

國立交通大學

生物資訊所

博士論文

計算結構生物學：

在構形亂度和蛋白質動態上的研究

Computational Structural Biology Studies on:

I. Conformational entropy

II. Protein dynamics

研究生：黃少偉

指導教授：黃鎮剛 教授

中華民國九十七年九月

計算結構生物學：在構形亂度和蛋白質動態上的研究

Computational Structural Biology Studies on:

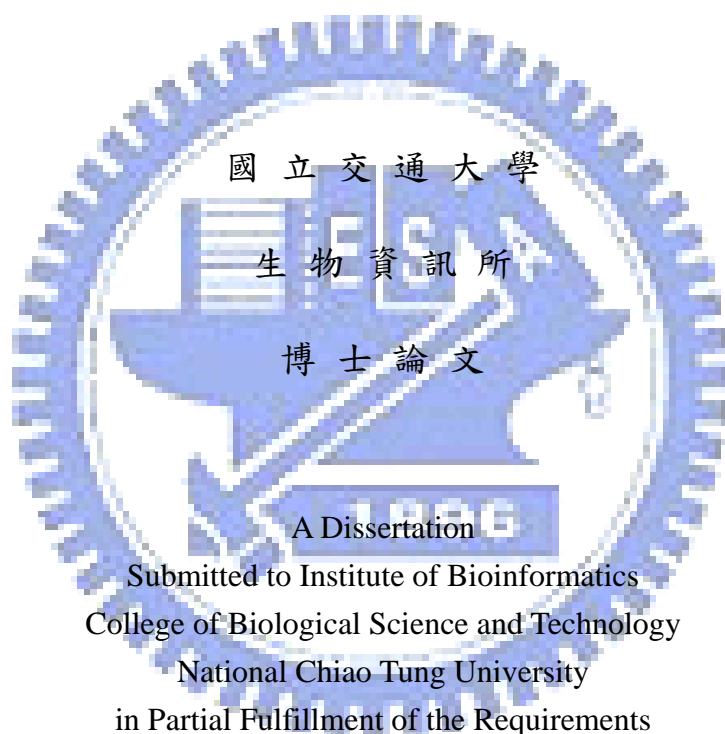
I. Conformational entropy II. Protein dynamics

研究生：黃少偉

Student : Shao-Wei Huang

指導教授：黃鎮剛

Advisor : Jenn-Kang Hwang



Submitted to Institute of Bioinformatics
College of Biological Science and Technology
National Chiao Tung University
in Partial Fulfillment of the Requirements

for the Degree of
PhD
in

Bioinformatics

September 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年九月

計算結構生物學：在構形亂度和蛋白質動態上的研究

學生：黃少偉

指導教授：黃鎮剛

國立交通大學生物資訊研究所博士班

摘要

一段蛋白質序列通常會構成一個獨特的立體結構，但前人的研究指出，不論是人工合成或是自然界存在的短序列片段，都發現到它們會在不同的結構環境中，形成不同的二級結構。這種短序列的特殊性質，最早是由 Kabsch 和 Sander 所發現的，在他們利用序列的相似性來預測蛋白質結構的研究中，發現了這個現象，並把有這類性質的短序列，命名為 chameleon (變色龍)。他們發現了許多長度為五個胺基酸的小序列片段，在不同的蛋白質裡面，形成了不同的二級結構。相反的，有些小序列片段，不管在什麼蛋白質中，都形成相同的二級結構，這樣子的序列片段，稱為有高度的結構保守性 (structure conservation)，而具 chameleon 性質的序列片段，其結構保守性則很低。在這部份的工作裡，我們利用 Support vector machine，開發了一種預測蛋白質短序列片段之結構保守性的方法。而我們將預測的結果，和 Hydrogen isotope exchange 實驗互相比較，發現了高度結構保守性的胺基酸，通常也具有緩慢的 Hydrogen isotope exchange 速率。

研究蛋白質的動態 (dynamics) 是分子生物學研究中，一個很重要

的課題。最典型的計算生物學方法是分子動態模擬 (molecular dynamics simulation, MD)，這種方法計算原子和原子間的交互作用：鍵能、電荷作用等等。雖然 MD 的計算很精確，但是缺點是需要大量的計算時間和參數調整。另一種方法是 Gaussian network model (GNM)，它把蛋白質結構轉換成一個 C α 原子連結而成的網路結構，GNM 可以估算出個別胺基酸的熱擾動 (thermal fluctuation) 以及胺基酸之間的動態相關性 (correlation of motion)。最近由本實驗室開發的方法 Protein-fixed-point (PFP) model，是一個非常簡單而準確的方法，所需要的資訊只有胺基酸 C α 原子的空間座標。首先決定整個蛋白質構造的重心 (center of mass) 座標，我們發現某一胺基酸的熱擾動大小，和它到蛋白質重心的距離平方成正比例的關係。另一個本實驗室開發的方法是 Weighted-contact number (WCN) model，它計算胺基酸周圍的原子數目，愈接近的原子可以得到較高的權重 (weight)，反之則影響較小。胺基酸若處於有很多原子集結的環境，則熱擾動的值較小。PFP 和 WCN model 的結果顯示，蛋白質的動態資訊，可以單純的由它的結構推算出來，並不需要做任何機械模型的假設 (如 GNM)。在這部份的研究，為了驗證 PFP 和 WCN model 的正確性，我們利用這兩個方法來預測蛋白質的 NMR order parameter，並和實驗值相比較。這兩種方法的預測結果都比前人所做的方法要好。

蛋白質的功能通常會牽涉到結構上大規模的運動 (large scale motion)。Normal mode analysis (NMA) 早在 1980 年開始，就被用來研究蛋白質的大規模運動，它最主要的特點是把蛋白質的動態拆解成很多不同頻率的運動，包含了頻率較低、較大規模的運動，以及頻率較高、較小規模的運動。生物學家通常感興趣的部份是頻率較低、較大規模的運動，因為通常這類的運動和蛋白質的功能表現有最直接的關

係。NMA 最早是把 MD 模擬當中的位能函數 (potential function) 做二次微分得到 Hessian matrix，再對它做對角化 (diagonalize) 運算後，得到蛋白質中各種不同頻率的運動。而這個方法的缺點是，在計算較大的蛋白質時，計算時間會變的非常龐大。另一種方法是 Elastic network model (ENM)，它把蛋白質結構轉換成一個 $C\alpha$ 原子連結而成的網路結構，基於這個網路結構，前人開發出了一種簡化版本的 NMA，它大幅減小了所需要的計算時間，並且也可以得到非常好的結果。現今最廣泛被使用的 ENM 方法是 Gaussian network model (GNM)。本實驗室所研發的 PFP model，是一種簡單，同時可以準確預測胺基酸熱擾動的方法，我們基於 PFP model，研發出了另一種 NMA 的方法，在這部份的研究裡，我們將 PFP model 的 NMA 結果，和 GNM 的結果相比較，並且發現了它們在所研究的例子裡，大致上有相吻合的結果。



Computational Structural Biology Studies on:
I. Conformational entropy II. Protein dynamics

student : Shao-Wei Huang

Advisors : Dr.Jenn-Kang Hwang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

A complete protein sequence usually determines a unique structure; however the situation is different for shorter subsequence. Studies found that both designed and nature occurring subsequences may have different secondary structures in different contexts. This feature of short sequence is called “chameleon” which was first reported by Kabsch and Sander when they used sequence homology to predict protein structures. They found that several pentapeptides which have identical sequence adopt different secondary structures in different protein structures. For nature occurring proteins, systematic search on PDB shows that identical subsequences could have very different conformations. Here we developed a method to compute structure conservation from protein sequence. During protein folding process, there are some structured regions which are similar to folded conformation. Hydrogen isotope exchange (HX) rate is usually used to identify those structured regions. We applied this method to a set of proteins with known HX rate data and found a strong correlation between structure conservation and slow HX

rate.

One of the most important topics in biological science is to understand the protein function. It is well-known that protein dynamics is closely related to the function of protein. Several computational methods have been developed to get the protein dynamics. Molecular dynamics (MD) simulation has been widely used in the study of protein function and dynamics. It simulates the interactions between each atom, bonding force, van der Waals force, charge-charge interaction, etc. The computation time is extremely long when the size of the protein is large and the selection of appropriate parameters of force field itself is a complicated problem. Gaussian network model (GNM) transfers the protein structure into a network in which each $C\alpha$ atom pair is connected together if their distance is smaller than a given cutoff value. Using this protein-converted network, GNM can compute the theoretical thermal fluctuation of each atom and correlation of motions between each atom pair. Recently we have developed a model to predict the thermal fluctuation from protein structure, which is called protein-fixed-point (PFP) model. The PFP model only uses the coordinates of $C\alpha$ atoms and simply determines the center of mass of the protein. We found that the thermal fluctuation is proportional to the squared distance from the atom to the center of mass of the structure. Another model called weighted contact number (WCN) model computes the number of neighboring atoms weighted by the inverse distance between each atom pair. The PFP and WCN model show that the protein dynamics can be extracted directly from the intrinsic property of protein structure without the use of any mechanical model. The order parameter obtained by the NMR experiment is widely used to study the dynamic-related protein functions. Here, we use the PFP and WCN model to predict the N-H S^2 order parameter directly from the protein structure. Our results show that the WCN model

can more accurately reproduce the experimental order parameter than previous publication.

The biological function of proteins is closely related to cooperative motions and correlated fluctuations which involve large portions of the structure. Normal Mode Analysis (NMA) had been used to study biomolecules since early 1980s. It decomposes the protein dynamics into a collection of motions which include large scale/low frequency and small scale/high frequency motions. Biologists usually focus on the large scale/low frequency motions which are relevant to protein functions. The major contribution of NMA to the biological research field is the ability to provide the information of large, domain-scale protein motions which is hard to compute by other methods. The classical approach of NMA is to diagonalize the Hessian matrix, i.e. the second derivative of the potential function of a molecular dynamics (MD) simulation. The major shortcoming of the classical NMA is that the sampling time increases dramatically with the size of the protein. The Elastic Network Model (ENM), which is able to describes protein dynamics without amino acid sequence and atomic coordinates, has been widely used in the studies of protein dynamics and structure-function relationship. The ENM views the protein structure as an elastic network, the nodes of which are the C α atoms of individual residues. Residue pairs within a cutoff distance are connected by springs which have a uniform force constant in the network. Based on ENM, a coarse-grained version of NMA is developed and widely used because of its low computation cost and the ability to extend the dynamics to longer timescale and larger motions. The coarse-grained NMA had been applied to various topics, for example, protein functions and catalytic residues. One of the most widely used ENM-based methods is the Gaussian network model (GNM). The protein-fixed-point (PFP)

model is a simple method to compute the protein dynamics only using the coordinates of $C\alpha$ atoms. Despite its simplicity, it has been shown to be able to accurately predict the B-factors for a dataset of 972 proteins. Here, we compared the results of NMA based on the PFP model with those by Gaussian network model (GNM).



誌謝

感謝我的指導教授黃鎮剛老師，從我大學三年級做專題開始，一直到現在七年來的耐心教導，啟發了我對生物資訊方面研究的興趣和熱忱。感謝大學的班導師彭慧玲老師，從大學到研究所一直關心班上同學的生活和發展，給我們很多鼓勵和各方面的幫助。謝謝楊進木老師、黃憲達老師、呂平江老師、林彩雲老師在學位考試的指導和對論文需要改進地方的建議，讓我了解到自己不足和需要加強的地方。

感謝玉菁學姐、春吟學姐在我做專題生時熱心的幫忙。大熊學長、勇欣學長、景盛學長、禎祥學長、涵堃學長對於研究方面的幫助。最照顧學弟妹的草霸、志鵬、鐵雄、志杰學長們，有好康的東西總不會忘記我們。碩班同時進實驗室的戰友，小操、建華、啟德、肇基、蔚倫，和你們一起修課、努力做研究的日子是最難忘的。感謝實驗室的學弟妹們：世瑜、思樸、書瑋、彥龍、小胖、啟文、阿壁、松桓、仙蕾、慧雯、子琳、小芬、人維、乃文，為我們這些老人家帶來許多歡樂。

讀研究所期間，感謝我的好朋友們在開心的時候一起分享歡樂、難過的時候互相陪伴，最會關心別人的健誠，常常陪我丟棒球的小操，沒事會招待我們去家裡玩的祐俊和怡伶夫婦；最會替別人著想的宗鳳、彥均、庭毓，僑牌高手 Vic 和小新、常常開車帶我們出去玩的明宏，老羅學長、昱佑、劭恆、小宏，感謝你們的陪伴。特別感謝子慧四年來容忍我難相處的個性，給我無止盡的包容和鼓勵。

最重要的感謝我的家人，爸爸和媽媽從小到大的照顧和關心，讓

我能夠專心學習並且追求自己的興趣，在決定要攻讀博士學位的時候，也一直給我支持和鼓勵。能夠專注在一件事情上，並且堅持完成的態度，都是從你們身上學習到的，感謝你們！



Index

中文摘要	i
英文摘要	iv
誌謝	viii
Index	x
Tables	xi
Figures	xii
Abbreviations	xiv
Chapter 1	Computation of Conformational Entropy from Protein Sequence	
1.1	Introduction.....	1
1.2	Methods.....	4
1.3	Results.....	10
1.4	Discussion.....	28
Chapter 2	NMR Order Parameter Prediction Using Protein-fixed-point Model and Weighted Contact Number Model	
2.1	Introduction.....	30
2.2	Methods.....	34
2.3	Results.....	36
2.4	Discussion.....	47
Chapter 3	Normal Mode Analysis by the Protein-fixed-point Model	
3.1	Introduction.....	49
3.2	Methods.....	51
3.3	Results.....	53
3.4	Discussion.....	64
Appendix		
A	67
B	68
C	69
D	71
E	72
References	77

Tables

Table 1	The predicted backbone N-H order parameters computed with the WCN, PFP, CM, and rCENM.....	36
Table 2	The Pearson correlation coefficient between the correlation maps computed by the PFP model and the GNM.....	53
Table 3	RS126 dataset.....	67
Table 4	Hydrogen exchange dataset.....	68
Table 5	Normal mode analysis dataset.....	71



Figures

Figure 1	The AVLAE sequence (red color) forms an α -helix in potassium channel (right, 1BL8) but forms a β -sheet in a transposase inhibitor (left, 1B7E).....	2
Figure 2	Flowchart of conformational entropy calculation from protein sequence.....	4
Figure 3	Maximum-margin hyperplanes for a SVM trained with samples from two classes. Samples along the hyperplanes are called the support vectors.....	5
Figure 4	Conformational entropy profile of hen egg-white lysozyme.....	10
Figure 5	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of hen egg-white lysozyme.....	11
Figure 6	Conformational entropy profile of chymotrypsin inhibitor 2.....	12
Figure 7	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of chymotrypsin inhibitor 2.....	13
Figure 8	Conformational entropy profile of cytochrome c.....	14
Figure 9	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of cytochrome c.....	15
Figure 10	Conformational entropy profile of each unfolding unit of cytochrome c and the corresponding average values.....	16
Figure 11	Conformational entropy profile of barnase.....	18
Figure 12	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of barnase.....	19
Figure 13	Conformational entropy profile of α -lactalbumin.....	20
Figure 14	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of α -lactalbumin.....	21
Figure 15	Conformational entropy profile of CTX III.....	22
Figure 16	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of CTX III.....	23
Figure 17	Conformational entropy profile of ribonuclease H.....	24
Figure 18	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of ribonuclease H.....	25
Figure 19	Conformational entropy profile of BPTI.....	26
Figure 20	Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of BPTI.....	27
Figure 21	Superposition of apo <i>Aquifex</i> Adk (red) and <i>Aquifex</i> Adk in complex (green).....	30
Figure 22	Predicted and experimental order parameter of β ARK1 PH domain....	37
Figure 23	Predicted and experimental order parameter of Calbindin.....	38

Figure 24	Predicted and experimental order parameter of Cold-shock protein A from <i>E-coli</i>	39
Figure 25	Predicted and experimental order parameter of Frenolicin acyl carrier protein.....	40
Figure 26	Predicted and experimental order parameter of Lysozyme.....	41
Figure 27	Predicted and experimental order parameter of SH2 domain of p85 subunit of phosphoinositide 3-kinase.....	42
Figure 28	Predicted and experimental order parameter of Ubiquitin.....	43
Figure 29	Predicted and experimental order parameter of Ketosteroid isomerase.....	44
Figure 30	Predicted and experimental order parameter of 4-oxalocrotonate Tautomerase.....	45
Figure 31	Predicted and experimental order parameter of Interleukin-4.....	46
Figure 32	The correlation maps of (a) Prot-glu methylsterase (1RPT), (b) Ribonuclease T2 (1BOL), and (c) Uridine nucleosidase (1EUG) computed with the PFP model (left) and GNM (right), respectively....	54
Figure 33	Distribution of displacements along the first mode, second mode, and third mode computed for thiol-endopeptidase (9PAP).....	57
Figure 34	The regions subject to opposite direction displacements computed by the PFP model (left) and the GNM (right) of (a) the first mode, (b) the second mode, and (c) the third mode for thiol-endopeptidase (9PAP)...	58
Figure 35	Ribbon diagram colored blue-yellow-red in the order of increasing mobility along the first mode of the PFP model (left) and GNM (right).....	59
Figure 36	Distribution of displacements along the first mode, second mode, and third mode computed for G/11 xylanase (1BVV).....	60
Figure 37	The regions subject to opposite direction displacements computed by the PFP model (left) and the GNM (right) of (a) the first mode, (b) the second mode, and (c) the third mode for G/11 xylanase (1BVV).....	61
Figure 38	The displacements of actinidin (1AEC) subject to opposite directions along the first mode.....	62
Figure 39	Ribbon diagram of actinidin colored blue-yellow-red in the order of increasing mobility along the first mode.....	63
Figure 40	The flowchart of computing normal mode motions by PFP model and GNM.....	65

Abbreviations

Chapter 1

BLAST	Basic local alignment search tool
BPTI	Bovine pancreatic trypsin inhibitor
CI2	Chymotrypsin inhibitor 2
HEWL	Hen egg-white lysozyme
HX	Hydrogen exchange
PSI-BLAST	Position-specific iterative BLAST
PSSM	Position-specific substitution matrix
RNaseH	Ribonuclease H
SVM	Support vector machine

Chapter 2

CM	Contact model
CspA	Cold-shock protein A
Frendicin ACP	Frendicin acyl carrier protein
GNM	Gaussian network model
MD	Molecular dynamics
PFP	Protein-fixed-point
rCENM	Reorientational contact-weighted elastic network model
WCN	Weighted contact number

Chapter 3

ENM	Elastic network model
GNM	Gaussian network model
MD	Molecular dynamics
NMA	Normal mode analysis
PFP	Protein-fixed-point

Computation of Conformational Entropy from Protein Sequence

Introduction

A complete protein sequence usually determines a unique structure; however the situation is different for shorter subsequence. Studies¹⁻⁴ on both designed and nature occurring subsequences found that they may have different secondary structures in different contexts. This feature of short sequence is called “chameleon” which was first reported by Kabsch and Sander¹ using sequence homology to predict protein structures. They found that several pentapeptides which have identical sequence adopt different secondary structures in different protein structures. For nature occurring proteins, systematic search^{1,3,4} on PDB⁵ shows that identical subsequences could have very different conformations. For example, the pentapeptide AVLAE forms an α -helix in a potassium channel but forms a β -sheet in a transposase inhibitor (Figure 1.1). Following the observation of Kabsch and Sander¹, Cohen et al.⁶ discovered eight pairs of hexapeptides sequences which adopt α -helix in one protein and β -strand in the other using a larger protein dataset.

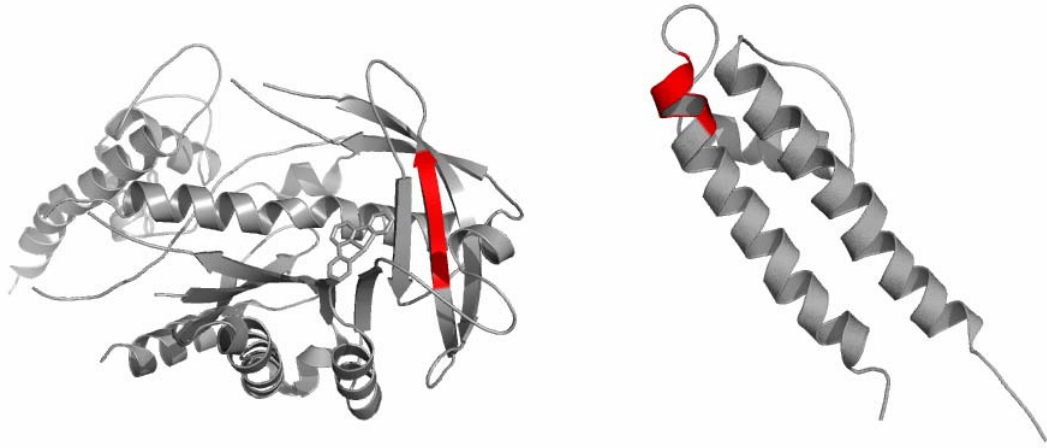
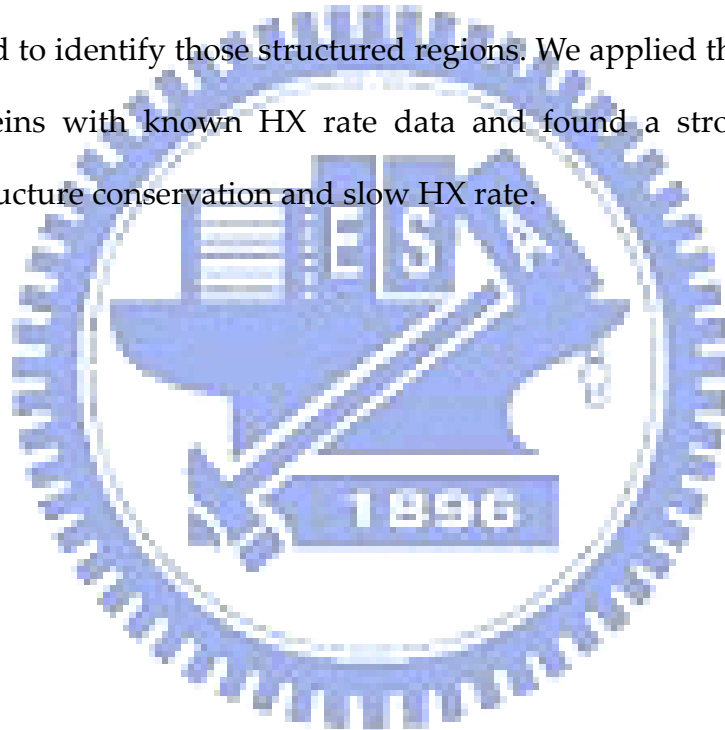


Figure 1.1 The AVLAE sequence (red color) forms an α -helix in potassium channel (right, 1BL8) but forms a β -sheet in a transposase inhibitor (left, 1B7E).

In addition to the studies focused on subsequences forming different secondary structures in different proteins, Minor and Kim² artificially put a short sequence into different locations of a single protein. They designed an 11-peptide sequence which folds as an α -helix in one context while a β -sheet in another. The designed subsequence (AWTVEKAFKTF) was replaced at different positions of IgG-binding domain of protein G (GB1), and the subsequence was determined experimentally to fold into different conformations.

However, some of the shorter subsequences are able to form unique conformation in different contexts. In the case of subsequence AALAE, the conformation remains the same α -helix in different proteins. This property of structural conservation or variability is determined by the local and non-local interactions.

