2011

A computational study of protein dynamics, structure ensembles, and functional mechanisms

Tu-liang Lin
Iowa State University

A computational study of protein dynamics, structure ensembles, and functional mechanisms

by

Tu-Liang Lin

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Guang Song, Major Professor
David Fernandez-Baca
Mark S. Hargrove
Vasant Honavar
Robert L. Jernigan

Iowa State University

Ames, Iowa

2011

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I am very grateful to Dr. Guang Song for his valuable guidance throughout this research and the writing of this thesis. This work would not have been finished without his constant support and kind encouragement.

I would also like to thank the other members of my committee, Dr. David Fernandez-Baca, Dr. Mark S. Hargrove, Dr. Vasant Honavar and Dr. Robert L. Jernigan, for giving their valuable time and inputs.

I am thankful to another member in our lab, Santhosh, for his kind and friendly association. I am also thankful to my friend, Dr. Ganesh Ram Santhanam, who gives useful suggestions to the thesis formation.

I am extremely grateful to my parents and family members for supporting me to pursue this program. Especially, my special thanks goes to my wife, Huey-may Wu, who always encourages and supports me during the difficult times in my life.

ABSTRACT

Proteins are vital parts of living organisms and involved in almost every single biological process. When participating in in-vivo reactions, proteins are constantly in motion and their dynamics is critical to the realization of their functions. Although the advancement of structure determination methods and computational approaches has opened up great opportunities for studying protein dynamics and functional mechanisms, much remains to be understood. In this thesis, I aim to establish some new computational methods for studying protein dynamics and functional mechanisms.

In the first half of this thesis, I will describe the new computational methods for protein dynamics that I have developed. One of the most common methods for obtaining the protein dynamics computationally is molecular dynamics (MD) simulations. Although MD simulations can provide atomic details of the protein dynamics, it is computationally expensive and is thus limited to short time scales, especially for large systems. In this thesis I focus on methods for studying protein dynamics that can circumvent such limitations. Two strategies are employed: (1) represent protein dynamics using weighted structure ensembles; (2) improve existing coarse-grained models with multi-body potentials using generalized spring tensors.

In the second half of the thesis, I investigate the functional mechanisms of ligand migration and allosteric communication using novel, dynamics-based methods. Specifically, two subgoals are defined and accomplished: (1)chart the ligand migration channels in heme proteins using different structure ensembles; (2) determine the allosteric communication pathways using dynamic motion correlations.

CHAPTER 1. OVERVIEW AND OBJECTIVES

The study of protein dynamics and functional mechanisms at the molecular level has been greatly advanced in recent years due to the development of various structure determination methods and computational approaches, but how to most effectively model protein dynamics and use it to understand functional mechanisms is still an important open question. Currently, it is well accepted that the functions of a protein are closely related to not only its structure but also its dynamics and there is strong evidence [Burra et al. (2009); Eisenmesser et al. (2005)] that a single structure is not sufficient for fully understanding the protein functions and some degree of mobility is necessary. In this thesis, I plan to address some of the difficulties in modeling the protein dynamics and in doing so gain a better understanding of the functional mechanisms. Two specific aims are set up to achieve this objective.

Aim # 1: Establish new computational methods for protein dynamics

One of the most common methods for studying the protein dynamics computationally is molecular dynamics (MD) simulation. Although MD simulation can provide atomic details of the protein dynamics, it is computationally expensive and is limited to short time scales, especially for large systems. In this thesis I focus on methods for studying protein dynamics that can circumvent such limitations.

Subgoal # 1.1: Represent protein dynamics using weighted structure ensembles

A structure ensemble can be used to represent the effects of protein dynamics and capture protein structural flexibility around the native states. Recently, some researchers point out that different structures of the same protein under different experimental conditions or of

proteins with high sequence similarity can form an ensemble that resembles the heterogeneity of the native state or a wide global landscape of protein dynamics [Burra et al. (2009); Best et al. (2006)]. An ensemble of conformations sampled by MD simulation, on the other hand, is usually limited to a local region of the conformation space as transitions among different conformation states can take a longer time than MD simulations can reach. Therefore, in this subgoal the abundant structures of the same protein in the Protein Data Bank [Berman et al. (2000)] are utilized and are combined with some conformational sampling approaches to obtain a more globally distributed ensemble.

An ensemble representation of the protein dynamics has the advantageous flexibility to incorporate the knowledge of protein dynamics from various sources, such as existing crystal structures, NMR ensembles, and conformations sampled by MD simulations. The method developed is able to obtain a more globally distributed dynamic ensemble than a traditional MD simulation under the same time constraint.

Subgoal # 1.2: Improve existing coarse-grained models with multi-body potentials using generalized spring tensors

In the last decade, various coarse-grained elastic network models (ENMs) have been developed to study the large scale motions of proteins and protein complexes where computer simulations using detailed all-atom models are not feasible. In order to achieve simplicity, these coarse-grained ENMs usually adopt only two-body Hookean-like potentials, such as in Gaussian Network Model (GNM) [Bahar et al. (1997)] and Anisotropic Network Model (ANM) [Atilgan et al. (2001)]. However, these two-body interactions are limited in fully representing the interactions between a pair of residues. Specifically, under the two-body potentials of ANM, the fluctuation of one residue relative to its interacting partner is only constrained longitudinally along the axis connecting them. Therefore, it is necessary to include multi-body interactions in the potential in order to have more realistic constraints. This goal is achieved by deriving the model from a physically more realistic multi-body G$\bar{o}$-like potentials [Clementi et al. (2000)].

Aim # 2: Determine the functional mechanisms of ligand migration and allostery communication

It has been pointed out that protein dynamics plays an essential role in understanding some functional mechanisms, such as allosteric communication [Tobi and Bahar (2005); Bahar et al. (2007); Bahar and Rader (2005); Yang et al. (2009)] and ligand migration [Ruscio et al. (2008); Bossa et al. (2004); Bourgeois et al. (2006); Ostermann et al. (2000)]. In this aim, the functional mechanisms of allosteric communication and ligand migration are investigated using protein dynamics models.

Subgoal # 2.1: Chart the ligand migration channels in heme proteins using different structure ensembles

Many biological reactions happen at a catalytic site that is hidden beneath the protein surface and can only be reached through some transient channels. For small ligands such as O_2, the opening and closing of these channels are mainly controlled by the dynamic fluctuations of the host protein [Cohen et al. (2006)]. For this reason, a complete map showing how a ligand migrates in the host protein cannot be obtained from a single static structure. MD simulations provide a way to identify these channels but are limited by the time scale that can be reached. In this subgoal, an approach to map the ligand migration channels in dynamic structure ensembles is presented. The ligand migration maps from different ensembles are further compared and the control mechanisms of the identified ligand migration channels are investigated.

Subgoal # 2.2: Determine the allosteric communication pathways using dynamic motion correlation

Allosteric regulation can be described as the binding of an effector at one site that switches the functionality of another site, often at distance. Although a wide variety of models have been proposed [Tang et al. (2007); Gandhi et al. (2008); Chennubhotla and Bahar (2006); Zheng et al. (2006); Zheng and Brooks (2005)], the underlying mechanisms of the allosteric

communication remain unclear. In this subgoal, I stress on the important role that protein dynamics plays in the allosteric communication and hypothesize that the allosteric communication between the allosteric site and catalytic site should be carried out along pathways of residues that have strongly correlated motions, so that information such as conformation change can be quickly transduced from one site to another. A simple and computationally inexpensive approach to identify the putative allosteric communication pathways using coarse-grained ENMs is provided. The results are validated by examining whether the residues along the predicted pathways are evolutionarily conserved.

Overall, I aim to provide novel and systematic computational approaches to model protein dynamics and to elucidate the functional mechanisms of allosteric communication and ligand migration. The concrete output of this thesis includes (1) a novel method for determining the relative populations of the conformation states within an ensemble; (2) a generalized spring tensor model; (3) an efficient algorithm for charting the ligand migration maps, and (4) an efficient algorithm for predicting the allosteric communication pathways.

Thesis Organization

The thesis is organized as follows:

Chapter 1: Overview and Objectives
This chapter gives a general introduction to the thesis, presenting the overall structure and the aims of this thesis.

Chapter 2: Determine the Populations of Protein Conformation States Using Experimental Residual Dipolar Coupling Data
In this chapter, the subgoal # 1.1 is fulfilled. A new dimension, the relative population at each structure, is added to the ensemble, and it greatly enhances the ensemble's ability to describe the conformation space. A novel computational method that determines the relative populations is developed by employing iterative least squares fitting methods to the experimental

RDC data. As a result, we are able to use this method to determine several ubiquitin ensembles with low Q-factors, which would not have been possible without the use of relative populations.

Chapter 3: Evaluate the Quality of Conformation Sampling Methods using Experimental Residual Dipolar Coupling Data

In this chapter, we further demonstrate the wide application of the iterative least squares fitting methods, developed in Chapter 2. Many computational approaches have been developed and used for sampling protein conformations near the native state. However, it has been difficult to evaluate the qualities of the conformations sampled or to compare them among the various sampling schemes. The developed approaches are applied to evaluate ubiquitin conformations generated from four widely-used conformational sampling approaches, namely, MD simulation, Elastic Network Model (ENM), CONCOORD, and tCONCOORD.

Chapter 4: Generalized Spring Tensor Models for Protein Fluctuation Dynamics and Conformation Changes

This chapter fulfills the subgoal # 1.2 and chapter 2-4 combined complete the fulfilment of Aim # 1. In order to achieve simplicity, the multi-body potential, which is known to be important in protein structure prediction and protein design etc., is neglected in most elastic network models. In this chapter, we address this insufficiency and introduce three-body and four-body potentials through bond bending and torsional interactions in the G$\bar{o}$-like potential.

Chapter 5: Efficient Mapping of Ligand Migration Channel Networks in Dynamic Proteins

In this chapter, subgoal # 2.1 is fulfilled. Ligand migration in a dynamic protein resembles closely a well-studied problem in robotics, namely, the navigation of a mobile robot in a dynamic environment. In this chapter, we present a novel robotic motion planning inspired approach that can map the ligand migration channel network in a dynamic protein. The dynamic behaviors are represented as structure ensembles. The dynamic ligand migration maps are charted using different structure ensembles.

Chapter 6: Predicting Allosteric Communication Pathways Using Motion Correlation Network

Subgoal # 2.2 is fulfilled in this chapter and Chapter 5 and 6 combined complete Aim # 2. In this chapter, we hypothesize that the allosteric communication between an allosteric site and its catalytic site is through pathways of residues that have strongly correlated motions. A weighted network from the coarse-grained elastic network model is formulated and graph search algorithms are used to identify allosteric communication pathways.

Chapter 7: Conclusion and Future Research

This final chapter presents conclusions and some future research directions.

CHAPTER 2. DETERMINE THE POPULATIONS OF PROTEIN CONFORMATION STATES USING EXPERIMENTAL RESIDUAL DIPOLAR COUPLING DATA

A paper to be submitted

Tu-Liang Lin, Santhosh Kumar Vammi and Guang Song

Abstract

Ensembles have been increasingly used to represent protein native states and structural heterogeneity. In this work, we add a new dimension, the relative population of each structure, to the ensemble, which greatly enhances the ensemble's ability to describe the conformation space. We develop a novel computational method that determines the relative populations by employing iterative least squares fitting methods to the experimental RDC data. We compare Q-factors among several ubiquitin ensembles. Our results show that ensembles with RDC derived populations significantly improve the agreement between the calculated and experimental RDCs. As a result, we are able to use this method to determine several ubiquitin ensembles with low Q-factors, which would not have been possible without the use of relative populations. These ensembles represent different solutions to the experimental constraints from the well-known EROS ensemble. As a result, different conclusions may be drawn regarding whether a structure is reached by conformation selection or induced-fit.

Introduction

Protein structure ensembles

The functions of a protein are closely related to not only its structure, but also its dynamics. For more and more proteins, it is becoming increasingly evident that the functional behavior of a protein is best represented not by one single static structure, but by the distribution and dynamic transition among a number of conformation states that form the native-state ensemble [Karplus and McCammon (2002)]. With the advancement of structure determination methods, protein structures are becoming increasingly available and for some well-studied proteins, tens and even hundreds of structures (of the same protein) have been determined. These structures have been shown to capture a representative subset of the native-state ensemble [Best et al. (2006)]. However, the relative populations of these conformation states in the conformation space are not known. Currently, several approaches for obtaining relative population in disordered protein were proposed [Fisher et al. (2010); Choy and Forman-Kay (2001)], but to the best of our knowledge no attempts at obtaining the relative populations have been made for folded protein ensembles even though such information is essential for describing the conformation space (other key information is the transition rates among these states). The aim of this work is to employ a novel iterative least-square fitting approach to determine the relative populations of these structures using experimental RDC data.

Significance of the work

There are many practical applications for conformation ensembles. Park et al. showed that structure ensembles can help improve docking, screening, and selectivity prediction for small nuclear receptors [Park et al. (2010)]. Friedland and Kortemme demonstrated the usefulness of conformation ensembles in computational protein design [Friedland and Kortemme (2010)]. Several research groups used conformation ensembles to explain the recognition process in biomolecular bindings [Lange et al. (2008); Boehr et al. (2009); Wlodarski and Zagrovic (2009)]. A conformation ensemble along with information regarding its relative populations can be used to enhance the aforementioned applications.

Methods for determining structure ensembles

A number of experimental [Scheek et al. (1991)], computational [Karplus and McCammon (2002); Shehu et al. (2006); de Groot et al. (1997); Seeliger and De Groot (2009)], or hybrid [Lange et al. (2008); Lindorff-Larsen et al. (2005); Kuriyan et al. (1991); Richter et al. (2007)] methods have been developed to capture protein structure heterogeneity and then represent it using ensembles. NMR spectroscopy is the most commonly used experimental technique for studying protein dynamics in solution. The technique is powerful even though it has a limitation on the size of the protein. Relaxation of nuclear magnetization can quantitatively probe fast protein dynamics (picoseconds to nanoseconds) or the dynamics in a much slower domain (microseconds to milliseconds) [Lange et al. (2008)]. Thus there exists a blind window ranging from nanoseconds to microseconds that is beyond the capacity of nuclear magnetization relaxation. MD simulation is the most commonly used computational approach for obtaining a structure ensemble. The main limitation of MD simulation is the limited timescale that can be reached, especially for large systems. It also has been shown recently that MD simulations beyond hundreds of nanoseconds might have the potential risk of running into high free energy states and staying there for a long time, thus incurring skewed populations [Lange et al. (2010)]. Therefore, obtaining the dynamics over a broader time-scale is a challenge. To overcome some of these difficulties, residual dipolar couplings (RDCs), which provide complementary dynamics information that is inaccessible to NMR relaxation methods, have been used as ensemble constraints for ensemble determination [Lange et al. (2008); Mittermaier and Kay (2006)].

One possible way to circumvent the aforementioned limitations in ensemble determination is to construct ensembles using known experimental structures only (which has its own limitation as well). Recently, it was suggested that the available structures of the same protein determined under different experimental conditions or of proteins with high sequence similarity should be useful in representing the heterogeneity of the native states and in understanding the functions of the protein [Best et al. (2006)]. Zoete et al. (2002) calculated the relative backbone fluctuations among an ensemble of HIV-1 protease structures and found it comparable with experimental B-factors. Zhang et al. (1995) investigated the agreement and discrepancy among

25 crystal structures of T4 lysozyme. Best et al. (2006) compared the dynamic properties of ensembles formed by different X-ray structures of the same proteins with NMR experimental data and found that the order parameters, scalar couplings, and residual dipolar couplings were all well reproduced.

Several approaches incorporate the RDC, NOE data, or order parameters as ensemble constraints in MD simulations to obtain dynamic ensembles. Lindorff-Larsen et al. (2005) first developed a dynamic-ensemble refinement (DER) method that incorporates NOE (Nuclear Over-hauser enhancement) data and generalized order parameters as ensemble constraints into the energy function. RDC data were first used as constraints in an innovative study by Lange et al. (2008), who carried out ensemble refinements using MD simulations in explicit solvent under restraints from NOE and RDC data. (Their approach was named EROS, standing for ensemble refinement with orientational restraints). Richter et al. (2007) adopted replica-simulations and presented a MD simulation protocol for generating protein ensembles under the ensemble-averaged NMR restraints back calculated from the reference ensemble.

Besides the aforementioned methods, there exist other computational approaches for constructing an ensemble, such as the loop prediction program [Shehu et al. (2006)], the geometric restriction checking program [de Groot et al. (1997)], and the chemical shift prediction algorithm [Jensen et al. (2010)].

In all of these methods, it was assumed that all structures within an ensemble contribute equally. The idea of adding the dimension of relative populations to the description of an ensemble was not explored, even though it is a highly significant feature of any ensemble. This is especially true for ensembles formed by experimental structures, about which little is known experimentally regarding their relative populations. Precise knowledge of the relative populations for ensembles generated from MD simulations is also not always available . Recently it was shown that MD simulations longer than hundreds of nanoseconds possess potential risks of running into high free energy states and resulting in more skewed populations and worse correlations with experimental RDCs than shorter simulations [Lange et al. (2010)].

Residual dipolar coupling (RDC)

The dipolar coupling is the interaction that exists between two magnetic nuclei. Under isotropic solution conditions, dipolar coupling averages to zero as a result of the effects of Brownian motion. The use of an alignment medium that can create a weak force on the protein and eventually lead to an incomplete averaging of anisotropic magnetic interactions makes it possible to measure dipolar coupling. Therefore, the residual dipolar coupling (RDC) between two spins represents the incomplete averaging of spatially anisotropic dipolar couplings. One intriguing property of RDC is that it can probe the blind spot of the conventional NMR relaxation. Conventional NMR relaxation has been used to probe the dynamics faster than the rotational correlation time of a system which is in the range from picoseconds to nanoseconds or to identify the dynamics slower than microseconds. However, many important biological processes happen in the blind spot of conventional NMR, between nanoseconds and microseconds, where RDC plays an important role in the protein dynamics studies. In this work, we develop an algorithm that can exploit the RDC data and output the relative populations of the structures within an ensemble that is in best agreement with the RDC data.

Induced-fit or conformation selection?

We apply the proposed algorithm to study the recognition mechanism in protein-protein interactions. For some well studied proteins such as Ubiquitin, there are many experimental structures available in the Protein Data Bank [Berman et al. (2000)]. These structures provide valuable information about protein structural heterogeneity since each determined structure may represent a valid conformation state in the energy landscape. Some structures are unbound free structures and some are bound structures, but how these structures distribute in the energy landscape (i.e., their relative populations) is not fully known. Because of this, two distinct models, induced-fit and conformation selection, are proposed to explain the protein-protein recognition mechanism. In the induced-fit model, the interactions between a protein and its binding partner induce a series of conformation changes in the protein. In conformation selection model, it is thought that unbound and bound conformations pre-exist even before the

interaction takes place. The interaction merely causes the bound state to be more favorable and the population to shift towards the bound state.

Table 2.1 Q factors and CCs between the experimental and calculated RDC of 1UBQ and 1YD8 ensemble.

leave out HC	NH RDC	HC RDC	NC RDC
Q factor for the equal weighted ensemble	0.4463	0.3908	0.3151
Q factor for the 88% 1UBQ vs. 12% 1YD8 weighted ensemble	0.3105	0.2826	0.2416
CC for the equal weighted ensemble	0.9104	0.9227	0.9608
CC for 88% 1UBQ vs. 12% 1YD8 weighted ensemble	0.9557	0.9595	0.9761

Results and Discussions

Case study using an artificial two-structure ubiquitin ensemble

To demonstrate the importance of having relative populations in an ensemble, we first apply our method to a ubiquitin ensemble consisting of only two structures (pdb-id: 1UBQ and 1YD8). 1UBQ is a free structure and 1YD8 is a bound ubiquitin structure in complex with human GGA3 GAT domain. In this case study, we will determine the relative populations between the two structures and show how much they can help improve the accuracy in RDC calculations. To this end, 56 Non-HC RDC datasets are used in the least square fitting (see Methods section) and the relative populations are determined - 88% and 12% respectively for 1UBQ (free) and 1YD8 (bound). We compute the Q factors and correlation coefficients (CCs) between the experimental and calculated RDC of the NH, HC (not used in population calculations), and NC datasets. Table 2.1 shows the differences in Q-factors and CCs with and without relative populations. The substantial improvements in Q factors and CCs clearly demonstrate the significant contribution that relative populations can bring to reproducing RDC data.

Ensembles with relative populations

In this section, we apply our approach to several well-known Ubiquitin ensembles. These include, an X-ray structure ensemble consisting of 46 X-ray ubiquitin structures, 4 ubiquitin

Table 2.2 Q-factors and CCs between the experimental and calculated NH RDC of different ensembles.

NH RDC (leave out HC)	1D3Z (NMR ensemble)	2NR2 (MUMO Ensemble)	1XQQ (DER Ensemble)	2K39 (EROS Ensemble)	46 X-ray Structures	46 X-ray Structures + 1D3Z
Q factor for the equal weights	0.1699	0.2744	0.3004	0.0917	0.2263	0.1945
Q factor for the RDC derived weights	0.1649	0.2017	0.2494	0.0886	0.1991	0.1511
CCs for the equal weights	0.9874	0.9628	0.9561	0.9961	0.9740	0.9809
CCs for the RDC derived weights	0.9875	0.9797	0.9689	0.9963	0.9800	0.9887

Table 2.3 Q-factors and CCs between the experimental and calculated HC RDC of different ensembles.

HC RDC (leave out HC)	1D3Z (NMR ensemble)	2NR2 (MUMO Ensemble)	1XQQ (DER Ensemble)	2K39 (EROS Ensemble)	46 X-ray Structures	46 X-ray Structures + 1D3Z
Q factor for the equal weights	0.2282	0.3502	0.4031	0.2249	0.2465	0.2305
Q factor for the RDC derived weights	0.2268	0.3117	0.3786	0.2224	0.2407	0.2135
CCs for the equal weights	0.9736	0.9340	0.9129	0.9727	0.9682	0.9721
CCs for the RDC derived weights	0.9738	0.9469	0.9225	0.9730	0.9704	0.9760

ensembles determined computationally (with experimental constraints), and a combined ensemble made up of X-ray ubiquitin structures and an NMR ensemble (1D3Z). In the first test, 56 non-HC RDC datasets are used to obtain the relative populations of the structures in the different ensembles. Table 2.2 shows the improvement in Q-factors from equal population to RDC derived populations in 36 NH RDC datasets. The Q-factor of the X-ray structure ensemble improves from 0.2263 to 0.1991 when RDC derived weights (i.e., populations) are used. The Q-factor of the X-ray and NMR combined structure ensemble improves to 0.1511 when relative populations are used and it outperforms the NMR ensemble or the X-ray ensemble alone. We observe improvements in Q-factors from equal population to RDC derived populations among all ensembles, with the MUMO ensemble (2NR2) having the greatest improvement. Table 2.3

shows the results of HC RDC, which is used here as a testing dataset since it is not used in determining the populations. The Q-factor of the X-ray and NMR combined structure ensemble improves from 0.2305 to 0.2135. The X-ray and NMR combined ensemble with RDC derived populations outperforms all the other ensembles in reproducing the HC RDC.

Table 2.4 Q factors and CCs between the experimental and calculated NC RDC of different ensembles.

NC RDC (leave out NC)	1D3Z (NMR ensemble)	2NR2 (MUMO Ensemble)	1XQQ (DER Ensemble)	2K39 (EROS Ensemble)	46 X-ray Structures	46 X-ray Structures + 1D3Z
Q factor for the equal weights	0.2711	0.3708	0.4645	0.2555	0.2474	0.2415
Q factor for the RDC derived weights	0.2697	0.3196	0.4219	0.2602	0.2465	0.2459
CCs for the equal weights	0.9790	0.9569	0.9377	0.9803	0.9825	0.9833
CCs for the RDC derived weights	0.9793	0.9697	0.9467	0.9796	0.9821	0.9827

Table 2.5 Correlation Coefficient between the weights derived from non-NC and non-HC data sets.

Ensemble	1D3Z (NMR ensemble)	2NR2 (MUMO Ensemble)	1XQQ (DER Ensemble)	2K39 (EROS Ensemble)	46 X-ray Structures	46 X-ray Structures + 1D3Z
Correlation Coefficient	0.9934	0.9793	0.9812	0.7813	0.9964	0.9949

Table 2.4 shows the same results as Table 2.3 except in this case NC RDC data, instead of HC RDC, are left out for validation. The Q factor for NC RDC using the X-ray and NMR structure combined ensemble slightly deteriorates when using the derived relative populations, which is likely due to the noise in RDC data. The X-ray and NMR combined ensemble with RDC derived populations again has the best Q-factor of 0.2459 in NC RDC and it outperforms all the other ensembles.

The relative populations of the structures with an ensemble as determined above using non-NC and non-HC datasets may not necessarily be consistent. To check how robust and consistent these populations are, we compute the correlation coefficients between the two sets

of populations that are derived from non-NC and non-HC datasets respectively and the results are shown in Table 2.5. For all the ensembles except EROS, the populations derived from the two datasets are strongly correlated and have a correlation coefficient over 0.97. The relatively low correlation coefficient between the populations found for the EROS ensemble may reflect the fact that the structures in this ensemble were originally intended to have equal populations and were optimized to fit the non-NC RDC data (with NC data left out for validation).

Table 2.6 Structures with non-zero populations in the X-ray and NMR combined ensemble.

PDB ID	Chain ID in X-ray Structure or Model Number in NMR Structure	Weights	Comments
1UBQ	A	0.1005	Free structure
1S1Q	B	0.0114	
1AAR	B	0.0523	Diubiquitin
1CMX	B	0.1125	Largest weight among the bound structures
1TBE	A	0.0504	Tetraubiquitin
1TBE	B	0.0422	
2C7N	H	0.0466	
2D3G	A	0.0419	
2FCQ	A	0.0051	
2FCQ	B	0.0193	
2G45	E	0.0246	
1YIW	C	0.0559	
1F9J	B	0.0158	Tetraubiquitin
1D3Z	2	0.2631	NMR Structure
1D3Z	7	0.1392	NMR Structure
1D3Z	10	0.0192	NMR Structure

Table 2.6 shows all the structures with non-zero populations in the X-ray and NMR combined ensemble and the corresponding populations. Although the original ensemble consists of 56 X-ray or NMR structures, only 16 structures end up having non-zero contribution to the final RDC. Among these 16 structures with non-zero populations, only three of them are NMR structures, but their populations sum up to over 40%. The ligand-free crystal structure 1UBQ has a population of about 10%, which together with that of NMR structures, adds up to a population of over 50% for free structures. Among the bound structures, 1CMX-B, which

binds to Ubiquitin c-terminal hydrolase L3, has the largest population of 11%. For the 40 structures with zero population, either they indeed represent a state that is not populated, or they are close to a conformation with significant population (one of 16 structures with non-zero populations) and their contribution is somehow overshadowed.

Conformation Selection or Induced Fit?

Lange et al. (2008) presented a RDC-derived ensemble (named EROS) for Ubiquitin. It was argued that conformation selection was sufficient to explain the recognition dynamics of Ubiquitin since all the 46 crystal structures are within less than 0.8 Å backbone root mean square distance (RMSD) away from at least one of the structures in the EROS ensemble. In the paradigm of relative populations presented in this work, one may question whether the EROS conformations that the 46 X-ray structures are closest to, are significantly populated, or have any population at all. How confident can one be about the necessity for including any single conformation in the ensemble? While the EROS ensemble as a whole may have captured most of the dynamics revealed in RDC data, it is a much more challenging task to establish that any single conformation is essential and even irreplaceable to the ensemble. After all, it is quite possible that there may exist other ensembles that are able to reproduce the same set of experimental data. In other words, there may exist multiple solutions to the same experimental constraints. In what follows next, we want to address two questions: the first one is, "Does there exist another solution?" The second question is, "If multiple solutions (ensembles) exist, do they have consensus on the recognition dynamics?" In other words, will we still see that all the 46 crystal structures of Ubiquitin are close to at least one of the conformations in the ensembles?

The existence of multiple ensemble solutions

The answer to the first question is affirmative. Indeed, when our algorithm is applied to the EROS ensemble to determine the relative populations, we are able to lower the NH Q-factor further. And the derived populations tell us that about 40% of the conformations in the

ensemble have zero population. In other words, one may pick a subset of the EROS ensemble, those with non-zero populations, and assign them with proper relative populations and that will represent an equally good, if not better, ensemble than the original EROS. Another solution may be the X-ray and NMR combined ensemble that contains only 16 conformations (3 NMR modes and 13 crystal structures). One big advantage of this ensemble is that it is composed of purely experimental structures, thus the quality of each conformation is more reliable. The NH Q-factor of this ensemble is 0.15, which is higher than that of EROS (which is 0.09). However, the NC and HC Q-factors are both lower than those of EROS. We construct yet another possible solution by combining X-ray/NMR ensemble with some of the conformations in EROS. We repeatedly select one conformation from the EROS ensemble and add it to the X-ray/NMR ensemble. The conformation being selected at each iteration is the one that can lower the most the NH Q-factor of the resulting ensemble. We select 11 EROS conformations in this way and the final ensemble, which has 17 experimental structures and the newly added 11 EROS conformations, has a NH Q-factor that is 0.0991, which is about the same as that of EROS.

Consensus?

To answer the second question, "do the multiple solution ensembles have a consensus regarding the recognition mechanism?" we plot the backbone root mean square distance (RMSD) from the 46 crystal structures to their closest conformations in the aforementioned four ensembles original EROS (118 structures), EROS subset with relative populations (68 structures), X-ray/NMR combined ensemble (16), and X-ray/NMR/EROS ensemble (28 structures).

From figure 2.1, we see that the minimum backbone RMSDs from the 46 X-ray structures to the original EROS ensemble all fall below roughly 0.8 Å. The ensemble using EROS subset with relative populations (dark green curve) has higher RMSD than EROS, which is expected since it contains fewer conformations than EROS. The backbone RMSD (i.e., the closest RMSD distance from an X-ray structure to any conformation in the ensemble) to the EROS subset ensemble (dark green curve in the figure) rises up for a number of crystal structures. Particu-

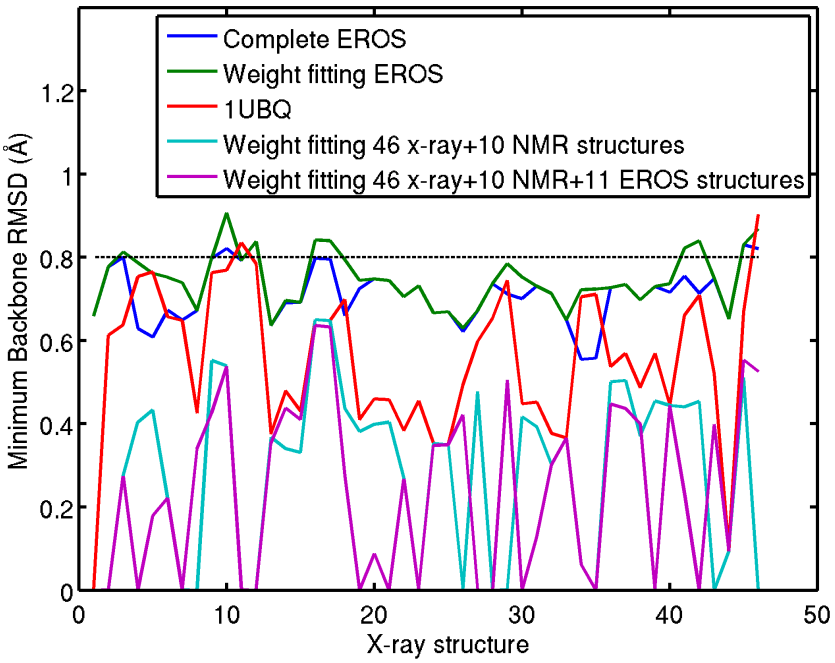


Figure 2.1 The minimum backbone RMSD between 5 ensembles and x-ray strcutures.

larly, it rises up to over 0.9 Å for crystal structure 1P3Q (chain V). The X-ray/NMR combined ensemble (cyan curve) and X-ray/NMR/EROS ensemble (purple curve), on the other hand, have even smaller backbone RMSD, falling below 0.5 Å for most of the crystal structures.

Based on these observations, can one say that all these four ensembles have consensus on the recognition dynamics, or that they are all in favor of conformation selection? In Lange et al. (2008), the authors concluded that the recognition dynamics is mostly conformation selection, based on the fact that all 46 x-ray structures fall within 0.8 Å backbone RMSD to at least one of the structures in the EROS ensemble. The threshold value 0.8 Å is critical in the interpretation. The conclusion would have been different if a lower value than 0.8 Å had been used. To put this in perspective, let us consider the backbone RMSD from a single free structure (1UBQ) to all the other 45 crystal structures. These values, shown in the red line in Figure 1, show how far the other structures, mostly bound structures, are from

the free structure. The values set the upper limit on the threshold value that can be used in interpreting the ensembles. As an extreme case, if one chooses a threshold value of 0.90 Å (which is the largest distance from a bound structure, 1F9J-B, to the free structure) and say any structure that falls with this distance to the free structure is considered conformation selection, then the recognition mechanism would by default be conformation selection for all the bound structures.

