's group, Prof. R.B. Freedman, are in the process of bringing forward such experiments. Their choice of fluorophore pairs based on the HCG and MD simulations is Atto550–Atto647N, Atto550–Atto665 and Dylight488–Atto665. Each pair has a mid point distance where the resolution is optimal. The mid points of the effective ranges are 65Å, 60Åand 39Årespectively and they are quantitative approximately 10Å either side of the mid point. Hence, with the three fluorophore pairs it is possible to explore inter-cysteine distances approximately between 29Å to 75Å. This will allow to put to the text the maximum inter-cysteine distance identified by the HCG and MD methods.

In summary, he HCG method provides a quick and versatile approach to explore protein conformational changes that would be very computationally costly for atomistic methods. Besides its advanced performance, it also provides the user with data that was not revealed by MD simulations and that could be of biological importance. Further, the HCG method advice on the type of experiments and experimental probes that could be most useful to corroborate the simulations.

Chapter 9

Outlook and further research

Despite the success in exploring the conformational space using normal modes while ensuring stereochemical constraints, there are some limitations of FRODA to consider [38, 82]. Firstly, a fairly common occurrence during FRODA simulations is a sudden abort of the simulation after the fitting procedure repeatedly failed to satisfy constraints. The enforcement of constraints procedure does not guaranty that the number of constraint violations are reduced at each step. For example, atoms in a crowded environment could face multiple overlaps at the same time or a group of overlapping atoms could simultaneously provoke alternate corrective distances that provoke recursively new violation of constraints. This could lead to the atoms being bounced back and forth from overlap to overlap, which could explain the limitations of the software in terms of how the jamming effects occur. Another issue relates to the rigid cluster templates that it incorporates from FIRST. The hydrogen bonds and hydrophobic contacts that are considered as rigid within are maintained rigid as the protein moves. Therefore, keeping the bonds distance and orientation fix (unable to rotate) so that the residues within the rigid clusters are prevented from readjusting limits the motion of the protein artificially. Hence, a large scale motion may be inhibited or even blocked out if the geometries of the hydrogen and hydrophobic bonds are not allowed to re-arrange. This hypothesis is supported by a very recent simulation for yeast PDI structure (2B5E) I carried out using FRODAN at the "www.pathways.asu.edu". The results shown in Figure 9.1 indicate that the inter-cysteine distance ranges from the initial structure $d_{min}^{(FrodaN)} \simeq 26.51$Å to a $d_{max}^{(FrodaN)} \simeq 61.16$Å. This result is closer to the maximum inter-cystein values obtained from the MD simulations than the ones obtained with FRODA. However, the most striking result is that the E_{cut} was chosen to be $E_{\text{cut}} = -0.1$ kcal/mol, an energy cutoff that is much higher than the ones FRODA showed the largest mobil-

