

2009

# Chemical Information Based Elastic Network Model: A Novel Way To Identification Of Vibration Frequencies In Proteins.

Sharad K. Raj

*University of Massachusetts Amherst*

Follow this and additional works at: <https://scholarworks.umass.edu/theses>

 Part of the [Molecular Biology Commons](#), and the [Structural Biology Commons](#)

---

Raj, Sharad K., "Chemical Information Based Elastic Network Model: A Novel Way To Identification Of Vibration Frequencies In Proteins." (2009). *Masters Theses 1911 - February 2014*. 261.

Retrieved from <https://scholarworks.umass.edu/theses/261>

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).

CHEMICAL INFORMATION BASED ELASTIC NETWORK MODEL: A NOVEL  
WAY TO IDENTIFICATION OF VIBRATION FREQUENCIES IN PROTEINS.

A Thesis Presented

by

SHARAD K. RAJ

Submitted to the Graduate School of the  
University of Massachusetts Amherst in partial fulfillment  
of the requirements for the degree of

MASTER OF SCIENCE

December 2008

Mechanical and Industrial Engineering

CHEMICAL INFORMATION BASED ELASTIC NETWORK MODEL: A NOVEL  
WAY TO IDENTIFICATION OF VIBRATION FREQUENCIES IN PROTEINS.

A Thesis Presented

by

SHARAD K. RAJ

Approved as to style and content by:

---

Moon K. Kim, Chair

---

Byung H. Kim, Co-chair

---

Robert W. Hyers, Member

---

Mario Rotea, Department Head  
Mechanical & Industrial Engineering

## ACKNOWLEDGEMENTS

I would like to thank Professor Moon Kim for his constant support, patience, and guidance throughout my academic career at the University of Massachusetts Amherst. Professor Moon Kim is a remarkable personal and it has been a privilege to both work and study under him. I would also like to thank Professor Byung Kim for his continuous guidance and directions during my stay here and Professor Hyers for his assistance and insight along the way. I would also like to extend a special thanks to my lab mate and dear friend Ming-Wen Hu for his constant help, support and assistance. Lastly, I would like to thank my family and friends. I can't thank you enough for all you've done for me.

## ABSTRACT

### **CHEMICAL INFORMATION BASED ELASTIC NETWORK MODEL: A NOVEL WAY TO IDENTIFICATION OF VIBRATION FREQUENCIES IN PROTEINS.**

DECEMBER 2008

SHARAD K. RAJ, B.S., UNIVERSITY OF MUMBAI, INDIA

M.S. UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Moon K. Kim

A novel method of analysis of macromolecules has been worked upon through this research. In an effort to understand the dynamics of macromolecules and to further our knowledge, pertaining specifically to the low frequency domain and also to elucidate certain important biological functions associated with it, a theoretical technique of chemical information based Normal Mode Analysis has been developed. These simulations render users with the ability to generate animations of modeshapes as well as key insight on the associated vibration frequencies. Harmonic analysis using atomistic details is performed taking into account appropriate values of masses of constituent atoms of a given macromolecule. In order to substantiate the applicability of such a technique, simple linear molecules were first worked upon. Subsequently, this technique has been applied to relatively more complex structures like amino acids, namely Cysteine. Consequently, this approach was extended to large macromolecules like Lactoferrin. Animations of modeshapes from simulations suggest a one to one correspondence with other computational techniques reported by other researchers.

Computed  $\beta$ -factors are also in close agreement with the experimentally observed values of the same. Hence, as opposed to a simple  $C_\alpha$  coarse grained model, our method with right masses and reasonable force fields yields not only the correct modeshapes but also provides us with useful information on wavenumbers that can be used to extract useful information about the frequency domain. Moreover, as opposed to conventional Molecular Dynamics' simulations and Laser spectroscopy, the proposed methodology is significantly faster, cheaper and efficient.

## TABLE OF CONTENTS

	<b>Page</b>
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x
<b>CHAPTER</b>	
<b>1. BACKGROUND OF PROTEINS AND THEIR DYNAMICS.....</b>	<b>1</b>
1.1 Proteins.....	1
1.2 Protein Dynamics.....	11
1.3 Applications of Spectroscopy in Biomolecules.....	11
<b>2. CHEMICAL INFORMATION BASED NORMAL MODE ANALYSIS.....</b>	<b>13</b>
2.1 Introduction.....	13
2.2 Linking Matrix.....	15
2.3 Stiffness and Mass matrices.....	18
<b>3. MODESHAPE ASSIGNMENT AND FREQUENCY DETERMINATION IN LINEAR MOLECULES.....</b>	<b>20</b>
3.1 Introduction.....	20

3.2 Methodology.....	22
3.2.1 Force constants in Acetylene.....	22
3.2.2 Optimization of computed force constants.....	23
3.3 Results and discussions.....	25
3.4 Conclusions.....	27
<b>4. APPLICATION OF CHEMICAL INFORMATION BASED NMA TO AMINO ACIDS.....</b>	<b>29</b>
4.1 Introduction.....	29
4.2 Methodology.....	30
4.3 Results.....	32
4.4 Discussions.....	36
4.4.1 Force field parameterization.....	36
4.4.2 Sensitivity to the cutoff distance.....	38
4.5 Conclusions.....	41
<b>5. ANALYSES OF MACROMOLECULES USING CHEMICAL INFORMATION BASED NMA.....</b>	<b>43</b>
5.1 Introduction.....	43
5.2 Methodology.....	44
5.3 Results and discussions.....	44



5.4 Sensitivity analysis.....	51
5.5 Computational complexity.....	57
5.6 Conclusions.....	61
<b>6. HYBRID NORMAL MODE ANALYSIS USING CHEMICAL INFORMATION BASED ELASTIC NETWORK MODEL.....</b>	<b>62</b>
6.1 Introduction.....	62
6.2 Methodology.....	63
6.3 Results and discussions.....	67
6.4 Conclusions.....	70
<b>7. CONCLUSIONS AND FUTUREWORK.....</b>	<b>72</b>
7.1 Conclusions.....	72
7.2 Future work.....	73
<b>APPENDICES</b>	
A: THE LINKING MATRIX CODE, ALL ATOM NMA.....	75
B: ALL ATOM NMA CODE.....	82
C: HYBRID NMA CODE.....	84
<b>BIBLIOGRAPHY.....</b>	<b>91</b>

## LIST OF TABLES

Table	Page
3.1: A list of bond specific force constants.....	25
3.2: Comparison between experimental and predicted frequencies.....	26
4.1: Cartesian coordinates of a nominal Cysteine structure used for running the simulations.....	31
4.2: Comparison between experimental and predicted frequencies for L-Cysteine.....	32
4.3: Computed wavenumbers based on both generalized and bond specific force constants of $7 \times 10^5$ dynes/cm, a non-bonded force constant of $6 \times 10^3$ dynes/cm, and a cutoff distance of $8\text{\AA}$ were applied when computing vibration frequencies.....	34
4.4: Computed frequencies of Cysteine based on different values of cutoff distance.....	39
5.1: Represents a list of atoms observed to have high values of computed $\beta$ -factors numbers, their types and the amino acids they constitute.....	49
6.1: Represents one of the two clustering schemes used to run HNMA simulations on Lactoferrin.....	66
6.2: Represents the other clustering scheme used to run HNMA simulations on Lactoferrin.....	67

## LIST OF FIGURES

Figure	Page
1.1: Represents the 20 natural amino acids with their three letter and single letter abbreviations.....	4
1.2: Represents primary protein structure in a sequence of a chain of amino acids.....	5
1.3: Represents the peptide bond between consecutive amide and carboxyl groups along the backbone.....	6
1.4: Representation of protein structure as a coarse-grained elastic network.....	9
1.5: Representation of the linking matrix of 2BOH obtained from a $C\alpha$ coarse grained model based NMA.....	10
2.1: Represents the sequence of amino acids for Lactoferrin's open form, 1LFH.....	16
3.1: Schematic of acetylene (a) a ball and stick representation of the ENM model setup using the appropriate values of masses and spatial coordinates: (b) a chemical diagram of acetylene.....	22
3.2: Animation of $C\equiv C$ stretching mode in acetylene.....	23
3.3: Optimization of spring constants.....	24
4.1: Modeshape animations of Cysteine with their corresponding frequencies for the first four computed modes.....	33
4.2: A plot of normalized wavenumbers versus the mode number for the case of generalized force fields which elucidates the observed divergence at higher modes resulting from a variation in the non-bonded force constants.....	35
4.3: Plot of the ratio between two consecutive wavenumbers $\omega_{i+1} / \omega_i$ against mode number representing the variance of wavenumbers in terms of cutoff distance.....	40
5.1: First three modes for 1LFH obtained from simulations using chemical information based NMA.....	45
5.2: First three modes for 1LFH obtained from simulations using $C\alpha$ NMA.....	46

<b>5.3:</b> A plot showing experimental versus calculated $\beta$ -factors for Lactoferrin.....	47
<b>5.4:</b> Images of atoms on the outer periphery obtained from the conformation of Lactoferrin obtained from PDB.....	50
<b>5.5:</b> Semi-logarithmic plot of computed as well as experimental $\beta$ -factors for Lactoferrin from all-atom NMA simulation.....	51
<b>5.6: (a)</b> Plot of normalized wavenumbers for Lactoferrin against the mode number.....	52
<b>5.7:</b> First three modes for 1LFG obtained from simulations using chemical information based NMA.....	54
<b>5.8:</b> First three modes for 1LFG obtained from simulations using $C_{\alpha}$ NMA.....	55
<b>5.9:</b> Represents a plot of run time for generating the linking matrix in all NMA simulation, showing the variation in the same as a function of the number of atoms of the protein.....	59
<b>5.10:</b> Represents a plot of run time for running the all atom NMA simulation Showing the variation in the same as a function of the number of atoms of the protein.....	60
<b>6.1:</b> Schematic of the hybrid elastic network model for the complex structure which contains both rigid domains and flexible loop regions.....	63
<b>6.2:</b> A rigid-cluster model of the Lactoferrin structure.....	64
<b>6.3:</b> Represents animations of the first three modes obtained by running the the HNMA simulations on 1LFH.....	68
<b>6.4:</b> Represents the animations of the first three modes of 1LFH by running HNMA simulations on a model defined to have five clusters.....	69

## CHAPTER 1

### BACKGROUND OF PROTEINS AND THEIR DYNAMICS

#### 1.1 Proteins

The word “protein” is derived from the Greek word proteios, meaning “primary” or “first rank of importance.” This chapter discusses proteins and their three-dimensional structures, along with the basic amino acids that are known as the building blocks of Proteins.

Proteins are known to form the very basis of life. They perform a disparate set of functions and regulate a variety of activities in all living organisms; from the process of translation and transcription, i.e. replication of the genetic code to transporting oxygen, and are generally responsible for regulating the cellular machinery and consequently, the phenotype of an organism. In the form of skin, hair, callus, cartilage, muscles, tendons and ligaments, proteins hold together, protect, and provide structure to the body of a multi-celled organism. In the form of enzymes, hormones, antibodies, and globulins, they catalyze, regulate, and protect the body chemistry. Proteins accomplish their task by three-dimensional tertiary and quaternary interactions between various substrates such as DNA and RNA, and other proteins. Knowledge about the structure of the protein, enables scientists to further probe in to deciphering the observed phenomenon of protein folding that is associated with certain specific function.

In order to better understand a protein’s structure, primarily, most efforts are concentrated on the prediction of the sequence of the amino acids that compose any

given protein. While having the information on the constituent amino acids and their sequence is a necessary but insufficient requirement. The reason for this can be observed from the fact that almost every protein structure has known to exist in not a linear but complex 3-D structures. It is in this orientation that a protein holds the relevance of any folding that it may go through.

Proteins can be fold into a variety of 3-dimensional shapes. Experiments to unfold and refold proteins have shown that the amino acid sequence itself contains all the instructions needed for proper folding. Scientists across the world have been working on trying to understand the basic principles governing folding but have predicting a 3-dimensional structure merely from amino acids' sequence remains a challenging task. Researchers across the world have so far successfully studied and determined the functions of many proteins using a variety of methods. Yet, with the plethora of proteins that have been discovered post the Genome project, has implied that the study of protein still offers a variety of challenges to better comprehend its biological significance.

Work on the human genome has revealed that there are 20,000–25,000 genes [1]. It is fascinating if not impossible, to believe that considering the post-translational processes which yield close to a 100,000 proteins are essentially made up of just 20 amino acids. Each amino acid consists of a carboxylic acid group (COOH), an amino group (NH<sub>2</sub>), and one of twenty functional (R) groups. This R group, or side-chain, varies between amino acids from a simple hydrogen atom in the amino acid glycine to a complex structure found in tyrosine. Amino acids polymerize at the carboxylic acid group of one amino acid to the amino group of the next to form a peptide. A protein is a

long polypeptide chain. Every protein adopts its distinct function and structure from the unique sequence of the composing amino acids and their chemical properties. Removing one amino acid or changing it from the protein sequence can significantly alter its structure and the same would hold true with regard to its biological functioning. Since some of the natural amino acids are not synthesized by human metabolic processes, they are essential diet components. The best food source of these nutrients is protein, but it is important to recognize that not all proteins have equal nutritional value.

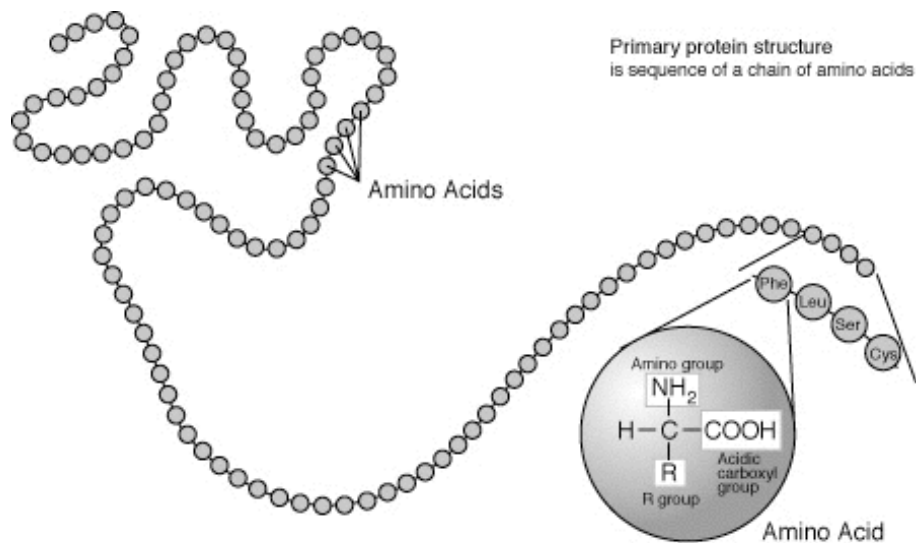
The **Figure 1.1** shows the essential natural amino acids. As discussed above, there are certain amino acids that are not synthesized during the metabolism process, and are marked in green. The usual nomenclature is to represent every amino acid with one and three letter abbreviations.

Name	Formula	Abbreviations	Name	Formula	Abbreviations
Glycine		Gly G	Cysteine		Cys C
Alanine		Ala A	Methionine		Met M
Valine		Val V	Lysine		Lys K
Leucine		Leu L	Arginine		Arg R
Isoleucine		Ile I	Histidine		His H
Phenylalanine		Phe F	Tryptophan		Trp W
Proline		Pro P	Aspartic Acid		Asp D
Serine		Ser S	Glutamic Acid		Glu E
Threonine		Thr T	Asparagine		Asn N
Tyrosine		Tyr Y	Glutamine		Gln Q

**Figure 1.1:** Represents the 20 natural amino acids with their three letter and single letter abbreviations. The ones represented in green are those that are not synthesized during the metabolic processes in the body.

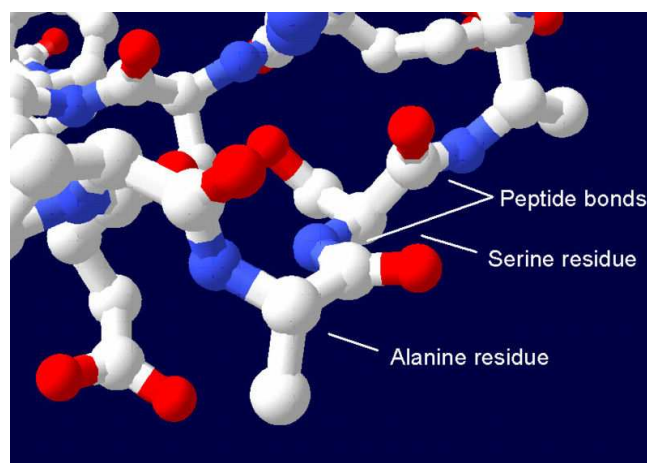
In any given structure of a protein, the amino acids are linked together by a peptide bond. The **Figure 1.2** below shows an arbitrary schematic representation of a given protein structure. The various amino acids that comprise this primary structure are connected to each other by a chemical bond between the  $C_{\alpha}$  and nitrogen atoms





**Figure 1.2:** Represents primary protein structure in a sequence of a chain of amino acids.

of the carboxyl groups and amino groups of two consecutive amino acids, respectively, and this in fact is referred to as a peptide bond. It is this link of peptide bonds that form in a proteins what is referred to as the back bone. The **Figure 1.3** below gives a representation of a peptide bond in any general amino acids' sequence for a protein structure.



**Figure 1.3:** Represents the peptide bond between consecutive amide and carboxyl groups along the back bone.

As discussed, it can be observed that a given protein structure and conformation are a direct effect of the sequence and the orientation of its constituent amino acids that ultimately are responsible for the a distinct folding characteristics that in turn are the most crucial factor in any observed biological function associated with a given protein. This phenomenon facilitates the study of motions of proteins to further our understanding of folding. This class of research is often referred to as protein dynamics. Numerous researchers and scientists across the world are working on experimental and theoretical platforms to exploit these interesting characteristics of proteins. Though this is a very vast field of research, certain important elements of the current work along with excerpts and an overview of both experimental as well as theoretical approaches in practice today, are discussed in the following section.

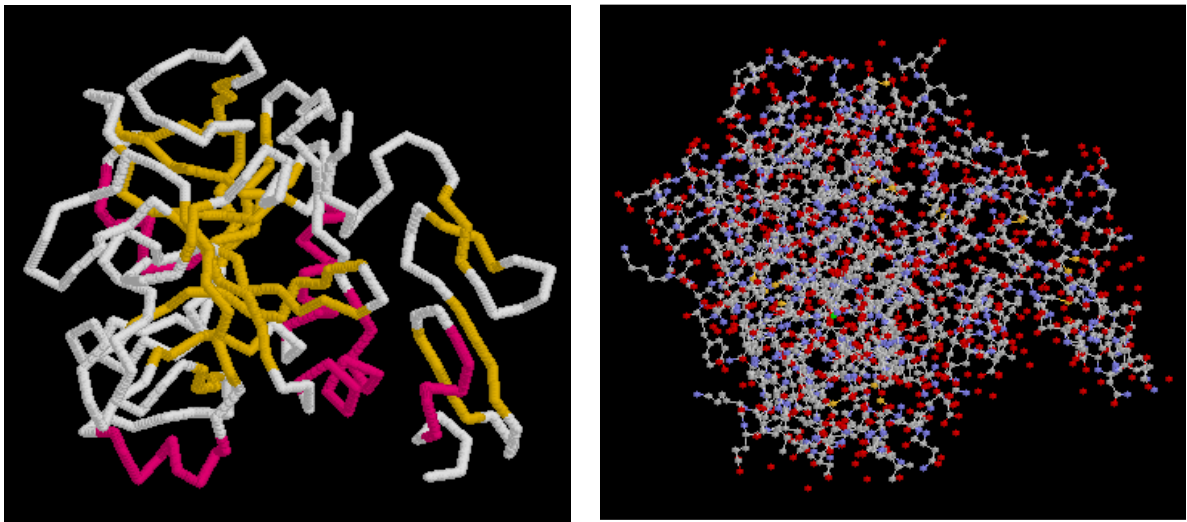
## **1.2 Protein Dynamics**

Protein molecules adopt different conformations depending upon whether its active centre is occupied or empty. With the advent of x-ray structural analysis, usually a protein's conformation can be classified as open and a closed conformation. A closed form is defined as the physical state of a protein after it has grabbed an atom. For example, Lactoferrin is responsible for transporting iron atoms throughout the body. So, co-ordinates of the form with an iron rich center would be classified as a closed form and vice versa. While X-ray crystallography can give structural information of the two forms, more rigorous experimental and theoretical techniques have been developed to expand the understanding of dynamics of proteins.

As a number of protein and nucleic acid structures have been obtained experimentally and deposited in the Protein Data Bank [2], biological research areas such as computational biology, bioinformatics, and protein dynamics have been growing rapidly. Among the many ways to analyze the dynamic characteristics of macromolecules, the conventional engineering disciplines such as kinematics and mechanical vibrations have been proved as a powerful tool to reduce computational cost and the generated results have shown a good agreement with the experimentally observed dynamics of macromolecules. From this fact, it is convincing that they can play an important role in developing much more computationally efficient methods than traditional molecular dynamics (MD) simulations, as well as establishing theoretical foundation for linking the structural information of macromolecules to their biological functions. MD simulations conventionally have been one of the most common tools directed towards the study of protein dynamics. With the advent of super computers and a much better estimation of empirical potential energies, MD inspite of its computational burden, remains a very precise and accurate method in this domain. As MD is based upon an energy minimization approach, the output from such an analysis is highly dependent on the step size or the interval. Hence, in addition to being dependent on empirical potential energy function, energy conservation can be violated in simulations because of an insufficiently short integration time step or an inaccurate representation of the intermolecular forces or the non-bonded interactions governed by the principles of Lennard-Jones potential or the electrostatic forces present in a given proteins.