For this reason, we probably should not increase the threshold value further than 0.8 Å. If this is the case, we could conclude that all the four ensembles have a consensus and favor conformation selection for most of the crystal structures. However, the recognition mechanism is unclear for the other structures whose backbone RMSD to at least one of the ensembles is higher than 0.8 Å (above the dashed line in the figure). These structures and their closest RMSD to the EROS subset ensemble (dark green curve in the figure) are: 1S1Q-D (0.81 Å), 1P3Q-V (0.91 Å), 1TBE-B (0.84 Å), 1YDB-U (0.84 Å), 1YDB-V (0.84 Å), 1YIW-A (0.82 Å), 1YIW-B (0.84 Å), 1F9J-A (0.83 Å), 1F9J-B (0.87 Å). On the other hand, probably a lower threshold value should be used. Because when one uses 0.8 Å as the threshold value for characterizing conformation selection, one has already assumed that all structures, except bound structures 1TBE-A and 1F9J-B (which are the peaks on the red line in the figure that surpass 0.8 Å), are in favor of conformation selection. However, if we use a lower threshold value, say 0.6 Å, then there is no consensus at all regarding the recognition mechanism among the four ensembles. The four ensembles represent four different solutions and they yield different conclusions on the matter.

Conclusion

In this work, we develop a novel computational method that is able to determine the relative populations of structures within an ensemble by employing the iterative least squares method to fit the experimental RDC data. We compare Q-factors among several ubiquitin ensembles and the results show that ensembles with RDC derived populations significantly improve the agreement between the calculated and experimental RDCs. Consequently, we are able to use

this method to determine several ubiquitin ensembles with low Q-factors, which would not have been possible if relative populations had not been used. These ensembles may represent different solutions to the same experimental constraints. Therefore, an interesting question that arises here is whether or not these ensembles have consensus regarding the recognition mechanism for ubiquitin. The answer to this depends heavily on the threshold value used for characterizing conformation selection, which is the backbone root mean square distance from a bound structure to the closest populated conformation state. When a high threshold value is used, all the four ensemble solutions are in favor of conformation selection for most of the bound structures, while a clear consensus cannot be reached if a lower threshold value is used instead. The low backbone RMS distances between the bound structures and the free structure 1UBQ strongly suggest that a lower threshold value should be used.

Materials and Methods

In this section, we will present the algorithm for deriving the relative populations of the structures within an ensemble and the cross validation method. The ensembles and RDC datasets used are also given.

Structural alignment

Structural alignment is used to align the structures within a given ensemble to the common coordinate system. Therefore, all the structures can be assumed to have the same molecular reference frame after the alignment.

Residual dipolar coupling (RDC) calculation of a single structure

Residual dipolar coupling comes from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field and is defined by Cornilescu et al. (1998)

$$D_{ij} = \frac{hr_i r_j}{(2\pi r)^3}\langle 3cos^2\theta - 1\rangle \tag{2.1}$$

where r_i and r_j are the nuclear magnetogyric ratios of the nuclei i and j, h is Planks constant, r is the internuclear distance between the two nuclei and θ is the angle between the internuclear

vector and the external magnetic field. The brackets signify the average. Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media.

With regards to the 3D structure, the RDC (D_{ij}) can be expressed according to the molecular frame. First, the elements of the Saupe matrix are defined as

$$S_{lm} = \langle \frac{3cos\beta_l cos\beta_m - k_{lm}}{2} \rangle \tag{2.2}$$

where β_l denotes the orientation of the l-th molecular axis with respect to the external magnetic field. The RDC (D_{ij}) can be reformulated in the molecular frame as

$$D_{ij} = \frac{hr_i r_j}{(2\pi r)^3} \begin{pmatrix} \alpha_y^2 - \alpha_x^2 & \alpha_z^2 - \alpha_x^2 & 2\alpha_x\alpha_y & 2\alpha_x\alpha_z & 2\alpha_y\alpha_z \end{pmatrix} \begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \tag{2.3}$$

where α_x , α_y and α_z are the cosines of the angles between the bond vector of the two nuclei and the x, y and z axes of the molecular frame. Let $\alpha_{x,k}$, $\alpha_{y,k}$ and $\alpha_{z,k}$ represent the k-th α_x, α_y and α_z. When all the bond vectors are considered, we will have the following formula.

$$D_{exp} = \frac{hr_i r_j}{(2\pi r)^3} \begin{pmatrix} \alpha_{y,1}^2 - \alpha_{x,1}^2 & ... & 2\alpha_{y,1}\alpha_{z,1} \\ : & : & : \\ \alpha_{y,N}^2 - \alpha_{x,N}^2 & ... & 2\alpha_{y,N}\alpha_{z,N} \end{pmatrix} \begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \tag{2.4}$$

where D_{exp} is the experimental D_{ij} of all interactions and N is the total number of interactions in a protein structure. Equation 2.4 can be rewritten in the following matrix form:

$$D_{exp} = cAS \tag{2.5}$$

where c is the constant $\frac{hr_i r_j}{(2\pi r)^3}$ and A is the Nx5 matrix in the equation 2.4 and the S is the five element vector. Basically, the S and D_{calc} can be calculated from the Moore-Penrose

pseudoinverse of matrix A.

$$S = A^{-1} D_{exp} \tag{2.6}$$

$$D_{calc} = A A^{-1} D_{exp} \tag{2.7}$$

Residual dipolar coupling (RDC) calculation of an ensemble

The RDC calculation method for a single structure can be extended to take ensemble averaging into account so that the ensemble D_{calc} can be obtained. First consider the assumption that all structures have equal contributions toward the experimental RDC: D_{exp} . When an ensemble with equal weights is considered, we will have the following formula.

$$\left(\frac{A_1}{n} + \frac{A_2}{n} + ... + \frac{A_k}{n} + ... \frac{A_n}{n} \right) S = D_{exp} \tag{2.8}$$

where A_k is the A matrix obtained from the k-th structure. S can be obtained from the following formula.

$$S = \left(\frac{A_1}{n} + \frac{A_2}{n} + ... + \frac{A_k}{n} + ... \frac{A_n}{n} \right)^{-1} D_{exp} \tag{2.9}$$

Now consider another assumption that different structures may have different populations and thus different contributions toward the D_{exp} and can be combined linearly. Therefore, weights (representing the relative populations) are given to different structures and the following formula is used to represent the combination:

$$\left(w_1 A_1 + w_2 A_2 + ... + w_k A_k + ... + w_n A_n \right) S = D_{exp} \tag{2.10}$$

where n is the total number of structures and w_k and A_k are the relative population (or weight) and A matrix of the k-th structure. Thus, S can be obtained from the following formula.

$$S = \left(w_1 A_1 + w_2 A_2 + ... + w_k A_k + ... + w_n A_n \right)^{-1} D_{exp} \tag{2.11}$$

The definition of our problem is thus to find the optimal relative populations of the structures within the ensemble such that the experimental RDC is best reproduced.

Iterative least squares fitting for optimal populations for a single RDC data set

In the process of back-calculating the residual dipolar coupling (RDC) from a protein structure or ensemble, singular value decomposition is used to obtain a least square solution for the alignment tensor. We apply the same technique iteratively to obtain the optimal relative populations for a given ensemble. Due to the assumption linearity, the weights can be obtained via iterative least squares fitting. First, an equal value is given for all populations and Equation 2.11 is used to obtain S. After S is obtained, it is used to determine the w_ks via least squares fitting. The process is iterated until the weights converge. In the end, each structure has either positive or zero population, since the weights are derived with nonnegative constraints [Lawson and Hanson (1995)]. The following algorithm gives the detailed implementation of the iterative least squares fitting for a single RDC data set.

Iterative Least Squares Fitting $([A_1 A_2 \cdots A_n], D_{exp})$

for $i = 1$ to n **do**

$new\_weights(i) \leftarrow \frac{1}{n}$

end for

repeat

$old\_weights \leftarrow new\_weights$

$A \leftarrow old\_weights(1) * A_1 + \cdots + old\_weights(n) * A_n$

$S \leftarrow pseudo\_inverse(A) * D_{exp}$

$AS \leftarrow [A_1 S \; A_2 S \cdots A_n S]$

$new\_weights \leftarrow non\_negative\_least\_squares(AS, D_{exp})$

until *old_weights* and *new_weights* converge

return *new_weights*

Iterative least squares fitting for optimal relative populations for multiple RDC data sets

In the case of multiple RDC data sets, different alignment tensors are calculated for different media. The optimal weight combination (the relative populations) is obtained by least squares

fitting to all the RDC data sets. The following algorithm gives the detailed implementation of the iterative least squares fitting for multiple RDC data sets.

Iterative Least Squares Fitting Multiple RDCs $([A_1 A_2 \cdots A_n], [D_1, D_2 \cdots D_m])$

for $i = 1$ to n **do**

$new\_weights(i) \leftarrow \frac{1}{n}$

end for

repeat

$old\_weights \leftarrow new\_weights$

$A \leftarrow old\_weights(1) * A_1 + \cdots + old\_weights(n) * A_n$

for $i = 1$ to m **do**

$S(i) \leftarrow pseudo\_inverse(A) * D_i$

$AS(i) \leftarrow [A_1 S(i)\ A_2 S(i) \cdots A_n S(i)]$

end for

$$AS\_all \leftarrow \begin{pmatrix} AS(1) \\ AS(2) \\ . \\ . \\ . \\ AS(m) \end{pmatrix}$$

$$D\_all \leftarrow \begin{pmatrix} D_1 \\ D_2 \\ . \\ . \\ . \\ D_m \end{pmatrix}$$

$new\_weights \leftarrow non\_negative\_least\_squares(AS\_all, D\_all)$

until *old_weights* and *new_weights* converge

return *new_weights*

Table 2.7 The selected ensembles.

Ensemble	Structure determination method	Number of structures
X-ray structures	X-ray crystallography	46
1D3Z	NMR spectroscopy [Cornilescu et al. (1998)]	10
2NR2	Minimal Under-restraining Minimal Over-restraining (MUMO) [Richter et al. (2007)]	144
1XQQ	Dynamic Ensemble Refinement (DER) [Lindorff-Larsen et al. (2005)]	128
2K39	Ensemble Refinement with Orientational restraints (EROS) [Lange et al. (2008)]	116
X-ray structures+ 1D3Z	X-ray crystallography and NMR spectroscopy	56

Table 2.8 The RDC data sets for obtaining the weights and validation.

Source	RDC Type	Number of RDC data sets
Lakomek et al. (2008)	NH	13
Lakomek et al. (2006)	NH	5
	NC'	4
	HC'	4
Ottiger and Bax (1998)	NH	2
	CaC'	2
	CaHa	2
	NC'	2
	HC'	2
Tolman (2002)	NH	9
Ruan and Tolman (2005)	NH	7
Kontaxis and Bax (2001)	Methyl	10

The ubiquitin ensemble and RDC data set for obtaining the weights and validation

Ubiquitin has long been used as a model protein to probe the protein dynamics and several ubiquitin ensembles exist in the PDB and were generated from different structure determination methods to satisfy both dynamic and structural constraints. In this work, we select 4 well-known ubiquitin ensembles and an X-ray structure ensemble consisting of 46 X-ray structures for our study. We also form a combined ensemble of X-ray and NMR structures. Table 2.7 shows the selected ensembles. A total of 62 RDC data sets, including NH, NC', HC', CαC', CαHα and side chain methyl, are used to obtained the weight combinations (i.e., relative populations) of the 6 selected ubiquitin ensembles. Table 2.8 shows the types and references

for these 62 RDC data sets.

The cross validation

Q factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is calculated by the following formula [Cornilescu et al. (1998)],

$$Q = \frac{\sqrt{\sum (D_{calc} - D_{exp})^2}}{\sqrt{\sum D_{calc}^2}} \tag{2.12}$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC. We also use the correlation coefficient to measure the agreement, which is calculated as the following,

$$\rho = \frac{(D_{calc} - \overline{D_{calc}})(D_{exp} - \overline{D_{exp}})}{\sqrt{\sum (D_{calc} - \overline{D_{calc}})^2 \sum (D_{exp} - \overline{D_{exp}})^2}} \tag{2.13}$$

We performed two rounds of cross validations. In the first round, all 6 HC RDC data sets were left out of the weight evaluation process and were used for validation only. The alignment tensors used to compute the HC RDCs in different alignment media were determined from non-HC RDCs. In the second round, NC RDC data sets were left out for validation and the alignment tensors for NC RDCs in this round are obtained from non-NC RDCs.

Acknowledgements

Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged.

CHAPTER 3. EVALUATING THE QUALITY OF CONFORMATION SAMPLING METHODS USING EXPERIMENTAL RESIDUAL DIPOLAR COUPLING DATA

A paper to be submitted

Tu-Liang Lin, Santhosh Kumar Vammi and Guang Song

Abstract

Many computational approaches have been developed and used for sampling protein conformations near the native state. However, it has been difficult to evaluate the qualities of the conformations sampled or to compare them among the various sampling schemes. In this work, we develop a novel method for evaluating the quality of conformation ensembles and apply it to evaluate ubiquitin conformations generated from four widely-used conformation sampling approaches, namely, MD simulation, Elastic Network Model (ENM), CONCOORD, and tCONCOORD. We choose ubiquitin because there exists abundant experimental residual dipolar coupling (RDC) data on this protein. RDC data contains rich ensemble-averaged information about a given protein and thus provide tight constraints that can be used for probing what conformations should make up the protein ensemble. Our results demonstrate that the conformations generated by MD simulations are the best among all sampling methods. Specifically, MD simulation performs significantly better than the other methods in capturing the side chain motions. The backbone flexibility modeled and sampled by tCONCOORD comes quite close, with CONCOORD and ENM trailing behind.

Introduction

Conformational sampling

It is now well accepted that the functions of a protein are closely related to not only its structure but also its dynamics. There is strong evidence showing that the best representation of the native states of a protein should be an ensemble of structures [Karplus and McCammon (2002)]. Therefore, numerous computational approaches have been developed and used to sample the conformational space around the native state in hope that the distribution and dynamic transition among the conformation states can be well studied and understood.

MD simulation is one of the most commonly used computational approaches for conformation sampling. The main limitation of MD simulation is its high computation cost, which greatly limits the timescale that can be reached, especially for large systems. Recently it also has been shown that MD simulations beyond hundreds of nanoseconds might have the potential risk of running into high free energy states and staying there for a long time, thus skewing the populations [Lange et al. (2010)]. Therefore, obtaining the correct sampling in a broad time-scale is still a challenge. To overcome such difficulties, residual dipolar couplings (RDCs) data, which provide complementary dynamics information that is inaccessible to NMR relaxation methods, have been used as constraints in conformation sampling [Lange et al. (2008); Mittermaier and Kay (2006)].

Elastic Network Model (ENM) is another choice for conformation sampling. In the last decade various coarse-grained elastic network models have been developed to study the large-scale motions of proteins and protein complexes where computer simulations using detailed all-atom models are not feasible. Among these models, the Gaussian Network Model (GNM) and the Anisotropic Network Model (ANM) have been widely used [Bahar et al. (1997); Atilgan et al. (2001)] due to their simplicity. Specifically, the analytic solutions to residue fluctuations and motion correlations can be easily derived. In this work, we use one of the recently developed ENMs, the Torsional Network Model (TNM), which uses the backbone torsional angles as the essential degrees of freedom of the protein [Mendez and Bastolla (2010)]. The major advantage

of TNM is that the covalent bond geometry, such as bond lengths and bond angles, are naturally conserved in the motions represented by the modes. CONCOORD uses geometrical constraints to sample the conformational space. From a given input structure, a geometric description of the structure is calculated and this geometric description can be used to generate hundreds of structures [de Groot et al. (1997)]. Later, CONCOORD is re-implemented as tCONCOORD to allow the sampling of conformational transitions of a protein under geometrical constraints [Seeliger et al. (2007); Seeliger and De Groot (2009)]. The advantage of the CONCOORD and tCONCOORD over MD is computational efficiency.

With the development of various structure determination methods, protein structures are becoming increasingly more available and for some well-studied proteins, tens and even hundreds of structures (of the same protein) have been determined and are available in the PDB. These structures have been shown to capture a representative subset of the native-state ensemble [Best et al. (2006)]. Therefore, an ensemble formed by experimental structures from the PDB can be regarded as samplings too and can be used as a reference frame to examine the other conformation sampling methods.

For most conformation sampling approaches, the relative populations of the sampled conformations in the conformation space are not known. Currently, few methods exist for obtaining such relative populations even though they are essential information for describing the conformation space. Recently, we developed a method for determining the relative populations of protein conformation states using experimental residual dipolar coupling data and applied it to study the protein recognition mechanism of ubiquitin [Lin and Song (2011)].

The abundant ensemble-averaged information included in RDC data also provides tight constraints that can be used for probing what conformations should make up the ensemble. Thus, the aim of this work is to evaluate the sampling ability of different computational approaches. For a given conformation sampling method, we will examine how well the conformations it generates can reproduce the experimental RDC data. Our hypothesis is that the samplings that best approximate the conformation space should also reproduce the RDC data with the highest fidelity.

Besides the aforementioned methods, there exist other computational approaches for constructing an ensemble, such as the loop prediction program [Shehu et al. (2006)], the geometric restriction checking program [de Groot et al. (1997)], and the chemical shift prediction algorithm [Jensen et al. (2010)].

Significance of the work

Although many conformation sampling methods have been proposed, to the best of our knowledge there is not much effort to compare the sampling qualities of these methods. The algorithm presented in this work provides a way to measure the sampling quality of a given conformation sampling method, and to identify its strengths and weaknesses. Current work uses RDC data as the constraints to evaluate an ensemble. However, this can be easily extended to include other experimental data, such as NMR order parameters.

The residual dipolar coupling (RDC)

The residual dipolar coupling is an interaction that exists between two magnetic nuclei. Under isotropic solution condition, dipolar coupling averages to zero as a result of the effects of Brownian motion and tumbling of the molecule. The use of an alignment medium that can create a weak force on the protein and eventually leads to an incomplete averaging of anisotropic magnetic interactions makes the measurement of residual dipolar coupling possible. Therefore, the residual dipolar coupling (RDC) between two spins represents the incomplete averaging of spatially anisotropic dipolar couplings. One intriguing property of RDC is that it can probe the blind spot of the conventional NMR relaxation. Conventional NMR relaxation has been used to probe the dynamics faster than the rotational correlation time of a system which can range from picoseconds to nanoseconds or to identify the dynamics slower than microseconds. However, many important biological processes happen in the blind spot of the conventional NMR, between nanoseconds to microseconds, where RDC plays an important role in protein dynamics studies. In this work, we develop an algorithm that can exploit the RDC data and use them as constraints in evaluating the sampling qualities of different conformation sampling

approaches.

Results

Case study using an artificial two-structure ubiquitin ensemble

Protein structure ensembles are represented by a collection of conformations. In Lin and Song (2011), we have shown that, in order to better describe the conformation space using an ensemble, it is helpful to introduce information about the relative population of the ensemble, i.e., an ensemble that is described not only by a set of conformations, but also by the relative population of each conformation. Such relative populations are not readily available most times when an ensemble is generated but can be determined by iterative least square fitting as described in Lin and Song (2011). For example, in a case study performed on an artificial ubiquitin ensemble consisting of only two structures (pdb-id: 1UBQ and 1YD8, with 1UBQ being a free structure and 1YD8 a bound ubiquitin structure in complex with human GGA3 GAT domain), we showed that introducing relative populations between the two structures greatly improved the accuracy in reproducing the experimental RDC data (as compared to the case without relative populations) [Lin and Song (2011)].

The results of a 50-ns MD simulation of ubiquitin

In this section, we examine the trend of Q-factor changes over time based on conformations cumulated in a 50-ns MD simulation (see Figure 3.1). At the beginning, the Q-factors of the backbone RDCs, NH, NC and HC RDC, are below 0.45, but at the end of 50-ns simulation, the Q factors increase to above 0.6. The Q-factor of the side chain RDC is very high at the beginning, but it decreases to 0.5 around 5 ns. The Q-factor of NH RDC also reaches its minimum at 2 ns. The results imply that longer simulations do not necessarily result in better correlations between the experimental and calculated RDCs. This phenomenon has also been observed in recent work by Lange et al. in several 1000-ns MD simulations using different force fields [Lange et al. (2010)].

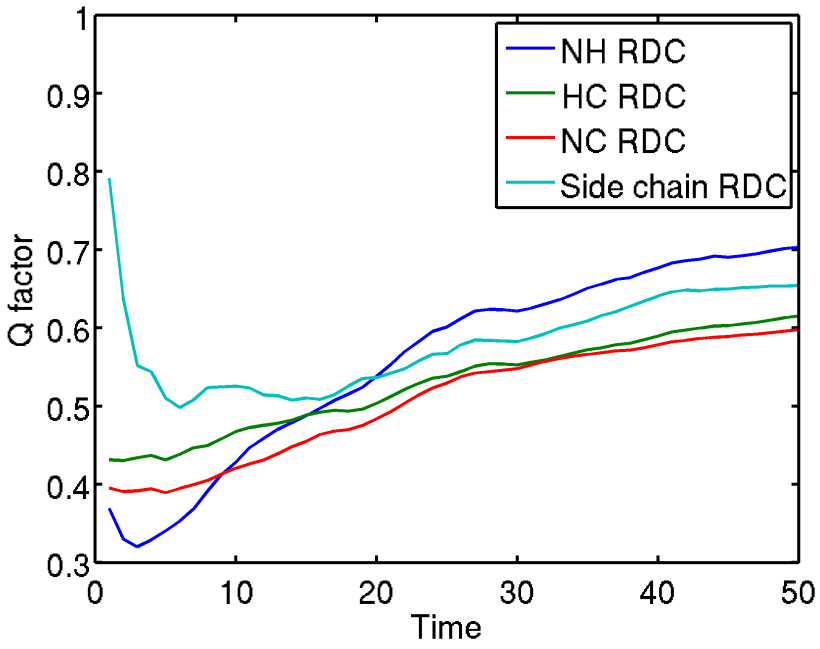


Figure 3.1 The changes of Q factors over time.

The trends suggest that as simulation progresses in time, the accumulated population may not necessarily increase in their accuracies in approximating the conformation space. There could be several reasons for this. First, the simulation has not been run long enough and some conformation states have not been reached. Secondly, the simulation has not been run long enough and the samplings are biased by what the starting structure is and what the initial simulation conditions are. Lastly, the force fields may not be accurate enough to give the right proportion of populations to the different conformation states. If the cause is one or both of the latter two, the problem can be solved by re-assigning a relative population to each conformation. If it is the first one, then a different sampling scheme has to be applied to reach those conformation states.

Our newly developed iterative least square RDC fitting algorithm [Lin and Song (2011)] is applied to the 50-ns MD ensemble. After re-assigning relative populations, the Q-factors of all the RDCs are lowered significantly (see Table 3.1), but they remain quite high. This indicates the simulation, though 50-ns long, has not reached some important conformation

Table 3.1 Overall Q-factors and cross correlations (CCs) between the experimental and calculated RDCs using 50-ns MD ensemble.

50-ns MD	NH RDC	HC RDC	NC RDC	Side chain RDC
Q factor for the equal weighted ensemble	0.70	0.62	0.60	0.65
Q-factor for ensemble with relative population	0.27	0.38	0.33	0.34
CC for the equal weighted ensemble	0.81	0.83	0.86	0.84
CC for ensemble with relative population	0.96	0.92	0.95	0.95

states that significantly contribute to the observed RDCs. In the sections that follow, we will show this problem can be mostly alleviated by running MD simulations from multiple starting structures.

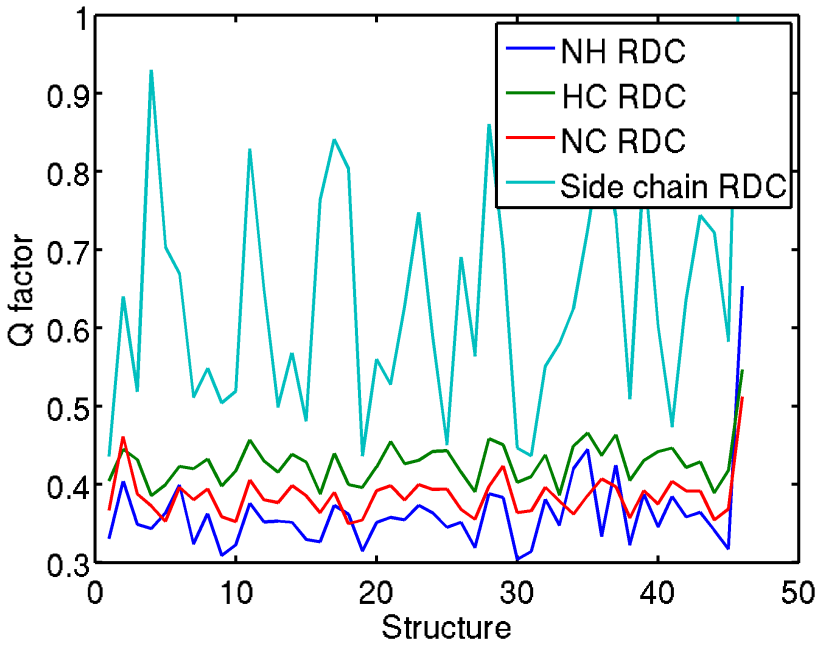


Figure 3.2 The fluctuations of Q-factors on the starting structure of MD simulations.
46 different X-ray structures, shown in the abscissa, are used as the starting points in the simulations.

Multiple shorter MD simulations using different X-ray structures as starting points

In this section, we will investigate the quality of the conformations generated from multiple shorter (2-ns each) MD simulations using different structures as the starting points. 46 X-

ray ubiquitin structures are chosen and 2-ns MD simulations are performed for each chosen structure. For each structure, 1,000 conformations (one per every 2-ps) are collected in the corresponding MD simulation and are used to compute the Q-factors. Figure 3.2 shows the fluctuations of Q-factors between the calculated and experimental RDCs when each of these 46 structures is used as the starting point of the simulation. Side chain RDC varies extensively in the range between 0.4 and 1. Only one structure (1F9J chain B) gives backbone RDCs that are higher than 0.5.

Table 3.2 Average Q factors and CCs between the experimental and calculated RDCs of 46 2-ns MD ensembles.

2-ns MDs of 46 xray structures	NH RDC	HC RDC	NC RDC	Side chain RDC
Q factor for the equal weighted ensemble	0.3625	0.4274	0.3854	0.6388
Q-factor for ensemble with relative population	0.2679	0.3617	0.3180	0.5051
CC for the equal weighted ensemble	0.9399	0.9044	0.9389	0.8103
CC for ensemble with relative population	0.9657	0.9305	0.9570	0.8747

The iterative least square RDC fitting algorithm (see [Lin and Song (2011)] or a review of the algorithm in the Methods section) is then applied to all the 46 2-ns MD ensembles. Table 3.2 shows the average Q-factor of the equal weighted ensemble (without relative populations) and RDC weighted ensemble (with relative populations). The Q-factors of all the RDCs are reduced by about 20-30% after the fitting.

Table 3.3 Q-factors of the conformation ensembles generated from different sampling approaches (without relative populations).

Sampling methods	MD	ENM	CONCOORD	tCONCOORD	46 X-ray Structural Ensemble*
Q factor of NH RDC	0.223	0.326	0.261	0.197	0.224
Q factor of HC RDC	0.328	0.438	0.297	0.268	0.255
Q factor of NC RDC	0.288	0.350	0.246	0.223	0.247
Q factor of side chain RDC	0.281	0.383	0.361	0.307	0.294

*The ensemble of the original 46 X-ray structures (last column) serves as a reference.

The results from MD, ENM, CONCOORD and tCONCOORD

In this section, we will examine and compare the conformations generated by four different sampling approaches, namely, MD, ENM, CONCOORD and tCONCOORD. Since it is much more difficult to sample well the conformation space from a single structure, as the sampling process can be easily trapped in a local energy well, we choose to use 46 known X-ray structures of ubiquitin as the starting points when evaluating all the aforementioned conformation sampling methods. When evaluating the sampling ability of a given method, a sub-ensemble of 200 structures is generated around each of the 46 X-ray structures (see Methods section for details on how the conformations are generated by the different methods). As a result, the final ensemble contains 46 sub-ensembles, each of which in turn contains 200 conformations, and thus has a total of 9,200 conformations. To evaluate how well such an ensemble can potentially reproduce the RDC data, we need to first determine the relative population of each conformation. However, due to the large size of the ensemble, it is not feasible to compute their relative populations directly using the iterative least square fitting method, as there would be more parameters than the number of data points. Therefore, we first determine the relative populations of the conformations within each sub-ensemble by fitting the ensemble to the experimental RDCs. Then the top ten conformations with largest populations are selected from each sub-ensemble to form the final ensemble, which now has only 460 conformations.

Table 3.4 Q-factors of the conformation ensembles generated from different sampling approaches (with relative populations).

Sampling methods	MD	ENM	CONCOORD	tCONCOORD
Q factor of NH RDC	0.154	0.210	0.191	0.161
Q factor of HC RDC	0.281	0.350	0.271	0.259
Q factor of NC RDC	0.228	0.282	0.23	0.215
Q factor of side chain RDC	0.215	0.262	0.258	0.241

Table 3.3 shows the Q-factors computed from the ensembles generated from different conformational sampling approaches. For all four ensembles, the 460 conformations in the ensemble are given an equal population. It is worth noting that tCONCOORD has the best performance for backbone RDCs among the four sampling approaches, while the MD ensemble has the best

performance for side chain RDCs. Table 3.4 shows the Q-factor results after assigning relative populations to the 460 conformations in each ensemble. Interestingly, the MD ensemble now outperforms all the other ensembles, including tCONCOORD, on both NH and side chain RDC data sets. On NC and HC RDCs, for both of which the experimental data are much sparser, MD ensemble performs about equally well as the other methods. This indicates that the MD ensemble contains conformations that resemble more closely the experimental populated conformation states of ubiquitin than any other ensembles. The reason why this is not obvious in Table 3 when equal populations are used is probably due to the fact that the simulations had not been run long to correctly reproduce the population distribution. The accuracies of force fields may have had some influences as well.

Discussions and Conclusion

In this work, we develop a novel method for evaluating the quality of conformation ensembles and apply it to evaluate ubiquitin conformations generated from four widely-used conformation sampling approaches, namely, MD simulation, Elastic Network Model (ENM), CONCOORD, and tCONCOORD. A protein's conformation space has very high dimensions and there exist many local energy minima. For proteins like Ubiquitin, abundant evidence exists showing that the protein has multiple well-populated conformation states that are separated by high energy barriers. The transition times among some of these states can be on the order of microseconds. Thus, it is infeasible to produce a good coverage of the conformation space starting from a single structure. Besides, most conformation sampling methods are suited only for local sampling, with MD simulation being perhaps the only exception. However, there are limitations with MD simulations as well. Conformations sampled by MD are more prone to be biased towards some regions of the conformation space and it takes extremely long simulations to remove the bias. In addition, it is possible that the force fields may lead the conformations away from the desired conformation states found in nature. Fortunately, most of these limitations can be alleviated by applying our recently developed RDC-based method that is able to re-assign the relative population of each conformation. For the aforementioned rea-

sons, we use 46 (instead of 1) known crystal structures of Ubiquitin as the starting points for generating conformations when evaluating each sampling method. Our results demonstrate that the conformations generated by MD simulations are still the best among all sampling methods. However, this is only true if the populations of the conformations in the ensemble have been readjusted. Specifically, MD simulations perform significantly better than the other methods in capturing the side chain motions, even before relative population readjusting is applied. The backbone flexibility modeled and sampled by tCONCOORD comes quite close, with CONCOORD and ENM trailing behind. Longer MD simulations probably are required to further justify the sampling ability of MD in the backbone motion and this will be investigated in the future research.

Materials and Methods

In this section, we give a detailed description of the processes by which conformations are generated using the four different sampling methods. For convenience, we also include here the algorithm [Lin and Song (2011)] for computing and reassigning relative populations to conformations within an ensemble so that they can best reproduce the experimental residual dipolar coupling (RDC) data. The RDC datasets used in this work and how RDC can be computed from a structure or an ensemble are also given.

Structural alignment

Structural alignment is used to align the structures within a given ensemble to the common coordinate system. Therefore, all the structures can be assumed to have the same molecular reference frame after the alignment.

Residual dipolar coupling (RDC) calculation of a single structure

Residual dipolar coupling comes from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field and is defined as Cornilescu et al. (1998)

$$D_{ij} = \frac{hr_i r_j}{(2\pi r)^3}\langle 3cos^2\theta - 1\rangle \tag{3.1}$$

where r_i and r_j are the nuclear magnetogyric ratios of the nuclei i and j, h is Planks constant, r is the internuclear distance between the two nuclei and θ is the angle between the internuclear vector and the external magnetic field. The brackets signify the average. Normally, the residual dipolar coupling is reduced to zero because of isotropic tumbling. The anisotropic measurement is obtained by the aid of various types of liquid crystalline media.

