Distance-based NMR structure determination and refinement

Feng Cui
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Biophysics Commons, and the Biostatistics Commons

Recommended Citation
Cui, Feng, "Distance-based NMR structure determination and refinement" (2005). Retrospective Theses and Dissertations. 1857. https://lib.dr.iastate.edu/rtd/1857

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
NOTE TO USERS

This reproduction is the best copy available.

UMI®
This is to certify that the doctoral dissertation of
Feng Cui
has met the dissertation requirements of Iowa State University
DEDICATION

To my family and my fiancé Ning
# TABLE OF CONTENTS

**LIST OF TABLES** .......................................................... viii

**LIST OF FIGURES** ......................................................... x

**CHAPTER 1. General Introduction** ..................................... 1

1.1 Introduction ................................................................. 1

1.2 Nuclear Magnetic Resonance (NMR) ..................................... 2

1.2.1 General Principle ....................................................... 2

1.2.2 COSY and NOESY ...................................................... 4

1.2.3 Fundamental problem in NMR structure modeling .................. 5

1.3 Numerical methods for NMR structure determination ................. 5

1.3.1 EMBED algorithm ..................................................... 6

1.3.2 Geometric build-up algorithm ...................................... 8

1.3.3 Simulated annealing-based methods ................................ 10

1.4 Knowledge-based NMR structure refinement .......................... 12

1.4.1 Refinement with solvent ............................................. 12

1.4.2 Database-derived knowledge-based potentials ...................... 13

1.5 Software development for macromolecular modeling ................. 14

1.6 Dissertation organization .............................................. 14

1.7 References ................................................................. 16

**CHAPTER 2. A Knowledge-Based, Structural Bio-informatics Approach**

to the Refinement of NMR-Determined Protein Structures ............... 24

2.1 Abstract ................................................................. 24

2.2 Introduction ............................................................. 24
4.3.3 Comparisons with NMR and X-Ray structures of $PrP^c$ wild types ... 69
4.3.4 Ramachandran plots of residues of loop regions .................. 71
4.3.5 Backbone structure superimposition and side-chain packing of loop regions 72
4.4 Discussion .......................................................... 73
4.5 References .......................................................... 74

CHAPTER 5. A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse and Inexact Distance Data . . 87
5.1 Abstract .............................................................. 87
5.2 Introduction ......................................................... 88
5.3 The algorithm for full sets of exact distances and metrization ........... 91
5.4 The algorithm for sparse sets of inexact distances ....................... 93
5.5 Computational issues .............................................. 94
5.6 Computational results ............................................. 100
5.7 Concluding remarks ............................................... 102
5.8 References .......................................................... 103

CHAPTER 6. MTMM – A MATLAB Toolbox for Macromolecular Modeling 110
6.1 Abstract .............................................................. 110
6.2 Introduction ......................................................... 110
6.3 Distance-based structural modeling .................................. 113
6.4 Potential energy minimization ...................................... 116
6.5 Structural comparison and transformation .......................... 117
6.6 Implementation ..................................................... 119
   6.6.1 Reading and updating PDB files .............................. 119
   6.6.2 DME computation ............................................. 119
   6.6.3 RMSD computation ........................................... 120
   6.6.4 Structural transformation .................................... 121
   6.6.5 Solving the molecular distance geometry problem .......... 124
   6.6.6 Energy minimization functions ............................. 125
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Deviations of distances in NMR-determined structures*</td>
<td>39</td>
</tr>
<tr>
<td>2.2</td>
<td>Incorrect cross-residue inter-atomic distances*</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>RMSD values of the ensembles of refined NMR structures</td>
<td>40</td>
</tr>
<tr>
<td>2.4</td>
<td>RMSD values of refined NMR structures compared to X-ray structures</td>
<td>40</td>
</tr>
<tr>
<td>3.1</td>
<td>Data Sets</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>Numbers of incorrectly formed inter-atomic distances verses affected residue pairs in ensembles of NMR structures*</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>RMSD values of ensembles of refined NMR structures against mean structures</td>
<td>63</td>
</tr>
<tr>
<td>3.4</td>
<td>RMSD values of refined NMR structures against their X-ray structures*</td>
<td>63</td>
</tr>
<tr>
<td>3.5</td>
<td>Ramachandran plots of NMR structures of 1CRP*</td>
<td>64</td>
</tr>
<tr>
<td>3.6</td>
<td>RMSD values of hydrogen atoms, backbone atoms, and non-hydrogen atoms of NMR structures compared with the mean structures*</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Experimental restraints</td>
<td>85</td>
</tr>
<tr>
<td>4.2</td>
<td>Average RMSD from experimental restraints</td>
<td>85</td>
</tr>
<tr>
<td>4.3</td>
<td>Average RMSD from idealized geometries</td>
<td>86</td>
</tr>
<tr>
<td>4.4</td>
<td>Potential energy of different types</td>
<td>86</td>
</tr>
<tr>
<td>4.5</td>
<td>Percentage of residues in different Ramachandran plot regions</td>
<td>86</td>
</tr>
</tbody>
</table>
Table 6.1  Wall clock times and floating-point operations of selected MTMM routines for protein 1HMV (with 4200 atoms) on a PC Pentium III .......................... 136
## LIST OF FIGURES