Here we developed a method to compute structure conservation from protein sequence. The main idea is to predict, for each residue, the possibilities distribution of each secondary structure type. Based on the distribution, a conformational entropy value is computed for each residue and is assumed to be related to the structural conservation. The low conformational entropy parts indicate highly structural conserved regions. During protein folding process, there are some structured regions which is similar to folded conformation⁷⁻¹⁰. Hydrogen isotope exchange (HX) rate¹¹⁻¹³ is usually used to identify those structured regions. We applied this method to a set of proteins with known HX rate data and found a strong correlation between structure conservation and slow HX rate.



Methods

Overview

The structure conservation (**conformational entropy**) of each residue is calculated from predicted secondary structure distribution from protein sequence. We use the evolutionary information from PSI-BLAST¹⁴ as training data of support vector machine (SVM) to predict the secondary structure distribution. Figure 1.2 shows the flowchart of our method.

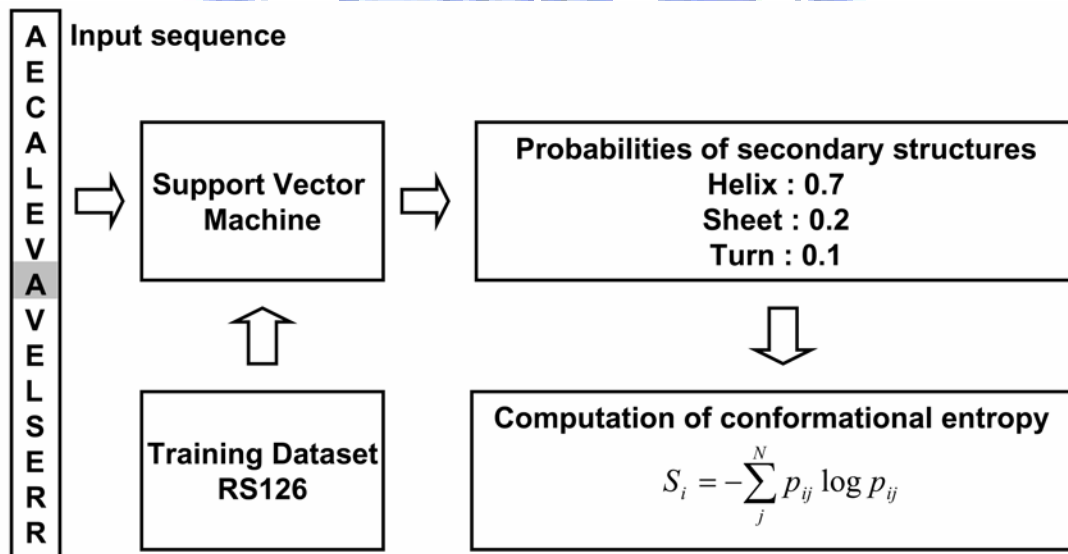


Figure 1.2 Flowchart of conformational entropy calculation from protein sequence.

The Support Vector Machine

The idea of support vector machine method (SVM)¹⁵ is to find the separating hyperplane with largest distance between two classes. However in many cases, the data can not be separated linearly. To solve this problem, SVM uses kernel functions to nonlinearly transform the input space into higher dimensional feature space, and the data may be well separated in the higher dimensional space.

Consider the data points of the form: $(x_1, c_1), (x_2, c_2), \dots, (x_n, c_n)$ where the c_i is the class to which the data point x_i belongs (either 1 or -1 in this example). Each x_i is a n -dimensional vector of input training data which we would like SVM to distinguish, by means of the dividing hyperplane, which takes the form

$$w \times x - b = 0 \tag{1.1}$$

We want this hyperplane to have maximum distance (called *margin*) to the closest data points from both classes, as shown in figure 1.3.

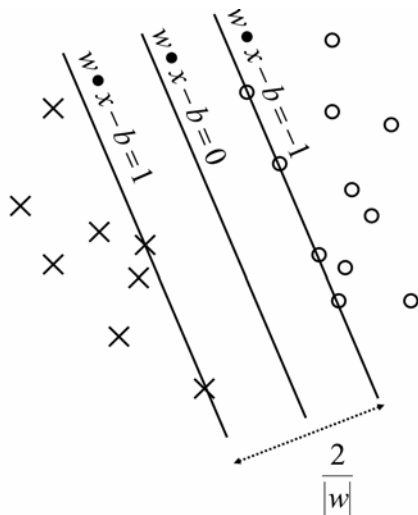


Figure 1.3 Maximum-margin hyperplanes for a SVM trained with samples from two classes. Samples along the hyperplanes are called the support vectors.

The maximum-margin hyperplane can be defined by finding the parallel hyperplanes closest to the support vectors in either class. These parallel hyperplanes can be described by equations

$$\begin{aligned} w \times x - b &= 1 \\ w \times x - b &= -1 \end{aligned} \quad (1.2)$$

We would like these hyperplanes to have maximum distance from the dividing hyperplane and no data points between them. To exclude data points, all i must satisfy either

$$w \times x_i - b \geq 1 \quad (1.3)$$

or

$$w \times x_i - b \leq -1 \quad (1.4)$$

We can rewrite this to

$$\begin{aligned} c_i (w \times x_i - b) &\geq 1 \\ 1 &\leq i \leq n \end{aligned} \quad (1.5)$$

So we want to maximize the distance between the hyperplanes ($2/|w|$) subject to the constraint (1.5). After training, we can use SVM to classify test data points by following decision rule

$$\hat{c} = \begin{cases} 1, & \text{if } w \times x + b \geq 0 \\ -1, & \text{if } w \times x + b \leq 0 \end{cases} \quad (1.6)$$

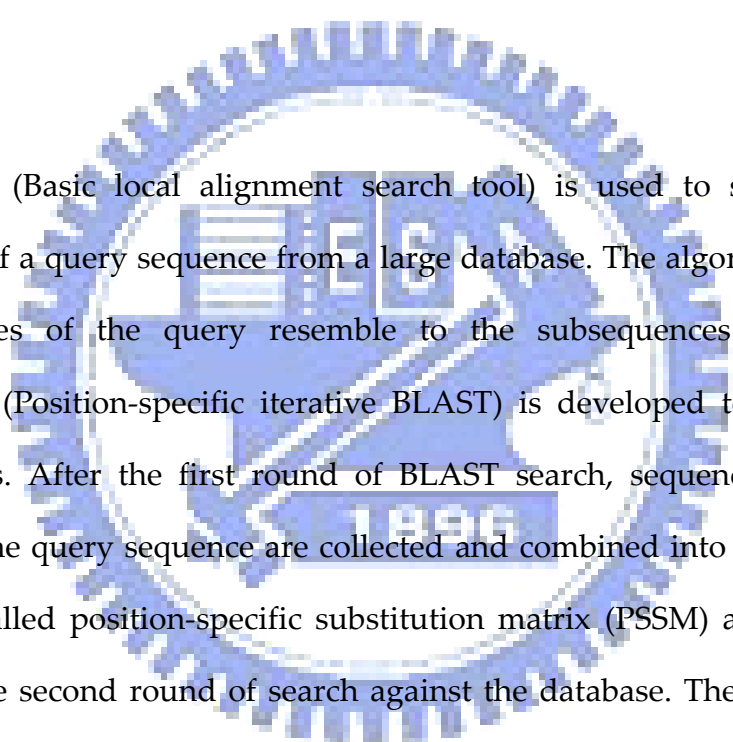
SVM had been successfully applied to secondary structure prediction¹⁶⁻¹⁸, solvent accessibility prediction¹⁹, fold assignment^{20,21}, subcellular localization^{22,23} and other computation biology problems²⁴⁻²⁶. In this work, the

software package LIBSVM²⁷ was used.

Computation of Conformational Entropy

The conformational entropy of each residue is calculated from predicted secondary structure distribution from protein sequence. We use the evolutionary information from PSI-BLAST¹⁴ as training data of SVM to predict the secondary structure distribution.

PSI-BLAST



BLAST (Basic local alignment search tool) is used to search similar sequences of a query sequence from a large database. The algorithm will find subsequences of the query resemble to the subsequences of database. PSI-BLAST (Position-specific iterative BLAST) is developed to find distant homologues. After the first round of BLAST search, sequences which are similar to the query sequence are collected and combined into a profile. This profile is called position-specific substitution matrix (PSSM) and is used as query in the second round of search against the database. The rebuilding of profile and searching procedure is then repeated. PSSM contains the information of relate sequences and is more sensitive in finding distant relate sequences.

Secondary structure definition

In this work, the secondary structure is defined by DSSP²⁸ program. DSSP assigns secondary structures of a given 3D structure according to the atom coordinates and hydrogen bonding patterns. Eight classes of secondary

structures are defined in DSSP

- H = α -helix
- B = residue in isolated β -bridge
- E = extended strand, participates in β ladder
- G = 3/10 helix
- I = π -helix
- T = hydrogen bonded turn
- S = bend
- U = undefined

Conformational entropy calculation

We compute the probability distribution of the secondary structure by SVM. The inputs of SVM are in the form of $W * 20$ PSSM profiles, where W is the sliding window size. The sliding window is used to include the information of neighboring residues on sequence. In the work, window size is chosen to be 15. PSSM profile is obtained after 3 iterations of PSI-BLAST search against a non-redundant database with the E -value threshold set to $1 * 10^{-3}$.

Each element in PSSM profile represents the score of a particular residue substitution at that position. These elements are usually in the range ± 7 and are normalized to the range $[0, 1]$ by the following function

$$\pi(x) = \begin{cases} 0.0 & \text{if } x \leq -5 \\ 0.5 + 0.1x & \text{if } -5 \leq x \leq 5 \\ 1.0 & \text{if } x \geq 5 \end{cases} \quad (1.7)$$

where x is the original matrix element value. The output of SVM for the target residue is an eight-element vector $O = (o_1, o_2, \dots, o_8)$, where o_i is the decision value of secondary structure type i . The output of SVM does not give the probability for each classification, so we use a function $[\arctan(o_i) + \pi]$ to transform the decision value to probability in the range $[0, 1]$. With $P = (p_1, p_2, \dots, p_8)$, where p_i is the probability of secondary type i of target residue, we calculate conformational entropy by the following equation

$$S_i = -\sum_j^N p_{ij} \log p_{ij} \quad (1.8)$$

where S_i is the conformational entropy of residue i , and p_{ij} is the probability of residue i classified to be secondary type j . N is the number of secondary structure types.

Training Dataset

We train SVM using a standard non-redundant dataset RS126²⁹, with sequence identity less than 25% over a length of more than 80 residues. See appendix A for more detail.

Results

Hen egg-white lysozyme

Hen egg-white lysozyme (HEWL) contains two sub-domains: the α -domain which has four α -helices (A, B, C, and D), two 3_{10} helices, and the β -domain composed of three β -strands. Hydrogen exchange study³⁰ on HEWL showed that the slowest exchange amide protons are located in the α -domain: helix A (M12), helix B (W28-A31), and helix C (A95, K96, I98); and the next slowest in strand β_3 (I58).

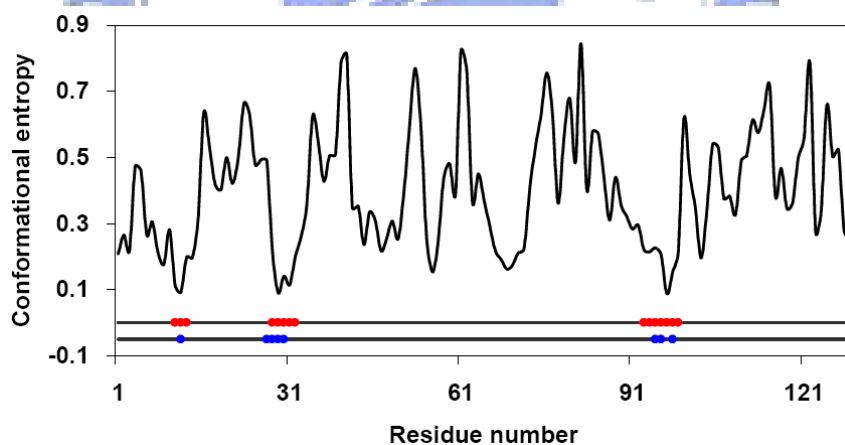


Figure 1.4 Conformational entropy profile of hen egg-white lysozyme. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.4 shows the conformational entropy profile of HEWL. The lowest entropy regions are A11-K13 (helix A), W28-K33 (helix B), and N93-V99 (helix C), respectively. These amino acids overlap with the slow exchange amide protons in helices A, B, and C. Note the residues in helix D (V109-R114) have relatively higher entropy. This result is supported by the experiment that amino acids in helix D have much higher amide exchange rate. Figure 1.5 is the ribbon diagrams of HEWL colored by low conformational entropy and slow exchange regions.

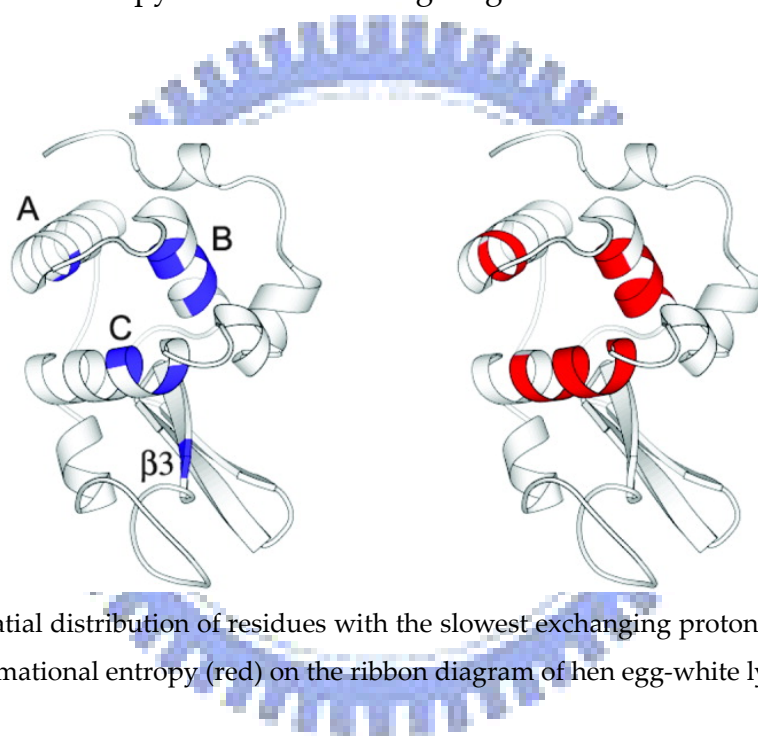


Figure 1.5 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of hen egg-white lysozyme.

Chymotrypsin inhibitor 2

Chymotrypsin inhibitor 2 (CI2) is a monomeric protein of 64 residues and folds by simple two-state kinetics. The only well developed region during intermediate state is an N-terminal α -helix and some distant residues in sequence which contact with it. The α -helix packs with the β -sheet to form the hydrophobic core. Studies^{31,32} showed that the slowest exchange-rate residues are located in the C-terminus of the α -helix (I20-I21) and the central strand of β -sheet (V47, L49-V51); the other slowest exchange amide protons are K11(β 2), I30 and L32 (β 3).

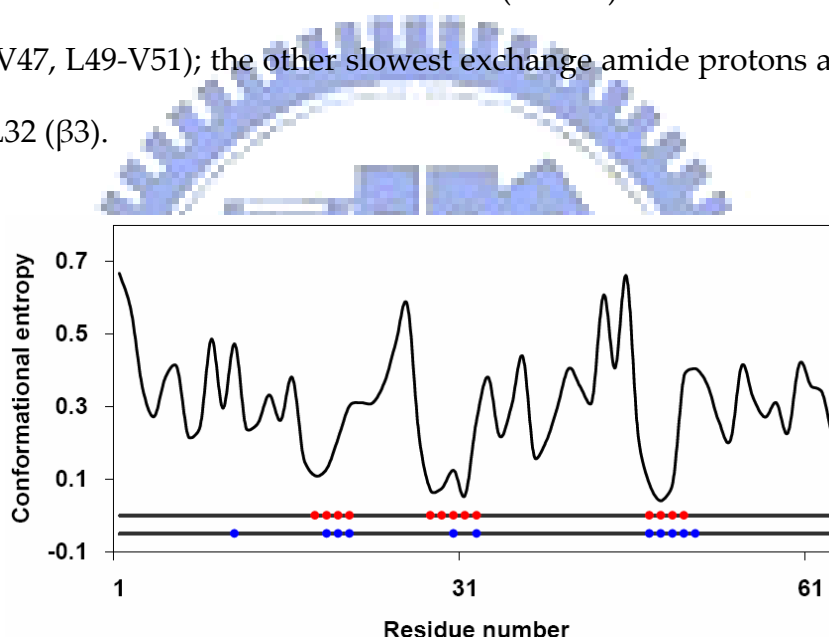


Figure 1.6 Conformational entropy profile of chymotrypsin inhibitor 2. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.6 shows that most residues having the lowest conformational entropy overlap with those of the slowest hydrogen exchange rates, except K11. Most of them are on the α -helix, β -strands 3 and 4. Figure 1.7 compares the spatial distribution of the lowest conformational entropy and the slowest exchange-rate regions on ribbon diagram of CI2.

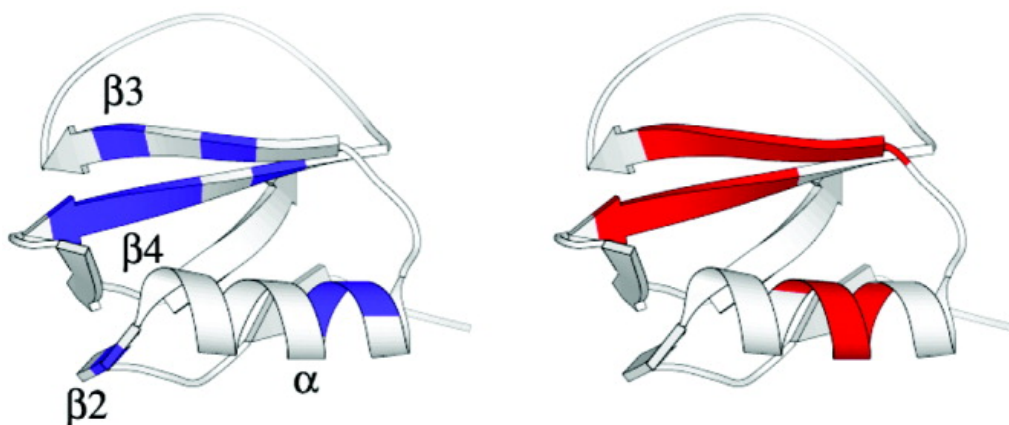


Figure 1.7 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of chymotrypsin inhibitor 2.



Cytochrome c

Hydrogen exchange study³³ shows that the slowest exchanging amide protons are located in the three major helical segments of cytochrome c. F10 (N-helix) and L94-K99 (C-helix) carry the slowest exchanging amide protons.

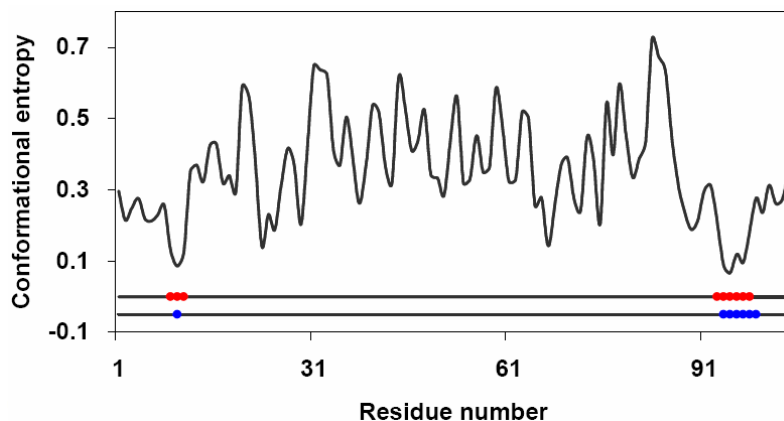


Figure 1.8 Conformational entropy profile of cytochrome c. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.8 shows the conformational entropy profile of cytochrome c. I9-V11 and D93-L98 are the residues having the lowest entropy values which match the slowest exchanging residues on N-helix and C-helix suggested by experiment. Note that the 60's helix also has relative low entropy values. This is consistent with the experimental results that the next slowest exchanging amide protons are located in the 60's helix. Figure 1.9 shows the spatial distribution of residues with the slowest exchanging protons and the lowest conformational entropy on ribbon diagram of cytochrome c.

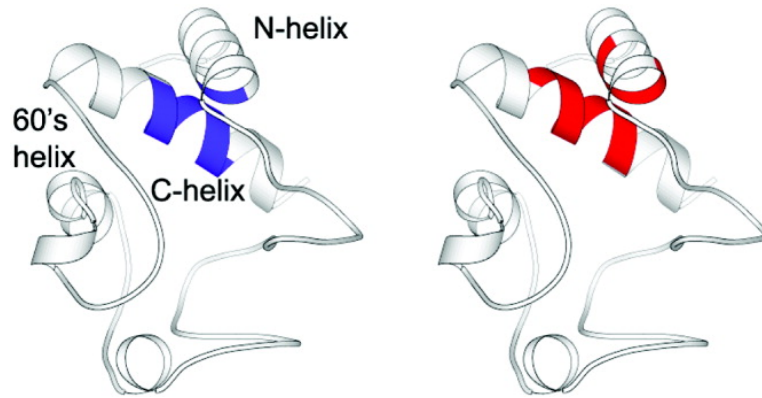


Figure 1.9 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of cytochrome c.



Equilibrium Protein Folding

Cytochrome c is a good model system in studies³⁴⁻³⁶ of protein folding and unfolding. Under native conditions, most proteins exist in their unique native conformation. However some of them also exist in higher energy states and continue to cycle through totally unfolded states and partially folded states. These non-native forms are usually hard to be detected because of the abundant native conformational signals. Hydrogen exchange experiment can be used to detect these partially folded conformations and define the unfolding units of proteins.

Through HX studies³⁴⁻³⁶, there are four cooperative unfolding units defined in cytochrome c: the blue bi-helix (B), the green Ω loop and the 60's helix (G), the yellow (Y) and the red Ω loop (R) in the order of decreasing unfolding free energy. These unfolding units may define the folding and unfolding pathways of cytochrome c by forming various intermediates through different combinations.

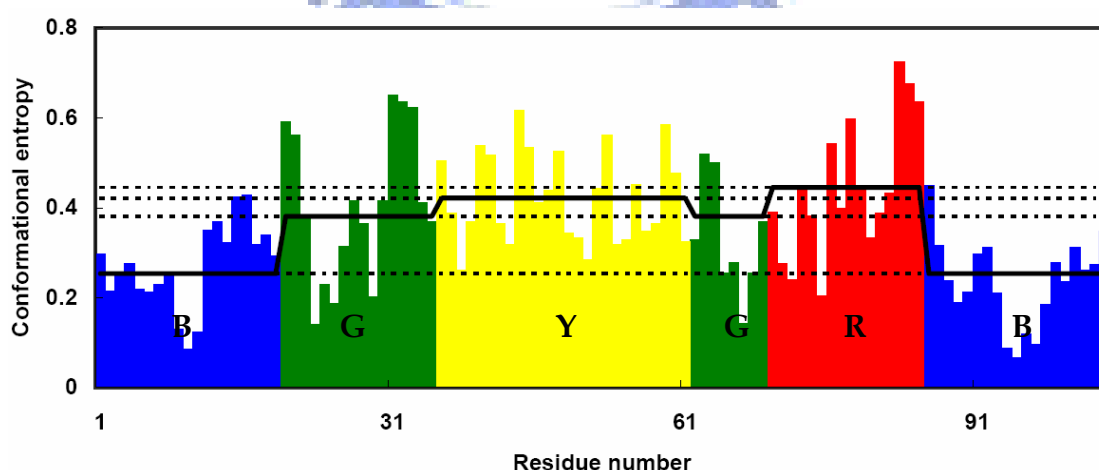


Figure 1.10 Conformational entropy profile of each unfolding unit of cytochrome c and the corresponding average values.

