RICE UNIVERSITY

Hierarchical Normal Mode Refinement of Anisotropic Thermal Parameters

by

Zhenwei Luo

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE

Jianpeng Ma, Chair
Professor of Bioengineering, Rice University and Lodwick T. Bolin Professor of Biochemistry, Baylor College of Medicine

Ching-Hwa Kiang
Associate Professor of Physics and Astronomy and Bioengineering, Rice University

Robert M. Raphael
Associate Professor of Bioengineering, Rice University

HOUSTON, TEXAS
December 2016

ABSTRACT

Hierarchical normal mode refinement of anisotropic thermal parameters for supramolecular complexes

by

Zhenwei Luo

In this thesis, we report a novel normal-mode based protocol for modeling anisotropic thermal motions of supramolecular complexes in x-ray crystallographic refinement, named HNMRef. The method models not only the global movements of the whole complex but also the deformational patterns of substructures. Compared with another widely adopted anisotropic thermal parameters refinement method—multi-group TLS, HNMRef delivers much more accurate thermal parameters for the complex and greatly simplifies the choice of substructure partition schemes. The effectiveness of the procedure is demonstrated on the refinements of a set of complexes with moderate resolutions. This protocol was shown to be able to significantly reduce the values of R_{free} and improve the electron density maps. Moreover, the distribution of anisotropic thermal ellipsoids was much more consistent throughout the whole structure and agreed with the functional structure movements. We expect this protocol to be very effective in the anisotropic refinements of very large and flexible complexes with low or moderate-resolution x-ray diffraction data.

Acknowledgments

I would like to thank to my advisor, Dr. Jianpeng Ma, for his patience, enthusiasm and continuous support of my study and research. As one of the best and smartest scientists, Dr. Ma always points out the keys to the problem which guide the research direction to a correct way. Without his enlightenment, the new ideas in this thesis would never come into reality.

My sincere thanks also go to the rest of my thesis committee: Dr. Kiang and Dr. Raphael for their encouragement and insightful advices.

I would also like to thank all of my current and former colleagues in Dr.Ma's lab, Tianwu Zang, Linling Yu, and Mingyang Lu, and my friends Chun He and Ye Zhang for all the stimulating discussions and enjoyable time together. Part of the work in my thesis was supported by Davinci at Rice University.

Last but most important, I would like to thank my parents. Without their love and support, this thesis would have never been possible.

Contents

Acknowledgments ..... **iii**

Contents ..... **iv**

List of Figures ..... **v**

List of Tables ..... **vi**

List of Equations ..... **vii**

Nomenclature ..... **viii**

Introduction **1**

Normal Mode Analysis **1**

2.1. Hessian Matrix 1

2.2. fSUB 3

Crystallographic Refinement of ADP **1**

3.1. Crystallographic Refinement of Atomic Displacements 2

3.2. TLS Method 4

3.3. Normal Mode Based Crystallographic Refinement 4

3.4. HNMRef 5

3.5. Test Criteria 6

Results **1**

4.1. Refinement Protocols 1

4.2. $\boldsymbol{R}\mathbf{free}$ and Ramachandran Favored Region 3

4.3. Electron Density Map 4

4.4. $\boldsymbol{B}\mathbf{KL}$ 5

Conclusions **1**

Notes **2**

References **3**

Appendix A **6**

List of Figures

Figure 1 Improvements of *R*free Values for different proteins 7

Figure 2 *B*KL profile for each refined structure..... 7

Figure 3 Electron density map and structure comparison for 2CG9 8

Figure 4 Electron density map and structure comparison for 3I59..... 9

Figure 5. The open structure of an E.coli mechanosensitive channel at 3.45 Å showing 50% probability ellipsoids refined with (A) HNMRef and (B) multi-group TLS methods.....10

List of Tables

Table 1 *R*free(%) and Ramachandran Favored regions for 20 protein structures with different refinement methods (The *R*free values and their improvements for HNMRef refined structures over TLS are shown in percent.) **4**

List of Equations

No table of figures entries found.

Nomenclature

ADP	Atomic Displacement Parameters
NMA	Normal Mode Analysis
TLS	Translation Libration and Screw
HNMRef	Hierarchical Normal Mode based Refiement
NCS	Non-crystallographic symmetry

Chapter 1

Introduction

The diffraction intensity data collected in x-ray crystallography experiment is not a static snapshot of a single structure, but a time and space average of a set of conformers (Painter and Merritt, 2006a). This information contained in experimental data is routinely modelled by atomic displacement parameters (ADP), which give the mean square deviation of each atom from its average position, or the variance of atomic position assuming it is distributed according to gaussian. The various conformers of a protein mainly originate from its motions, which range from high frequency atomic vibrations to low frequency collective movements. Among those forms of motions, the lowest frequency collective displacements contribute the most to a conformal change (Petrone and Pande, 2006). For large complexes comprised of highly flexible components, their low frequency movements are often hierarchical and anisotropic (Lu and Ma, 2005; Painter and Merritt, 2006a). The hierarchy of complex motions means that the conformers of

subunits vary according to their own low frequency modes and the whole structure also oscillates by its low frequency modes. Both of them contribute nontrivially to the large scale conformal changes. While the anisotropy mainly results from the orientation-specific nature of the low frequency modes of whole structure, such as hinge-bending motion between two subunits. A side effect of those large scale movements is that they deteriorate the resolutions that crystals can diffract to. To properly model the position uncertainties resulting from those motions in large complexes, first of all, anisotropic ADPs should be used. However, a full-scale individual anisotropic refinement requires three positional and six thermal parameters for each atom, thus significantly lowering the data to parameter ratio and increasing the risk of overfitting. Secondly, the ADPs of large complexes should represent the various hierarchies of their motions.

To address the overfitting problem in the anisotropic refinement of supramolecular complexes with moderate resolution, a number of methods have been proposed to reduce the number of parameters required for anisotropic refinement. These methods were inspired by the fact that the motions of a group of atoms can be well approximated by the linear combination of low frequency modes of this group (Brooks and Karplus, 1983; Go et al., 1983; Levitt et al., 1983). This formalism allows us to optimize the linear coefficients of low frequency modes when performing anisotropic refinement, instead of refining six ADPs for each atom. Two commonly used models in crystallography are TLS (translation/libration/screw) (Schomaker and Trueblood, 1968) and normal mode (Diamond, 1990; Kidera and Go, 1990; Kidera and Gō, 1992, 1992; Kidera et al.,

1992, 1994; Poon et al., 2007). Their difference is that TLS model only utilizes the six zero-frequency modes of the structure, while normal mode based model takes other low frequency modes into account. Another challenge for modeling the ADPs for large complexes is describing the various hierarchies of their movements. This was first systematically addressed by multi-group TLS method (Painter and Merritt, 2006a). The main concept of multi-group TLS method is to partition the complex into multiple groups, each described by a set of TLS parameters, and refine the TLS parameters for each group individually. Though this method has shown to improve the cross validation score (R_{free}) (Brünger, 1997) of refined structures, it also has certain limitations. First of all, it introduces an additional problem, generating an optimal partition scheme. Secondly, as it is highlighted in later analysis, the ADPs refined by the multi-group TLS method exhibits great discrepancy between the ADPs of atoms on the boundary of different groups even though those groups are physically bonded. Thirdly, the multi-group TLS method refines the ADPs of different groups individually and neglects the interactions between those groups.