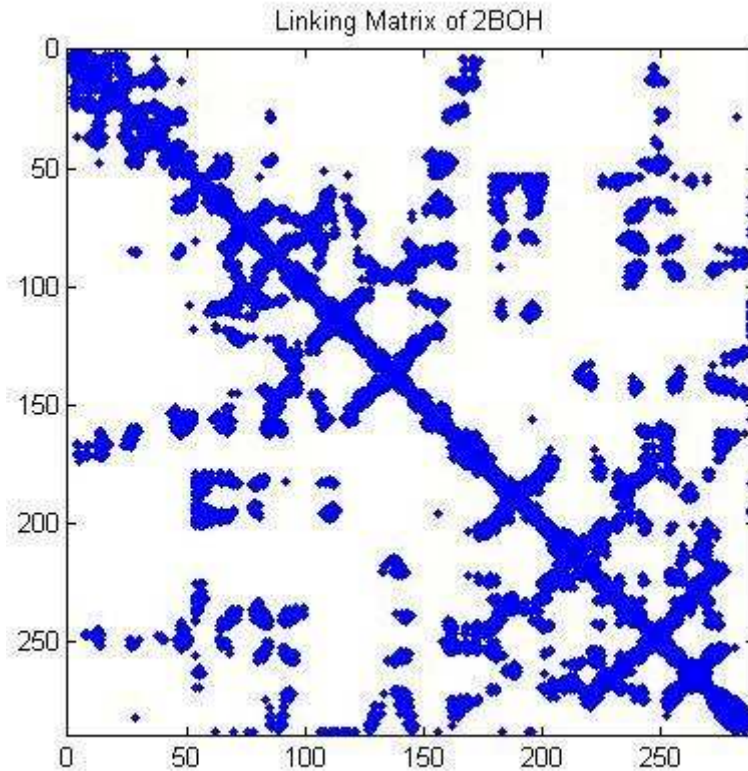
To overcome these drawbacks various coarse-graining approaches have been developed. Amidst many approaches, a  $C_{\alpha}$  coarse-grained Elastic Network Model (ENM) based Normal Mode Analysis (NMA) has been a broadly used as a harmonic analysis tool. In this approach, any given macromolecule is represented as a spring mass system. Various atoms that compose any given macromolecule are considered as point masses, and their chemical interactions with the neighboring atoms are represented by linear massless springs. This representation leads to setting up of a mathematical model of a given system which is referred to as an ENM. In ENM, This concept of ENM was first introduced by Tirion and other researchers [3]. In a conventional  $C_{\alpha}$  coarse grained model, a representative  $C_{\alpha}$  atom is selected from each amino acid in the sequence. So far, as the models considered have been based on representative  $C_{\alpha}$  atoms, and thus, though computationally efficient, they have certain limitations and limited applications. Traditionally, the point masses have been considered as unity masses while the spring constants analogous to the force constants have been represented using different schemes mainly; binary assignment: assigning 1 to represent the presence of a bond, or 0 to represent the absence of a bond. This is done without any consideration of the type of chemical interaction between atoms, i.e. covalently bonded or non-covalently bonded. There have also been some assignment schemes that propose incorporating bonded force constants as a ratio of the non bonded force constant, or using a certain cutoff distance to replicate the real scenario of atomic interactions by making connections with the neighboring atoms such that from a given atom, any other atom within a certain range of distance in the 3-D space would be assigned a bond with. This range is defined by the

defined cutoff distance. Hence, this type of an approach has worked well in the low frequency domain, to visualize modeshapes but as previously discussed, provide little information on the frequency associates with these modeshapes. Therefore, the most profound limitation of a coarse grained NMA tool is that the set of eigenvalues computed that represents the wavenumbers of the various observed modeshapes does not possess any physical meaning since only representative alpha Carbon atoms with unity mass are used. Also, the linking matrix representing the connection of the various alpha Carbon atoms in accordance with the amino acid sequence for a given protein being based on a distance cutoff method has proven to be yet another limiting factor in computing the wavenumbers.



**Figure 1.4:** Representation of protein structure as a coarse-grained elastic network. The all atom model (PDB code: 2BOH) is illustrated with a ball and stick representation (right). On the other hand, only  $C\alpha$  atoms are selected as representatives and the spring connections between atoms within a cutoff distance of  $8\text{\AA}$  are represented (left).

The Figure 1.4 shows one of the many representation schemes of crystal structure of factor XA in complex with compound "1". The image on the left represents the model with just the  $C\alpha$  atoms connected to each other. Consecutive  $C\alpha$  atoms here yield what is referred to as previously mentioned, the backbone. This type of an ENM is utilized to perform NMA. On the other hand, the image to the right is a ball and stick representation of the same protein, showing all the constituent atoms. These images were generated by using RASMOL. By default, Oxygen, Carbon and Nitrogen atoms are represented by the balls of colors red, white and blue, respectively. Once NMA is performed in either case, an eigenvector set is generated, and subsequently, the positional co-ordinates of the original PDB file are altered to generate images of perturbation along the computed eigenvectors to yield animations for various modeshapes. Figure 1.5 shows the generated linking matrix, which gives a representation of the connectivity of the constituent atoms. It is densely packed around the diagonal since atoms are bound to be connected to other atoms in its near vicinity.



**Figure 1.5:** Representation of the linking matrix of 2BOH obtained from a  $C_{\alpha}$  coarse grained model based NMA.

### 1.3 Applications of Spectroscopy in Biomolecules.

When atoms and molecules are incident upon by electromagnetic radiation, absorption, emission, or scattering phenomenon of such radiation is observed. These observed phenomena are quantified to study numerous such atoms or molecules, or to study their physical processes. The interaction of radiation with matter can cause redirection of the radiation and/or transitions between the energy levels of the atoms or molecules. A transition from a lower level to a higher level with transfer of energy from the radiation field to the atom or molecule is called absorption. A transition from a

higher level to a lower level is called emission if energy is transferred to the radiation field or nonradioactive decay if no radiation is emitted. Redirection of light due to its interaction with matter is called scattering, and may or may not occur with transfer of energy, i.e., the scattered radiation has a slightly different or the same wavelength. Hence, these observed effects are of great importance in the study of numerous molecules and their properties.

Spectroscopy has been widely used in the study of biomolecules and along with other significant areas of applications. It can be observed that many such efforts have been directed towards disparate domains of structural analysis or in deciphering biological functions of numerous biomolecules [4-9]. Raman spectroscopy is a very commonly used method in the field of spectroscopy. It has conventionally been widely used in fundamental chemistry since vibrational information is very specific for the chemical bonds in molecules [10-13]. One of its important biological applications has been to study changes in chemical bonding when a substrate is added to an enzyme. In solid state physics, spontaneous Raman spectroscopy is used to, among other applications, to characterize materials, measure temperature, and find the crystallographic orientation of a sample. In addition to its applicability in large molecules and complex biomolecules, in single molecules, it is often used to determine and identify phonon modes.



## CHAPTER 2

### CHEMICAL INFORMATION BASED NORMAL MODE ANALYSIS

#### 2.1 Introduction

As mentioned in the previous section, ENM has been widely used for analyzing macromolecular dynamics in frequency domain [14-21]. In chemical information based NMA, like conventional  $C\alpha$  coarse grained ENM based NMA, a biomolecule can be treated as a mass-spring system. Thereby, each atom is an individual point mass and is connected to other atoms by virtual springs. In accordance with the present methodology, the positional coordinates of a given macromolecule are obtained from the Protein Data Bank. It is then modeled to be consisting of point masses, values of which were consistent with the atomic masses of the constituent atoms. These point masses are connected to each other by massless springs, representing the interactions between different atomic pairs, with values of stiffness analogous to the force constant between any two given atoms.

With regard to more complex structures of molecules, in the scope of the undertaken modeling scheme and force field parameterization, the effect of non-bonded interactions are more dominant than what is observed for linear molecules due to the orientation and structural conformation of most biomolecules, for example, like that for Cysteine, a certain distance cutoff scheme is adopted to replicate the non-bonded interactions. In such a system, the total kinetic energy  $T$  and potential energy  $V$  in a network of  $n$  point masses can be presented as

$$T = \frac{1}{2} \sum_{i=1}^n m_i \|\dot{\bar{x}}_i(t)\|^2, \quad (2.1)$$

$$V = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n k_{i,j} \left\{ \|\bar{x}_i(t) - \bar{x}_j(t)\| - \|\bar{x}_i(0) - \bar{x}_j(0)\| \right\}^2, \quad (2.2)$$

where  $\bar{x}_i(t)$  is the position of the  $i^{\text{th}}$  atom at time  $t$ ,  $m_i$  is the mass of the  $i^{\text{th}}$  atom,

and  $k_{i,j}$  is called “linking matrix” which is defined as

$$k_{i,j} = \begin{cases} C & \|\bar{x}_i - \bar{x}_j\| \leq l \\ C_x & l \leq \|\bar{x}_i - \bar{x}_j\| \leq d, \\ 0 & \|\bar{x}_i - \bar{x}_j\| \geq d \end{cases} \quad (2.3)$$

where  $d$  is a certain cutoff distance such that for a given atom, any atom which is farther than the distance  $d$  is considered to not have a connection with the atom under consideration, additionally,  $l$  is a lower limit on the distance between any two atoms and it is greater than the average covalent bond length.  $C$  is a non-bonded force constant that is assigned between any two atoms that are spatially proximate and around the generally observed covalent bond length. When this computed distance between two atoms is observed to be between  $l$  and  $d$ , a function  $C_x$ , which expresses the force constant as a function of distance between the two atoms is used to compute the value to be assigned. In addition to the above mentioned scheme of assignment which is employed to include non-bonded interactions, bonded atoms are sorted and subsequently, bond specific force constants are assigned. In physical terms, this replicates the actual force fields. We also define  $\bar{\delta}_i(t)$  as a vector of small displacement and the global mass matrix  $M$  for the whole network system such that

$$\bar{x}_i(t) = \bar{x}_i(0) + \bar{\delta}_i(t), \quad (2.4)$$

$$T = \frac{1}{2} \dot{\bar{\delta}}^T M \dot{\bar{\delta}}, \quad (2.5)$$

Where,  $\bar{\delta} = [\bar{\delta}_1^T, \dots, \bar{\delta}_n^T]^T \in R^{3n}$ . If we assume that the deformations are very small,  $V$  becomes a classical quadratic potential energy function. Then Eq. (2) becomes

$$V = \frac{1}{2} \bar{\delta}^T K \bar{\delta}. \quad (2.6)$$

Here  $K$  is the stiffness matrix for the whole network. In the end, the equation of motion for ENM can be simply represented by the following equation

$$M \ddot{\bar{\delta}} + K \bar{\delta} = 0. \quad (2.7)$$

The above equation represents a mathematical model with consistent spring constants proportional to the force constants, resulting in a global stiffness matrix  $K$  and a mass matrix  $M$  which represents the masses of individual constituent atoms, and is solved for computing the eigenvectors and the eigenvalues [14, 15].

## 2.2 Linking Matrix

In order to perform chemical information based ENM, the most important task is to accurately construct the linking matrix. Unlike a  $C_\alpha$  coarse grained based NMA, linking matrix in the current methodology represents not only the connectivity between atoms in coherence with the all atom approach, but it also stores the values of the corresponding force constants. Therefore, we try to incorporate both bonded and non bonded interactions between all the constituent atoms. Hence, our linking matrix is

composed of suitable force constants between covalently bonded atoms within an amino acids and also reasonably approximate force constants that represent inter and intra molecular non bonded interactions. In order to generate the linking matrix, the positional co-ordinates of the constituent atoms and also the sequence of the amino acids is required, Figure 2.1. This data is obtained from the Protein Data Bank (PDB) which records the structural data such as that obtained from X-ray crystallography or NMR.

SEQRES	1 A	691	GLY	ARG	ARG	ARG	SER	VAL	GLN
SEQRES	2 A	691	PRO	GLU	ALA	THR	LYS	CYS	PHE
SEQRES	3 A	691	ARG	LYS	VAL	ARG	GLY	PRO	PRO
SEQRES	4 A	691	ASP	SER	PRO	ILE	GLN	CYS	ILE
SEQRES	5 A	691	ARG	ALA	ASP	ALA	VAL	THR	LEU
SEQRES	6 A	691	GLU	ALA	GLY	LEU	ALA	PRO	TYR
SEQRES	7 A	691	ALA	GLU	VAL	TYR	GLY	THR	GLU
SEQRES	8 A	691	TYR	TYR	ALA	VAL	ALA	VAL	VAL
SEQRES	9 A	691	GLN	LEU	ASN	GLU	LEU	GLN	GLY
SEQRES	10 A	691	GLY	LEU	ARG	ARG	THR	ALA	GLY
SEQRES	11 A	691	THR	LEU	ARG	PRO	PHE	LEU	ASN
SEQRES	12 A	691	PRO	ILE	GLU	ALA	ALA	VAL	ALA
SEQRES	13 A	691	CYS	VAL	PRO	GLY	ALA	ASP	LYS
SEQRES	14 A	691	CYS	ARG	LEU	CYS	ALA	GLY	THR
SEQRES	15 A	691	PHE	SER	SER	GLN	GLU	PRO	TYR
SEQRES	16 A	691	PHE	LYS	CYS	LEU	LYS	ASP	GLY
SEQRES	17 A	691	ILE	ARG	GLU	SER	THR	VAL	PHE
SEQRES	18 A	691	ALA	GLU	ARG	ASP	GLU	TYR	GLU
SEQRES	19 A	691	THR	ARG	LYS	PRO	VAL	ASP	LYS
SEQRES	20 A	691	ALA	ARG	VAL	PRO	SER	HIS	ALA
SEQRES	21 A	691	ASN	GLY	LYS	GLU	ASP	ALA	ILE
SEQRES	22 A	691	ALA	GLN	GLU	LYS	PHE	GLY	LYS

**Figure 2.1:** Represents the sequence of amino acids for Lactoferrin's open form, 1LFH. It has 691 residues and is ordered from left to right and top to bottom in an ascending order.

Subsequently, this information is stored as a data file in MATLAB. The PDB assigns a numbering scheme for atoms in accordance with the sequence of the amino

acids. This data is then sorted for both  $C_{\alpha}$  and Nitrogen atoms in order to create the peptide bond along the backbone. An executable m-file is then generated which uses the information on sequence of amino acids and assigns the bonded force constants between constituent atoms of the Protein. However, while performing this assignment, it is often observed that due to an insufficient resolution or possibly due to the orientation of the crystallized specimen, the PDB can not provide the positional coordinates of all the atoms and this result in some missing atoms and inconsistent numbers for atoms in a given amino acid. As a direct consequence of this, the linking matrix assignment is affected and can be rendered incorrect. For example, in the case of 1LFH, as illustrated in Fig 1.1, Arginine can be observed to notice that there are 11 constituent atoms. During the modeling of Lactoferrin, some residues of Arginine were also observed to have five atoms. This observation prevailed for many other amino acid residues. And so, in addition to following a scheme of assignment of force constants based on just the sequence of amino acids, the number of atoms in a given amino acid was also recorded and utilized at the time for allocating the values to ensure a correct linking matrix. In addition to the bonded interactions between atoms in a molecule, force fields also consist of non-bonded interactions, and as the name implies, these interactions exist between atoms which are not linked by covalent bonds. Hence, these non-bonded interactions are a result of intra molecular and intermolecular forces. Broadly, force fields can be defined to be composed of non-bonded interactions of the following two types: electrostatic interactions and Van der Waals interactions. Within a force field framework, the Van der Waals interaction is considered to consist of the all the interactions between atoms (or

molecules) that are not covered by the electrostatic interaction. Hence, in addition to the bonded interactions, non bonded interactions also need to be accounted for in the linking matrix. In order to do so, the non bonded force constant between any two atoms is expressed as a function of distance between them. An exponential function is employed into use such that. For a distance less than 2Å, a constant value of 6000 dynes/cm is assigned as the force constant. This is done so, since typical covalent bond lengths vary between 1.2Å to 2Å. For a value of the computed distance between 2Å and 8Å, the force constant is computed in accordance with the equation (2.8) mentioned below:

$$F_{nb}=F\_constant*\exp(-(dis-2)) \quad (2.8)$$

Where,

exp: inbuilt operator in MATLAB to compute  $e^{f(x)}$

dis: distance between any two atoms  $X_i$  and  $X_j$

F\_constant= Non bonded force constant.

For a value of distance greater than 8Å, the non bonded force constant was found out to be significantly small, and so, no spring was assigned between the  $i_{th}$  and  $j_{th}$  atoms in this scenario.

### 2.3 Stiffness and Mass matrices

Once the linking matrix has been setup, in accordance with a methodology explained in the previous section, stiffness and mass matrices have to be then setup to perform chemical information based NMA. Using the generated linking matrix the global stiffness matrix K for the system is computed. The mass matrix M consists of an

n x n array, of 3 x 3 symmetric blocks, where n is the total number of atoms in a given protein. The values of masses are aligned along the diagonal, representing the independence about the three principal co-ordinate axes. Once these matrices are obtained, a matrix S is defined and obtained, such that,

$$S = M^{-1/2} \times K \times M^{-1/2} \quad (2.9)$$

The eigenvalues and the eigenvectors of the matrix S are then computed in accordance with the Eq. (2.7). The eigenvector set computed is used to generate the modeshapes, and the corresponding frequencies of these modes are calculated from the eigenvalues.

The wavenumber for a given mode is computed by the equation given below:

$$W_n = \sqrt{\frac{\pi \times d}{2 \times C}} \quad (2.10)$$

Where,

$W_n$  : Wavenumber.

$d$  : Eigenvalue.

$C$ : Speed of light in cm/s.

The eigenvalues' set then computed consists of 3n-6 non-zero values, where the first six eigenvalues correspond to translation and rotation about the three principal co-ordinate axes. Subsequently, animations are generated for small arbitrary perturbations about the given conformation along the corresponding eigenvectors.

## CHAPTER 3

### MODESHAPE ASSIGNMENT AND FREQUENCY DETERMINATION IN LINEAR MOLECULES

#### 3.1 Introduction

Determination of force constants in molecules has been a widely studied subject [22-26]. For instance, various authors have reported that atomic force constants depend on the assignment of the vibration spectrum [27-31]. It is therefore desirable to examine the experimental data for vibration assignment which can be utilized to identify bond specific spring constants that are analogous to force constants in molecules. It is noted that the covalent bonds of molecules are not rigid, but are more like mass less pseudo springs that can be stretched and bent. At ordinary temperatures, these bonds vibrate in a variety of ways, and the vibrational energies of molecules may be assigned to quantum levels in the same manner as are their electronic states. Transitions between vibrational energy states may be induced by absorption of energy. For example, vibration frequency of a diatomic molecule can be calculated by Eq. (3.1) which describes the major factors that influence the stretching frequency of a bond between two atoms of masses  $m_1$  and  $m_2$ , respectively.

$$\nu = \left( \frac{1}{2\pi C} \right) \left[ \sqrt{\frac{f(m_1 + m_2)}{m_1 m_2}} \right] \quad (3.1)$$

Where  $\nu$  is a frequency ( $\text{cm}^{-1}$ ),  $f$  is a force constant, and  $C$  is the speed of light. In the analogy of a mass-spring system, the force constant  $f$  corresponds to the spring's



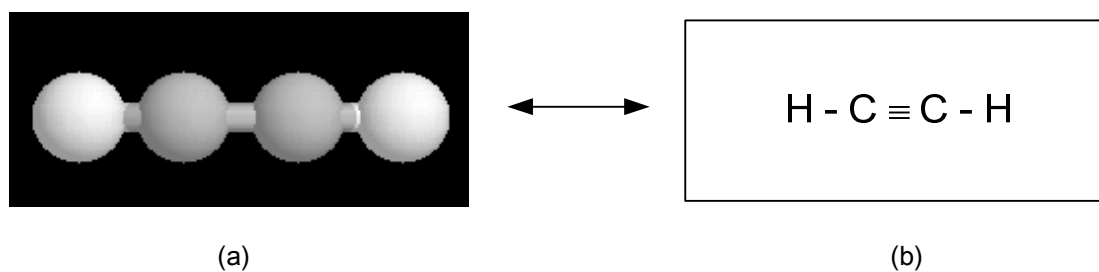
stiffness proportional to the strength of the bond linking  $m_1$  and  $m_2$ . Although such a quantum mechanics enables us to determine precise force constants (or fields) by electronic structure calculation, it is limited to small chemical compounds due to its computational complexity. So it is not feasible to study macromolecular dynamics using this methodology. The relationship between elastic network and internal coordinates such as bond angle and torsion angle was already discussed elsewhere [19]. One can directly assign spring constants obtained from Eq. (3.1) to covalent bonds while other stiffness values of virtual springs should be adjusted by comparison in frequency domain between experimental data and the result of NMA based on ENM. Here we propose a novel method based on an amalgamation of experimental vibrational frequency with the computational approach of NMA which incorporates atomistic details to calculate the virtual spring constants between various atoms. To validate the proposed method, we study four linear molecules; Acetylene, Cyanogen, Ethylene isocyanide and Diacetylene, and construct their ENMs by assigning computed spring constants and the appropriate mass values for atoms, respectively. Then we calculate vibration frequencies and modes from NMA and compare them with experimental data. These computed modeshapes can be assigned to the corresponding frequency spectrum obtained by spectroscopy experiment with which so far we have only detected vibration frequencies as a specific signature identifying only a molecule itself (not its motions).

### 3.2 Methodology

In a general 3-dimensional 3D structure, non-bonded atoms interactions are more predominant due to close proximity of atoms in the spatial domain. Moreover, atoms in the same molecule can occasionally become very close to each other, leading to large values for the non-bonded energy and forces, especially Van der Waals', so special measures are sometimes needed to accommodate this effect. A linear molecule is favorable for an initial analysis using chemical information based NMA as the effects of non-bonded force constants is minimal on modes involving simple stretching. This assumption confers with the theoretical findings since simple stretching can be considered as a strong function of the force constant between the two atoms involved in stretching.

#### 3.2.1 Force constants in Acetylene

The structure in Fig. 3.1 represents a molecule of acetylene showing the four atoms; two hydrogen and two carbons, their orientation, and the bonds between them.



**Figure 3.1:** Schematic of acetylene (a) a ball and stick representation of the ENM model setup using the appropriate values of masses and spatial coordinates: (b) a chemical diagram of acetylene.