Figure 1.10 shows the conformational entropy profile of each unfolding unit and the corresponding average values. Unit B has the lowest average conformational entropy, unit G has second lowest entropy, and then Y, and R. The order of increasing conformational entropy follows the order of decreasing unfolding free energy.



Barnase

Barnase has three α -helices and a five-stranded β -sheet. The first helix packs onto the β -sheet to form the major hydrophobic core of the protein. The second and the third α -helices pack onto another side of the β -sheet to form a smaller hydrophobic core.

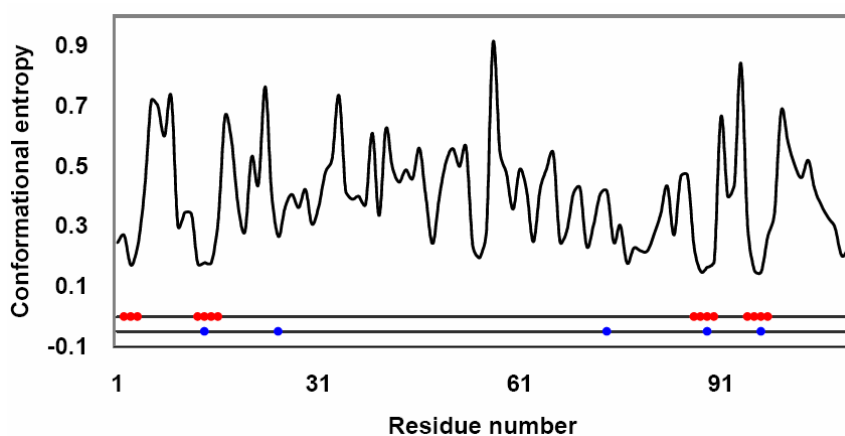


Figure 1.11 Conformational entropy profile of barnase. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.11 shows the conformational entropy profile of barnase. Previous study^{37,38} showed that L14, I25, A74, L89 and Y97 have the lowest exchange rates. L14 on the first α -helix and I25, A74 and L89 on the β -sheet are located in the major hydrophobic core of the protein. Y97 is in the center of the smaller hydrophobic core formed by the two smaller helices and part of the β -sheet. Figure 1.12 shows the spatial distribution of residues with the slowest exchanging protons and the lowest conformational entropy on the ribbon diagram of barnase.

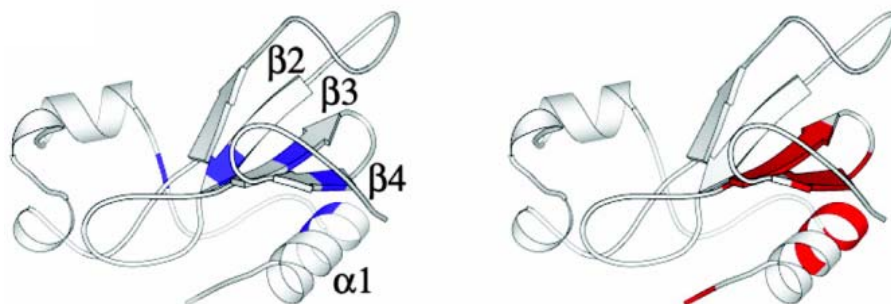


Figure 1.12 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of barnase.



α -Lactalbumin

α -lactalbumin is a calcium-binding protein which consists of a β -sheet domain and a α -helical domain. The helical domain is composed of four helices: helix A (C6-E11), helix B (P24-S34), helix C (D87-D97) and helix D (D102- L105). Previous studies^{39,40} showed that in the helical domain, the C-helix is the most protected, having the lowest exchanging rate, followed by the B and then the A-helix.

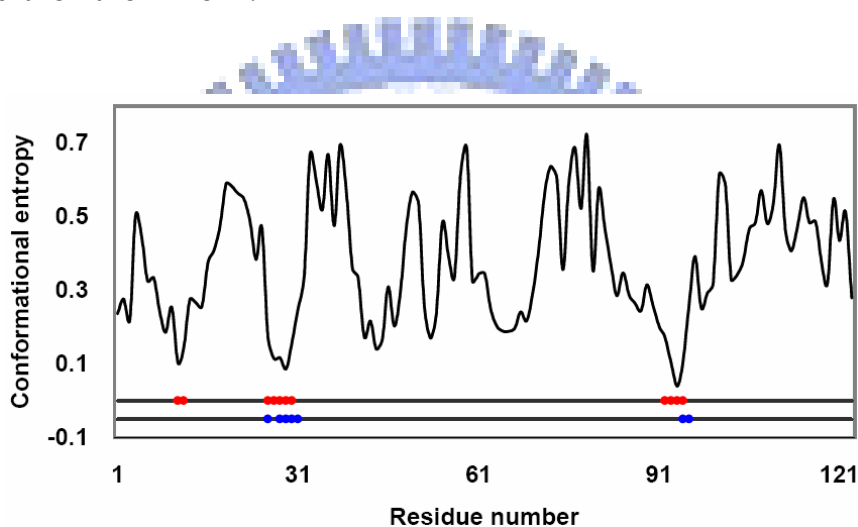


Figure 1.13 Conformational entropy profile of α -lactalbumin. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.13 shows the profile of the conformational entropy. The residues in C-helix have the lowest conformational entropy and the residues which have second lowest conformational entropy are located in helix B and helix A. The regions which have low conformational entropy are consistent with those having slow exchanging rate^{39,40}. Note that a previous study⁴⁰ showed that helix D exchanges too fast and the exchanging rates are not measurable. Our calculation also indicates that helix D has high conformational entropy. Figure 1.14 shows the spatial distribution of residues with the slowest exchanging

protons and the lowest conformational entropy on the ribbon diagram of α -lactalbumin.

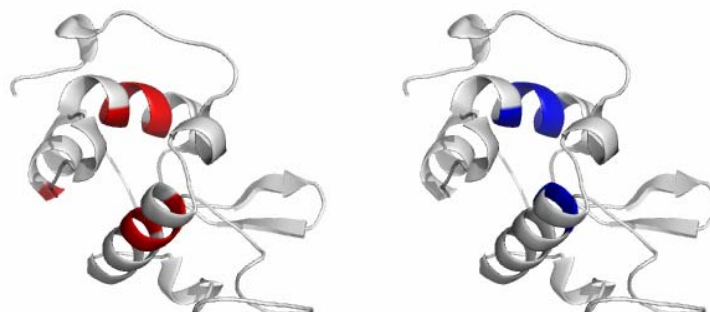


Figure 1.14 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of α -lactalbumin.



Cardiotoxin III

Cardiotoxin III (CTX III) is a 60-amino acid, all β -sheet protein. The secondary structure of CTX III includes five β -strands forming double and triple-stranded anti-parallel β -sheets. Hydrogen exchange study⁴¹ on CTX III showed that residues K23, I39, and Y51-N55 constitute the hydrophobic cluster of the protein.

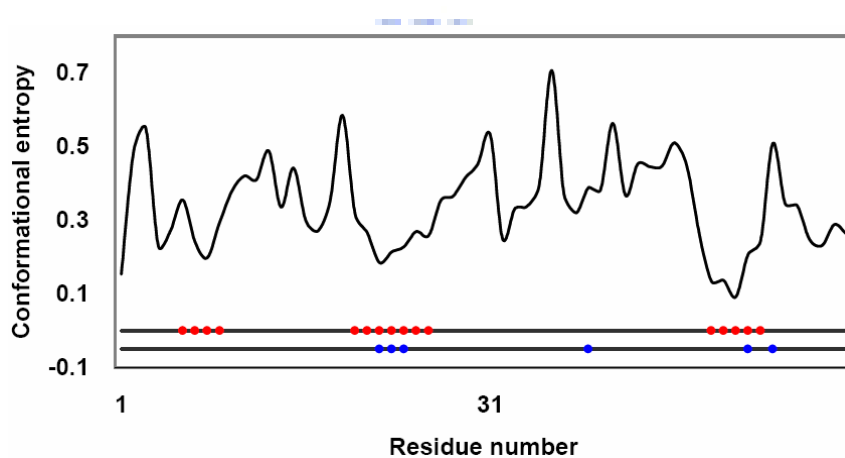


Figure 1.15 Conformational entropy profile of CTX III. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.15 shows the conformational entropy profile of CTX III. Our calculation of low conformational entropy residues covers most of the slow exchange residues (Y22-M24, V52). Figure 1.16 shows the spatial distribution of residues with the slowest exchanging protons and the lowest conformational entropy on the ribbon diagram of cardiotoxin III.

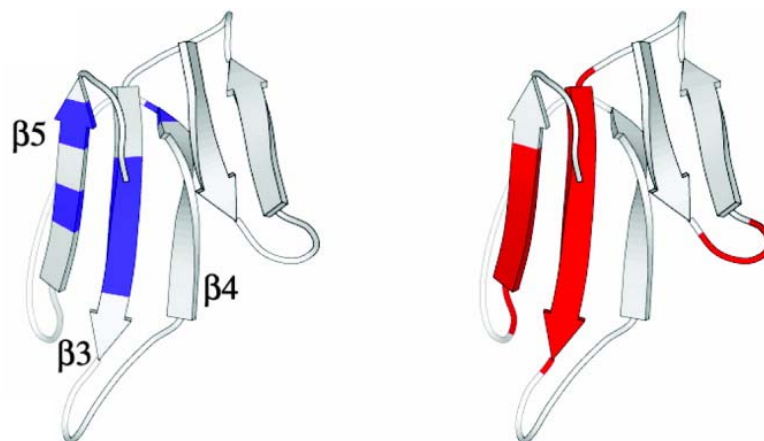


Figure 1.16 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of CTX III.



Ribonuclease H

Ribonuclease H (RNaseH) is a 155-residue protein with four α -helices packing with a five-stranded β -sheet. A previous study⁴² showed that helix A (T43-L56) and helix D (V101-L111) are the most stable regions in the protein.

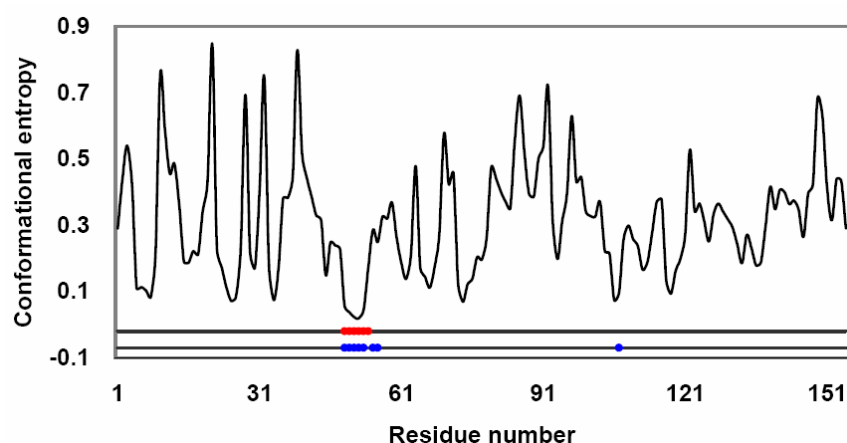


Figure 1.17 Conformational entropy profile of ribonuclease H. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.17 shows the conformational entropy profile of RNaseH. The slowest exchanging residues are L49-I53 and A55-L56 which are located on helix A. The conformational entropies of L49-V54 are the lowest from our calculation. L107 on helix D also has slow exchanging rate and its conformational entropy is relative low. Figure 1.18 shows the spatial distribution of residues with the slowest exchanging protons and the lowest conformational entropy on the ribbon diagram of RNaseH.

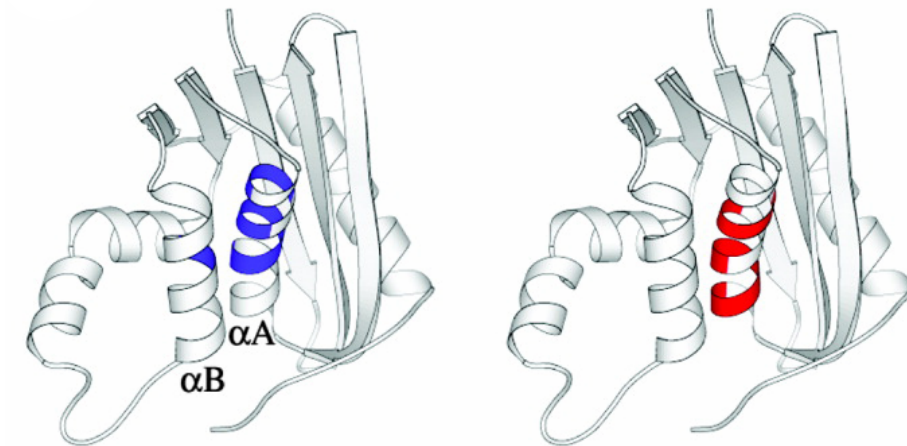


Figure 1.18 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of ribonuclease H.



Bovine pancreatic trypsin inhibitor

Bovine pancreatic trypsin inhibitor (BPTI) is a 58-residue protein composed of two α -helices and a two-stranded anti-parallel β -sheet. Hydrogen exchange study⁴³ showed that the slowest exchanging residues are R20-Y23, Q31, F33, and F45.

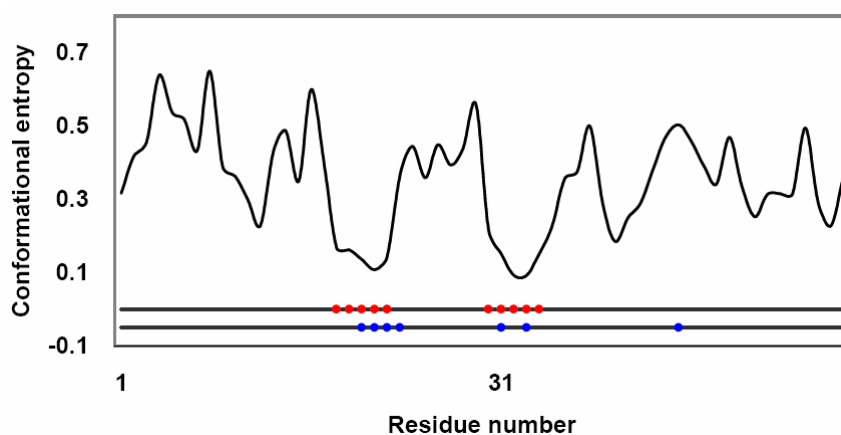


Figure 1.19 Conformational entropy profile of BPTI. Residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) are labeled.

Figure 1.19 shows the conformational entropy profile of BPTI. The residues with the lowest conformational entropy are I18-F22 and C30-V34 which are located on the two β -strands. The slow exchanging residues and low conformational entropy residues overlap well. Figure 1.20 shows the spatial distribution of residues with the slowest exchanging protons and the lowest conformational entropy on the ribbon diagram of BPTI.

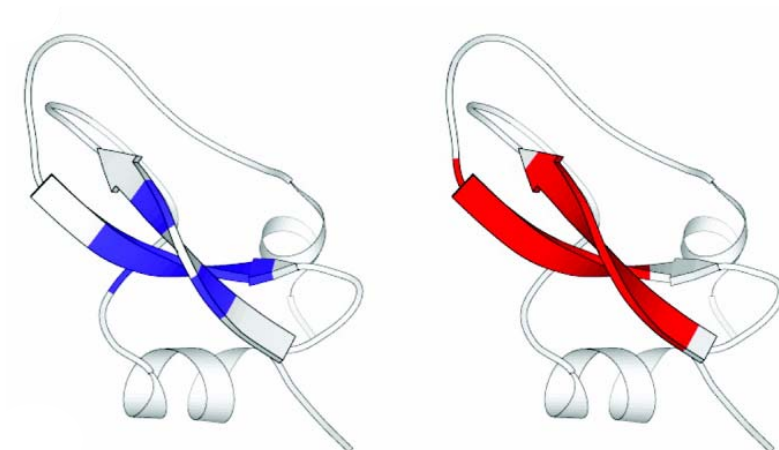


Figure 1.20 Spatial distribution of residues with the slowest exchanging protons (blue) and the lowest conformational entropy (red) on the ribbon diagram of BPTI.



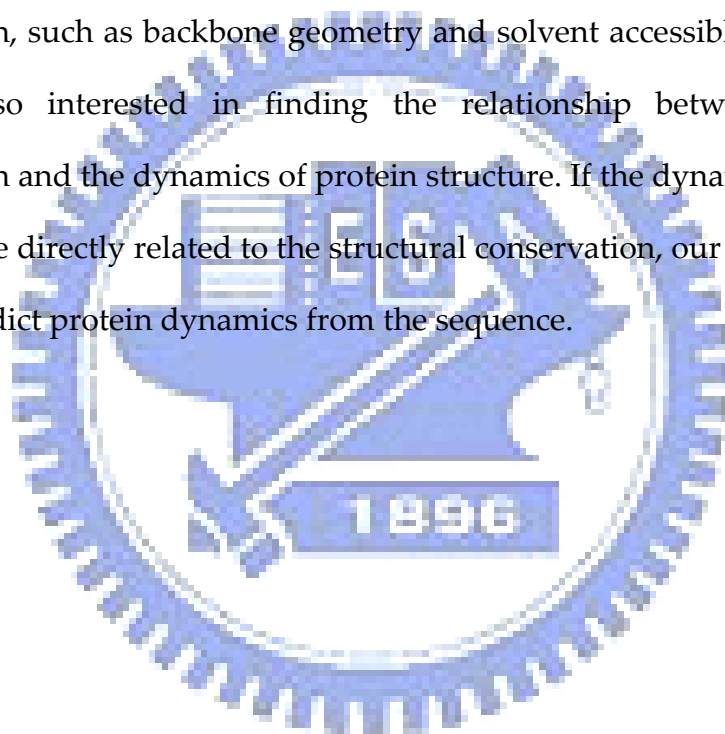
Discussion

We developed a method to compute structure conservation from protein sequence. The basic idea is to know whether the secondary structure of a residue is predictable from its context. Although chameleon sequences are usually thought to be a problem in the prediction of secondary structures using sequence homology^{1,6,44}, we instead use this kind of property to develop our method to compute structure conservation. The residual property of the predictability is quantified to a value, the conformational entropy. The secondary structure of the structure-conserved residues can be predicted more easily than that of the structurally variable residues. As a consequence, the structure-conserved residues have lower conformational entropy.

We applied our method to a set of proteins with known hydrogen-exchange rate data and found a strong correlation between structure conservation and slow hydrogen-exchange rate. The hydrogen-exchange rates are measurement of the exchanging rate of backbone amide protons. They are involved in the formation of secondary structures and related to the local structure stability. For example, in the typical α -helices, strong hydrogen bonds are formed between the backbone oxygen atom of residue i and the backbone nitrogen atom of residue $i+3$. The atoms which are located in rigid environment or involving in more complicated hydrogen-bond networks are usually observed to have slow

hydrogen-exchange rates. Our results show that these atoms are more structurally conserved.

Our method only considered the local interactions. On the other hand, structure conservation is also affected by long-range interactions. It would be interesting to know how long-range interactions contribute to the property of structure conservation. In addition to secondary structures, there are other attributes of protein structure can be used to describe the structure conservation, such as backbone geometry and solvent accessible surface area. We are also interested in finding the relationship between structural conservation and the dynamics of protein structure. If the dynamics of protein structure are directly related to the structural conservation, our method can be used to predict protein dynamics from the sequence.



CHAPTER TWO

NMR Order Parameter Prediction Using Protein-fixed-point Model and Weighted Contact Number Model

Introduction

One of the most important topics in biological science is to understand the protein function. It is well-known that protein dynamics is closely related to the function of protein. For example, the Aquifex adenylate kinase changes its conformation for the phosphotransfer function. Figure 2.1 shows the structure of the protein before (red) and after (green) the conformational change. The two lids must close to exclude bulk water from the active site and to bring the substrate into position for phosphotransfer.

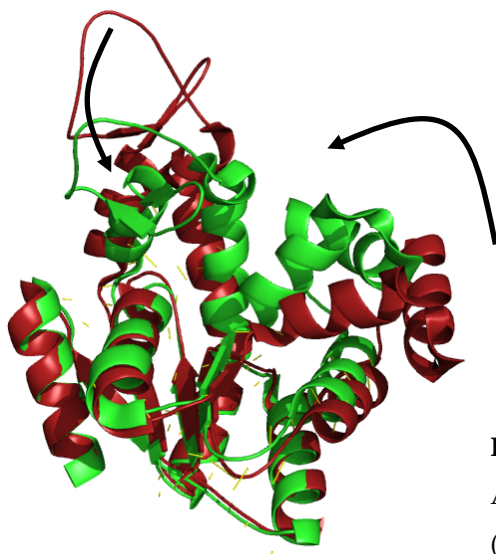


Figure 2.1 Superposition of apo *Aquifex* Adk (red) and *Aquifex* Adk in complex (green)

The advance of X-ray diffraction technology gives us the information about how the protein moves, i.e. B-factor or temperature factor, in the experiment environment. However, the development of computational theories to predict protein dynamics directly from structure is still needed:

1. Only a small fraction of protein structures are solved (52013 solved structures in PDB⁵ due to July 31, 2008).
2. The X-ray B-factors are heavily affected by the experimental conditions: temperature, solvent, or biological unit.

Several computational methods have been developed to determine the protein dynamics. Molecular dynamics (MD) simulation has been widely used in the study of protein function and dynamics. It simulates the interactions between each atom, bonding force, van der Waals force, charge-charge interaction, etc. The computation time is extremely long when the size of the protein is large and the selection of appropriate parameters of force field itself is a complicated problem. Gaussian network model (GNM)⁴⁵ transfers the protein structure into a network in which each C α atom pair is connected together if their distance is smaller than a given cutoff value. Using this protein-converted network, GNM can compute the theoretical thermal fluctuation of each atom and the correlation of motions between each atom pair. Recently we have developed a model to predict the thermal fluctuation from protein structure, which is called protein-fixed-point (PFP) model⁴⁶. The PFP model only uses the coordinates of C α atoms and simply determines the center of mass of the protein. We found that the thermal fluctuation is

proportional to the squared distance from the atom to the center of mass of the structure. Another model called weighted contact number (WCN) model⁴⁷ computes the number of neighboring atoms weighted by the inverse distance between each atom pair. The PFP and WCN models show that the protein dynamics can be extracted directly from the intrinsic property of protein structure without the use of any mechanical model.

The order parameter according to the NMR experiment is widely used to study the dynamic-related protein functions. Zhang and Brushweiler⁴⁸ used a contact model to predict the order parameter from protein structure. They expressed the backbone S^2 order parameter as a function of close contacts between the amide proton and the carbonyl oxygen of the preceding amino acid and the surrounding protein atoms, i.e.,

$$S_i^2 = \tanh[\alpha(\sum_k e^{-r_{i-1,k}^O/\rho} + \beta e^{-r_{i,k}^H/\rho})] + \gamma \quad (2.1)$$

where $r_{i-1,k}^O$ is the distance between the carbonyl oxygen of residue $i-1$ and heavy atom k and $r_{i,k}^H$ is the distance between the amide proton and heavy atom k . The parameter α , β and γ were determined empirically and ρ is set to 1 Å. We will refer to this method the contact model (CM). Though Eq. 2.1 is quite simple, it predicts quite accurate order parameters. Later, to take into account of motional correlation effects, Bruscheweiler and coworker developed a hybrid between the CM and the elastic network model (ENM)^{45,49} referred to as reorientational contact-weighted elastic network model (rCENM)⁵⁰.

Here, we use the PFP and WCN model to predict the N-H S^2 order parameter directly from the protein structure. Our results show that the WCN

model can more accurately reproduce the experimental order parameter than the previous published results.