As a goal to fully capture the characteristics of the movements of molecular complexes, in this study, we propose a hierarchical normal mode based anisotropic refinement method (HNMRef). This method is based on a novel normal mode analysis (fSUB (Lu et al., 2012)), which employs a hierarchical normal mode approach, developed in our group recently. In fSUB, the normal mode calculation on the whole complex is performed in two stages. In the first stage, the normal modes of substructures are calculated with substructures in isolation. The normal modes for the whole complex are then constructed on the basis of the substructure modes.

By employing fSUB, we are able to obtain the motions of various hierarchies, such as the deformations within a substructure and the movements between substructures, in a supramolecular complex. Our HNMRef also consists of two stages to model the various hierarchies of complex movements. The ADPs resulting from the low frequency motions of whole complex is first obtained by conventional normal mode refinement method, where the whole structure is refined with a single set of parameters and its low frequency modes. The whole structure is then divided into multiple substructures, each of which is modelled with a set of parameters and the low frequency substructure modes. The correlations between the motions of substructures are considered by introducing a new restraint which penalizes the differences between the ADPs obtained in this stage and the first stage. Unlike multi-group TLS, of which partition schemes are generated by *post hoc* analysis (Painter and Merritt, 2006b), this new method requires minimal considerations about the partition scheme; partitioning the whole structure according to chain division is good enough. Besides, this method addresses the movements of substructures and their correlations simultaneously.

In this thesis, we first introduce the theory of HNMRef. We then assess the new HNMRef by comparing it with the widely adopted multi-group TLS method on a wide range of proteins, using the diffraction data and structures from PDB.

Chapter 2

Normal Mode Analysis

Normal mode analysis (NMA) is a powerful tool for describing the global, collective and functional motions of protein complexes (Brooks and Karplus, 1983; Go et al., 1983; Levitt et al., 1983; Ma, 2005). In this approach, the potential energy function of a protein is assumed to be harmonic so that protein motions can be described as a linear combination of a set of independent harmonic modes. Consequently, the low frequency modes obtained from NMA can serve as the basis to model the atomic displacement parameters of protein complexes in x-ray crystallographic refinement. In this chapter, we introduce the basic ideas behind NMA and present a novel NMA for supramolecular complexes, which is particularly suitable for describing the various hierarchies of the motions of supramolecular complexes.

2.1. Hessian Matrix

In this section, we briefly explain the Hessian matrices of protein structures and their roles in determining the motions of protein structures. Assuming the

potential energy of protein complexes is harmonic around the local minimum represented by the native structure, the potential energy of a conformation whose coordinates deviate from the coordinates of native structure by a small vector $\boldsymbol{\xi}$ can be expressed as

$$U \approx U_0 + \frac{\partial U}{\partial q_i}\xi_i + \frac{1}{2}\frac{\partial^2 U}{\partial q_i \partial q_j}\xi_i\xi_j = \frac{1}{2}\frac{\partial^2 U}{\partial q_i \partial q_j}\xi_i\xi_j, \quad (1)$$

where $\frac{\partial U}{\partial q_i}$ are the first order derivatives of the potential energy function and equal to zero at the local minimum, and $\frac{\partial^2 U}{\partial q_i \partial q_j}$ are the second order derivatives of the potential energy function at the local minimum. The matrix comprised of the all-atom second order derivatives is often called the Hessian matrix of protein complexes. Denote the mass matrix of a protein complex where each diagonal element represents the mass of atom as **M** and the Hessian matrix as **H**, the equation of motion of this protein is

$$\mathbf{M}\ddot{\boldsymbol{\xi}} + \mathbf{H}\boldsymbol{\xi} = \mathbf{0}. \quad (2)$$

where $\ddot{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ are the accelerations and displacements of atoms in this protein, respectively. Hence, the key to solve the equation of motion is to calculate the eigenvectors and eigenvalues for the Hessian matrix. Since all-atom Hessian matrix is of size $3N \times 3N$ for a molecule of N atoms, solving the eigenvalue equation for this matrix when N is large incurs great computational cost. A number of approaches have been proposed to reduce the computational cost. The most notable methods include the rotations-translations blocks (RTB) method (Tama et al., 2000), and elastic network models (Atilgan et al., 2001; Haliloglu et al., 1997; Tirion, 1996).

2.2. fSUB

fSUB is a novel normal mode analysis method which is devised to model the various hierarchies of the motions of supramolecular complexes. In fSUB, a complex structure is divided into n substructures. For each substructure i, the j^{th} normal mode is denoted as $\mathbf{x}_i^j$. A normal mode $\mathbf{y}$ of the whole complex can be divided into n parts, and $\mathbf{y}_i$ is the portion of the normal mode $\mathbf{y}$ corresponding to the substructure i. The normal mode $\mathbf{y}_i$ can be approximated by a linear combination of the first k_i lowest frequency modes of the substructure i, which can be expressed as

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{v}_i, \tag{3}$$

where $\mathbf{X}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^{k_i}\}$, each of which is obtained by solving $\mathbf{H}_i \mathbf{x}_i^{k_i} = \lambda_i^{k_i} \mathbf{x}_i^{k_i}$, and $\mathbf{v}_i$ is a $k_i \times 1$ vector that contains the coefficients for this combination. The eigenvalue and eigenvector problem of the whole complex can be written as

$$\mathbf{H}\boldsymbol{y} = \lambda \boldsymbol{y}, \tag{4}$$

where $\mathbf{H}$ is the hessian matrix for the whole complex. By projecting the original hessian matrix $\mathbf{H}$ into the linear space spanned by substructure modes, namely, substitution Eq. **3** into Eq. **4**, we can express the eigenvalue and eigenvector problem as

$$\mathbf{HPv} = \lambda \mathbf{Pv}, \tag{5}$$

where $\mathbf{P} = \begin{pmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_\mathrm{n} \end{pmatrix}$ is the projection matrix and $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_n^T)^T$ is a collection of coefficients. Since $\mathbf{P}$ is a unitary matrix, multiplication of both sides of Eq. **5** with $\mathbf{P}^T$ yields

$$\mathbf{H}_{\mathrm{fSUB}}\mathbf{v} = \lambda\mathbf{v}, \tag{6}$$

where $\mathbf{H}_{\mathrm{fSUB}} = \mathbf{P}^T\mathbf{H}\mathbf{P}$. Consequently, $\mathbf{v}$ is an eigenvector of the projected Hessian matrix for the whole complex. The eigenvectors of the whole complex can be easily converted from $\mathbf{v}$ to $\mathbf{P}$ by Eq. **3**. The hessian matrix construction and calculation in this study are performed by MGR (Lu and Ma, 2011).