The balls represent the atoms with their appropriate mass values, while the sticks represent the bond between atoms with the appropriate values of force constant between them. In order to setup the model elucidated above, the positional coordinates of the constituent atoms were obtained from NIST [32]. Fig. 3.2 represents the  $C\equiv C$  stretching in acetylene. Animation was utilized to represent the perturbed positions of the atoms in the given mode. With an approach identical to one mentioned above, animations were generated for both C-H symmetric as well as asymmetric stretching. Thereby, these were utilized to correlate mode numbers to their modeshapes and wave numbers.

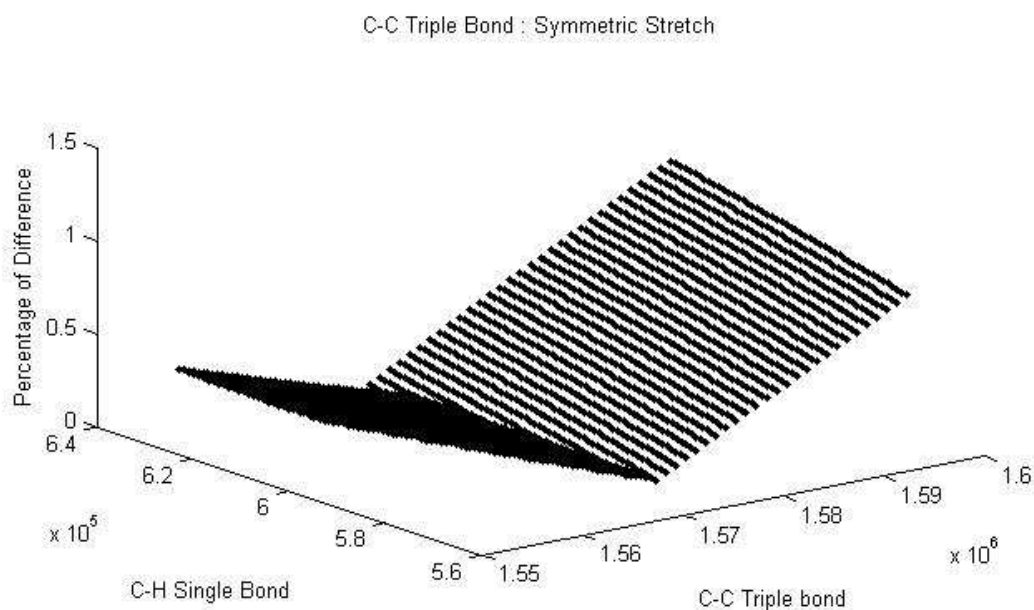


**Figure 3.2:** Animation of  $C\equiv C$  stretching mode in acetylene. Weighted eigenvectors are added to its original coordinates to visualize compression and extension of  $C \equiv C$  bond in (a) and (b), respectively.

### 3.2.2 Optimization of computed force constants

Since, wavenumbers are a function of the force constants and mass, and also, that the mass values were constant, the force constants' values were optimized in order to compute wavenumbers in closest agreement to the literature as possible. In the case of acetylene, it can be observed that there are two types of bond as represented in Fig. 3.1(b). Hence, a 3D plot of difference in a wavenumber against the combination of the two spring constants utilized for computing it can be generated, where difference is

defined as a percentage difference between the computed and the experimental wavenumbers. Fig. 3.3 shows an example plot corresponding to  $C\equiv C$  stretching mode in which we can determine the final spring constants based on a minimum value of percentage difference in the wavenumber. Two more similar plots were generated for C-H symmetric and asymmetric stretching modes, respectively (not displayed here). Based on the inference from these plots, the values for force constants in acetylene were computed and tabulated in Table 3.1. An identical methodology was then employed into use for Cyanogen ( $N\equiv C-C\equiv N$ ). Its results are also summarized in the Table 3.1.



**Figure 3.3:** Optimization of spring constants. The variation of percentage of difference between the computed and the experimental wavenumbers is represented along the Z axis with respect to the two bonds (i.e. spring constants of  $C\equiv C$  and  $C-H$  bonds along the X and Y axes, respectively).

**Table 3.1:** A list of bond specific force constants. Comparison between experimental and predicted frequencies yields specific force constants of each pair of atoms. The obtained force constants can be assigned to appropriate spring constants in ENM.

Molecule	Bond	Experimental Frequency (1/cm)	Predicted Frequency (1/cm)	Difference (%)	Computed Force Constants ( $10^5$ dynes/cm)
Acetylene	$C \equiv C$	1974	1974	0.00	15.71
	C - H	3289(Asym.)	3290	0.03	5.94
	C - H	3374(Sym.)	3370	0.12	5.94
Cyanogen	C - C	846	846	0.00	6.99
	$C \equiv N$	2150(Asym.)	2150	0.00	17.64
	$C \equiv N$	2330(Sym.)	2424	4.03	17.64

### 3.3 Results and discussions

As mentioned in the previous section, in our approach, we examine two chemical compounds; Acetylene and Cyanogen, to first determine the force constants between their constituent atoms. The vibration assignments for different bonds' stretching were obtained from NIST. In accordance with the methodology elaborated in this paper, our ENM based NMA was then performed to compute the vibration frequency of the various stretching modes in both the molecules and animations of the corresponding modeshapes were generated using RASMOL. Spring constants in ENM can be considered to be analogous to the bond force constants and these values for different bonds represented in Fig. 3.1 were adjusted till a reasonably accurate eigenvalue set was obtained. The predicted frequencies and the computed force constants for the constituent bonds have been summarized in the Table 3.1. The values in Table 3.1 for the computed force constants have been recorded and utilized for setting

up the ENM model for Ethynyl isocyanide (C<sub>3</sub>HN) and Diacetylene (C<sub>4</sub>H<sub>2</sub>) and subsequently, NMA has been performed. The positional coordinates of the constituent atoms were also determined from the NIST webpage. The predicted frequency from NMA along with the experimental data for vibration assignment has been summarized in the Table 3.2. It suggests that the computed force constants give a good prediction of the frequencies for various stretching modes in Diacetylene and Ethynyl isocyanide. Various authors have reported a substantially vast range of force constant values in Diacetylene and Ethynyl isocyanide/Cyanoacetylene. Wu and Shen [33] determined the value of

**Table 3.2:** Comparison between experimental and predicted frequencies. Vibration modes in Ethynyl isocyanide and Diacetylene are predicted, respectively, using the computed bond specific force constants listed in Table I and then they are compared with nominal experimental values obtained experimentally.

Molecule	Bond	Experimental Frequency (1/cm)	Predicted Frequency (1/cm)	Difference (%)
N ≡ C - C ≡ C - H (Ethynyl isocyanide)	C - H	3327	3333	0.19
	C ≡ C	2079	2085	0.23
	C - C	864	850	1.58
	C ≡ N	2274	2381	4.74
H - C ≡ C - C ≡ C - H (Diacetylene)	C - C	2020	2038	0.89
	C ≡ C	2184	2318	6.14
	C - H	3293	3319	0.79
	C - H	3329	3335	0.18

stretching force constant for carbon-carbon single bond to be  $3.58 \times 10^5$  dynes/cm. Similar low values have been reported by Meister [34,35] (C-C force constant:  $3.234 \times 10^5$  dynes/cm), Herzberg [36], and Kovner [37]. A range of values for the C-C bond have been reported, spanning from  $4.5 \times 10^5$  dynes/cm to  $6.7 \times 10^5$  dynes/cm

depending up on the compound containing the bond. Moreover, Jones [38] also pointed out the values for  $C\equiv C$ ,  $C-H$ ,  $C-C$  to be  $15.12\times 10^5$  dynes/cm,  $5.82\times 10^5$  dynes/cm and  $7.14\times 10^5$  dynes/cm, respectively, which are in close agreement with the values obtained from NMA, but unlike our hypothesis, the author postulates that these force constants are specific for Diacetylene. Similarly, Turrell [39] reported a force constant of  $7.83\times 10^5$  dynes/cm for  $C-C$  stretching in Diacetylene based on an observed (Raman active) symmetrical stretching of  $874\text{ cm}^{-1}$  reported by Jones. NMA computes similar force constants to what have been mentioned in the literature, but unlike some other contemporary studies, it can be stated, that for a particular bond, the value of the corresponding force constant can be unique and any sort of invariance can be attributed to the surrounding atoms and the interactions with them. The closeness of the computed frequencies to those obtained from experimental data reaffirms this finding.

### 3.4 Conclusions

In nature, the force field consists of both bonded and non-bonded interactions. However for a linear molecule, results obtained suggest that stretching modes be predominantly a strong function of bonded interactions and non-bonded interactions have minimal effects on both symmetric as well as asymmetric stretching vibration modes. This enables us to perform NMA without considering these effects, and the observed outcome does supplement our hypothesis. As a result, for linear molecules, the force constant between any two atoms can be calculated and represented by a unique number such that it yields the same value of frequency for a mode as that from

experiment in any other given linear molecule. This methodology can be employed as a new vibration assignment scheme. Since the modeshapes computed from ENM based NMA can also be animated, it will enable us to not only compute vibration spectrum of a given macromolecule but also visualize the corresponding modeshapes. In the near future, we will apply this scheme to determine precise spring constant values of various bonds engaged in protein structures resulting in identification or prediction of vibration frequencies and modeshapes, which is one of the most important topics in computational structural biology to elucidate biological function of a protein from its structural information.



## CHAPTER 4

### APPLICATION OF CHEMICAL INFORMATION BASED NMA TO AMINO ACIDS

#### 4.1 Introduction

Cysteine has been a widely studied amino acid due to the availability of substantial crystallographic data, computationally generated low energy metastable states elucidating the structural details of its numerous possible conformations [40-47], and the relative simplicity of the orientation of the constituent atoms. Disparate studies discussing the vibration spectrum assignment in Cysteine have been conducted and reported by several authors [48-51]. These efforts have been necessitated by the fact that vibration assignment is predominantly governed by the structure of a given amino acid, which in turn is dependent on the orientation of the constituent molecules. As mentioned above, in nature, like other amino acids, Cysteine has also been found to exist with different possible conformations, assignment of a unique vibration spectrum to a generalized coordinate set could be considered to be a rather naive approach. On the other hand, with the structural complexities of the examined structures and the realization of a huge set of possible solutions coupled with the limited structural information from experimental techniques such as X-ray crystallography, accurate vibration assignment almost seems like a daunting task. Some authors have also reported that a detailed study of the constituent force fields and internal coordinates in a molecule to substantiate a suitable representation of the intramolecular and intermolecular (e.g. in

case of dimers) forces can be utilized to compute accurate vibration frequencies. They can then be used subsequently for modeling of such and more macromolecules and other complex biological systems and simulations can be run for obtaining the results used for vibration spectrum assignment [52-56]. The low-frequency vibrations are typically dominated by non-covalent, intermolecular interactions such as electrostatic, Van der Waals, and hydrogen bonds [57]. These lower modes have also been observed to be more global in nature, even in the case of a single amino acid residue like in large macromolecules, involving motions between non-bonded constituent atoms that can be characterized to be more complex, arising from intramolecular interactions. In contrast, higher modes involve more of individual bond stretching between a pair or pairs of atoms, which therefore depends on bonded force constants which is a characteristic of the covalently bonded atoms [58]. Using this approach of identifying and distinguishing bonded and non-bonded interactions, to facilitate a suitable and approximate replication of the naturally existing force fields, we examine Cysteine to simulate its vibration spectra which yields us with the information to generate animations of an ordered set of modeshapes and their corresponding frequencies. In order to do so, we use the method of Chemical information based Elastic Network Model (ENM) to perform the required simulations for Normal Mode Analysis (NMA).

## 4.2 Methodology

In accordance with the present methodology, the positional coordinates of Cysteine listed in Table 4.1 were obtained from BMRB [59] for determining the initial

conformation and setting up of the ENM [32]. Then Cysteine was modeled to be consisting of point masses, values of which were consistent with the atomic masses of the constituent atoms.

**Table 4.1:** Cartesian coordinates of a nominal Cysteine structure used for running the simulations.

Atom Number	Atom	Positional Co-ordinates in Å		
		X	Y	Z
1	N	1.559	-0.060	-0.596
2	C	0.088	-0.071	-0.544
3	C	-0.402	1.298	-0.579
4	O	-0.132	2.024	-1.528
5	C	-0.383	-0.930	0.661
6	S	0.184	-2.644	0.606
7	O	-1.099	1.790	0.302
8	H	1.841	0.464	-1.427
9	H	1.896	0.457	0.219
10	H	-0.256	-0.563	-1.455
11	H	-1.474	-0.964	0.694
12	H	-0.030	-0.497	1.600
13	H	-0.397	-2.925	1.803
14	H	-1.385	2.621	0.243

These point masses were connected to each other by massless springs, representing the interactions between different atomic pairs, with values of stiffness analogous to the force constant between any two given atoms. Although a generalized force constant of  $7 \times 10^5$  dynes/cm was primarily used for representing the bonded interaction, the force constants for these bonds were computed by other smaller molecules (see Table 3.1) which consisted of the bonds present in Cysteine. The use of a generalized force

constant was prompted by the observed results for Cysteine. The output for wavenumbers for Cysteine with two different inputs for force constants for bonded atoms being precise (as that obtained by performing NMA on linear molecules) and a generalized value were identical. It could be inferred from such a result that the absolute values of wavenumbers were dependent on the ratio of the force constants of bonded and non bonded atoms. Hence, as long as the model incorporated a certain order of difference, the wavenumbers were not altered.

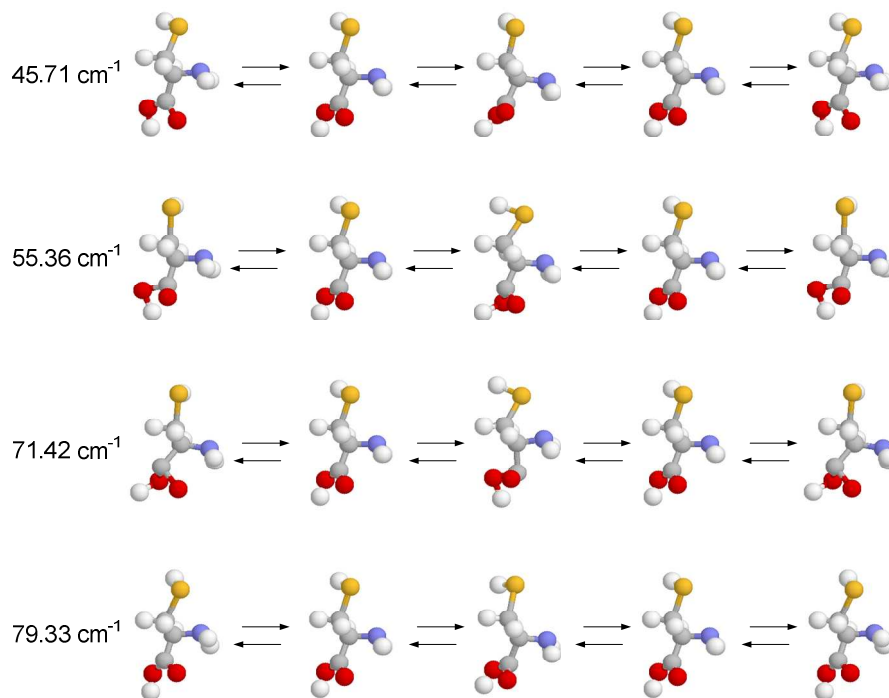
With regard to more complex structures of molecules, in the scope of the undertaken modeling scheme and force field parameterization, the effect of non-bonded interactions are more dominant than what is observed for linear molecules due to the orientation and structural conformation of most biomolecules, like that for Cysteine in this case, and hence, a distance cutoff scheme is adopted to replicate the non-bonded interactions.

### 4.3 Results

**Table 4.2:** Comparison between experimental and predicted frequencies for L-Cysteine. A generalized force constant of  $7 \times 10^5$  dynes/cm, a non-bonded force constant of  $6 \times 10^3$  dynes/cm, and a cutoff distance of  $8\text{\AA}$  were applied when computing vibration frequencies

Molecule	Experimental Frequency (1/cm)	Predicted Frequency (1/cm)
L- Cysteine	46	46.79
	56	55.53
	71	71.64
	80	79.81

As can be seen from the results in Table 4.2, the values for frequencies predicted from NMA using chemical information based ENM are almost identical to the ones reported in the literature. Also, Fig. 4.1 illustrates the animations that were generated for the first four modes. They are identical to the modeshapes obtained from literature.

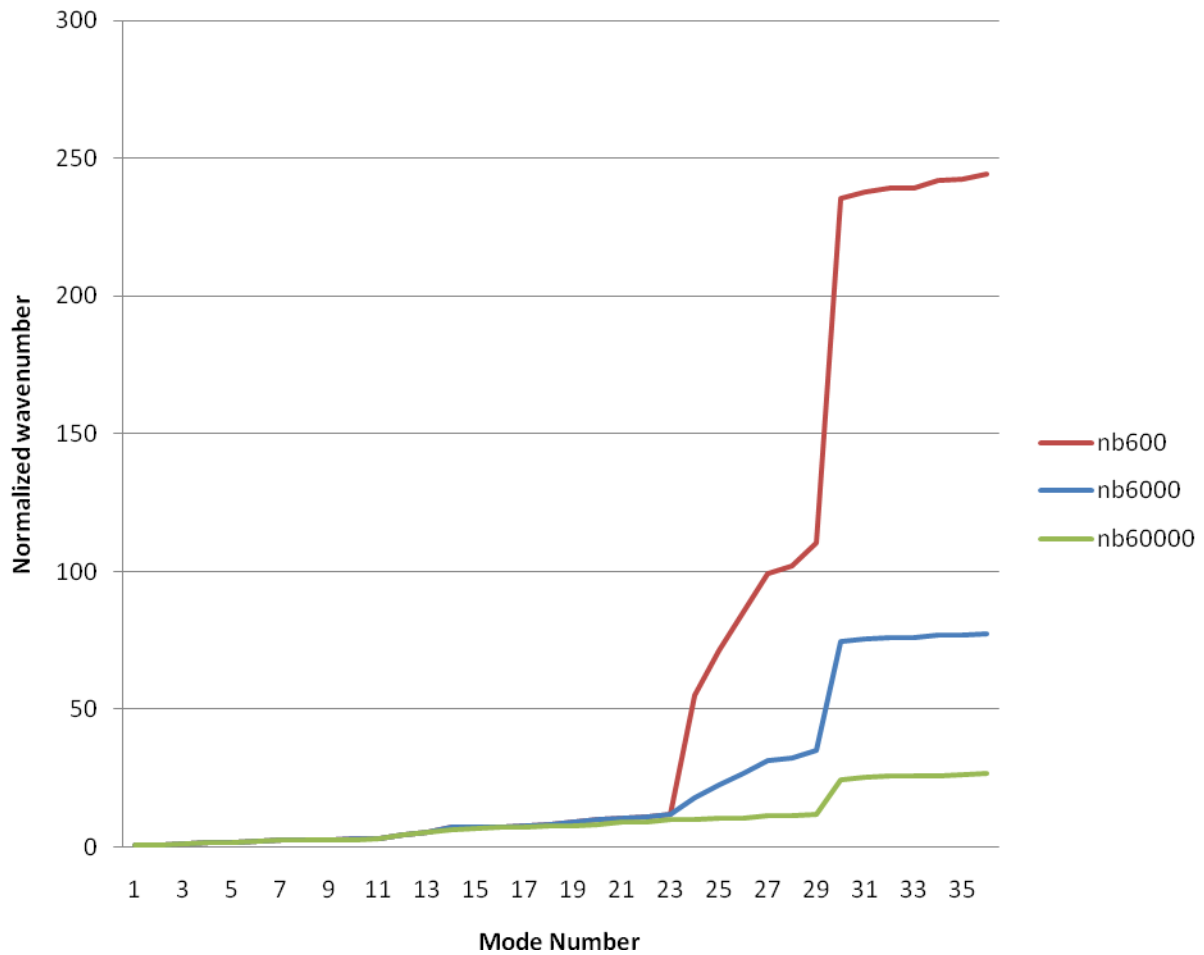


**Figure 4.1:** Modeshape animations of Cysteine with their corresponding frequencies for the first four computed modes.

Hence, in accordance with the hypothesis put forward in the previous section, it can indeed be observed from the results summarized in Table 4.3 and Fig. 4.2 that the lower frequencies, more global modes, are strongly dependent on the connectivity of the model which can be determined with the scheme of force fields used for experiments.

**Table 4.3:** Computed wavenumbers based on both generalized and bond specific force constants in addition to the three different values of non-bonded force constants referred to as cases; **A:**  $6 \times 10^2$  dynes/cm, **B:**  $6 \times 10^3$  dynes/cm, **C:**  $6 \times 10^4$  dynes/cm. The reported values of wavenumbers are in  $\text{cm}^{-1}$ .