| Figure 2.1 | Sample distance distribution functions. The graph on the left is for the distance between the two $C_\alpha$ atoms in two ALA residues separated by one residue, and the one on the right is for the distance between the two backbone atoms $N$ and $C$ in two adjacent ALA residues. | 35 |
| Figure 2.2 | Example acceptance rates of refined NMR structures. The graphs show the acceptance rates for two ensembles of NMR structures, 1E8L on the left and 1IGL on the right, refined with (green line) and without (blue line) using database distance constraints. | 36 |
| Figure 2.3 | Example residue-residue comparisons between refined NMR and X-ray structures. The graphs show the residue RMSD values for an accepted structure (left) and an averaged and energy minimized structure (right) of 2IGG refined with (red line) and without database distance constraints (blue line). | 37 |
| Figure 2.4 | NMR and X-ray crystal structures of 2IGG. The NMR structures are refined with (green line) and without (red line) using additional database distance constraints. They are compared against the structure determined by X-ray crystallography (blue line). | 38 |
| Figure 3.1 | Example acceptance rates of refined NMR structures. The graphs show the acceptance rates for the ensembles of structures of (a) 1E8L and (b) 1IGL, refined with NMR_NOE and NMR_TOR constraints (black), NMR_NOE constraints only (blue), and NMR_NOE and DB_DIST constraints (green) of proteins. | 56 |
Figure 3.2  Example residue-residue comparisons between refined NMR and X-ray structures. The graphs show the residue-residue RMSD values for the backbone atoms (N, C, and Ca) of the averaged and energy-minimized structures of 1CRP, obtained using NMR.NOE and NMR.TOR constraints (green), NMR-NOE constraints only (magenta), and NMR.NOE and DB.DIST constraints (blue). .......................... 57

Figure 3.3  Superimposition of NMR and X-ray crystal structures of 1CRP. The averaged and energy-minimized structures of 1CRP, obtained using NMR.NOE and NMR.TOR constraints (blue) and NMR.NOE and DB.DIST constraints (red), are superimposed to the X-ray structure 1IAQ Chain A (green). .......................... 58

Figure 3.4  Ramachandran plot of the averaged and energy-minimized structure of 1CRP, obtained using NMR.NOE and NMR.TOR constraints. The plot was generated using PROCHECK. .......................... 59

Figure 3.5  Ramachandran plot of the averaged and energy-minimized structure of 1CRP, obtained using NMR.NOE and DB.DIST constraints. The plot was generated using PROCHECK. .......................... 60

Figure 3.6  Example distance distributions. (a) Distribution of distances between atom C in residue ARG and atom O in residue ILE. The residues are adjacent. (b) Distribution of distances between atom Cβ in ALA and atom N in LEU. The residues are separated by one residue. .......................... 61

Figure 4.1  Sample distance distribution functions. (1a) (left) Distances between the Cβ atom of GLU at position i and the Cβ atom of ASP at position i + 1. (1b) (right) Distances between the Cβ atom of ALA at position i and the Cα atom of GLU at position i + 2. .......................... 79
Figure 4.2 Residue-residue comparisons between the refined mutant and wild-type structures. The graphs show the residue RMSD values (with the backbone atoms, $N$, $C_{\alpha}$, $C$, and $O$) for the average and energy-minimized structures of $(E200K)^{NMR+D}$ (magenta line) and $(E200K)^{NMR}$ (green line) when compared against the same types of structures of (a) the wild-type hPrPc NMR structure (1QM0) at mildly acidic condition (pH 4.5); (b) the wild-type hPrPc NMR structure (1HJM) at neutral condition (pH 7.0); (c) the wild-type shPrPc X-ray structure (1UW3) in a monomeric form; (d) the wild-type hPrPc X-ray structure (1I4M) in a dimeric form.

Figure 4.3 Detailed residue RMSD plots for helix and loop regions. The graphs show the detailed residue RMSD values (of the backbone atoms, $N$, $C_{\alpha}$, $C$, and $O$) for the average and energy-minimized structures of $(E200K)^{NMR+D}$ (magenta line) and $(E200K)^{NMR}$ (green line) at (a) N-terminal of helix 2 (residues 172-190) when compared with the hPrPc X-ray structure (1I4M); (b) C-terminal of helix 2 and the loop between helix 2 and helix 3 (residues 191-199) when compared with the shPrPc X-ray structure (1UW3); and (c) helix 3 (residues 201-228) when compared with the hPrPc X-ray structure (1I4M).