Based on the concept of normal mode analysis, the instantaneous displacements of atoms from their equilibrium positions can be expressed in terms of normal modes and the normal mode variables $\boldsymbol{\sigma}$ as

$$\Delta\mathbf{r} = \mathbf{E}\boldsymbol{\sigma}, \tag{7}$$

where each column of the matrix $\mathbf{E}$ is an eigenvector which represents a pattern of collective motion of atoms. The dimension of matrix $\mathbf{E}$ is $3N\times M$, $M \in [6,3N]$, where $3N$ is the size of all the coordinates of atoms and M is the number of low-frequency modes, including six zero-frequency modes.

Chapter 3

Crystallographic Refinement of ADP

X-ray diffraction data from a protein crystal is a space and time average of various conformers of this protein. These conformers are mainly generated by the constant movements of the protein structure. Hence, an important part of crystallographic refinement is to model the uncertainties of atomic coordinates resulting from its movements. Accurately modelling of the atomic displacement parameters of the protein structure can improve the accuracy of calculated structure factors, which in turn allows the improvement of the structural model. This chapter is devoted to interpreting the role of atomic displacement parameters in crystallographic refinement and introducing common methods to model ADP. At the end of this chapter, we also discuss the metrics to evaluate the quality of refined ADPs.

3.1. Crystallographic Refinement of Atomic Displacements

Since the atom is oscillating around the average position in the 3 dimensional space, the probability of finding an atom in a specific position is usually modeled by a trivariate Gaussian distribution. Thus, the average electron density is given by a convolution between the static atomic electron density and its distribution (Trueblood et al., 1996)

$$\langle \rho_{\text{atom}}(\mathbf{r}) \rangle = \rho_{\text{atom,static}}(\mathbf{r}) * P(\mathbf{u}), \tag{8}$$

where $P(\mathbf{u}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\sigma}|^{1/2}} \exp\{-\frac{1}{2}\mathbf{u}^{\boldsymbol{T}}\boldsymbol{\sigma}^{-1}\mathbf{u}\}$ is the trivariate Gaussian distribution for modelling position uncertainty. The average structure factor can be expressed as the Fourier transform of this convolution, which is the product of the Fourier transform of each function (Trueblood et al., 1996),

$$\langle f(\mathbf{h}) \rangle = f(\mathbf{h})T(\mathbf{h}). \tag{9}$$

Here $T(\mathbf{h})$, the Fourier transform of the gaussian distribution $P(\mathbf{u})$, is,

$$T(\mathbf{h}) = \exp\{-2\pi^2\mathbf{h}^T\boldsymbol{\sigma}\mathbf{h}\}. \tag{10}$$

The variance-covariance matrix $\boldsymbol{\sigma}$, which is also known as atomic displacement parameters (ADP), is often denoted as $\mathbf{U}$ in crystallographic literature. We will use this notation in the following derivation. The previous discussion focus on the structure factor of an atom. For a protein structure with N atoms, its structure factor is a sum of those atomic structure factors, that is,

$$F(\mathbf{h}) \approx \sum_{k=1}^{N} f_k(\mathbf{h})T_k(\mathbf{h}) \exp(2\pi i\mathbf{h} \cdot \mathbf{r}_k), \tag{11}$$

where $\mathbf{r}_k$ is the coordinate for the kth atom in the structure.

The crystallographic refinement is devised to optimize the parameters of the protein model, thus making the structure factors calculated from the optimized model fit with the observed ones. The objective function for optimizing is often in the form,

$$E = \underset{\mathbf{r},\mathbf{U}}{\operatorname{argmin}} \sum_{\boldsymbol{h}} w(\mathbf{h}) l(|F_{\text{calc}}(\mathbf{h},\mathbf{r},\mathbf{U})| - |F_{\text{obs}}(\mathbf{h})|) + N(\mathbf{r}) + N(\mathbf{U}), \tag{12}$$

where l is a convex function, $w(\mathbf{h})$ is the weight factor for the reflection with index $\mathbf{h}$, $|F(\mathbf{h})|$ represents the magnitude of the structure factor, and N are the restraint of the atomic coordinates $\mathbf{r}$ and the ADPs $\mathbf{U}$, respectively. Though this objective function is nonconvex, it can be minimized by the standard techniques for nonlinear optimization in practice.

Assuming the contributions to anisotropy are independent, the ADP can be constructed as (Afonine et al., 2012; Winn et al., 2001)

$$\mathbf{U} = \mathbf{U}_{\text{group}} + \mathbf{U}_{\text{local}} + \mathbf{U}_{\text{cryst}}. \tag{13}$$

$\mathbf{U}_{\text{group}}$ results from the concert motions of atoms within the same group. $\mathbf{U}_{\text{local}}$ is mainly due to individual atomic displacements, such as local vibrations. Thus, it can be well approximated by harmonic oscillator assumption and modeled with the isotropic B factor. $\mathbf{U}_{\text{cryst}}$ originates from the overall displacement of the crystal and some additional experimental anisotropic effects. This effect contributes equally to all atoms, and is treated as a common parameter.

3.2. TLS Method

The TLS method (Schomaker and Trueblood, 1968) is a popular way to model anisotropic thermal parameters in x-ray refinement. In the TLS method, the ADP matrix $\mathbf{U}_{\text{group}}$ is parameterized as

$$\mathbf{U}_{\text{group}} = \mathbf{T} + \mathbf{S}^T \times \mathbf{r}^T - \mathbf{r} \times \mathbf{S} - \mathbf{r} \times \mathbf{L} \times \mathbf{r}^T, \quad (14)$$

where **T**, **L**, and **S** are the translation, libration, and screw matrices, respectively, **r** is the atomic displacement. **T** and **L** are symmetric matrices which can be described by 6 parameters, while **S** is not usually symmetric and includes 8 parameters. In the multi-group TLS formalism, the ADPs of atoms in each group are described by such a set of 20 parameters. The TLS method is available in *phenix.refine* (Afonine et al., 2012). The partition scheme for multi-group TLS method is generated from TLSMD web server (Painter and Merritt, 2006b).