Mode number	Generalized force fields			Bond specific force fields		
	Non-Bonded force constants			Non-bonded force constants		
	A	B	C	A	B	C
1	14.802394	<b>46.799885</b>	147.712688	14.802314	<b>46.797360</b>	147.638190
2	17.566925	<b>55.533519</b>	175.077050	17.566746	<b>55.527973</b>	174.926561
3	22.664127	<b>71.646103</b>	225.852934	22.663856	<b>71.637691</b>	225.620437
4	25.254733	<b>79.811407</b>	250.871235	25.253977	<b>79.787510</b>	250.114341
5	29.285772	<b>92.601344</b>	292.587999	29.285743	<b>92.600420</b>	292.560266
6	34.182133	107.936899	337.075490	34.178455	107.823686	334.211115
7	41.013705	129.609085	406.864171	41.013091	129.588954	405.977264
8	42.532467	134.394182	421.819645	42.531559	134.365878	421.042915
9	43.903574	138.607706	431.865613	43.901880	138.554803	430.496069
10	46.375018	146.169890	449.902193	46.370693	146.039695	447.236901
11	52.427631	165.222906	507.203778	52.423861	165.107598	504.330310
12	71.587008	226.081154	705.826376	71.584706	226.008734	703.592034
13	86.121486	271.888231	846.031591	86.118432	271.791872	809.447766
14	108.636961	343.009237	943.108720	108.628612	342.706927	845.274628
15	109.996546	346.324759	1052.229650	109.990271	346.071547	938.774141
16	112.031108	353.376916	1074.452445	112.007725	352.610718	1048.230581
17	119.409673	376.420552	1082.831708	119.403099	376.202982	1066.804573
18	124.608217	393.786443	1162.356936	124.604080	393.655277	1089.122709
19	140.516103	443.777081	1203.040409	140.506756	443.442714	1179.282466
20	152.250916	480.756301	1235.868040	152.240044	480.341573	1206.614674
21	157.390350	496.990751	1354.495399	157.379099	496.605723	1242.218098
22	167.845776	530.116714	1395.683648	167.838817	529.862081	1384.002953
23	178.168274	562.210643	1502.630376	178.156303	561.683791	1463.387947
24	821.788961	841.273981	1509.137248	709.677795	724.998621	1489.469310
25	1060.668824	1071.857468	1554.712950	831.855189	848.704285	1545.709602
26	1262.838116	1273.631650	1567.580000	1085.017298	1098.451706	1604.310588
27	1469.864604	1473.664242	1683.451411	1382.885984	1391.363310	1698.371076
28	1512.725668	1520.153586	1724.991412	1553.175553	1560.397049	1768.002661
29	1639.036406	1644.975163	1787.334597	1960.770002	1963.795820	1996.252085
30	3485.078103	3495.689946	3613.038798	2263.474405	2280.148660	2495.856523
31	3521.269155	3552.057489	3757.088950	3096.413436	3131.749653	3352.327592
32	3537.214699	3565.454913	3804.375740	3155.340705	3172.107638	3481.567577
33	3541.403251	3570.520791	3807.817929	3258.678771	3294.682605	3536.424115
34	3578.319565	3603.386225	3850.755895	3297.099562	3332.880148	3629.611845
35	3588.716823	3611.544141	3884.923918	3331.132151	3349.634119	3705.081566
36	3615.824138	3632.857618	3963.330542	3722.849729	3745.789734	3977.608170



**Figure 4.2:** A plot of normalized wavenumbers versus the mode number for the case of generalized force fields which elucidates the observed divergence at higher modes resulting from a variation in the non-bonded force constants.

Variation on non-bonded force constants under the same connectivity does not affect too much the low mode frequencies (not absolute values but normalized ones) and their modeshapes. One can also recognize that there is significant variation on higher modes due to the simplification of force constants for various covalent bonds as a generalized force constant. More precise force fields would be required to accurately determine the frequencies for higher modes with predominantly local vibrations.

Nonetheless, the observed unique mode number can be utilized to provide us with a sequentially arranged modeshapes. Namely, although the values of frequencies for certain higher modes might not be precise yet, the information on sequence of animations is very useful to give an ordered set of modeshapes assorted in ascending values of their corresponding frequencies. In other words, the model in itself can provide the user with a unique mode number below which swapping of modes cannot take place, for the given force fields. Animations therefore can be suitably utilized for mode shape assignment.

## **4.4 Discussions**

### **4.4.1 Force field parameterization**

The robustness of the fundamental behavior of the modeling scheme adopted to generate the reported results has to be ascertained to safely assume repeatability and consistency in the outcome. In order to do so, a sensitivity analysis is mandatory to understand the effects of variations in input parameters on the output. There are primarily three input variables; bonded and the non-bonded force constants as well as the connectivity of the constituent atoms. As mentioned in the previous section, the bonded force constants were determined by performing NMA on linear molecules, so as to evaluate the values for the same by comparing the output for bond stretching from that observed from the reported values in the literature. For instance, force constants for Acetylene and Cyanogen have been reported in the Table 3.1. Similar methodology was adopted to compute the force constants of numerous other pair of atoms (not displayed



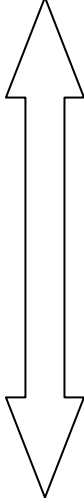

here). Hence, the approach adopted during this work was to run simulations with different combinations of these parameters (i.e. to optimize the outcome based on the proposed hypothesis of using a generalized set of force fields resulting in Table 4.2). Firstly, ratios of bonded and non-bonded force constants were varied by orders of difference. Subsequently, eigenvector set for the perturbed conditions were compared and animations for modeshapes were generated to observe any discrepancies. With regard to the same, normalized wavenumbers were plotted for different test conditions. Normalized wavenumbers in the scope of this study have been defined as the ratio of frequencies of modes to the slowest mode. This suggests that, while the first mode is considered as unity, all the subsequent modes can then be expressed in terms of the first mode. Similarly, the value of the limiting distance utilized to set up the connectivity matrix, which replicates the interacting pairs of atoms was also varied to further examine the dependence of the model's behavior. Values for vibration frequencies for Cysteine with different ratios of bonded and non-bonded force constants have been summarized in the Table 4.3. These values were obtained from both, a generalized force constant of  $7 \times 10^5$  dynes/cm, as well as bond specific force constants which were obtained by performing NMA on smaller molecules. These were used in combination with different non-bonded force constants of  $6 \times 10^2$  dynes/cm,  $6 \times 10^3$  dynes/cm and  $6 \times 10^5$  dynes/cm. Several researchers have developed and postulated different techniques of force field determination for such analyses [60, 61]. In addition to the above reported values, Fig. 4.2 illustrates the plot of normalized wavenumbers versus the modenumbers.

Though the absolute values of normalized wavenumbers are of little significance since they are merely ratios of all the mode frequencies with respect to the first, slowest mode, yet, the plots generated as illustrated in Fig. 4.2 provide insight in to model's behavior. Such an analysis was required to substantiate the effect of variations in non-bonded force constants as well as to establish the correctness of the scheme adopted. Moreover, animations suggest that these induced variations in the input parameters do not alter the modeshape up to a certain critical mode number of 23 as that suggested by the plot in the Fig. 4.2. This observed characteristic of the model can be attributed to the variations induced by the ratio of bonded to the non-bonded force constants, as a result of which, only the absolute values of wavenumbers were rendered different. This analysis suitably explained the effect on output parameters by such intended variations in the inputted values, and facilitated the use of a generalized force field parameterization scheme in addition to the specified non-bonded force constant of an exponentially decreasing function, and yielded results which were analogous to experimentally found values.

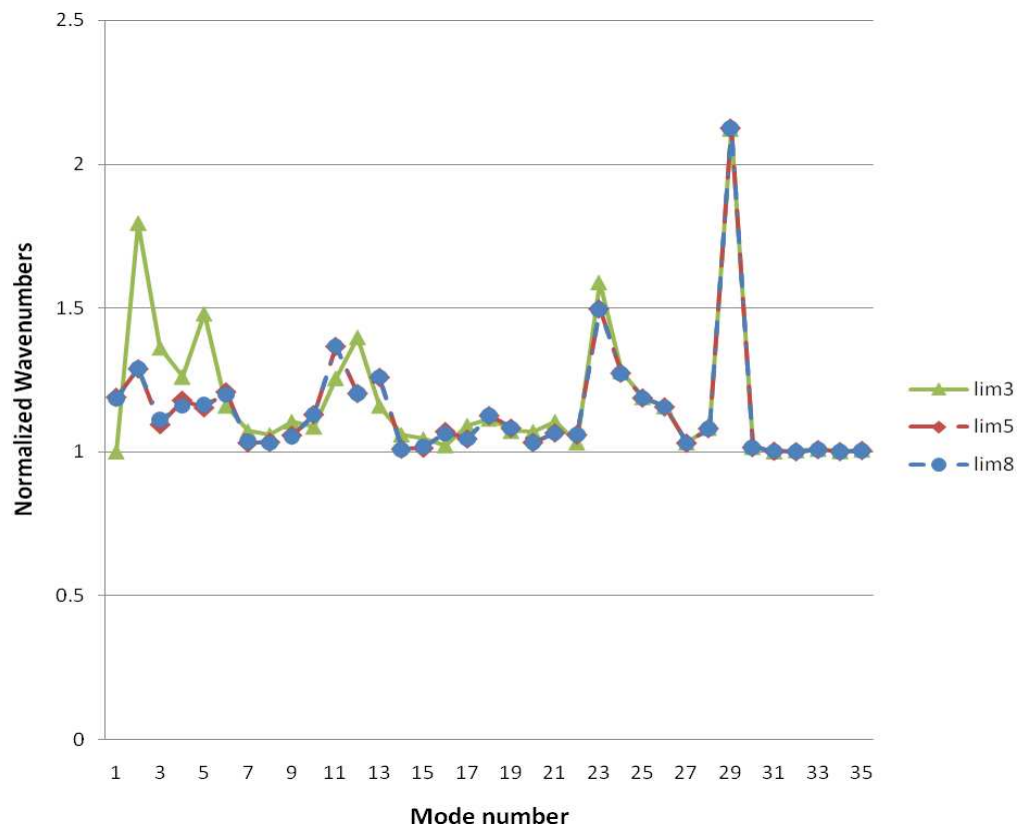
#### **4.4.2. Sensitivity to the cutoff distance**

As mentioned in the previous section, in addition to understanding the effect of force field parameterization used, it was imperative to identify and replicate the non-bonded pairs of atoms. In order to do so, a certain cutoff distance was used to determine the contacting pairs, and subsequently represent this in the linking matrix. The results obtained from such an analysis have been summarized in the Table 4.4.

**Table 4.4:** Computed frequencies of Cysteine based on different values of cutoff distance. Underestimation of non-bonded interactions with relatively short cutoff distance results in unrealistic lower frequencies while higher modes are less sensitive to non-bonded interactions.

Mode number	Cutoff			
	3Å	5Å	8Å	
1	0.149826	46.621149	46.799885	<b>Sensitive to Cutoff</b> 
2	19.922917	55.506206	55.533519	
3	35.772108	71.562895	71.646103	
4	48.707778	78.421667	79.811407	
5	61.383380	92.543589	92.601344	
6	90.858375	106.800687	107.936899	
7	105.373115	129.133670	129.609085	
8	112.993592	133.326458	134.394182	
9	119.609614	137.933739	138.607706	
10	132.189067	146.030844	146.169890	
11	143.687886	165.103283	165.222906	
12	180.427150	225.731958	226.081154	
13	252.291337	271.762231	271.888231	
14	292.514752	342.372586	343.009237	
15	309.831336	346.101460	346.324759	
16	323.663682	350.949408	353.376916	
17	330.835542	376.088296	376.420552	
18	360.953140	393.514943	393.786443	
19	402.323910	443.527899	443.777081	
20	431.092529	480.487184	480.756301	
21	461.399415	496.798032	496.990751	
22	509.867902	529.974083	530.116714	
23	525.997832	561.880237	562.210643	
24	<b>835.948406</b>	<b>841.195905</b>	<b>841.273981</b>	<b>Insensitive to Cutoff</b> 
25	<b>1070.192281</b>	<b>1071.850569</b>	<b>1071.857468</b>	
26	<b>1272.792792</b>	<b>1273.611250</b>	<b>1273.631650</b>	
27	<b>1472.984176</b>	<b>1473.639862</b>	<b>1473.664242</b>	
28	<b>1519.518035</b>	<b>1520.148893</b>	<b>1520.153586</b>	
29	<b>1644.223534</b>	<b>1644.973737</b>	<b>1644.975163</b>	
30	<b>3490.111614</b>	<b>3495.089660</b>	<b>3495.689946</b>	
31	<b>3546.749467</b>	<b>3551.817114</b>	<b>3552.057489</b>	
32	<b>3549.325707</b>	<b>3564.792005</b>	<b>3565.454913</b>	
33	<b>3565.681246</b>	<b>3570.519510</b>	<b>3570.520791</b>	
34	<b>3598.229896</b>	<b>3603.125237</b>	<b>3603.386225</b>	
35	<b>3604.031512</b>	<b>3611.543399</b>	<b>3611.544141</b>	
36	<b>3628.020041</b>	<b>3632.857604</b>	<b>3632.857618</b>	

In addition to computing the wavenumbers, animations for the corresponding modeshapes were generated to understand the effects of a certain cutoff value on the model's behavior. While the wavenumbers for cutoff values of 8Å and 5Å are almost identical, these values are different for a cutoff value of 3Å. This observed finding can be attributed to the fact that global motions are dependent on non-bonded acting pairs, and in turn, on the linking matrix which represents the connectivity.



**Figure 4.3:** Plot of the ratio between two consecutive wavenumbers  $\omega_{i+1}/\omega_i$  against mode number representing the variance of wavenumbers in terms of cutoff distance. Since the first non-zero mode with a cutoff of 3Å is unrealistically small, the computed value of  $\omega_2/\omega_1$  was discarded and the arbitrary number of 1 is instead chosen to match scales.

As for the higher modes, the values tend to converge shown in both Table 4.5 and Fig. 4.3 as these vibrations can be characterized as more local vibrations pertaining to bond stretching and bending. Hence, this result is much in agreement with the proposed hypothesis. As discussed; higher modes which consist of more local vibrations are less dependent to non-bonded intra-atomic interactions. Hence, as expected, the sensitivity analysis with different values of cutoff distance yields disparate linking matrices so that the modeshapes can be also different for the lower modes while subsequent animations suggest that the higher modes be identical.

#### 4.5 Conclusions

Our implementation of a chemical information based ENM approach for performing NMA simulations has been applied to Cysteine with results that are comparable to that observed from reported values obtained from terahertz spectroscopy [62]. The results obtained for both, animations as well as the corresponding frequencies suggest a favorable trend, and the proposed hypothesis is supported by the precise determination of frequencies and modeshapes for the first few slower modes. Though other researchers have also predicted and reported values for vibration frequencies of Cysteine through experimental as well as theoretical approaches, they all seem to exhibit a common trend of lack of information for frequencies lower than  $100\text{cm}^{-1}$  and thereby a poor resolution in the slow frequency domain [63-65]. Moreover, for Cysteine, it has been observed that the computed frequencies are indeed fundamental modes, and in fact, are not characterized by lattice vibrations as proposed by some authors [62]. A deviation

in these results for higher modes was an expected outcome due to the use of approximate force fields. Hence, because of the excellent resolution that NMA provides to compute results comparable with that obtained from Terahertz spectroscopy, it is imperative to analyze the behavior of such a modeling scheme with more precise force fields. While a more accurate force field parameterization, accounting for both bonded and non-bonded interactions will facilitate a better prediction of vibrational frequencies in amino acids in a manner elucidated for Cysteine, the statistical data for macromolecules suggest that a majority of biological functions are observed in the low frequency domain and can be associated with a conformational change. This would result in a more appropriate connectivity matrix, and as the results suggest, the connectivity is a crucial factor in the determination of low modes. Understanding these parameters is a worthy effort for exploiting the possibility of implementation of such an analysis.

**CHAPTER 5**  
**ANALYSIS OF MACROMOLECULES USING CHEMICAL INFORMATION**  
**BASED NMA**

**5.1 Introduction**

Lactoferrin has been a widely studied Protein. Due to ability to bind iron, and its natural anti-bacterial, anti-fungal and anti-viral properties render it useful for a number of product applications [66-70]. Numerous studies have been conducted for vibration spectrum assignment in Lactoferrin [71-73]. Advanced research focusing on spectroscopy methods for enhanced signals has been conducted on human and animal Lactoferrin to better comprehend its biological functions for varied applications [74-76]. The current methods of vibration spectrum assignment like Raman and Infrared spectroscopy lack significant information on frequencies lower than a  $100\text{ cm}^{-1}$ . On the other hand, there is little data available on its vibration frequencies and techniques facilitating the visualization of its numerous important modeshapes. In the current scope of its study, most widely used computational techniques, like Normal Mode Analysis; models with reduced degrees of freedom, in addition to simplistic approach of generating an Elastic Network Model are used. It has been statistically been observed that most of the important biological functions are associated with more global, local modes. Animations from current NMA methodology do not provide the users with an ordered set of modes and their corresponding frequencies, but still suffice in assisting biologists in associating these modeshapes with some crucial biological functions.

Results obtained from analysis of linear molecules as well as simple amino acids prompt a possible applicability of all atom NMA for large macromolecules, like Lactoferrin that has been used in this study. In a manner similar to that elucidated for the case of simple linear molecules as well as Cysteine, simulations were also performed for Lactoferrin's two forms; 1LFH and 1LFG. While animations from such an analysis show a one to one correspondence with modeshapes obtained from conventional harmonic analysis, it is observed that a model with reduced degrees of freedom often yields in swapping of modes. This observed effect is overcome by undertaking the mentioned approach of an all atom Normal Mode Analysis.

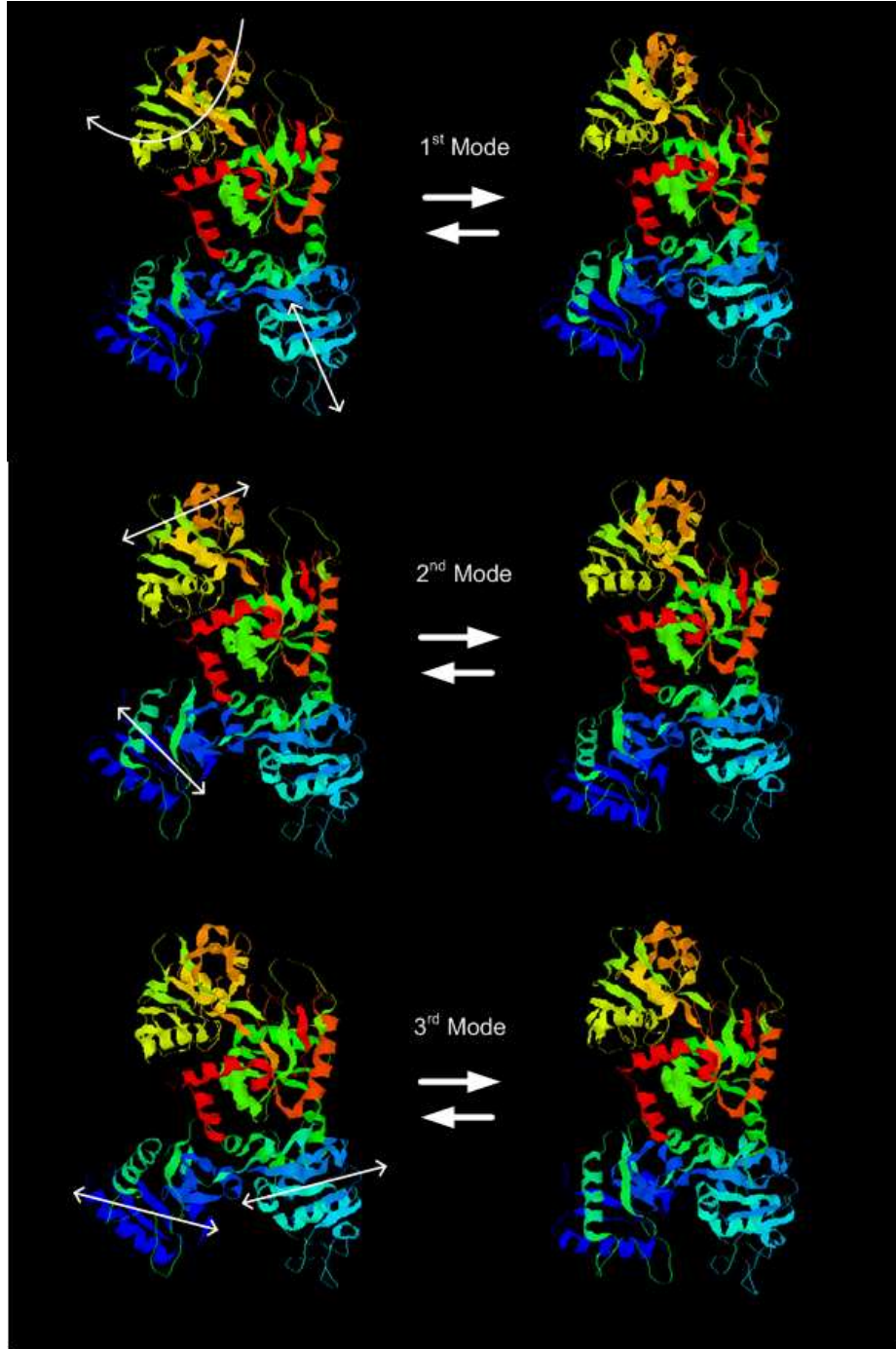
## 5.2 Methodology

Fundamentally, setting up the model for performing chemical information based NMA for Lactoferrin is identical to the approach adopted for previous, simpler cases. Due to the relative complexity of Lactoferrin, an automation scheme has to be incorporated. As discussed in chapter 2, the key requirements for the undertaken modeling scheme involve setting up the mass matrix as well as a linking matrix.