Methods

Protein-fixed-point Model (PFP)

Let \mathbf{X}_0 be the center of mass of the protein,

$$\mathbf{X}_0 = \frac{\sum_k m_k \mathbf{X}_k}{\sum_k m_k} \quad (2.2)$$

where m_k and \mathbf{X}_k are the mass and the crystallographic position of atom k , respectively. The squared distance of atom i from the center of mass of the protein is computed by

$$r_i^2 = (\mathbf{X}_i - \mathbf{X}_0)(\mathbf{X}_i - \mathbf{X}_0) \quad (2.3)$$

Each protein of size N will have its distinct distribution given by $(r_1^2, r_2^2, \dots, r_N^2)$, referred as the r^2 profile. The squared distance r_i^2 is directly related to the predicted order parameter. In order to confine the computed order parameter S^2 to lying between 0 and 1, we apply the hyperbolic tangent function to r_i^2 .

$$S_i^2 \sim 1 - \tanh^2 r_i^2 \quad (2.4)$$

We will refer this method as the protein-fixed-point (PFP) model.

Weighted-Contact Number Model (WCN)

We define the weighted protein contact number as

$$v_i = \sum_{j \neq i}^N 1/r_{ij}^2 \quad (2.5)$$

where r_{ij} is the distance between C α atoms of residue i and j . This equation defines the number of neighboring C α atoms surrounding that of the i^{th} residue -- the contribution of each surrounding atom j to the central atom i is scaled down by the factor $1/r_{ij}^2$. In order to confine the computed order parameter S^2 to lying between 0 and 1, we apply the hyperbolic tangent function to v_i .

$$S_i^2 \sim \tanh^2 v_i \quad (2.6)$$

We will refer this method as the weighted protein contact-number (WCN) model.

Dataset

We used the datasets of Zhang and Bruschweiler⁴⁸ and Ming and Bruschweiler⁵⁰. However, our dataset is not completely identical with the original one, since we deleted some of the order parameters that are not consistent with those of the original dataset, and we also added some new ones from current literatures. Our current dataset is larger than the original dataset and comprises β ARK1 PH domain (1BAK), calbindin (4ICB), CspA (3MEF), frenolicin acyl carrier protein (1OR5), lysozyme (1JEF), P85 α SH2 domain (1BFJ), ubiquitin (1UBQ), ketosteroid isomerase (8CHO), tautomerase (4OTA), and interleukin-4 (1HIK).

Results

We computed S^2 values for the following proteins: β ARK1 PH domain (1BAK), calbindin (4ICB), CspA (3MEF), frenolicin acyl carrier protein (1OR5), lysozyme (1JEF), P85_SH2 domain (1BFJ), ubiquitin (1UBQ), ketosteroid isomerase (8CHO), tautomerase (4OTA), and interleukin-4 (1HIK). Table 2.1 summarizes the Pearson correlation coefficients between the experimental and the computed N-H S^2 order parameters by the WCN, PFP, CM, and rCENM.

Table 2.1 The experimental backbone N-H order parameter data: β ARK1 PH domain⁵¹, Calbindin⁵², CspA⁵³, Frenolicin acyl carrier protein⁵⁴, Lysozyme⁵⁵, P85 α SH2 domain of phosphoinositide 3-kinase⁵⁶, Ubiquitin⁵⁷, Ketosteroid isomerase⁵⁸, Tautomerase⁵⁹, and Interleukin-4⁶⁰. * The average correlation coefficient over the 7 structures for WCN, PFP, CM and rCENM are 0.82, 0.74, 0.73 and 0.81, respectively.

Protein	PDB	WCN	PFP	CM	rCENM
β ARK1 PH domain	1BAK	0.83	0.83	0.53	0.84
Calbindin	4ICB	0.75	0.79	0.65	0.72
CspA	3MEF	0.78	0.74	0.71	--
Frenolicin acyl carrier protein	1OR5	0.85	0.81	0.89	0.87
Lysozyme	1JEF	0.83	0.76	0.72	0.68
P85 α SH2 domain	1BFJ	0.86	0.76	0.79	--
Ubiquitin	1UBQ	0.96	0.92	0.96	0.97
Ketosteroid isomerase	8CHO	0.82	0.75	0.57	0.78
Tautomerase	4OTA	0.51	0.28	0.44	--
Interleukin-4	1HIK	0.71	0.57	0.81	0.81
Average		0.79	0.72	0.71	--*

The average correlation coefficient of WCN model is 0.79, while that of the CM is 0.71. The average correlation coefficient of rCENM is 0.81 (for the 7 available proteins), while that of WCN is 0.82. In general, the WCN model

performs better than the CM, except two cases, 1OR5 and 1HIK. The results of WCN is comparable to the rCENM, however, there is no available software for rCENM which can be used to make a complete comparison. The following sections discuss each case in the dataset.

β ARK1 PH domain

β ARK1 PH domain has the same topology as other PH domains, which are characterized by several β -strands forming a β -sandwich flanked on one side by an extended C-terminal α -helix that behaves as a molten helix⁵¹. The Pearson correlation coefficient between the computed N-H order parameter S_{WCN}^2 and the experimental one S_{NMR}^2 is 0.83. Figure 2.2 shows the experimental and predicted order parameters for β ARK1 PH domain by using WCN, CM, and PFP models.

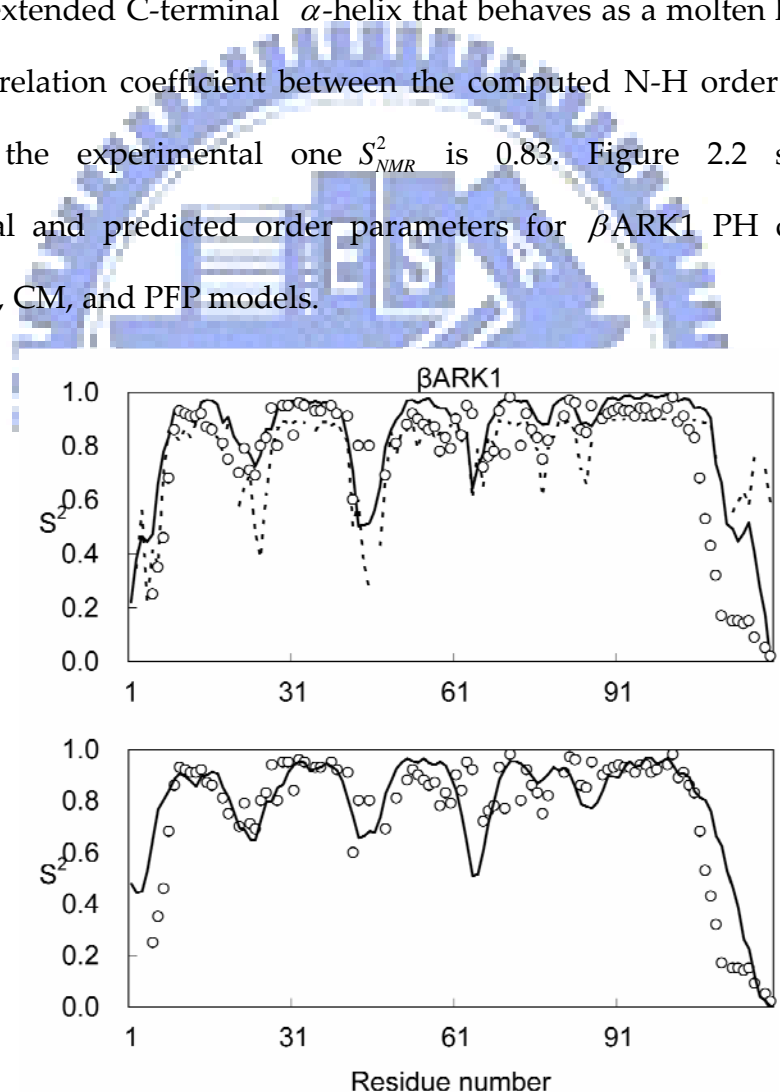


Figure 2.2 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Calbindin

Calbindin D_{9k}⁵² is composed of four α -helices, the N-terminal (E17-S24) and C-terminal Ca²⁺-binding loops (D54-S62), and the linker loop. Our prediction correctly identifies the most mobile linker loop and the C-terminal Ca²⁺-binding loop which have significant lower S^2 values. The rigid helical regions are also predicted to have higher order parameters. The experimental data does not produce the S^2 value of the residue P20 on the N-terminal Ca²⁺-binding loop because of the limit of NMR relaxation experiment. However P20 shows higher temperature factor than its neighboring residues in the X-ray⁶¹, which is consistent with our prediction. Despite the missing data of P20, the correlation coefficient is still high ($r=0.79$). Figure 2.3 shows the experimental and predicted order parameters for calbindin by using WCN, CM, and PFP model.

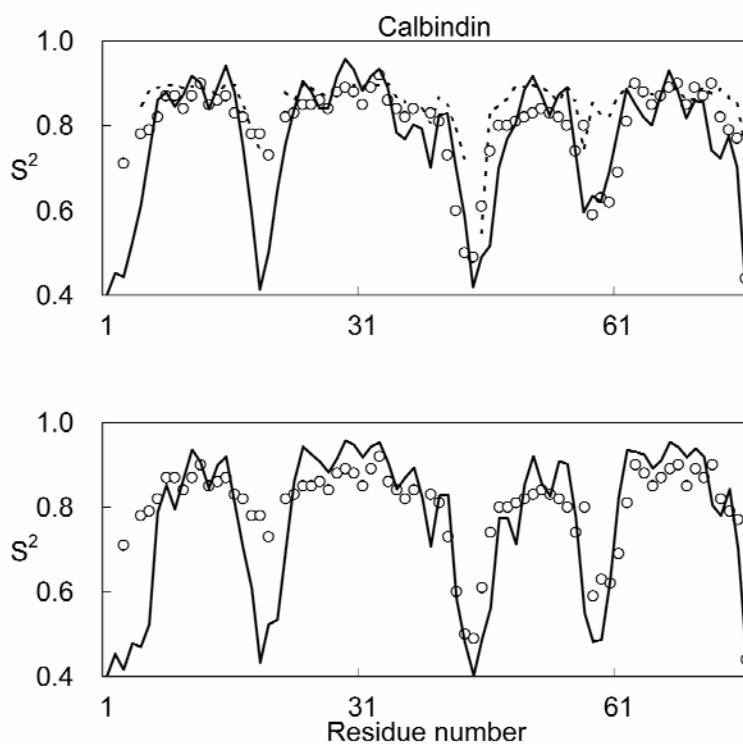


Figure 2.3 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Cold-shock protein A from *E-coli*

Cold-shock protein A from *E-coli* (CspA)⁵³ is a Greek-key β -barrel protein. The segment of residues N39 to Y42 between two β -strands is identified to be partially disordered in the crystallization environment⁵³. Our method successfully predicts it to be the most mobile region in the protein except the N-terminal loop. However there is an disagreement between the predicted value and the experimental value on residue D46 (S_{exp}^2 : 0.58, S_{WCN}^2 : 0.84) which did not fit well with any model in the NMR experiment⁵³. The correlation coefficient increases to 0.78 (from $r=0.74$) if the data of residue D46, which is less reliable, is removed. Figure 2.4 shows the experimental and predicted order parameters for CspA by using WCN, CM, and PFP model.

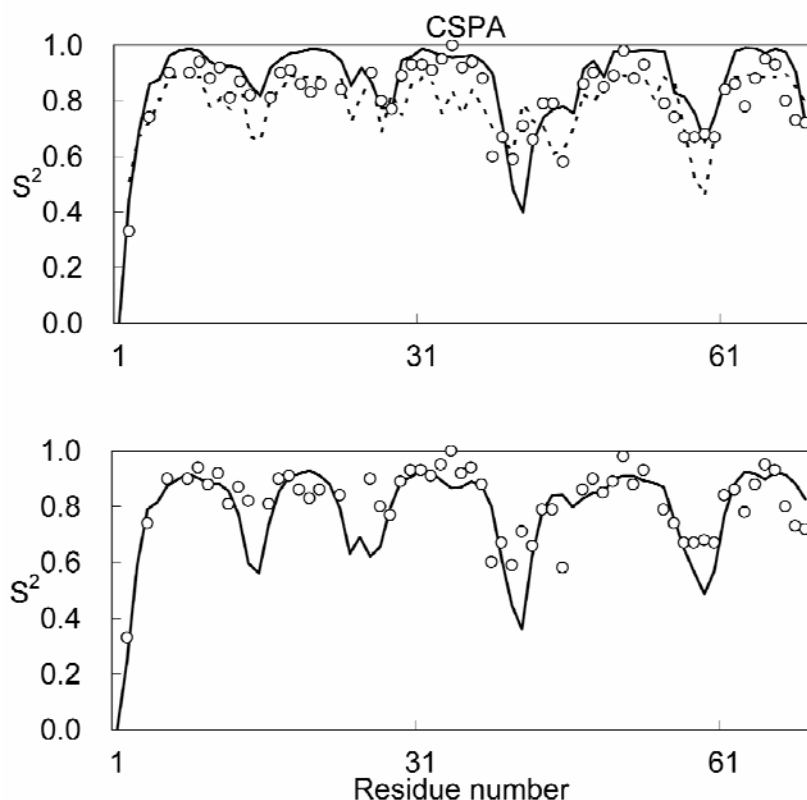


Figure 2.4 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Frenolicin acyl carrier protein

Frenolicin acyl carrier protein (frenolicin ACP)⁵⁴ is comprised of a three-helix bundle structure and have a high correlation coefficient between prediction and experiment ($r=0.81$). The average value of the order parameters of the three helices is 0.844, which is consistent with our prediction having high S^2 values. We also correctly predict that the C-terminal residues and the long loop (G17-D23) connecting two helices have the first and second lowest average S^2 values respectively (0.358 and 0.492). Figure 2.5 shows the experimental and predicted order parameters for frenolicin ACP by using WCN, CM, and PFP model.

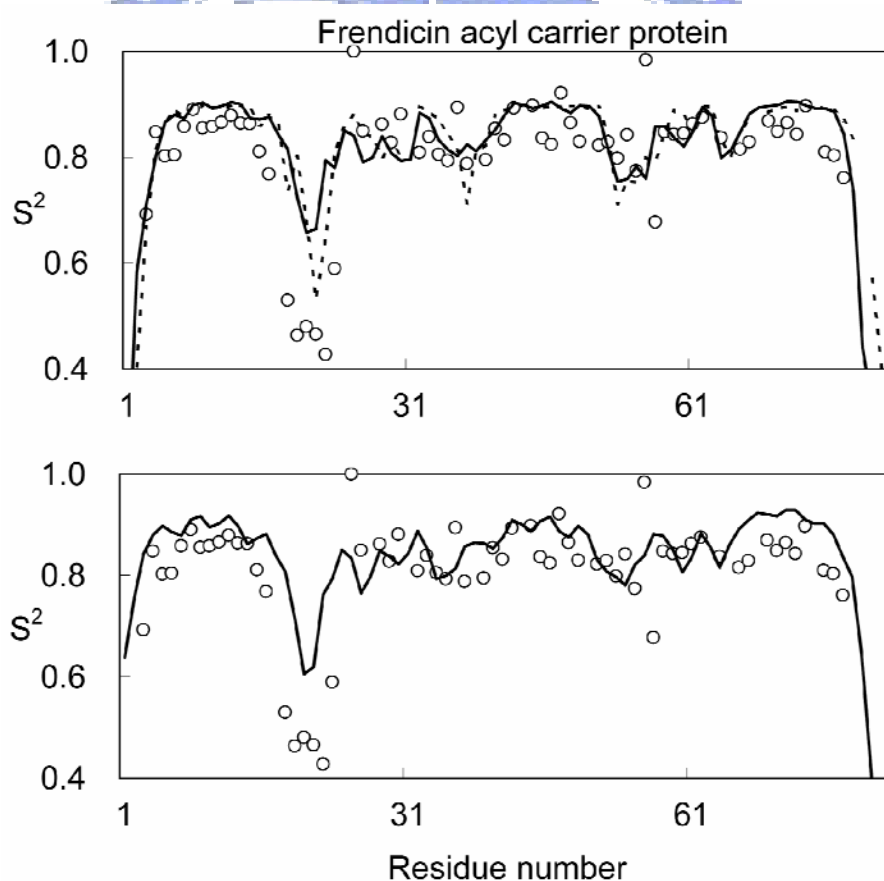


Figure 2.5 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Lysozyme

The correlation coefficient of the prediction for lysozyme⁵⁵ is 0.83. The order parameter of residue P70 located on the most flexible loop is not available from NMR relaxation. According to the X-ray structure of lysozyme, P70 has the sixth highest temperature factor (28.37) in the whole protein⁶² (the four residues having the highest temperature factors are on the C-terminal region). Our prediction also shows that P70 is the most flexible except a few residues located on the C-terminal. Figure 2.6 shows the experimental and predicted order parameters for lysozyme by using WCN, CM, and PFP model.

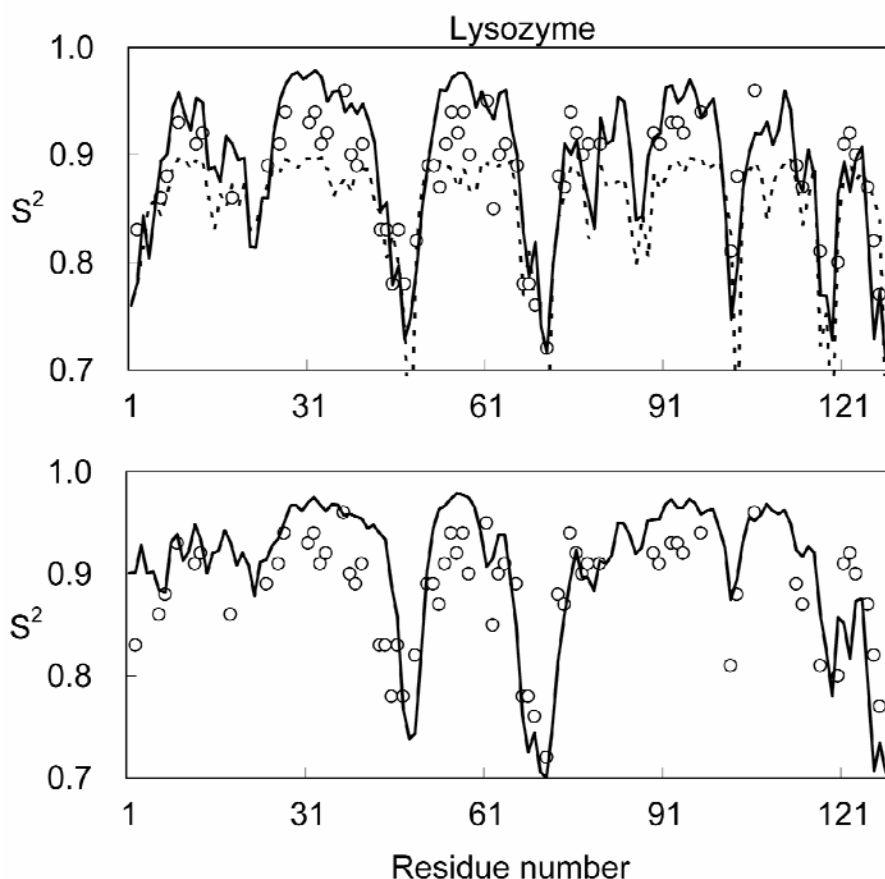


Figure 2.6 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

SH2 domain of p85 α subunit of phosphoinositide 3-kinase

The SH2 domain of the p85 α subunit of phosphoinositide 3-kinase contains 111 residues forming a structure with 3 α -helices and a 3-strand anti-parallel β -sheet. The correlation coefficient between the experimental value and predicted value of WCN model is 0.86, which is higher than that of the PFP model ($r=0.76$) and the CM ($r=0.79$). Figure 2.7 shows the experimental and predicted order parameters for the protein by using WCN, CM, and PFP model. The result for this protein predicted with the WCN model is basically excellent.

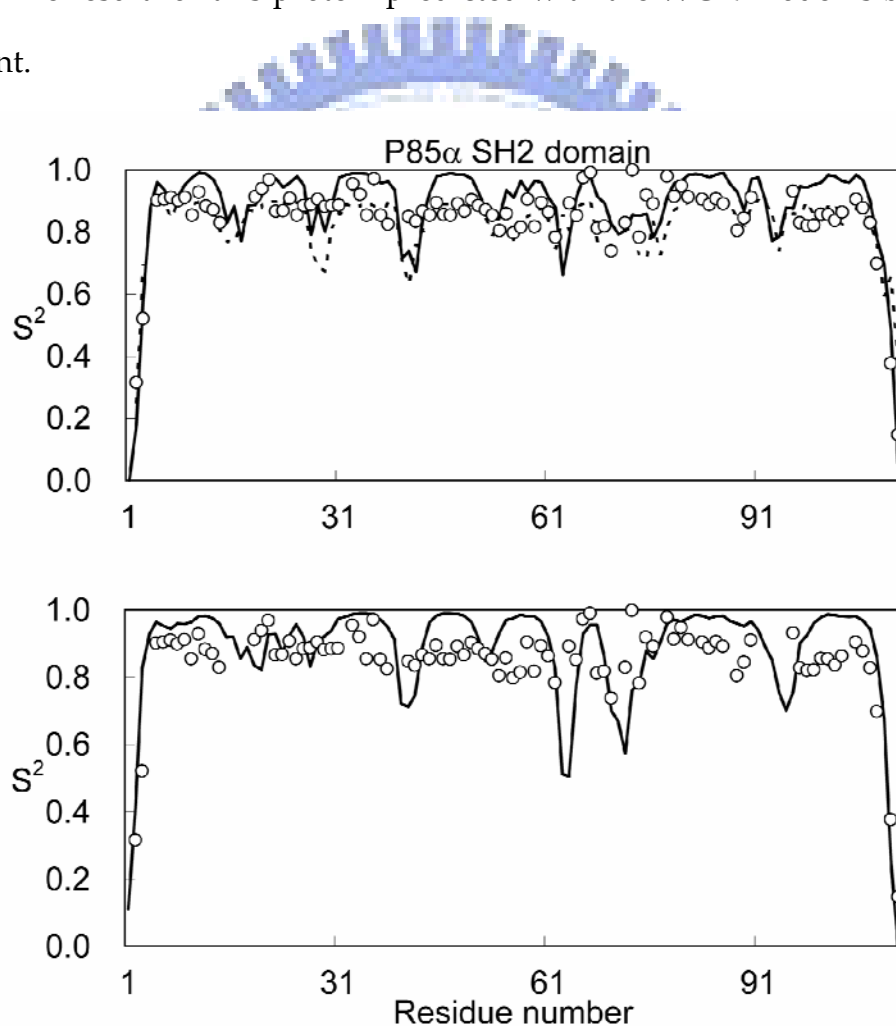


Figure 2.7 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Ubiquitin

Ubiquitin⁵⁷ is a small single-domain protein with 76 residues containing both an α -helix and a β -sheet. The agreement between S_{WCN}^2 and S_{NMR}^2 of ubiquitin is excellent ($r=0.96$). Figure 2.8 shows the experimental and predicted order parameters for ubiquitin by using the WCN, CM, and PFP model. The CM and PFP model also yield excellent results with correlation coefficients of 0.96 and 0.92, respectively.