3.3. Normal Mode Based Crystallographic Refinement

In normal-mode-based crystallographic refinement, $\mathbf{U}_{\text{group}}$ is expanded in terms of the effective normal modes. Since the instantaneous atomic displacement of an atom *j* is expressed in terms of normal modes as Eq. **7**, the group ADP for atom *j* becomes

$$\mathbf{U}_{\text{group}} = \langle (\Delta \mathbf{r}_j)^{\mathbf{2}} \rangle = \mathbf{E}_j \langle \boldsymbol{\sigma} \boldsymbol{\sigma}^T \rangle \mathbf{E}_j^T = \mathbf{E}_j \boldsymbol{\Pi} \mathbf{E}_j^T, \quad (15)$$

where $\mathbf{E}_j$ represents the rows in matrix **E** corresponding to atom *j*, and $\boldsymbol{\Pi}$ is the variance-covariance matrix for normal mode variables. In our method, the positive

semidefiniteness of matrix $\mathbf{\Pi}$ is guaranteed by decomposing it into two triangular matrices, that is,

$$\mathbf{\Pi} = \mathbf{\Omega}\mathbf{\Omega}^T. \quad (16)$$

Here, $\mathbf{\Omega}$ are the independent parameters to be optimized against experimentally determined amplitudes in the refinement. The size of the parameter set $\mathbf{\Omega}$ is determined by the number of normal modes. If n normal modes are used, $\mathbf{\Omega}$ includes $\frac{n(n+1)}{2}$ parameters.

3.4. HNMRef

To model the contributions to $\mathbf{U}_{\text{group}}$ from both the complex and substructure motions, HNMRef consists of two stages. In the first stage, the whole complex is assigned with a single set of parameters, $\mathbf{\Omega}$, and the eigenvector matrix $\mathbf{E}$ is made up of the normal modes of the whole complex. Upon the completion of the first stage, the ADPs, $\mathbf{U}_{\text{whole}}$, obtained from this stage are stored for further use. In the second stage, the whole complex is divided into multiple substructures. The division scheme usually is based on the hierarchical structure of complex. For example, each chain in a complex can be designated as a substructure. Each substructure has its own set of parameters $\mathbf{\Omega}_{\text{i}}$. The eigenvector matrix $\mathbf{E}_{\text{i}}$ which serves as the basis to model ADPs for atoms in substructure i is composed of the normal modes of this substructure. To take the motion of whole complex into account, we add a new restraint to the crystallographic target function,

$$E_{\mathrm{nm}} = \sum_{j} \left(\mathbf{U}_{j} - \mathbf{U}_{\mathrm{whole},j}\right)^{2}, \tag{17}$$

where $\mathbf{U}_{\mathrm{whole},j}$ is the ADP for atom j computed in the first stage by optimizing with the normal modes of the whole structure, and $\mathbf{U}_{j}$ is the ADP for atom j evaluated in current stage. The new ADPs then are calculated by minimizing this new crystallographic target function. With this target function, an optimal set of ADPs obtained on the substructure modes should be able to fit well with the observed diffraction intensity while accounting the ADPs originates from the motions of whole complex. We implemented the HNMRef with the *phenix.refine* (Afonine et al., 2012) framework.

3.5. Test Criteria

A widely used technique for assessing the quality of refined model is cross validation. In crystallographic literature, this metric is often called R_{free} (Brünger, 1997). To calculate the R_{free} value for a refined model, the whole dataset of observed reflections should be randomly split in to two sets at the beginning of refinement. One dataset is named as validation set which will not involve in the optimization. The other dataset is training set and used to optimize the model. After the model is refined to converge, the R_{free} value is calculated on the validation set by the following equation,

$$R_{\mathrm{free}} = \frac{\sum_{\mathbf{h}\in \mathrm{validation}} ||F_{\mathrm{obs}}(\mathbf{h})| - |F_{\mathrm{calc}}(\mathbf{h})||}{\sum_{\mathbf{h}\in \mathrm{validation}} |F_{\mathrm{obs}}(\mathbf{h})|}, \tag{18}$$

where $F_{\text{calc}}(\mathbf{h})$ is the structure factor calculated from the refined model. A good model should have a low R_{free} factor, which means that the difference between the data predicted by the model and the experimentally observed data is small. Thus the model with low R_{free} value has small generalization error and superior predictive power.

Another useful criterion for evaluating the refined model is the electron density map, which can be obtained from the inverse Fourier transform of the structure factor. The electron density map obtained from the structure factors calculated by the model is the model itself, thus unable to provide any useful information for us. We here consider some other types of electron density maps which incorporate the experimentally observed intensities. Since the x-ray diffraction data only contains intensities but not phases for the structure factors, all the information about phases in x-ray structure determination is often biased by the model. But with this limited information, we are still able to calculate different types of electron density maps which can reveal the errors of model. The mostly used one is the $(2|F_{\text{obs}}| - |F_{\text{calc}}|)\exp(i\alpha_{\text{calc}})$ map, where the phases are still calculated from the model, while the amplitudes are replaced by the difference between $2|F_{\text{obs}}|$ and $|F_{\text{calc}}|$. This approximately equals with $2\rho_{\text{true}} - \rho_{\text{calc}}$, where ρ_{true} is the true electron density map and ρ_{calc} is the electron density map of the model. Therefore, the difference map has additional densities at where the model has missing atoms, and the difference map has missing densities at where the model atomic positions are incorrect. A similar but more robust electron density map is the weighted difference map proposed by Randy Read (Read, 1986). This map is often denoted as $2mF_o -$

dF_c, where the m and d are weighted coefficients for maps obtained by maximum likelihood method.