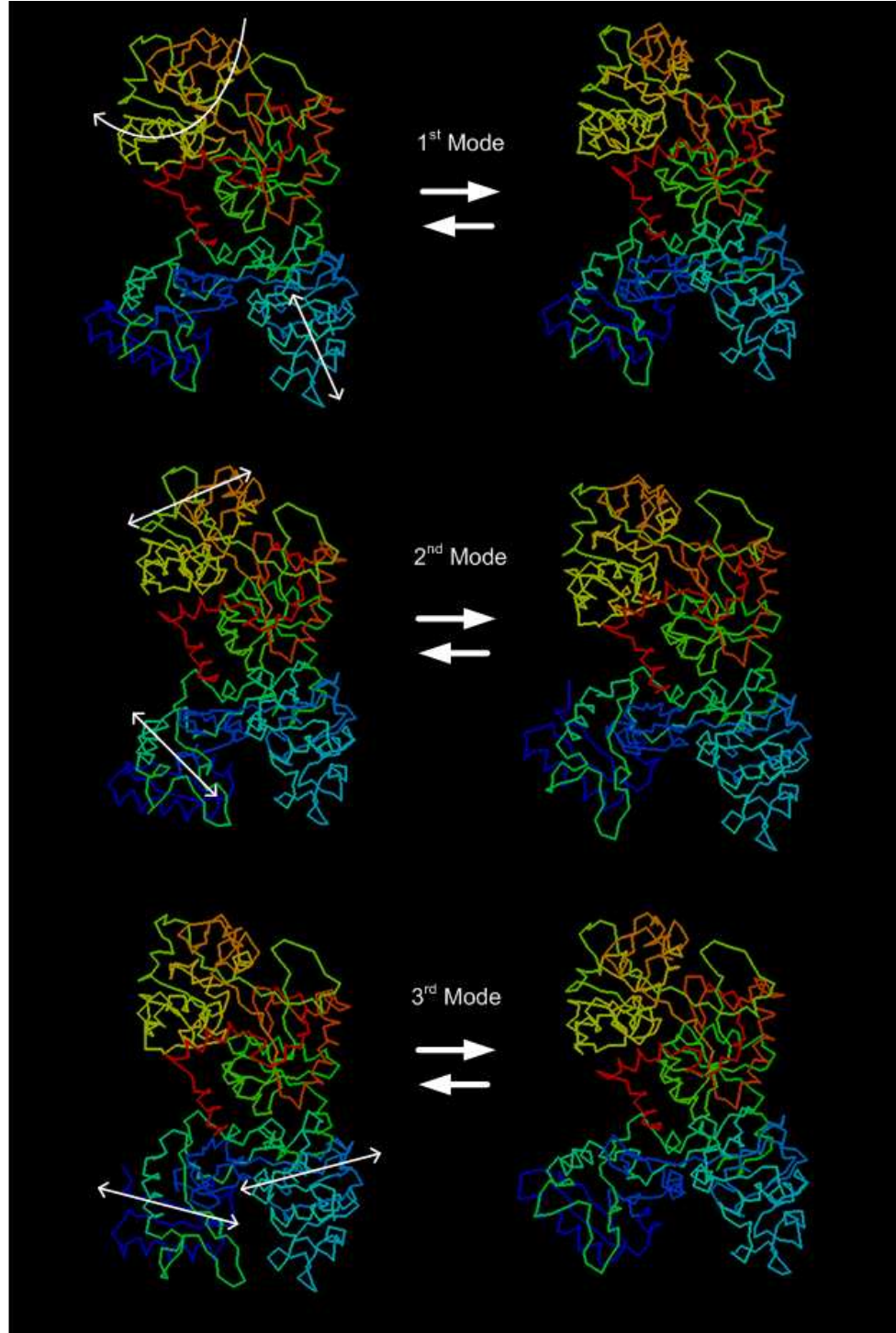
## 5.3 Results and Discussions

Figure 5.1 illustrates the modeshape of the first, second and third modes of the open form of Lactoferrin (1LFH) obtained from the animations subsequent to the NMA simulations. Figure 5.2 represents the animations generated using a  $C_\alpha$  NMA. Results from both suggest a one to one correspondence.



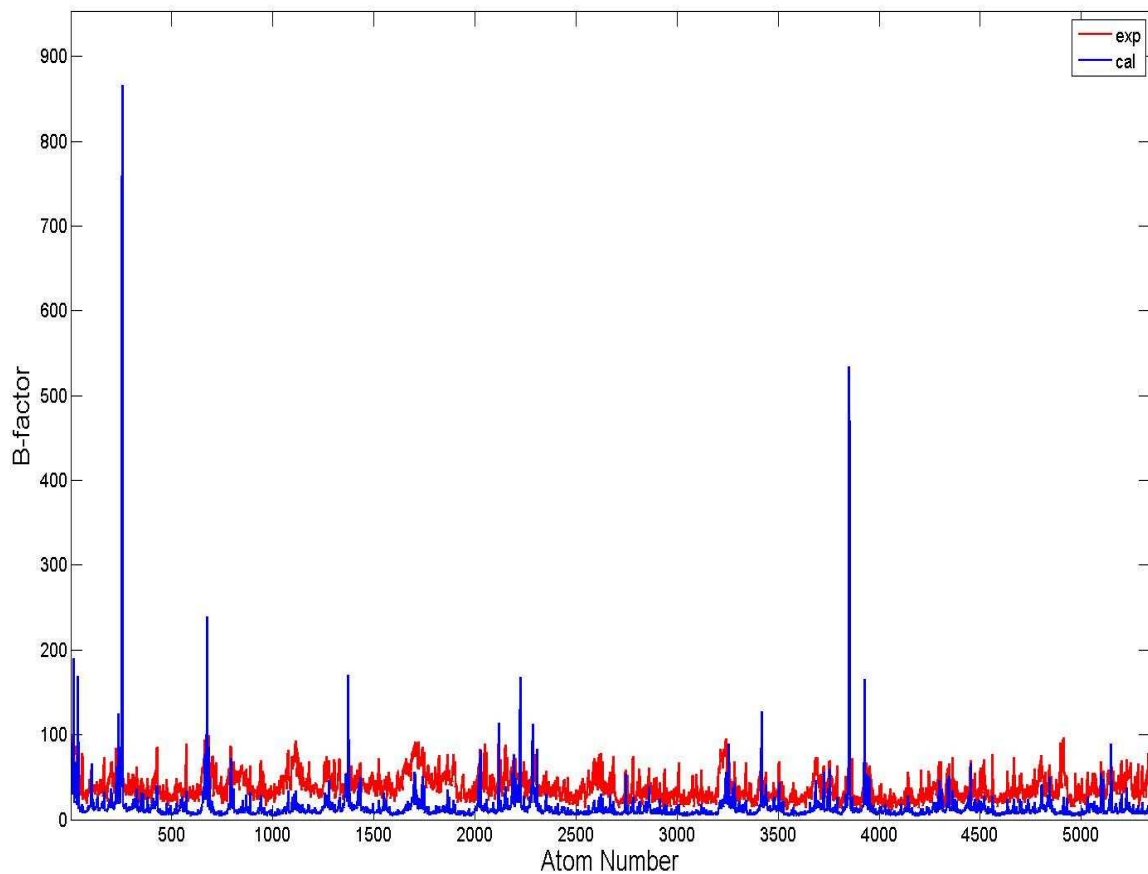


**Figure 5.1:** First three modes for 1LFH obtained from simulations using chemical information based NMA. They all suggest the global motion of what is commonly referred to as the head and the two lobes.



**Figure 5.2:** First three modes for 1LFH obtained from simulations using  $C_{\alpha}$  NMA. They all have been represented using RASMOL in a wireframe representation

As opposed to certain rigid cluster NMA modeling schemes [14-16], while in the current approach, no distinction has been defined, the constituent residues can be classified into certain flexible and rigid domains, leading to a hinging motion of the defined lobes. In addition to the outcome of the simulations reported, the  $\beta$ -factors computed from NMA simulations have been plotted along with the experimentally obtained values from the Protein Data Bank.



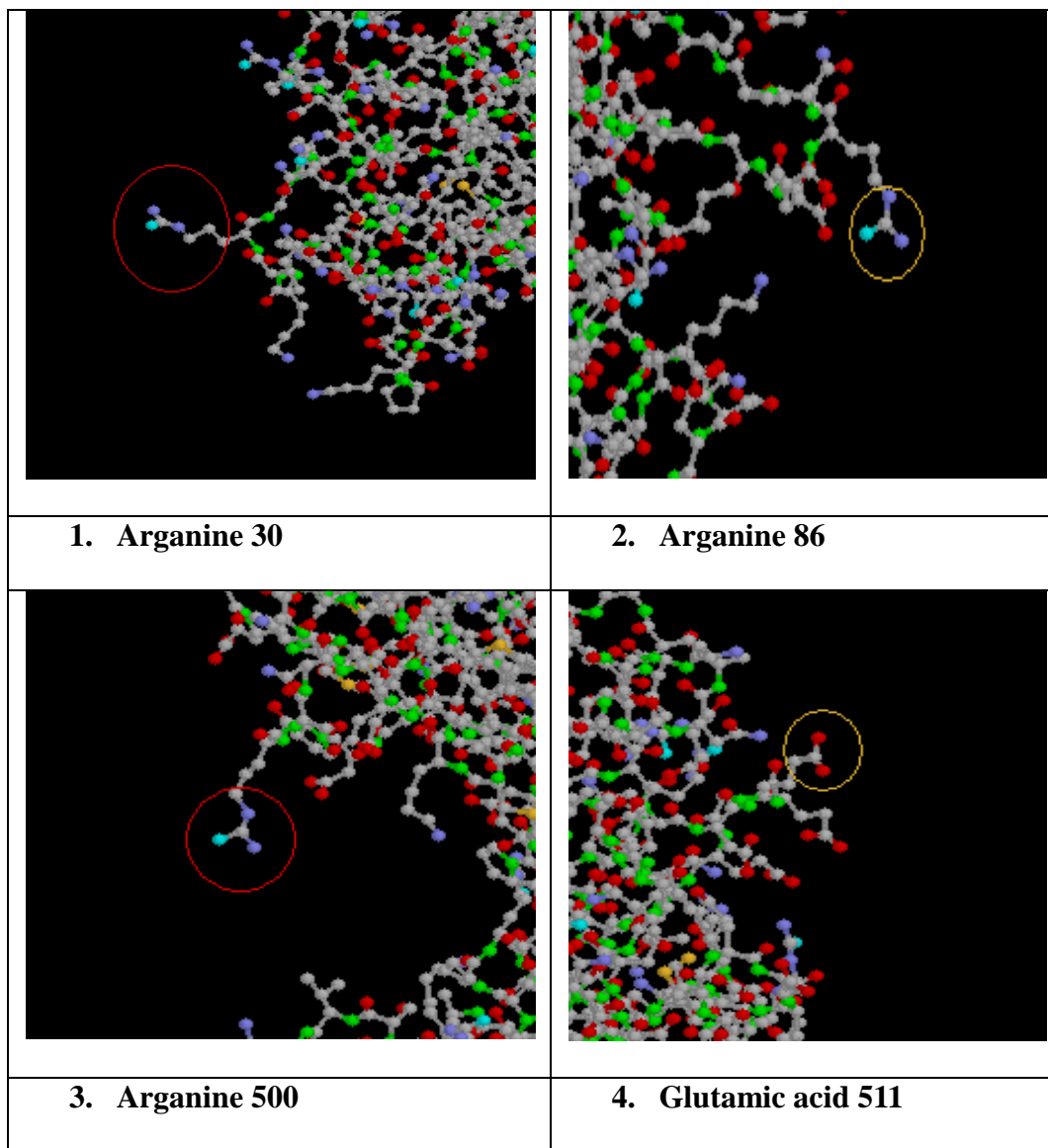
**Figure 5.3:** A plot showing experimental versus calculated  $\beta$ -factors for Lactoferrin. The calculated values are represented by the blue line, and the red line represents the experimentally reported values obtained from the Protein Data Bank.

From Figure 5.3, a good agreement in the values of the  $\beta$ -factors yet again affirms the suitability of using an all atom NMA methodology. With regards to the  $\beta$ -factors, there are some interesting characteristics that have been observed, and can be associated with the model's behavior; primarily, the observed peaks are found for atoms in the outer periphery of the main cluster. It can be explained because we use spring constants now instead of unity, so these specific atoms are connected to other atoms through stronger bonded force constant, and weak non bonded force constants to other non-covalently bonded atoms. This directly alters the stiffness matrix and causes much greater amplitude for  $\beta$ -factors for these specific atoms. With regards to the observed peaks, that is extremely high values of computed  $\beta$ -factors, as mentioned, can be attributed to the orientation of the structure of Lactoferrin. Certain atoms were found to be on the outer periphery, and constitute the amino acids that are generally observed to be hydrophilic like Arginine and Glutamic acid, and tend to be orient along the outer periphery. As a result, the atoms that constitute these residues, due to greater values of mutual distances with other atoms, not only have minimal non bonded interactions. This is an effect that gets induced in to the ENM by the value of cutoff distance and also the values for non bonded force constants between such atoms. Hence, in the chemical information based ENM, these typical atoms are rendered such that they have strong covalent bonds and weak non bonded force constants. This results in a much higher value of  $\beta$ -factors for these specific atoms. The Table 5.1 represents a set of such atoms that were identified in Lactoferrin, and their structural details, elucidating their orientation are represented in Figure 5.4.

**Table 5.1:** Represents a list of atoms observed to have high values of computed  $\beta$ -factors numbers, their types and the amino acids they constitute.

Atom Number	Atom type	Residue name and number
253	N	Arganine-30
254	C	
255	N	
256	N	
676	N	Arganine-86
1374	O	Glutamic acid-178
3851	C	Arganine-500
3852	N	
3853	N	
3929	O	Glutamic acid-511

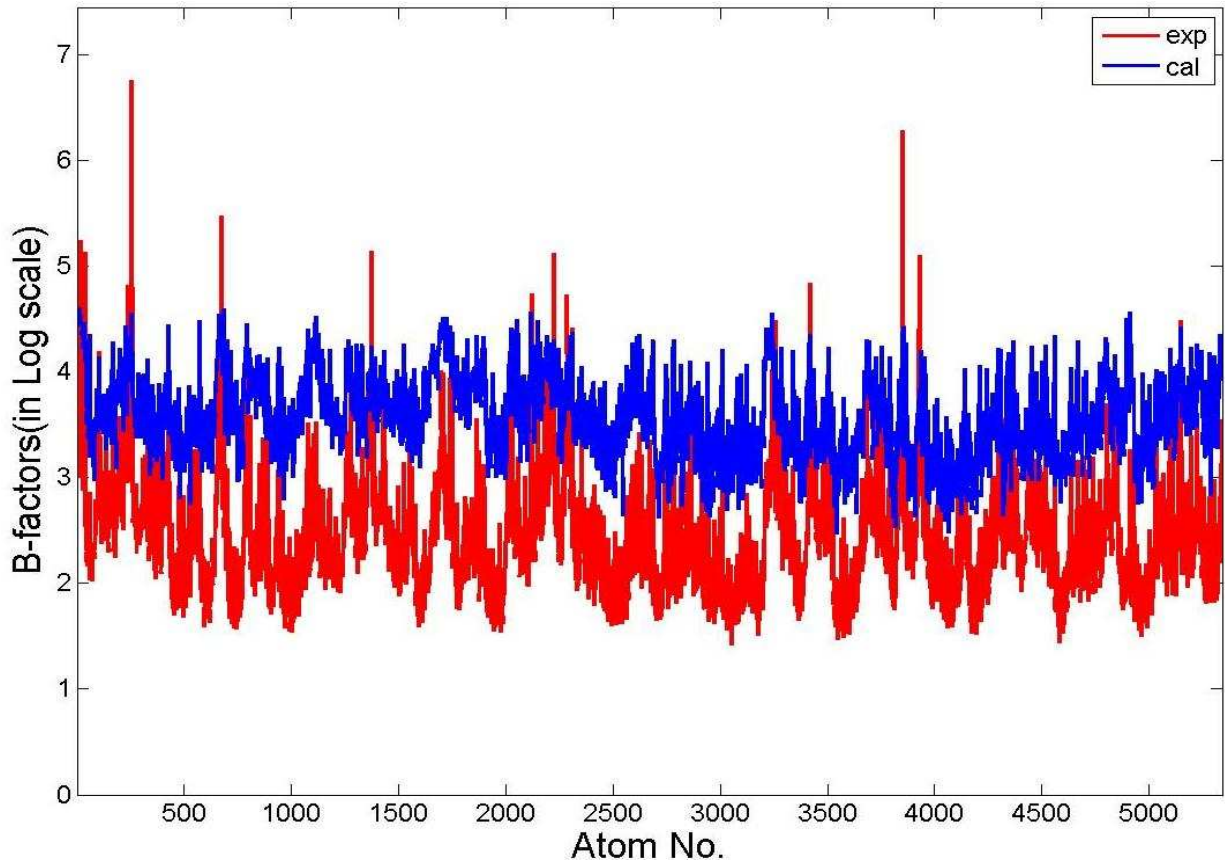
Another interesting observation from the plot of  $\beta$ -factors is that, the NMA simulations are performed on an isolated molecule of Lactoferrin. This is the primary reason for the observed peaks. As opposed to this, the reported values of  $\beta$ -factors in the PDB are experimental values that are obtained from analysis of a crystal and not of an isolated molecule as in the simulations. As a result of these, the peripheral atoms would be surrounded by similar atoms from the neighboring Lactoferrin molecule. As a result of this, there would be more non bonded interactions in a real system.



**Figure 5.4:** Images of atoms on the outer periphery obtained from the conformation of Lactoferrin obtained from PDB. The ENM represents the connections of these atoms with the surrounding atoms.

Figure 5.5 shows the plot of the computed values of  $\beta$ -factors from NMA simulations and the experimental values on a semi-logarithmic scale. This is done so to observe the proximity of the trend shown by the computed as well as the experimental values.

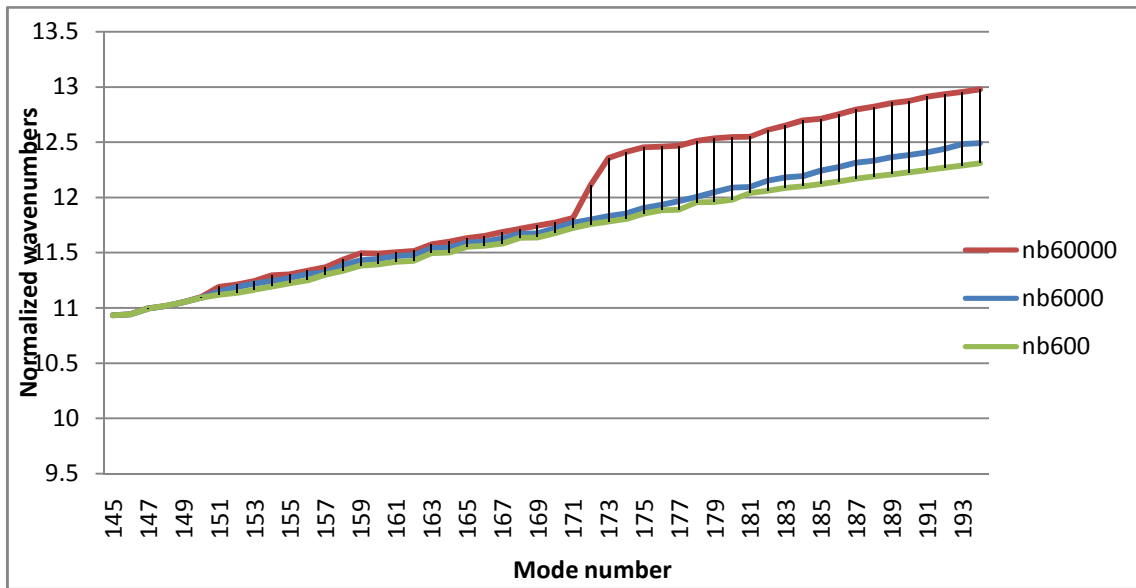
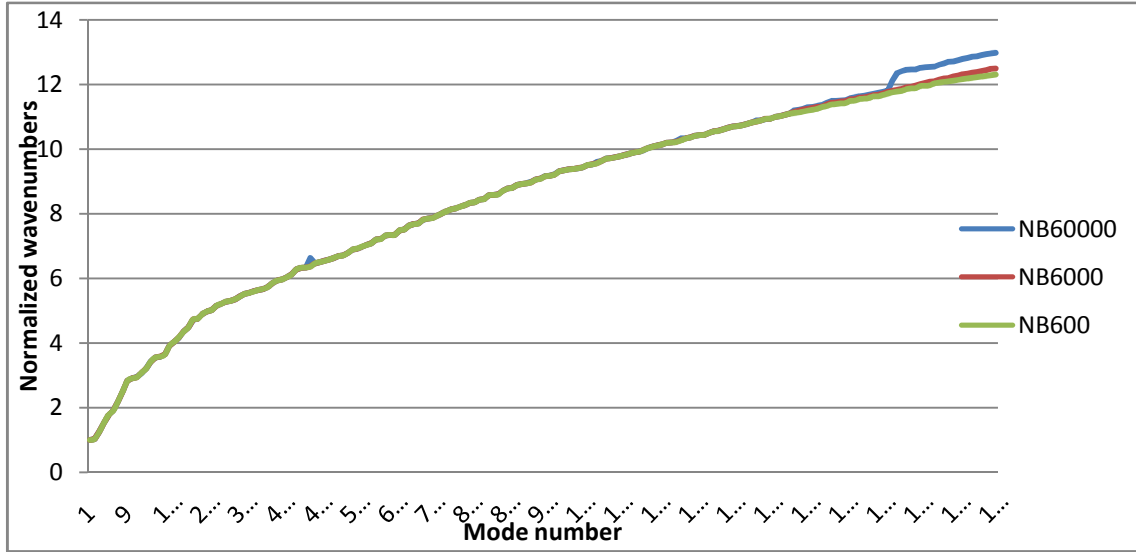




**Figure 5.5:** Semi-logarithmic plot of computed as well as experimental  $\beta$ -factors for Lactoferrin from all-atom NMA simulation.

#### 5.4 Sensitivity Analysis

As in the case for amino acids, a similar sensitivity analysis of studying the effect of changes in the input parameters on the model's behavior in the case of Lactoferrin has also been carried out. To establish the robustness of the model, the non-bonded force constants were varied between, 600 dynes/cm to 60000dynes/cm. Moreover, the connectivity matrix was also altered by changing the cutoff distance that in turn altered the linking matrix and the overall stiffness of the system.