Figure 4.4 The $\psi$ and $\phi$ angles of the E200K residues. Those for $(E200K)^{NMR+D}$ are represented by magenta squares and for $(E200K)^{NMR}$ by green squares.
Figure 4.5 **Ramachandran plots.** (a) Ramachandran plot showing the values of ($\psi$, $\phi$) angles of the average and energy-minimized structure of $\langle E200K \rangle^{NMR}$ (green squares); (b) Ramachandran plot of $\langle E200K \rangle^{NMR+D}$ (magenta squares). The residues in the loop between the $\beta$-sheet and helix 2 (residues 167-171) are shown. In (a), a number of the residues are found outside the most favorable (red) regions, while in (b), most of the residues lie in most favorable regions.

Figure 4.6 **Superimposition of tertiary structures.** Tertiary structures of the average and energy-minimized structures of $\langle E200K \rangle^{NMR+D}$ (small magenta cylinders) and $\langle E200K \rangle^{NMR}$ (small green cylinders) are superimposed in (a) the backbone of the whole protein; (b) the backbone of the loop, residues 167-171, between $\beta$-sheet and helix 2; (c) the backbone of the loop, residues 195-199, between helix 2 and helix 3; (d) the backbone and sidechain of the loop between $\beta$-sheet and helix 2 (residues 167-171); (e) the backbone and sidechain of the loop between helix 2 and helix 3 (residues 195-199).

Figure 5.1 **The atom $i$ can be determined with its distances to four base atoms (atoms 1, 2, 3, and 4).**

Figure 5.2 **The coordinates of the atom 3 can be estimated by the four points surrounding it.**

Figure 5.3 **A geometric build-up algorithm for problems with sparse sets of inexact distances**

Figure 5.4 **Flowchart of EMBED algorithm and Geometric Build-Up algorithm for distance data with bounds**

Figure 5.5 **Performance comparison between EMBED algorithms (full metrization and partial metrization) and the geometric build-up algorithm.**
| Figure 5.6 | CPU total time and part time for difference processes of EMBED algorithm | 107 |
| Figure 5.7 | CPU total time and part time for different processes of the geometric build-up algorithm | 108 |
| Figure 5.8 | Comparison of error values of structures generated by EMBED algorithm (4-atom metrization) and the geometric build-up algorithm. The error value is defined in Equation (5.35) | 108 |
| Figure 5.9 | Poly-alanine chains (50mer) generated by EMBED algorithm with 4-atom metrization (cyan) and the geometric build-up algorithm (green) | 109 |
| Figure 6.1 | The fifth atom is determined with the distances between the first four atoms and the fifth atom | 131 |
| Figure 6.2 | The coordinates of the atoms can be determined by a build-up procedure: First, determine the coordinates of atoms 1, 2, 3 and 4 with the distances among them. Then, determine the coordinates of atom 5 (and 6) and the first four atoms. Finally, determine the coordinates of atom 7 with the distances between atom 7 and atoms 1, 3, 5 and 6. | 131 |
| Figure 6.3 | The function is evaluated on a set of points at each temperature of the simulated annealing algorithm. A point is accepted if it satisfies the Metropolis criterion and rejected otherwise. The algorithm converges to the global minimum of the function with probability one as \( n \rightarrow \infty \) and \( T \rightarrow 0 \). | 132 |
| Figure 6.4 | The Cartesian coordinates, inter-atomic distances, bond-angles, and dihedral angles can be defined in terms of each other. | 132 |
| Figure 6.5 | A structure can be transformed from one coordinate system to another by using an appropriate transformation routine. | 133 |
Figure 6.6  Bond angle $\alpha_{ijk}$, $\alpha_{jkl}$, $\alpha_{ikl}$ and dihedral angle $\alpha_{ijkl}$ can be derived from the Cartesian coordinates of the atoms.  

Figure 6.7  The coordinates of the fourth atom $x_l$ can be determined, given the coordinates of $x_i$, $x_j$ and $x_k$, bond angles $\alpha_{ijk}$, $\alpha_{ikl}$ and $\alpha_{jkl}$, and the dihedral angle $\alpha_{ijkl}$.  

Figure 6.8  The Matlab pseudo code for $X = d g \_ s v d(D)$. A Matlab function $s v d s$ is used for the singular-value decomposition.  

Figure 6.9  A pseudo code for the geometric build-up algorithm for solving the molecular distance geometry problem with exact distances, a small linear system needs to be solved in every for-loop.  

Figure 6.10  A pseudo code for simulated annealing energy minimization; the temperature is lowered in 20 steps.
CHAPTER 1. General Introduction

1.1 Introduction

X-ray crystallography\(^1\) and nuclear magnetic resonance (NMR)\(^2\) spectroscopy are two widely used experimental techniques for protein structure determination. In Protein Data Bank (PDB)\(^3-4\), about 85% of deposited protein structures are determined by X-ray crystallography. The rest of the structures are determined by NMR spectroscopy. The main difference between these two approaches lies in the state of protein samples to which they are applied: for X-ray crystallography, a protein has to be in the crystalline state while for NMR, it may be in the solution state.