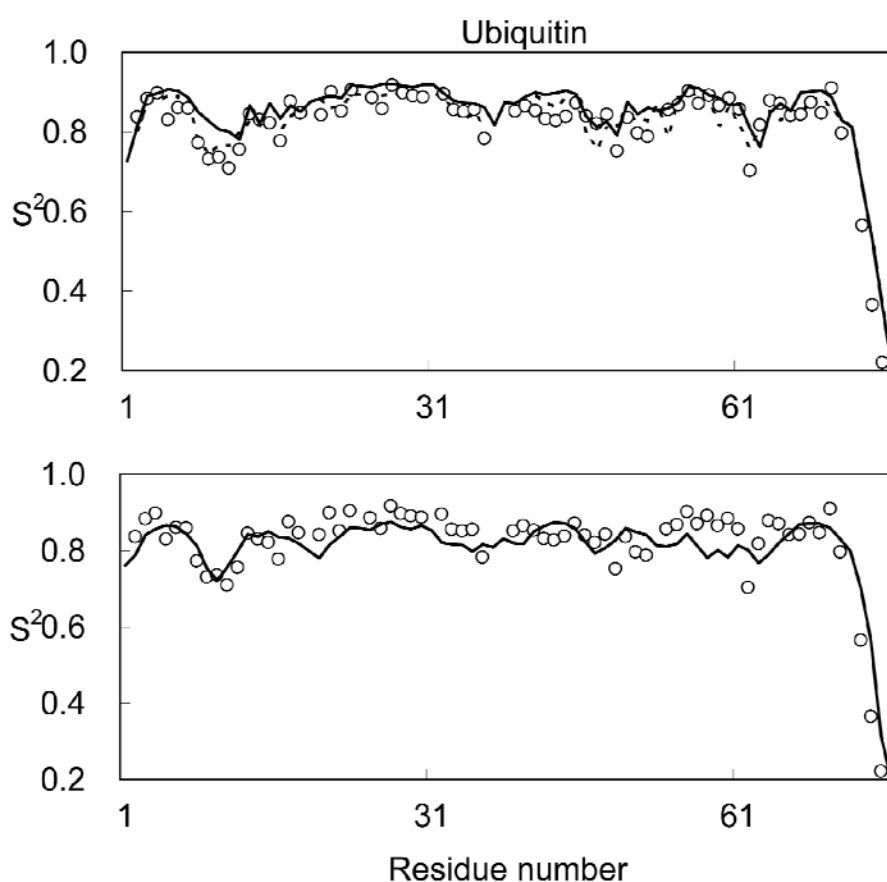


Figure 2.8 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Ketosteroid isomerase

The ketosteroid isomerase is a 125-residue protein comprising 2 α -helices and 2 anti-parallel β -strands. The correlation coefficient between experimental value and predicted value of the WCN model is 0.82, which is higher than that of the PFP model ($r=0.43$) and CM ($r=0.57$). Figure 2.9 shows the experimental and predicted order parameters for ketosteroid isomerase by using the WCN, CM, and PFP model. Both the PFP and WCN model predict the Y88-K92 region to be flexible (even the CM). The region is actually an isolated loop connecting two β -strands. The reason for the disagreement between experimental and predicted value may be that the protein forms a dimer under experimental condition and the loop is located and stabilized at the interface between two subunits.

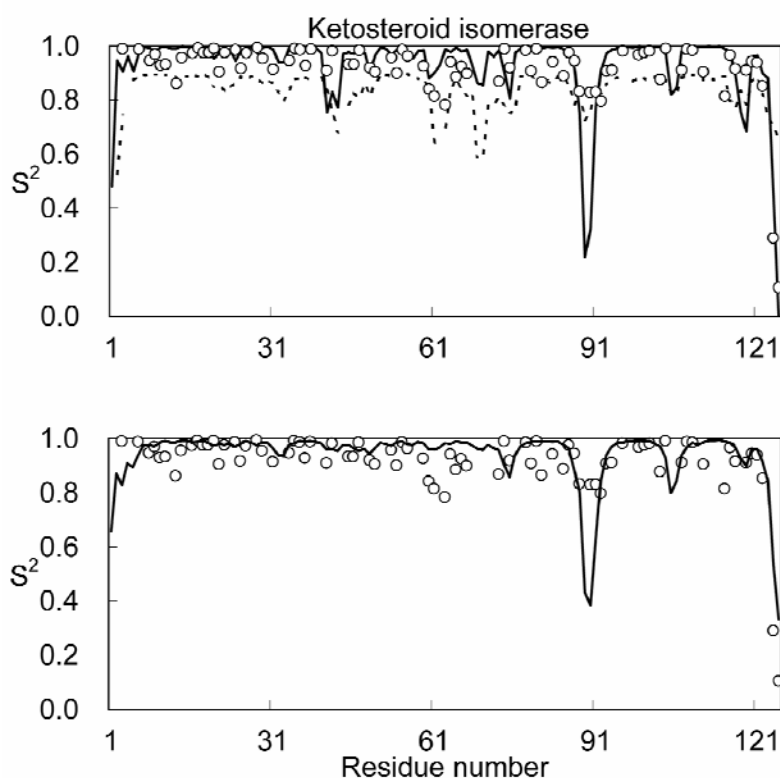


Figure 2.9 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

4-oxalocrotonate tautomerase

The 4-oxalocrotonate tautomerase is a homohexamer with 62 residues per unit, consisting of an α -helix, 2 β -strands, a β -hairpin, 2 loops, 2 turns, and a C-terminal coil. The correlation coefficient between the experimental and predicted value of the WCN model is 0.51, which is higher than that of the PFP model ($r=0.28$) and CM ($r=0.44$). The PFP model predicts the S28-R39 loop region to be flexible, however, the region is actually highly ordered according to the NMR experiment. The reason may be that the loop is away from the center of the protein but the environment around it is crowded. The WCN model correctly predicts the loop region to be highly ordered. Figure 2.10 shows the experimental and predicted order parameters for 4-oxalocrotonate tautomerase by using the WCN, CM, and PFP model.

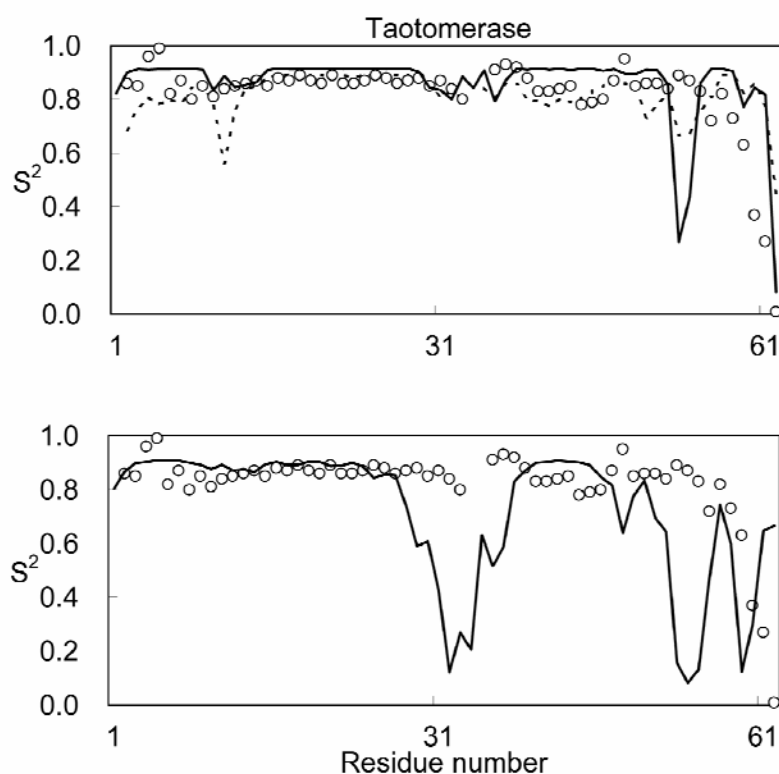


Figure 2.10 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Interleukin-4

In the original paper for the experimental NMR data of interleukin-4⁶⁰, the authors mentioned an uncommon loop region (H59-A70) which has high order parameters. The observed average order parameter for the loop is 0.87 which is significantly higher than those of other loops and is only slightly lower than that observed for the secondary structure regions. The uncommon loop is predicted to have low order parameter for the WCN model, however, the correlation coefficient increases to 0.78 if the loop region is removed. Figure 2.11 shows the experimental and predicted order parameters for Interleukin-4 by using the WCN, CM, and PFP model.

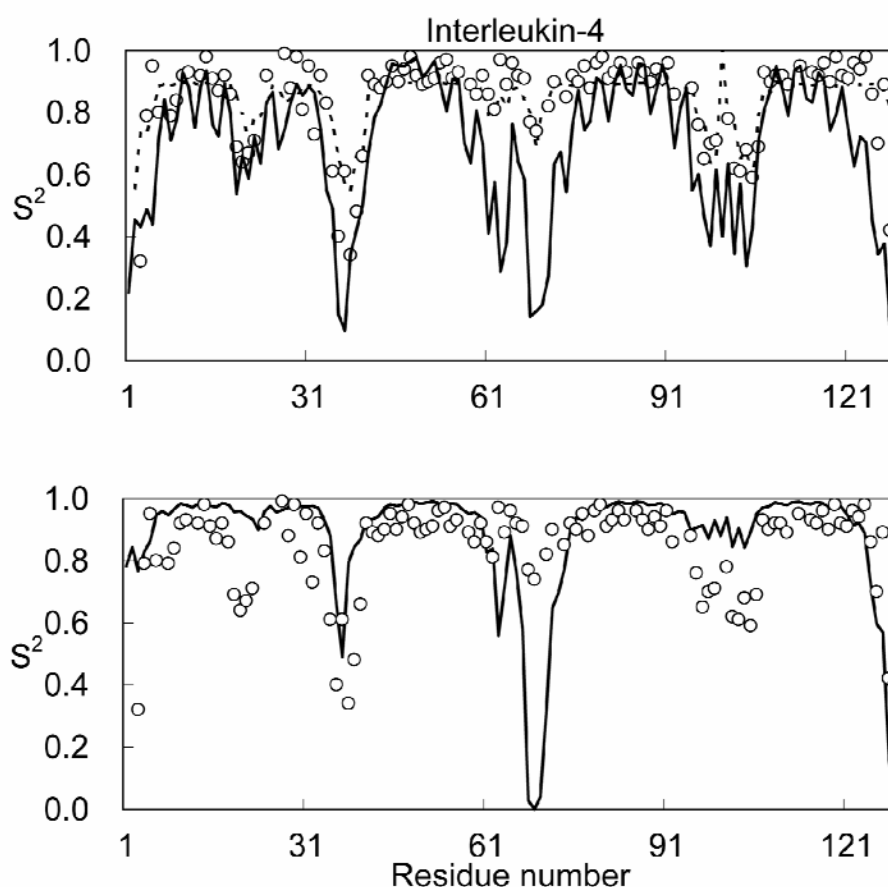


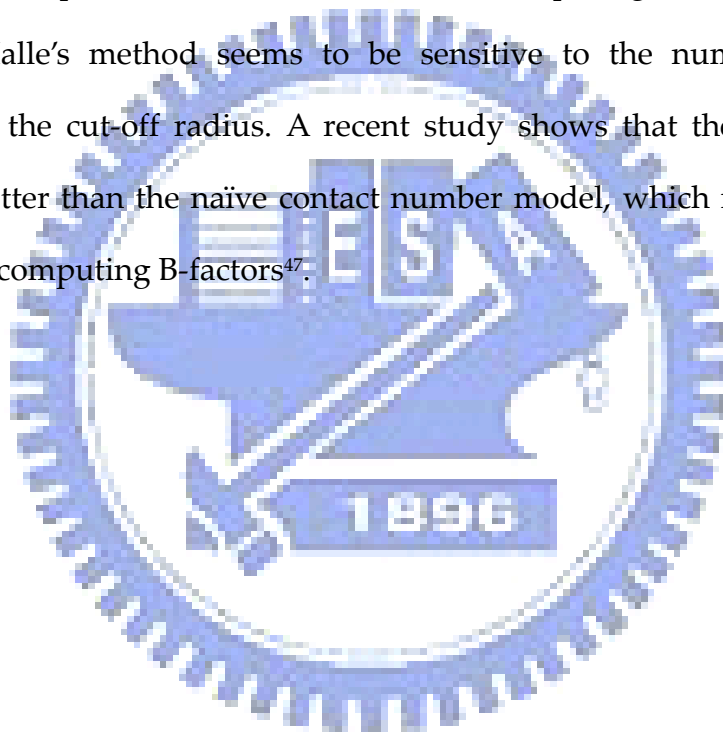
Figure 2.11 Upper part: computed S^2 values of WCN model (solid line) and CM (dotted line) and experimental S^2 values (circle). Lower part: computed S^2 values of PFP model (solid line) and experimental S^2 values (circle).

Discussion

Our results show that the backbone dynamics of protein structures can be directly inferred from the static structural properties without the assumption of any additional mechanical model. Since we can compute the order parameters directly from the topological properties (such as protein contact number) of protein structures, our study underscores a very direct link between protein topological structure and its dynamics. In addition, since our model uses only $C\alpha$ atoms, our results indicate that protein dynamics (such as the order parameters) can be determined without the knowledge of protein sequences. As increasing numbers of protein structures are solved, this method offers an efficient way to determine backbone motions with high accuracy and is practical in the study of protein function-dynamics relationship and structural genomics.

To check the effects of the additional information of side-chain groups on the computed order parameters, we compared the computed S^2 values with the side-chain information. However, the inclusion of the side-chain groups deteriorates the performance of the WCN model (\bar{r} goes down from 0.79 to 0.73). The reason for this is not clear. It may be that the flexible side-chain conformations, though conveying more detailed information about the atomic environments, introduce undesirable noises that overshadow the supposedly useful information of the former in the computation of the order parameters.

A recent study by Halle⁶³ found a linear correlation between the contact number of non-covalent neighboring atoms and B-factors. Their result shows that the B-factor is inversely proportional to the contact density, i.e., the number of non-covalent neighbors. They assume that the contact density of covalent neighbors for each $C\alpha$ atom is similar. Therefore, only the non-covalent neighbors are needed to be considered. Our method has the advantage of not using the cut-off distance, while Halle's method needs to determine the optimal cut-off distance when computing contact number. In addition, Halle's method seems to be sensitive to the number of atoms included in the cut-off radius. A recent study shows that the WCN model performs better than the naïve contact number model, which needs a cut-off distance, in computing B-factors⁴⁷.



CHAPTER THREE

Normal Mode Analysis by the Protein-fixed-point Model

Introduction

The biological function of proteins is closely related to the cooperative motions and the correlated fluctuations which involve large portions of the structure⁶⁴⁻⁶⁷. Normal Mode Analysis (NMA) had been used to study biomolecules since early 1980s^{68,69}. It decomposes the protein dynamics into a collection of motions which include large scale/low frequency and small scale/high frequency motions. Biologists usually focus on the large scale/low frequency motions that are relevant to protein functions. The major contribution of NMA to the biological research field is the ability to provide the information of large, domain-scale protein motions which is hard to compute by other methods. The classical approach of NMA is to diagonalize the Hessian matrix, i.e. the second derivative of the potential function of a molecular dynamics (MD) simulation. The major shortcoming of the classical NMA is that the sampling time increases dramatically with the size of the protein.

The Elastic Network Model (ENM)⁴⁹, which describes protein dynamics without amino acid sequence and atomic coordinates, has been widely used in the studies of protein dynamics and structure-function relationship. The ENM views the protein structure as an elastic network, the nodes of which are

the $C\alpha$ atoms of individual residues. Residue pairs within a cutoff distance are connected by springs which have a uniform force constant in the network. Based on ENM, a coarse-grained version of NMA is developed and widely used because of its low computation cost and the ability to extend the dynamics to longer timescale and larger motions. The coarse-grained NMA had been applied to various topics, for example, protein functions and catalytic residues. One of the most widely used ENM-based methods is the Gaussian network model (GNM)^{45,70,71}.

The protein-fixed-point (PFP) model⁴⁶ is a simple method to compute the protein dynamics only using the coordinates of $C\alpha$ atoms. Despite its simplicity, the PFP model has been shown to be able to accurately predict the B-factors for a dataset of 972 proteins⁷². The PFP model simply computes the vector r_i between residue i and the center of mass of the protein. The B-factor of residue i has been shown to be correlated to the inner product of r_i itself, i.e. the square of the length of r_i . However, the ability to compute the normal mode motions, i.e. decomposition of dynamics to motions of different frequencies, is more important than calculating theoretical B-factor of single residues. Since the thermal fluctuation of residue i correlates to the auto product of vector r_i , we can get the correlation of motion between residue i and j by computing the inner product of vector r_i and r_j . By simply diagonalizing the correlation matrix, in which the elements are the correlation of motion between residue pairs, the normal mode motions can be obtained. Here, we compared the results of NMA based on the PFP model with those by Gaussian network model (GNM).

Methods

NMA by Protein-fixed-point Model

Let \mathbf{X}_0 be the center of mass of the protein,

$$\mathbf{X}_0 = \frac{\sum_k m_k \mathbf{X}_k}{\sum_k m_k} \quad (3.1)$$

where m_k and \mathbf{X}_k are the mass and the crystallographic position of atom k , respectively. The distance of atom i from the center of mass of the protein is computed by

$$r_i^2 = (\mathbf{X}_i - \mathbf{X}_0)(\mathbf{X}_i - \mathbf{X}_0) \quad (3.2)$$

Each protein of size N will have its distinct distribution given by $(r_1^2, r_2^2, \dots, r_N^2)$, referred as the r^2 profile. The r^2 profile has been shown to be closely related to the B-factors (temperature factors) from X-ray experiment. To compute the correlation of motion r_{ij} between atom i and atom j ,

$$r_{ij} = (\mathbf{X}_i - \mathbf{X}_0)(\mathbf{X}_j - \mathbf{X}_0) \quad (3.3)$$

Each protein of size N will have a correlation matrix of size $N \times N$ which denoted as $C = (r_{11}, r_{12}, \dots, r_{1N}, \dots, r_{NN})$. The correlation matrix is then diagonalized,

$$C = R^T \lambda^{-1} R \quad (3.4)$$

where λ and R are the matrix of eigenvalue and eigenvector, respectively. The eigenvalues are the frequencies and the eigenvectors are the amplitudes of motions of each corresponding modes. Note that the diagonal term r_{ii} of the correlation matrix is equal to the r^2 profile (equation 3.2) which is closely correlated to the atomic thermal fluctuations. In equation 3.4, the eigenvalue λ reversely contributes to r_{ii} , hence, the low frequency modes (with small eigenvalues) dominate the thermal fluctuations.



Results

Comparison of Correlation Maps

We first compared the correlation maps computed by the PFP model and the GNM using a dataset of 18 proteins (Appendix D). Table 3.1 summarized the Pearson correlation coefficient between the correlation maps by the aforementioned two methods.

	CC*		CC
1A16	0.81	1CHD	0.69
1B6A	0.73	1CTT	0.74
1BIO	0.68	1DNK	0.71
1BOL	0.75	1EUG	0.71
1BR6	0.71	1LBA	0.69
1BTL	0.72	1RPT	0.72
1BVV	0.72	1YTW	0.70
1BWP	0.71	8TLN	0.81
1BXO	0.78	9PAP	0.73
Total average			0.73

Table 3.1
The Pearson correlation coefficient between the correlation maps computed by the PFP model and the GNM

*Pearson correlation coefficient

Figure 3.1 shows some cases in the dataset (complete results can be found in Appendix E). Each element in the correlation map represents the correlation of motion of each residue pair. The color blue-green-yellow-red is in the order from negative correlation to positive correlation. In general, the correlation maps computed by the PFP model and the GNM displays similar patterns (average Pearson correlation coefficient: 0.73).

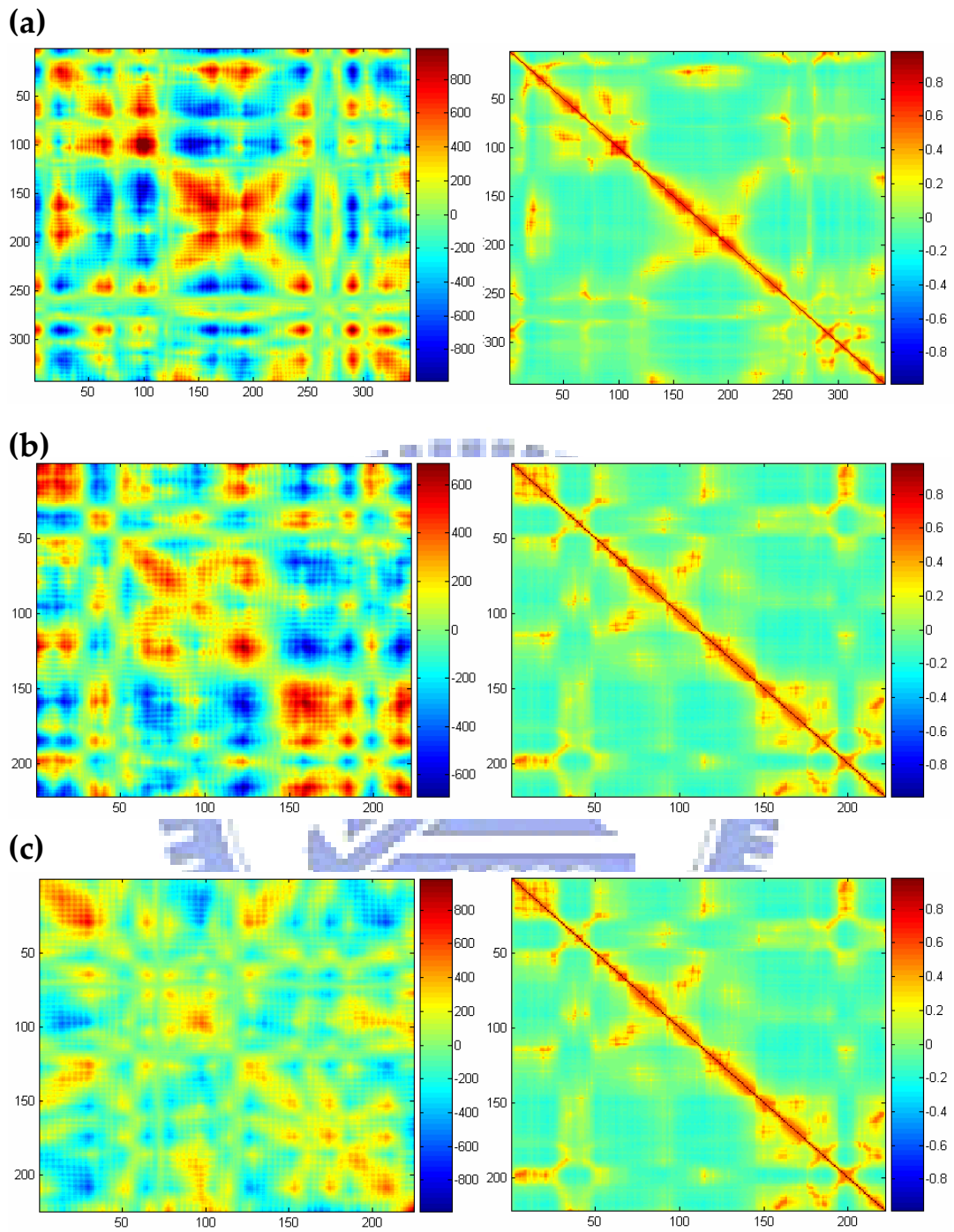


Figure 3.1 The correlation maps of (a) Prot-glu methyltransferase (1RPT), (b) Ribonuclease T2 (1BOL), and (c) Uridine nucleosidase (1EUG) computed with the PFP model (left) and GNM (right), respectively.

Comparison of Normal Mode Motions by the PFP model and GNM

We compared the results of NMA by the PFP model and the GNM. Table 3.2 summarized the correlation coefficient between the eigenvector of the three slowest modes of PFP model and GNM. The average correlation coefficients of the 18 proteins for the three slowest modes are 0.84, 0.63, and 0.45 respectively. In general, the PFP model correlates well with GNM in the first mode, except 1bio ($r=0.58$), 1bv_v ($r=0.50$), and 1chd ($r=0.56$). However, in the second and third modes, the differences between these two models increase.

Table 3.2 The correlation coefficient between the eigenvectors of PFP model and that of GNM in the three slowest modes.