With regards to the 3D structure, the RDC (D_{ij}) can be expressed according to the molecular frame. First, the elements of Saupe matrix is defined as

$$S_{lm} = \langle \frac{3cos\beta_l cos\beta_m - k_{lm}}{2} \rangle \tag{3.2}$$

where β_l denotes the orientation of the l-th molecular axis with respect to the external magnetic field. The RDC (D_{ij}) can be reformulated in the molecular frame as

$$D_{ij} = \frac{hr_i r_j}{(2\pi r)^3} \begin{pmatrix} \alpha_y^2 - \alpha_x^2 & \alpha_z^2 - \alpha_x^2 & 2\alpha_x\alpha_y & 2\alpha_x\alpha_z & 2\alpha_y\alpha_z \end{pmatrix} \begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \tag{3.3}$$

where α_x , α_y and α_z are the cosines of the angles between the bond vector of the two nuclei and the x, y and z axes of the molecular frame. Let $\alpha_{x,k}$, $\alpha_{y,k}$ and $\alpha_{z,k}$ represent the k-th α_x, α_y and α_z. When all the bond vectors are considered, we will have the following formula.

$$D_{exp} = \frac{hr_i r_j}{(2\pi r)^3} \begin{pmatrix} \alpha_{y,1}^2 - \alpha_{x,1}^2 & ... & 2\alpha_{y,1}\alpha_{z,1} \\ : & : & : \\ \alpha_{y,N}^2 - \alpha_{x,N}^2 & ... & 2\alpha_{y,N}\alpha_{z,N} \end{pmatrix} \begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \tag{3.4}$$

where D_{exp} is the experimental D_{ij} of all interactions and N is the total number of interactions in a protein structure. Equation 3.4 can be rewritten in the following matrix form:

$$D_{exp} = cAS \tag{3.5}$$

where c is the constant $\frac{hr_ir_j}{(2\pi r)^3}$ and A is the Nx5 matrix in the equation 3.4 and the S is the five element vector. Basically, the S and D_{calc} can be calculated from the Moore-Penrose pseudoinverse of matrix A.

$$S = A^{-1}D_{exp} \tag{3.6}$$

$$D_{calc} = AA^{-1}D_{exp} \tag{3.7}$$

Residual dipolar coupling (RDC) calculation of an ensemble

The RDC calculation method for a single structure can be extended to take ensemble averaging into account so that the ensemble D_{calc} can be obtained. First consider the assumption that all structures have equal contributions toward the experimental RDC: D_{exp} . When an ensemble with equal weights is considered, we will have the following formula.

$$\left(\frac{A_1}{n} + \frac{A_2}{n} + ... + \frac{A_k}{n} + ...\frac{A_n}{n}\right) S = D_{exp} \tag{3.8}$$

where A_k is the A matrix obtained from the k-th structure. S can be obtained from the following formula.

$$S = \left(\frac{A_1}{n} + \frac{A_2}{n} + ... + \frac{A_k}{n} + ...\frac{A_n}{n}\right)^{-1} D_{exp} \tag{3.9}$$

Now consider another assumption that different structures may have different populations and thus different contributions toward the D_{exp} can be combined linearly. Therefore, weights (representing the relative populations) are given to different structures and the following formula is used to represent the combination:

$$(w_1A_1 + w_2A_2 + ... + w_kA_k + ... + w_nA_n)\, S = D_{exp} \tag{3.10}$$

where n is the total number of structures and w_k and A_k are the relative population (or weight) and A matrix of the k-th structure. Thus, S can be obtained from the following formula.

$$S = (w_1A_1 + w_2A_2 + ... + w_kA_k + ... + w_nA_n)^{-1}\, D_{exp} \tag{3.11}$$

The definition of our problem is thus to find the optimal relative populations of the structures within the ensemble such that the experimental RDC is best reproduced.

Iterative least squares fitting for optimal populations for a single RDC data set

In the process of back-calculating the residual dipolar coupling (RDC) from a protein structure or ensemble, singular value decomposition is used to obtain a least square solution for the alignment tensor. We apply the same technique iteratively to obtain the optimal relative populations for a given ensemble. Due to assuming linearity, the weights can be obtained via iterative least squares fitting. First, equal values are given for all populations and Equation 3.11 is used to obtain S. After S is obtained, it is used to determine the w_ks via least squares fitting. The process is iterated until the weights converge. In the end, each structure has either positive or zero population, since the weights are derived under the nonnegative constraints [Lawson and Hanson (1995)]. The following algorithm gives the detailed implementation of the iterative least squares fitting for a single RDC data set.

Iterative Least Squares Fitting $([A_1 A_2 \cdots A_n], D_{exp})$

for $i = 1$ to n **do**

$new\_weights(i) \leftarrow \frac{1}{n}$

end for

repeat

$old\_weights \leftarrow new\_weights$

$A \leftarrow old\_weights(1) * A_1 + \cdots + old\_weights(n) * A_n$

$S \leftarrow pseudo\_inverse(A) * D_{exp}$

$AS \leftarrow [A_1 S \; A_2 S \cdots A_n S]$

$new\_weights \leftarrow non\_negative\_least\_squares(AS, D_{exp})$

until *old_weights* and *new_weights* converge

return *new_weights*

Iterative least squares fitting for optimal relative populations for multiple RDC data sets

In the case of multiple RDC data sets, different alignment tensors are calculated for different media. The optimal weight combination (the relative populations) is obtained by least squares

fitting to all the RDC data sets. The following algorithm gives the detailed implementation of the iterative least squares fitting for multiple RDC data sets.

Iterative Least Squares Fitting Multiple RDCs $([A_1 A_2 \cdots A_n], [D_1, D_2 \cdots D_m])$

for $i = 1$ to n **do**

$new\_weights(i) \leftarrow \frac{1}{n}$

end for

repeat

$old\_weights \leftarrow new\_weights$

$A \leftarrow old\_weights(1) * A_1 + \cdots + old\_weights(n) * A_n$

for $i = 1$ to m **do**

$S(i) \leftarrow pseudo\_inverse(A) * D_i$

$AS(i) \leftarrow [A_1 S(i)\ A_2 S(i) \cdots A_n S(i)]$

end for

$$AS\_all \leftarrow \begin{pmatrix} AS(1) \\ AS(2) \\ . \\ . \\ . \\ AS(m) \end{pmatrix}$$

$$D\_all \leftarrow \begin{pmatrix} D_1 \\ D_2 \\ . \\ . \\ . \\ D_m \end{pmatrix}$$

$new\_weights \leftarrow non\_negative\_least\_squares(AS\_all, D\_all)$

until *old_weights* and *new_weights* converge

return *new_weights*

Table 3.5 The RDC data sets for obtaining the weights and validation.

Source	RDC Type	Number of RDC data sets
Lakomek et al. (2008)	NH	13
Lakomek et al. (2006)	NH	5
	NC'	4
	HC'	4
Ottiger and Bax (1998)	NH	2
	CaC'	2
	CaHa	2
	NC'	2
	HC'	2
Tolman (2002)	NH	9
Ruan and Tolman (2005)	NH	7
Kontaxis and Bax (2001)	Methyl	10

The ubiquitin ensemble and RDC data set for obtaining the weights and validation

Ubiquitin has long been used as a model protein to probe protein dynamics. In this work, we select 46 X-ray structures to form an X-ray structure ensemble for comparison with the ensembles obtained from MD, ENM, Concoord and tConcoord.

Conformational Sampling from MD

NAMD [Phillips et al. (2005)], a parallel molecular dynamics simulation program, is used to conduct the MD simulations. Periodic boundary conditions are applied in all simulation processes under the CHARMM27 force field. Each starting structure is solvated in a 10Å water box and each simulation starts with energy minimization and then equilibrium. One 50 ns MD simulation starting from PDB structure 1UBQ and 46 2-ns MD simulations are conducted using a HPC cluster computer.

Conformational Sampling using ENM

In ENM conformational sampling, we used all the atoms of protein in Hessian matrix calculations but only backbone torsional angles are considered free to rotate. The steps done to generate new structures from TNM (Torsional Network Model) are : (1) Torsional modes

from TNM are computed. (2) A random linear sum among the first ten low frequency modes weighted by their corresponding eigen frequencies are used to obtain torsional changes. (3) New structures are generated by rotating the original protein chain along the torsional changes obtained from step 2. (4) Newly generated structures are then energy minimized to remove any atom clashes.

Conformational Sampling using CONCOORD and tCONCOORD

46 CONCOORD and 46 tCONCOORD ensembles are generated from the program CONCOORD version 2.1 and tCONCOORD version 1.0 respectively. OPLS-AA parameters are used for VdW parameters and Engh-Huber parameters are used for the bonded parameters. RDC data set used A total of 62 RDC data sets, including NH, NC', HC', CαC', CαHα and side chain methyl, are used to obtained the weight combinations (i.e., relative populations) of the 6 selected ubiquitin ensembles. Table 3.5 shows the types and references of these 62 RDC data sets.

A total of 62 RDC data sets, including NH, NC', HC', CaC', CaHa and side chain methyl, are used to obtained the weight combinations (i.e., relative populations) of the 6 selected ubiquitin ensembles. Table 3.5 shows the types and references of these 62 RDC data sets.

The validation

Q factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is calculated by the following formula [Cornilescu et al. (1998)],

$$Q = \frac{\sqrt{\sum (D_{calc} - D_{exp})^2}}{\sqrt{\sum D_{calc}^2}} \tag{3.12}$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC. We also use the correlation coefficient to measure the agreement, which is calculated as the following,

$$\rho = \frac{(D_{calc} - \overline{D_{calc}})(D_{exp} - \overline{D_{exp}})}{\sqrt{\sum (D_{calc} - \overline{D_{calc}})^2 \sum (D_{exp} - \overline{D_{exp}})^2}} \tag{3.13}$$

Acknowledgements

Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged.

CHAPTER 4. GENERALIZED SPRING TENSOR MODELS FOR PROTEIN FLUCTUATION DYNAMICS AND CONFORMATION CHANGES

A paper published in the BMC Structural Biology

Tu-Liang Lin and Guang Song

Abstract

Background: In the last decade, various coarse-grained elastic network models have been developed to study the large-scale motions of proteins and protein complexes where computer simulations using detailed all-atom models are not feasible. Among these models, the Gaussian Network Model (GNM) and Anisotropic Network Model (ANM) have been widely used. Both models have strengths and limitations. GNM can predict the relative magnitudes of the fluctuations well, but due to its isotropic assumption, it can not be applied to predict the directions of the fluctuations. In contrast, ANM adds the ability to do the latter, but it loses a significant amount of precision in the prediction of the magnitudes. In this article, we develop a generalized spring tensor model (STeM) that is able to predict well both the magnitudes and the directions of the fluctuations.

Results: The new STeM is able to reproduce the mean square fluctuations for a set of 111 X-Ray structures with significantly better B factor correlations than ANM. The average correlation coefficient is 0.60 as compared to 0.53 by ANM and 0.59 by GNM. Despite the use of a more sophisticated potential, the performance of STeM is about the same as the performance of the GNM in experimental B factor prediction. However, the new model preserves the

anisotropic information, just like ANM, and greatly improves the magnitude prediction ability. Also the overlaps and correlations between the observed conformational changes and the most involved mode derived from STeM is about 4.5% improvement than the mode derived from ANM for a set of 20 pairs "open" and "closed" conformations. The frequency of the lowest mode identified as the most involved mode is also higher in STeM than in ANM.

Conclusions: STeM outperforms ANM in explaining protein conformation changes. All of these are accomplished without sacrificing the essential features that have made ANM and GNM attractive.

Introduction

It is now well accepted that the functions of a protein are closely related to not only its structure but also its dynamics. With the advancement of the computational power and increasing availability of computational resources, function-related protein dynamics, such as large-scale conformation transitions, has been probed by various computational methods at multiple scales. Among these computational methods, coarse-grained models play an important role since many functional processes take place over time scales that are well beyond the capacity of all-atom simulations [Voth (2009)]. One type of coarse-grained models, the elastic network models (ENMs), have been particularly successful and widely used in studying protein dynamics and in relating the intrinsic motions of a protein with its functional-related conformation changes over the last decade [Bahar et al. (1997); Atilgan et al. (2001); Bahar and Rader (2005); Ma (2005)].

The reason why ENMs have been well received as compared to the traditional normal mode analysis (NMA) lies at its simplicity to use. ENMs do not require energy minimization and therefore can be applied directly to crystal structure to compute the modes of motions. On the other, minimization is required for carrying out normal mode analysis (NMA). The problematic aspect of energy minimization is that it normally shifts the protein molecule away from its crystal conformation by about 2 Å. In addition, in ENMs analytic solutions to residue

fluctuations and motion correlations can be easily derived. Although several variations of ENMs exist, the ENMs basically have the potential similar to the following harmonic form:

$$V = \frac{\gamma}{2} \sum_{all\ qualified\ pairs\ (i,j)} (|\mathbf{r_{ij}}| - |\mathbf{r_{0,ij}}|)^2 \tag{4.1}$$

where γ is the force constant and $|\mathbf{r_{ij}}|$ and $|\mathbf{r_{0,ij}}|$ are the instantaneous and equilibrium distances respectively between residues i and j. All qualified pairs are the pairs that are within some distance cutoff. The simplicity of traditional ENMs has been regarded as an advantage because the second derivative can be obtained analytically. On the other hand, the simplicity leaves much room for improvement and many new models has been proposed [Ming and Brschweiler (2006); Song and Jernigan (2006, 2007); Lu et al. (2006); Yan (); Zheng (2008)].

The two most widely used ENM models are Gaussian Network Model (GNM) and Anisotropic Network Model (ANM). They have been used to predict the magnitude or direction of the residue fluctuations from a single structure and have been applied in many research areas, such as domain decomposition and allosteric communication [Lin and Song (2009); Bahar et al. (2007); Zheng and Brooks (2005); Yang et al. (2009); Bahar and Rader (2005); Zheng and Brooks (2005); Kundu et al. (2007); Tama and Sanejouand (2001)]. Both models have their own advantages and disadvantages. GNM can predict the relative magnitudes of the fluctuations well, but due to its isotropic assumption, it can not be applied to predict the directions of the fluctuations. In contrast, ANM adds the ability to do the latter, but it loses a significant amount of precision in the prediction of the magnitudes.

Gaussian Network Model. Gaussian Network Model (GNM) was first introduced in Bahar et al. (1997) under the assumption that the separation between a pair of residues in the folded protein is Gaussianly distributed. The model gives a good agreement between the theoretical and experimental crystallographic B-factors. The model describes a protein structure as a cluster of C_α atoms. The connectivity among the C_α's is expressed in Kirchhoff matrix Γ (see Eq. (4.2)). Two C_α's are considered to be in contact if their distance falls within a certain cutoff distance. The cutoff distance between a pair of residues is the only parameter in the model and is normally set to be 7 Å to 8 Å. Let $\mathbf{\Delta r_i}$ and $\mathbf{\Delta r_j}$ represent the instantaneous fluctuations from equilibrium positions of residue i and j and r_{ij} and $r_{0,ij}$ be the respective

instantaneous and equilibrium distances between residue i and j. The Kirchhoff matrix $\mathbf{\Gamma}$ is:

$$\Gamma_{ij} = \begin{cases} -1 & if \quad i \neq j \cap r_{0,ij} \leq r_c \\ 0 & if \quad i \neq j \cap r_{0,ij} > r_c \\ \sum_{j,j\neq i}^{N} \Gamma_{ij} & if \quad i = j \end{cases} \tag{4.2}$$

where i and j are the indices of the residues and r_c is the cutoff distance.

The simplicity of the Kirchhoff matrix formulation results from the assumption that the fluctuations of each residue are isotropic and Gaussian distributed along the X, Y and Z directions. The expected value of residue fluctuations, $< \mathbf{\Delta r_i}^2 >$, and correlation, $< \mathbf{\Delta r_i} \cdot \mathbf{\Delta r_j} >$, can be easily obtained from the inverse of the Kirchhoff matrix:

$$< \mathbf{\Delta r_i}^2 >= \frac{3k_BT}{\gamma}(\mathbf{\Gamma^{-1}})_{ii}, \tag{4.3}$$

$$< \mathbf{\Delta r_i} \cdot \mathbf{\Delta r_j} >= \frac{3k_BT}{\gamma}(\mathbf{\Gamma^{-1}})_{ij}, \tag{4.4}$$

where k_B is the Boltzmann constant and T is the temperature. The $< \mathbf{\Delta r_i}^2 >$ term is directly proportional to crystallographic B-factors.

Anisotropy Network Model. GNM provides only the magnitudes of residue fluctuations. To study the motions of proteins in full details, especially to determine the directions of the fluctuations, normal mode analysis (NMA) is needed. Traditional NMA is all-atom based and requires a structure to be first energy-minimized before the Hessian matrix and normal modes can be computed, which was rather cumbersome. Even after the energy minimization, the derivation of the Hessian matrix is not easy due to the complicated all-atom potential. In Tirion's pioneering work [Tirion (1996)] the energy minimization step was removed and a much simpler Hookean potential was used, and yet it was shown that the low frequency normal modes remained mostly accurate. Since then, the Hookean spring potentials have been used in coarse-grained C_α models [Hinsen (1998); Tama and Sanejouand (2001)] as well. Such models are best known as Anisotropy Network Model (ANM) [Atilgan et al. (2001)] since they have directional information of the fluctuations and the fluctuations are anisotropic. The potential in ANM has the simplest harmonic form similar to that is in Eq. 4.1. Assuming that a given

structure is at equilibrium, the Hessian matrix (3N×3N) can be derived analytically from such a potential [Atilgan et al. (2001)]. The 3N×3N Hessian matrix $\mathbf{H_{ANM}}$ can be repartitioned into N×N super elements and each super element is a 3×3 tensor.

$$\mathbf{H_{ANM}} = \begin{bmatrix} \mathbf{H_{1,1}} & \mathbf{H_{1,2}} & ... & \mathbf{H_{1,N}} \\ \mathbf{H_{2,1}} & \mathbf{H_{2,2}} & ... & \mathbf{H_{2,N}} \\ : & : & : & : \\ \mathbf{H_{N,1}} & \mathbf{H_{N,2}} & ... & \mathbf{H_{N,N}} \end{bmatrix} \tag{4.5}$$

where $\mathbf{H_{i,j}}$ is the interaction tensor between residue i and j and can be expressed as:

$$\mathbf{H_{i,j}} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{4.6}$$

Let $\mathbf{H}^+$ be the pseudo inverse of Hessian matrix $\mathbf{H_{ANM}}$. The mean square fluctuation $< \mathbf{\Delta r_i}^2 >$ and correlation can be calculated by summing the fluctuations over the $X-$, $Y-$ and $Z-$ directions.

$$< \mathbf{\Delta r_i}^2 >= \frac{3k_BT}{\gamma}(H^+_{3i-2,3i-2} + H^+_{3i-1,3i-1} + H^+_{3i,3i}) \tag{4.7}$$

$$< \mathbf{\Delta r_i} \cdot \mathbf{\Delta r_j} >= \frac{3k_BT}{\gamma}(H^+_{3i-2,3j-2} + H^+_{3i-1,3j-1} + H^+_{3i,3j}) \tag{4.8}$$

Strengths and Limitations of GNM and ANM. The advantages of ANM/GNM over the conventional NMA lie in several aspects: (i) it is a coarse-grained model and uses the C_α's to represent the residues in a structure; (ii) it does not require energy minimization and thus can be applied directly to crystal structures to compute the modes of motions; (iii) it provides analytic solutions to the mean square fluctuations and motion correlations.

The limitations of the GNM model. GNM provides only information of the magnitudes of residue fluctuations but no directional information. Therefore, the modes of GNM should not be interpreted as protein motions or components of the motions, since the potential of GNM is not rotational invariant [Thorpe (2007)].

The limitations of the ANM model. In contrast to GNM, ANM adopts the Hookean springs and the potential is simply a sum over Hookean potentials so the interaction potential is now

rotational invariant. And thus, the modes of ANM do represent the possible modes of protein motions. Yet, when applied to compute B-factors, ANM has a significantly poorer performance than GNM. The reason is that, in GNM, the fluctuations in the separation between a pair of residues are assumed to be Gaussian distributed and isotropic, while in ANM, because only a Hookean spring is attached between a pair of residue i and j, the fluctuation of residue j is only constrained longitudinally along the axis from i to j. The fluctuation is uncontrained transversely. The interaction spring tensor $\mathbf{H_{i,j}}$ between residue i and j in Eq. (4.6) becomes the following in the local frame (where the Z axis is along the direction from i to j):

$$\mathbf{H_{i,j}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.9}$$

Because the fluctuation of residue j is unconstrained transversely relative to residue i, the fluctuations given by ANM is less realistic than that by GNM. Such unrealistic-ness of ANM is an artifact due to the simplistic potential of ANM. In reality, the transverse fluctuation of residue j could have been constrained by i through bond bending interaction or torsional interaction, both of which are missing in ANM.

Our Contributions. To overcome the limitations of ANM and GNM, we develop a generalized spring tensor model for studying protein fluctuation dynamics and conformation changes. It is called generalized spring tensor model, or STeM, for the reason that the interaction between a pair of residue i and j is no longer a linear Hookean spring as is in Eq. (4.9), but takes a generalized tensor form that can provide proper transverse constraints on a residue's fluctuations relative to its neighbours. We obtain the generalized tensor form by deriving the Hessian matrix from a more physically realistic coarse-grained potential, the G$\bar{o}$-like potential [Clementi et al. (2000)], which has been successfully used in many MD simulations to study the protein folding processes and conformational changes [Clementi et al. (2000); Koga and Takada (2001, 2006)]. In additional to the Hookean spring interactions, the potential includes bond bending and torsional interactions, both of which had been found to be helpful in removing the "tip effect" of the ANM model [Lu et al. (2006)]. The inclusion of the bond bending and

torsional interactions is reflected in the generalized tensor spring interaction between i and j, in such a way that the tensor now includes not only the two-body interaction between i and j, but also three-body and four-body interactions that involve i and j.

The STeM model is able to integrate all the aforementioned attractive features of ANM and GNM and overcome their limitations. Results on predicting B-factors and conformation changes in Section 4 demonstrate that the STeM model is more accurate in predicting the directions of motions than ANM and in predicting the magnitudes of fluctuations than GNM. This is accomplished without incurring significantly more computational cost. STeM is naturally rotational-invariant since it is derived from a rotation-invariant potential function.

In addition, STeM has the following advantages that neither GNM nor ANM has. The potential in the STeM model includes bond bending and torsional interactions. Such rotational springs are desirable but are missing in most ENMs. In doing this, STeM naturally includes three-body and four-body interactions explicitly, which have been shown to be important, for example, in structure predictions [Feng et al. (2007)]. Most ENMs, on the other hand, only explicitly use two-body potential although the method of calculating collective modes may take account of the coupling between all nodes in the network and thus implicitly include the three and four body interactions. STeM is still more physically realistic by including bond bending and torsional rotations explicitly since they capture the chain behavior of protein molecules, which is neglected in most elastic network models where the protein is treated as an elastic rubber. Therefore, we have reasons to expect this model will further distinguish itself in studying protein dynamics where a correct modeling of bond bending or torsional rotations is crucial.

Results and discussion

Crystallographic B-factor Prediction

Table 1 shows the correlation coefficients between the experimental and calculated B factors of 111 proteins. The mean values of the correlation coefficients of ANM, GNM, and STeM are 0.53, 0.59, and 0.60 respectively. Hence, the STeM provides the directional information of the

Table 4.1 The correlation coefficient between the experimental and calculated B factors among different models.

Column R(Å) gives the resolution of each structure.

Protein	R(Å)	ANM	GNM	STeM	Protein	R(Å)	ANM	GNM	STeM	Protein	R(Å)	ANM	GNM	STeM
1AAC	1.31	0.7	0.71	0.76	1ADS	1.65	0.77	0.74	0.71	1AHC	2.00	0.79	0.68	0.61
1AKY	1.63	0.56	0.72	0.6	1AMM	1.20	0.56	0.72	0.55	1AMP	1.80	0.62	0.59	0.68
1ARB	1.20	0.78	0.76	0.83	1ARS	1.80	0.14	0.43	0.41	1ARU	1.60	0.7	0.78	0.79
1BKF	1.60	0.52	0.43	0.5	1BPI	1.09	0.43	0.56	0.57	1CDG	2.00	0.65	0.62	0.71
1CEM	1.65	0.51	0.63	0.76	1CNR	1.05	0.34	0.64	0.42	1CNV	1.65	0.69	0.62	0.68
1CPN	1.80	0.51	0.54	0.56	1CSH	1.65	0.44	0.41	0.57	1CTJ	1.10	0.47	0.39	0.62
1CUS	1.25	0.74	0.66	0.76	1DAD	1.60	0.28	0.5	0.42	1DDT	2.00	0.21	-0.01	0.49
1EDE	1.90	0.67	0.63	0.75	1EZM	1.50	0.56	0.6	0.58	1FNC	2.00	0.29	0.59	0.61
1FRD	1.70	0.54	0.83	0.77	1FUS	1.30	0.4	0.63	0.61	1FXD	1.70	0.58	0.56	0.7
1GIA	2.00	0.68	0.67	0.69	1GKY	2.00	0.36	0.55	0.44	1GOF	1.70	0.75	0.76	0.78
1GPR	1.90	0.65	0.62	0.66	1HFC	1.50	0.63	0.38	0.35	1IAB	1.79	0.36	0.42	0.53
1IAG	2.00	0.34	0.52	0.44	1IFC	1.19	0.61	0.67	0.53	1IGD	1.10	0.18	0.44	0.27
1IRO	1.10	0.82	0.51	0.85	1JBC	1.15	0.72	0.7	0.73	1KNB	1.70	0.63	0.66	0.54
1LAM	1.60	0.53	0.63	0.71	1LCT	2.00	0.52	0.57	0.61	1LIS	1.90	0.16	0.43	0.3
1LIT	1.55	0.65	0.62	0.76	1LST	1.80	0.39	0.72	0.73	1MJC	2.00	0.67	0.67	0.61
1MLA	1.50	0.59	0.57	0.54	1MRJ	1.60	0.66	0.49	0.5	1NAR	1.80	0.62	0.76	0.74
1NFP	1.60	0.23	0.48	0.41	1NIF	1.70	0.42	0.58	0.61	1NPK	1.80	0.53	0.55	0.64
1OMP	1.80	0.61	0.63	0.65	1ONC	1.70	0.55	0.7	0.58	1OSA	1.68	0.36	0.42	0.55
1OYC	2.00	0.78	0.73	0.77	1PBE	1.90	0.53	0.61	0.63	1PDA	1.76	0.6	0.76	0.58
1PHB	1.60	0.56	0.52	0.59	1PHP	1.65	0.59	0.63	0.65	1PII	2.00	0.19	0.44	0.28
1PLC	1.33	0.41	0.47	0.42	1POA	1.50	0.54	0.66	0.42	1POC	2.00	0.46	0.52	0.39
1PPN	1.60	0.61	0.64	0.67	1PTF	1.60	0.47	0.6	0.54	1PTX	1.30	0.65	0.51	0.62
1RA9	2.00	0.48	0.61	0.53	1RCF	1.40	0.59	0.63	0.58	1REC	1.90	0.34	0.5	0.49
1RIE	1.50	0.71	0.25	0.52	1RIS	2.00	0.25	0.24	0.47	1RRO	1.30	0.08	0.31	0.36
1SBP	1.70	0.69	0.72	0.67	1SMD	1.60	0.5	0.62	0.67	1SNC	1.65	0.68	0.71	0.72
1THG	1.80	0.5	0.53	0.5	1TML	1.80	0.64	0.64	0.58	1UBI	1.80	0.56	0.69	0.61
1WHI	1.50	0.12	0.33	0.38	1XIC	1.60	0.29	0.4	0.47	2AYH	1.60	0.63	0.73	0.82
2CBA	1.54	0.67	0.75	0.8	2CMD	1.87	0.68	0.6	0.62	2CPL	1.63	0.61	0.6	0.72
2CTC	1.40	0.63	0.67	0.75	2CY3	1.70	0.51	0.5	0.67	2END	1.45	0.63	0.71	0.68
2ERL	1.00	0.74	0.73	0.85	2HFT	1.69	0.63	0.79	0.72	2IHL	1.40	0.62	0.69	0.72
2MCM	1.50	0.78	0.83	0.79	2MHR	1.30	0.65	0.52	0.64	2MNR	1.90	0.46	0.5	0.47
2PHY	1.40	0.54	0.55	0.68	2RAN	1.89	0.43	0.4	0.31	2RHE	1.60	0.28	0.38	0.33
2RN2	1.48	0.68	0.71	0.75	2SIL	1.60	0.43	0.5	0.51	2TGI	1.80	0.69	0.71	0.73
3CHY	1.66	0.61	0.75	0.68	3COX	1.80	0.71	0.71	0.72	3EBX	1.40	0.22	0.58	0.4
3GRS	1.54	0.44	0.57	0.59	3LZM	1.70	0.6	0.52	0.66	3PTE	1.60	0.68	0.83	0.77
4FGF	1.60	0.41	0.27	0.43	4GCR	1.47	0.73	0.81	0.75	4MT2	2.00	0.42	0.37	0.46
5P21	1.35	0.4	0.51	0.45	7RSA	1.26	0.42	0.63	0.59	8ABP	1.49	0.61	0.82	0.62

residue fluctuations as ANM and has an accuracy even slightly better than GNM in B-factor prediction, but whether STeM is better than GNM in B factor prediction is questionable. Due to the small difference between the GNM and STeM result, the two models are comparable in B factor prediction if we consider the statistical error limits.

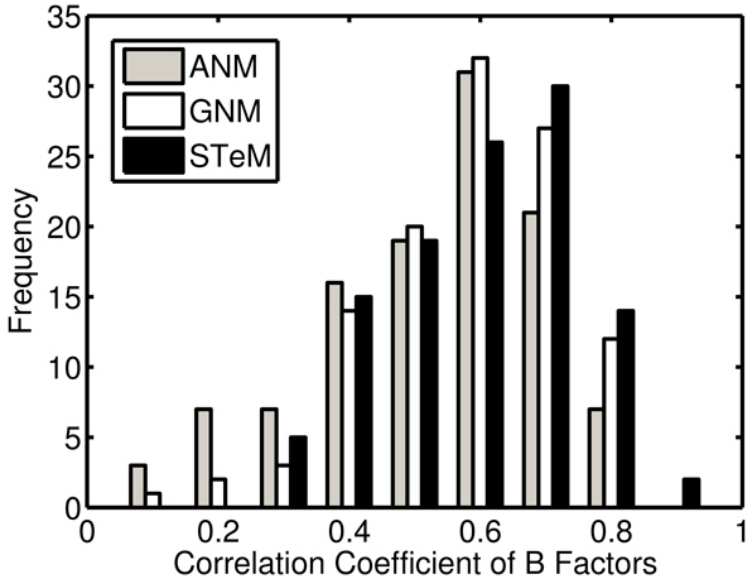


Figure 4.1 The distribution of the correlation coefficient between the experimental and calculated B factors.

Figure 4.1 shows the distribution of the correlation coefficients between the predicted B-factors and the experimental B-factors. STeM is the only model that there are instances where the correlation coefficient is above 0.85 and no instances where the correlation coefficient is below 0.25. This implies that the performance of STeM is more steady than either ANM or GNM. The scatter plot of the correlation coefficients between ANM and STeM in Figure 2 shows that STeM performs better than ANM for 80% of the proteins in the data set.

A particular interesting case is diphtheria toxin (1DDT) where previous research [Kundu et al. (2002)] indicated that the low correlation of the GNM (-0.01) was due to the lack of the crystal packing effect. When the effect of crystal neighbours was taken into account, the correlation increased to 0.6 [Kundu et al. (2002)]. Using the STeM model, which includes the effect of bond bending and torsional rotations, the correlation coefficient improves to 0.49

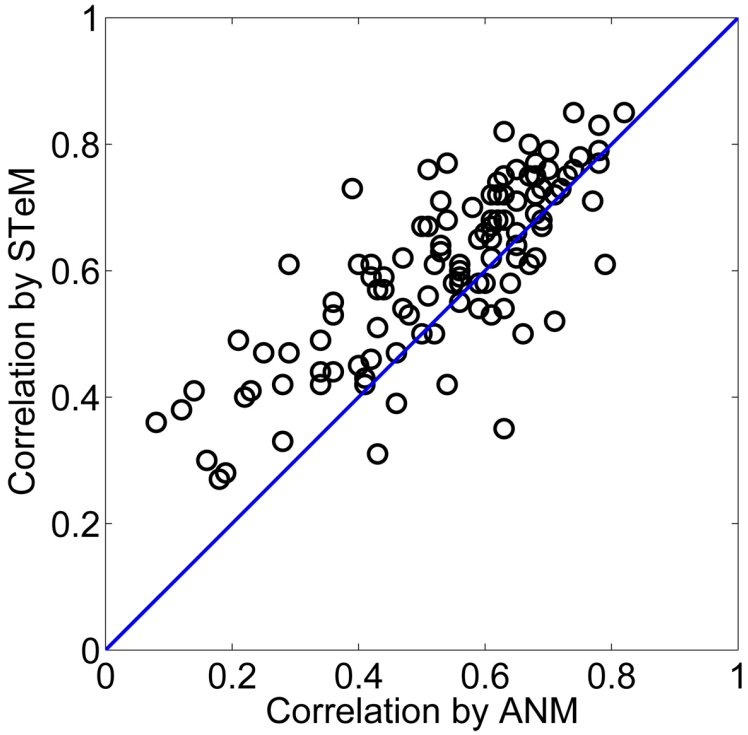


Figure 4.2 The scatter plot of the correlation coefficients from ANM and that from STeM.