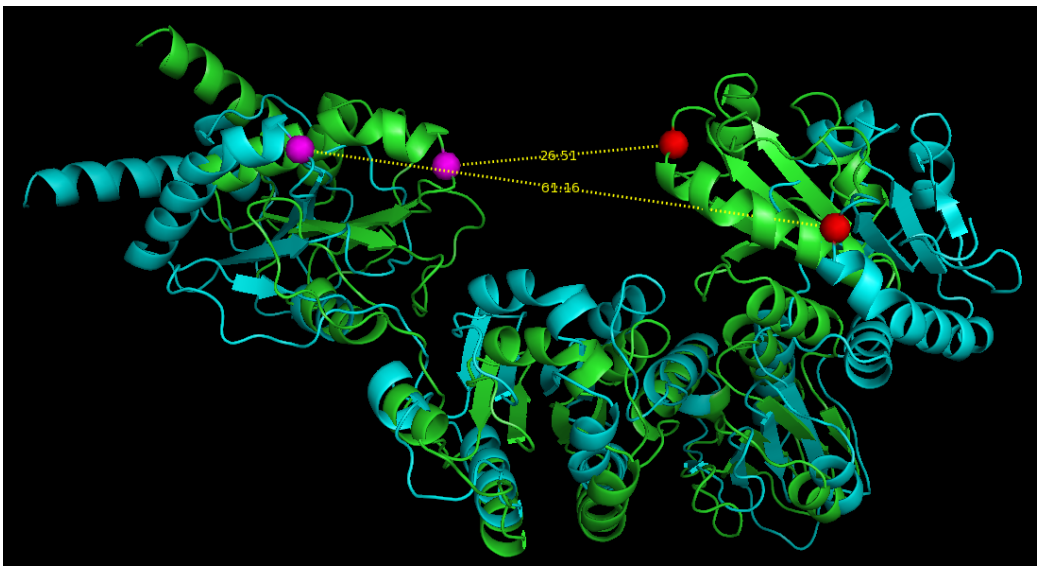


Figure 9.1: Superimposed yeast PDI structures during FRODAN simulations. The structure with the crystal coordinates is shown in green colour and the most open structure is shown in blue. The cysteine residues of interest are shown in: red for the cysteine in domain ***a***, i.e. α-carbon 599 within residue 61, and pink for the cysteine in domain ***a'***, i.e. α-carbon 5979 within residue 406. The inter-cysteine distance for the two conformers is shown in yellow and it ranges from $d_{min}^{(FrodaN)} \simeq 26.51$Å to $d_{max}^{(FrodaN)} \simeq 61.16$Å. The E_{cut} was purposely chosen as $E_{\text{cut}} = 0.1$ kcal/mol.

ity, i.e. $E_{\text{cut}} \leq 0.522$ kcal/mol. This is surprising since the inter-cystein distances for similar E_{cut} values using FRODA are $d_{E_{cut}=(-0.133)}^{(Froda)} \simeq 43$Å. This means that for at best considering the same number of hydrogen bond constrains or E_{cut}, the use of FRODA appears to restrict protein mobility with respect both MD and FRODAN simulations. Hence, these results supports [38, 82] at questioning the robustness and limitations of the geometric simulation method FRODA. Nevertheless, the method does bring useful insights when protein motion is biased using normal modes and it is advisable to use, especially to complement other techniques. In order to identify the limitations and robustness of the method new simulations and experiments will be invaluable to bring a better understanding. The investigation carried out during this thesis has brought a wealth of possible new avenues to continue with the research from this thesis. There are, of course, many systems that would be interesting to investigate using the HCG method but also new conceptual developments that could be applied to new coarse graining models. Our understanding about proteins, their structure and dynamics has evolved towards proteins as "dynamic networks" challenging old concepts for a better understanding to emerge.

Chapter 10

Appendix

10.1 Appendix

Here I compile several graphs for which there was not much space in the main text. These graphs are showing the scalar product between initial modes and the ones obtained for each conformer as discussed in chapter 4.