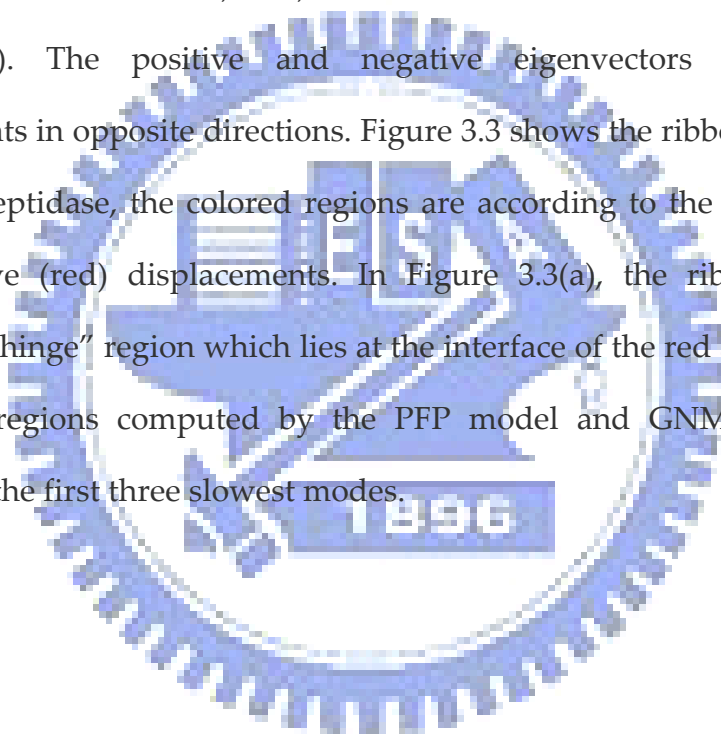
	mode 1	mode 2	mode 3		mode 1	mode 2	mode 3
1A16	0.92	0.57	0.34	1CHD	0.56	0.78	0.93
1B6A	0.91	0.61	0.58	1CTT	0.93	0.87	0.69
1BIO	0.58	0.56	0.45	1DNK	0.95	0.65	0.48
1BOL	0.96	0.91	0.54	1EUG	0.92	0.30	0.25
1BR6	0.88	0.44	0.10	1LBA	0.72	0.66	0.43
1BTL	0.94	0.32	0.44	1RPT	0.94	0.90	0.40
1BVV	0.50	0.43	0.46	1YTW	0.84	0.65	0.33
1BWP	0.74	0.24	0.14	8TLN	0.96	0.64	0.18
1BXO	0.93	0.92	0.55	9PAP	0.97	0.90	0.85
			Total average		0.84	0.63	0.45

Here, we focus on the low frequency motions (also called global or large scale motions) which are opposite to high frequency motions (local motions). The contribution to fluctuation of each mode is scaled by the inverse of the mode's frequency. Hence, the low frequency modes contribute to most of the fluctuations (see equation 3.4).

In the following section, we discuss some cases in the dataset.

Thiol-endopeptidase (9PAP)

Thiol-endopeptidase is a single-domain protein of 212 residues, consisting of 7 α -helices and an anti-parallel β -sheet. Figure 3.2 shows the eigenvector of each residue along the first mode, second mode, and third mode for the PFP model and GNM. The agreement between them is excellent (correlation coefficient: 0.97, 0.90, and 0.85 for the first three slowest modes, respectively). The positive and negative eigenvectors represent the displacements in opposite directions. Figure 3.3 shows the ribbon diagrams of thiol-endopeptidase, the colored regions are according to the positive (blue) and negative (red) displacements. In Figure 3.3(a), the ribbon diagrams illustrate a “hinge” region which lies at the interface of the red and blue parts. The hinge regions computed by the PFP model and GNM are basically identical in the first three slowest modes.



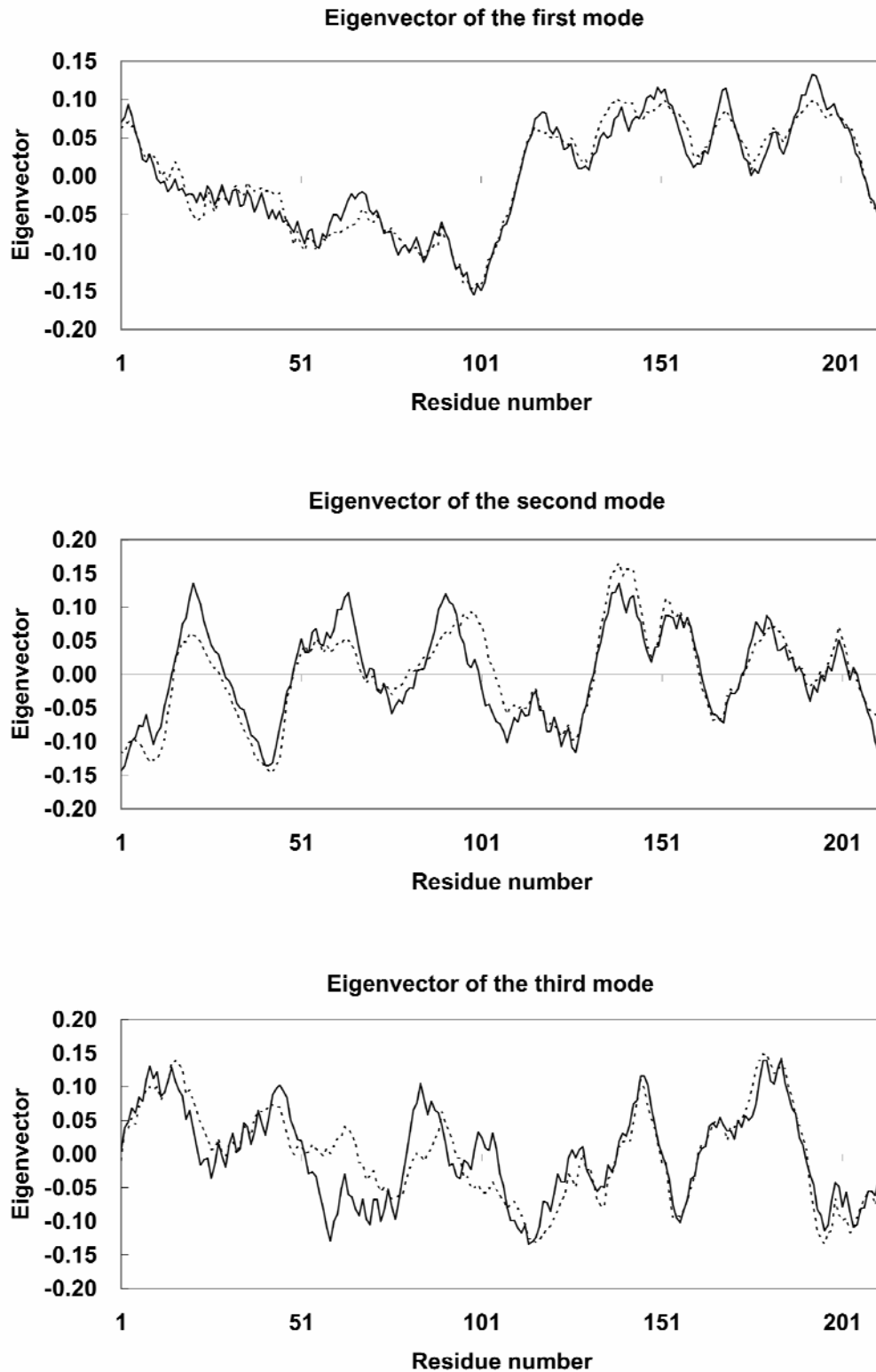


Figure 3.2 Distribution of displacements along the first mode, second mode, and third mode computed for thiol-endopeptidase (9PAP). (solid line: PFP model, dotted line: GNM)

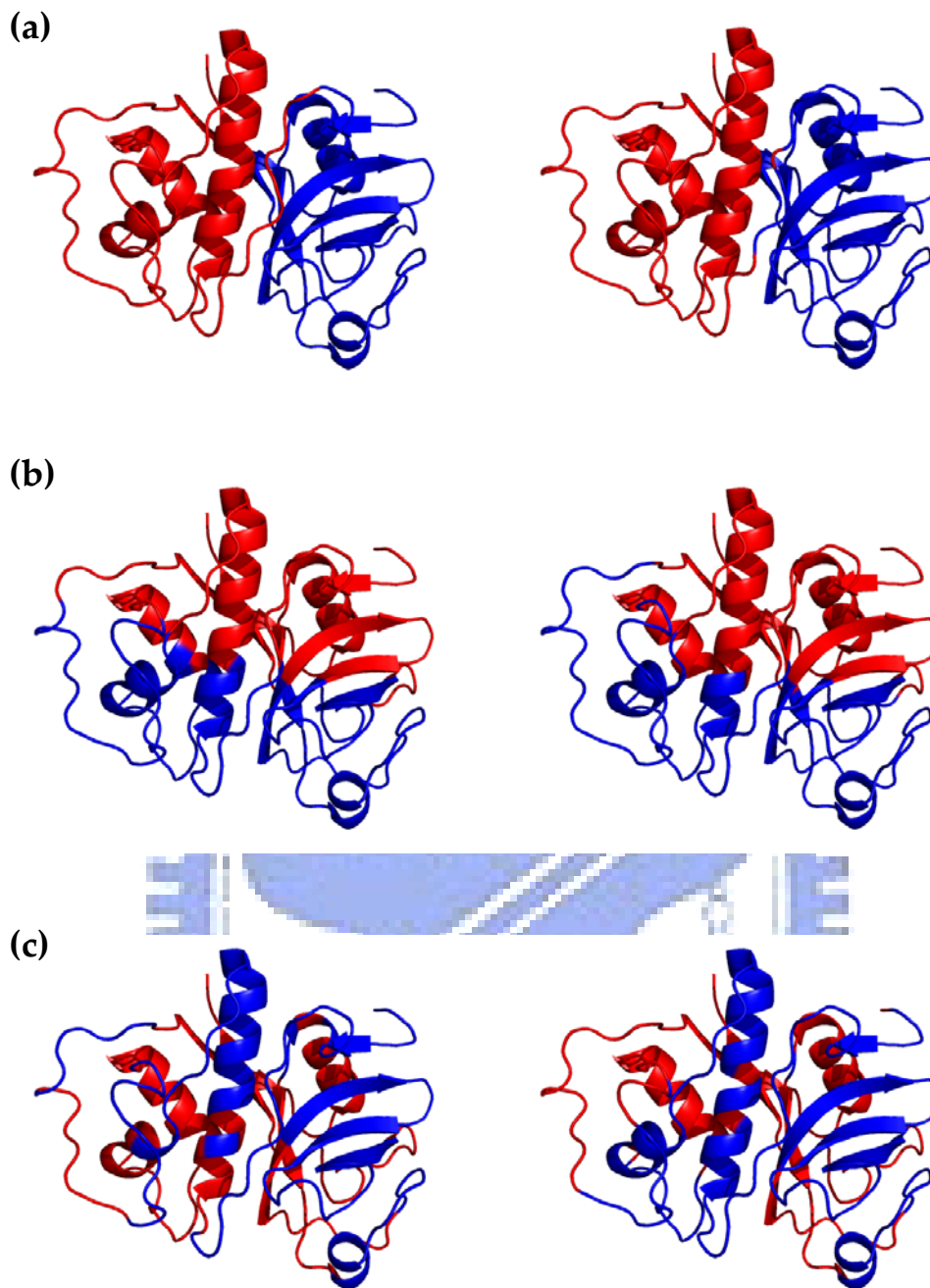


Figure 3.3 The regions subject to opposite direction displacements computed by the PFP model (left) and the GNM (right) of (a) the first mode, (b) the second mode, and (c) the third mode for thiol-endopeptidase (9PAP). Regions colored in blue and red correspond to positive and negative displacements respectively.

G/11 xylanase (1BVV)

Figure 3.5 shows the eigenvector of each residue along the first mode, second mode, and third mode for the PFP model and GNM. The correlation coefficients are 0.50, 0.43, and 0.46 for the first three slowest modes, respectively. In the first mode, the major disagreement between these two methods is a loop region (residue T109-W129). Figure 3.4 illustrates the ribbon diagram colored blue-yellow-red in the order of increasing mobility along the first mode. The loop region is highly mobile in the calculation of GNM; however, from the computation of PFP model, the loop is less flexible in the first mode.

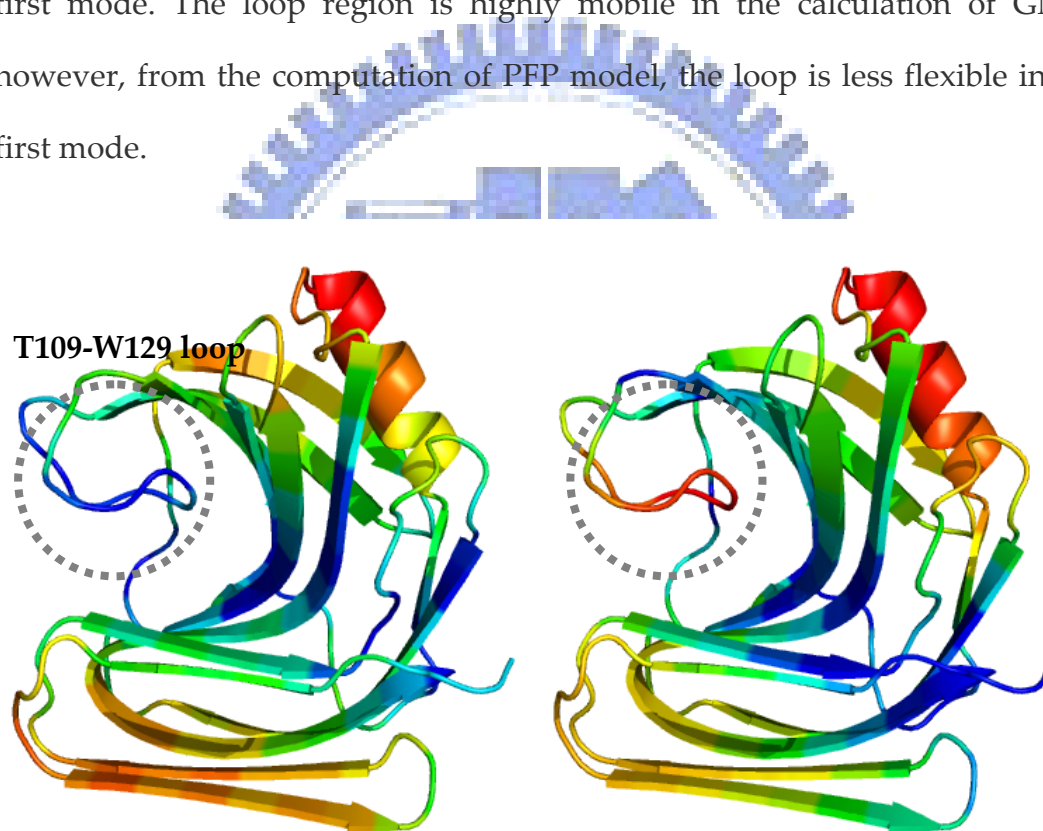


Figure 3.4 Ribbon diagram colored blue-yellow-red in the order of increasing mobility along the first mode of the PFP model (left) and GNM (right). The T109-W129 loop is surrounded by dotted circle.

Figure 3.6 shows the ribbon diagrams of G/11 xylanase, the colored regions are according to the positive (blue) and negative (red) displacements.

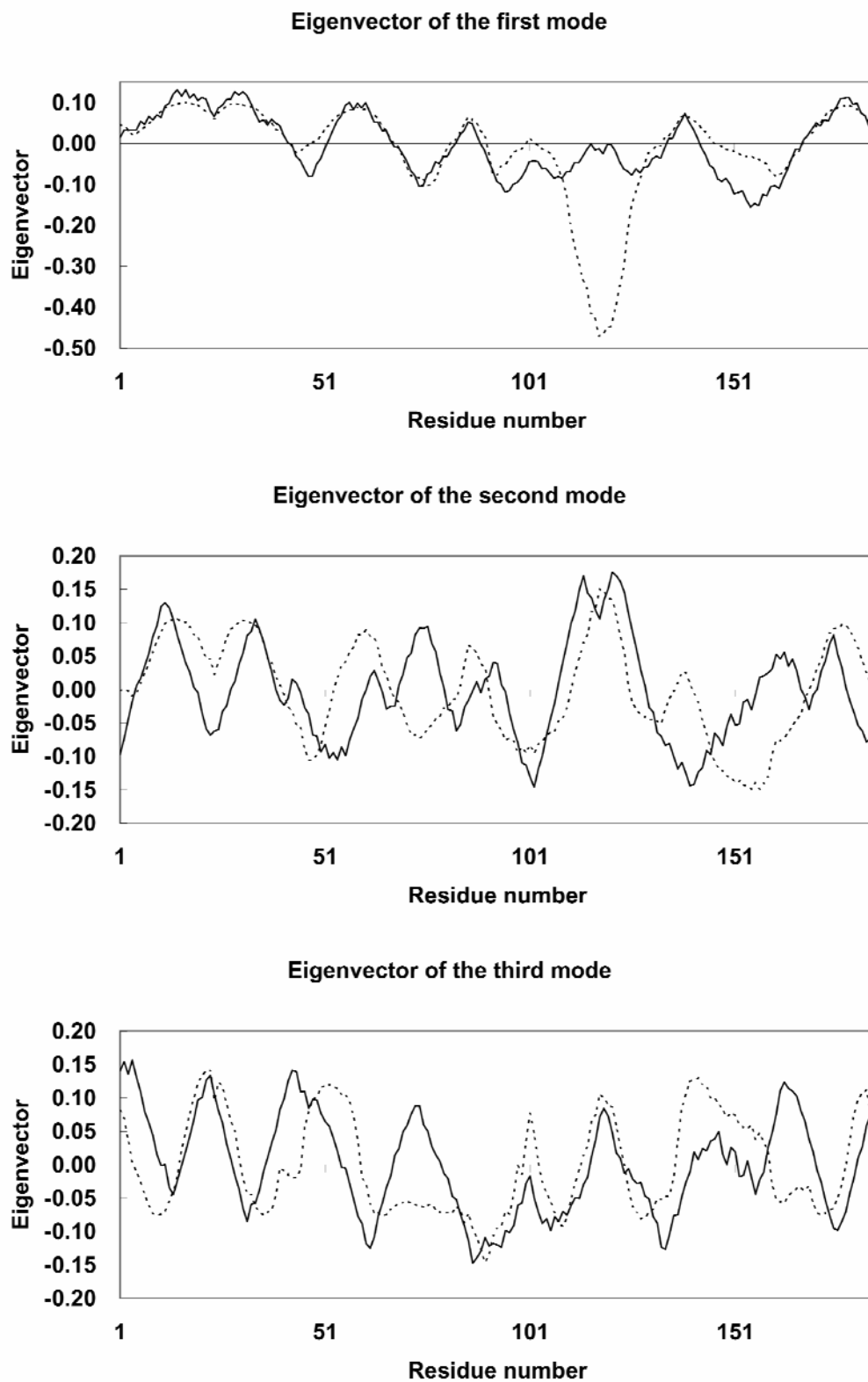


Figure 3.5 Distribution of displacements along the first mode, second mode, and third mode computed for G/11 xylanase (1BVV). (solid line: PFP model, dotted line: GNM)

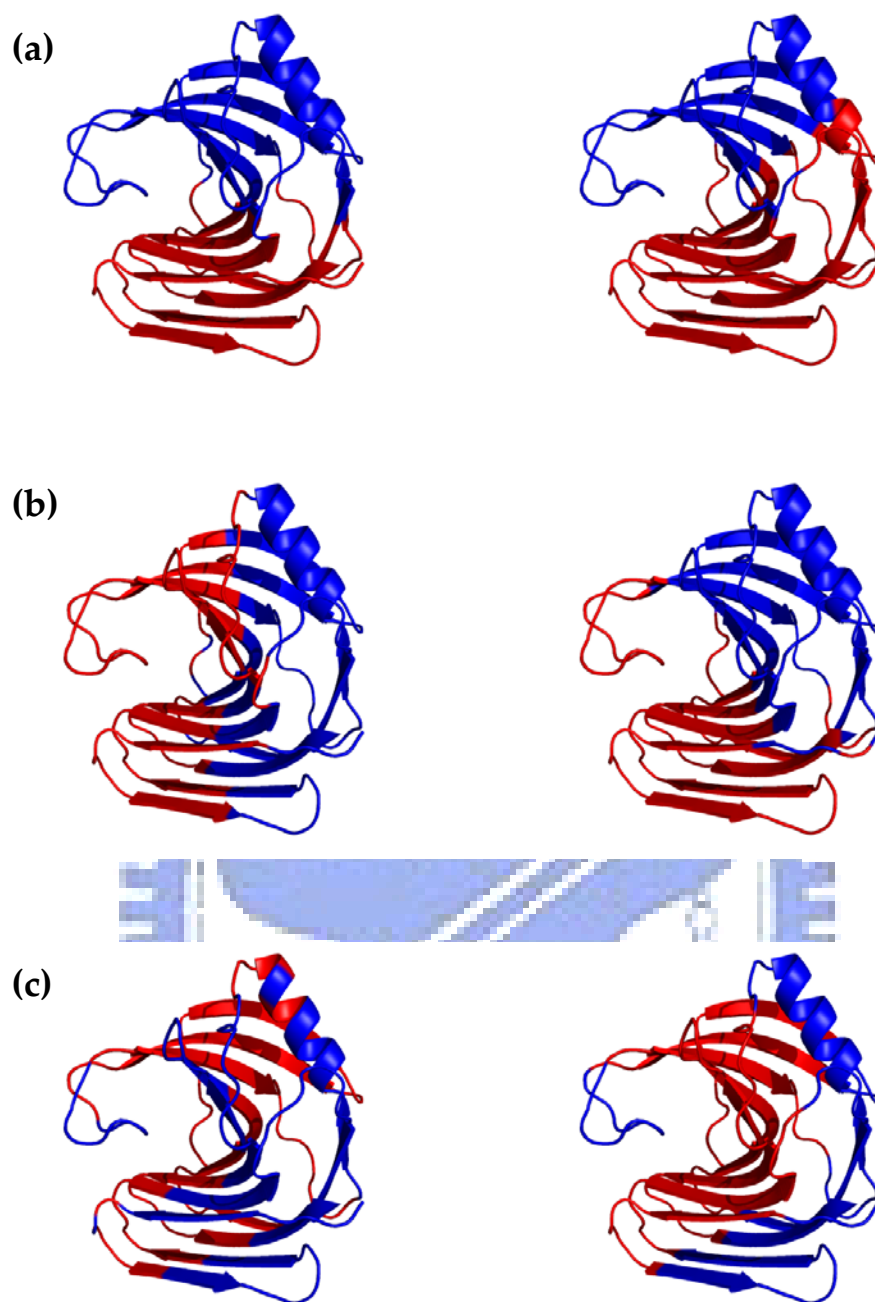


Figure 3.6 The regions subject to opposite direction displacements computed by the PFP model (left) and the GNM (right) of (a) the first mode, (b) the second mode, and (c) the third mode for G/11 xylanase (1BVV). Regions colored in blue and red correspond to positive and negative displacements respectively.

Catalytic residues

Recent studies⁷³ show that the catalytic residues are usually immobilized in order to maintain the delicate arrangement of function groups. The catalytic residues are found to be located at the center^{74,75} or have relatively lower B-factors comparing to other residues in the protein⁷⁶. Based on GNM, Yang et al.⁷³ analyzed the normal mode motions of several proteins and found that the catalytic residues are located near the hinge region, i.e. the crossover region between the parts of oppositely correlated movement, of the most large scale motion. Here, we show a case of cysteine protease, actinidin (1AEC). Actinidin is a 218-residue protein; the catalytic residues are C25, H162, and N182. Figure 3.7 shows the normal mode displacements of actinidin along the first mode. The catalytic residues shown in the stick model are in the crossover region between the two substructures undergoing oppositely correlated motions.