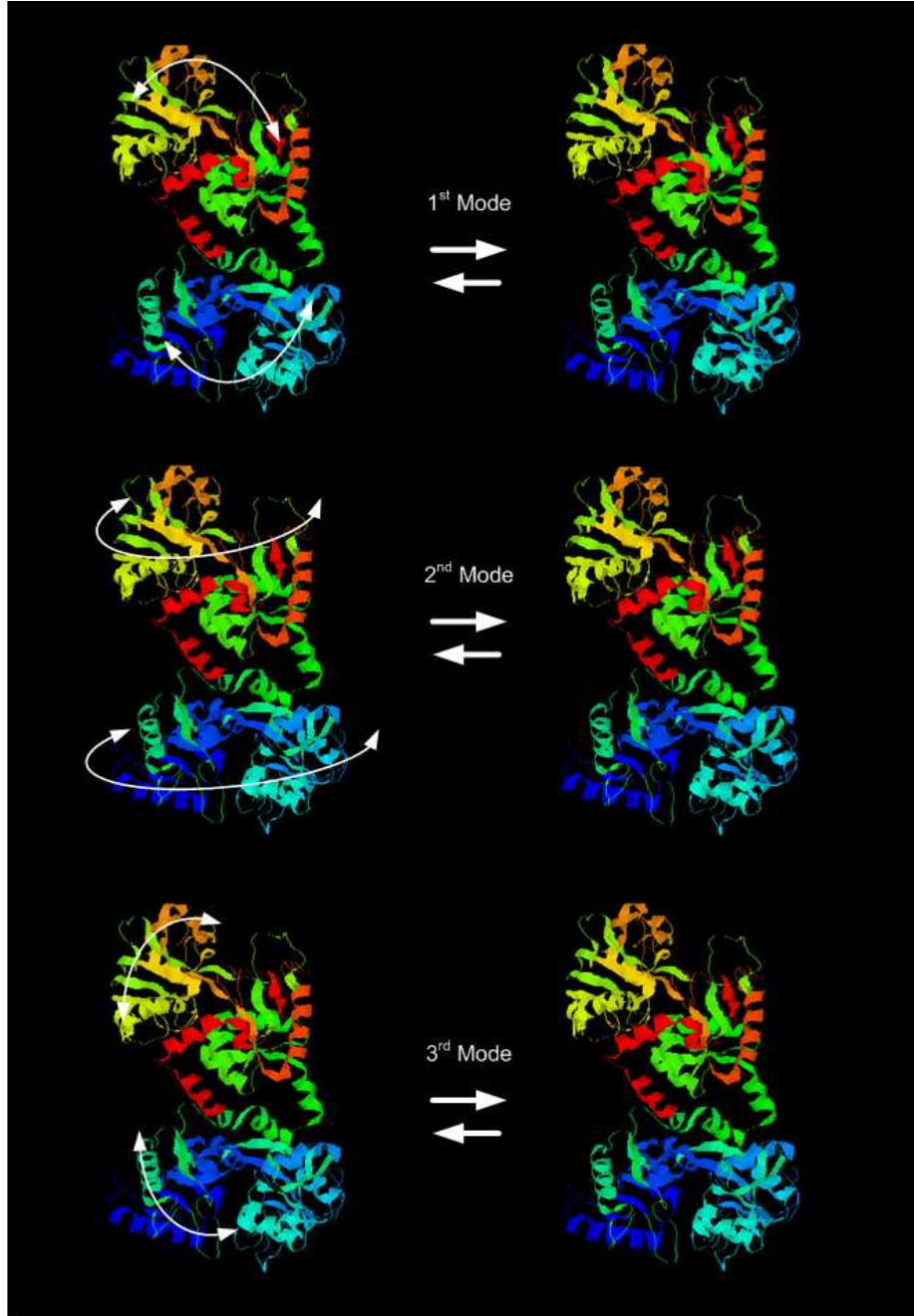


**Figure 5.6:** (a) Plot of normalized wavenumbers for Lactoferrin against the mode number. (b) Normalized wavenumbers for Lactoferrin from mode numbers 145 to 194, elucidating the convergence up to the 150<sup>th</sup> mode.

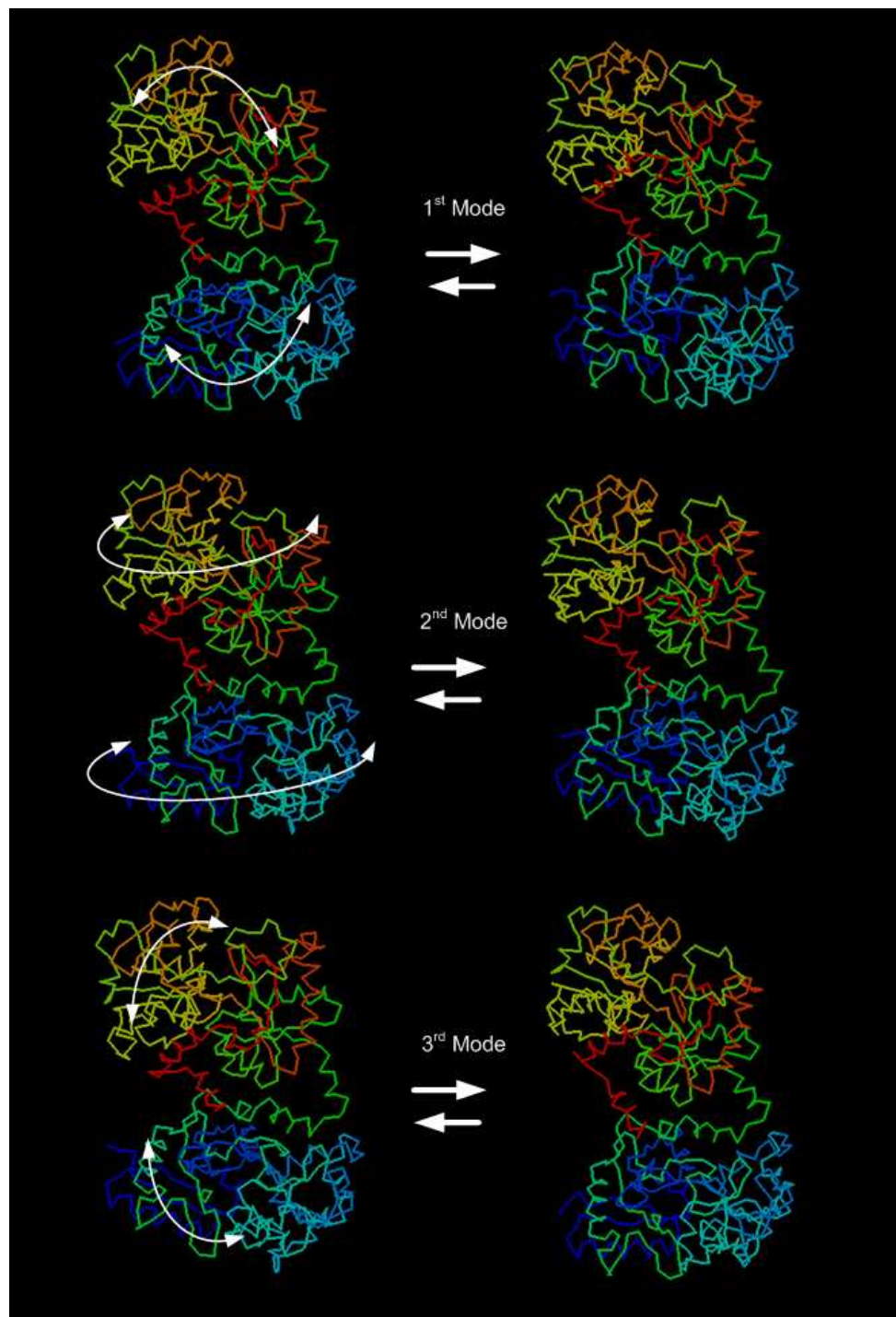
A plot for normalized wavenumbers was then generated in a manner similar to the one adopted for Lactoferrin as reported in Figure 5.6 (a) and (b). It is interesting to note however that through this analysis, macromolecules insinuate a greater sensitivity



to the connectivity in the low frequency domain. It implies that as long as the total connectivity, which is determined by the defined cutoff distance is not altered, the animations suggest no change in the modeshapes. But the high frequency, more local modes, show a greater dependency on the absolute value of force constant, both bonded as well as non-bonded assigned between atoms. In addition to the results for 1LFH as illustrated in this section, simulations for 1LFG were performed. On comparing the modeshapes and wavenumbers in both these cases, the animations for 1LFG, like in the case for 1LFH show one to one correspondence with the existing literature. The animations for 1LFG for all atom case have been summarized in Figure 5.7.



**Figure 5.7:** First three modes for 1LFG obtained from simulations using chemical information based NMA. They all suggest the global motion of what is commonly referred to as the head and the two lobes.



**Figure 5.8:** First three modes for 1LFG obtained from simulations using  $C_{\alpha}$  NMA. They all have been represented using RASMOL in a wireframe representation.

Subsequent to running all atom simulations for ILFG,  $C_{\alpha}$  coarse grained model was simulated and animations were generated to ascertain the credibility of the new modeling scheme. These animations for all-atom NMA and  $C_{\alpha}$  NMA are represented in Figure 5.7 and Figure 5.8, respectively. As can be seen from the animations, they are in agreement. Moreover, these animations were also compared with results generated in another lab that is involved in study of biomolecules (Bahar Lab: School of Medicine, University of Pittsburgh) which uses an Anisotropic Network Model to explore the relationship between dynamics and function for many proteins [77, 78]. Like the  $C_{\alpha}$  coarse grained model, it uses Elastic Network methodology and represents the system in the residue level. The macromolecule is thus represented as a network, or graph. Each node is the  $C_{\alpha}$  atom of a residue and the overall potential is simply the sum of harmonic potentials between interacting nodes.

With regards to the wavenumbers, the one hindrance with comparing computed wavenumbers with spectroscopy data is that the available experimental frequency assignment does not provide us with the same resolution as that obtained from full atom NMA. Therefore, the focus in this work has been on incorporating the sensitivity analysis scheme to stress upon the fact that modes shapes are not altered in the low frequency domain, and the ratios of wavenumbers are also unaffected by variations in the input, indicating that once the lowest frequency is obtained, it can be matched with the results from full atom NMA, since it would mean scaling up and down the inputted force constants. So, until the inputted force constants that exactly replicate the real physical system are obtained, these normalized eigenvalues can be utilized for exploiting

model's characteristics as opposed to absolute wavenumbers; like in the case of Cysteine, where in such a scaling was done once the lowest wavenumber was obtained and subsequently, the higher wavenumbers were determined by matching the lowest wavenumber from simulation.

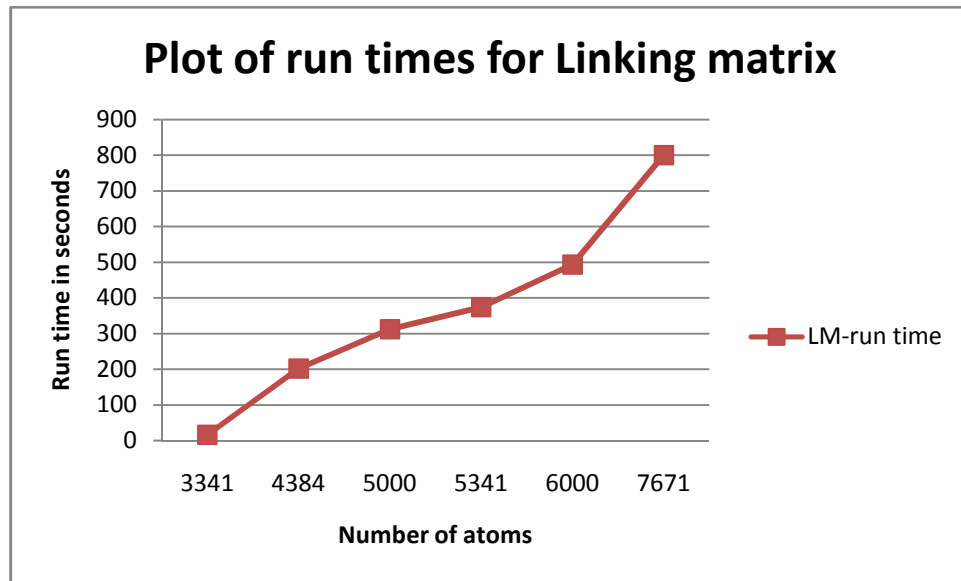
### **5.5 Computational complexity**

A full atom NMA requires a far more computational effort than a conventional  $C\alpha$  coarse grained model. These simulations have so far been computationally prohibitive due to the high number of degrees-of-freedom required to capture motions of large structures. Understandably, the primary difference between the two methods is the linking matrix. While in a  $C\alpha$  NMA, the linking matrix only represents a binary scheme of assignment, 1's indicating the presence of a bond and 0's representing the absence of the same, the linking matrix in an all atom NMA was required to store specific information of force constants between all interacting pairs of atoms. Hence, this directly prompted the identification of intramolecular interactions within a residue and also the peptide bond between carboxyl and the amide groups of interacting amino acid residues. While employing this assignment scheme into use, certain aberrations were observed, due to absence of positional co-ordinates of all the atoms of Lactoferrin in the PDB file, possibly due to a poor resolution. As a direct consequence of this, a given residue was found to consist of different number of atoms. For example, Arginine was found to have 5 as well as 11 atoms. Initially, this caused improper assignment of force constants and rendered the entire linking matrix wrong. Hence, to overcome this, in addition to

identifying the amino acid, a distinction of number of atoms in the same was introduced. Apart from the challenge of assigning appropriate force constant, determination of the same was also an involving task. As explained in the previous chapters, NMA simulations were performed on simple linear molecules to determine the force constants between numerous acting pairs of atoms. Once all the possible types of bonds present in Lactoferrin were identified, these simple linear molecules were selected depending on the presence of the required bond.

So far we have discussed the ground work that was required in computing the inputs. In addition to this, understandably, full atom NMA required much more computational effort. Significantly large dimensions of input matrices; mass ( $3n \times 3n$ ), linking ( $n \times n$ ) and stiffness ( $3n \times 3n$ ) were observed. As opposed to conventional NMA, which required only 691  $\alpha$  atoms, one from each amino acid, full atom required handling of approximately 5300 atoms. The dimensionalities of the matrices were significantly affected by this. Performing inversion and other matrix operations on such huge matrices required special inclusions in the code. In order to give an estimation of run times for various simulations as a function of the number of atoms of the system, simulations for different proteins were run, in order to generate their linking matrices and also the NMA simulations. The run time for all these simulations was recorded in MATLAB, and using regression, a polynomial expression was determined to compute the computational time for both, the linking matrix as well as the NMA simulations. This was done so, since these two codes of simulations require the most of computational effort that goes into such an analysis. By running different simulations, these runtimes

were recorded and plots have been generated to understand the relationship between number of atoms of a system and the proportional time for performing its simulations. This would enable user to predetermine the computational effort and time that would go into such an analysis.



**Figure 5.9:** Represents a plot of run time for generating the linking matrix in all atom NMA simulation, showing the variation in the same as a function of the number of atoms of the protein.

The Figure 5.9 shows the values of runtimes observed while generating linking matrices for different Proteins with disparate number of atoms. As mentioned, in MATLAB, using regression a polynomial expression was obtained to compute the time required for generating the linking matrix of a given protein with ‘N’ atoms, such that;

$$L(t) = 0.0000009603711 \times N^2 + 0.17050365 \times N - 564.5358 \quad (5.1)$$

Where,

$L(t)$ : Time to generate the linking matrix

$N$ : No. of atoms in the given protein.

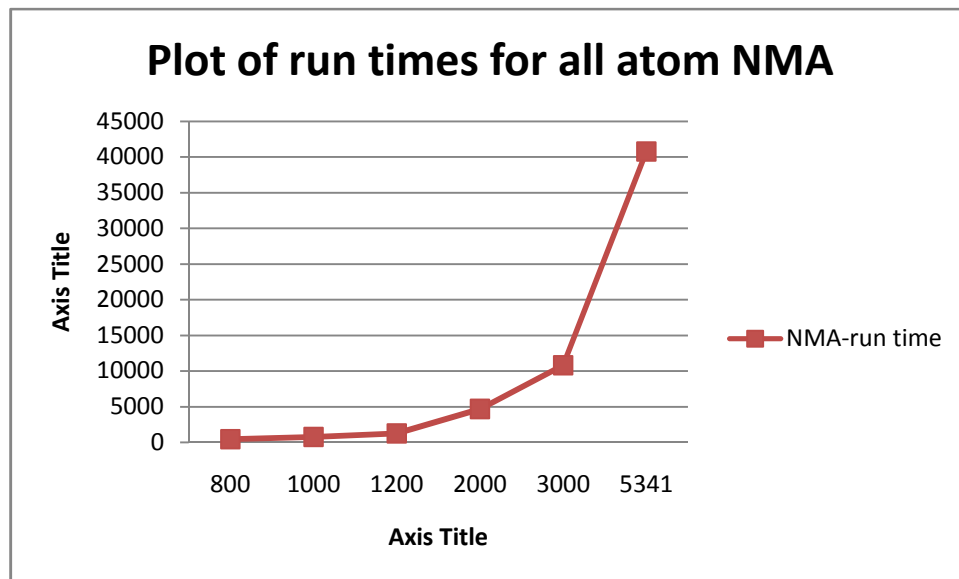
Principally, though conventional NMA is analogous to all atom NMA but the run times of the simulations are significantly different. This is due to the fact that matrix operations like the computation of stiffness matrix from linking matrix requires more time in its computation. Also, the equation 2.9 involves multiplication of stiffness matrix and mass matrices of the orders of 15300x15300 and subsequent computation of the eigenvector and eigenvalue sets from the ‘S’ matrix. Like Figure 5.9, runtimes of NMA simulations were also recorded to express them as a polynomial in ‘N’, and a plot was generated as illustrated in the Figure 5.10.

$$P(t) = 0.0017685 \times N^2 - 1.951 \times N + 746.48 \quad (5.2)$$

Where,

$P(t)$ : Time to run all atom NMA simulation

$N$ : Number of atoms in the given protein.



**Figure 5.10:** Represents a plot of run time for running the all atom NMA simulation, showing the variation in the same as a function of the number of atoms of the protein.



The eigenvector set in particular, computed from all atom NMA was also over 2GB of memory. Hence, while these simulations could be performed on a personal computer with 6GB of ram, MATLAB version 7.8 was required so that the huge input and output matrices could be dealt with.

## 5.6 Conclusions

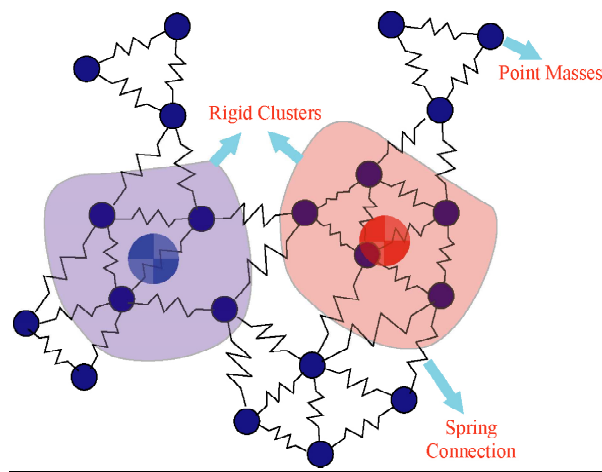
Through the output computed from our analysis, it can be deciphered that an all atom model based NMA indeed provides much larger information pertaining to the biologically relevant and important low frequency domain. With the incorporation of atomistic details, the modeling scheme tested is a better representation of the actual Lactoferrin's behavior. While in the scope of the current study, results have been established for Lactoferrin alone, with information on the structural details of other biomolecules available on various databases like the Protein Data Bank, the modeling scheme is sufficiently flexible to incorporate any observed aberrations by assigning specific values for force constants in the linking matrix, if any. Further insight into the complex field involving the determination of precise force fields would result in enabling the undertaken modeling scheme to have a better input, thereby enabling its users to determine even more precise wavenumbers of the observed modeshapes. It can thereby be proposed that by comparing this data with results from other experimental and computational approaches like spectroscopy, significant information that would help in elucidating the biological function of a macromolecule can be explained using chemical information based NMA.

## CHAPTER 6

### HYBRID NORMAL MODE ANALYSIS USING CHEMICAL INFORMATION BASED ELASTIC NETWORK MODEL.

#### 6.1 Introduction

In large macromolecules, slower more global motions are observed to consist of collective motions of the constituent atoms. In other words, the globally collective motions of the system are dominantly ruled by a few of the slowest modes. Statistical mechanics also predicts that the contribution to the corresponding eigenvalue occurs naturally favorable in the low-frequency modes. This means that the low frequency modes are naturally favorable to occur. Information from such unison motion of a large set of atoms can be used to identify certain rigid domains and flexible loops within a conformation. This implies that certain residues in a protein act as hinges about which the collective motions of atoms take place. Knowledge of such dynamic behavior of the system can be incorporated in the way that a protein is modeled. This understanding has lead to the development of a Hybrid Normal Mode Analysis (HNMA). In this methodology, broadly, the constituent atoms of a protein are classified to be either a part of a cluster or independent point masses in the spatial domain. Hence, clusters are defined, consisting of certain fixed number of atoms, and these clusters are modeled to be connected to neighboring clusters by certain defined point masses. A pictorial depiction of such a methodology is given in Figure 6.1.



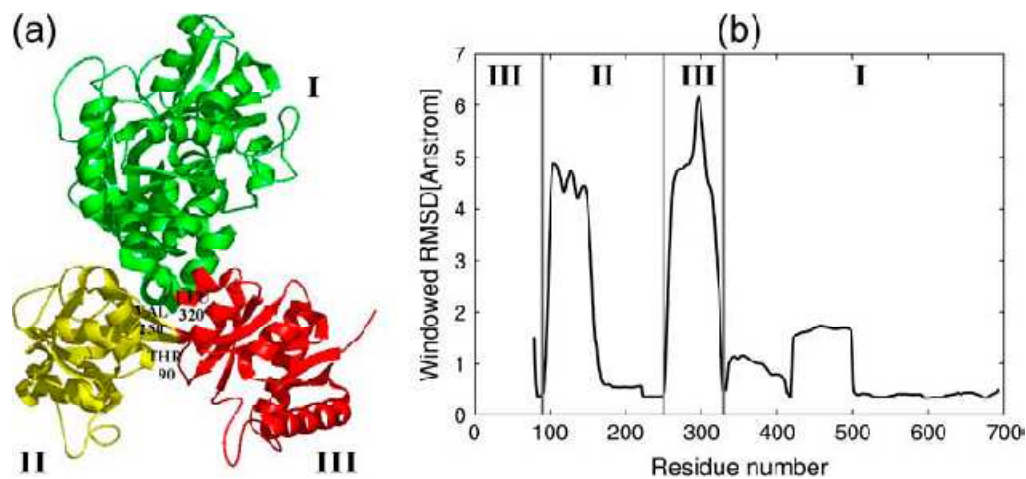
**Figure 6.1:** Schematic of the hybrid elastic network model for the complex structure which contains both rigid domains and flexible loop regions.