For 80% of the proteins listed in Table 1, STeM does better.

without taking into account any crystal packing effect. Therefore, much of the discrepancy observed between experimental and calculated B-factors for this protein may have not been due to crystal packing, but in a larger degree, to a proper modeling of intramolecular interactions.

Protein structures of higher resolution have more accurate information of coordinates and B-factors. We investigate whether our model's performance can be further improved when the dataset used is limited to structures with higher resolution. We select the 12 structures with resolution better than 1.3 Å from the first dataset. The mean values of the correlation coefficients of these 12 structures are 0.56, 0.62, and 0.63 for ANM, GNM, and STeM, respectively, which give the same 3% increase for all three models.

Since the G$\bar{o}$-like potential has several terms contributing to it, including bond stretching, bending, dihedral rotations, and the non-bonded interactions, we also investigate the contribution of different terms to the agreement with experimental B factors. Bonded stretching term

Table 4.2 The contribution of different terms to the experimental B factor predictions.

$\mathbf{H_{ANM}}$ is the Hessian matrix from ANM. $\mathbf{H_{V1}}$, $\mathbf{H_{V2}}$, $\mathbf{H_{V3}}$ and $\mathbf{H_{V4}}$ are the Hessian matrices from bond stretching (V1), bond bending (V2), torsional rotation (V3) and non-local interaction (V4) terms, respectively.

Hessian matrices used	Correlation Coefficient with B factors	Improvement with respect to ANM
$\mathbf{H_{ANM}}$	0.53	0.00
$\mathbf{H_{V4}}$	0.55	0.02
$\mathbf{H_{V4} + H_{V1}}$	0.57	0.04
$\mathbf{H_{V4} + H_{V2}}$	0.57	0.04
$\mathbf{H_{V4} + H_{V3}}$	0.56	0.03
$\mathbf{H_{V4} + H_{V1} + H_{V2}}$	0.59	0.06
$\mathbf{H_{V4} + H_{V1} + H_{V3}}$	0.58	0.05
$\mathbf{H_{V4} + H_{V2} + H_{V3}}$	0.57	0.04
$\mathbf{H_{V4} + H_{V1} + H_{V2} + H_{V3}(= H_{STeM})}$	0.60	0.07
$\mathbf{H_{ANM} + H_{V1}}$	0.54	0.01
$\mathbf{H_{ANM} + H_{V2}}$	0.54	0.01
$\mathbf{H_{ANM} + H_{V3}}$	0.54	0.01
$\mathbf{H_{ANM} + H_{V1} + H_{V2} + H_{V3}}$	0.56	0.03

(V_1), bending term (V_2), or torsional rotation term (V_3) alone (see Methods section) is not enough to constraint some proteins in which the hessian matrix becomes close to singular and will have more than 6 zero eigenvalues when doing the eigen value decomposition for normal mode analysis. Only non-bonded interaction term (V_4) is able to provide enough constraints which can enforce the hessian matrix to have only 6 zero eigenvalues. Therefore, V_4 is the base term for the comparison of different terms to the agreement with experimental B factors with respect to ANM results. Table 2 shows the comparison of the contributions of these bonded and nonbonded interactions in resulting in an improvement of B factor predictions with respect to ANM results. Our results agree with the previous literatures which indicated that the non-bonded interaction plays a dominant role [Bahar et al. (1997)]. Table 2 shows that the contribution to the agreement with experimental B factors mainly comes from the non-bonded interaction and we found that the result even surpass the correlation coefficient obtained from ANM when non-boned interaction term is used alone.

It has been pointed out that the performance of B factor predictions can be improved by replacing the single force constant in ANM or GNM with a force constant depends on the inverse square of pairwise distance [Yan ()]. The Taylor expansion of the non-bonded interaction term (V4) obtained from the G$\bar{o}$-like potential has a force constant as $\frac{120\varepsilon}{r_{0,ij}^2}$ which depends on the inverse square of pairwise distance and this might explain why the correlation

Table 4.3 The overlap and correlation of the observed conformational change and the most involved mode among different models in open conformations.

Protein	Overlap in ANM	Correlation in ANM	Overlap in STeM	Correlation in STeM
Adenylate kinase	0.49(1)	0.62(1)	0.55(1)	0.63 (1)
Alcohol dehydrogenase	0.69(3)	0.54(9)	0.73 (2)	0.65 (30)
Annexin V	0.33(1)	0.60(32)	0.33 (1)	0.56 (22)
Aspartate aminotransferase	0.56(9)	0.63(9)	0.68 (6)	0.67 (6)
Calmodulin	0.44(5)	0.62 (77)	0.48 (1)	0.62 (16)
Che Y protein	0.46(1)	0.78(12)	0.40(1)	0.74(1)
Citrate synthase	0.48(7)	0.72(26)	0.49(5)	0.63(5)
Dihydrofolate reductase	0.71(1)	0.65(1)	0.73(1)	0.66(1)
Diphtheria toxin	0.43(1)	0.69(2)	0.50(2)	0.73(2)
Enolase	0.31(1)	0.45(34)	0.32(1)	0.49(53)
HIV-1 protease	0.67(1)	0.78 (10)	0.85 (1)	0.90(1)
Immunoglobulin	0.68(3)	0.57(3)	0.66(3)	0.58(3)
Lactoferrin	0.48(1)	0.64(24)	0.48(1)	0.70(36)
LAO binding protein	0.81(1)	0.74(1)	0.87(1)	0.80(1)
Maltodextrin binding protein	0.77(2)	0.66(2)	0.80(2)	0.70(2)
Seryl-tRNA synthetase	0.21(4)	0.59(10)	0.21(4)	0.60(37)
Thymidylate synthase	0.37(4)	0.69(9)	0.44(3)	0.68(9)
Triglyceride lipase	0.35(15)	0.50(25)	0.30(14)	0.56(24)
Triose phosphate isomerase	0.15(38)	0.28(11)	0.14(7)	0.30(8)
Tyrosine phosphatase	0.41(2)	0.57(27)	0.42(1)	0.59(25)

coefficient with experimental B factors from the non-bonded interaction alone can perform better than the correlation coefficient from ANM by 2%.

Although the contribution to the agreement with experimental B factors of the bonded interactions, including stretching, bending and dihedral rotations, are much smaller than non-bonded interactions, they still account for about 5% increase when considering all the three bonded terms. When we look at the individual contributions, we found that the bond stretching contributes about 2% increase, the bond bending also contributes 2% increase, and the torsinal rotation contributes about 1% increase.

We also change the base Hessian matrix to $\mathbf{H_{ANM}}$ and observe the contributions regarding to these three bonded terms. The correlation coefficient increases 3% when all the three bonded terms are added to the $\mathbf{H_{ANM}}$

Conformational Change Evaluation

It is known that the modes derived from the open form of a structure have better overlaps and correlations with the direction of a protein's conformation change than the ones derived from the closed form [Tama and Sanejouand (2001)]. Here we apply the STeM model to study

the conformation changes between the open and closed forms of 20 proteins and the open forms are used to calculate the normal modes. Table 3 lists the overlaps and correlations of the observed conformational changes and the indices of the modes that are most involved in the conformation changes. GNM is not considered since it does not provide directional information. The mean values of the overlaps and correlation coefficients of ANM are 0.49 and 0.61 respectively, and for STeM, 0.52 and 0.64 respectively. The performance increase gained by STeM is about 4.5% on both overlap and correlation. In both the overlap and correlation calculations, the modes that are most involved in the conformation change tend to have lower indices in STeM than in ANM, suggesting the modes of STeM may be of higher quality.

Conclusions

In this work, we develop a generalized spring tensor model for protein fluctuation dynamics and conformation changes. The new STeM model is able to reproduce the mean square fluctuations for a set of 111 X-ray structures with significantly better correlations with the experimental B factors than ANM by about 13%. The overlaps and correlations between the observed conformational changes and the most involved modes improve by 4.5% when using STeM over ANM for a set of 20 proteins. Therefore, STeM maintains the anisotropic information as is available in ANM, and improves the accuracy significantly in predicting the magnitudes of the fluctuations. Although the performance of STeM slightly surpasses the performance of GNM in experimental B factor predictions by 0.01, the difference is within the statistical error limits. Therefore, the two models are comparable in the experimental B factor predictions, but GNM is unable to provide the directions of the fluctuations.

Although the derivation of Hessian matrix in STeM is more complicated than the ANM or GNM, the computational complexity is almost the same as ANM. It means that with only slight increase in the computational time, STeM is able to provide much more accurate results than ANM.

The STeM is based on a physically sounder potential. The G$\bar{o}$-like potential has been used frequently in studying the protein folding processes and in other MD simulations [Clementi

et al. (2000); Koga and Takada (2001, 2006)]. The force parameters of the G$\bar{o}$-like potential, K_r, K_θ, K_ϕ and ε, are taken from the previous literature [Clementi et al. (2000)] without any tuning process. Therefore, it can be expected that these force parameters can be further tuned for different applications or within different context, such as under the crystal environment, and hence increase the predicting accuracy.

Other than a linear Hookean spring that is in ANM, the residue interactions in STeM take a generalized spring tensor form. It is foreseeable that other spring tensors can be used to model residue-residue interactions, for example, by deriving from other forms of potential functions, and consequently, many other variations of spring tensor models can be developed.

Chain breaking, such as that due to missing residues, has a more felt impact on STeM than on ANM or GNM, since the first, second, and third terms of the potential used to derive the model are all related to the continuity of the chain. We have not evaluated such impact in the current work but this could be a future research direction and our STeM model would be a proper tool for evaluating the impact of chain breaking on protein motions. As another part of future research, we will evaluate the importance of each term in the potential function and determine exactly how each term contributes to the protein motions. STeM does not always outperform ANM in B-factor predictions - it does better than ANM for 80% of the proteins studied. it would be interesting to find out why this is so. Crystal packing has been known to impact B-factor predictions. Therefore, a proper inclusion of crystal packing effects may further enhance STeM's performance. Since STeM takes into account bond bending and torsional interactions, it is expected that it will further distinguish itself in studying protein dynamics where a correct modeling of bond bending or torsional rotations is crucial, such as in predicting S^2 order parameters of NMR structures.

Methods

In this section we will show the derivations of the Hessian matrix from a G$\bar{o}$-like potential proposed by Clementi, Nymeyer and Onuchic [Clementi et al. (2000)] and we call the potential CNO (Clementi, Nymeyer and Onuchic) G$\bar{o}$-like potential.

The G$\bar{o}$-like potential

The CNO G$\bar{o}$-like potential [Clementi et al. (2000)] takes the non-native and native (equilibrium) conformations as input and it can be divided into four terms. The first term of this G$\bar{o}$-like potential (defined as V_1 for later use) preserves the chain connectivity. The second (V_2) and third terms (V_3) define the bond angle and torsional interactions respectively and the last term (V_4) is nonlocal interactions. The G$\bar{o}$-like potential has the following expression.

$$\begin{aligned} V(X, X_0) &= \sum_{bonds} V_1(r, r_0) + \sum_{angles} V_2(\theta, \theta_0) \\ &\quad + \sum_{dihedral} V_3(\phi, \phi_0) + \sum_{i<j-3} V_4(r_{ij}, r_{0,ij}) \\ &= \sum_{bonds} K_r(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \\ &\quad + \sum_{dihedral} \{K_\phi^{(1)}[1 - cos(\phi - \phi_0)] \\ &\quad + K_\phi^{(3)}[1 - cos3(\phi - \phi_0)]\} \\ &\quad + \sum_{i<j-3} \varepsilon[5(\frac{r_{0,ij}}{r_{ij}})^{12} - 6(\frac{r_{0,ij}}{r_{ij}})^{10}] \end{aligned} \tag{4.10}$$

In equation (4.10), r and r_0 represent respectively the instantaneous and equilibrium distances of the virtual bond formed by the C_α's of two consecutive residues. Similarly, the θ (θ_0) and ϕ (ϕ_0) are respectively the instantaneous (equilibrium) virtual bond angles formed by three consecutive residues and virtual dihedral angles formed by four consecutive residues. The r_{ij} and $r_{0,ij}$ represent respectively the instantaneous and equilibrium distances between two non-consecutive residue i and j. This G$\bar{o}$-like potential is physically more accurate than the Hookean potential that is used in ANM.

The G$\bar{o}$-like potential (equation 4.10) includes several force parameters (K_r, K_θ, $K_\phi^{(1)}$, $K_\phi^{(3)}$ and ε) and the value of these force parameters are taken directly from the previous literature [Clementi et al. (2000)] without any tuning process. The values of these parameters

are $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\phi^{(1)} = \varepsilon$, $K_\phi^{(3)} = 0.5\varepsilon$ and $\varepsilon = 0.36$.

Anisotropic fluctuations from the second derivative of the G$\bar{o}$-like potential

Similar to ANM, STeM has a 3N×3N Hessian matrix and the Hessian matrix can be decomposed into N×N super-elements (equation 4.6). Each super-element in STeM is a summation of four 3×3 matrices. The first 3×3 matrix is the contribution from bond stretching. The second and third 3×3 matrices are the contributions from bond bending and torsional rotations respectively. The fourth 3×3 matrix is the contribution from nonlocal contacts.

$$
\begin{aligned}
\mathbf{H_{i,j}} \quad = \quad & \begin{bmatrix} \frac{\partial^2 V_1(r,r_0)}{\partial X_i \partial X_j} & \frac{\partial^2 V_1(r,r_0)}{\partial X_i \partial Y_j} & \frac{\partial^2 V_1(r,r_0)}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_1(r,r_0)}{\partial Y_i \partial X_j} & \frac{\partial^2 V_1(r,r_0)}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_1(r,r_0)}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V_1(r,r_0)}{\partial Z_i \partial X_j} & \frac{\partial^2 V_1(r,r_0)}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_1(r,r_0)}{\partial Z_i \partial Z_j} \end{bmatrix} + \\
& \begin{bmatrix} \frac{\partial^2 V_2(\theta,\theta_0)}{\partial X_i \partial X_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial X_i \partial Y_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Y_i \partial X_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Z_i \partial X_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_2(\theta,\theta_0)}{\partial Z_i \partial Z_j} \end{bmatrix} + \\
& \begin{bmatrix} \frac{\partial^2 V_3(\phi,\phi_0)}{\partial X_i \partial X_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial X_i \partial Y_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Y_i \partial X_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Z_i \partial X_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_3(\phi,\phi_0)}{\partial Z_i \partial Z_j} \end{bmatrix} + \\
& \begin{bmatrix} \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial X_i \partial X_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial X_i \partial Y_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Y_i \partial X_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Z_i \partial X_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_4(r_{ij},r_{0,ij})}{\partial Z_i \partial Z_j} \end{bmatrix}
\end{aligned}
\tag{4.11}
$$

The Hessian matrix is derived from the second derivative of the overall potential (equation 4.10). Let us first consider the first term of the G$\bar{o}$-like potential and let (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j) be the Cartesian coordinates of two consecutive residue i and j.

$$
\begin{aligned}
V_1(r, r_0) &= K_r(r - r_0)^2 \\
&= K_r\{[(X_j - X_i)^2 + (Y_j - Y_i)^2 +
\end{aligned}
$$

$$(Z_j - Z_i)^2]^{1/2} - r_0\}^2 \tag{4.12}$$

The first and second partial derivatives of V_1 with respect to the X-direction of residue i are

$$\frac{\partial V_1}{\partial X_i} = -2K_r(X_j - X_i)(1 - r^0/r) \tag{4.13}$$

$$\frac{\partial^2 V_1}{\partial X_i^2} = 2K_r(1 + r^0(X_j - X_i)^2/r^3 - r^0/r) \tag{4.14}$$

We will get similar results for the Y- and Z-directions of residue i. Since we focus only on the equilibrium fluctuations, we can have $r \cong r^0$ at equilibrium and the first and second partial derivatives of V_1 can be further simplified to the following expressions.

$$\frac{\partial V_1}{\partial X_i} = 0 \tag{4.15}$$

$$\frac{\partial^2 V_1}{\partial X_i^2} = 2K_r(X_j - X_i)^2/r^2 \tag{4.16}$$

In a similar way, the second cross-derivatives have the following form:

$$\frac{\partial^2 V_1}{\partial X_i \partial Y_j} = -2K_r(X_j - X_i)(Y_j - Y_i)/r^2 \tag{4.17}$$

Equations 4.16 and 4.17 give the elements of the first 3×3 matrix for the super element $\mathbf{H_{ij}}$ in equation 4.11. For the diagonal super elements $\mathbf{H_{ii}}$, equations 4.16 and 4.17 are substituted by the following:

$$\frac{\partial^2 V_1}{\partial X_i^2} = -\sum_{j=i-1,i+1} 2K_r(X_j - X_i)^2/r^2 \tag{4.18}$$

$$\frac{\partial^2 V_1}{\partial X_i \partial Y_i} = \sum_{j=i-1,i+1} 2K_r(X_j - X_i)(Y_j - Y_i)/r^2 \tag{4.19}$$

Now let's consider the second term of the G$\bar{o}$-like potential and let (X_i, Y_i, Z_i), (X_j, Y_j, Z_j) and (X_k, Y_k, Z_k) be the Cartesian coordinates of three consecutive residue i, j and k. Suppose

θ is the virtual bond angle formed by these three consecutive residues. The second term of the CNO G$\overline{o}$-like potential is $V_2 = K_\theta(\theta - \theta_0)^2$ and the first and second partial derivative of V_2 are

$$\frac{\partial V_2}{\partial X_i} = 2K_\theta(\theta - \theta_0)\frac{\partial \theta}{\partial X_i} \tag{4.20}$$

$$\frac{\partial^2 V_2}{\partial {X_i}^2} = 2K_\theta(\frac{\partial \theta}{\partial X_i})^2 + 2K_\theta(\theta - \theta_0)\frac{\partial^2 \theta}{\partial {X_i}^2} \tag{4.21}$$

Since θ equals θ_0 at equlibrium, $\frac{\partial^2 V_2}{\partial {X_i}^2}$ can be further simplified as

$$\frac{\partial^2 V_2}{\partial {X_i}^2} = 2K_\theta(\frac{\partial \theta}{\partial X_i})^2 \tag{4.22}$$

Likewise, $\frac{\partial^2 V_2}{\partial X_i \partial X_j}$ becomes

$$\frac{\partial^2 V_2}{\partial X_i \partial X_j} = 2K_\theta(\frac{\partial \theta}{\partial X_i})(\frac{\partial \theta}{\partial X_j}) \tag{4.23}$$

Let $\mathbf{p} = (X_i - X_j, Y_i - Y_j, Z_i - Z_j)$ and $\mathbf{q} = (X_k - X_j, Y_k - Y_j, Z_k - Z_j)$. We define G as the following.

$$G = \frac{(\mathbf{p}.\mathbf{q})}{|\mathbf{p}||\mathbf{q}|} \tag{4.24}$$

The θ can be expressed as

$$\theta = cos^{-1}(\frac{(\mathbf{p}.\mathbf{q})}{|\mathbf{p}||\mathbf{q}|}) = cos^{-1}(G) \tag{4.25}$$

The partial derivatives of θ are

$$\frac{\partial \theta}{\partial X_i} = \frac{-1}{\sqrt{1 - G^2}}\frac{\partial G}{\partial X_i} \tag{4.26}$$

$$\frac{\partial \theta}{\partial X_j} = \frac{-1}{\sqrt{1 - G^2}}\frac{\partial G}{\partial X_j} \tag{4.27}$$

$$\frac{\partial \theta}{\partial X_k} = \frac{-1}{\sqrt{1 - G^2}}\frac{\partial G}{\partial X_k} \tag{4.28}$$

The derivative of G is

$$\frac{\partial G}{\partial X_i} = \frac{\partial}{\partial X_i}\frac{(\mathbf{p}.\mathbf{q})}{|\mathbf{p}||\mathbf{q}|} = \frac{(X_k - X_j)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{q}|}{|\mathbf{p}|}(X_i - X_j)}{(|\mathbf{p}||\mathbf{q}|)^2} \tag{4.29}$$

We can also get $\frac{\partial G}{\partial X_j}$ and $\frac{\partial G}{\partial X_k}$.

$$\begin{aligned}\frac{\partial G}{\partial X_j} \quad = \quad & \frac{(2X_j - X_i - X_k)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{q}|}{|\mathbf{p}|}(X_j - X_i)}{(|\mathbf{p}||\mathbf{q}|)^2} \\ & - \frac{(\mathbf{p}.\mathbf{q})\frac{|\mathbf{p}|}{|\mathbf{q}|}(X_j - X_k)}{(|\mathbf{p}||\mathbf{q}|)^2}\end{aligned} \tag{4.30}$$

$$\frac{\partial G}{\partial X_k} = \frac{(X_i - X_j)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{p}|}{|\mathbf{q}|}(X_k - X_j)}{(|\mathbf{p}||\mathbf{q}|)^2} \tag{4.31}$$

Combined eq (4.22),(4.26) and (4.29), we can get the following formula.

$$\frac{\partial^2 V_2}{\partial {X_i}^2} = \frac{2K_\theta}{1-G^2}(\frac{(X_k - X_j)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{q}|}{|\mathbf{p}|}(X_i - X_j)}{(|\mathbf{p}||\mathbf{q}|)^2})^2 \tag{4.32}$$

Similarly, Combined eq (4.23),(4.26), (4.27), (4.29) and (4.30), the second cross-derivative $\frac{\partial^2 V_2}{\partial X_i X_j}$ becomes

$$\begin{aligned}\frac{\partial^2 V_2}{\partial X_i X_j} &= \frac{2K_\theta}{1-G^2}(\frac{(X_k - X_j)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{q}|}{|\mathbf{p}|}(X_i - X_j)}{(|\mathbf{p}||\mathbf{q}|)^2}) \\ &\quad (\frac{(2X_j - X_i - X_k)|\mathbf{p}||\mathbf{q}| - (\mathbf{p}.\mathbf{q})\frac{|\mathbf{q}|}{|\mathbf{p}|}(X_j - X_i)}{(|\mathbf{p}||\mathbf{q}|)^2} \\ &\quad -\frac{(\mathbf{p}.\mathbf{q})\frac{|\mathbf{p}|}{|\mathbf{q}|}(X_j - X_k)}{(|\mathbf{p}||\mathbf{q}|)^2})\end{aligned} \tag{4.33}$$

Following the same approach, we are able to get $\frac{\partial^2 V_2}{\partial X_j X_k}$ and $\frac{\partial^2 V_2}{\partial X_k X_i}$ and these second cross-derivatives form the elements of the second 3×3 matrix of the super element $\mathbf{H_{ij}}$ in equation 4.11.

Due to the complexity of the derivation process of the third (dihedral angle) term of CNO G$\bar{o}$-like potential, we omit the derivation process here. Readers can refer to the appendix for further details.

Finally, let's consider the final (non-local contact) term.

$$V_4 = \varepsilon[5(\frac{r_{0,ij}}{r_{ij}})^{12} - 6(\frac{r_{0,ij}}{r_{ij}})^{10}] \tag{4.34}$$

A talor expansion will give us the following form.

$$V_4 = -\varepsilon + \frac{120\varepsilon}{r_{0,ij}^2}(r_{ij} - r_{0,ij})^2 \tag{4.35}$$

Equation 4.35 has the same harmonic form as the first term but with a different force constant, so the derivation process is the same as the first term. Therefore, we only give the derivation result here.

$$\frac{\partial^2 V_4}{\partial X_i \partial Y_j} = -\frac{240\varepsilon}{r_{0,ij}^2}(X_j - X_i)(Y_j - Y_i)/r_{ij}^2 \tag{4.36}$$

After combining the Hessian matrix from all four terms, we can calculate the pseudo inverse of the final Hessian matrix $\mathbf{H}$. The mean square displacement $< \mathbf{\Delta r_i}^2 >$ and inter residue correlation $< \mathbf{\Delta r_i} \cdot \mathbf{\Delta r_j} >$ can be calculated by summing the elements over the $X-$, $Y-$ and $Z-$ directions, as is in ANM.

$$< \mathbf{\Delta r_i}^2 >= \frac{k_B T}{\gamma}(H^+_{3i-2,3i-2} + H^+_{3i-1,3i-1} + H^+_{3i,3i}) \tag{4.37}$$

$$< \mathbf{\Delta r_i} \cdot \mathbf{\Delta r_j} >= \frac{k_B T}{\gamma}(H^+_{3i-2,3j-2} + H^+_{3i-1,3j-1} + H^+_{3i,3j}) \tag{4.38}$$

The Protein Sets Studied

To evaluate the STeM model, we apply it to compute thermal B-factors and to study protein conformation changes and compare the results with those computed from ANM and GNM. For B-factors computation, the protein data set is from Kundu et al. (2002) that contains 111 proteins. Two proteins, 1CYO and 5PTP, are removed from the data set because they no longer exist in the current Protein Data Bank [Berman et al. (2000)]. The proteins in the first dataset all have a resolution that is better than 2.0 Å. For conformation change studies, the data set is from Tama and Sanejouand (2001), which contains 20 pairs of protein structures. Each pair of protein structures have significantly large structure difference from each other.

Evaluation Techniques

We used the same evaluation techniques as have been applied before [Kundu et al. (2002); Tama and Sanejouand (2001)] to evaluate the STeM model. The following three numerical values are computed.

The correlation between the experimental and calculated B factors

The linear correlation coefficient between the experimental and calculated B factors are calculated according to the following formula.

$$\rho = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{[\sum_i^N (x_i - \overline{x})^2 \sum_i^N (y_i - \overline{y})^2]^{1/2}} \tag{4.39}$$

where x_i and y_i are respectively the experimental and calculated B factors of the C_α atom of the residue i and $\overline{x}$ and $\overline{y}$ are the mean values of the experimental and calculated B factors. N is the number of residues.

The overlap between the experimental observed conformational changes and the calculated modes

The overlap measures the directional similarity between the conformational changes and a calculated mode. The formula for calculating the overlap is

$$I = \frac{|\sum_i^{3N} e_i r_i|}{[\sum_i^{3N} e_i^2 \sum_i^{3N} r_i^2]^{1/2}} \tag{4.40}$$

where e_i is the coordinate of residue i of a selected mode and r_i is the conformational displacement of residue i.

The correlation between the experimental observed conformational changes and the calculated modes

The correlation measures the magnitude similarity between the conformational changes and a calculated mode. The formula used for calculating the correlation is the same as equation (4.39), with different meaning for x_i and y_i.

$$\rho = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{[\sum_i^N (x_i - \overline{x})^2 \sum_i^N (y_i - \overline{y})^2]^{1/2}} \tag{4.41}$$

where x_i is the amplitude of the displacement of residue i in the conformational change and y_i is the amplitude of the displacement of residue i in a calculated mode. $\overline{x}$ and $\overline{y}$ are the corresponding mean values.

List of abbreviations used

ENM: Elastic Network Model

GNM: Gaussian Network Model

ANM: Anisotropic Network Model

STeM: Spring Tensor Model

Competing interests

The authors declare that they have no competing interests.

Authors contributions

Tu-Liang Lin collaborated with Guang Song on the ideas of the Spring Tensor Model and applied to the prediction of the B factors and conformation change. Gunag Song conceived the idea of using more sophisticated potential and suggested the G$\overline{o}$-like potential. Most of the implementation was done by Tu-Liang Lin under the supervision of Guang Song. Both authors read and approved the manuscript.

CHAPTER 5. EFFICIENT MAPPING OF LIGAND MIGRATION CHANNEL NETWORKS IN DYNAMIC PROTEINS

A paper accepted by Proteins: Structure, Function, and Bioinformatics

Tu-Liang Lin and Guang Song

Abstract

For many proteins such as myoglobin, the binding site lies in the interior and there is no obvious route from the exterior to the binding site in the average structure. Although computer simulations for a limited number of proteins have found some transiently-open channels, it is not clear if there exist more channels elsewhere or how the channels are regulated. A systematic approach that can map out the whole ligand migration channel network is lacking. Ligand migration in a dynamic protein resembles closely a well-studied problem in robotics, namely, the navigation of a mobile robot in a dynamic environment. In this work, we present a novel robotic motion planning inspired approach that can map the ligand migration channel network in a dynamic protein. The method combines an efficient spatial mapping of protein inner space with a temporal exploration of protein structural heterogeneity, which is represented by a structure ensemble. The spatial mapping of each conformation in the ensemble produces a partial map of protein inner cavities and their inter-connectivity. These maps are then merged to form a super map that contains all the channels that open dynamically. Results on the pathways in myoglobin for gaseous ligands demonstrate the efficiency of our approach in mapping the ligand migration channel networks. The results, obtained in significantly less time than trajectory-based approaches, are in agreement with previous simulation results. In

addition, the method clearly illustrates how and what conformational changes open or close a channel.

Introduction

Proteins are one of the fundamental functional units of living systems. The desire to understand their functional mechanisms drives the progress in biological sciences and medicine. Many proteins function through the binding of a small ligand. Examples are substrate and/or cofactor binding in enzyme catalysis. For many proteins such as myoglobin, the binding sites lie in the interior and there is no obvious route connecting the binding site to solvent in the average structure. Although protein dynamics is known to open pathways dynamically, exactly how the dynamics exerts its control is not fully known. Our knowledge of how a ligand may enter or exit the binding site has advanced significantly thanks to the developments in mutagenesis studies, X-ray crystallography, time-resolved Laue X-ray diffraction, and molecular dynamics (MD) simulations. Yet, such knowledge remains incomplete even for the most studied proteins such as myoglobin. For example, though some pathways are revealed by the aforementioned methods, it is not clear if there exist alternative pathways or how these pathways are regulated. Moreover, it remains a challenge to extend the time-resolved Laue X-ray diffraction studies or MD simulations broadly to the study of ligand migration process in other proteins, due to the difficulties in experimental setup or limitations in computation. A method that can efficiently map the whole ligand migration network and elucidate how the protein dynamics regulates the channels is lacking.

Ligand migration in a dynamic protein resembles closely a problem that has been well studied in robotics, namely, the navigation of a mobile robot in a dynamic environment. Inspired by this observation, we develop an innovative motion planning based approach that can efficiently map the complete ligand migration channel network in a dynamic protein. The approach overcomes the computational barrier faced by existing methods by integrating an efficient geometric mapping component with the dynamic exploration of a protein's structure flexibility. By taking as input a structure ensemble of a given protein, which may be composed

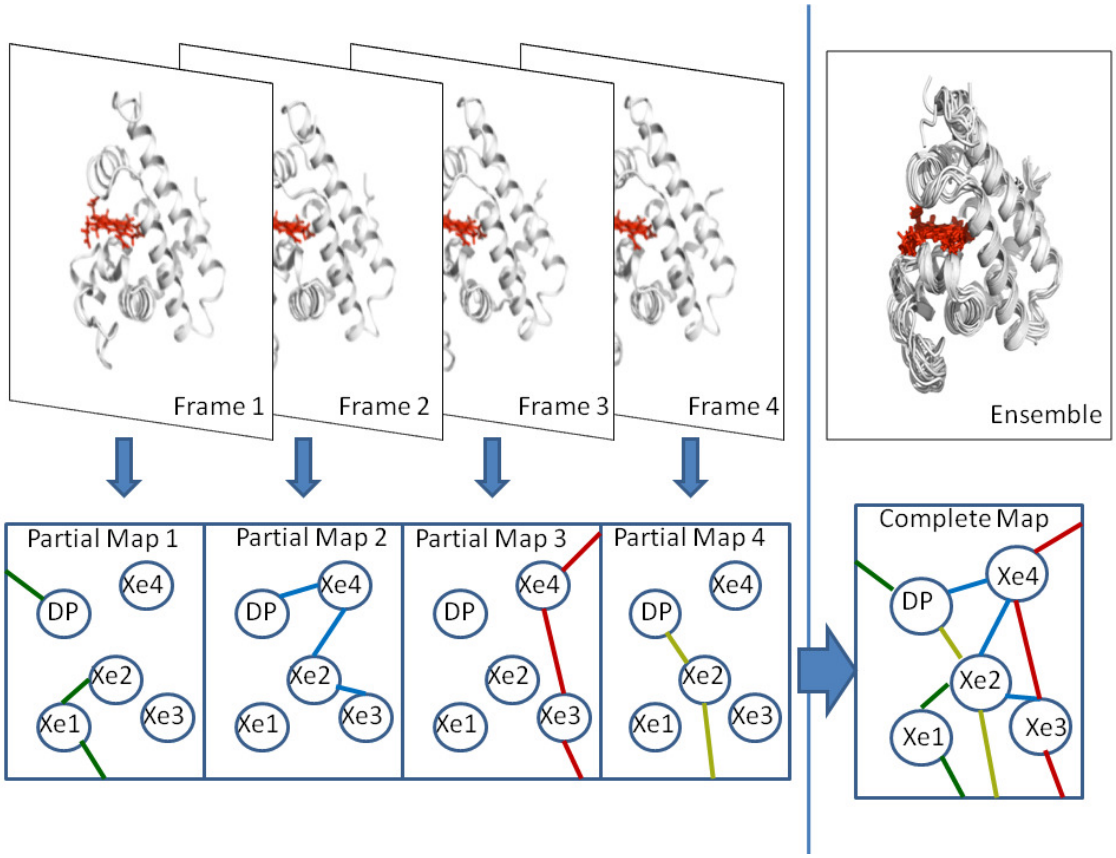


Figure 5.1 Overview of the method.