We evaluate the difference of refined ADPs for two nearby atoms that belong to the same TLS group or two different TLS groups according to a new metric proposed here. Since the ADPs are the covariance matrices describing the distributions of atomic positions, the difference between ADPs can be measured by the metrics evaluating the difference between two probability distributions. In probability theory, Kullback-Leibler (KL) divergence is often used as such a metric. We defined pair KL divergence, or P_{KL}, for an atom pair i and j as

$$P_{\mathrm{KL}} = \frac{D_{\mathrm{KL}}(i,j) + D_{\mathrm{KL}}(j,i)}{2r_{ij}}, \tag{19}$$

where $D_{\mathrm{KL}}(i,j)$ is the KL distance of the ADP$_{\mathrm{s}}$ of the atoms i and j, $D_{\mathrm{KL}}(j,i)$ is included to make P_{KL} commutative, and inverse of atomic distance r_{ij} is used as weight to emphasize the packed atoms. For two gaussian distributions with the same means, their KL divergence can be computed as

$$D_{\mathrm{KL}}(i,j) = -\frac{3}{2} + \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) + \frac{1}{2}\mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right), \tag{20}$$

where Σ_i is the covariance matrix for the distribution i, $|\Sigma_i|$ is the determinant of Σ_i, and Σ_j^{-1} is the inverse of Σ_j. In our test, the average P_{KL} is evaluated for all atom pairs within 5Å in the same TLS group (denoted as $\bar{P}_{\mathrm{KL}}^{\mathrm{inside}}$) or two different TLS groups (denoted as $\bar{P}_{\mathrm{KL}}^{\mathrm{boundary}}$), and the ratio of the average P_{KL} for these two categories were defined as the boundary indicator

$$B_{\mathrm{KL}} = \frac{\bar{P}_{\mathrm{KL}}^{\mathrm{boundary}}}{\bar{P}_{\mathrm{KL}}^{\mathrm{inside}}}. \quad (21)$$

The B_{KL} tests were also performed on HNMRef refined proteins using the same TLS group partition schemes.

Chapter 4

Results

4.1. Refinement Protocols

To examine the effectiveness of HNMRef method on different systems, we selected 20 deposited structures which represent a wide range of protein complexes from PDB. The resolutions of these structures are within the range 2.0~6.7 Å. To eliminate the possible bias of deposited models, a series of preprocessing was applied on them. First of all, we shook those proteins by running constant temperature molecular dynamics simulation on them for 200 cycles. In addition, the ADPs of atoms of those proteins were all reset to isotropic B-factor of size 20$Å^2$. The refinement protocols were organized in an automated fashion, thus eliminating the bias may be introduced by human intervention. They all consisted of twenty macro-cycles of individual coordinates and ADP refinement with rotamer fitting and peptide side-chain (NQH) flips being turned on. If NCS is available, the NCS restraints would be generated by *phenix.refine* and incorporated in positional refinement automatically. In real scenario, the ADP refinement cycle often consists of several micro-cycles, depending on the ADP parameterization method. In this

work, we considered both the contributions from $\mathbf{U}_{\text{local}}$ and $\mathbf{U}_{\text{group}}$. The ADP refinement cycle started by refining $\mathbf{U}_{\text{group}}$, which was modelled by either the HNMRef and the mutli-group TLS method. To model $\mathbf{U}_{\text{local}}$, we added a micro-cycle in which the local B-factor for each atom was refined with x-ray/ADP weight optimization turned on. As the multi-group TLS method in *phenix.refine* was followed by a cycle of unrestrained group B-factor refinement, to solely compare the effectiveness of HNMRef and TLS method in modelling the large-scale deformations of protein, we removed that cycle. To locate the optimal parameter set for both methods, a grid search where each grid represents a TLS grouping scheme or a combination of the number of complex or substructure modes was performed. The TLS group partition schemes for every peptide chain in the structure were generated by the TLSMD server (Painter and Merritt, 2006b). They differentiate from each other by the number of groups per chain. The number groups per chain was searched from 1 to 20 (or the maximum number of groups per chain can be partitioned) in our test. For the combination of the number of whole complex modes and the number of substructure modes, the number of whole complex modes n varied from 10 to 50 with a step size 5, and for a fixed number of whole complex modes n, the number of substructure modes p was searched from 10 till n with the same step size. In the HNMRef protocol, the weight factor between the new restraint E_{nm} and the target function in *phenix.refine* was set to 0.5.

4.2. R_{free} and Ramachandran Favored Region

As it was shown in Figure 1, 19 out of 20 HNMRef-refined structures had lower R_{free}-factor values comparing to the multi-group TLS structures. The largest improvement in R_{free} values is 2.96% (2I37). The averaged R_{free}-factor value of HNMRef-refined structure is 0.86% smaller than the averaged R_{free}-factor value of the multi-group TLS-refined structures. Besides, most of the $R_{\mathrm{free}} - R_{\mathrm{work}}$ values of HNMRef refined structures were lower than TLS refined ones. The average improvement over TLS refined structures was 0.59%. These two improvements in the cross validation scores demonstrated the HNMRef method significantly reduced overfitting in refinment.

In addition, in 15 out of 20 HNMRef-refinement structures, the percentage of residues in the most favored Ramachandran plot region are increased, with the average being 0.64% higher than those of TLS-refinement structures (Table 1).

PDB ID	Resolution(Å)	R_{free} (%)			$R_{free} - R_{work}$ (%)			Ramachandran Favored (%)		
		HNMRef	TLS	Improvement	HNMRef	TLS	Improvement	HNMRef	TLS	Improvement
4JM2	3.10	27.20	28.19	0.99	5.71	6.43	0.72	91.11	90.59	0.52
2VV5	3.45	30.40	30.82	0.42	3.17	5.26	2.09	87.40	87.63	-0.23
2BBJ	3.90	32.56	34.11	1.55	8.54	9.02	0.48	85.64	85.03	0.61
3CAP	2.90	26.69	26.83	0.14	2.96	2.93	-0.03	92.90	93.06	-0.16
2IOQ	3.51	31.37	30.87	1.26	7.8	7.51	-0.29	86.97	85.37	1.6
4LSN	2.98	29.13	29.56	0.43	4.32	4.52	0.2	92.18	91.87	0.31
2CG9	3.10	36.86	38.07	1.21	9.14	9.77	0.63	77.22	76.80	0.42
3SN6	3.20	25.74	26.86	2.02	4.89	4.93	0.04	90.66	89.58	1.08
2OAR	3.50	29.60	30.23	0.63	2.31	3.70	1.39	85.20	84.23	0.97
2I37	4.15	35.39	41.02	2.96	5.71	6.43	0.72	85.88	84.49	1.39
1QFW	3.50	26.63	27.19	0.56	9.16	9.44	0.28	83.12	82.48	0.64
4K0E	3.71	29.75	30.01	0.26	5.86	6.13	0.27	91.56	91.53	0.03
1PZN	2.85	28.59	29.11	0.52	4.44	4.50	0.06	92.38	92.79	-0.41
2WAQ	3.35	35.70	35.70	0.00	5.92	5.82	-0.07	82.55	81.79	0.76
3TT3	3.22	29.74	30.29	0.55	6.45	7.16	0.71	90.18	89.54	0.64
3E3J	6.70	38.79	30.23	0.44	5.55	7.16	1.61	85.39	85.09	0.30
3MJ9	2.95	31.48	31.72	0.24	4.89	7.3	2.41	89.48	88.51	0.97
3LOH	3.80	30.97	31.37	0.40	4.89	5.13	0.24	83.66	83.66	0.00
4K9E	2.70	29.66	30.55	0.89	8.32	8.88	0.56	95.50	93.09	2.41
3I59	2.29	28.09	29.91	1.82	4.98	4.67	-0.31	96.29	95.30	0.99
Average	3.50			0.86			0.59			0.64

Table 1 R_{free}(%) and Ramachandran Favored regions for 20 protein structures with different refinement methods (The R_{free} values and their improvements for HNMRef refined structures over TLS are shown in percent.)