Both approaches have their own pros and cons. For example, X-ray crystallography is a mature technique capable of providing more objective interpretation of data. This approach has various quality indicators such as resolution and R-factor\(^5-6\) to assess the structures. It can be applied to large molecules, \textit{e.g.}, virus particles, and produce a single model that is easy to visualize and interpret. Raw data processing is highly automatic. In contrast, NMR is a relatively new technique and provides more subjective interpretation of the data. It lacks established quality indicators of data and models. In addition, it is limited to determination of relatively small proteins (<20kDa) and produces an ensemble of possible structures rather than one model. Data sometimes have to be manually processed.

One the other hand, a protein has to form stable crystals for X-ray analysis, which could be time-consuming and often impossible. The crystalline state is not a natural and physiological environment for the protein either. In addition, X-ray crystallography is less useful for large flexible modular proteins. In contrast, the solution state of a protein is closer to biological conditions and relatively easy to prepare. NMR can provide information on dynamics and
identify individual side-chain motion, often used to monitor conformational change on ligand binding \(^7\)–\(^8\).

With the pros and cons, both approaches have undergone dramatic development during the past five years, especially for NMR \(^9\). Advances in data collection, spectra assignment and analysis, structure calculation and computer graphics bring no barrier among NMR spectra assignment process, NMR structure assessment and visualization. Many quality indicators such as bond length, angle and NOE violations (inter-atomic distances that lie outside of NOE ranges) have been developed and used for quality assessment of NMR structures \(^10\). Novel refinement schemes aimed at increasing the accuracy of the resulting structures have been proposed and tested \(^11\)–\(^14\). As a result, nowadays, proteins in size up to 30 kDa (about 260 residues) are routinely accessible by NMR spectroscopy with increased resolution, equivalent to approximately 2.5-Å resolution crystal structures \(^15\)–\(^16\). The research presented in following chapters constitutes part of this emerging effort.

The General Introduction section is organized as follows. NMR general principles as well as COSY (correlated spectroscopy) and NOESY (nuclear Overhauser enhancement spectroscopy) will be introduced, followed by numerical methods for NMR structure determination and for knowledge-based NMR structure refinement. Topics include molecular dynamics, EMBED, geometric build-up algorithms, structure refinement with solvent and with mean-force potential, etc. Lastly, organization for this dissertation will be described.

1.2 Nuclear Magnetic Resonance (NMR)

1.2.1 General Principle

Nuclei are tiny units with positive charges. Some of them have odd-numbered masses (e.g. \(^1\)H, \(^11\)B, \(^13\)C, \(^15\)N, \(^19\)F, \(^31\)P, etc) and behave as though they are spinning. Any spinning charged particle possesses a magnetic moment and generates a magnetic field. If this particle, say, proton (\(^1\)H), is placed in an external magnetic field, it will rotate around an axis that is either parallel or anti-parallel to the direction of the applied field. Different spin states correspond to upper and lower energy levels. For example, a proton spinning in a direction
parallel to that of the external field possesses lower energy while the energy is higher if it spins in the opposite direction. The difference between these two energy levels (ΔE) depends upon the particular nucleus type and the strength of the field (B₀) that the nucleus is immersed in. It is given by

$$\Delta E = \gamma h B_0 / 2\pi$$  \hspace{2cm} (1.1)

where $\gamma$ is called gyromagnetic ratio, a characteristic constant for a particular nucleus and $h$ is Planck’s constant ($6.63 \times 10^{-27}$ erg sec).

Under thermal equilibrium, the spin states of a nucleus (or many nuclei) are stable. Some spins can occasionally “flip” from one orientation to another. If energy equal to the energy difference between two spin states (ΔE) is applied to and absorbed by the spin system, more spins “flip” from one state to another, in other words, from upper energy level to lower energy level as well as from lower energy level to higher energy level. This phenomenon is called nuclear magnetic resonance (NMR) and can be detected as a free induction decay (FID).\(^{17}\)

If an ensemble of proton nuclei (e.g. a protein) is placed in an external field, each nucleus may not experience the exact same magnetic field (e.g., the applied field) because the local magnetic field experienced by each nucleus is affected by small induced fields generated by neighboring atoms. For example, if two proton nuclei are very close in space, the local magnetic field experienced by one proton would be affected by the induced field generated by another proton, leading to a different energy gap between two spin states. As a result, a proton may be activated by a different radiofrequency (RF) radiation depending upon a specific environment of the proton. The difference in RF (referred to as the chemical shift) gives rise to different positions of resonance peaks (signals) in the spectrum. In one-dimensional NMR experiments, the signals are represented as a function of one parameter, which is the chemical shift. In two-dimensional NMR experiments, the signals are represented as a function of two parameters. Both of them are chemical shift ranges, which are plotted in the first and second axes in a 3-D dimension. The third axis represents the magnitude (or intensities) of signals (cross peaks). The intensity is usually indicated using contour lines.
1.2.2 COSY and NOESY

COSY\textsuperscript{18} and NOESY\textsuperscript{19} are two different classes of 2-D NMR experiments. In correlated spectroscopy (COSY), the cross peaks occur between protons that are separated by no more than three covalent bonds. By tracing the bond connectivity, residues of a polypeptide chain can be sequentially identified and assigned. Therefore, COSY conveys structural information "through bond".