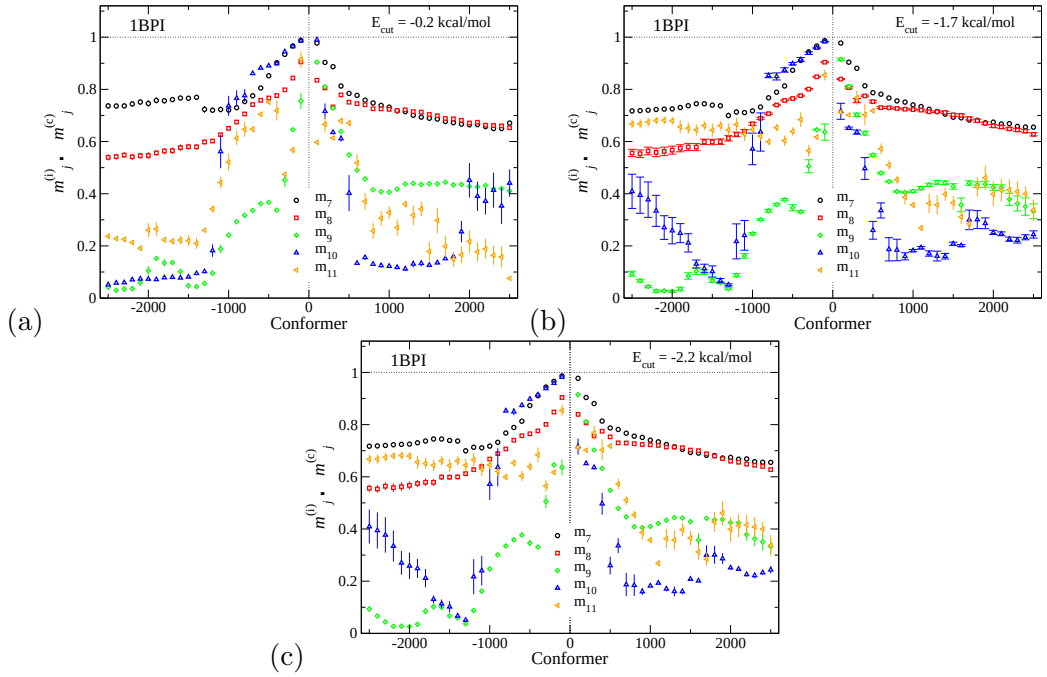


Figure 10.1: Dot product graph for BPTI (1BPI). The dot product $m_j^{(i)} \cdot m_j^{(c)}$ between an initial starting mode $m_j^{(i)}$ and its current mode $m_j^{(c)}$, $j = 7, \ldots, 11$ as the initial structure is projected along the initial mode. The current modes, $m_j^{(c)}$, are obtained from performing normal mode analysis on the current conformations as the initial structure is projected along an initial mode $m_j^{(i)}$. For clarity, dot products for only 25 conformations of each direction of motion are shown. The evolution of the dot product along the conformations is reported for different cutoff energies, which for BPTI are (a) $E_{\text{cut}} = -0.2$ kcal/mol, (b) $E_{\text{cut}} = -1.7$ kcal/mol and (c) $E_{\text{cut}} = -2.2$ kcal/mol. The horizontal and vertical dotted lines denote the largest possible value of $m_j^{(i)} \cdot m_j^{(c)}$ and the zero on the conformer axis, respectively.

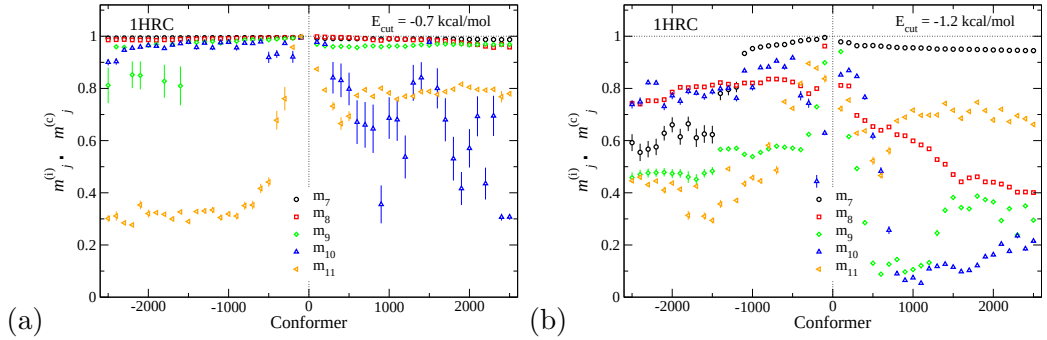


Figure 10.2: Dot product graph for cytochrome-c (1HRC) as described in Figure 10.1 but with E_{cut} values of (a) -0.7 kcal/mol and (b) -1.2 kcal/mol.