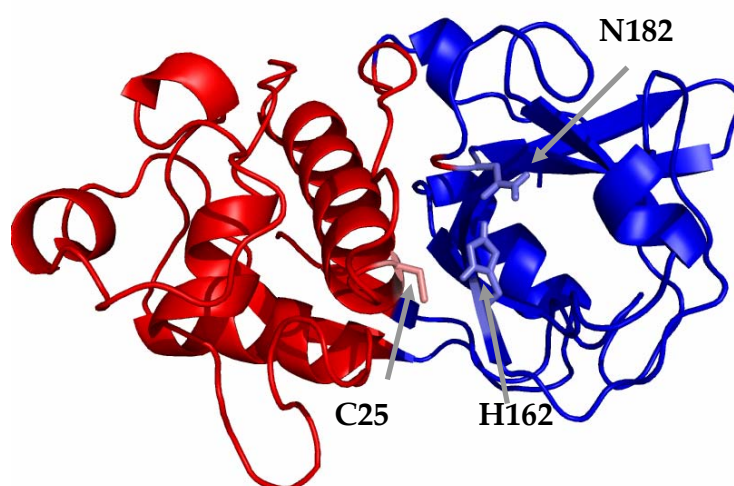


Figure 3.7 The displacements of actinidin (1AEC) subject to opposite directions along the first mode. The catalytic residues are shown in the stick model.

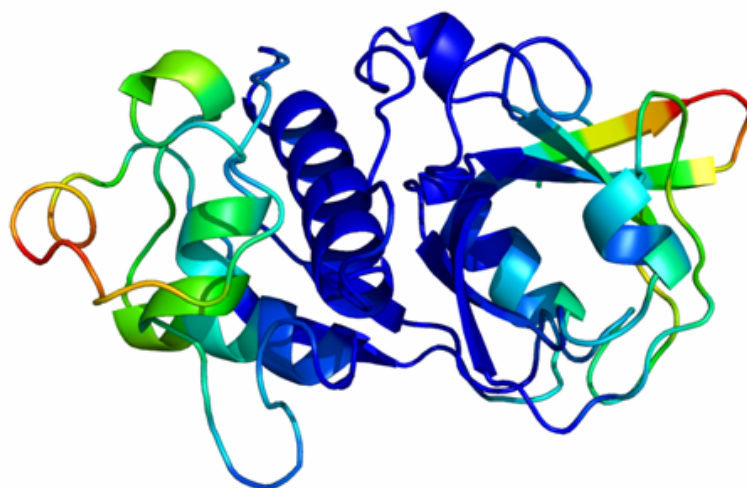


Figure 3.8 Ribbon diagram of actinidin colored blue-yellow-red in the order of increasing mobility along the first mode.

Figure 3.8 shows the ribbon diagram colored blue-yellow-red in the order of increasing mobility along the first mode. It clearly shows that the hinge region is the least mobile part. The result is consistent with previous studies that the catalytic residues are located on the most stable region in the protein structure.

Discussion

It had been shown that the PFP model can accurately compute the theoretical thermal fluctuations of individual residues^{46,72}. In this work, we further extend the model to compute the correlation of motion between residue pairs. In addition, the modes of motion which subject to different frequencies can be obtained by simply diagonalizing the correlation map. We compared the correlation maps and mode of motion by the PFP model and GNM, and found that the two models generally have similar results, especially in the largest scale, slowest mode. The PFP model simply calculates the distance between $C\alpha$ atoms and the center of mass of the protein; on the other hand, the GNM assumes the protein is a network in which the $C\alpha$ atoms are balls connected to each other by springs. Here, the results suggest that the dynamics properties, i.e. thermal fluctuation, correlation of motion, and normal mode motions, can be directly obtained from the protein structure without assuming any mechanical models.

Recent studies not only focus on the relationship between protein structure and dynamics but also on the link between protein dynamics and functions. Chennubhotla⁷⁷ uses elastic network model to study the allosteric mechanism and signal transduction pathways of chaperonin GroEL. Yang⁷³ proposed that protein catalytic sites are located on the “hinge” region based on the normal mode analysis. Liu⁷⁸ uses the correlation matrices to analyze the correlated mutations in HIV-1 protease. The widely use of NMA in recent

articles in the study of protein functions shows the growing importance of normal mode analysis. In addition to the prediction of thermal fluctuations, the PFP model has the ability to characterize the normal modes motions which are relevant to protein functions and catalytic mechanisms.

One of the differences between the PFP-based and GNM-based NMA is that the PFP-based method directly compute the correlation map from protein structure and diagonalize the correlation map to get the normal mode motions. On the other hand, the GNM needs to first build the Hessian matrix and the Hessian matrix is inverted to get the correlation map. To compute the normal mode motions, the GNM diagonalize the Hessian matrix instead of the correlation map. Figure 3.9 illustrates the differences between the PFP-based and GNM-based NMA.

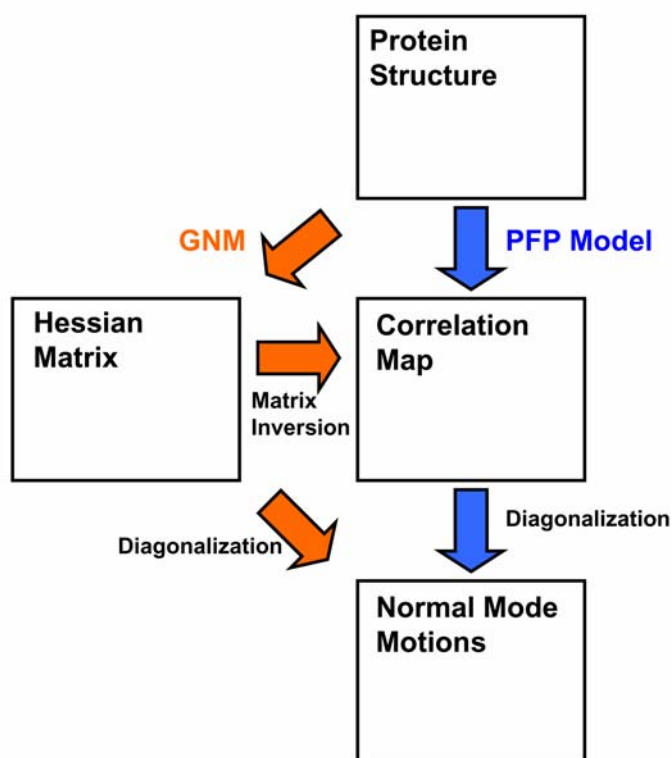
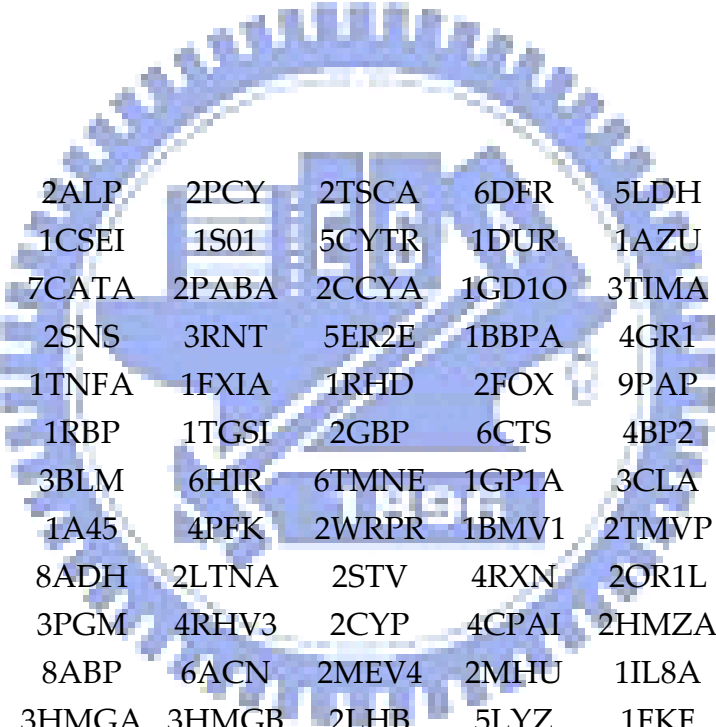


Figure 3.9 The flowchart of computing normal mode motions by PFP model and GNM

The direct computation of normal mode motions from the correlation map suggests that it is possible to skip the Hessian matrix and obtain the normal mode motion from correlation maps computed by any method.

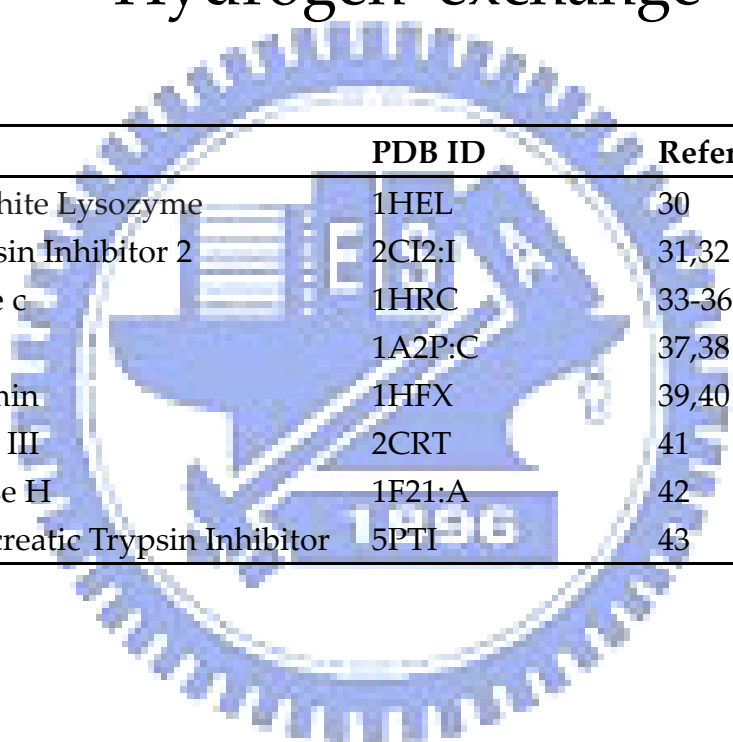


RS126 Dataset



1FC2C	2ALP	2PCY	2TSCA	6DFR	5LDH	2I1B
2TGPI	1CSEI	1S01	5CYTR	1DUR	1AZU	2PHH
1ACX	7CATA	2PABA	2CCYA	1GD1O	3TIMA	1PAZ
2UTGA	2SNS	3RNT	5ER2E	1BBPA	4GR1	4TS1A
1GDJ	1TNFA	1FXIA	1RHD	2FOX	9PAP	4RHV4
2LTNB	1RBP	1TGSI	2GBP	6CTS	4BP2	1CBH
6CPA	3BLM	6HIR	6TMNE	1GP1A	3CLA	
1BKSA	1A45	4PFK	2WRPR	1BMV1	2TMVP	
2AK3A	8ADH	2LTNA	2STV	4RXN	2OR1L	
2GN5	3PGM	4RHV3	2CYP	4CPAI	2HMZA	
4SGBI	8ABP	6ACN	2MEV4	2MHU	1IL8A	
1CRN	3HMGA	3HMGB	2LHB	5LYZ	1FKF	
7ICD	2RSPA	9APIB	1UBQ	3CLN	1BKSB	
1BMV2	1PYP	1BDS	2CAB	9APIA	4CMS	
2SODB	1CDTA	1PPT	1FDLH	1SH1	3CD4	
3AIT	256BA	1ECA	6CPP	1IQZ	1L58	
1HIP	1OVOA	3ICB	1LAP	4XIAA	1LMB3	
1ETU	9WGAA	7RSA	1R092	1FND	1G6NA	
1MRT	1CC5	5HVPA	1MCPL	9INSB	2GLSA	
1CYO	3EBX	4RHV1	2AAT	4CPV	4SDHA	

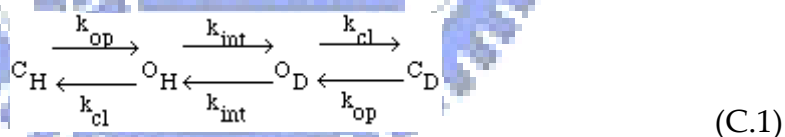
Hydrogen exchange dataset



Protein	PDB ID	References
Hen Egg-White Lysozyme	1HEL	30
Chymotrypsin Inhibitor 2	2CI2:I	31,32
Cytochrome c	1HRC	33-36
Barnase	1A2P:C	37,38
α -Lactalbumin	1HFX	39,40
Cardiotoxin III	2CRT	41
Ribonuclease H	1F21:A	42
Bovine Pancreatic Trypsin Inhibitor	5PTI	43

Hydrogen exchange experiment

During protein folding process, there are some structured regions which is similar to folded conformation⁷⁻¹⁰. Hydrogen isotope exchange rate¹¹⁻¹³ is usually used to identify those structured regions. The basic assumption of the hydrogen exchange experiment is that the exchange rate reflects the exposure of the particular amide to the solvent. Thus an amide which is buried inside the hydrophobic core of a protein will exchange slowly, while an amide on the surface will exchange rapidly. The two-step model of protein hydrogen exchange process commonly accepted is:



In this model C_H and C_D represent the protonated and deuterated closed (unable to exchange) forms of the residue. O_H is the protonated open, exchangeable form. k_{op} is the rate constant for the opening step and k_{cl} is the rate constant for the closing step. k_{int} is the rate constant for the exchange reaction or the exposed hydrogen.

The observed rate constant for the exchange process, k_{ex} , is given by

$$k_{ex} = \frac{k_{op} \times k_{int}}{k_{cl} + k_{int}}
 \tag{C.2}$$

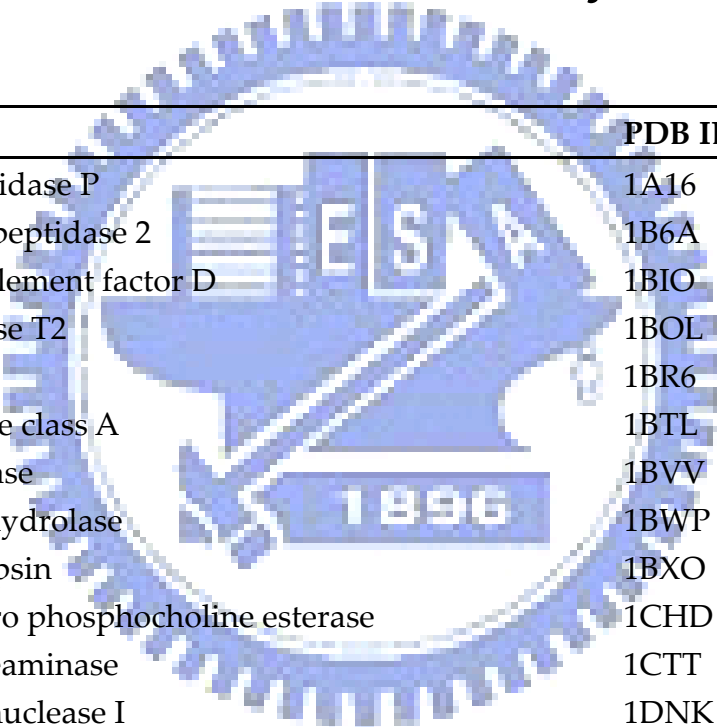
There are two limiting cases for this model. If $k_{cl} > k_{int}$ then

$$k_{ex} = \frac{k_{op}}{k_{cl}} \times k_{int} \quad (C.3)$$

and the process is said to follow EX2 kinetics. In the other limit, termed EX1, $k_{cl} < k_{int}$ and the rate limiting step in the process is the opening step since $k_{ex}=k_{op}$. Protons which exchange following EX2 kinetics are usually solvent exposed in the native state of the protein or require only smaller structural fluctuations to expose them. In contrast, protons exchanging with EX1 kinetics are generally deeply buried or have strong hydrogen bonds with surrounding residues^{79,80} and require a more global fluctuation in the structure of the protein to be exposed to the solvent.

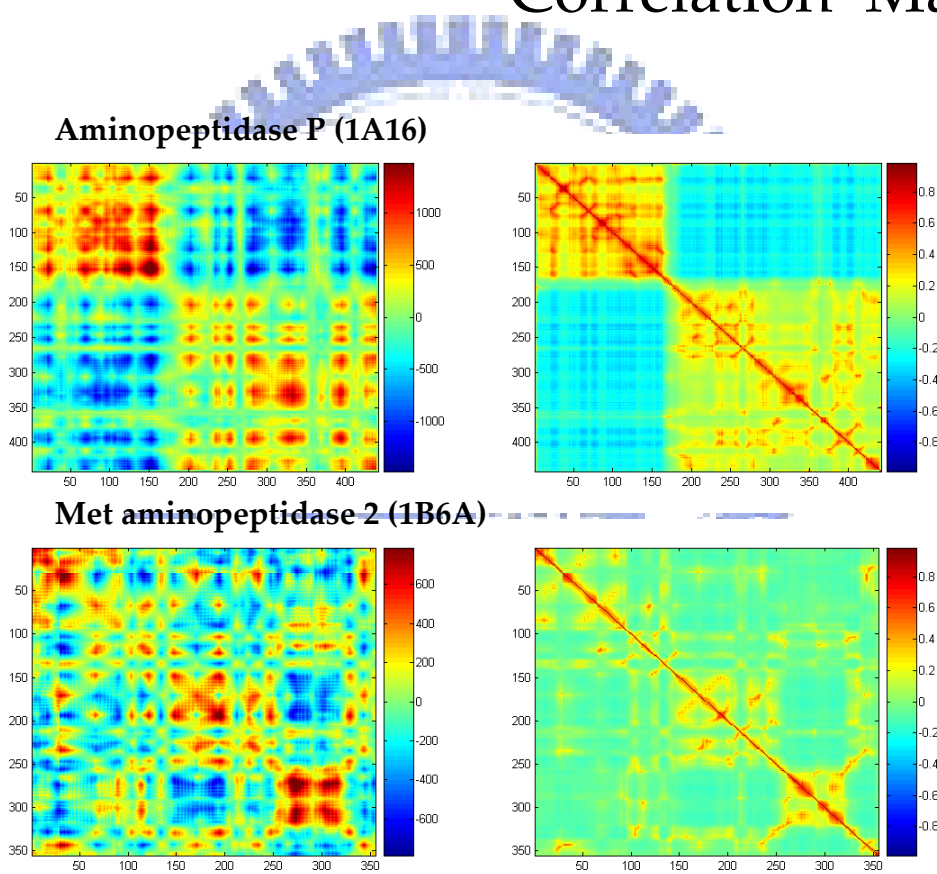


Normal mode analysis dataset



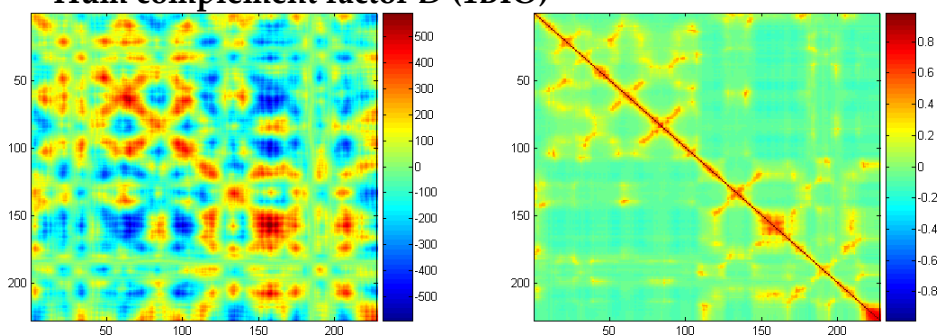
Protein	PDB ID
Aminopeptidase P	1A16
Met aminopeptidase 2	1B6A
Hum complement factor D	1BIO
Ribonuclease T2	1BOL
Ricin	1BR6
β -Lactamase class A	1BTL
G/11 xylanase	1BVV
2-acetyl-1-hydrolase	1BWP
Penicillopepsin	1BXO
Alkylglycero phosphocholine esterase	1CHD
Cytidine deaminase	1CTT
Deoxyribonuclease I	1DNK
Uridine nucleosidase	1EUG
T7 lysozyme	1LBA
Prot-glu methylesterase	1RPT
Yersinia protein tyrosine phosphatase	1YTW
Metalloproteinase M4	8TLN
Thiol-endopeptidase	9PAP

Correlation Maps

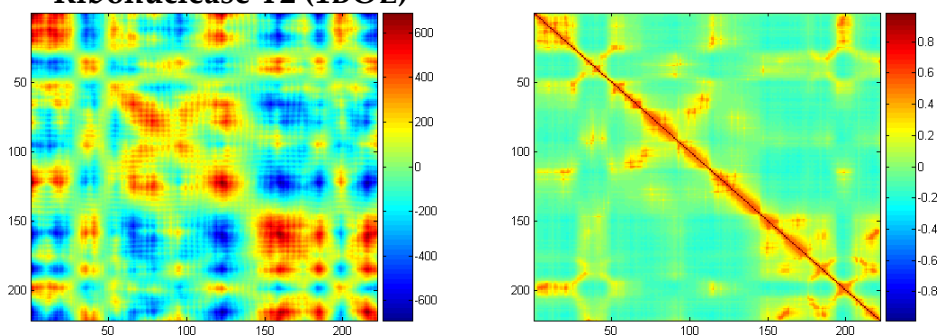


Appendix E Correlation maps computed with the PFP model (left) and GNM (right) for the dataset of 18 proteins

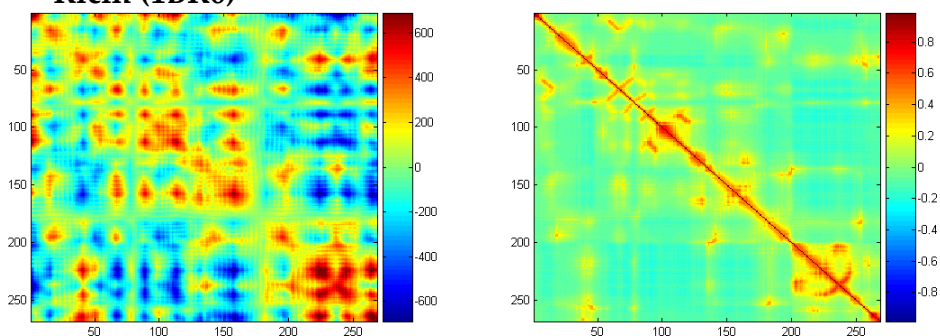
Hum complement factor D (1BIO)



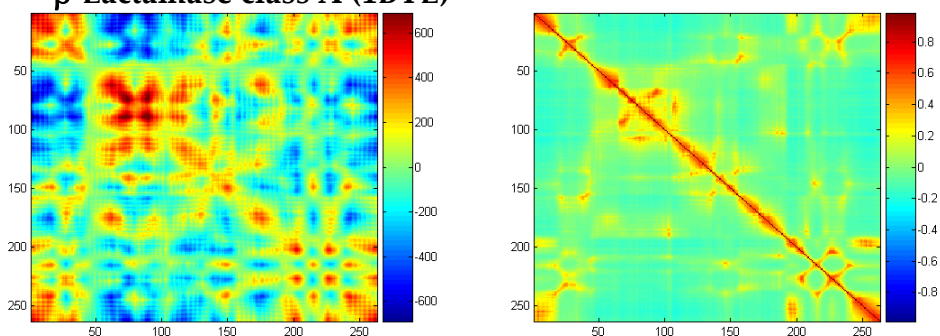
Ribonuclease T2 (1BOL)



Ricin (1BR6)