4.3. Electron Density Map

Other important improvements were found in the electron density maps and structures. Take the structure 3I59 as an example, we calculated $2mF_o - dF_c$ maps using the experimental intensities and phases from the HNMRef- or multi-group TLS- refined structures. Some typical changes in the structure and $2mF_o - dF_c$ map were shown in Figure 4. The Figure 4 (A) showed the omit $2mF_o - dF_c$ map of the HNMRef structure. Comparing with Figure 4 (B), when both contoured at 1.5σ, it had complete densities for sidechain. In addition, in the region of residues 39A-41A,

the HNMRef refined structure fitted well with its $2mF_o - F_c$ map, while some atoms of the multi-group TLS refined structure were located outside its $2mF_o - F_c$ map, which suggested a possible model error.

Figure 3 (C) and (D) presented another electron density map improvement from the HNMRef refined structure in the region of residues 145A-147A. Comparing with the multi-group TLS refined electron density map, the HNMRef refined electron density map had more complete electron densities at this region when both were contoured at 2σ. In this region, the multi-group TLS refined $2mF_o - F_c$ map had either missing or incomplete densities for sidechains. Moreover, in this region, the HNMRef refined structure had shifted to improve the map-coordinate consitency. Another example about the electron density map improvement was in 2CG9. As it was shown in

Figure 3, in the region of residues 329B-331B, the HNMRef refined $2mF_o - dF_c$ map had complete backbone densities, whereas the multi-group TLS refined map was weak and fragmented. Besides, there were some structural shifts in the HNMRef refined structure which made it fit well with the electron density map. The improvements in electron density maps indicated that the HNMRef method improved the phase accuracy, which in turn allowed improvement of the structural model.

4.4. B_{KL}

We computed the B_{KL} values, defined in Method, for the HNMRef refined and multi-group TLS refined structures. The B_{KL} value is the ratio between the average

pair KL divergences of two different types of nearby atom pairs, namely, those belong to the same TLS group and those belongs to two different TLS groups. Therefore, we can use it as an indicator to check the internal consistency of ADPs throughout the whole structure. The higher the B_{KL} value, the bigger the discrepancy between the differences of ADPs of intragroup atom pairs and intergroup atom pairs. As it was shown in Figure 2, the multi-group TLS refined structures all had higher B_{KL} values comparing with the HNMRef refined structures. This highlighted the inconsistency of ADPs throughout the whole structure for multi-group TLS refined structures. It is mainly caused by the arbitrary choice of TLS groups and the limitation of the multi-group TLS method, in that it models ADPs locally and independently across different groups. While HNMRef method doesn't suffer from such drawbacks as it treats each chain as an integral part and its ADPs are restrained by the global modes refined ADPs during refinement.

We also compared the ADPs generated by the HNMRef and multi-group TLS methods by visualizing them as ellipsoids. As it was shown in Figure 5, the atoms of 2VV5 after refinement with the HNMRef method showed greater anisotropy at the end of transmembrane regions, which was consistent with the rotation movement of helices at that region to open the channel (Wang et al., 2008). However, the ellipsoids of the structure refined by the multi-group TLS method didn't present much orientation preference, and failed to depict the highly anisotropic movements of that region.