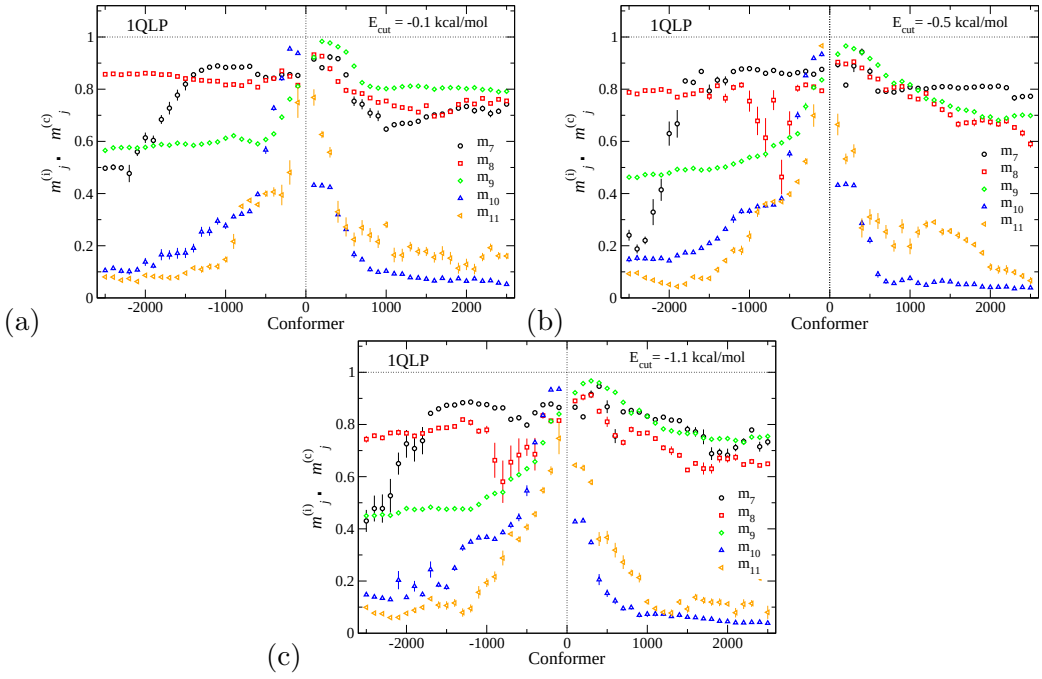


Figure 10.3: Dot product graph for α1-antitrypsin (1QLP) as described in Figure 10.1 but with E_{cut} values of (a) $E_{\text{cut}} = -0.1$ kcal/mol, (b) $E_{\text{cut}} = -0.5$ kcal/mol and (c) $E_{\text{cut}} = -1.1$ kcal/mol.

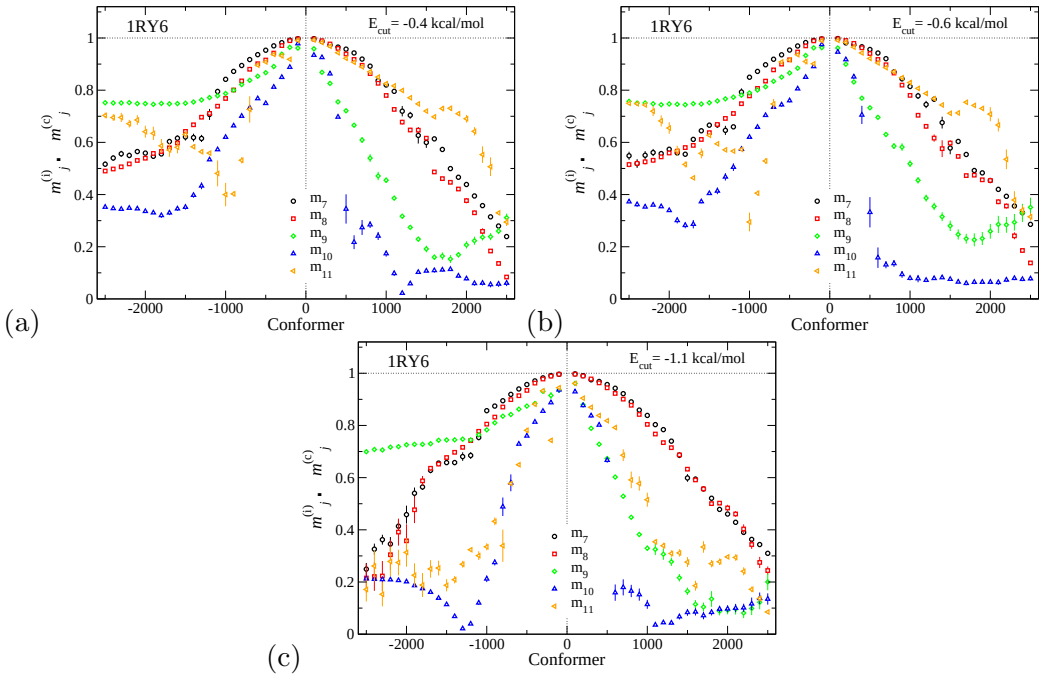


Figure 10.4: Dot product graph for internal kinesin motor domain (1RY6) as described in Figure 10.1 but with E_{cut} values of (a) -0.4 kcal/mol (b) -0.6 kcal/mol and (c) -1.1 kcal/mol.