In two-dimensional nuclear Overhauser enhancement spectroscopy (NOESY), the cross peaks occur between protons that are close to each other in real space. The NOE signal intensity, $I$, is inversely proportional to the distance between two protons $i$ and $j$, $r_{ij}$, by the equation,

$$I \propto r_{ij}^{-6} \quad (1.2)$$

Experimentally, NOE signals can be detected for distance $r_{ij}$ up to 5 Å. If the distance $r_{ij}$ is out of the 5-Å range, the NOE signals drop below the signal to noise ratio limit and thus cannot be detected. The NOESY experiment correlates all protons which are close in a polypeptide chain. It also correlates protons which are distant in the polypeptide chain but close in space due to tertiary structure. Therefore, NOESY conveys structural information "through space".

However, the equation (1.2) cannot provide an exact measurement of the distance $r_{ij}$ due to the fact that the NOE intensity $I$ can be affected by the internal mobility of a protein\textsuperscript{20}. Therefore, $r_{ij}$ is usually assigned to a rough range based on comparison between detected NOEs and pre-determined empirical NOEs, obtained from pairs of protons with known distance from each other. Typically, "strong" NOEs for which the intensity exceeds the empirical limits correspond to a distance shorter than 3.0 Å, "medium" NOEs to a distance shorter than 3.5 Å and "weak" NOEs to a distance shorter than 5 Å. The lower bounds for all distance ranges are set to 2 Å, corresponding to the sum of the van der Waals radii of two hydrogen atoms. Thus NOESY experiments provide a set of distance ranges for the distances between atoms of a protein. This is the most important information for the determination of protein structures.

It is important to note that the distance data derived from NOE experimental data comprise
an incomplete set of distance data. Long-range distances (e.g. > 5 Å) that are vitally important for determination of protein global conformation are missing.

1.2.3 Fundamental problem in NMR structure modeling

The fundamental problem in NMR structure modeling is stated as follows: Given a set of distance constraints obtained from NMR experiments in general and NOESY experiments in particular, find the coordinates of atoms and hence the structure of the protein satisfying the distance constraints. This problem can be formulated as a mathematical problem called the molecular distance geometry problem\textsuperscript{21-24}. A general molecular distance geometry problem can be stated as follows: Given a set of distances $d_{ij}$ for a pair of atoms $i$ and $j$, find the coordinates $x_i, x_j, \ldots$ for atom $a_i, a_j, \ldots$ such that

$$
\|x_i - x_j\| = d_{ij}
$$

(1.3)

If the distances for all pairs of atoms are given, the problem can be solved in polynomial time, for example, using a singular value-decomposition algorithm\textsuperscript{21} or geometric build-up algorithm\textsuperscript{25-26}. However, if only a subset of distances is given, either exact or inexact, the problem cannot be solved in polynomial time and proven to be \textit{NP-hard}\textsuperscript{27-29}.

1.3 Numerical methods for NMR structure determination

Many numerical methods have been developed for solving the molecular distance geometry problem, namely, the fundamental problem in NMR structure modeling, for example, the EMBED algorithm\textsuperscript{21}, the graph reduction algorithm\textsuperscript{28,30}, the alternating projection algorithm\textsuperscript{31-32}, the multidimensional scaling algorithm\textsuperscript{33-34}, and the global smoothing algorithm\textsuperscript{29, 35-39}. Among them, the EMBED algorithm is one of the most well-developed methods and has been implemented in several public and commercial software packages including DGII\textsuperscript{40}, CNS\textsuperscript{41}, X-PLOR\textsuperscript{42}, and Converter in InsightII (Accelrys)\textsuperscript{43} for NMR structure determination and chemical structure generation. However, this algorithm is not very efficient in terms of computing time. Some key steps of this algorithm such as bound smoothing and embedding spend
most of the computing time. Replacing these time-consuming steps with a geometric build-up module greatly reduces the computing time and makes the process of NMR structure determination at least 6 times faster (Chapter 5).

However, the EMBED algorithm or the EMBED-related methods have some drawbacks including for example inefficient sampling of the available conformational space, long computational time, etc. Inefficient sampling may not be able to locate the global minimum of a target function that is made up of stereo-chemical (e.g. covalent bond length, angles and improper angles) and experimental restraints (e.g. NOEs). This problem could be partially solved by approaches based on simulated annealing. These approaches incorporated various energy terms for covalent bonds and non-bonded interactions (van der Waals repulsion) with different weights. The whole system is initially heated up to overcome potential energy barriers and then cooled down gradually to locate the global minimum of the target function. During this process, restraints with different weights are applied.

A hybrid distance geometry-dynamical simulated annealing method has been developed. In this method, distance geometry methods provide coordinates for about one of the third atoms with correct folds. Missing atoms are placed close to the atoms from the same residues. The resulting structure is used as a starting structure for a dynamical simulated annealing calculation. This hybrid method has been implemented in CNS and X-PLOR as an alternative method for NMR structure determination.