## 6.2 Methodology

In accordance with the concept of HNMA and to implement a modeling technique that affirms the stated concept, the one essential prerequisite is to identify rigid clusters and point masses in a protein. This is done by studying the two conformations of Lactoferrin. Since the PDB provides the Cartesian co-ordinates of the conformations, a technique of Windowed Root Mean Square Distance (WMRSD) is used to classify the atoms as either point masses or a part of a cluster defined. Although many rigidity algorithms and theories have been introduced so far, there is still no unique way to define rigid clusters and point masses with given structures. In this context, first, rigid-clustering starts with the static comparison between two end structures. We can also count on the structural information defined by previous literature or experimentally observed rigid cluster domains. Next, the WRMSD is measured to define rigid cluster set. As expected, a certain window size is defined; such that, at a given instance, a set of residues are

compared in both the conformations are compared. For example, a window size of 10 would imply that the positional co-ordinates of the constituent atoms within the same residues in both the conformations would be compared. As the name suggests, it is a square root of the mean of the squares of the difference between the co-ordinates of the same atoms from both the conformations. This enables us to identify flexible and rigid parts. The residues that experience greater values of displacement have a high value of WMRSD. Similarly, certain residues that undergo small values of displacements can be considered as hinges, about which the hinging motion takes place. Hence, the size of windows should be small enough not to lose local flexibility of structures. For Lactoferrin, Windowed RMSD results suggested that it could be broadly classified to consist of three rigid domains:

Head, Right lobe, and Left lobe.



**Figure 6.2:** A rigid-cluster model of the Lactoferrin structure. (a) Lactoferrin is assumed to have three rigid clusters: head (green), left (yellow), and right (red) lobes. Two lobes are opened and closed by the hinge motion around Thr90 and Val250. RMSD between corresponding clusters in each conformation is displayed in (b)

The Figure 6.2a shows the schematic of a hybrid model for Lactoferrin with three clusters. The Figure 6.2b on the other hand gives the WMRSD values that were used to define these clusters. These rigid domains are connected to peripheral point masses, which in turn are connected to other point masses in the neighboring rigid domains. The Tables 6.1 and 6.2 discuss the way in which the clusters have been defined. In accordance with the underlying principle, the clusters are connected to each other by point masses, it was essential to define these point masses. Hence, to represent the connection between two clusters, certain atoms at the interface of any two given clusters were considered as individual atoms based on a certain cutoff distance. In order to do so, for a given cluster, the distances between its atoms from all the atoms of the interfacing cluster were computed. A certain predetermined cutoff distance was then used. For a pair of atoms, i.e. an atom from one cluster and the second atom from another cluster, any distance less than this cutoff distance's value, both the atoms were classified as point masses. In the case with three clusters, the cutoff distance used was  $6\text{\AA}$ , while in the case of five clusters the  $4\text{\AA}$  was the value of cutoff distance used. Based on the cutoff scheme, the number of point masses in the cases with three and five clusters was determined to be 560 and 1255, respectively. Use of such a distance cutoff scheme and from the knowledge of flexible domain from WMRSD calculations enabled the modeling of Lactoferrin to be composed of both rigid clusters connected by point masses. Once, the point masses and clusters were identified, the corresponding linking and the mass matrices were adjusted to accurately represent the reduced DOF model.

**Table 6.1:** Represents one of the two clustering schemes used to run HNMA simulations on Lactoferrin. Specific clusters with their constituent amino acids and corresponding atoms numbers are listed.

Cluster	Residue #	Atom #	#Atoms in a cluster	#Point masses
Right lobe	1-90 251-320	1 – 710 1950 - 2497	1115	560
Left lobe	91-250	711 - 1949	1054	
Head	321-691	2498 - 5341	2615	

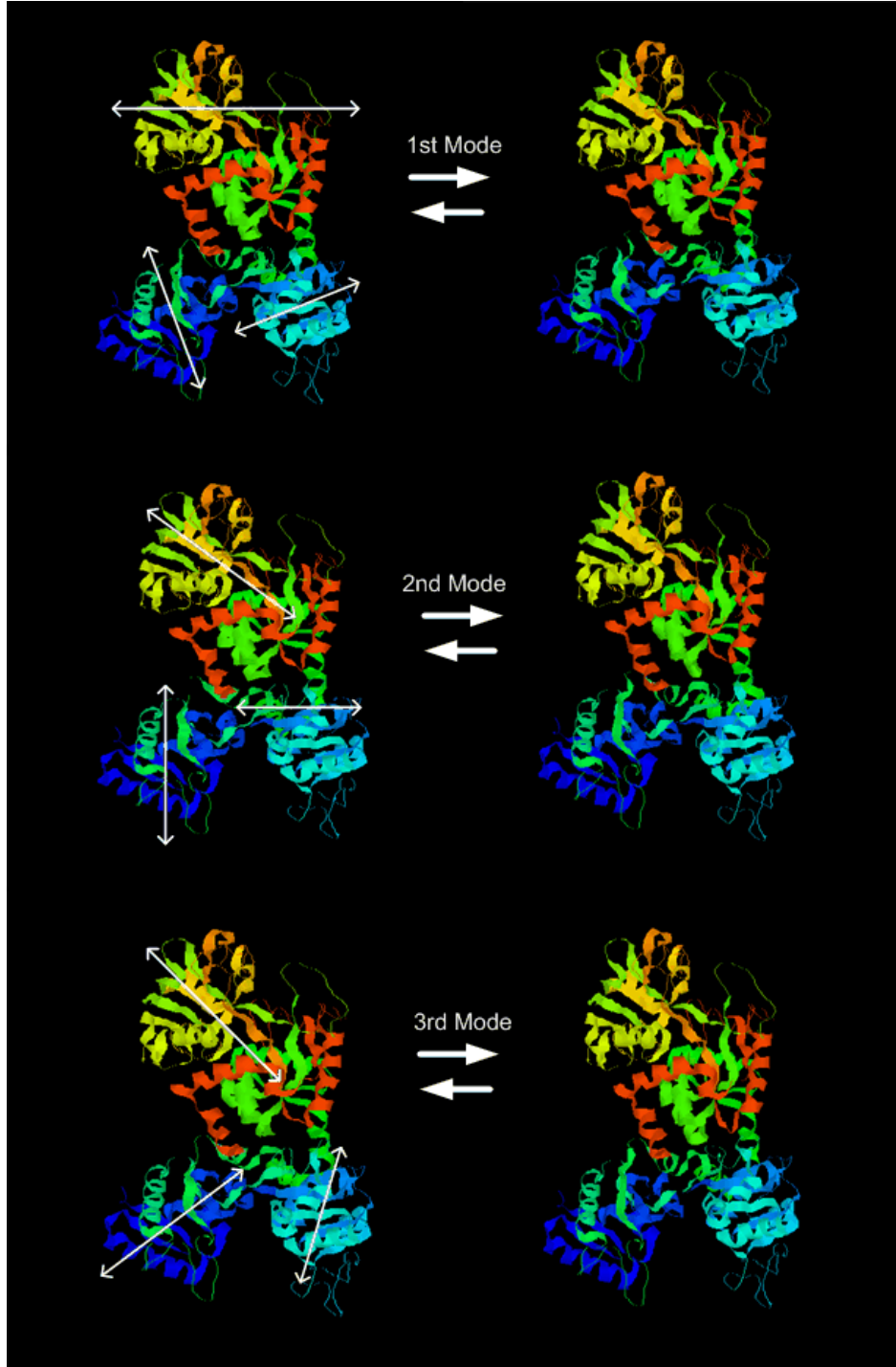
In order to perform HNMA simulations, the stiffness and the inertia matrix were generated and subsequently, the equation of motion was determined [18]. The primary efforts in determining key input parameters were essentially concentrated on determining force constants as explained in Chapter 3, and also in generating the linking matrix. Moreover, once the atoms were classified as either point masses or to be a part of a cluster, a sequential rearrangement was required and all the input parameters like the mass and the linking matrices had to be reordered to match the new sequence of the assorted set of atoms. Subsequently, the inertia and the stiffness matrices of the reduced DOF system had to be generated in order to perform HNMA simulations. The results for the modeshapes from such a clustering have been discussed in the next section. In addition to the tested clustering scheme, another scheme with five clusters was also carried out. This is represented in the Table 6.2. Using the clustering scheme mentioned in the Table 6.2, HNMA was performed again to observe the outputs of modeshapes.

**Table 6.2:** Represents the other clustering scheme used to run HNMA simulations on Lactoferrin. Specific clusters with their constituent amino acids and corresponding atoms numbers are listed.

Cluster	Residue #	Atom #	# Atoms in a cluster	#Point masses
Right lobe 1	1-90	1 – 710	532	1255
Right lobe 2	251-320	1950 – 2497	1180	
Left lobe	91-250	711 - 1949	364	
Head 1	321-520	2498 – 3966	1035	
Head 2	521-691	3967-5341	975	

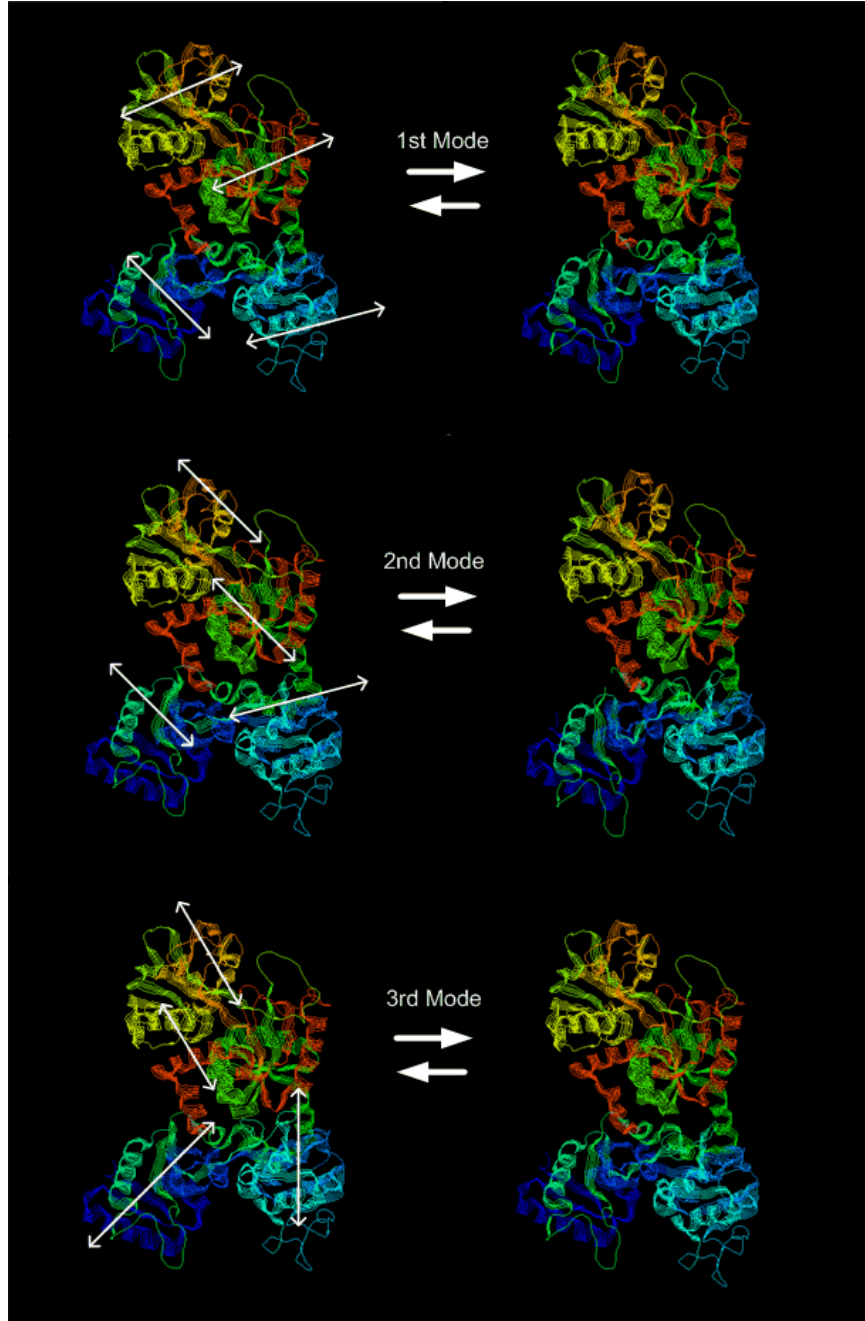
### 6.3 Results and discussions

Once the HNMA simulations with the initial clustering scheme were run, animations of first few lowest modes were generated. These have been represented in the Figure 6.3 below. On comparing these results with the results from that obtained from all atom NMA and coarse grained model, it was established that these modeshapes were rendered incorrect. While the real values of eigenvalues suggested that the code used for HNMA simulations was correct, it directly implied that the outputs were sensitive to the input parameters. It could be observed that the way in which the clusters are defined governs the dynamics.



**Figure 6.3:** Represents animations of the first three modes obtained by running the HNMA simulations on 1LFH.





**Figure 6.4:** Represents the animations of the first three modes of 1LFH by running HNMA simulations on a model defined to have five clusters.

In order to substantiate this hypothesis, a new clustering scheme as that elucidated in Table 6.2 was undertaken to observe the variation in the output of

modeshapes. The results obtained from HNMA simulations with such a clustering are represented in the Figure 6.4. So, while the modeshapes were indeed different from that obtained with a scheme of using three clusters, yet they do not match with the results obtained from all atom or coarse grained model, which have been verified to give results that match with those obtained from existing modeling and analysis schemes. This is understandable because, based on a clustering scheme; the inertia and the stiffness matrices are altered. Hence, while the code does return real and positive eigenvalues and eigenvectors, it can be observed that the way in which the clusters are defined in this study do not replicate the flexibility of the real system and so significantly affect the dynamics of the system.

#### **6.4 Conclusions**

In this part of the research, the possibility of using a Hybrid Elastic Network Model, which is mathematically more rigorous and computationally much more efficient method of modeling than the all atom ENM. Subsequently, NMA simulations have been performed. This is done so, as the results from application of a hybrid model to coarse grained models have yielded useful results pertaining to the low frequency domain and have been successfully implemented to animate the modeshapes. A general code that incorporates atomistic details has been successfully generated as a part of this research, and the observed real and positive eigenvalues suggest that with a better clustering, a better replication of the real physical system can be put in place to further exploit if the HNMA can give comprehensive results for modeshapes as that achieved by all atom

NMA. As a part of this study, the clustering undertaken was based on results from WMRSD calculations as explained in the previous sections. As the results summarized in this chapter can be utilized to establish the dependency of the outcome on the way in which these rigid clusters are defined such that local flexibility is not lost, more study would be required to be carried out in this domain to be able to further exploit the clustering schemes and ultimately the way in which the number and size of each cluster is defined.

## CHAPTER 7

### CONCLUSIONS AND FUTUREWORK

#### 7.1 Conclusions

The step by step approach of application of chemical information based NMA suggests that an all atom based ENM modeling scheme is a feasible option for the analysis of large macromolecules and to study their dynamics pertaining to the low frequency domain. While the results from the analysis of simple linear molecules suggest that with appropriate representation of force fields, modeshapes can be identified along with the corresponding vibrational frequencies, in more complex structures, like amino acids and proteins, due to the effect of non bonded chemical interactions between various molecules, the current methodology does provide accurate identification of modeshapes, and the distribution of the corresponding frequencies which has been explained by the concept of normalized wavenumbers. Hence, in simple molecules, this approach can be used as a vibration spectrum assignment scheme, and in large macromolecules, this enables us to generate an ordered set of modeshapes, with animations that provide insight into their global motions which is of great significance in the study of their dynamics to decipher any possible biological function or conformational changes associated with the same. With the unique ability of this technique to generate results pertaining to the frequency domain renders it as a good approach to be coupled along with results from numerous experimental approaches, thereby enabling us to exploit greater information out of the existing data at hand.

Moreover, the entire set of simulations that were required during this research has been generated on personal computers. This implies that as compared to some more expensive methods like Molecular Dynamics, chemical information based NMA is computationally less expensive, also, by incorporating more atomistic details than a  $C_\alpha$  coarse grained model, an all-atom modeling scheme is much more coherent to the actual physical system. As a result, this methodology has been established as a good intermediate approach that presents a fine balance of accuracy of the outcome while also providing its users with the relative ease of computational effort and time.

## 7.2 Future Work

While the current methodology has shown much better results than the existing methodologies, there are broadly two domains where in this work can be further improved upon. Firstly, with a more precise force field parameterization, more accurate results for wavenumbers can be expected. But this would primarily alter the analysis of small molecules. In macromolecules, the sensitivity analysis suggests invariance to these input parameters as illustrated by the normalized wavenumbers. Secondly, more rigorous mathematical modeling can be employed to further reduce down the computational effort required for these calculations. As mentioned, the major inputs in an-all atom modeling scheme are values of masses and representation of force fields. Hence, the force constants used for this study were developed by performing NMA simulations on linear molecules. While comparisons with literature suggest that these values are in reasonable agreement with the reported values, yet the results for absolute values of

wavenumbers obtained at the amino acids' or at protein's level indicate that by incorporating more precise force fields, accurately capturing the chemical interactions among all the atoms in a given macromolecule would result in even better results. With regards to the computational effort, as it has been observed and discussed, the low frequency modes are more global in nature. Hence, this implies that these modes involve a large number of atoms to move together in sync. As a result, like a Hybrid Normal Mode Analysis has been applied to  $C_{\alpha}$  coarse grained models, a similar successful application of defining these rigid domains in an all-atom modeling scenario would greatly reduce the time involved in performing these simulations. By defining such rigid clusters, some flexibility of the system is lost, and so, identification of such rigid domains is a crucial parameter that can affect the outcome. Hence, further work utilizing certain existing techniques like Windowed Root Mean Square Distance (WRMSD) will assist in obtaining this objective. As a result, further work on these two described factors is essential if not imperative to enhance the further applicability of such a methodology.

## APPENDIX A

### THE LINKING MATRIX, ALL-ATOM NMA

```
clear all
clc
format long
load resity
load atomno
load ca % Ca position
load n % N position
load a

i=1;

%bonded interaction coefficient
NC=7e5;% dynes/cm
N2C=7e5;
NH=7e5;
CH=7e5;
CC=7e5;
C2C=7e5; % double bond
CO=7e5; % single bond
C2O=7e5; % double bond
CS=7e5;

%non-bonded interaction coefficient
nb=6e3;
cutoff=2; % lower cutoff
lim=15; % upper cutoff
least=1e-12;
%based on LFH%
m=size(a,1);
rn=size(atomno,1);
% sparse linking matrix
k=sparse(zeros(m));
% for i=1:m
%   i
%   k_initial(i,i+1:m)=1e-12;
% end
% save k_initial k_initial
% check1=k_initial+k_initial;
% load k_initial
% k=k+k_initial;

%non-bonded interactions
for g=1:m-1
    g
    for h=g+1:m
        dis=norm(a(g,:)-a(h,:));
        if dis<=cutoff
            k(g,h)=nb;
        end
    end
end
```

```

elseif dis<=lim
    k(g,h)=nb*exp(-(dis-2));
end
end
end
% k_non_bond=k;
k_non_bond=k+k';
save k_non_bond k_non_bond
%peptide bond
for l=1:rn-1
    k(ca(l)+1,n(l+1))=NC;
end

stack=[1];
i=1;

for t=1:rn
    t
    if resity(t,:)=='GLY' & atomno(t,1)==4
        k(i,i+1)=NC;
        k(i+1,i+2)=CC;
        k(i+2,i+3)=C2O;
        i=i+4;
        stack=[stack;i];
    %
        break;
    else if resity(t,:)=='ARG' & atomno(t,1)==11
        k(i,i+1)=NC;
        k(i+1,i+2)=CC;
        k(i+2,i+3)=C2O;
        k(i+1,i+4)=CC;
        k(i+4,i+5)=CC;
        k(i+5,i+6)=CC;
        k(i+6,i+7)=NC;
        k(i+7,i+8)=NC;
        k(i+8,i+9)=NC;
        k(i+8,i+10)=N2C;
        i=i+11;
        stack=[stack;i];
    %
        break;
    else if resity(t,:)=='ARG' & atomno(t,1)==5
        k(i,i+1)=NC;
        k(i+1,i+2)=CC;
        k(i+2,i+3)=C2O;
        k(i+1,i+4)=CC;
        i=i+5;
        stack=[stack;i];
    %
        break;
    else if resity(t,:)=='SER' & atomno(t,1)==6
        k(i,i+1)=NC;
        k(i+1,i+2)=CC;
        k(i+2,i+3)=C2O;
        k(i+1,i+4)=CC;

```



```

k(i+4,i+5)=CO;
i=i+6;
stack=[stack;i];
%
break;
else if resity(t,:)=='VAL' & atomno(t,1)==7
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+4,i+6)=CC;
i=i+7;
stack=[stack;i];
%
break;
else if resity(t,:)=='GLN' & atomno(t,1)==9
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i+6,i+7)=C2O;
k(i+6,i+8)=NC;
i=i+9;
stack=[stack;i];
%
break;
else if resity(t,:)=='CYS' & atomno(t,1)==6
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CS;
i=i+6;
stack=[stack;i];
%
break;
else if resity(t,:)=='ALA' & atomno(t,1)==5
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
i=i+5;
stack=[stack;i];
%
break;
else if resity(t,:)=='ASN' & atomno(t,1)==8
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=C2O;
k(i+5,i+7)=NC;
i=i+8;

```

```

stack=[stack;i];
%
break;
else if resity(t,:)=='PRO' & atomno(t,1)==7
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i,i+6)=NC;
i=i+7;
stack=[stack;i];
%
break;
else if resity(t,:)=='GLU' & atomno(t,1)==9
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i+6,i+7)=CO;
k(i+6,i+8)=C2O;
i=i+9;
stack=[stack;i];
else if resity(t,:)=='GLU' & atomno(t,1)==5
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
i=i+5;
stack=[stack;i];
%
break;
else if resity(t,:)=='THR' & atomno(t,1)==7
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CO;
k(i+4,i+6)=CC;
i=i+7;
stack=[stack;i];
%
break;
else if resity(t,:)=='LYS' & atomno(t,1)==9
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i+6,i+7)=CC;
k(i+7,i+8)=NC;
i=i+9;