Efficient geometric mapping is applied to each conformation in the ensemble to obtain a partial map of how the internal cavities (shown here are the four xenon binding sites and a distal pocket (DP) of myoglobin) are connected to each other and to the solvent. These partial maps are then merged to form the entire dynamic migration network, which shows how the cavities can be accessed dynamically.

of existing experimental structures and/or conformations generated from MD simulations, the approach carries out an efficient spatial mapping of the protein's inner space at each conformation in the ensemble (see Figure 5.1). The spatial mapping reveals the partial connectivity among the cavities and the solvent. All partial maps are then merged to form a super-graph that represents the whole migration channel network that is accessible to the ligand, both *spatially* and *dynamically*. The super-graph can then be used to identify all the migration pathways that exist in the ensemble. This is much more efficient than most other methods that identify only one pathway at a time. The quality of the structure ensemble is critical

to the success of this method. A good ensemble needs to represent sufficiently the structural heterogeneity of the host protein. Our method is flexible and a structure ensemble can be constructed using crystal structures of the same protein, or NMR ensembles, or conformations generated from MD simulations, or a combination of all above. The benefit of having such flexibility is that it facilitates the integration of existing structure and dynamics information about a given protein. Dynamic conformational ensembles have also been found useful in understanding molecular recognition [Boehr et al. (2009)].

We apply this method to myoglobin and cytochrome P450cam and the results demonstrate the feasibility of our approach in efficiently mapping the ligand migration channel network and in identifying migration pathways that are in agreement with previous experimental and simulation results. Moreover, the method clearly illustrates how and what conformational changes open or close a channel. Such close association between the variations of the channel clearances (the effect) and conformational changes (the cause) may provide the needed information to decipher the regulation mechanism.

Myoglobin(Mb) is a single chain globular protein rich in muscle tissues and is responsible for oxygen storage. The backbone of Mb forms eight alpha-helices wrapped around a heme and the heme iron can bind with small gaseous ligands such as Xenon, O_2, CO, or NO. A number of experimental and computational methods have been developed and applied to study the dynamics of ligand migration and binding in myoglobin [Frauenfelder et al. (2009)]. Here we highlight some of the results from mutagenesis, X-ray crystallography, time-resolved Laue X-ray crystallography, and molecular dynamics. Related work on protein geometry calculations and robotics-inspired methods are also reviewed.

Flash photolysis and Mutagenesis. The pioneering work of Austin et al. (1975) introduced the flash photolysis technique that is still widely used today in studying the recombination kinetics in heme proteins. The geminate recombination, or the re-binding of the ligand before it exits to the solvent, reveals the heterogeneity of the conformational substates and the existence of several binding pockets. The existence of these binding pockets was further confirmed by xenon binding sites [Tilton et al. (1984)]. Ligands were also found to reside in

the distal pocket near the binding site for some short time period after photo-dissociation at low temperature [Teng et al. (1997)]. In addition to the four xenon (Xe) sites and the distal pocket, another two cavities where ligands may temporarily reside were identified in a 90-ns MD simulation [Bossa et al. (2004)]. Recently, some cavities that are smaller than the xenon sites were suggested to be located at the branching points of the migration channels by another MD simulation [Ruscio et al. (2008)]. Xenon gas was also used as the perturbing agent to migration pathways [Tetreau et al. (2004)]. The effects of xenon on ligand binding, with 25 mutants, were studied Scott and Gibson (1997). The residues that surround these cavities were found to be highly conserved [Frauenfelder et al. (2001)], which indicates that the cavities should have functional roles. Site-directed mutagenesis of 27 residues was used to map out the ligand pathways [Scott et al. (2001)]. Random mutagenesis studies conducted by Huang and Boxer (1994) showed that single mutations of several other clusters of residues far away from the pathways were discovered to profoundly affect the ligand-binding kinetics.

X-ray Crystallography. The abundance of crystal structures (over 200 for myoglobin for example) in the PDB [Berman et al. (2000)] presents a valuable structural basis for studying protein inner cavities and their role in the rebinding kinetics. Especially worth mentioning are the high-resolution near-atomic resolution crystal structures, many of which capture structure substates [Kachalova et al. (1999); Vojtechovsky et al. (1999)]. Studies of the conversion among substates clearly demonstrated that dynamics plays a key role in the realization of protein functions [Frauenfelder et al. (2001); Teeter (2004)]. Base on X-ray crystallography, structure changes observed before and after photolysis confirmed that Xe_1, the proximal cavity site, is the secondary ligand-docking site [Chu et al. (2000); Nienhaus et al. (2005); Ostermann et al. (2000)]. Such transition from the distal site to the proximal binding site was used to infer significant thermal fluctuations that are necessary to open the channel between the two sites [Nienhaus et al. (2005); Ostermann et al. (2000)]. On the other hand, none of the crystal structures alone presents a complete picture of how a ligand may migrate inside the protein matrix. It has been widely believed that one [Scott et al. (2001)] or several dynamic pathways [Cohen et al. (2006)] can be created via the thermal fluctuations and that such

dynamic pathways cannot be identified from a single static structure.

Time-resolved X-ray Crystallography. A great leap in the understanding of the migration process was made possible by the advances in time-resolved Laue diffraction studies, thanks to the pioneering work of Moffat and co-workers [Srajer et al. (2001, 1996)]. The technique is a perfect tool to study light-sensitive protein dynamics, such as that of heme proteins [Bourgeois et al. (2007)]. It literally allows one to trace a photo-disassociated ligand as it migrates through the protein, as well as structure relaxation, over a broad range of timescales, from a few nanoseconds to as long as a few milliseconds [Schotte et al. (2003); Bourgeois et al. (2006, 2003); Schmidt et al. (2005)]. It has provided direct insight about how the correlated motions of the backbone and side chains provide a gating mechanism for ligand migration [Schotte et al. (2003)]. Particularly, Srajer et al. (2001) observed the ligand appearing at two locations, the distal pocket and the Xe_1 binding site. Schotte et al. (2003) observed the ligand translocation in the L29F mutant MbCO between the binding site and the Xe_1 site and found that the Xe_4 site was probably the intermediate stop site. Bourgeois et al. (2006) reported the sub-nanosecond time-resolved Laue X-ray diffraction results on the triple mutant YQR-Mb and observed the photolyzed CO move to the Xe_4 site before reaching the Xe_1 site. All of these Laue X-ray diffractions showed only the Xe_1 site was occupied by the photolyzed ligand at around 100-ns and, it was suggested that an intermediate stop at Xe_4 site might take place during the ligand's migration between the binding site and the Xe_1 site.

Molecular dynamics. MD was first used as early as the late 70's by Case and Karplus (1979) to study the dynamics of ligand binding in myoglobin, followed by work such as Elber and Karplus (1990), and is still widely used today. Nutt and Meuwly (2004) applied classical MD and QM-MD to capture some ligand pathways and were able to reproduce the infrared spectrum data. Hummer et al. (2004) did both MD simulations and time-resolved X-ray, and the results from time-resolved X-ray validated the MD trajectories. Similarly, Anselmi et al. (2008) and Bossa et al. (2004, 2005)'s MD simulations reproduced CO diffusion and kinetics that agreed with experimental data, especially that of the time-resolved Laue X-ray diffractions. Perhaps the most extensive MD results were obtained by Ruscio et al. (2008), whose cumulative

7-μs simulations were used to identify many different trajectories and entry/exit portals on the protein surface. Besides the aforementioned work in MD, Cohen et al. (2006) proposed an implicit ligand sampling method to image the migration pathways without the presence of the ligand. The approach was able to map out important cavity sites as well as pathways among them and predicted additional exit pathways that were not easy to be probed by experiments.

Robotics-inspired methods. Robotics-inspired methods have been applied to study many problems in biology, such as protein loop closure [Canutescu and Dunbrack (2003); Kolodny et al. (2005); Manocha et al. (1995)] , structure determination [van den Bedem et al. (2005)], protein backbone motions [Noonan et al. (2005)], protein flexibility and conformation sampling [Shehu et al. (2006, 2009); Yao et al. (2008); Kazerounian et al. (2005); Chirikjian (2011)], and conformation transitions [Kim et al. (2005); Schuyler et al. (2009)]. Particularly, as maps of the environments are often built to facilitate the path planning of robots, roadmap-based motion planning methods [Kavraki et al. (1996)] have been successful in studying the motions of proteins [Apaydin et al. (2003); Apaydin (2004); Chiang et al. (2007); Chodera et al. (2007); Singhal et al. (2004); Amato and Song (2002); Song (2003); Thomas et al. (2005, 2007)] and ligands [Apaydin et al. (2002); Bayazit et al. (2001)].

Voronoi graphs, alpha shapes, and computations on protein geometry. Voronoi graphs have been used to study protein geometry and packing since the 70's, in works such as those by Richards (1974) and Finney (1975). Voronoi graphs or generalized Voronoi graphs are widely used in robotic motion planning [Latombe (1991); Choset et al. (2005); LaValle (2006)] since they produce maximum-clearance roadmaps [Choset and Burdick (2000); Wilmarth et al. (1999)] that can be used to guide the navigation of a robot. Some recently developed tools to identify channels in proteins, such as Petrek et al. (2007) and Yaffe et al. (2008), are based on Voronoi graphs and alpha shapes [Edelsbrunner and Mucke (1994)]. Alphas shapes are useful to compute molecular surface area and volume [Liang et al. (1998a)], as well as to identify internal cavities [Liang et al. (1998b)] and surface pockets [Liang et al. (1998c)].

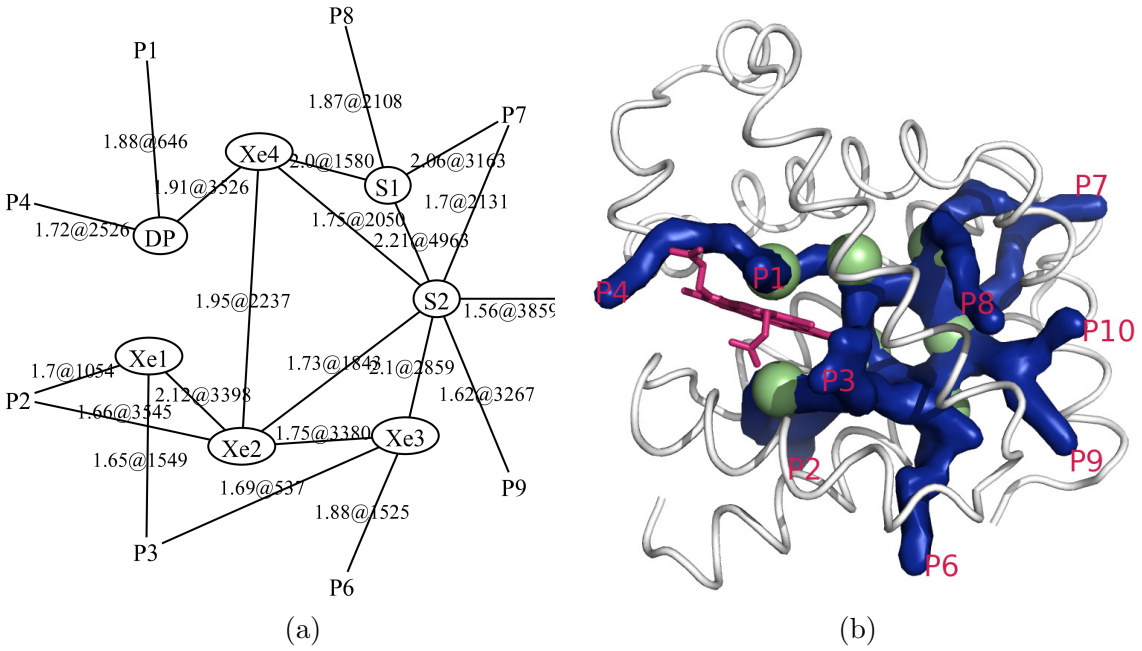
Methods

Overview of Dynamic Map Ensemble (DyME)

Compared to the existing computational methods reviewed above, our method, Dynamic Map Ensemble (DyME), is unique in several respects. First, it is ensemble-based – the dynamics and structure flexibility of the host protein is represented by an ensemble of conformations. An ensemble in the DyME approach can be composed of crystal structures, NMR structures, conformations generated in one MD simulation, or even conformations generated in multiple MD simulations using multiple crystal structures as the starting points. Secondly, Voronoi graphs computed for every conformation in the ensemble are used to identify the locations of the internal cavities and surface portals (i.e., entry/exit points on the protein surface that link the internal cavities to the solvent) as well as the channels connecting them. The channels are the maximum clearance paths among cavity sites or between a cavity site and the solvent. A channel is considered open when its clearance reaches a certain threshold. As a result, for each conformation, a map is produced that shows the current partial connectivity of the ligand migration channel network and the extent of openness of the channels. Thirdly, the partial maps are connected together at the common cavity sites and surface portals to form a super-graph (see Figure 5.1). This final super-graph represents the whole dynamic migration channel network and contains all the ligand migration pathways existing in the ensemble. Lastly, because each conformation in the ensemble is fully mapped and the channel clearances at each conformation are known, there is a direct mapping between conformations and channel clearances, or between conformational changes and variations in the channel clearances. Such direct correlations, which are difficult to acquire with any other methods, can be collected and analyzed to identify the key conformational changes that regulate these channels.

Geometric mapping methods. Voronoi graphs [De Berg et al. (1997)] are used to map the protein inner space. All atoms are approximated by spheres of the same radius of 1.6 Å. Since the vast majority of the atoms in a protein are carbons, oxygens, or nitrogens, and have a van der Waal's radius [Bondi (1964)] of 1.7 Å, 1.52 Å, or 1.55 Å respectively, the uniform sphere of 1.6 Å introduces only a small error when computing the clearances of the channels.

Hydrogen atoms are not included in the model. In future work, one improvement can be made is to represent each atom by its actual vdW radius. Consequently, additively-weighted Voronoi graphs will be needed, which, unfortunately, is computationally much more expensive. An alternative approach is to use the alpha shapes [Edelsbrunner and Mucke (1994); Liang et al. (1998b); Edelsbrunner et al. (1998)]. However, since the alpha shapes are based on power diagrams, not the additively-weighted Voronoi graphs that are needed for the exact calculation of the channel clearances, it may introduce some errors too when computing channel clearances. In the current work, a uniform sphere of 1.6 Å is used.



(a) (b)

Figure 5.2 The ligand migration channel network of myoglobin using the MD ensemble.

(a) 2-D view and (b) 3-D view. The internal cavities (shown in circles in the 2D plot and colored green on the 3D plot) consist of four xenon binding sites (Xe1-4), two sites (S1 and S2) discovered by MD [5], and the DP (distal pocket). The entry/exit portals are marked by P followed by the portal number. Between a given pair of vertices, say between DP and Xe_4, there may exist an open channel (an edge on the 2D, colored blue on 3D plot) in a subset of the conformations. The conformation (the frame number from MD) whose channel has the largest clearance as well as the clearance value itself is labeled on the edge on the 2-D figure. 2-D figures are created using software Graphviz [Ellson et al. (2004)] and 3-D figures using PyMol [Schrödinger, LLC (2010)] (www.pymol.org).

Internal cavities and the identification of their locations. For myoglobin, besides

the four xenon binding sites and the distal pocket, two additional cavities (named Ph1 and Ph2 in [Bossa et al. (2004)] and renamed as S1 and S2 here) that are large enough for a ligand to transiently stay are included in our list of cavities. The locations of these cavities relative to the protein structure are known (the green spheres in Figure 5.2(b)). Such information is used to identify which vertices in the computed Voronoi graphs correspond to cavity regions.

Channels, channel clearances, and entry/exit portals. The clearances of the channels connecting the cavities and the solvent are extracted from the Voronoi graphs computed at every conformation. The solvent ends of the channels reveal the locations of the entry/exit portals on the protein surface. A channel may be open at one conformation while closed at another. The effect of protein conformational changes is manifested in the fluctuations of the channel clearances. In other words, the channels are regulated by the conformational changes of the protein.

Implementation Details of the Dynamic Map Ensemble (DyME) approach

Voronoi diagram basics. Voronoi diagram is used to map the free space inside the protein at every conformation in the ensemble. Let $P = \{p_1, p_2, \cdots, p_n\}$ be the atom centers. The Voronoi diagram subdivides the space into polytopes, one for each atom, such that any point q lying within a polytope is nearer to the atom inside the polytope than to any other atoms. Since the subdivision is not necessarily a closed space, an infinity vertex is introduced to complete the subdivision. The subdivision can be formulated as the following mathematic definition:

$$for\ all\ q\ in\ the\ polytope\ VD(i), dist(q, p_i) < dist(q, p_j) \tag{5.1}$$

Normally uniform spheres for all atoms are assumed and Euclidean distances are used. When different atoms are given different radii, then weighted distances are used.

The Voronoi diagram has many useful properties. Each Voronoi vertex is equi-distant to its four closest atom neighbors. Two Voronoi vertices share an edge if they have three atom neighbors in common. Most attractively, the Voronoi diagram represents the maximum clearance paths

among the atoms, which makes it an ideal geometric construct for finding ligand migration pathways.

There are four major steps in the dynamic map ensemble (DyME) approach. Figure 5.3 shows the flow chart.

Step 1: Construct the maximum clearance graph and identify solvent vertices. In the first step of DyME, the Voronoi diagram is computed and the clearance of each Voronoi vertex is assigned based on its distance to its closest atom neighbors. Similarly the clearances of the edges of the Voronoi diagram can be computed quickly. Vertices as well as edges that have low clearances are then removed. The remaining Voronoi graph is further reduced to a maximum clearance tree (which is the same as a maximum spanning tree) that has no loops. Now since an actual ligand migration channel network may contain loops, some deleted edges are added back. We name this procedure "adding loops". The edge added back at each iteration has to satisfy two criteria: i) it has the largest clearance among all the deleted edges, and ii), it bridges a gap between two vertices which otherwise would have a long shortest-path distance between each other. (Specifically, a distance is long if it is longer than a threshold value, which is a parameter in the model. The performance of the model is insensitive to this parameter.) It turns out that only a few edges need to be added back as the shortest-path distances between many other pairs of vertices quickly decrease as edges are added back.

Vertices outside the protein surface are identified as solvent vertices, if they are reachable by the probe sphere from the infinity vertex. In the end, the vertices on the maximum clearance "tree" (which is now a graph with some loops) are divided into two categories, solvent or non-solvent. Among non-solvent vertices, those with high clearance (higher than 1.8 Å) are marked as candidate cavity vertices.

Step 2: Identify cavity regions.

In this step, candidate cavity vertices identified in step 1 for every conformation in the ensemble are superposed. This is done by aligning the corresponding conformations. To identify common cavity regions from these vertices, two reasonable assumptions are made.

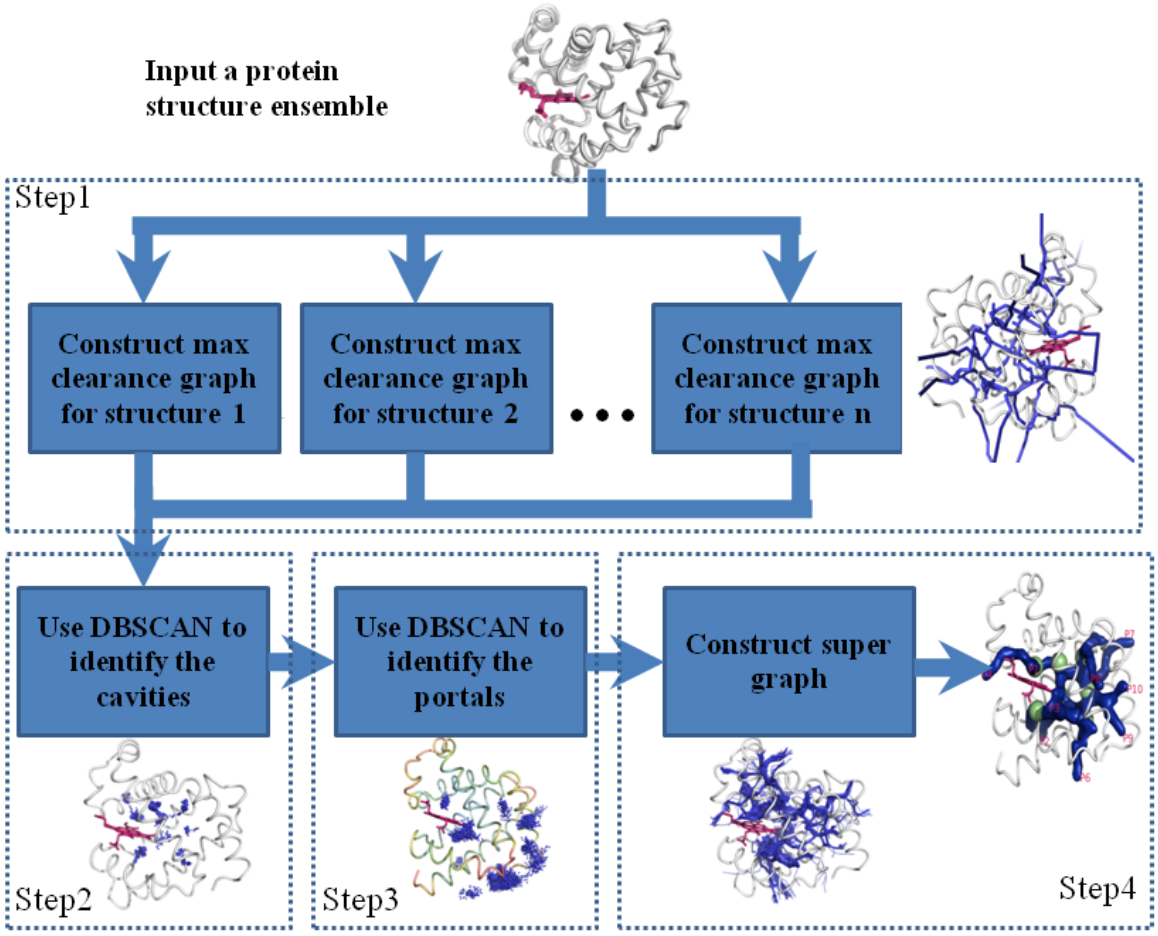


Figure 5.3 Flow chart of the DyME method.

At step 1, efficient geometric mapping is applied to each conformation in the ensemble and a maximum clearance graph that best represents the internal free space of the protein and its connectivity is computed. In steps 2 and step 3, cavity and portal vertices are identified by DBSCAN [Ester et al. (1996)] clustering and marked out on the graph. Finally at step 4, the maximum clearance graphs are combined to form a super graph (or a map ensemble), which contains the dynamic ligand migration channel network and can be used to search for open channels between cavities and the solvent.

The first assumption is that a cavity is at least some distance (a parameter in the model) away from the solvent. This is to distinguish a cavity from a surface pocket. The second assumption is that the cavities should appear more frequently and with higher clearance than the channels connecting them. This assumption allows us to distinguish cavities from open channels. A density based clustering algorithm (DBSCAN [Ester et al. (1996)]) is applied on the cavity candidate vertices and the resulting clusters are identified as cavity regions and each cavity is given a unique label. The identified cavity regions are then mapped back to each individual conformation and its associated Voronoi graph. As a result, for each individual Voronoi graph, the vertices that fall into the cavity regions are known. Within each cavity region, the vertex with maximum clearance is chosen as the representative cavity vertex for that cavity. Thus, by the end of this step, all the cavity vertices on the graphs are marked out, along with the solvent vertices outside the protein surface.

There are two parameters for DBSCAN, K and ϵ. K is the minimal number of vertices required to form a cluster and ϵ is the neighborhood radius within which two vertices are considered to be in the same cluster. An alternative approach to the clustering algorithm is to provide the cavity centers directly using known experimental data, specifically the Xenon binding sites if they are available.

Step 3: Identify portal regions.

In the context of ligand migration into and out of a host protein, a portal is defined as a ligand entry or exit point on the protein surface. They are the channel openings on the surface. In our maximum clearance graph, a portal vertex is a non-solvent vertex that connects directly with a solvent vertex and is used to connect the solvent with one of the cavities. To identify portal vertices on the graph, we recursively remove all the leaf vertices on the graph that are neither cavity vertices nor the infinity vertex. A vertex is a leaf if it has only one edge connecting to it. In the end, only cavity vertices, the infinity vertex and the vertices that are on the paths that connect them are kept. All the non-solvent vertices that have a direct edge to a solvent vertex are now identified as portal vertices. These portal vertices represent all the possible entry/exit points on the protein surface via which cavities and the solvent are

connected in that conformation. The graph also contains the maximum clearance paths among the cavities.

Next, the graphs from all the conformations in the ensemble are superposed again and DBSCAN [Ester et al. (1996)] is applied to all the portal vertices found in the graphs. The clustering identifies locations on the protein surface, or the portal sites, where portal vertices are consistently located. Each identified portal site is given a unique portal number. At this point of the method, the graphs from all the conformations in the ensemble have in common the same cavity sites and surface portal sites.

Step 4: Construct the super graph

Finally, merge the corresponding cavity vertices of the same cavity sites across all the graphs/ conformations. This results in a super graph that contains not only the spatial but also the temporal mapping of the free space inside the protein. Within this super graph, the channels among the cavity sites or between the cavities and the solvent (via the portal sites) may reach their maximum clearances at different conformations. Thus, a maximum clearance path that goes through the protein may be composed of a series of channel segments for which the maximum clearance are found at different conformations.

MD Simulation

NAMD [Phillips et al. (2005)], a parallel molecular dynamics simulation program, was used to conduct the MD simulations. Periodic boundary conditions were applied in the simulation process with the CHARMM27 force field. The starting structure (PDB ID: 1A6G) was solvated in a 10 Å water box. The simulation started with 100 ps energy minimization, followed by a 10-ns equilibration. MD simulation was conducted using a HPC cluster computer. 5,000 conformations were extracted from the MD trajectory at 2ps intervals.

Structure Ensemble Preparation

To apply the dynamic map ensemble (DyME) approach, we first construct ensembles of

Table 5.1 The execution times of DyME on three ensembles.

The results are obtained by executing DyME on a Linux PC with 2.53 GHz CPU and 2 GB memory using MATLAB.

Ensemble	**Execution time [sec]**
Myoglobin crystal structure ensemble (227 structures)	1264
10-ns Myoglobin MD ensemble (5,000 conformations)	59,351
Cytochrome P450cam crystal structure ensemble (120 structures)	8,387

conformations that represent the protein structure flexibility near the native state.

The DyME approach is ensemble-based and it provides the flexibility to study the effect of protein dynamics expressed in different forms. When applied to an ensemble of crystal structures, it can provide results on ligand migration pathways that are based purely on experimental structures. It has been shown [Best et al. (2006)] that "even a modest set of structures of a protein determined under different conditions, or with small variations in sequence, captures a representative subset of the true native state ensemble." For myoglobin, there are over 200 crystal structures available. Such a collection of experimental structures of the same protein present a valuable sample of the native state ensemble that is not simulation based [Yang et al. (2008)]. New NMR-based approaches are also being proposed that promise to generate structure ensembles with richer dynamics [Lindorff-Larsen et al. (2005); Richter et al. (2007); Lange et al. (2008)]. Such ensembles will be ideal inputs for DyME when it is used to study the effect of protein dynamics on ligand migration.

We have constructed three ensembles in this work. The first ensemble contains all the myoglobin structures in PDB [Berman et al. (2000)] that have a resolution better than 2.0 Å and their substates. There are 227 such structures (counting the substates). These structures provide valuable information on myoglobin's structural flexibility that comes directly from experimental data. The second ensemble is composed of conformations generated in a 10-ns molecular dynamics simulation of the myoglobin (PDB-id: 1A6G) in explicit water. The third ensemble contains 79 cytochrome P450cam PDB structures that have a resolution better than 3.0 Å and their substates, and thus a total of 120 conformations. The DyME approach was applied to all three ensembles and the execution times are given in Table 5.1.

Results

Mapping the Ligand Migration Channel Network of Myoglobin

One advantage of the DyME approach over most other methods is that it produces a complete mapping of all the space that is *spatially* and *dynamically* accessible to a ligand and thus the whole dynamic ligand migration channel network.

Figure 5.2 shows the channel network mapped by DyME, using the ensemble of the MD-generated conformations. The vertices in the network represent internal cavities or surface portals via which cavities are connected to the solvent. The edges represent the channels connecting them. In addition, each edge label on the 2-D plot (Figure 5.2(a)) shows the maximum clearance reached for that channel as well as the conformation (the frame number from the MD simulation) at which the maximum clearance is reached. Remarkably, from this single ensemble of conformations generated from a 10-ns MD simulation, all the known significant internal cavities are mapped out and most entry/exit channels connecting the inner cavities to the solvent are identified. In comparison, most existing MD approaches simulate the ligand explicitly and each simulation takes tens and perhaps even hundreds of nanoseconds to generate one pathway or sometimes even none. Multiple simulations have to be repeated in order to find other pathways. The advantage of mapping approaches such as DyME over the trajectory-based approaches has also been reported for protein folding pathway studies [Amato and Song (2002); Song (2003)].

The dynamic channel network portrayed in Figure 5.2(a) also demonstrates clearly that protein dynamics is critical to the migration of a ligand – the channels open at the different conformations and a ligand's migration in the protein has to be in sync with the conformational changes of the protein. Figure 5.2(b) shows the corresponding 3-D plot of the migration channel network inside the protein matrix.

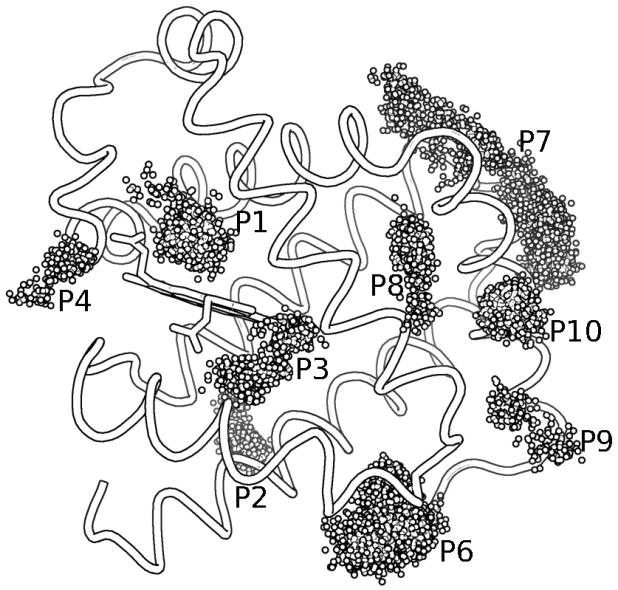


Figure 5.4 Portal clusters identified based on the MD ensemble.

Each small circle represents a channel opening on the protein surface at a particular conformation. All the channel openings (circles) are overlaid and a clustering algorithm is applied to identify the surface portals via which a ligand may enter or exit the protein matrix.

A Close Examination of the Surface Portals of the Channel Network

Surface Portal Identification

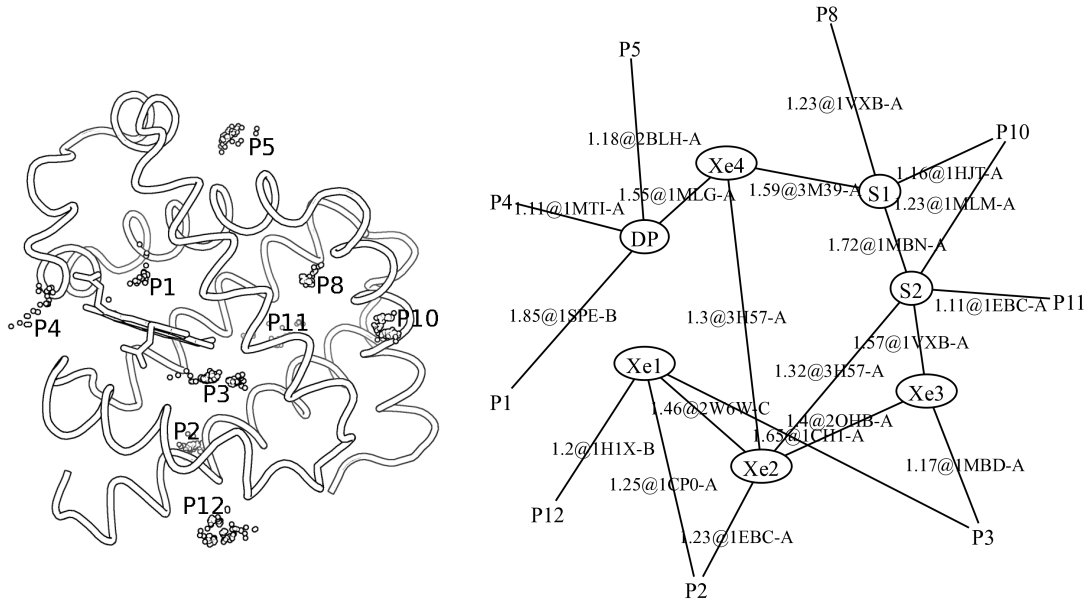
The geometric mapping of each conformation in the ensemble reveals how the cavities are inter-connected and how they are connected to solvent. Figure 5.4 shows all the channel endpoints on the surface after mapping all the conformations in the ensemble and structurally aligning them to the mean structure. The points fall into clusters (DBSCAN clustering algorithm is employed for this purpose [Ester et al. (1996)]), each of which represents a populated region where a ligand may enter or exit the protein. Such regions are named entry/escape portals by Ruscio et al. (2008), who applied extensive MD simulations to identify these channel openings. Remarkably, most of the portals (except P5) found by their cumulative 7-μs MD simulations [Ruscio et al. (2008)] are identified by our method using the ensemble constructed from a 10-ns MD simulation, which is about 3 orders of magnitude shorter than the cumulative

Table 5.2 Surface residues surrounding the entry/exit portals.