1.3.1 EMBED algorithm

Recall that in NOESY experiments, only proton-proton distances within 5 Å are detectable. In other words, an incomplete set of distances between atoms is obtainable experimentally. In addition, the given distances are not exact. Instead, they are given in ranges with upper bounds and lower bounds. Mathematically, it has been proven that solving the protein structure so that the given distance ranges are satisfied is \( NP \)-hard. However, if exact distances between every pair of atoms, namely, a complete and exact distance matrix, are given, this problem can be solved in polynomial time. So the idea behind the EMBED algorithm is to prepare
a complete and exact distance matrix from an initial matrix, which is an incomplete matrix with distance ranges.

Both experimental restraints (NOEs) and covalent bond length are incorporated into the initial matrix. The missing distance ranges are computed through a so-called bound smoothing process. The idea is the following. For example, assume distance ranges for the distances $d_{ik}$ and $d_{jk}$ are given (e.g. $u_{ik}$, $l_{ik}$, $u_{jk}$, $l_{jk}$), but not for the distance $d_{ij}$. The upper bound $u_{ij}$ and lower bound $l_{ij}$ on the distance $d_{ij}$ can be calculated via the triangle inequality:

$$d_{ij} \leq d_{ik} + d_{jk}$$

for all triple $i$, $j$, $k$. It follows that if $d_{ik} \leq u_{ik}$ and $d_{jk} \leq u_{jk}$, then we have

$$d_{ij} \leq d_{ik} + d_{jk} \leq u_{ik} + u_{jk}$$

If $u_{ik} + u_{jk} \leq u_{ij}$, then $u_{ij}$ can be replaced by the upper limit $u_{ik} + u_{jk}$ on $d_{ij}$. The expression of $l_{ij}$ can be obtained similarly. If $l_{ij} \leq l_{ik} - u_{jk}$, then $l_{ij}$ can be replaced by $l_{ik} - u_{jk}$. Exhaustive application of these relations leads to tightest bounds possible on the values of all the distances. This problem could be efficiently solved by using Floyd’s shortest-path algorithm. The complexity of this algorithm is $O(n^3)$, meaning that the amount of computer time it requires can increase as, at most, the cube of the number of atoms $n$.

With the complete matrix of distance bounds, distances could be independently chosen between the given upper bounds and lower bounds with a uniform distribution. However, some chosen distances may not obey the triangle inequality. A process known as metrization ensures that every chosen distance is consistent with earlier choices and the triangle inequality rule is followed. The idea behind metrization is simple. Once a distance is chosen between its upper bound and low bound, the two bounds are set to this chosen value. All other distance ranges are adjusted so that the requirement for the triangle inequality is met. So the next picked distance would be consistent with the earlier choices in terms of the triangle inequality.

With the complete matrix of exact distances, a structure can be calculated to best fit the given distances by eigenvalue methods. This is so-called the embedding process. Assume an $n \times n$ distance matrix $D$ is given, where $n$ is the number of atoms. If $D = XX^T$, where $X$
is the coordinate sets of the atoms, which is an \( n \times 3 \) matrix, then the rank of \( D \) must be at most 3. Since \( D \) is a positive and symmetric matrix, \( D \) is diagonalizable and subject to singular value decomposition. So we have \( D = U\Sigma U^T \), where \( U \) is \( n \times 3 \) orthogonal matrix, \( \Sigma \) is a \( 3 \times 3 \) matrix with the diagonal elements being the largest three eigenvalues of the matrix \( D \). Therefore, the coordinate set \( X = U\Sigma^{1/2} \).

The resulting structures may contain violations to the given restraints or stereo-chemical criteria, which needs to be removed. This involves minimizing an error function that measures the violations of restraints by the coordinates\(^{40} \). The error function is given by

\[
F(X) = \sum_{(i,j)} A_{ij}^2(X) + \sum_{(i,j)} B_{ij}^2(X) + \sum_{(i,j)} C_{ij}^2(X)
\]

where \( A_{ij} \) is used for enforcing the hard-sphere lower bounds, \( B_{ij} \) for satisfying the given lower and upper bounds, and \( C_{ij} \) for maintaining the chirality of a molecule.

### 1.3.2 Geometric build-up algorithm

EMBED algorithm ends up to solve an eigenvalue problem for a huge matrix by singular value decomposition. This method requires at least \( O(n^2 \sim n^3) \), where \( n \) is the number of atoms, meaning that the computing time increases as the square or the cube of the number of atoms \( n \). If \( n \) is very big, say over 10,000, then this method becomes very inefficient. Another method called geometric build-up algorithm has been developed to make the computational cost affordable.