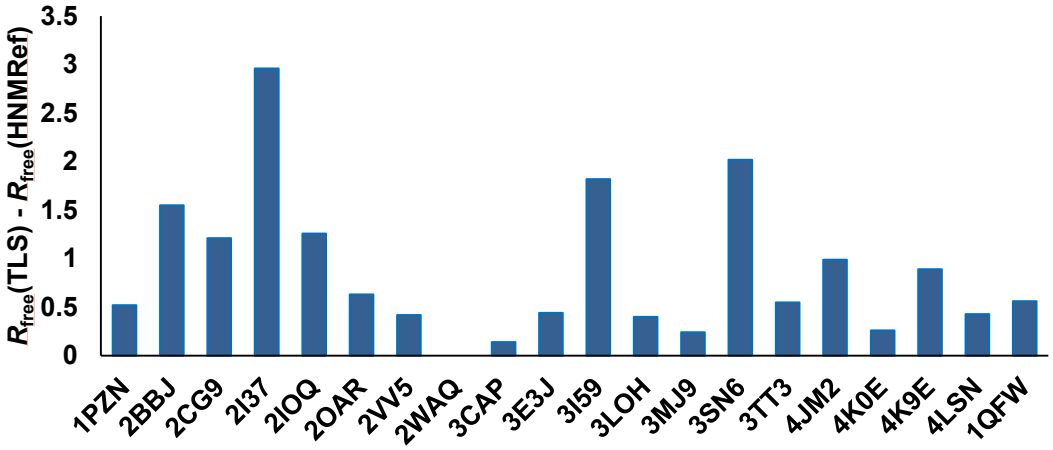

Figure 1 Improvements of R_{free} Values for different proteins

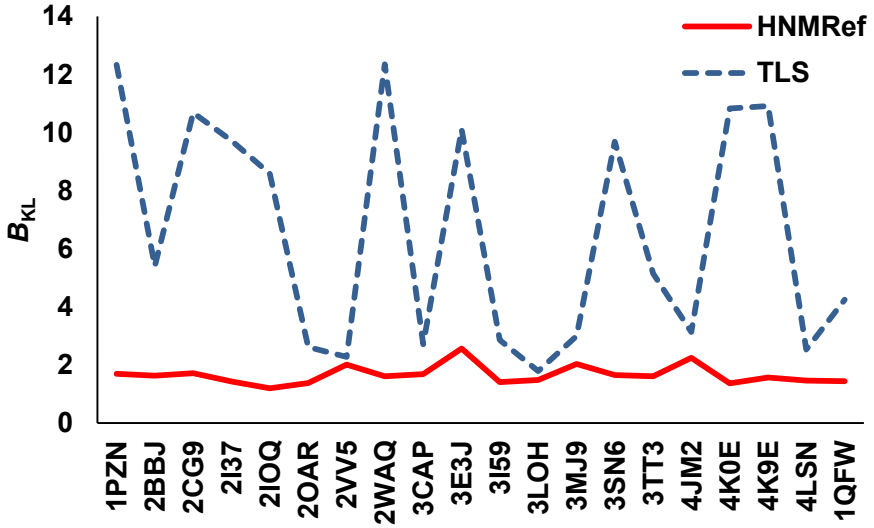

Figure 2 B_{KL} profile for each refined structure

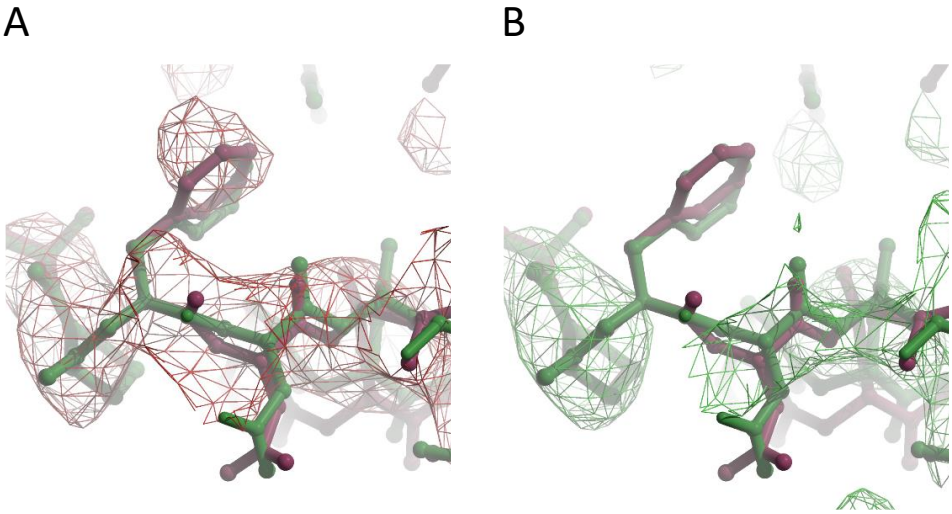

Figure 3 Electron density map and structure comparison for 2CG9

The left shows the electron density map calculated from the HNMRef refined structure, and the right shows the electron density map calculated from the multi-group TLS refined structure. Both $2mF_o - dF_c$ maps are contoured at 2σ. The HNMRef refined structure is colored with red, while the multi-group TLS refined structure is colored with green.

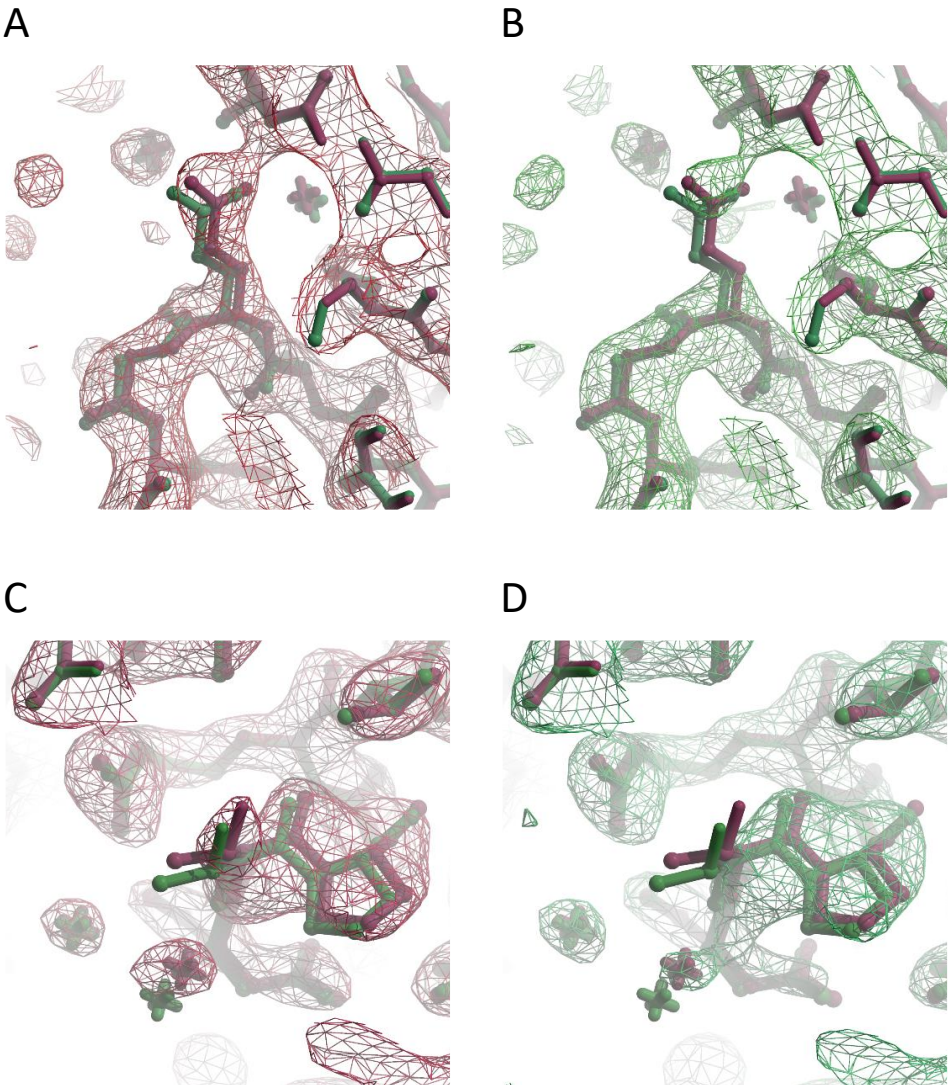

Figure 4 Electron density map and structure comparison for 3I59.

The left shows the electron density maps calculated from the HNMRef refined structure, and the right shows the electron density maps calculated from the multi-group TLS refined structure. The $2mF_o - dF_c$ maps in (A) and (B) are contoured at 1.5σ, and the $2mF_o - dF_c$ maps in (C) and (D) are contoured at 2.0σ. The HNMRef refined structures are colored with red, while the multi-group TLS refined structures are colored with green.