The common residues between Ruscio et al. (2008) and ours are colored red. Residues in bold are those identified as important for ligand binding kinetics by Huang and Boxer (1994) mutagenesis study. Portals 10 to 12 are new portals (marked with *) predicted by our method.

Portal	Portal Residues from Crystal Ensemble	Portal Residues from the 10ns MD Ensemble	Portal Residues from Ruscio et al. (2008)
1	**64, 67**, Hem	**64, 67**, Hem	**64, 67**, Hem
2	101, 105, 143	101, 139, 143	101, 104, 139, 143, **146**
3	**71**, 85, 89	67, 70, **71**, 85, 88, 89	**71**, 74, 75, 82, 85, 89
4	42, **44**, 96	**44**, 97	**43, 45**
5	**26**, 27, **56**		**34**, 51, 55
6		1, 79, 80, 83, 86, **137**, 141, **145**	7, 134, **137**
7		12, 13, **16**, 17, 115, **119**, 122	**16**, 20, **24**, 118, **119**
8	18, 21, 70	18, 70, 74, 77, 78	18, 21, 70
9		7, 11, 77,79	8, 11, 79
10*	11, **14**, 77	11, 77	
11*	128, 132		
12*	83, 141, **144**, 148		

7-μs simulation [Ruscio et al. (2008)] and 1 order of magnitude shorter than what is normally required for the conventional MD simulation to find just one ligand migration pathway in myoglobin. The residues that surround these portals are listed in Table 5.2. From the table it is seen that portals 6, 7, and 9 are absent in the crystal structure ensemble, indicating these portals may become visible only when the protein is in motion. On the other hand, putative portals 11 and 12 suggested by the crystal ensemble are not observed in the MD ensemble or by Ruscio et al.'s trajectory-based simulations [Ruscio et al. (2008)]. Since a lower channel clearance requirement was used for open channels in the crystal ensemble, it is possible that portal 11 or 12 is not a true portal. Another possible explanation is that these two portals open up much less frequently than the others.



(a) (b)

Figure 5.5 Portal clusters and the ligand migration channel network predicted solely from the ensemble of crystal structures.

Direct Experimental Evidence of Ligand Channels

The DyME approach employs structure ensembles to represent protein structural flexibility and relies on efficient geometric mapping to identify dynamic ligand migration channels. When applied to the experimental structures of a given protein, it can weave partial information existing in each individual structure together to form a more complete picture of the ligand migration channel network that is based solely on experimental data, untainted by uncertainties or errors that are likely to be introduced in simulations.

Figure 5.5(a) shows the ligand entry/exit portals based on the crystal structure ensemble. Figure 5.5(b) shows the 2-D view of the ligand migration channel network. Each edge represents a channel between two cavities or between a cavity and the solvent, with an edge label indicating the maximum clearance reached in that channel as well as the structure PDB-id for which the maximum clearance is reached. Since our goal here is to predict where channels may potentially open, the clearance threshold for the channels is lowered and set to be 1.1 Å. Remarkably, many of the channels found in MD simulations (see Figure 5.2 and results in Ruscio et al. (2008))

can be predicted by simply analyzing the existing crystal structures. Ruscio et al. (2008) found that there are two discrete dynamic pathways in myoglobin, a "major" pathway that is more frequently used by the ligand, and a "minor" pathway (see Ruscio et al. (2008)). Interestingly, all the portals (portals 1 to 5) on the major pathway can be identified by DyME from the crystal structure ensemble, and only one of the four portals (portals 6 to 9) is identified on the minor pathway, which provides a structure-based explanation for the less frequent usage of the minor pathway as was observed by Ruscio et al. (2008). Most entry/exit channels shown in Figure 5.5(b) have a clearance less than 1.4 Å. However, it is foreseeable that thermal fluctuation can further increase the clearances of these channels, causing them to be truly open. The scale of fluctuations in channel clearance that the thermal fluctuations of a protein can bring about is examined in the next section.

Table 5.3 Numbers of conformations at which a portal is open or may potentially open.

From the MD ensemble (5000 conformations in total), the first two most frequently open channels are P6 and P1, the latter of which corresponds to the Histidine gate. For the crystal structure ensemble (223 structures in total), a lower threshold (1.1 Å instead of 1.5 Å) is used to identify channels that may potentially open.

Ensemble	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
MD (clearance $\geq$ 1.5 Å)	538	14	15	7	0	3526	119	44	1	5	0	0
Crystal (clearance $\geq$ 1.1 Å)	28	97	66	1	1	0	0	4	0	7	1	5

Table 5.3 lists the number of crystal structures or MD conformations at which a given portal is open (or nearly open, in the case of crystal structures). One portal (P6) opens significantly more frequently than others. A close examination shows that many residues at or near this portal are highly flexible. Figure 5.6 shows the root mean square fluctuation (RMSF) of the residues throughout the 10-ns MD simulation. Many of the residues near portal P6, marked out with an 'x' on the plot, have significantly higher RMSF values. The second most frequently opened channel during the MD simulation is from DP to portal P1, which corresponds to the "Histidine gate". This is not surprising since the "Histidine gate" had long been recognized as a major channel [Scott et al. (2001); Ringe et al. (1984); Olson et al. (1988)] to the binding

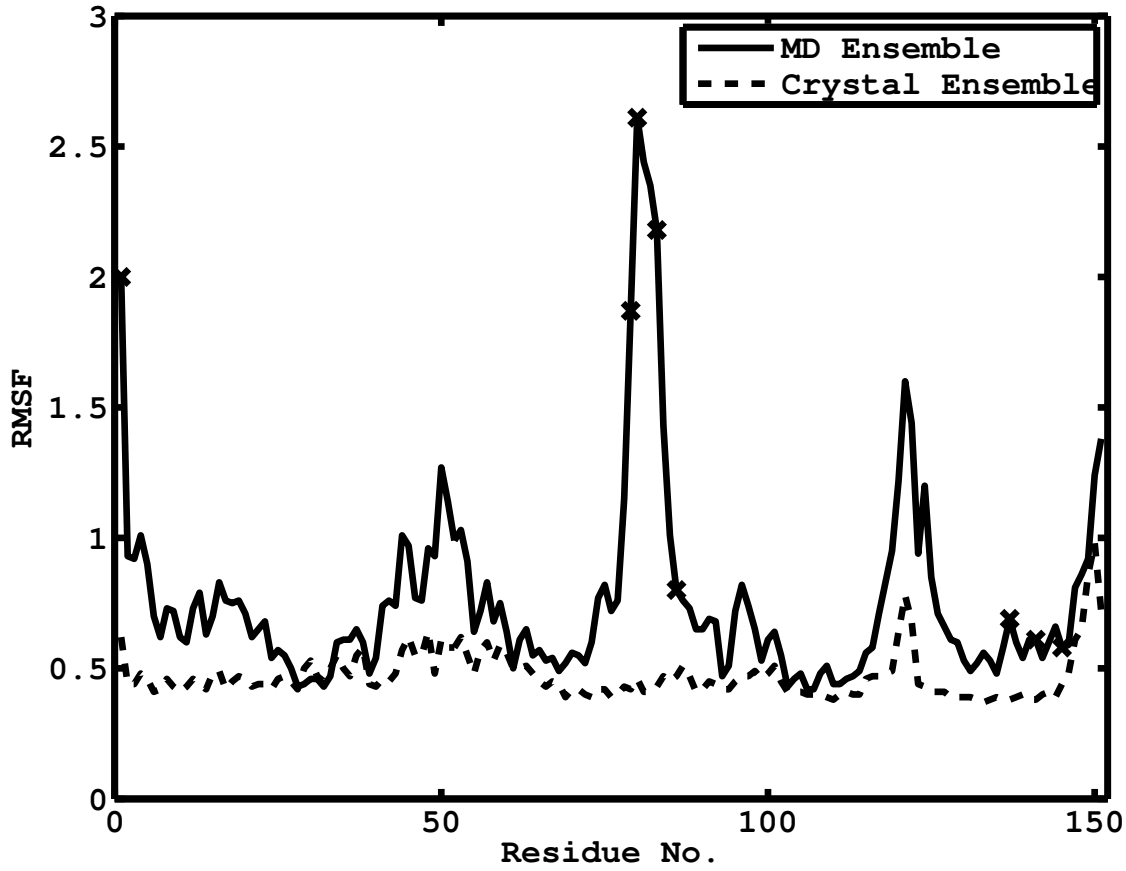


Figure 5.6 The root mean square fluctuations of the residues based on the MD ensemble.

Residues near portal P6 are marked out by 'X'. Several residues near P6 are highly flexible, providing a dynamics basis for the opening of this channel.

site. It is worth noting, however, that for these two most frequently opened channels, the opening mechanisms are quite different. The His gate is mainly controlled by the swing of the side chain of Histidine 64, while for portal P6, the large backbone motion in the loop region between helices E and F is the main contributor to the opening of this channel.

The Dynamics of the Channels

Because the ligand migration channel network is individually mapped at each and every conformation in the ensemble, another advantage of the DyME approach is that it provides direct information about the dynamic fluctuations of each channel in the network, and particularly, how often a channel is open and to what extent a channel's clearance fluctuates.

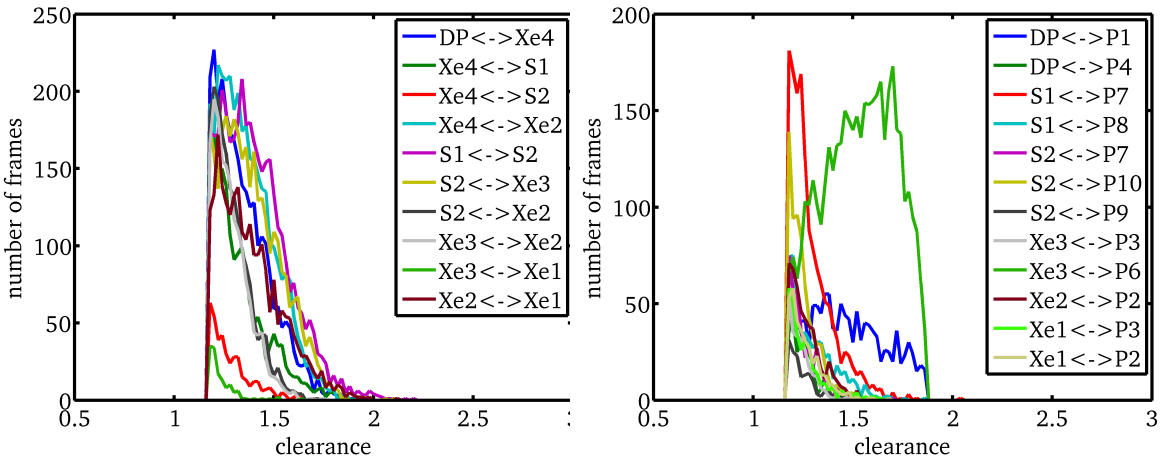


Figure 5.7 Clearance distributions of the channels that are, (a) between the cavities, and (b) between cavities and solvent (via the portals), based on the MD ensemble.

Since a channel is no longer recognizable when its clearance is too small, the distributions all begin at 1.2 Å. The channels via portals P6 and P1 are the top two most frequently opened channels, as clearly displayed in the clearance distributions of these two channels.

The Fluctuations of the Channels

Figure 5.7 shows the fluctuations of the channel clearances within the conformation ensemble generated from the 10-ns MD simulation. Specifically, Figure 5.7(a) shows the clearance distributions of the channels between the cavities and Figure 5.7(b) the channels between cavities and the exterior. There are several observations. First, the distributions of the channel clearances are close to Gaussian distributions. Particularly, each distribution has a tail that decreases rapidly in magnitude. The scale of the fluctuation, i.e., the difference between the maximum and minimum clearances, ranges from 0.4 Å for some channels to as high as 1.0 Å for some others. The fluctuation is contributed by the thermal fluctuations of the structure. It is thus perceivable that such fluctuation should be able to cause the near-open channels observed in the crystal structure ensemble (see Figure 5.5) to truly open up. Importantly, the inter-cavity channels are found to open wider and more frequently than the cavity-to-solvent channels. This may have a functional reason, for example, to regulate the rate of reactions of the ligands and to prevent expulsion of ligand.

Table 5.4 Channels identified by our method using the MD ensemble and the residues lining the channels.

Residues in bold are those identified by Huang and Boxer (1994) using mutagenesis data to be important for ligand binding kinetics.

Channels	Residues Lining the Channel
DP $\longleftrightarrow$ Xe_4	Hem **107 68 29**
DP$\longleftrightarrow$P1	Hem **67 64**
DP$\longleftrightarrow$P4	Hem **107** 32 **29 64 43 33 46 44** 97
Xe_4 $\longleftrightarrow$S1	111 72 **69 14**
Xe_4 $\longleftrightarrow$S2	111 72 **69 14** 131
Xe_4 $\longleftrightarrow$ Xe_2	111 72 **68** Hem 135
S1$\longleftrightarrow$S2	115 111 17 **14** 131 72
S1$\longleftrightarrow$P7	114 28 **24** 17 115 **119**
S1$\longleftrightarrow$P8	**69** 28 **24** 17 25 21 **14** 18 70 73 74 77
S2$\longleftrightarrow$ Xe_3	135 134 131 76 72 **138**
S2$\longleftrightarrow$ Xe_2	134 131 10 76 135 **138** 72
S2$\longleftrightarrow$P7	131 115 **14** 10 123 13 17 **119 16** 122
S2$\longleftrightarrow$P10	135 131 76 **14** 10 77 11
S2$\longleftrightarrow$P9	134 131 76 10 **14** 11 7 77
Xe_3 $\longleftrightarrow$ Xe_2	**138** 76 75 72 135
Xe_3 $\longleftrightarrow$P3	**138** 76 75 72 86 Hem **71** 89 85
Xe_3 $\longleftrightarrow$P6	134 76 75 1 **137** 141 79 86 80
Xe_2 $\longleftrightarrow$ Xe_1	Hem 135 **108** 139 **138** 142 104
Xe_1 $\longleftrightarrow$P3	Hem 142 89 **138** 86 75 **71** 85 **67**
Xe_1 $\longleftrightarrow$P2	**146** 142 104 101 139

Validating Ligand Migration Channels with Mutagenesis Data

The dynamic map ensemble stores the information about all the channels, for example, the residues and atoms that line the channels. Table 5.4 lists the channels identified by our method using the 10-ns MD ensemble and the residues that line these channels. The random mutagenesis study by Huang and Boxer (1994) found that a surprisingly large number of mutants exhibited different ligand-binding kinetics from the wild-type protein. In our study, we find that the mutated residues that are significant for ligand-binding kinetics appear frequently on the surface of the ligand migration channels identified by our method (shown in bold in Table 5.4). For every channel, there is at least one residue that was found important for the ligand-binding kinetics. Among all migration channels, the channels connecting to the DP site (i.e., DP to Xe_4, DP to P1, DP to P4, see the map in Figure 5.2(a)) must have a significant role in ligand-binding kinetics since most of the residues lining these channels were identified as important residues by Huang and Boxer (1994).

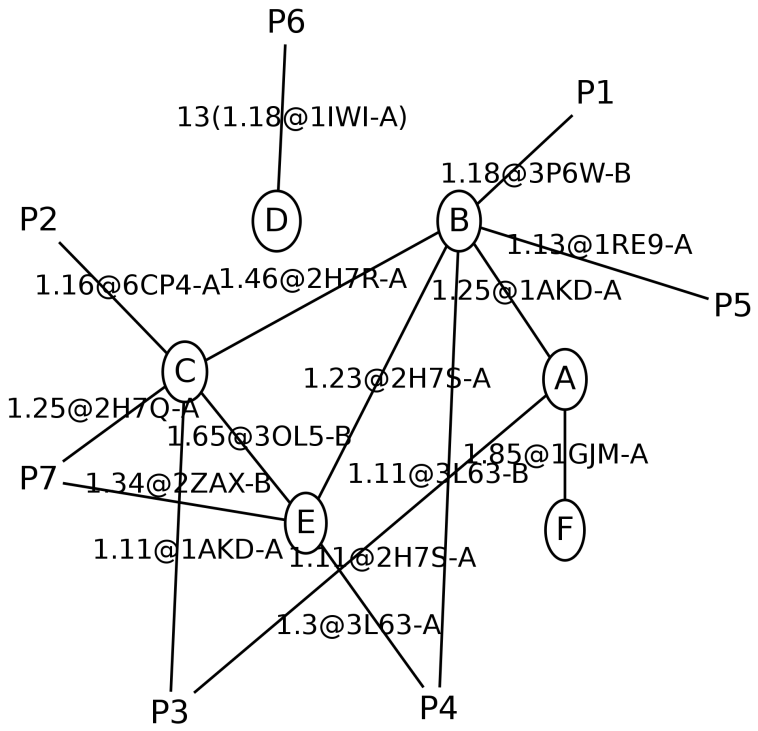


Figure 5.8 2-D view of the ligand migration channel network of cytochrome P450cam using an ensemble of 120 crystal structures.

The internal cavities (shown in circles) consist of six cavities identified by DyME. The entry/exit portals are marked by P followed by the portal number. An edge represents a putative open channel in a subset of the conformations. The conformation whose channel has the largest clearance as well as the clearance value itself is labeled on the edge.

Mapping the Ligand Migration Channel Network of cytochrome P450cam

The mapping approach presented in this work is applicable to other proteins as well. Here we apply it to a cytochrome P450cam ensemble that contains 120 X-ray structures/substates. Cytochrome P450cam was among the first solved X-ray structures in the cytochrome P450 superfamily [Winn et al. (2002)]. Like myoglobin, the structure has a deeply-buried heme pocket. A substrate, camphor, is bound to the active site. Previously, eight persistent cavities were identified [Mouawad et al. (2007)] from 36 trajectories of 1-ns MD simulations. However,

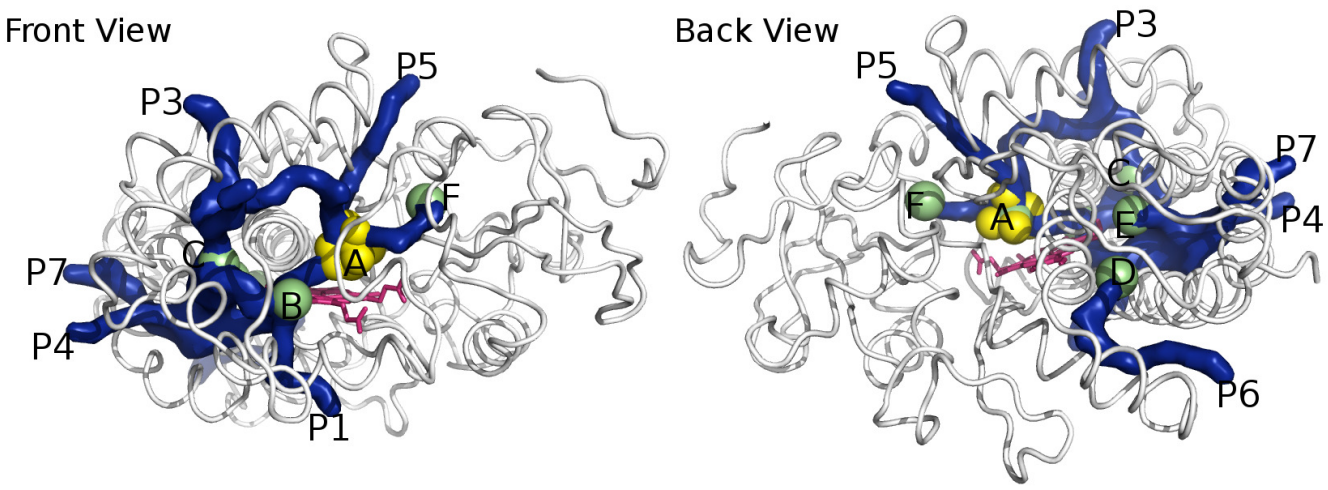


Figure 5.9 3-D view of the ligand migration channel network of cytochrome P450cam using an ensemble of 120 crystal structures.

The internal cavities, colored green, consist of six cavities identified by DyME. Putative open channels are colored blue. Heme is colored red and camphor is colored yellow.

only five of them, namely cavities A-E, seemed to possess functional importance. Cavity A is where camphor resides, and cavities B, C and D are where the Xenon gases were found. We identify six cavities from the X-ray ensemble and find a high frequency of appearance for all of the five aforementioned cavities, i.e., cavities A-E. Figure 5.8 and Figure 5.9 are respectively the 2D and 3D representations of the ligand migration channel network, using 1.1 Å as the clearance threshold for putative open channels. Six portals also are identified. Previous work using MD simulations identified three main pathways, i.e. pathways 1-3, for substrates or small ligands to migrate into and out of the protein [Mouawad et al. (2007); Wade et al. (2004)]. Our results show similar network patterns. The paths going through portals P4 and P7 correspond to pathway 1, while the paths going through P5 and P3 correspond to pathways 2 and 3, respectively. From the 2D graph in Figure 5.8, it is seen that the most feasible path from the active site to the solvent is the path that starts at cavity A, going through cavities B and E, and enters the solvent at portal P7, which corresponds to pathway 1 identified before.

Discussion

Since it is difficult to observe the migration pathways of a ligand experimentally, simulation-based methods, particularly molecular dynamics (MD), have been the choice of approach in studying ligand migration pathways. Compared with the conventional usage of MD, DyME has several clear advantages. It has its own limitations too. Understanding its strengths and limitations is necessary to properly apply the method.

First, the DyME approach maps the whole ligand migration channel network of the host protein and delineates how the clearances of the network channels may vary as the host protein changes its conformations under thermal fluctuations. It establishes direct correlations between the conformations of the host protein and the channel clearances, as each conformation in the ensemble is individually mapped and the channel clearances at each conformation are known. The analysis of such correlations may reveal how the channel clearances are regulated by the dynamics, or the conformational changes, of the host protein. The dynamic nature of the ligand migration channel network, and the dynamic variation of the channel clearances, are most clearly portrayed in DyME. It shows that different parts of the channels may open at different conformations, which explains why there is no route from the solvent to the buried binding site for the average structure, but many routes may exist when the dynamics of the protein is taken into account. If we treat the geometric mapping of the host protein at each conformation as a spatial exploration to the free space available to the ligand, and the dynamics of the host proteins as a temporal exploration of protein flexibility, we will see that there are no open routes within the space inside the host protein at any single time, but there are many routes in the combined space-time. Trajectory-based MD simulations trace the trajectory of a ligand in this space-time, while the DyME approach decouples the time-space exploration of the whole system into a separated temporal exploration of protein structural heterogeneity and a spatial exploration of protein inner space.

MD-like methods approach the ligand migration problem by integrating over the time evolution of the whole system that includes the protein, the ligand, and solvent. The trajectory of the ligand is traced through the simulation process. Each simulation usually produces just

one pathway. If a different pathway is desired, then the simulation has to be repeated, possibly for multiple times or for a much longer time. Since the motion of the ligand is driven by diffusion and is stochastic in nature, such trajectory-based approaches are not best suited to trace out the complete ligand migration channel network. A mapping approach, such as the one developed in DyME, is the better choice for such a purpose.

Another way to appreciate the difference between a mapping approach and a trajectory-based one is to realize that it requires a large extent of coordination between the motions of the ligand and the protein in order for the ligand to migrate through the protein. For example, when a transient channel is opened, the ligand needs not only to be there but also to have the right momentum in order to take advantage of the transiently opened channel to move through the channel to the next site. A transient channel may have to open and close many times before the ligand can pass through. Such coordination, accomplished through random diffusion, inevitably lengthens the time needed to simulate explicitly a ligand's transition from one site to another, making it many times longer than the time it takes for the channel to open just once (which would be sufficient for a mapping approach to capture it).

The computational efficiency gained by a mapping approach over trajectory-based methods does come with some compromises. MD simulations produce time-dependent trajectories of the ligand. In the DyME approach, the exact trajectories that a ligand may take and their time-dependent information are no longer known. What is known instead is the exact shapes and sizes of the network channels and how they vary as the host protein fluctuates. The extent of opening and the fluctuations of each channel and the wait time until it takes place for the protein to make the needed conformational changes, however, can be used to indirectly determine the rate of ligand migration along the channel. The estimated transition rates among the cavities and the solvent can then be used to write down a master equation of the whole system, the solution to which will give the population kinetics at each cavity.

The interaction between the ligand and the protein is automatically included in explicit-ligand trajectory-based approaches. In the DyME approach, since the trajectory of the ligand is not explicitly followed and the network channels are the focus and are mapped instead, care

must be taken to adequately address the effect that a ligand would have on the host protein and thus the channel network. Since the size of diatomic ligands is small and uncharged, it is thus not unreasonable, as the zero-order approximation, to map the ligand migration channel network without the presence of the ligand. Indeed, work such as that by Tilton et al. (1984) had found that the presence of the ligand has only a small effect on the protein structure and cavity volumes. The idea of using an implicit ligand was also exploited Cohen et al. (2006), where the migration pathways were imaged using an implicit ligand model by calculating the mean free energy. In the case where the zero-order approximation is not sufficient, the following first-order approximation to the effect of the ligand can be adopted. In contrast to the zero-order approximation where the protein flexibility and dynamics is explored without the presence of any ligand, at the first-order approximation a ligand can be placed in turn at each cavity site or the solvent and multiple MD simulations can be run to generate ensembles of conformations. In this way, the interactions between the ligand and the protein and their effect on the channel network can be mostly included.

In summary, the DyME approach has the following advantages:

- Mapping the whole ligand migration channel network. A trajectory-based MD simulation run can produce only one single transition pathway a time for the ligand. The dynamic map ensemble constructed by DyME, on the other hand, represents the whole ligand migration channel network that is accessible to the ligand, *spatially* and *dynamically*.

- Computational Efficiency. Because the trajectory of the ligand is not explicitly followed, the requirement for the coordination between the motions of the ligand and the protein through random diffusion is removed. We now need to examine the dynamics of protein only long enough to see the transient channels open once. This is many times shorter than the simulation time needed to see a ligand actually traversing through these channels.

- Clear display of the effect of protein dynamics or conformational changes on ligand migration. Since the protein inner space at each conformation is individually mapped in DyME, the effect of protein conformational changes is directly reflected in the fluctuations

of channel clearances. It is thus clearly shown in DyME how protein dynamics may open one channel and at the same time close another. The scales of fluctuations of channel clearances are also readily available in DyME.

- Facilitating the study of the control mechanism. Though protein dynamics is well known to contribute the opening of transient channels, it is not clear how the control is carried out at the atomic level and what residues are key in regulating the channels. Since the dynamic map ensembles directly link protein conformational changes to the variations in channel clearances, analysis of the correlation between conformational changes at the residue level and channel clearances may pinpoint the key regulatory residues, as is expected that the motions of the key channel-controlling residues should have a strong correlation with the channel clearances.

- Flexibility and broad applicability. In DyME, the dynamics of a protein is represented by an ensemble of conformations. Consequently, DyME can be applied to study the effect of protein dynamics in general, not limited to conformations generated in MD simulations. For example, it can be applied to an ensemble of crystal structures (of the same protein) and/or NMR structure ensembles. When used in this way, it becomes a method for studying ligand migration pathways that is purely experimental structure-based. This is novel. It is significant too because it provides direct experimental evidence for how ligand may migrate in the host protein. Furthermore, the flexibility of DyME allows the use of conformations of mixed origins. This allows one to form an ensemble by combining experimental structures, conformations generated from one or more simulations, and/or conformations sampled using normal modes. The combined information may reveal a more complete picture of the ligand migration process than any subset of the conformations can. DyME can also be applied to conformation substates, such as the three well-known substates (A0, A1, and A3) of myoglobin, to study how substate inter-conversion affects ligand migration [Hummer et al. (2004)].

Conclusions

To conclude, we have developed a novel computational framework called DyME that can efficiently map the ligand migration channel networks in dynamic proteins. Results on the pathways in myoglobin for gaseous ligands demonstrate that the method is able to map out the ligand migration channel network in a much shorter time than what is required for conventional MD simulations. DyME is unique in that it integrates an efficient spatial mapping of protein inner space with a temporal exploration of protein structural heterogeneity, and produces an exact mapping of all the space reachable by a ligand, *spatially* and *dynamically*. Moreover, DyME provides direct information on the correlation between protein conformational changes and the fluctuations of the channel clearances. Such correlation information may be critical in determining the molecular mechanism by which the ligand migration channels are regulated.

Acknowledgements

The authors would like to thank Robert Jernigan and Jie Liang for valuable discussions. Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged.

CHAPTER 6. PREDICTING ALLOSTERIC COMMUNICATION PATHWAYS USING MOTION CORRELATION NETWORK

A paper published in Asia-Pacific Bioinformatics Conference 2009

Tu-Liang Lin and Guang Song

Abstract

Background: Allosteric regulation can be described as the binding of an effector at one site switches the functionality of another site, often at distance. Although a wide variety of models have been proposed, the underlying mechanism of the allosteric communication remains unclear. In this work, we hypothesize that the allosteric communication between the allosteric site and catalytic site should be carried out along pathways of residues that have strongly correlated motions, so that information such as conformation change can be quickly transduced from one site to another. Results: (i) The intramolecular communication pathways of 10 out of 15 myosin proteins derived from our Motion Correlation Network (MCN) model agree with the pathways derived from multiple sequence alignment (MSA) in a very high statistically significant level ($< 1.0E-08$). (ii) The pathways of the remaining 5 myosin proteins, which all fall in the post-rigor state, are completely different from the pathways obtained from MSA and the disagreement suggests the possibility of the existence of a different route in the post-rigor state. (iii) The intramolecular communication pathways of thrombin derived from our method agree with the pathways derived from electron density maps in a high statistically significant level ($< 1.0E-05$). Conclusions: We provide a simple and computationally inexpensive approach to identify the putative allosteric communication pathways. The excellent agreement between our results and previous works supports our hypothesis that the most efficient allosteric

communication is through pathways of residues that have strongly correlated motions. Such an agreement also implies that sequence conservation, which has been used to identify allosteric communication pathways, may have a dynamics origin.

Background

The importance of allosteric communication

A number of important biological processes, such as cell signaling or metabolic activities, employ allosteric regulation to mediate the processes. Allosteric regulation can be described as the binding of an effector at one site causing a conformational change at a distant catalytic site. The conformational change might switch the functionality of the proteins, e.g., from T (tense) state to R (relaxed) state or from R state to T state [Monod et al. (1963)]. Many proteins adopt the allostery to control their functions, e.g., myosin, G protein-coupled receptors and phosphofructokinase [Gether (2000); Houdusse et al. (2000); Shirakihara and Evans (1988)]. Although some of the allosteric communication details have been revealed in the past few years [Swain and Gierasch (2006)], the mechanism that governs the allosteric regulation is still a mystery.

Previous methods for allosteric communication pathway identification

The identification of allosteric communication pathways in the allosteric proteins remains a difficult task and has received a lot of attention recently [Swain and Gierasch (2006)]. Several experimental and theoretical methods have been proposed, including statistical analysis of sequence conservation [Suel et al. (2003); Tang et al. (2007)], molecular dynamics simulation [Rousseau and Schymkowitz (2005)], elastic network model [Zheng and Brooks (2005)], and experimental methods such as studying electron density map changes [Gandhi et al. (2008)] and targeted allosteric mutant sites [Kimmel and Reinhart (2000)].

On the theoretical side, formulating the proteins as network structures is a favorable trend in recent computational studies of allosteric communication [Tang et al. (2007); Daily et al. (2007);

Chennubhotla and Bahar (2006).] Tang et al. (2007) used the residue contact network with the weights derived from sequence conservation scores of multiple sequence alignments. Daily et al. (2007) formulated a contact rearrangement network between two different states (active and inactive). Chennubhotla and Bahar (2006) employed Markov propagation of information.