This method is based on a simple relationship between distances and coordinates, i.e., every time when given the coordinates of four atoms and four distances between the four atoms and the fifth atom, which is undetermined, the coordinates of the fifth atom can immediately be determined by solving a small system of algebraic equations (1.7). More specifically, let \( x_1, x_2, x_3 \) and \( x_4 \) be the coordinate vector of the four atoms and the fifth atom, say atom \( i \). Let the distances between the four atoms and the fifth atom be denoted by \( d_{i,j} \) for \( j = 1, 2, 3, 4 \). If the four atoms are not in a plane, the coordinate vector \( x_i \) of atom \( i \) is determined uniquely
by solving the following system of equations,

\[
\begin{align*}
\|x_i - x_1\| &= d_{i,1} \\
\|x_i - x_2\| &= d_{i,2} \\
\|x_i - x_3\| &= d_{i,3} \\
\|x_i - x_4\| &= d_{i,4}
\end{align*}
\]

which can further be reduced to a linear system of equations and solved efficiently in constant time\textsuperscript{25}.

An overall geometric build-up procedure can be constructed as follows. Given a molecule of \(n\) atoms and a full set of exact distances between all pairs of atoms, assume that at least four atoms (base atoms) in the molecule whose coordinates have already been fixed (the coordinates of the base atoms can always be determined if the distances among them are available). Then, examine each of the remaining unfixed atoms to find four fixed atoms such that the distances between any of the four fixed atoms and the unfixed one are given. If such four base atoms are found, the coordinates for the unfixed atom can immediately be determined by using the four distances between the unfixed atom and the base atoms. Iterative application of this process leads to determination of as many atoms as possible, one atom at a time. For each atom, only a linear system of equations needs to be solved, which takes constant time. In other words, the complexity is \(O(1)\). If a structure of \(n\) atoms needs to be solved, this method can solve the whole structure in \(O(n)^{25}\) instead of \(O(n^2 \sim n^3)\) for EMBED algorithm\textsuperscript{21}.

If a sparse set of exact distance is given, the algorithm can be implemented to solve the problem either complete or partially, depending on given distances. In case that the problem can only be solved partially, a partial structure rather than a full structure is produced\textsuperscript{26}. Another issue is that if the distance data is sparse, the initial set of fixed atoms may not be enough to fix all the remaining atoms. Therefore the set of base atoms that are used to determine new atoms has to be changed as the build-up process proceeds. Some atoms determined in the previous step may have to be used to determine the unfixed atoms in the current step. The numerical errors are generated in every step during the process of atom
determination, and accumulated and passed to the next step. As the process continues, the accumulated errors may eventually result in an inaccurate structure. In order to avoid the errors, instead of using the previously calculated coordinates, the coordinates for the base set of atoms are re-determined by using the distances among them whenever they are available. Then this set of coordinates is used to fix a set of new atoms. Once finished, all these atoms are moved back to the original location through appropriate translations and rotations.50

Careful analysis of the complexity of the EMBED algorithm shows that bound smoothing other than embedding is the most time-consuming. A variant of the geometric build-up algorithm is developed and used as a module to replace the bound smoothing and embedding processes in EMBED algorithm. This hybrid approach can significantly reduce the computational cost by at least six times (Chapter 5).

1.3.3 Simulated annealing-based methods

The EMBED algorithm is generally able to provide the correct folding of a protein structure. However, it has its own drawbacks such as inefficient sampling, long computational time, and relatively poor stereochemistry. Inefficient sampling across all conformation space may not be able to locate the global minimum of the target function. As a result, the protein conformation is "trapped" at a local minimum of the target function. To overcome energy barriers, the system usually needs to be heated up, like in restrained molecular dynamics, which is a commonly used method in protein structure determination. However, this method cannot guarantee to provide correct folding of a protein. So, both approaches cannot reliably locate the global minimum of the target function.

A simulated annealing algorithm can be used to search for the global minimum of a function. The algorithm is expected to be able to find the global minimum of the function since physical annealing can often bring a physical system successfully to its ground state.

A physical annealing process starts from a high temperature, and then cools down gradually to the zero temperature where the system reaches its ground state. The process usually proceeds slowly so that at each cooling stage the system has enough time to reach equilibrium.
Otherwise, it is trapped in a local state.

The simulated annealing algorithm mimics this process by considering the function for a global minimization problem as the energy function of a simulated system. In the original description of this algorithm, the Metropolis algorithm was used to simulate the system. Specifically, a temperature parameter is introduced and decreased stage by stage. At each stage, the function values are randomly sampled. When a point within the function domain with a lower function value is found, it is accepted as the current point. Otherwise the point is accepted and rejected randomly using the Metropolis criterion, which depends on the temperature: If the temperature is higher, the probability of accepting the points is also higher. This property allows the algorithm to accept more points at high temperature and gradually settle down at lower temperature to small regions where the point with the lowest function value may be located.

Instead of the Metropolis algorithm, another approach similar to molecular dynamics has been incorporated into the simulated annealing scheme and used to locate the global minimum of a target function.

In this approach, the target function $F_{\text{tot}}$ is defined as follows:

$$F_{\text{tot}} = F_{\text{covalent}} + F_{\text{repel}} + F_{\text{NOE}}$$

where $F_{\text{tot}}$ represents total potential energy, $F_{\text{covalent}}$ for covalent structures, $F_{\text{repel}}$ for non-bonded contacts, and $F_{\text{NOE}}$ for NOE distance restraints.