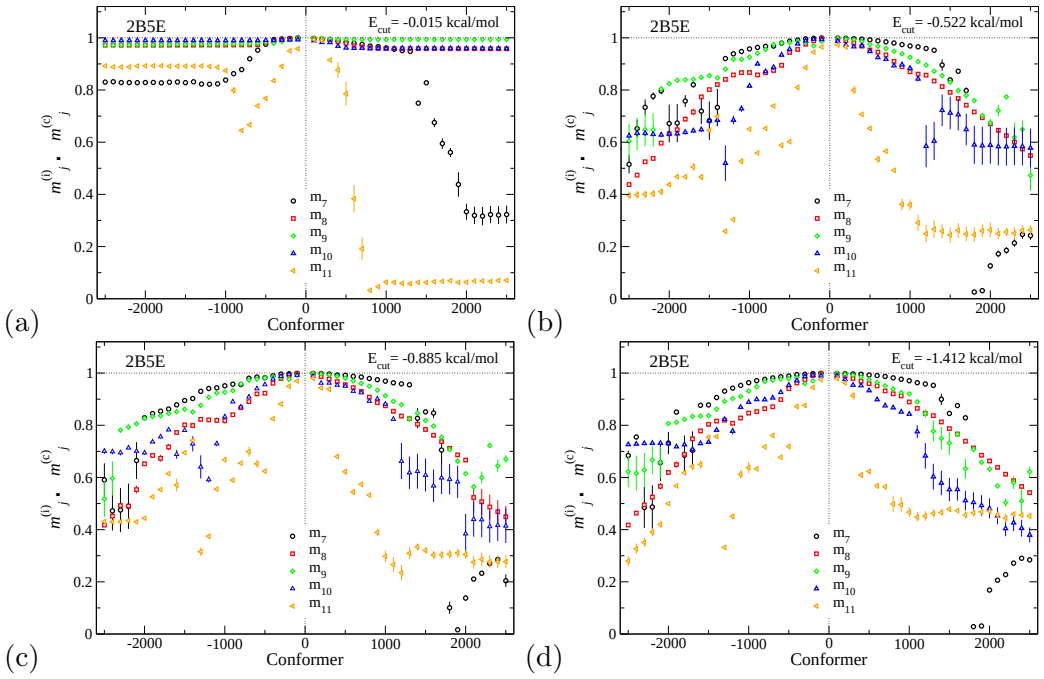


Figure 10.5: Dot product graph for yeast PDI (2B5E) as described in Figure 10.1 but with E_{cut} values of (a) $E_{\text{cut}} = -0.015$ kcal/mol, (b) $E_{\text{cut}} = -0.522$ kcal/mol, (c) $E_{\text{cut}} = -0.885$ kcal/mol and (d) $E_{\text{cut}} = -1.412$ kcal/mol.

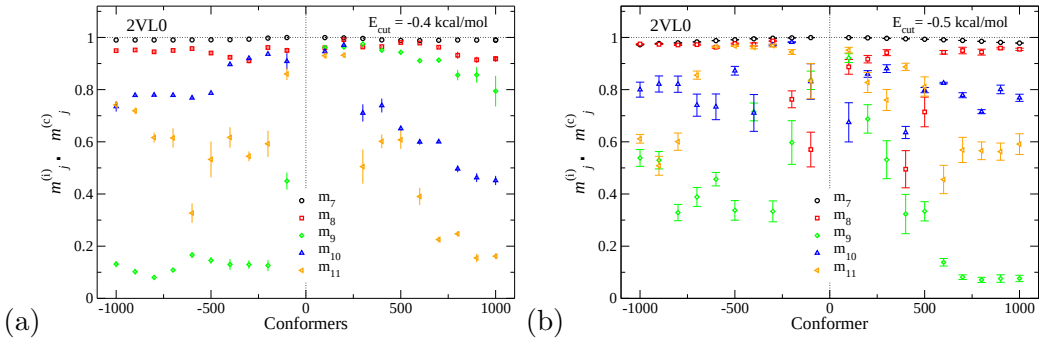


Figure 10.6: Dot product graph for a ligand gated ion channel protein (2VL0) as described in figure 10.1but with E_{cut} values of (a) $E_{\text{cut}} = -0.4$ kcal/mol and (b) at $E_{\text{cut}} = -0.5$ kcal/mol.

Bibliography

[1] Rabi II, Millman S, Kusch P, and Zacharias JR. The molecular beam resonance method for measuring nuclear magnetic moments. Phys. Rev, 53(495):318, 1938.

[2] Förster T. Zwischenmolekulare energiewanderung und fluoreszenz. Annalen der Physik, 437(1-2):55–75, 1948.

[3] Hawkins HC, de Nardi M, and Freedman RB. Redox properties and cross-linking of the dithiol/disulphide active sites of mammalian protein disulphide-isomerase. Biochemical journal, 275(Pt 2):341, 1991.

[4] Alder BJ and Wainwright TE. Studies in molecular dynamics. i. general method. The Journal of Chemical Physics, 31:459, 1959.