```

```

stack=[stack;i];
break;
%
else if resity(t,:)=='PHE' & atomno(t,1)==11
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=C2C;
k(i+5,i+7)=CC;
k(i+6,i+8)=CC;
k(i+8,i+10)=C2C;
k(i+7,i+9)=C2C;
k(i+9,i+10)=CC;
i=i+11;
stack=[stack;i];
break;
%
else if resity(t,:)=='MET' & atomno(t,1)==8
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CS;
k(i+6,i+7)=CS;
i=i+8;
stack=[stack;i];
break;
%
else if resity(t,:)=='ILE' & atomno(t,1)==8
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+4,i+6)=CC;
k(i+5,i+7)=CC;
i=i+8;
stack=[stack;i];
break;
%
else if resity(t,:)=='ASP' & atomno(t,1)==8
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CO;
k(i+5,i+7)=C2O;
i=i+8;
stack=[stack;i];
break;
%
else if resity(t,:)=='LEU' & atomno(t,1)==8
k(i,i+1)=NC;

```

```

k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i+5,i+7)=CC;
i=i+8;
stack=[stack;i];
break;
%
else if resity(t,:)=='TYR' & atomno(t,1)==12
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=C2C;
k(i+5,i+7)=CC;
k(i+6,i+8)=CC;
k(i+7,i+9)=C2C;
k(i+8,i+10)=C2C;
k(i+9,i+10)=CC;
k(i+10,i+11)=CO;
i=i+12;
stack=[stack;i];
break;
%
else if resity(t,:)=='HIS' & atomno(t,1)==10
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=NC;
k(i+5,i+7)=C2C;
k(i+7,i+9)=NC;
k(i+6,i+8)=N2C;
k(i+8,i+9)=NC;
i=i+10;
stack=[stack;i];
break;
%
else % TRP
k(i,i+1)=NC;
k(i+1,i+2)=CC;
k(i+2,i+3)=C2O;
k(i+1,i+4)=CC;
k(i+4,i+5)=CC;
k(i+5,i+6)=CC;
k(i+5,i+7)=CC;
k(i+6,i+8)=N2C;
k(i+8,i+9)=NC;
k(i+7,i+9)=C2C;
k(i+7,i+10)=CC;
k(i+9,i+11)=CC;

```



## APPENDIX B

### ALL-ATOM NMA CODE

```
clear all
clc
format long
load k1 %Linking Matrix
load M
load a
data=a;
m=size(data,1);

GP=sparse(zeros(3*m));
for i=1:m-1
    i
    for j=i+1:m
        if k1(i,j)>0
            dx=data(i,:)-data(j,:);
            GP(3*(i-1)+1:3*i,3*(j-1)+1:3*j)=k1(i,j)*dx*dx'/norm(dx)^2;
        end
    end
end
GP=GP+GP';
save GP GP
K_R=-GP;
disp('GP saved')

for i=1:m
    i
    temp=zeros(3);
    for l=1:m
        temp=temp+GP(3*(i-1)+1:3*i,3*(l-1)+1:3*l);
    end
    K_R(3*(i-1)+1:3*i,3*(i-1)+1:3*i)=temp;
end

save K_R K_R
S_R=M*K_R*M;

save S_R S_R
disp('K_R M S_R saved')

clear all
```

```

clc
format long
disp('computing eigenvalues and the eigenvectors')
load S_R
load a
load M
n=3*size(a,1);
KK=full(S_R);
1
[v,d]=eig(full(KK));
disp('step1 done')
d=diag(d);
[d,index]=sort(d);
for i=1:n
new_v(:,i)=v(:,index(i));
end
Vx=M*new_v;
save Vx Vx -V6
save d d
disp('step2 done')

for i=1:n
wn_lfhr_gen(i,1)=sqrt(d(i))/2/pi/3e10; %#ok<AGROW>
end
save wn_lfhr_gen wn_lfhr_gen

```

## APPENDIX C

### HYBRID NMA CODE

```
clear all
clc
load an
load k1n
load nc
load mn
% load mass1

load ctofmass1

data=an;

k1=k1n;

m=size(data,1);

num_of_pm=560;      %adjustment
num_of_cluster=5;  %adjustment

offset=3*num_of_pm;
offset1=6*num_of_cluster;
offset2=offset+offset1;

for i=1:num_of_cluster
    eval(['load c',num2str(i)])
end

KT=zeros(offset2);
MT=zeros(offset2);

count=0;

%%%%%%%%%%%% Point Mass NMA %%%%%%%%%%%%%
GP=sparse(zeros(offset));
for i=1:num_of_pm-1
    for j=i+1:num_of_pm
        if k1(i,j)>0
            dx=data(i,:)-data(j,:);
            GP(3*(i-1)+1:3*i,3*(j-1)+1:3*j)=k1(i,j)*dx*dx'/norm(dx)^2;
            count=count+1;
        end
    end
end
```



```

        end
    end
end
GP=GP+GP';
save GP GP
disp('GP saved')
KP=-GP;

for i=1:num_of_pm
    temp=zeros(3);
    for l=1:num_of_pm
        temp=temp+GP(3*(i-1)+1:3*i,3*(l-1)+1:3*l);
    end
    KP(3*(i-1)+1:3*i,3*(i-1)+1:3*i)=temp;
end

for r=1:1028
    for c=r+1:1029
        KP(c,r)=KP(r,c);
    end
end
save KP KP
disp('KP saved')
%break

KT(1:offset,1:offset)=KP;
h=1;
for i=1:num_of_pm
    MT(3*i-2:3*i,3*i-2:3*i)=mn(i,1)*eye(3);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%End of Point mass %%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Rigid cluster nma %%%%%%%%%
KC=zeros(offset1);
MC=zeros(offset1);

for i=1:num_of_cluster-1 % first summation symbol
    eval(['num_sample1=size(c',num2str(i),'1)'])
    eval(['ca=c',num2str(i),';']) % the first cluster

```



```

end
end

KT(offset+1:offset2,offset+1:offset2)=KC;
MT(offset+1:offset2,offset+1:offset2)=MC;

save KC KC
save MC MC
save MT MT
disp('KC MC MT saved')

%%%%%%%%%%%%%%Hybrid NMA %%%%%%%%%%%%%%%
KH=zeros(offset2);
MH=zeros(offset2);

for i=1:num_of_pm % first summation symbol - point mass
    for j=1:num_of_cluster % second summation symbol - rigid cluster
        eval(['num_sample=size(c',num2str(j),',1);'])
        eval(['cb=c',num2str(j),';'])
        for w=1:num_sample % thrid summation symbol - residues in cluster
            beta=cb(w,1); % actual residue number of the second point
            if k1(i,beta) > 0
                gap=data(i,:)-data(beta,:);
                Y=gap'*gap/(norm(gap)^2);
                Q=Q4N(j,w); %% check here
                S=k1(i,beta)*Q'*Y*Q;

                Ma=S(1:3,1:3);
                Mb=S(1:3,4:9);
                Mbt=S(4:9,1:3);
                Mc=S(4:9,4:9);

                KH(3*(i-1)+1:3*i,3*(i-1)+1:3*i)=KH(3*(i-1)+1:3*i,3*(i-1)+1:3*i)+Ma;
                KH(3*(i-1)+1:3*i,6*(j-1)+1+offset:6*j+offset)=KH(3*(i-1)+1:3*i,6*(j-1)+1+offset:6*j+offset)+Mb;
                KH(6*(j-1)+1+offset:6*j+offset,3*(i-1)+1:3*i)=KH(6*(j-1)+1+offset:6*j+offset,3*(i-1)+1:3*i)+Mbt;
                KH(6*(j-1)+1+offset:6*j+offset,6*(j-1)+1+offset:6*j+offset)=KH(6*(j-1)+1+offset:6*j+offset,6*(j-1)+1+offset:6*j+offset)+Mc;
            end
        end
    end
end
end
end

```

```

KT=KT+KH;

%%%%% Eliminate Truncation Error to Make The Matrix Symmetric %%%%%%%%%

for r=1:1046
    for c=r+1:1047
        KT(c,r)=KT(r,c);
    end
end

save KT KT

%break
x1=MT^(-1/2);
for r=1:1046
    for c=r+1:1047
        x1(c,r)=x1(r,c);
    end
end
save x1 x1
clear all
load KT
load x1

ST=x1*KT*x1;
for r=1:1046
    for c=r+1:1047
        ST(c,r)=ST(r,c);
    end
end

save ST ST
disp('KT MT ST saved')

%%%%%%%%%%%%End of hybrid NMA %%%%%%%%%

clear all
clc
load an
load k1n
load nc
load mn
% load mass1

```

```

load ctomass1

data=an;

k1=k1n;

m=size(data,1);

num_of_pm=560;      %adjustment
num_of_cluster=5;  %adjustment

offset=3*num_of_pm;
offset1=6*num_of_cluster;
offset2=offset+offset1;

load ST
load x1
[v1,d]=eig(full(ST));
save v1 v1
save d d

[Y,I]=sort(diag(d));
v_sort=[];
for i=1:offset2
    v_sort=[v_sort;v1(:,I(i))];
end
v_sort=v_sort';
d=Y;
v=x1*v_sort;

save v_sort v_sort
save v v
save d d

for i=1:offset2
wn_lfh(i,1)=sqrt(d(i))/2/pi/3e10;
end
save wn_lfh wn_lfh

%%%%%%%%%%%%% converting into Cartesian coordinates %%%%%%%%%%%%%%

for w=1:offset2

```

```

for i=1:num_of_cluster
    eval(['load c',num2str(i)])
    eval(['c=c',num2str(i),'])
    k=size(c,1);
    cdelta=v(6*(i-1)+1+offset:6*i+offset,w);
    trans=cdelta(1:3,1);
    orient=cdelta(4:6,1);
    R=expm(Jmat(orient));
    for j=1:k
        data_new(c(j),:)=(data(c(j),:)-ctofmass1(i,:))*R'+ctofmass1(i,:)+trans';
    end
end
% %
for i=1:num_of_pm
    data_new(i,:)=data(i,:)+v(3*(i-1)+1:3*i,w)';
end
% %
for i=1:m
    delta1(3*(i-1)+1:3*i,w)=(data_new(i,:)-data(i,:))';
end
%
end

% % Grandschmidt and convert to original order for comparison % %
save delta1 delta1
delta2=gramschmidt(delta1);
save delta2 delta2
eig_converter('delta2')

```

## BIBLIOGRAPHY

- [1] International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*. 431, 931–945.
- [2] Berman, H. M, Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E., 2000. The Protein Data Bank. *Nucleic Acids Res.*, vol.28, pp.235-242.
- [3] Tirion, M. M., 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 77:1905–1908.
- [4] A. L. Jenkins, R. A. Larsen, T. B. Williams, 2005. Characterization of amino acids using Raman spectroscopy. *Spectrochimica Acta Part A* 61 1585–1594.
- [5] M. Pézolet,\* M. E. Rousseau, T. Lefèvre, L. Beaulieu, 2004. Raman Microspectroscopy: An Ideal Technique to Study the Conformation and Orientation of Proteins in Silkworm and Spider Silk Fibers. *Microsc Microanal* 10(Suppl 2), 1314-1315.
- [6] Chitra Murli\_, Susy Thomas, Sugandhi Venkateswaran, Surinder M Sharma, 2005. Raman spectroscopic investigation of  $\alpha$ -glycine at different temperatures. *Physica B* 364 (2005) 233–238.
- [7] S. Petersen, O.F. Nielsen, H.G.M. Edwards, D.H. Christensen, D.W. Farwell, J.P. Hart Hansen, M. Gniadecka, A.R. David, P. Lambert and H.C. Wulf, *J. Raman Spectrosc.* 34, 357 (2003).
- [8] H.G.M. Edwards and F.R. Perez, *Biospectrosc.* 5, 47 (1999)
- [9] D.L.A. de Faria, H.G.M. Edwards, M.C. de Afonso, R.H. Brody and D.L. Morais, *Spectrochimica Acta, Part A*, 60 ,1505 (2004).
- [10] Y Repelin, E. Husson, F. Bennani, C. Proust, 1999. Raman spectroscopy of lithium niobate and lithium tantalate. Force field calculations. *Journal of Physics and Chemistry of Solids*, Volume 60, Number 6, pp. 819-825(7).
- [11] Y Guan and G J Thomas, 1996. Vibrational analysis of nucleic acids. V. Force field and conformation-dependent modes of the phosphodiester backbone modeled by diethyl phosphate. *Biophys J.* 71(5): 2802–2814.

- [12] L. V. Stepakova, M. Y. Skripkin, L. V. Chernykh, G. L. Starova, L. Hajba, J. Mink. M. Sandstro, 2008 . Vibrational spectroscopic and force field studies of copper(II) chloride and bromide compounds, and crystal structure of KCuBr<sub>3</sub>. *J. Raman Spectrosc.* Volume 39: 16–31.
- [13] S. Dasgupta, K. A. Brameld, C. Fan, W. Goddard, 1997 .Ab initio derived spectroscopic quality force fields for molecular modeling and dynamics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Volume 53, Number 8, pp. 1347-1363(17).
- [14] M.K. Kim, M.K., G.S. Chirikjian, and R.L. Jernigan. *J. Mol. Graph. Model* 21, 151 (2002).
- [15] M.K. Kim, R.L. Jernigan, and G.S. Chirikjian. *Biophys. J.* 83, 1620 (2002).
- [16] M.K. Kim, Li. W., B.A. Shapiro, and G.S. Chirikjian. *J. Biomol. Struct. and Dyn.* 21, 395 (2003).
- [17] M.K. Kim, R.L. Jernigan, and G.S. Chirikjian. *J. Struct. Biol.* 143, 107 (2003).
- [18] M.K. Kim, R.L. Jernigan, and G.S. Chirikjian. *Biophys. J.* 89, 43 (2005).
- [19] Moon K. Kim, Yunho. Jang, and Jay I. Jeong. *Int. J. Control Autom. Syst.* 4, 382 (2006).
- [20] Yunho. Jang, Jay I. Jeong, and Moon K. Kim. *Nucleic Acids Research* 34, W57 (2006).
- [21] Jay I. Jeong, Yunho Jang, and Moon K. Kim. *J. Mol. Graph. Model.* 24, 296 (2006).
- [22] D.J.Millen. *Can. J. Chem.* 63 (1985).
- [23] P.Phillipson. *J. Chem. Phy.* 39 (1963).
- [24] K.Novosadov, B. J. *Struc. Chem.* 16 (1975).
- [25] V. I.Bazhanov, *J. Stru Chem.* 27 (1986).
- [26] A.Navarro,M.P.Fernandez-Lienres,A.Ben Altabef, M.Fernandez Gomez, J. Gonzalez , and R.Escribano. *J. Mol. Struct.* 482, 601 (1999).
- [27] K.Venkateswarlu and S.Sundaram. *Proc. Phys. Soc. A.* 69, 180 (1955).



- [28] O. Alvarez-Bajo, M. Sanchez-Castellanos, C.A. Amezcua-Eccius and R. Lemus. *J. Mol. Spec.* 237, 247 (2006).
- [29] L. Bizzocchi, C. Degli Esposti, A. Mazzavillani, and F. Tamassia. *J. Mol. Spec.* 221, 213 (2003).
- [30] A. E. Özel, Y. Büyükmurat, and S. Akyüz. *J. Mol. Struct.* 661-662, 455 (2001).
- [31] Phang Dinh Kieng, A. B. Kovrikov, and A. I. Komyak. *J. App. Spec.* 18 (1973).
- [32] NIST. National Institute of Standards and Technology.
- [33] T.Y. Wu, S.T. Shen, *Chin. J. Phys.* 2, 128 (1936).
- [34] S.M. Ferigle, F.F. Cleveland, and A.G. Meister. *Phys. Rev.* 81, 302 (1951).
- [35] S.M. Ferigle, F.F. Cleveland, and A.G. Meister. in symposium on Mol. Struct. Spect., Columbus, Ohio (1951).
- [36] G.Herzberg. *Infrared and Raman spectra of polyatomic molecules*, New York (1945).
- [37] M.A. Kovner, and V.T. Salosin. *Acad. Sci.* 75, 651 (1950).
- [38] A.V.Jones. in symposium on Mol. Struct. Spect. , Ohio State University (1951).
- [39] G.C.Turrell. *J. Chem. Phys.* 26 (1957).
- [40] J. Cz. Dobrowolski, J. E. Rode, J. Sadlej. *J. Mol. Struct.* 810 (2007).
- [41] A. Fernandez-Ramosa, E. Cabaleiro-Lagoa, J.M. Hermida-Ramona, E. Martinez-Nuneza, A. Pena-Gallegoa. *J. Mol. Struct.* 498 (2000).
- [42] Ching-Cheng Wang, Jui-Hung Chen, Shih-His Yin, Woei-Jer Chuang, *PROTEINS: Structure, Function, and Bioinformatics.* 63 (2006).
- [43] M. Noguera, L. Rodriguez-Santiago, M. Sodupe, J. Bertran. *J. Mol. Struct.* 537 (2001).
- [44] Shu-Zhen Liu, Hong-Qi Wang, Zheng-Yu Zhou, Xiu-Li Dong, Xiao-Li Gong. *International J. Quant. Chem.* 105 (2005).
- [45] K. A. Kerr, J. P. Ashmore and T. F. Koetzle. *Acta Cryst.* B31 (1975).

- [46] P. Tarakeshwar and S. Manogaran. *J. Mol. Struct.* 305 (1994).
- [47] S. Gronert, R. A. J. O'Hair. *J. Am. Chem. Soc.* 117 (1995).
- [48] P. Tarakeshwar, S. Manogaran, *Spectrochimica Acta.* 51A, No. 5 (1995).
- [49] D.A. Sapozhnikov, I.N. Smirnova, A.P. Shkurinov, N.V. Sumbatyan. *Vibrational Spectroscopy.* 47 (2008).
- [50] A. Barth. *Prog. Biophys. Mol. Biol.* 74 (2000).
- [51] N. Wright. *J. Biol. Chem.* 120 (1937).
- [52] V. Krishnakumar, N. Prabavathi, S. Muthunatesan. *Spectrochimica Acta Part A* 70 (2008).
- [53] V. Krishnakumar, R. John Xavier. *Spectrochimica Acta Part A* 60 (2004).
- [54] F. P. Ureña, M. F. Gómez, J. L. González, E. M. Torres. *Spectrochimica Acta Part A* 59 (2003).
- [55] V. P. Morozov, V. V. Belokopytov, G. D. Zegzhda, V. N. Moiseenko. *J. Struct. Chem.* 39(1998).
- [56] M. E. Tuttolomondo, L. E. Fernández, A. Navarro, E. L. Varetti, A. Ben Altabef. *Spectrochimica Acta Part A* 60 (2004).
- [57] N. Go, T. Noguti, T. Nishikawa. *Proc. Nati Acad. Sci. USA* 80 (1983).
- [58] A. Barth, C. Zscherp. *Quarterly Reviews of Biophysics* 35 (2002).
- [59] Biological Magnetic Resonance Data Bank.
- [60] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman. *J. Am. Chem. SOC.* 117 (1995).
- [61] R. B. Gerber, G. M. Chaban, S. K. Gregurick, B. Brauer. *Biopolymers.* 68 (2003).
- [62] T.M. Korter, R. Balu, M.B. Campbell, M.C. Beard, S.K. Gregurick, E.J. Heilweil. *Chemical Physics Letters* 418 (2006).
- [63] D. Chakraborty, S. Manogaran. *J. Mol. Struct.* 429 (1998).
- [64] E. Podstawka, Y. Ozaki, L. M. Proniewicz. *Applied Spectroscopy.* 58 (2004).

- [65] A. Pawlukojs, J. Leciejewicz, A.J. Ramirez-Cuesta, J. Nowicka-Scheibe. *Spectrochimica Acta Part A* 61 (2005).
- [66] F. Groot, T. B. H. Geijtenbeek, R. W. Sanders, C. E. Baldwin, M. Sanchez-Hernandez, R. Floris, Y. Kooyk, E. C. de Jong, B. Berkhout<sup>1</sup>, J. Viro, *Mar.* 2005, p. 3009–3015, Vol.79.
- [67] T. G. Kanyshkova, V. N. Buneva, G. A. Nevinsky, *Biochemistry (Moscow)*, Vol. 66, No. 1, 2001, pp. 1-7.
- [68] B. Waarts, O. J.C. Aneke, J. M. Smit, K. Kimata, R. Bittman, D. K.F. Meijer, J. Wilschut, *Virology* 333 (2005) 284–292.
- [69] B.W.A. van der Strate, L. Beljaars, G. Molema, M.C. Harmsen, D.K.F. Meijer, *Antiviral Research* 52 (2001) 225–239.
- [70] B. Berkhout, R. Floris, I. Recio, S. Visser, *BioMetals* 17: 291–294, 2004.
- [71] J. Filik, N. Stone, *Analyst*, 2007, 132, 544–550.
- [72] E. W. Ainscough, A. M. Brodie, J. E. Plowman, S. J. Bloor, J. S. Loehr, T. M. Loehr, *Biochemistry*, 1980, 19 (17), 4072-4079.
- [73] C. L. Day, K. M. Stowell, E. N. Baker, and J. W. Tweedell, *J. Bio. Chem.*, Vol ,267, 13857-1386.
- [74] C. Constantinescu, A. Palla-Papavlu, A. Rotaru, P. Florian, F. Chelu, M. Icriverzi, A. Nedelcea, V. Dinca, A. Roseanu, M. Dinescu, *APSUSC-17576*, 2008.
- [75] A. K. Gupta, A. S. G. Curtis, *Biomaterials* Volume 25, Issue 15, 2004, 3029-3040.
- [76] Bruno Teuwissen, P. L. Masson, P. Osinski, J. F. Heremans, *J. Biochem.* 35, 366-371 (1973).
- [77] Doruker, P, Atilgan, AR & Bahar, I. *Proteins* 40, 512-524, (2000).
- [78] Atilgan, AR, Durrell, SR, Jernigan, RL, Demirel, MC, Keskin, O. & Bahar, I. *Biophys. J.* 80, 505-515, (2001).