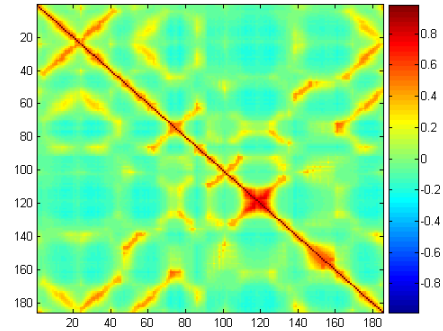
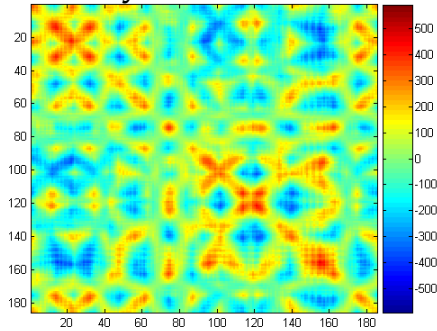


β -Lactamase class A (1BTL)

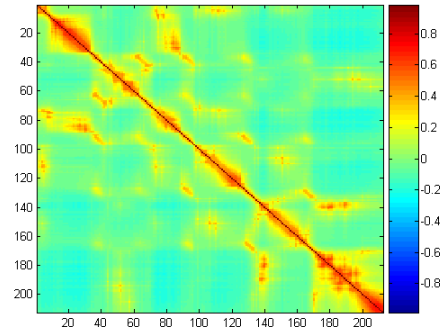
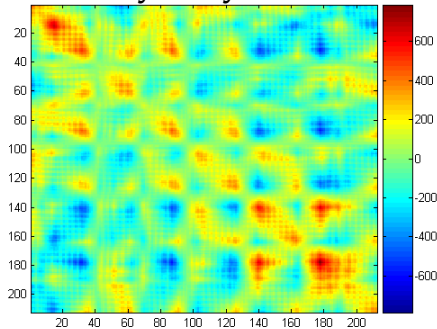


Appendix E (continued) Correlation maps computed with the PFP model (left) and GNM (right) for the dataset of 18 proteins

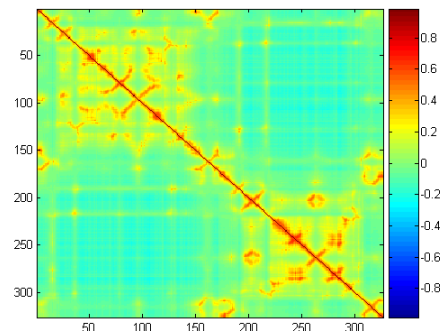
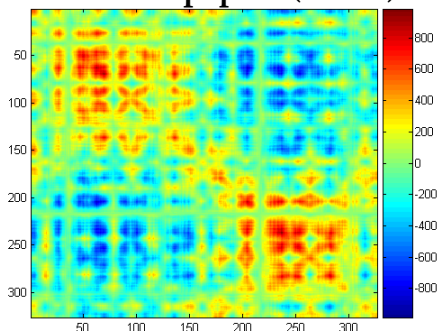
G/11 xylanase (1BVV)



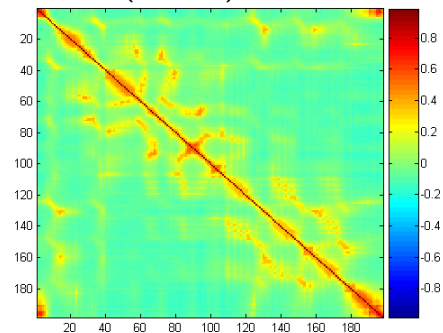
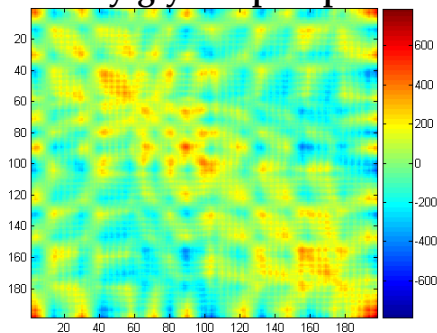
2-acetyl-1-hydrolase (1BWP)



Penicillopepsin (1BXO)

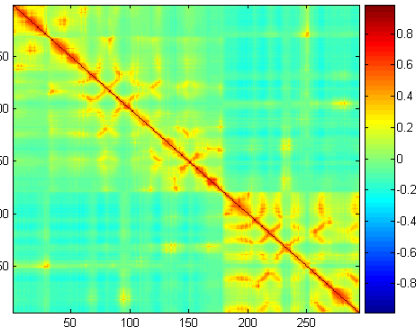
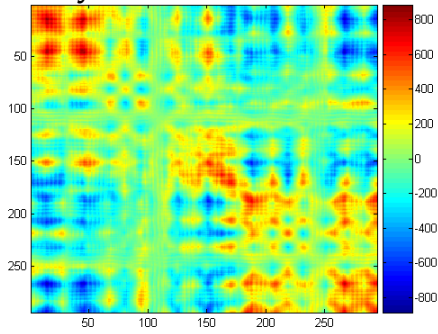


Alkylglycero phosphocholine esterase (1CHD)

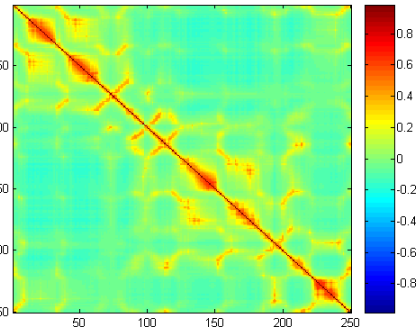
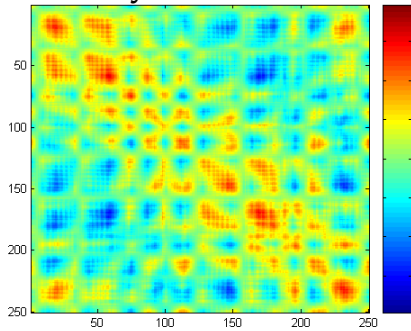


Appendix E (continued) Correlation maps computed with the PFP model (left) and GNM (right) for the dataset of 18 proteins

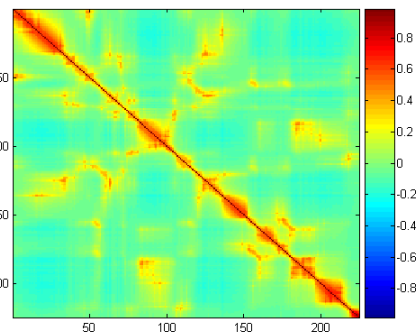
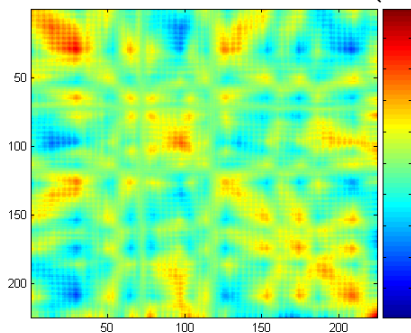
Cytidine deaminase (1CTT)



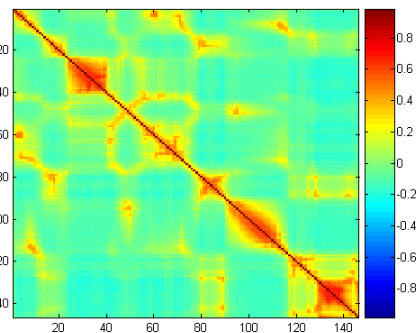
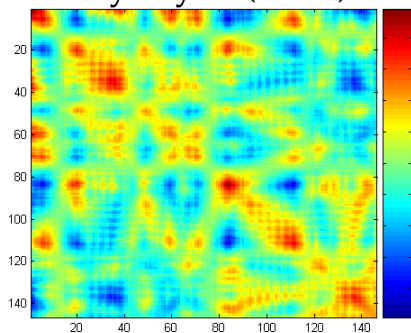
Deoxyribonuclease I (1DNK)



Uridine nucleosidase (1EUG)

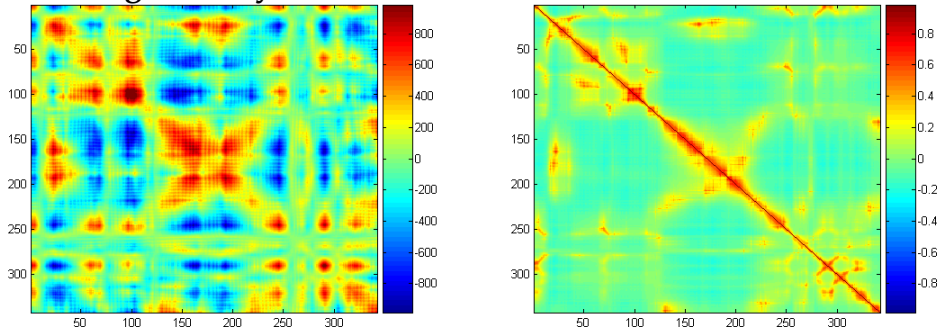


T7 lysozyme (1LBA)

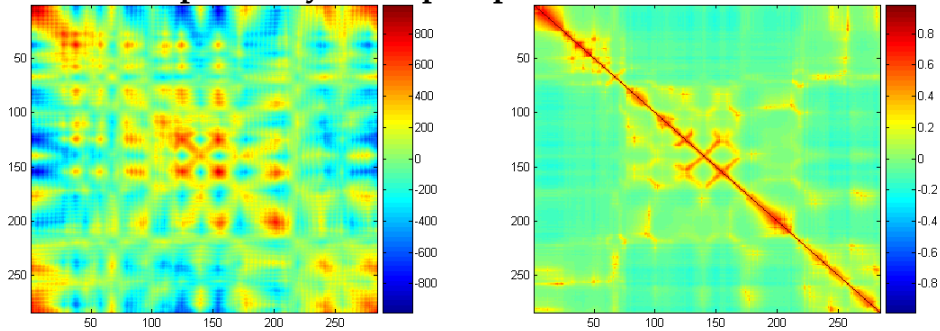


Appendix E (continued) Correlation maps computed with the PFP model (left) and GNM (right) for the dataset of 18 proteins

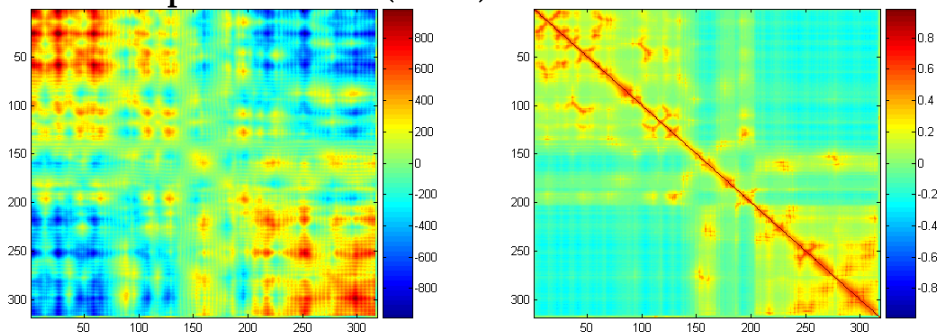
Prot-glu methylesterase (1RPT)



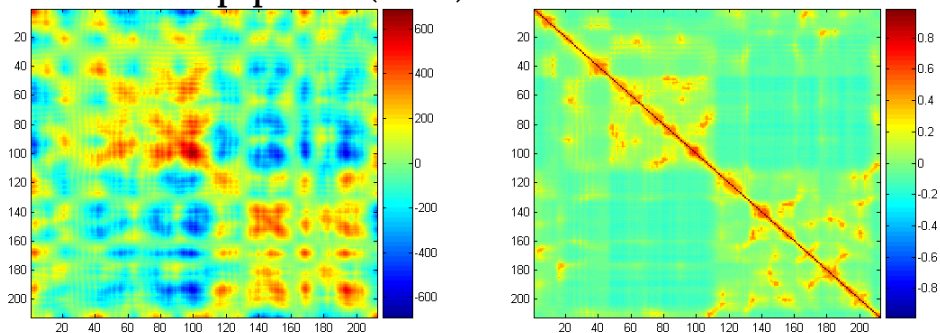
Yersinia protein tyrosine phosphatase (1YTW)



Metalloproteinase M4 (8TLN)



Thiol-endopeptidase (9PAP)



Appendix E (continued) Correlation maps computed with the PFP model (left) and GNM (right) for the dataset of 18 proteins

References

1. Kabsch, W. & Sander, C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA* **81**, 1075-8 (1984).
2. Minor, D.L., Jr. & Kim, P.S. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730-4 (1996).
3. Mezei, M. Chameleon sequences in the PDB. *Protein Eng* **11**, 411-4 (1998).
4. Sudarsanam, S. Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. *Proteins* **30**, 228-31 (1998).
5. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).
6. Cohen, B.I., Presnell, S.R. & Cohen, F.E. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci* **2**, 2134-45 (1993).
7. Leopold, P.E., Montal, M. & Onuchic, J.N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci USA* **89**, 8721-5 (1992).
8. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-95 (1995).
9. Shoemaker, B.A., Wang, J. & Wolynes, P.G. Structural correlations in protein folding funnels. *Proc Natl Acad Sci U S A* **94**, 777-82 (1997).
10. Onuchic, J.N. & Wolynes, P.G. Theory of protein folding. *Curr Opin Struct Biol* **14**, 70-5 (2004).
11. Betz, S.F. et al. Unusual effects of an engineered disulfide on global and local protein stability. *Biochemistry* **35**, 7422-8 (1996).
12. Hilser, V.J. & Freire, E. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J Mol Biol* **262**, 756-72 (1996).
13. Clarke, J. & Itzhaki, L.S. Hydrogen exchange and protein folding. *Curr Opin Struct Biol* **8**, 112-8 (1998).
14. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
15. Vapnik, V.N. *The Nature of statistical learning theory*, (New York Springer, 1995).

16. Hua, S. & Sun, Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol* **308**, 397-407 (2001).
17. Kim, H. & Park, H. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng* **16**, 553-60 (2003).
18. Ward, J.J., McGuffin, L.J., Buxton, B.F. & Jones, D.T. Secondary structure prediction with support vector machines. *Bioinformatics* **19**, 1650-5 (2003).
19. Kim, H. & Park, H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* **54**, 557-62 (2004).
20. Ding, C.H. & Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17**, 349-58 (2001).
21. Yu, C.S. et al. Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets. *Proteins* **50**, 531-6 (2003).
22. Yu, C.S., Lin, C.J. & Hwang, J.K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* **13**, 1402-6 (2004).
23. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**, 721-8 (2001).
24. Brown, M.P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* **97**, 262-7 (2000).
25. Dobson, P.D. & Doig, A.J. Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol* **330**, 771-83 (2003).
26. Chen, Y.C., Lin, Y.S., Lin, C.J. & Hwang, J.K. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins* **55**, 1036-42 (2004).
27. Chang, C.C. & Lin, C.J. LIBSVM: a library for support vector machines. Software available from <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (2001).
28. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637 (1983).
29. Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**, 584-99 (1993).

30. Radford, S.E., Buck, M., Topping, K.D., Dobson, C.M. & Evans, P.A. Hydrogen exchange in native and denatured states of hen egg-white lysozyme. *Proteins* **14**, 237-48 (1992).
31. Itzhaki, L.S., Neira, J.L. & Fersht, A.R. Hydrogen exchange in chymotrypsin inhibitor 2 probed by denaturants and temperature. *J Mol Biol* **270**, 89-98 (1997).
32. Neira, J.L., Itzhaki, L.S., Otzen, D.E., Davis, B. & Fersht, A.R. Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis. *J Mol Biol* **270**, 99-110 (1997).
33. Jeng, M.F., Englander, S.W., Elove, G.A., Wand, A.J. & Roder, H. Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry* **29**, 10433-7 (1990).
34. Bai, Y., Sosnick, T.R., Mayne, L. & Englander, S.W. Protein folding intermediates: native-state hydrogen exchange. *Science* **269**, 192-7 (1995).
35. Bai, Y. & Englander, S.W. Future directions in folding: the multi-state nature of protein structure. *Proteins* **24**, 145-51 (1996).
36. Hoang, L., Bedard, S., Krishna, M.M., Lin, Y. & Englander, S.W. Cytochrome c folding pathway: kinetic native-state hydrogen exchange. *Proc Natl Acad Sci USA* **99**, 12173-8 (2002).
37. Matouschek, A., Serrano, L., Meiering, E.M., Bycroft, M. & Fersht, A.R. The folding of an enzyme. V. H/2H exchange-nuclear magnetic resonance studies on the folding pathway of barnase: complementarity to and agreement with protein engineering studies. *J Mol Biol* **224**, 837-45 (1992).
38. Perrett, S., Clarke, J., Hounslow, A.M. & Fersht, A.R. Relationship between equilibrium amide proton exchange behavior and the folding pathway of barnase. *Biochemistry* **34**, 9288-98 (1995).
39. Chyan, C.L., Wormald, C., Dobson, C.M., Evans, P.A. & Baum, J. Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: a hydrogen exchange study. *Biochemistry* **32**, 5681-91 (1993).
40. Schulman, B.A., Redfield, C., Peng, Z.Y., Dobson, C.M. & Kim, P.S. Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human alpha-lactalbumin. *J Mol Biol* **253**, 651-7 (1995).
41. Sivaraman, T., Kumar, T.K., Chang, D.K., Lin, W.Y. & Yu, C. Events in the kinetic folding pathway of a small, all beta-sheet protein. *J Biol*

- Chem* **273**, 10181-9 (1998).
42. Chamberlain, A.K., Handel, T.M. & Marqusee, S. Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat Struct Biol* **3**, 782-7 (1996).
 43. Woodward, C.K. & Hilton, B.D. Hydrogen isotope exchange kinetics of single protons in bovine pancreatic trypsin inhibitor. *Biophys J* **32**, 561-75 (1980).
 44. Zhou, X., Alber, F., Folkers, G., Gonnet, G.H. & Chelvanayagam, G. An analysis of the helix-to-strand transition between peptides with identical sequence. *Proteins* **41**, 248-56 (2000).
 45. Bahar, I., Atilgan, A.R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des* **2**, 173-81 (1997).
 46. Shih, C.H. et al. A simple way to compute protein dynamics without a mechanical model. *Proteins* **68**, 34-8 (2007).
 47. Lin, C.P. et al. Deriving protein dynamical properties from weighted protein contact number. *Proteins* **72**, 929-35 (2008).
 48. Zhang, F. & Bruschweiler, R. Contact model for the prediction of NMR N-H order parameters in globular proteins. *J Am Chem Soc* **124**, 12654-5 (2002).
 49. Tirion, M.M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* **77**, 1905-1908 (1996).
 50. Ming, D. & Bruschweiler, R. Reorientational contact-weighted elastic network model for the prediction of protein dynamics: comparison with NMR relaxation. *Biophys J* **90**, 3382-8 (2006).
 51. Pfeiffer, S., Fushman, D. & Cowburn, D. Simulated and NMR-derived backbone dynamics of a protein with significant flexibility: a comparison of spectral densities for the betaARK1 PH domain. *J Am Chem Soc* **123**, 3021-36 (2001).
 52. Akke, M., Skelton, N.J., Kordel, J., Palmer, A.G., 3rd & Chazin, W.J. Effects of ion binding on the backbone dynamics of calbindin D9k determined by ¹⁵N NMR relaxation. *Biochemistry* **32**, 9832-44 (1993).
 53. Feng, W., Tejero, R., Zimmerman, D.E., Inouye, M. & Montelione, G.T. Solution NMR structure and backbone dynamics of the major cold-shock protein (CspA) from *Escherichia coli*: evidence for conformational dynamics in the single-stranded RNA-binding site. *Biochemistry* **37**, 10881-96 (1998).
 54. Li, Q., Khosla, C., Puglisi, J.D. & Liu, C.W. Solution structure and

- backbone dynamics of the holo form of the frenolicin acyl carrier protein. *Biochemistry* **42**, 4648-57 (2003).
55. Buck, M. et al. Structural determinants of protein dynamics: analysis of ¹⁵N NMR relaxation measurements for main-chain and side-chain nuclei of hen egg white lysozyme. *Biochemistry* **34**, 4041-55 (1995).
 56. Kristensen, S.M., Siegal, G., Sankar, A. & Driscoll, P.C. Backbone dynamics of the C-terminal SH2 domain of the p85[alpha] subunit of phosphoinositide 3-kinase: effect of phosphotyrosine-peptide binding and characterization of slow conformational exchange processes. *Journal of Molecular Biology* **299**, 771-788 (2000).
 57. Tjandra, N., Feller, S.E., Pastor, R.W. & Bax, A. Rotational diffusion anisotropy of human ubiquitin from ¹⁵N NMR relaxation. *J. Am. Chem. Soc.* **117**, 12562-12566 (1995).
 58. Yun, S., Jang, D.S., Kim, D.H., Choi, K.Y. & Lee, H.C. ¹⁵N NMR relaxation studies of backbone dynamics in free and steroid-bound Delta 5-3-ketosteroid isomerase from *Pseudomonas testosteroni*. *Biochemistry* **40**, 3967-73 (2001).
 59. Stivers, J.T., Abeygunawardana, C. & Mildvan, A.S. ¹⁵N NMR relaxation studies of free and inhibitor-bound 4-oxalocrotonate tautomerase: backbone dynamics and entropy changes of an enzyme upon inhibitor binding. *Biochemistry* **35**, 16036-47 (1996).
 60. Redfield, C., Boyd, J., Smith, L.J., Smith, R.A. & Dobson, C.M. Loop mobility in a four-helix-bundle protein: ¹⁵N NMR relaxation measurements on human interleukin-4. *Biochemistry* **31**, 10431-7 (1992).
 61. Svensson, L.A., Thulin, E. & Forsen, S. Proline cis-trans isomers in calbindin D9k observed by X-ray crystallography. *J Mol Biol* **223**, 601-6 (1992).
 62. Harata, K. & Muraki, M. X-ray structure of turkey-egg lysozyme complex with tri-N-acetylchitotriose. Lack of binding ability at subsite A. *Acta Crystallogr D Biol Crystallogr* **53**, 650-7 (1997).
 63. Halle, B. Flexibility and packing in proteins. *Proc Natl Acad Sci USA* **99**, 1274-9 (2002).
 64. Horiuchi, T. & Go, N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins* **10**, 106-16 (1991).
 65. Garcia, A.E. & Harman, J.G. Simulations of CRP:(cAMP)₂ in noncrystalline environments show a subunit transition from the open to the closed conformation. *Protein Sci* **5**, 62-71 (1996).

66. Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins* **33**, 417-29 (1998).
67. Kitao, A. & Go, N. Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol* **9**, 164-9 (1999).
68. Go, N., Noguti, T. & Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A* **80**, 3696-700 (1983).
69. Ichiye, T. & Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **11**, 205-17 (1991).
70. Jernigan, R.L. & Bahar, I. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* **6**, 195-209 (1996).
71. Bahar, I. & Jernigan, R.L. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol* **266**, 195-214 (1997).
72. Lu, C.H. et al. On the relationship between the protein structure and protein dynamics. *Proteins* **72**, 625-34 (2008).
73. Yang, L.W. & Bahar, I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13**, 893-904 (2005).
74. Ben-Shimon, A. & Eisenstein, M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol* **351**, 309-26 (2005).
75. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci* **15**, 2120-8 (2006).
76. Yuan, Z., Zhao, J. & Wang, Z.X. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* **16**, 109-14 (2003).
77. Chennubhotla, C., Yang, Z. & Bahar, I. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol Biosyst* **4**, 287-92 (2008).
78. Liu, Y., Eyal, E. & Bahar, I. Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* **24**, 1243-50 (2008).
79. Woodward, C. & Li, R. The slow-exchange core and protein folding. *Trends Biochem Sci* **23**, 379-81 (1998).
80. Li, R. & Woodward, C. The hydrogen exchange core and protein

folding. *Protein Sci* 8, 1571-90 (1999).