[5] Go N and Abe H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. biop, 20:991–1011, 1981.

[6] Abe H and Go N. Noninteracting local-structure model of folding and unfolding transition in globular proteins ii, application to two-dimensional lattice proteins. biop, 20:1013–1031, 1981.

[7] Rohl CA, Strauss CE, Misura KM, and Baker D. Protein structure prediction using rossetta. Methods Enzymol, 383:66–93, 2004.

[8] Bahar I, Atilgan AR, and Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold. Des, 2:173–181, 1997.

[9] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, and Bahar I. Anisotropy of fluctuation dynamics of proteins with an elaxtic network model. Biophys. J., 80(505-515), 2001.

[10] Tama F, Gadea FX, Marques O, and Sanejouand YH. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins, 41:1–7, 2000.

[11] Tama F and Brooks CL 3rd. The mechanism and pathways of ph induced swelling in cowpea cholorotic mottle virus. J. Mol. Biol., 318:733–747, 2002.

[12] Warshel A and Levitt M. Folding and stability of helical proteins: carp myogen. J. Mol. Biol., 106:421–437, 1976.

[13] Baker D. A surprising simplicity to protein folding. Nat, 405:39–42, 2000.

[14] Tama F and Sanejouand Y H. Conformational change of proteins arising from normal mode calculations. pe, 1:1–6, 2001.

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

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

[17] Jacobs DJ, Rader AJ, Kuhn LA, and Thorpe MF. Protein flexibility predictions using graph theory. Prot: Struct. Func. Gen., 44:150–165, 2001.

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

[19] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, and Bahar I. Anisotropy of fluctuation dynamics of proteins with an elaxtic network model. Biophys. J., 80(505-515), 2001.

[20] Krebs WG, Alexandrov V, Wilson CA, Echols E, Yu H, and Gerstein M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins, 48:682–695, 2002.

[21] Ahmed A and Gohlke H. Multi-scale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins, 63:1038–1051, 2006.

[22] Tama F, Wriggers W, and Brooks CL 3rd. Exploring global distortions of biological macromolecules and assemblies from low- resolution structural information and elastic network theory. J. Mol. Bio, 321:297–305, 2002.

[23] Delarue M and Sanejouand YH. Simplified normal mode analysis of conformational transitions in DNA-dependant polymerases:the elastic network model. J. Mol. Biol., 320:1011–1024, 2002.

[24] Petrone P and Pande VS. Can conformational change be described by only a few normal modes? Biophys J., 90:p1583–1593, 2006.

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

[26] Bahar I, Lezon TR, Bakan A, and Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev, 110:1463–1497, 2010.

[27] Ma J. Usefulness and limitations of normal mode analysis in modelling dynamics of biomolecular complexes. Structure, 13:373–380, 2005.

[28] Rueda M, Chacón P, and Orozco M. Through validation of protein normal mode analysis: A comparative study with essential dynamics. Structure, 15:565–575, 2007.

[29] Nakasako M, Maeno A, Kurimoto E, Harada T, Yamaguchi Y, Oka T, Takayama Y, Iwata, and Kato K. Redox-dependent domain rearrangement of protein disulfide isomerase from a thermophilic fungus. Biochemistry, 49:6953–6962, 2010.

[30] Wells SA, Menor S, Hespenheide BM, and Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. Phys. Biol., 2:S127–S136, 2005.

[31] Dove MT, Harris MJ, Hannon AC, Parker JM, Swainson IP, and Gambhir M. Floppy modes in crystalline and amorphous silicates. Physical review letters, 78(6):1070–1073, 1997.

[32] Hammonds KD, Heine V, and Dove MT. Insights into zeolite behaviour from the rigid unit mode model. Phase Transitions: A Multinational Journal, 61(1-4):155–172, 1997.

[33] Giddy AP, Dove MT, Pawley GS, and Heine V. The determination of rigid-unit modes as potential soft modes for displacive phase transitions in framework crystal structures. Acta Crystallographica Section A: Foundations of Crystallography, 49(5):697–703, 1993.

[34] Dove MT, Trachenko KO, Tucker MG, and Keen DA. Rigid unit modes in framework structures: theory, experiment and applications. Reviews in Mineralogy and Geochemistry, 39(1):1, 2000.

[35] Wells SA. Geometric analysis of structural polyhedra. 2003.