Theoretical models for protein allostery

Historically, two major models, MWC concerted [Monod et al. (1965)] and KNF sequential model [Koshland et al. (1966)], have been proposed to explain the protein allostery. For an oligomeric protein, MWC concerted model assumes all subunits are functionally identical and the change is all-or-none whereas KNF sequential model assumes each individual subunit can undergo its own conformational change. These two models are still under a lot of debate. However, it has been widely believed that the fundamental mechanism of undergoing a conformational change in response to an effector binding is intrinsic to proteins [Tang et al. (2007)]. Some studies show that the intrinsic dynamics of enzymes in the unbound state is related to allosteric regulation, and proteins have the ability to sample conformations that meet functional requirement under native state conditions [Bahar et al. (2007)].

The ability of sampling conformations in the absence of ligands suggests that extracting the allosteric communication pathway from a single unligand structure is possible. Our conjecture comes from some recent researches which attempt to link the equilibrium fluctuations and catalytic functions [Bahar et al. (2007); Kern and Zuiderweg (2003)]. Also recent researches provided mounting evidence that allosteric conformational changes not only been observed upon ligand binding but also in the equilibrium fluctuations when the ligand is absent [Fetler et al. (2007); Tobi and Bahar (2005); Hammes-Schiffers and Benkovic (2006)]. Due to the increase support of the linkage between the intrinsic dynamics and protein functions, it provides the possibility of predicting the allosteric pathway from the intrinsic dynamics encoded in the unbounded structure. Based on this conjecture, we propose a method to extract the allosteric communication pathway using a single unligand structure.

Our method is based on the hypothesis that the most efficient allosteric communication path-

way between the allosteric site and the catalytic site is through residues that have strongly correlated motions. We derive the scales of motion correlation among the residues from Elastic Network Model (ENM) and name our model as Motion Correlation Network (MCN). Elastic Network Model (ENM) [Bahar et al. (2007)] is one of the successful methods for studying protein dynamics and has been applied to study the intramolecular communication pathway in several allosteric proteins, e.g., myosin, cycling-dependent kinase II and GroEL chaperonin [Zheng et al. (2006); Gu and Bourne (2007); Zheng and Brooks (2005)].

Our contributions

In this paper, we introduce a novel approach to identifying the potential communication pathway of protein allostery. Our method differs from previous studies in several aspects. First, we formulate the intramolecular communications as a network structure and the communications is through correlated motions, which is an innovative idea. Second, our approach requires only a single protein structure and our results are comparable to other methods that used multiple structures, such as MSA [Tang et al. (2007)]. Third, our approach can generate not only single paths, but also path ensembles. Fourth, the motion correlation between two residues is calculated from correlation patterns [Yesylevskyy et al. (2006)] that have been used successfully in defining the relative motions in the domain decomposition problem. The motion correlation network (MCN) structure that we construct in this work is purely from ENM and we assume that the allosteric communications between two sites are accomplished through correlated motions.

The Markov propagation of allosteric effects proposed by Chennubhotla and Bahar (2006) is perhaps the closest to our work. However, they formulated the communication probability between a pair of residues to be a function of their spatial overlap, e.g., the percentage of atoms that are in contact. Therefore, the hypothesis behind their approach and ours is fundamentally different. In addition, the correlation between two residues in our approach is more robust to small changes in protein structures.

Results and discussion

The intramolecular communication pathway of myosin

Myosins are motor proteins which transform the chemical energy, such as that from ATP hydrolysis, into movement. The energy transduction process can be seen as an allosteric communication, where the chemical energy is propagated from the binding site to a distant region and a movement is induced. Crystallographic and biochemical researches have identified three main conformational states, pre-stroke state, rigor state and post-rigor state, which correspond to the different stages in the actomyosin cycle [Houdusse et al. (2000)]. We divide the myosin structure into motor head and tail (lever arm). At the beginning of the actomyosin cycle, the ATP hydrolyzes into ADP and phosphate (Pi), both of which are bound to the motor head and the lever arm is up. The release of the ATP hydrolysis energy will eventually cause movement of the lever arm. The pre-stroke state is the state before the movement occurs. After the power stroke, the ADP and phosphate are released from the binding site and the lever arm swings to the down position. The nucleotide-free structure is the rigor or near-rigor state. When another ATP binds to the myosin, the molecular structure goes back to the post-rigor state.

In deriving the allosteric communication paths, we use the same start and end points as in Tang et al. (2007) to represent the allosteric site and the catalytic site. The source of the allosteric communication is set to be the $\gamma-$ phosphate of ATP and the end residue T742, which is a residue that resides on the surface of the converter near the lever arm. According to Tang et al. (2007), T742 has the highest rank in the communication efficiency compared with other residues in the lever arm area and the communication efficiency is obtained through the Cartesian distance divided by the path conservation weight. Also, it has been pointed out by Zheng and Brooks (2005) that T742 plays an important role in the hinge motion of the converter. In our work, we incorporate the $\gamma-$phosphate into the network graph and assign a node to represent it, just as C_α atom is used to represent each residue (see Methods section). Depending on the presence of the nucleotide, the start node might be $\gamma-$phosphate or $\gamma-$phosphate analog when ATP is present, or $\beta-$phosphate when ADP is present, or the equivalent 181S when the structure is nucleotide-free.

For comparison, Table 6.1 includes the paths obtained by Tang et al. (2007), who adopted the MSA (multiple sequence alignment) to derive allosteric pathways. The p-values in Table 6.1 are the results of the statistical test on the similarity of two paths: 10 out of 15 path pairs are highly statistically significant. For the rest 5 path pairs, the two methods give different results, suggesting there may exist an alternative communication route.

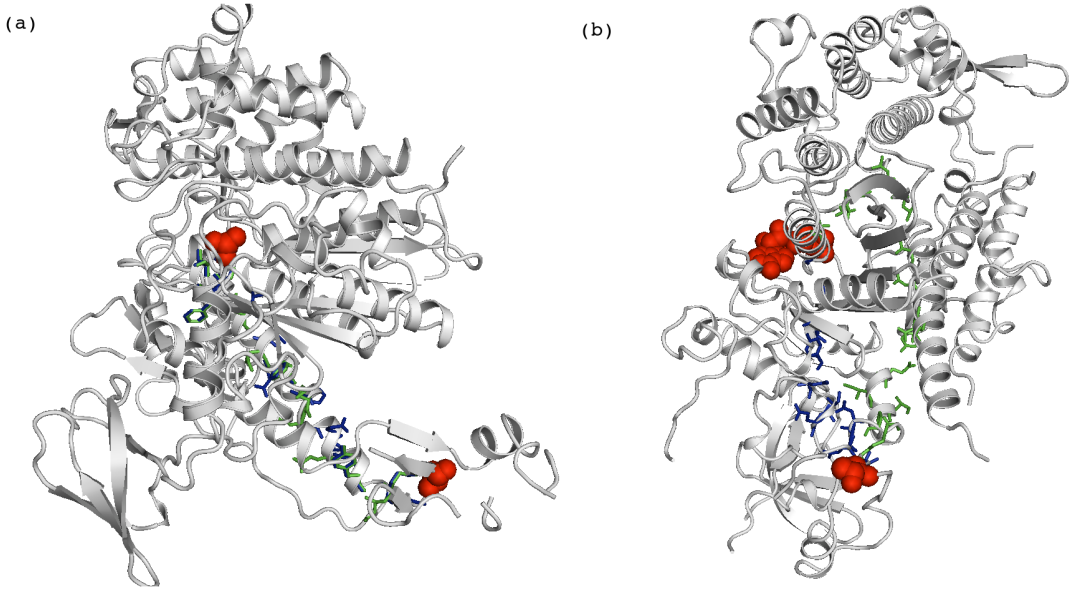


Figure 6.1 Allosteric communication paths of Myosin in two different states.

The allosteric communication paths derived from our MCN are colored blue and paths derived from MSA are green.The start and end points are marked as red beads in both subfigures. (a) is the prestroke state (1VOM) and (b) is the post-rigor state (1MMA). 1VOM (a) shows high similarity between the two paths derived from MCN and MSA. The two paths in 1MMA (b) are divergent.

Figure 6.1.(a)-(b) are the comparisons of the paths derived from MCN and MSA in two different states. The paths in the prestroke state are very similar (see Figure 6.1.(a)). In the post-rigor state, the paths derived from the two methods are similar for some species (not shown), while for the others, such as Dictyostelium (pdb:1MMA), the paths are divergent (see Figure 6.1.(b)).

Table 6.1 The comparison of the allosteric communication paths of myosin family derived from MCN and MSA.

For comparison, we include the paths obtained by MSA (multiple sequence alignment) [Tang et al. (2007)]. In order to perform comparison between different species, the original residue numbers are mapped to the reference residue numbers that are obtained from the alignment with Dictyostelium myosin II from which a common numbering system for different species is defined. The p-value is the Fisher's exact test for the similarity of the two paths from the two different methods, MCN and MSA. A cutoff distance of 7.0 Å is used in determining residue contacts (see Methods section).

State	PDB ID	Method	Allosteric communication path from ATP binding site to the lever arm	P-value
Pre-stroke	1VOM	MCN	VO4 S181 G457 F458 I471 T474 K477 Q480 F482 H484 K488 E490 E492 Y494 E497 G740 T742	2.0E-10
		MSA	VO4 S181 (G457) F458 N475 Q479 F482 M486 F487 Q491 Y494 I499 E497 T742	
	1YV3	MCN	VO4 S181 G457 F458 N475 L478 F481 N483 H485 L489 Q491 Y494 E497 G740 T742	1.0E-10
		MSA	VO4 S181 (G457) F458 N475 Q479 F482 M486 F487 Q491 E493 E497 T742	
	1QVI	MCN	VO4 S181 G457 F458 I471 N472 T474 K477 Q479 F481 N483 H484 H485 K488 E490 Q491 E493 Y494 L495 I499 E497 I741 G740 T742	6.3E-09
		MSA	VO4 G457 N475 Q479 F482 M486 F487 Q491 Y494 E497 T742	
	1LKX	MCN	VO4 S181 S456 G457 C470 T474 E476 L478 Q480 F482 H484 F487 L489 Q491 E493 Y494 I499 T742	4.9E-08
		MSA	VO4 S181 G457 F458 E467 Q468 N472 N475 L478 Q491 E492 G498 F745 T742	
	1BR2	MCN	AIF4 G182 F458 E459 N472 Y473 N475 L478 Q479 F481 F482 H484 H485 M486 K488 E490 Q491 E492 Y494 L495 K496 K498 I499 T742	4.3E-08
		MSA	AIF4 F458 E476 Q479 F482 M486 Q491 Y494 G498 T742	
	1W9J	MCN	AIF4 S181 G457 F458 I471 T474 K477 Q479 F481 N483 H485 K488 E490 E492 Y494 E497 I741 T742	1.0E-10
		MSA	AIF4 S181 (G457) F458 N475 Q479 F482 M486 F487 Q491 E493 E497 T742	
Post-rigor	1MMA	MCN	ADP G184 C655 V124 Y118 D90 S92 E93 N694 R695 F745 I744 T742	0.26
		MSA	ADP T186 S237 R238 L263 L262 I471 N475 Q479 Y573 E683 I685 I687 G691 P693 I744 K743 T742	
	1MMD	MCN	BeF3 A183 C655 V124 Y118 D90 M91 E93 S95 A748 R747 F746 F745 I744 T742	1.00
		MSA	BeF3 K185 G179 G457 N475 Q479 F482 M486 E490 Y494 E497 T742	
	1W7J	MCN	BeF3 E180 G457 F458 N475 K477 Q479 F481 F482 H484 M486 K488 E490 Q491 E493 L495 E497 G740 T742	1.4E-08
		MSA	BeF3 K185 D454 I455 G457 E459 N472 E476 N475 Q479 F482 V486 E490 Y494 E497 T742	
	1W9I	MCN	BeF3 G182 A183 I657 P658 E668 V671 D674 L676 G680 L682 E683 R686 I687 T688 R689 P693 F746 F745 I744 K743 T742	0.36
		MSA	BeF3 K185 E180 S181 E459 N472 E476 N475 Q479 F482 M486 E490 Y494 K498 E497 T742	
	1FMW	MCN	ATP A183 C655 V124 Y118 D90 M91 E93 S95 A748 R747 F746 F745 I744 T742	1.00
		MSA	ATP K185 D454 I455 S456 N475 Q479 L478 F482 M486 E490 Y494 E497 T742	
	1KK7	MCN	S181 E180 G457 F458 N472 N475 K477 Q479 F481 N483 H485 F487 L489 Q491 E493 L495 E497 G740 T742	7.0E-10
		MSA	K185 E180 E459 E457 (F458) N475 Q479 F482 M486 E490 Y494 E497 T742	
	1FMV	MCN	S181 A183 C655 V124 Y118 D90 M91 E93 S95 A748 R747 F746 F745 K743 T742	1.00
		MSA	K185 G179 G457 N475 Q479 F482 M486 E490 Y494 E497 T742	
Rigor	2AKA	MCN	S181 E180 G457 F458 N475 K477 Q480 N483 H484 F487 L489 Q491 E493 L495 E497 G740 T742	4.5E-08
		MSA	K185 G184 G182 S181 N233 N235 S236 R238 E459 N472 N475 Q479 F482 M486 E490 Y494 I499 E497 T742	
	1OE9	MCN	S181 E180 G457 F458 N475 K477 Q479 F481 N483 H485 M486 K488 E490 E492 E493 L495 E497 I741 T742	1.0E-10
		MSA	K185 G184 G182 S181 S237 R238 E459 F458 G457 N475 Q479 L478 F482 V486 E490 E493 E497 T742	

The allosteric pathways in pre-stroke and rigor structures are overall similar across all species. First, the communication starts from the $\gamma-$phosphate and travels through P loop (179-186) and then connects to switch II (454-459). It indicates that the motions between P loop and switch II are highly correlated. After the message passes through switch II, it connects to relay helix (466-518). At the final stage, the communication jumps from relay helix directly to the end point in the converter. Figure 6.2.(a) shows the overlapped paths of the prestroke states among different species and the paths are highly similar across different species.

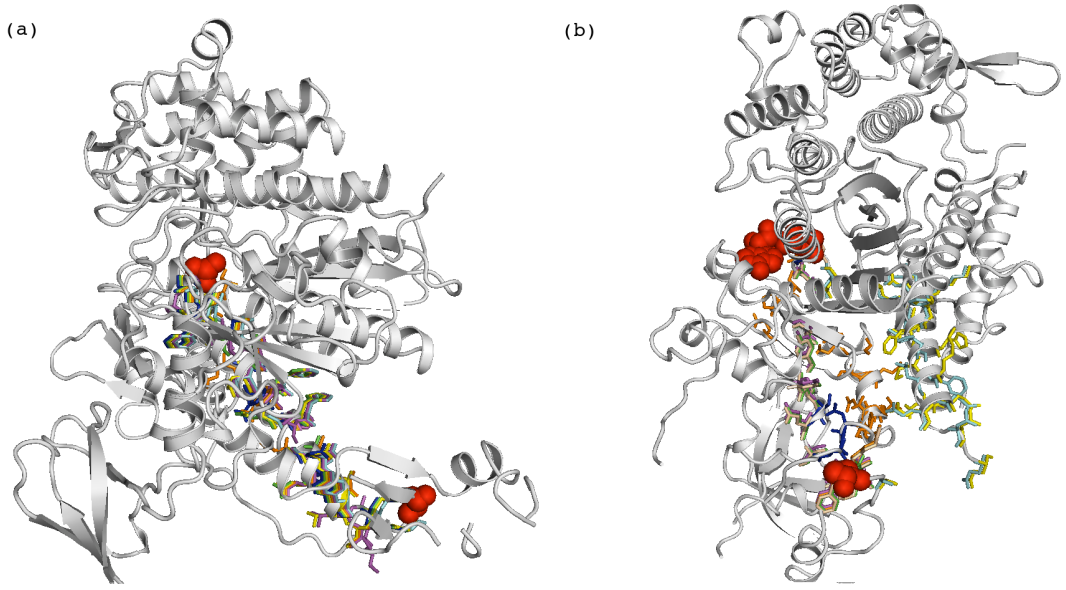


Figure 6.2 The overlapped paths of the prestroke and post-rigor states derived from MCN for myosin family.

(a) is the overlapped paths of the prestroke state and (b) is the overlapped paths of the post-rigor state. The start and end points are marked as red in both subfigures. The pre-stroke state shows high consistency among the paths of different species. The post-rigor state exhibits two different communication routes among the paths of different species.

The allosteric pathways in the post-rigor state are more divergent. It shows two types of paths in the post-rigor state. The first type is similar to the path obtained in the pre-stroke and rigor states. The allosteric communications of Chicken myosin V with ADP-BeF3 (1W7J) and nucleotide-free scallop myosin II (1KK7) belong to the first type. The second type of path goes through an interface (90-124) between N-Terminal subdomain and SH1 he-

lix and jumps from N-terminal subdomain to the converter. The Dictyostelium myosin II (1MMA,1MMD,1FMW,1FMV) at post-rigor state appears to have the second type of communication path. Figure 6.2 shows the overlapped paths among different species and the paths in the post-rigor state (Figure 6.2.(b)) exhibit two different routes. Therefore, we speculate that probably two pathways exist in the post-rigor state among different species. In our study, we find that when the link between the P loop (179-186) and switch II (454-459) becomes disconnected due to the stroke movement in some species, the communications between the $\gamma-$phosphate and the lever arm might have to go through a complete new route.

Further analysis of the difference between the MCN and MSA derived paths in the post-rigor state, reveals that the difference indeed comes from the breakage of the interface between P loop and switch II subdomain. When the gap between the P loop and switch II subdomain becomes larger and the message cannot pass the gap directly, the MSA derived path go through a small extra detour and connect back to the swith II subdomain,but the MCN derived path go through a complete different route which will bypass the switch II. While it is difficult to conclude which one is more accurate, the results do indicate that the allosteric communication paths are more prone to change in the post-rigor state.

In a recent study of the allosteric communications in Myosin V, Cecchini et al. (2008) proposed a transition pathway from the rigor to the post-rigor state by using a linear interpolation of the two states. Their study focused on the allosteric communication between the nucleotide binding site and the U50/L50 cleft, which is different from ours that is between the nucleotide binding site and the converter. Yet still some similarities can be found from the two studies. They found that only small fluctuations are present in the N, U50 and L50 subdomains except for the relay group (467-493) which is coupled to the converter. Our results also reveal the important role of the relay helix. All the MCN derived paths in the rigor state contain residues from the relay helix. Cecchini et al. (2008) also pointed out that the hinge was located at residues 763-769 in the rigor state, but changed to 696-697 in the post-rigor state. Due to the different locations of the hinge, they suggested that the converter subdomain might be more independent of the head in the post-rigor state than the rigor. This difference may explain why

two possible communication pathways exist in the post-rigor state, since some of the second type paths in the post-rigor state contain the residues near the hinge.

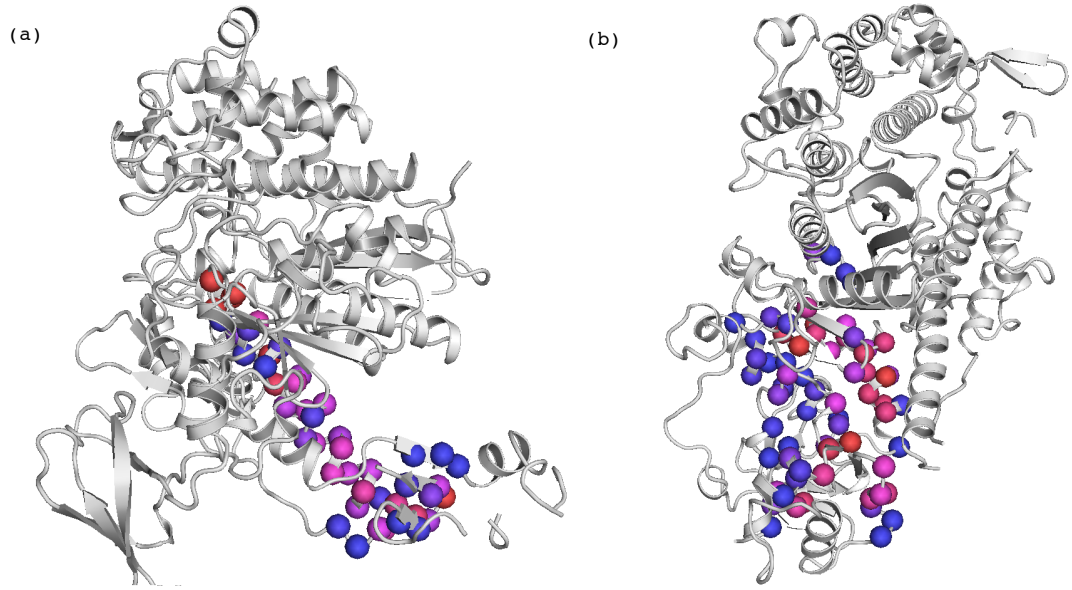


Figure 6.3 The allosteric communication path ensembles of the prestroke and post-rigor states.

(a) The allosteric communication path ensemble derived from the top 50000 shortest paths of MCN in the prestroke state (1VOM). (b) The allosteric communication path ensemble derived from the top 50000 shortest paths of MCN in the post-rigor state (1MMD). The residues are colored according to their frequency in the top 50000 shortest paths and the color scale is BMR. Blue (B) means less frequent, red (R) means more frequent and magenta (M) is somewhere in the middle.

The allosteric communication path ensemble

We also performed the k shortest path search on MCN. We set the k=50000 and obtained the path ensembles for the prestroke and post-rigor states. Figure 6.3.(a) displays the path ensemble of the prestroke state and Figure 6.3.(b) the path ensemble of the post-rigor state. The area visited by the top 50000 shortest paths is much narrower in the pre-stroke state than in the post-rigor state, again suggesting the communication paths in the post-rigor state should be more divergent.

After further examination of the 1VOM path ensemble, we found that F458, F482, I499 and their nearby residues are visited more frequently and these residues has been identified to

be important mutant sites. The mutant F458A has been showed to have a large impact on ATPase activity [Sasaki et al. (1998)] and the mutant F482A can increase the actin affinity in the presence of ATP [Ito et al. (2003)]. The mutant I499A does not change the kinetic of ATP hydrolysis, but it completely loses the motor functions [Sasaki et al. (2003)]. As the path ensemble of 1MMD, we found that G680 and G684 and their nearby residues are visited more frequently and the mutations of G680A and G684A also have been showed to have effects on the Pi release rate and nucleotide binding and release kinetics [Ito et al. (2003); Batra et al. (1999)].

The conservation of intramolecular communication pathway

Ten out of fifteen myosin structures have similar pathways when comparing the pathways obtained from MCN and MSA. The p-values of these 10 structures are less than 10^{-8}, which means that the similarities between the pathways from the two methods in these 10 structures are statistically very significant. It's remarkable that the results of such two completely different approaches can reach such high consistency. Since the paths from MSA are composed of residues that are highly conserved, it implies that the motion correlation is related to the evolutionary conservation, and sequence conservation may have a dynamics origin.

The intramolecular communication pathway of thrombin

Thrombin is a serine protease and have received lots of attention due to its importance in homeostasis. Two forms, fast and slow, exist at the equilibrium state. The slow form is Na^+-free and is responsible for anticoagulation. The fast form is Na^+-bound and is responsible for coagulation [Enrico et al. (2007)]. Two major allosteric pathways exist in the thrombin molecule. One relates to the binding of Na^+, which promotes the activity of procoagulant. The other involves the exosite I and the binding of thrombomodulin to exosite I promotes the activity of anticoagulant, protein C [Gandhi et al. (2008)]. Recently an allosteric communication pathway between the exosite I and the active site was revealed by the work of Gandhi et al. (2008). They proposed a plausible pathway which is consistent with the documented electron density map changes of thrombin mutant D102N. The free form of

thrombin mutant D102N (PDB ID code 3BEI) is stabilized in an inhibited state. The binding of protease activated receptor PAR1 fragment to exosite I causes a conformational change to the active site that is 30 Å away from the exosite I.

The allosteric communication pathway that Gandhi et al. (2008) proposed comprises four layers. The residues of the first layer are in direct contact with PAR1 fragment and they are F34 and R73. The second layer contains M32 and Q151. The third layer comprises the interactions between two $\beta-$strands,141-146 and 191-193, and involves W141, N143 and E192. The final layer contains a disulfide bonded C191 and C220 and E146, where the interactions are transmitted to the active site.

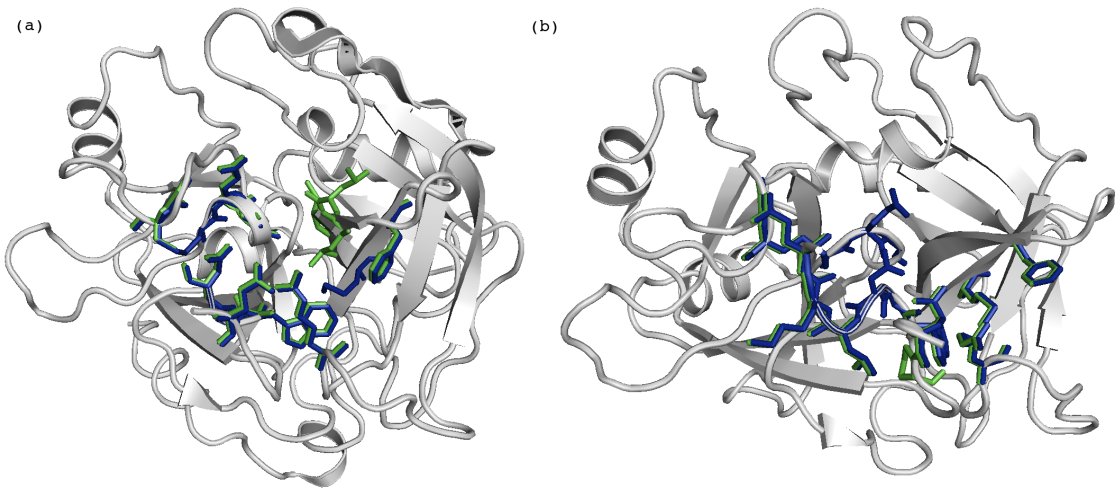


Figure 6.4 The allosteric communication paths of thrombin in two different states.

(a)The allosteric communication path in the inactive state (3BEI). (b)The allosteric communication path in the active state (3BEF). The comparison of allosteric communication paths derived from MCN (blue) and electron density map changes (green) in the two states shows the paths are highly similar.

The agreements between the MCN intramolecular communication pathway and electron density map

We studied the allosteric communication pathways of thrombin mutant D102N using MCN and found that the pathways elicited by MCN using atomic level contacts with cutoff distance set to 4.0 Å are similar to the pathway obtained by Gandhi et al. (2008). Table 6.2 shows the

Table 6.2 The allosteric communication paths of thrombin mutant D102N

The allosteric communication paths of thrombin mutant D102N between exosite I and the active site derived from MCN using atomic level contacts with cutoff distance set to 4.0 Å.

State	Allosteric communication paths between exosite I and the active site	P-Value
Inactive	F34 L41 C42 G193 E192 G219	9.5E-05
	F34 L40 W141 N143 C220	
	R73 Q151 N143 C220 G219	
	R73 Q151 N143 C220	
Active	F34 M32 W141 Q151 L144 K145 E146 G219	1.1E-05
	F34 M32 W141 Q151 L144 K145 E146 C220	
	R73 P152 L144 K145 E146 G219	
	R73 P152 L144 K145 E146 C220	

allosteric communication paths of thrombin mutant D102N between exosite I and active site derived from our MCN model and Figure 6.4 shows the good agreement between paths obtained from our method and those obtained by studying electron density map changes [Gandhi et al. (2008)].

The allosteric communication starts from F34 and reaches the C220 loop via L40 and the 141-146 $\beta-$strand and then reaches the 215-219 $\beta-$strand via L41, C42 and 191-193 $\beta-$strand. The other communication path starts from R73 and reaches C220 and the 215-219 $\beta-$strand via Q151 and 141-146 $\beta-$strand. Basically, there are four important components in the allosteric communication paths of thrombin mutant D102N, the M32 residue, the Q151 residue, the 141-146 $\beta-$strand and 191-193 $\beta-$strand. Our approach successfully identifies three of them except the M32, for which we are able to successfully identify its closeby neighbors, L40 or L41, which is about 3.8 Å away from M32 when considering atomic level contacts.

Conclusions

We provide a simple and computationally inexpensive approach to identifying the putative allosteric communication pathways and path ensembles. We compare the intramolecular communication pathways derived from our approach with the pathways derived from the statistical analysis of sequence conservation and experimental data and find very good agreement. The excellent agreement supports our hypothesis that the allosteric communication between the

allosteric site and the catalytic site is through pathways of residues that have strongly correlated motions.

Several important features exist in our approach. First, the pathways derived from our MCN model are strongly conserved. Most of the pathways derived from MCN agree with the pathways derived from statistical analysis of sequence conservation and those pathways that do not agree also show high conservation characteristics. The link between dynamics and sequence conservation has also been pointed out by Zheng et al. (2006). Second, the pathways derived from MCN agree with the pathways derived from electron density map. Third, our approach can generate allosteric path ensembles, which may be more meaningful than single paths.

Among the available computational approaches to allosteric pathway prediction, our method has the lowest requirement. Sequence conservation approach requires at least a sufficient amount of homologous to extract statistically significant data. Molecular dynamic simulation is computationally expensive. Our approach requires only a single pdb structure file to obtain the putative intramolecular communication pathway or path ensemble and the results are comparable with those obtained from other approaches.

Although we have observed some relations among motions, allostery and sequence conservation in this work, the underlying mechanism of allostery is still not fully understood and requires further investigation.

Methods

As a network approach, Motion Correlation Network (MCN) consists of three parts: nodes, edges and weights. Each node represents one residue of the protein, the edges represent the contacts between the residues, and the weights are the motion correlation between two residues that are in contact. The complete approach will involve graph generation, edge weight derivation and network exploration.

Graph generation

To identify the putative set of residues that involve in the allosteric communication, we formulate the graph structure of the MCN based on the protein 3D structure. Each protein is model as an undirected weighted graph $G = (V, E, W)$. The nodes of the graph $V = v_i | i = 1, 2, ..., n$ represent the residues, the edges of the graph $E = e_{ij}$ represent the contacts between residue i and j, and the non-negative weight $W = w_{ij}$ represents the motion correlation between two residues. Two kinds of network graph, residue level and atomic level, are used. In both graphs, each vertex represents a residue and an edge between two vertices represents the contacts between the two residues. The difference between the two kinds of graph is on how the contacts are defined. In the residue level graph, we draw an edge between two vertices if the distance between the C_α atoms of the two residues are within 7 Å. In the atomic level graph, we draw an edge between two vertices if the distance between any two heavy atoms of the two residues is below 3.5 Å.

Edge weight derivation

The edge weights define the motion correlation between residue pairs. The calculation of the motion correlation is based on Gaussian Network Model (GNM) [Bahar et al. (1997)], and can be extended to other kinds of elastic network models, e.g., Anisotropic Network Model (ANM) [Atilgan et al. (2001)].

Gaussian network model

Gaussian Network Model (GNM) was first introduced in Bahar et al. (1997). Inspired by Tirion (1996) in atomic level normal mode analysis, Bahar et al. (1997) extend the model to residue level and obtained an acceptable agreement between the theoretical and experimental crystallographic B-factors. The Gaussian network model describes a 3D protein structure as a cluster of C_α atoms connected by harmonic springs within a certain cutoff distance. The cutoff distance between residue pairs is the only parameter of the model and normally is from 7 Å to 8 Å. Let ΔR_i and ΔR_j represent the instaneous fluctuations from equilibrium positions

of residue i and j and ΔR_{ij} the difference between the instaneous fluctuations of two residues. As an elastic network model, the GNM has the following harmonic potential (6.1).

$$V_{GNM} = \frac{\gamma}{2}\Delta R^T \Gamma \Delta R = \frac{\gamma}{2}\sum_{i<j}^{N}\Gamma_{ij}(\Delta R_i - \Delta R_j)^2 \tag{6.1}$$

where γ is the force constant and Γ_{ij} is the Laplacian matrix (Hessian matrix) which can be derived directly from the 3D structure and is defined as (6.2).

$$\Gamma_{ij} = \begin{cases} -1 & if \quad i \neq j \cap R_{ij} \leq r_c \\ 0 & if \quad i \neq j \cap R_{ij} > r_c \\ \sum_{j,j\neq i}^{N}\Gamma_{ij} & if \quad i = j \end{cases} \tag{6.2}$$

where i and j are the indices of the residues and r_c is the cutoff distance.

The simplicity of the Hessian matrix formulation results from the assumption that the fluctuations of each residue are isotropic and Gaussian distributed along the X, Y and Z directions. The expectation value of residue fluctuations, $< \Delta R_i^2 >$ (6.3) , and correlation, $< \Delta R_i \cdot \Delta R_j >$ (6.4), can be easily obtained from the inverse of the Laplacian matrix under the isotropic and Gaussian assumption.

$$< \Delta R_i^2 >= \frac{k_B T}{\gamma}(\Gamma^{-1})_{ii} \tag{6.3}$$

$$< \Delta R_i \cdot \Delta R_j >= \frac{k_B T}{\gamma}(\Gamma^{-1})_{ij} \tag{6.4}$$

The k_B is the Boltzmann constant and T is the temperature. The $< \Delta R_i^2 >$ term is directly related to crystallographic B-factors.