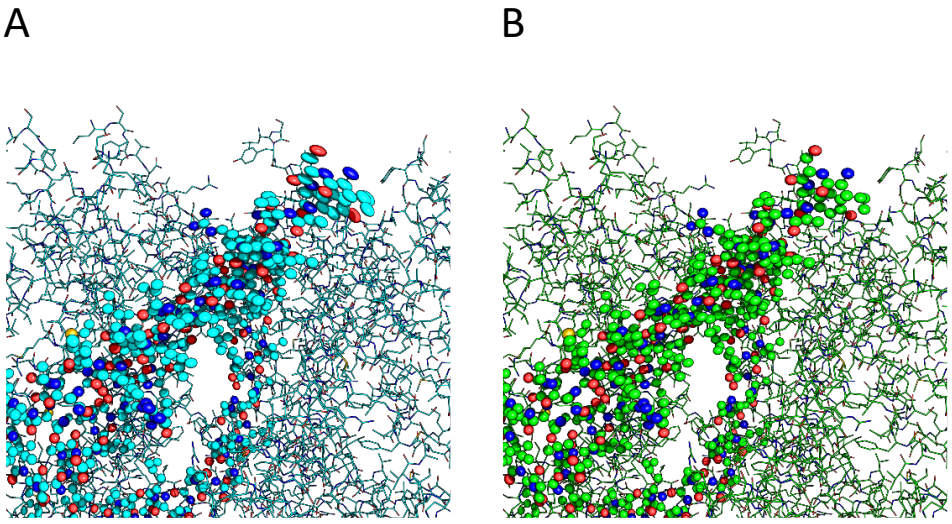

Figure 5. The open structure of an E.coli mechanosensitive channel at 3.45 Å showing 50% probability ellipsoids refined with (A) HNMRef and (B) multi-group TLS methods.

Conclusions

We here presented a new hierarchical normal mode based anisotropic refinement method (HNMRef). Comparing with its precursors, this new method has shown to greatly reduce the generalization errors of models (lower R_{free} values). This is an evidence for the superior ability of our new method to accurately model the motions of large complexes. The improvements of models have been further validated by inspecting their electron density maps. We also compared the ADPs generated by the HNMRef and multi-group TLS methods using a new metric B_{KL}. This comparison highlighted the unrealistic part of the multi-group TLS method; there is a great discrepancy between the differences of ADPs of intergroup atom pairs and intragroup atom pairs. This defect is caused by the problematic chain partition schemes of the multi-group TLS method and the negligence of correlations exist among the movements of different groups. In contrast, the HNMRef has managed to model the various hierarchies of complex motions without introducing inconsistency to the ADPs throughout the whole structure. Another benefit of the HNMRef method is that it minimizes the effort to search an optimal partition scheme. Therefore, we expect our method to be deployed in scenario like the anisotropic refinement of large complex with moderate resolution. Our method is expected to expedite the refinement process by delivering more accurate and physical model for those complexes.

References

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., and Adams, P.D. (2012). Towards automated crystallographic structure refinement with phenix. refine. Acta Crystallographica Section D: Biological Crystallography *68*, 352–367.

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

Brooks, B., and Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proceedings of the National Academy of Sciences *80*, 6571–6575.

Brünger, A.T. (1997). Free R value: cross-validation in crystallography. Methods in Enzymology *277*, 366.

Diamond, R. (1990). On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. Acta Crystallographica Section A: Foundations of Crystallography *46*, 425–435.

Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proceedings of the National Academy of Sciences *80*, 3696–3700.

Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian dynamics of folded proteins. Physical Review Letters *79*, 3090.

Kidera, A., and Go, N. (1990). Refinement of protein dynamic structure: normal mode refinement. Proceedings of the National Academy of Sciences *87*, 3718–3722.

Kidera, A., and Gō, N. (1992). Normal mode refinement: crystallographic refinement of protein dynamic structure: I. Theory and test by simulated diffraction data. Journal of Molecular Biology *225*, 457–475.

Kidera, A., Inaka, K., Matsushima, M., and Gō, N. (1992). Normal mode refinement: crystallographic refinement of protein dynamic structure applied to human lysozyme. Biopolymers *32*, 315–319.

Kidera, A., Matsushima, M., and Gō, N. (1994). Dynamic structure of human lysozyme derived from X-ray crystallography: normal mode refinement. Biophysical Chemistry *50*, 25–31.

Levitt, M., Sander, C., and Stern, P.S. (1983). The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. International Journal of Quantum Chemistry *24*, 181–199.

Lu, M., and Ma, J. (2005). The Role of Shape in Determining Molecular Motions. Biophysical Journal *89*, 2395–2401.

Lu, M., and Ma, J. (2011). Normal mode analysis with molecular geometry restraints: Bridging molecular mechanics and elastic models. Archives of Biochemistry and Biophysics *508*, 64–71.

Lu, M., Ming, D., and Ma, J. (2012). fSUB: Normal Mode Analysis with Flexible Substructures. The Journal of Physical Chemistry B *116*, 8636–8645.

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

Painter, J., and Merritt, E.A. (2006a). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallographica Section D Biological Crystallography *62*, 439–450.

Painter, J., and Merritt, E.A. (2006b). TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography *39*, 109–111.

Petrone, P., and Pande, V.S. (2006). Can Conformational Change Be Described by Only a Few Normal Modes? Biophysical Journal *90*, 1583–1593.

Poon, B.K., Chen, X., Lu, M., Vyas, N.K., Quiocho, F.A., Wang, Q., and Ma, J. (2007). Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-\AA crystallographic resolution. Proceedings of the National Academy of Sciences *104*, 7869–7874.

Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallographica Section A: Foundations of Crystallography *42*, 140–149.

Schomaker, V., and Trueblood, K.N. (1968). On the rigid-body motion of molecules in crystals. Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry *24*, 63–76.

Tama, F., Gadea, F.X., Marques, O., and Sanejouand, Y.-H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. Proteins: Structure, Function, and Bioinformatics *41*, 1–7.

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

Trueblood, K., Bürgi, H.-B., Burzlaff, H., Dunitz, J., Gramaccioli, C., Schulz, H., Shmueli, U., and Abrahams, S. (1996). Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. Acta Crystallographica Section A: Foundations of Crystallography *52*, 770–781.

Wang, W., Black, S.S., Edwards, M.D., Miller, S., Morrison, E.L., Bartlett, W., Dong, C., Naismith, J.H., and Booth, I.R. (2008). The structure of an open form of an E. coli mechanosensitive channel at 3.45 \AA resolution. Science *321*, 1179–1183.

Winn, M.D., Isupov, M.N., and Murshudov, G.N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallographica Section D: Biological Crystallography *57*, 122–133.