[36] Suhre K and Sanejouand Y-H. On the potential of normal mode analysis for solving difficult molecular replacement problems. Acta Cryst D, 60:796–799, 2004.

[37] Lei M, Zavodszky MI, Kuhn LA, and Thorpe MF. Sampling protein conformations and pathways. Journal of computational chemistry, 25(9):1133–1148, 2004.

[38] Farrell DW, Speranskiy K, and Thorpe MF. Generating stereochemically-acceptable protein pathways. Proteins, 78:2908–2921, 2010.

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

[40] Word JM, Lovell SC, Richardson JS, and Richardsonzhed DC. Asparagine and glutamine: Using hydrogen atoms contacts in the choice of side-chain amide orientation. J. Mol. Biol., 285:1735–1747, 1999.

[41] DeLano WL. The PyMOL molecular graphics system. http://www.pymol.org, 2002.

[42] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE. The protein data bank. Nucl. Acids Res., 28:235–242, 2000. http://www.rcsb.org.

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

[44] Hespenheide BM, Rader AJ, Thorpe MF, and Kuhn LA. Identifying protein folding cores: Observing the evolution of rigid and flexible regions during unfolding. J. Mol. Graph. & Model., 21:195–207, 2002.

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

[46] Rader AJ, Hespenheide BM, Kuhn LA, and Thorpe MF. Protein unfolding: rigidity lost. Proc, Natl. Acad. Sci. USA, 99:3540–3545, 1999.

[47] Sartbaeva A, Wells SA, Huerta A, and Thorpe MF. Local structural variability and the intermediate phase window in network glasses. Phys. Rev. B, 75:224204, 2007.

[48] Wells SA, Jimenez-Roldan JE, and Römer RA. Comparative analysis of rigidity across protein families. Phys. Biol., 6(4):046005–11, 2009.

[49] Hemberg M, Yaliraki SN, and Barahona M. Stocastik kinetics of viral capsid assembly based on detailed protein structures. Biophysical journal, 90:3029–3042, 2006.

[50] Jolley CC, Wells SA, Hespenheide BM, Thorpe MF, and Fromme P. Docking of photosystem I subunit c using a constrained geometric simulation. J. Am. Chem. Soc., 128(27):8803–8812, 2006.

[51] Jolley CC, Wells SA, Fromme P, and Thorpe MF. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. Biophys. J., 94:1613–1621, 2008.

[52] Macchiarulo A, Nuti R, Bellochi D, Camaioni E, and Pellicciari R. Molecular docking and spatial coarse graining simulations as tools to investigate substrate recognition, enhancer binding and conformational transitions in indoleamine-2,3-dioxygenase (ido). Biochim. et Biophys. Acta- Proteins and Proteomics, 1774:1058–1068, 2007.

[53] M. Sun, M. B. Rose, S. K. Ananthanarayanan, D. J. Jacobs, and C. M. Yengo. Characterisation of the pre-force-generation state in the actomyosin cross-bridge cycle. Proc. Nat. Acad. Sci., 105:8631–8636, 2008.

[54] Yaliraki SN and Barahona M. Chemistry across scales: from molecules to cells. Phil Trans R Soc A Math Phys Eng Sci, 365:2921–2934, 2007.

[55] Sherwood P, Brooks BR, and Sansom MSP. Multiscale methods for macromolecular simulations. Curr Opin Struct Biol, 18:630–640, 2008.

[56] Dykeman EC and Sankey OF. Normal mode analysis and applications in biological physics. J. Phys. Condens. Matter, 22:423202, 2010.

[57] Wlodawer A, Walter J, Huber R, and Sjolin L. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. J. Mol. Biol., 180(2):301–329, 1984.

[58] Amir D and Haas E. Reduced bovine pancreatic trypsin inhibitor has a compact structure. Biochemistry, 27(25):8889–8893, 1988.

[59] K Shipley, M Hekmat-Nejad, J Turner, C Moores, R Anderson, R Milligan, R Sakowicz, and R Fletterick. Structure of a kinesin microtubule depolymerization machine. EMBO Journal, 23(7):1422–1432, 2004.

[60] Elliott PR, Pei XY, Dafforn TR, and Lomas DA. Topography of a 2.0 $\AA$ structure of α-1-antitrypsin reveals targets for rational drug design to prevent conformational disease. Prot. Sci., 9(7):1274–1281, 2000.