Motion correlations

The cross correlation between the fluctuations of residues i and j can be written as (6.5).

$$c_{ij} = \frac{< \Delta R_i \cdot \Delta R_j >}{\sqrt{< \Delta R_i^2 > \cdot < \Delta R_j^2 >}} \tag{6.5}$$

The c_{ij}'s form a $n \times n$ correlation matrix and n is the total number of residues of the protein. Now consider the i^{th} and j^{th} columns, r_i and r_j, and define the motion correlation as the

correlation of these two columns [Yesylevskyy et al. (2006)]:

$$mc_{ij} = \frac{(r_i - \overline{r_i})(r_j - \overline{r_j})'}{\sqrt{(r_i - \overline{r_i})(r_i - \overline{r_i})'}\sqrt{(r_j - \overline{r_j})(r_j - \overline{r_j})'}} \tag{6.6}$$

mc_{ij} thus describes the motion correlation between residues i and j. The edge weight w_{ij} is related to mc_{ij} by,

$$w_{ij} = -log(|mc_{ij}|), \tag{6.7}$$

which means two strongly (weakly) motion-correlated residues have a low (high) weight between them. By doing this, the search for pathways of residues that have strongly correlated motions becomes equivalent to finding the shortest paths.

Network exploration

Yen's algorithm (finding the first k shortest paths)

Dijkstra's algorithm [Dijkstra (1959)] allows us to find the shortest path between two nodes in a graph, but sometimes the second shortest path, the third shortest path,..., until the kth shortest path are within our interests. In the allosteric communication, the first k shortest paths might form a path ensemble which includes more information than a single shortest path solution. A path ensemble can reveal the degree of involvement of a given residue in the path ensemble, which is not available in a single path solution. In this work, we applied the Yen's algorithm [Yen (1971); Martins and Pascoal (2003)] for finding the top k shortest paths. Only loopless paths are determined by the Yen's algorithm. The Yen's algorithm ranks the k shortest paths between a pair of nodes by constructing a tree of paths. The complexity of the Yen's algorithm is $O(k|V|(|E| + |V|\log|V|))$, which takes $k|V|$ times longer than Dijkstra's algorithm. Since it's computationally more expensive than finding the shortest path, we only applied the Yen's algorithm to myosin.

The statistical significance of the derived paths (ensembles)

In order to evaluate the paths derived from MCN, we adopt the Fisher's exact test [Samuels and Witmer (2002)] to test the statistical significance of the consistency of the MCN derived

path and the compared path, the path derived from other methods , such as MSA in myosin family or electron density maps in thrombin. The Fisher's exact test allows us to examine whether the appearance of the same residues in both MCN derived path and the compared path is pure random or not. The null hypothesis is that the residues that appear in the MCN derived path cannot be found in the compared path and the alternative hypothesis is that the residues that appear in the MCN derived path can be found in the compared path. In short, that is the similarity between the MCN derived path and the compared path. The p-values are calculated from the total number of residues of the protein, the total number of residues of the compared path, the total number of residues of the MCN derived path (ensemble) and the total number of residues or their immediate sequence neighbors that appear in both the MCN derived path (ensemble) and the compared path. The smaller the p-value, the more likely the residues that appear in the MCN derived path can be found in the compared path and the more similar between these two paths.

Table 6.3 The Allosteric Communication Pathway Test Set.

Protein	Conformational State	PDB Code
Myosin [Tang et al. (2007)]	Pre-stroke	1VOM, 1YV3, 1QVI, 1BR2, 1W9J ,1LKX
	Post-rigor	1MMA, 1MMD, 1W7J, 1W9I, 1FMW, 1KK7,1FMW
	Rigor	2AKA, 1OE9
thrombin [Gandhi et al. (2008)]	Inactive Active	3BEI 3BEF

Testing data set

We have derived the allosteric communication pathways in two protein members: myosin and thrombin. The names and Protein Data Bank (PDB) [Berman et al. (2000)] codes of these proteins are given in Table 6.3. The testing data set is compiled based on two previous works: Tang et al. (2007) and Gandhi et al. (2008). The results of Tang et al. (2007) are from statistical analysis of evolutionary conservation and the results of Gandhi et al. (2008) are from polar interactions and are consistent with the changes documented in the electron density map. The reason for comparing the derived allosteric communication pathways from

different approaches is due to the lack of a standard benchmark in allosteric communication pathway identification.

List of abbreviations used

MCN: Motion Correlation Network

MSA: Multiple Sequence Alignment

ENM: Elastic Network Model

GNM: Gaussian Network Model

ANM: Anisotropic Network Model

Competing interests

The authors declare that they have no competing interests.

Authors contributions

Tu-Liang Lin collaborated with Guang Song on the ideas of the network construction from motions and applied to the prediction of the allosteric communication. Gunag Song conceived the idea of path ensemble and suggested the k-shortest path algorithm. Most of the implementation was done by Tu-Liang Lin under the supervision of Guang Song. Both authors read and approved the manuscript.

Acknowledgements

We thank Prof. Ernesto Martins and Prof. David Eppstein who generously made the k-shortest path source code available.

CHAPTER 7. CONCLUSION AND FUTURE RESEARCH

In this thesis, I have made several contributions to the field of computational biology, especially in the area of computational studies of protein dynamics and functional mechanisms. Since a better understanding of protein dynamics lends to a better understanding of the protein functions, studies on protein dynamics are presented first.

In my work on protein dynamics, two computational methods are developed to overcome some of the limitations found in molecular dynamics simulations or elastic network models. First, I introduce relative population into the description of ensembles and develop a novel computational method that is able to determine the relative populations of structures within an ensemble using the experimental RDC data. I compare Q-factors among several ubiquitin ensembles and the results show that ensembles with relative populations significantly improve the agreement between the calculated and experimental RDCs, implying that an ensemble with relative populations is more efficient and can better describe the native states while using fewer parameters. Therefore, such ensemble representation should be the choice in future ensemble determination using the still limited experimental data. Secondly, I develop a new coarse-grained model with multi-body potentials using generalized spring tensors. This generalized spring tensor model is able to integrate in a single model the attractive features of two widely-used models, ANM and GNM, and to overcome their limitations.

As the second half of this thesis work, I develop two novel computational approaches for studying the functional mechanisms of allosteric communication and ligand migration. The ligand migration pathways of myoglobin and cytochrome P450cam are determined from a set of conformational ensembles with rich dynamics. The allosteric communication pathways of myosin and thrombin are determined from the dynamics information derived from elastic

network models.

Two important contributions are made in the second part of this thesis. First, an efficient approach for mapping the whole ligand migration channel network is presented. In contrast to a trajectory-based MD simulation run that can produce only one single trajectory at a time, the developed approach is more efficient and maps the whole ligand migration channel network. Second, the excellent agreement between allosteric communication pathways derived in our studies and the pathways derived from multiple sequence alignments implies that sequence conservation may have a dynamics origin.

With the establishment of the aforementioned frameworks, future research directions are multiple, including but not limited to: (1)incorporate weighted Voronoi diagram related methods such as alpha shapes to obtain more precise mappings of the ligand migration networks; (2)utilize the newly developed RDC fitting method to construct more accurate conformational sampling methods; (3)extend the RDC fitting method to other experimental data, such as chemical shifts or order parameters; (4)apply the RDC fitting method to study disordered proteins.

BIBLIOGRAPHY

Amato, N. M. and Song, G. (2002). Using motion planning to study protein folding pathways. *Journal of Computational Biology*, 9(2):149–168.

Anselmi, M., Di Nola, A., and Amadei, A. (2008). The kinetics of ligand migration in crystallized myoglobin as revealed by molecular dynamics simulations. *Biophysical Journal*, 94(11):4277–81.

Apaydin, M. S. (2004). *Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion*. Phd, Stanford University, Stanford, CA.

Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D., Latombe, J. C., and Varma, C. (2003). Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. *Journal of Computational Biology*, 10(3-4):257–281.

Apaydin, M. S., Guestrin, C. E., Varma, C., Brutlag, D. L., and Latombe, J. C. (2002). Stochastic roadmap simulation for the study of ligand-protein interactions. *Bioinformatics*, 18:S18–S26.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80:505–515.

Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–73.

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 91:3589–3599.

Bahar, I., Chennubhotla, I., and Tobi, D. (2007). Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Current Opinion in Structural Biology*, 17:633–640.

Bahar, I. and Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*, 15(5):586–592.

Batra, R., Geeves, M. A., and Manstein, D. J. (1999). Kinetic analysis of dictyostelium discoideum myosin motor domains with glycine-to-alanine mutations in the reactive thiol region. *Biochemistry*, 38:6126–6134.

Bayazit, O. B., Song, G., and Amato, N. M. (2001). Ligand binding with obprm and user input. *2001 Ieee International Conference on Robotics and Automation, Vols I-Iv, Proceedings*, pages 954–959.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28:235–242.

Best, R. B., Lindorff-Larsen, K., DePristo, M. A., and Vendruscolo, M. (2006). Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(29):10901–6.

Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5(11):789–96.

Bondi, A. (1964). Van der waals volumes and radii. *Journal of Physical Chemistry*, 68(3):441–51.

Bossa, C., Amadei, A., Daidone, I., Anselmi, M., Vallone, B., Brunori, M., and Di Nola, A. (2005). Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped co on the structure and dynamics of cavities. *Biophysical Journal*, 89(1):465–74.

Bossa, C., Anselmi, M., Roccatano, D., Amadei, A., Vallone, B., Brunori, M., and Di Nola, A. (2004). Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophysical Journal*, 86(6):3855–3862.

Bourgeois, D., Schotte, F., Brunori, M., and Vallone, B. (2007). Time-resolved methods in biophysics. 6. time-resolved laue crystallography as a tool to investigate photo-activated protein dynamics. *Photochemical & Photobiological Sciences*, 6(10):1047–56.

Bourgeois, D., Vallone, B., Arcovito, A., Sciara, G., Schotte, F., Anfinrud, P. A., and Brunori, M. (2006). Extended subnanosecond structural dynamics of myoglobin revealed by laue crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):4924–4929.

Bourgeois, D., Vallone, B., Schotte, F., Arcovito, A., Miele, A. E., Sciara, G., Wulff, M., Anfinrud, P., and Brunori, M. (2003). Complex landscape of protein structural dynamics unveiled by nanosecond laue crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8704–9.

Burra, P. V., Zhang, Y., Godzik, A., and Stec, B. (2009). Global distribution of conformational states derived from redundant models in the pdb points to non-uniqueness of the protein structure. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10505–10.

Canutescu, A. A. and Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12(5):963–972.

Case, D. A. and Karplus, M. (1979). Dynamics of ligand binding to heme proteins. *Journal of Molecular Biology*, 132(3):343–68.

Cecchini, M., Houdusse, A., and Karplus, M. (2008). Allosteric communication in myosin v: from small conformational changes to large directed movement. *PLoS Computational Biology*, 4:e1000129.

Chennubhotla, C. and Bahar, I. (2006). Markov propagation of allosteric effects in biomolecular systems: application to groel-groes. *Molecular System Biology*, 2(36):1–13.

Chiang, T. H., Apaydin, M. S., Brutlag, D. L., Hsu, D., and Latombe, J. C. (2007). Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: Folding rates and phi-values. *Journal of Computational Biology*, 14(5):578–593.

Chirikjian, G. S. (2011). Modeling loop entropy. *Methods Enzymol*, 487:99–132.

Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., and Swope, W. C. (2007). Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *Journal of Chemical Physics*, 126(15):155101.

Choset, H. and Burdick, J. (2000). Sensor-based exploration: The hierarchical generalized voronoi graph. *International Journal of Robotics Research*, 19(2):96–125.

Choset, H. M., Lynch, K. M., Hutchinson, S., Kantor, G., Burgard, W., Kavraki, L. E., and Thrun, S. (2005). *Principles of robot motion : theory, algorithms, and implementation*. MIT Press, Cambridge, Mass.

Choy, W. Y. and Forman-Kay, J. D. (2001). Calculation of ensembles of structures representing the unfolded state of an sh3 domain. *Journal of Molecular Biology*, 308:1011–1032.

Chu, K., Vojtchovsky, J., McMahon, B. H., Sweet, R. M., Berendzen, J., and Schlichting, I. (2000). Structure of a ligand-binding intermediate in wild-type carbonmonoxy myoglobin. *Nature*, 403(6772):921–3.

Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000). Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of Molecular Biology*, 298:937–953.

Cohen, J., Arkhipov, A., Braun, R., and Schulten, K. (2006). Imaging the migration pathways for o2, co, no, and xe inside myoglobin. *Biophysical Journal*, 91(5):1844–57.

Cornilescu, G., Marquardt, J. L., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *Journal of the American Chemical Society*, 120(27):6836–6837.

Daily, M. D., Upadhyaya, T. J., and Gray, J. J. (2007). Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Structure, Function, and Bioinformatics*, 71:455–466.

De Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. (1997). *Computational geometry : algorithms and applications*. Springer, Berlin ; Heidelberg ; New York.

de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins : Structure, Function, and Bioinformatics*, 29(2):240–51.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Edelsbrunner, H., Facello, M., and Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88(1-3):83–102.

Edelsbrunner, H. and Mucke, E. P. (1994). 3-dimensional alpha-shapes. *ACM Transactions on Graphics*, 13(1):43–72.

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–21.

Elber, R. and Karplus, M. (1990). Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *Journal of the American Chemical Society*, 112:9161–9175.

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., and Woodhull, G. (2004). Graphviz

and dynagraph - static and dynamic graph drawing tools. *Graph Drawing Software*, pages 127–148 378.

Enrico, D. C., Michael, J. P., Alaji, B., Leslie, A. B., and Laura, C. G. (2007). Thrombin allostery. *Physical Chemistry Chemical Physics*, 9:1292–1306.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. W. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E., Han, J. H., and Fayyad, U. M., editors, *the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press.

Feng, Y., Kloczkowski, A., and Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins : Structure, Function, and Bioinformatics*, 68:57–66.

Fetler, L., Kantrowitz, E. R., and Vachette, P. (2007). Direct observation in solution of a preexisting structural equilibrium for a mutant of the allosteric aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences of the United States of America*, 104:495–500.

Finney, J. L. (1975). Volume occupation, environment and accessibility in proteins. the problem of the protein surface. *Journal of Molecular Biology*, 96(4):721–32.

Fisher, C. K., Huang, A., and Stultz, C. M. (2010). Modeling intrinsically disordered proteins with bayesian statistics. *Journal of the American Chemical Society*, 132(42):14919–27.

Frauenfelder, H., Chen, G., Berendzen, J., Fenimore, P. W., Jansson, H., McMahon, B. H., Stroe, I. R., Swenson, J., and Young, R. D. (2009). A unified model of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5129–34.

Frauenfelder, H., McMahon, B. H., Austin, R. H., Chu, K., and Groves, J. T. (2001). The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of

myoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 98(5):2370–4.

Friedland, G. D. and Kortemme, T. (2010). Designing ensembles in conformational and sequence space to characterize and engineer proteins. *Current Opinion in Structural Biology*, pages 377–84.

Gandhi, P. S., Chen, Z., Mathews, F. S., and Cera, E. D. (2008). Structural identification of the pathway of long-range communication in an allosteric enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1832–1837.

Gether, U. (2000). Uncovering molecular mechanisms involved in activation of g protein-coupled receptors. *Endocrine Reviews*, 21:90–113.

Gu, J. and Bourne, P. E. (2007). Identifying allosteric fluctuation transitions between different protein conformational states as applied to cyclin dependent kinase 2. *BMC Bioinformatics*, 8(45):1–13.

Hammes-Schiffers, S. and Benkovic, S. J. (2006). Relating protein motion to catalysis. *Annual Review of Biochemistry*, 75:519–541.

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins : Structure, Function, and Bioinformatics*, 33:417–429.

Houdusse, A., Szent-Gyorgyi, A. G., and Cohen, C. (2000). three conformational states of scallop myosin s1. *Proceedings of the National Academy of Sciences of the United States of America*, 97:11238–11243.

Huang, X. and Boxer, S. G. (1994). Discovery of new ligand binding pathways in myoglobin by random mutagenesis. *Nature Structural & Molecular Biology*, 1(4):226–9.

Hummer, G., Schotte, F., and Anfinrud, P. A. (2004). Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15330–4.

Ito, K., Uyeda, T. Q., Suzuki, Y., Sutoh, K., and Yamamoto, K. (2003). Requirement of domain-domain interaction for conformational change and functional atp hydrolysis in myosin. *Journal of Biological Chemistry*, 278:31049–31057.

Jensen, M. R., Salmon, L., Nodet, G., and Blackledge, M. (2010). Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *Journal of the American Chemical Society*, 132(4):1270–2.

Kachalova, G. S., Popov, A. N., and Bartunik, H. D. (1999). A steric mechanism for inhibition of co binding to heme proteins. *Science*, 284(5413):473–6.

Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652.

Kavraki, L. E., Svestka, P., Latombe, J. C., and Overmars, M. H. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, 12(4):566–580.

Kazerounian, K., Latif, K., Rodriguez, K., and Alvarado, C. (2005). Nano-kinematics for analysis of protein molecules. *Journal of Mechanical Design*, 127(4):699–711.

Kern, D. and Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, 13:748–757.

Kim, M. K., Jernigan, R. L., and Chirikjian, G. S. (2005). Rigid-cluster models of conformational transitions in macromolecular machines and assemblies. *Biophysical Journal*, 89(1):43–55.

Kimmel, J. L. and Reinhart, G. D. (2000). Reevaluation of the accepted allosteric mechanism of phosphofructokinase from bacillus stearothermophilus. *Proceedings of the National Academy of Sciences of the United States of America*, 97:3844–3849.

Koga, N. and Takada, S. (2001). Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. *Journal of Molecular Biology*, 313:171–180.

Koga, N. and Takada, S. (2006). Folding-based molecular simulations reveal mechanisms of the rotary motor f1-atpase. *Proceedings of the National Academy of Sciences of the United States of America*, 103:5367–372.

Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. (2005). Inverse kinematics in biology: The protein loop closure problem. *International Journal of Robotics Research*, 24(2-3):151–163.

Kontaxis, G. and Bax, A. (2001). Multiplet component separation for measurement of methyl 13c-1h dipolar couplings in weakly aligned proteins. *Journal of Biomolecular NMR*, 20(1):77–82.

Koshland, D. E., Nemethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5:365–385.

Kundu, S., Melton, J. S., Sorensen, D. C., and Phillips Jr, G. N. (2002). Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophysical Journal*, 83:723–732.

Kundu, S., Sorensen, D. C., and Phillips Jr, G. N. (2007). Automatic domain decomposition of proteins by a gaussian network model. *Proteins : Structure, Function, and Bioinformatics*, 57:725–733.

Kuriyan, J., Osapay, K., Burley, S. K., Brunger, A. T., Hendrickson, W. A., and Karplus, M. (1991). Exploration of disorder in protein structures by x-ray restrained molecular-dynamics. *Proteins-Structure Function and Genetics*, 10(4):340–358.

Lakomek, N. A., Carlomagno, T., Becker, S., Griesinger, C., and Meiler, J. (2006). A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR*, 34(2):101–15.

Lakomek, N. A., Walter, K. F., Fares, C., Lange, O. F., de Groot, B. L., Grubmuller, H., Bruschweiler, R., Munk, A., Becker, S., Meiler, J., and Griesinger, C. (2008). Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. *J Biomol NMR*, 41(3):139–55.

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–5.

Lange, O. F., van der Spoel, D., and de Groot, B. L. (2010). Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophysical Journal*, 99(2):647–55.

Latombe, J.-C. (1991). *Robot motion planning*. Kluwer Academic Publishers, Boston.

LaValle, S. M. (2006). *Planning algorithms*. Cambridge University Press, Cambridge ; New York.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*. Classics in applied mathematics 15. SIAM, Philadelphia.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998a). Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins-Structure Function and Bioinformatics*, 33(1):1–17.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998b). Analytical shape computation of macromolecules: II. inaccessible cavities in proteins. *Proteins-Structure Function and Bioinformatics*, 33(1):18–29.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998c). Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, 7(9):1884–1897.

Lin, T. and Song, G. (2009). Predicting allosteric communication pathways using motion correlation network. In *Proceedings of the 7th Asia Pacific Bioinformatics Conference (APBC)*, pages 588–598. Tsinghua University, China.

Lin, T. and Song, G. (2011). Determine the populations of protein conformation states using experimental residual dipolar coupling data. *submitted*.

Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–32.

Lu, M., Poon, B., and Ma, J. (2006). A new method for coarse-grained elastic normal-mode analysis. *Journal of Chemical Theory and Computation*, 2:464–471.

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373–380.

Manocha, D., Zhu, Y. S., and Wright, W. (1995). Conformational-analysis of molecular chains using nano-kinematics. *Computer Applications in the Biosciences*, 11(1):71–86.

Martins, E. and Pascoal, M. (2003). A new implementation of yen?s ranking loopless paths algorithm. *4OR: A Quarterly Journal of Operations Research*, 1:121–133.

Mendez, R. and Bastolla, U. (2010). Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Physical Review Letters*, 104(22):228103.

Ming, D. and Brschweiler, R. (2006). Reorientational contact-weighted elastic network model for the prediction of protein dynamics: Comparison with NMR relaxation. *Biophysical Journal*, 90:3382–388.

Mittermaier, A. and Kay, L. E. (2006). New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–8.

Monod, J., Changeux, J. P., and Jacob, F. (1963). Allosteric proteins and cellular control system. *Journal of Molecular Biology*, 6:306–329.

Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *Journal of Molecular Biology*, 12:88–118.

Mouawad, L., Tetreau, C., Abdel-Azeim, S., Perahia, D., and Lavalette, D. (2007). Co migration pathways in cytochrome p450(cam) studied by molecular dynamics simulations. *Protein Science*, 16(5):781–794.

Nienhaus, K., Ostermann, A., Nienhaus, G. U., Parak, F. G., and Schmidt, M. (2005). Ligand migration and protein fluctuations in myoglobin mutant l29w. *Biochemistry*, 44(13):5095–105.

Noonan, K., O'Brien, D., and Snoeyink, J. (2005). Probik: Protein backbone motion by inverse kinematics. *International Journal of Robotics Research*, 24(11):971–982.

Nutt, D. R. and Meuwly, M. (2004). Co migration in native and mutant myoglobin: atomistic simulations for the understanding of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):5998–6002.

Olson, J. S., Mathews, A. J., Rohlfs, R. J., Springer, B. A., Egeberg, K. D., Sligar, S. G., Tame, J., Renaud, J. P., and Nagai, K. (1988). The role of the distal histidine in myoglobin and haemoglobin. *Nature*, 336(6196):265–6.

Ostermann, A., Waschipky, R., Parak, F. G., and Nienhaus, G. U. (2000). Ligand binding and conformational motions in myoglobin. *Nature*, 404(6774):205–8.

Ottiger, M. and Bax, A. (1998). Determination of relative n-h-n n-c', c-alpha-c', andc(alpha)-h-alpha effective bond lengths in a protein by NMR in a dilute liquid crystalline phase. *Journal of the American Chemical Society*, 120(47):12334–12341.

Park, S. J., Kufareva, I., and Abagyan, R. (2010). Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *Journal of Computer-Aided Molecular Design*.

Petrek, M., Kosinova, P., Koca, J., and Otyepka, M. (2007). Mole: a voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, 15(11):1357–63.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of Molecular Biology*, 82(1):1–14.

Richter, B., Gsponer, J., Varnai, P., Salvatella, X., and Vendruscolo, M. (2007). The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *Journal of Biomolecular NMR*, 37(2):117–35.

Ringe, D., Petsko, G. A., Kerr, D. E., and Ortiz de Montellano, P. R. (1984). Reaction of myoglobin with phenylhydrazine: a molecular doorstop. *Biochemistry*, 23(1):2–4.

Rousseau, F. and Schymkowitz, J. (2005). A systems biology perspective on protein structural dynamics and signal transduction. *Current Opinion in Structural Biology*, 15:23–30.

Ruan, K. and Tolman, J. R. (2005). Composite alignment media for the measurement of independent sets of NMR residual dipolar couplings. *Journal of the American Chemical Society*, 127(43):15032–15033.

Ruscio, J. Z., Kumar, D., Shukla, M., Prisant, M. G., Murali, T. M., and Onufriev, A. V. (2008). Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27):9204–9.

Samuels, M. L. and Witmer, J. A. (2002). *Statistics for the life sciences*. Prentice Hall, Upper Saddle River, New Jersey.

Sasaki, N., Ohkura, R., and Sutoh, K. (2003). Dictyostelium myosin II mutations that uncouple the converter swing and atp hydrolysis cycle. *Biochemistry*, 42:90–95.

Sasaki, N., Shimada, T., and Sutoh, K. (1998). Mutational analysis of the switch II loop of dictyostelium myosin II. *Journal of Biological Chemistry*, 273:20334–20340.

Scheek, R. M., Torda, A. E., Kemmink, J., and Vangunsteren, W. F. (1991). Structure determination by NMR - the modeling of NMR parameters as ensemble averages. *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*, 225:209–217.

Schmidt, M., Nienhaus, K., Pahl, R., Krasselt, A., Anderson, S., Parak, F., Nienhaus, G. U., and Srajer, V. (2005). Ligand migration pathway and protein dynamics in myoglobin: a time-resolved crystallographic study on l29w mbco. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11704–9.

Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, J. S., Phillips, G. N., J., Wulff, M., and Anfinrud, P. A. (2003). Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300(5627):1944–7.

Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1.

Schuyler, A. D., Jernigan, R. L., Qasba, P. K., Ramakrishnan, B., and Chirikjian, G. S. (2009). Iterative cluster-nma: A tool for generating conformational transitions in proteins. *Proteins : Structure, Function, and Bioinformatics*, 74(3):760–76.

Scott, E. E. and Gibson, Q. H. (1997). Ligand migration in sperm whale myoglobin. *Biochemistry*, 36(39):11909–17.

Scott, E. E., Gibson, Q. H., and Olson, J. S. (2001). Mapping the pathways for o2 entry into and exit from myoglobin. *The Journal of Biological Chemistry*, 276(7):5177–88.

Seeliger, D. and De Groot, B. L. (2009). tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules. *Journal of Computational Chemistry*, 30(7):1160–6.

Seeliger, D., Haas, J., and de Groot, B. L. (2007). Geometry-based sampling of conformational transitions in proteins. *Structure*, 15(11):1482–92.

Shehu, A., Clementi, C., and Kavraki, L. E. (2006). Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins : Structure, Function, and Bioinformatics*, 65(1):164–79.

Shehu, A., Kavraki, L. E., and Clementi, C. (2009). Multiscale characterization of protein conformational ensembles. *Proteins : Structure, Function, and Bioinformatics*, 76(4):837–51.

Shirakihara, Y. and Evans, P. R. (1988). Crystal structure of the complex of phosphofructokinase from escherichia coli with its reaction products. *Journal of Molecular Biology*, 204:973–994.

Singhal, N., Snow, C. D., and Pande, V. S. (2004). Using path sampling to build better markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *Journal of Chemical Physics*, 121(1):415–25.

Song, G. (2003). *A Motion Planning Approach to Protein Folding*. Phd, Texas A&M University, College Station, TX.

Song, G. and Jernigan, R. L. (2006). An enhanced elastic network model to represent the motions of domain-swapped proteins. *Proteins : Structure, Function, and Bioinformatics*, 63:197–09.

Song, G. and Jernigan, R. L. (2007). vGNM: a better model for understanding the dynamics of proteins in crystals. *Journal of Molecular Biology*, 369:880–93.

Srajer, V., Ren, Z., Teng, T. Y., Schmidt, M., Ursby, T., Bourgeois, D., Pradervand, C., Schildkamp, W., Wulff, M., and Moffat, K. (2001). Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved laue x-ray diffraction. *Biochemistry*, 40(46):13802–15.

Srajer, V., Teng, T., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M., and Moffat, K. (1996). Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. *Science*, 274(5293):1726–9.

Suel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10:59–69.

Swain, J. F. and Gierasch, L. M. (2006). The changing landscape of protein allostery. *Current Opinion in Structural Biology*, 16:102–108.

Tama, F. and Sanejouand, Y. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Engineering Design and Selection*, 14:1–6.

Tang, S., Liao, J. C., Dunn, A. R., Altman, R. B., Spudich, J. A., and Schmidt, J. P. (2007). Predicting allosteric communication in myosin via a pathway of conserved residues. *Journal of Molecular Biology*, 273:1361–1373.

Teeter, M. M. (2004). Myoglobin cavities provide interior ligand pathway. *Protein Sci*, 13(2):313–8.

Teng, T. Y., Srajer, V., and Moffat, K. (1997). Initial trajectory of carbon monoxide after photodissociation from myoglobin at cryogenic temperatures. *Biochemistry*, 36(40):12087–100.

Tetreau, C., Blouquit, Y., Novikov, E., Quiniou, E., and Lavalette, D. (2004). Competition with xenon elicits ligand migration and escape pathways in myoglobin. *Biophysical Journal*, 86(1 Pt 1):435–47.

Thomas, S., Song, G., and Amato, N. M. (2005). Protein folding by motion planning. *Physical Biology*, 2(4):S148–S155.

Thomas, S., Tang, X. Y., Tapia, L., and Amato, N. M. (2007). Simulating protein motions with rigidity analysis. *Journal of Computational Biology*, 14(6):839–855.

Thorpe, M. F. (2007). Comment on elastic network models and proteins. *Physical Biology*, 4:60–3.

Tilton, R. F., J., Kuntz, I. D., J., and Petsko, G. A. (1984). Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 a. *Biochemistry*, 23(13):2849–57.

Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, 77:1905–908.

Tobi, D. and Bahar, I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proceedings of the National Academy of Sciences of the United States of America*, 102:18908–18913.

Tolman, J. R. (2002). A novel approach to the retrieval of structural and dynamic information from residual dipolar couplings using several oriented media in biomolecular NMR spectroscopy. *Journal of the American Chemical Society*, 124(40):12020–30.

van den Bedem, H., Lotan, I., Latombe, J. C., and Deacon, A. M. (2005). Real-space protein-model completion: an inverse-kinematics approach. *Acta Crystallographica Section D-Biological Crystallography*, 61:2–13.

Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R. M., and Schlichting, I. (1999). Crystal structures of myoglobin-ligand complexes at near-atomic resolution. *Biophysical Journal*, 77(4):2153–74.

Voth, G. A. (2009). *Coarse-graining of condensed phase and biomolecular systems*. CRC Press.

Wade, R. C., Winn, P. J., Schlichting, E., and Sudarko (2004). A survey of active site access channels in cytochromes p450. *Journal of Inorganic Biochemistry*, 98(7):1175–1182.

Wilmarth, S. A., Amato, N. M., and Stiller, P. F. (1999). Maprm: A probabilistic roadmap planner with sampling on the medial axis of the free space. *Icra '99: Ieee International Conference on Robotics and Automation, Vols 1-4, Proceedings*, pages 1024–1031 3286.

Winn, P. J., Ludemann, S. K., Gauges, R., Lounnas, V., and Wade, R. C. (2002). Comparison of the dynamics of substrate access channels in three cytochrome p450s reveals different

opening mechanisms and a novel functional role for a buried arginine. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5361–5366.

Wlodarski, T. and Zagrovic, B. (2009). Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences of the United States of America*, 106(46):19346–51.

Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D., and Nussinov, R. (2008). Molaxis: efficient and accurate identification of channels in macromolecules. *Proteins : Structure, Function, and Bioinformatics*, 73(1):72–86.

Yang, L., Song, G., Carriquiry, A., and Jernigan, R. L. (2008). Close correspondence between the motions from principal component analysis of multiple hiv-1 protease structures and elastic network modes. *Structure*, 16(2):321–30.

Yang, Z., Mjek, P., and Bahar, I. (2009). Allosteric transitions of supramolecular systems explored by network models: Application to chaperonin groel. *PLoS Computational Biology*, 5:e1000360.

Yao, P., Dhanik, A., Marz, N., Propper, R., Kou, C., Liu, G., van den Bedem, H., Latombe, J. C., Halperin-Landsberg, I., and Altman, R. B. (2008). Efficient algorithms to explore conformation spaces of flexible protein loops. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(4):534–45.

Yen, J. Y. (1971). Finding the k shortest loopless paths in a network. *Management Science*, 17:712–716.

Yesylevskyy, S. O., Kharkyanena, V. N., and Demchenko, A. P. (2006). Hierarchical clustering of the correlation patterns: new method of domain identification in proteins. *Biophysical Chemistry*, 119:84–93.

Zhang, X. J., Wozniak, J. A., and Matthews, B. W. (1995). Protein flexibility and adaptability seen in 25 crystal forms of t4 lysozyme. *Journal of Molecular Biology*, 250(4):527–552.

Zheng, W. (2008). A unification of the elastic network model and the gaussian network model for optimal description of protein conformational motions and fluctuations. *Biophysical Journal*, 94:3853–857.

Zheng, W. and Brooks, B. (2005). Idnetification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. *Journal of Molecular Biology*, 346:745–759.

Zheng, W., Brooks, B., and Thirumalai, D. (2006). Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences of the United States of America*, 6:7664–7669.

Zoete, V., Michielin, O., and Karplus, M. (2002). Relation between sequence and structure of hiv-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *Journal of Molecular Biology*, 315(1):21–52.