[61] Freedman RB, Klappa P, and Ruddock LW. Protein disulfide isomerases exploit synergy between catalytic and specific binding domains. EMBO Journal, 15:136–140, 2002.

[62] Tian G, Kober F, Lewandrowski U, Sickmann A, Lennarz WJ, and Schindelin H. The catalytic activity of protein-disulfide isomerase requires a conformationally flexible molecule. J. Biol. Phys., 283:33630–33640, 2008.

[63] Tian G, Xiang S, Noiva R, Lennarz WJ, and Schindelin H. The crystal structure of yeast protein disulfide isomerase suggests cooperativity between its active sites. Cell, 124:61–73, 2006.

[64] Hilf RJC and Dutzler R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. Nature, 452(7185):375–379, 2008.

[65] Jimenez-Roldan JE, Wells SA, Freedman RB, and RoemerRA. Integration of FIRST, FRODA and NMM in a coarse grained method to study protein disulphide isomerase conformational change. In Journal of Physics: Conference Series, volume 286, page 012002. IOP Publishing, 2011.

[66] Jimenez-Roldan JE, Bhattacharya M, Vishveshwara S, Freedman RB, and Roemer RA. in preperation.

[67] Belfield W, Cole D, and Payne M. in preparation.

[68] Cai H, Wang CC, and Tsou CL. Chaperone-like activity of protein disulfide isomerase in the refolding of a protein with no disulfide bonds. Journal of Biological Chemistry, 269(40):24550–24552, 1994.

[69] Wetterau JR, Combs KA, McLean LR, Spinner SN, and Aggerbeck LP. Protein disulfide isomerase appears necessary to maintain the catalytically active structure of the microsomal triglyceride transfer protein. Biochemistry, 30(40):9728–9735, 1991.

[70] Edman JC, Ellis L, Blacher RW, Roth RA, and Rutter WJ. Sequence of protein disulphide isomerase and implications of its relationship to thioredoxin. 1985.

[71] Farquhar R, Honey N, Murant SJ, Bossier P, Schultz L, Montgomery D, Ellis RW, Freedman RB, and Tuite MF. Protein disulfide isomerase is essential for viability in saccharomyces cerevisiae. Gene, 108(1):81–89, 1991.

[72] Freedman RB, Hirst TR, and Tuite MF. Protein disulphide isomerase: building bridges in protein folding. Trends in biochemical sciences, 19(8):331–336, 1994.

[73] Darby NJ, Kemmink J, and Creighton TE. Identifying and characterizing a structural domain of protein disulfide isomerase. Biochemistry, 35(32):10517–10528, 1996.

[74] Kemmink J, Darby NJ, Dijkstra K, Nilges M, and Creighton TE. Structure determination of the n-terminal thioredoxin-like domain of protein disulfide isomerase using multidimensional heteronuclear 13c/15n nmr spectroscopy. Biochemistry, 35(24):7684–7691, 1996.

[75] Kemmink J, Darby NJ, Dijkstra K, Nilges M, and Creighton TE. The folding catalyst protein disulfide isomerase is constructed of active and inactive thioredoxin modules. Current Biology, 7(4):239–245, 1997.

[76] Vuori K, Myllylä R, Pihlajaniemi T, and Kivirikko KI. Expression and site-directed mutagenesis of human protein disulfide isomerase in escherichia coli. this multifunctional polypeptide has two independently acting catalytic sites for the isomerase activity. Journal of Biological Chemistry, 267(11):7211, 1992.

[77] Darby NJ and Creighton TE. Functional properties of the individual thioredoxin-like domains of protein disulfide isomerase. Biochemistry, 34(37):11725–11735, 1995.

[78] I Bahar and A J Rader. Coarse-grained normal mode analysis in structural biology. Current opinion in structural biology, 15(5):586–592, 2005.

[79] Heal JW, Wells SA, Jimenez-Roldan JE, Freedman RF, and Röemer RA. Rigidity analysis of hiv-1 protease. In Journal of Physics: Conference Series, volume 286, page 012006. IOP Publishing, 2011.

[80] Heal JW, Jimenez-Roldan JE, Wells SA, Freedman RB, and Römer RA. Inhibition of hiv-1 protease: the rigidity perspective. Bioinformatics, 28(3):350–357, 2012.

[81] Farrell DW and Thorpe M. http://flexweb.asu.edu. 2009.

[82] Farrell DW. Generating stereochemical acceptable pathways. PhD thesis, page 151, 2010.