The $F_{\text{covalent}}$ is used for maintaining correct bond lengths, angles, planes and chirality, and given by

$$F_{\text{covalent}} = \sum_{\text{bonds}} K_b(r - r_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{improper}} K_\phi(\phi - \phi_0)^2 + \sum_\omega K_\omega(\omega - \omega_0)^2$$

The non-bonded interaction is simply described by a single repulsion term, $F_{\text{repel}}$, for computational efficiency. It is given by

$$F_{\text{repel}} = \begin{cases} 0 & \text{if } r \geq r_{\text{min}} \\ k_{\text{edW}}(s^2r_{\text{min}}^2 - r^2)^2 & \text{if } r < r_{\text{min}} \end{cases}$$
The simplification of non-bonded contacts leads to unrealistic treatment of electrostatic and van der Waals (vdW) interactions, which may affect packing and hydrogen bonding. It will be taken care in the subsequent refinement steps.

A square-well potential is used to represent the NOE distance restraints, $F_{\text{noe}}$, which is given by

$$F_{\text{noe}} = \begin{cases} 
    k_{\text{noe}}(r_{ij} - r_{ij}^u)^2 & \text{if } r_{ij} > r_{ij}^u \\
    0 & \text{if } r_{ij}^l \leq r_{ij} \leq r_{ij}^u \\
    k_{\text{noe}}(r_{ij} - r_{ij}^l)^2 & \text{if } r_{ij} < r_{ij}^l
\end{cases}$$

(1.11)

where $r_{ij}^u$ and $r_{ij}^l$ are the values of upper and lower limits of the target distances, respectively.

1.4 Knowledge-based NMR structure refinement

1.4.1 Refinement with solvent

1.4.1.1 Explicit solvent

Because non-bonded interactions are crudely approximated in structure determination steps, great care has to be taken in refinement steps to ensure that non-realistic treatment of non-bonded contacts does not occur. Previous study showed that structure quality (e.g. packing, hydrogen bond donors or acceptors) can be significantly improved if refined in a thin layer of solvents such as water and DSMO$^{58-61}$. After structure calculation by the standard simulated annealing protocol of ARIA 1.2$^{62}$, 25 structures with lowest energies are chosen as starting structures from the ensemble of 100 structures. The structures are immersed in a 7.0 Å $\sim$ 12.5 Å shell of solvent molecules for refinement. A full non-bonded representation including Lennard-Jone vdW and electrostatic interaction from the OPLS force field$^{61}$ or CHARMM$^{59}$ force field is used. A restrained molecular dynamics protocol is applied for the refinement.
1.4.1.2 Implicit solvent

Refinement of NMR structure with explicit solvent molecules can be very costly, especially for big proteins. To reduce the computational cost, implicit solvent models have been developed to incorporate realistic solvent effects into structure refinement\textsuperscript{60,63}. With the models, NMR structures have noticeable improvement on Ramachandran plot\textsuperscript{64}, hydrogen bond donors or acceptors, comparable to those obtained by refinement with explicit waters.

However, both explicit solvent models and implicit solvent models may have effects on the quality of NMR structures only when experimental restraints are not sufficient. Force field has minimal influence on the final structures with sufficient restraints. For these cases, knowledge-based potentials derived from structure databases could be helpful for further refinement.

1.4.2 Database-derived knowledge-based potentials

Deriving protein structure information from databases of high-resolution structures is a well-established approach. The derived information has been used to construct knowledge-based mean force potential terms for many different purposes. For example, Sippl derived mean-force potential terms from inter-atomic distances and used them for fold recognition and structure validation\textsuperscript{65–66}. Miyazawa and Jernigan developed contact potentials between amino acids for studying protein dynamics and threading\textsuperscript{67–68}.

Some derived structure-based information have been used to refine X-ray and NMR structures, for example, inter-atomic distances for X-ray structure refinement\textsuperscript{69}, torsion angles\textsuperscript{13} and hydrogen bonding pattern\textsuperscript{70} for NMR structure refinement. Distributions of these structural properties in a large amount of high-resolution structures are analyzed and used to derive mean force potentials by best-fitting probability curves. However, distance-based information has not been applied to refining NMR structures. Little is known on whether the database-derived distance information can bring NMR final structures closer to corresponding X-ray structures. Here, we derived the most probable ranges for some short-range inter-atomic distances from structural databases and used them as additional constraints for refining NMR structures (Chapter 2). We also examined whether these short-range distances are able to
replace experimentally generated torsion angle restraints, aimed at reducing experimental and labor cost (Chapter 3). Lastly, we applied these derived distances to refining prion, a biologically important protein responsible for the Mad Cow Disease (Chapter 4).

1.5 Software development for macromolecular modeling

Many different software packages have been developed for macromolecular modeling, for example, CNS$^{41}$, X-PLOR$^{42}$, CHARMM$^{71-72}$, AMBER$^{73}$, etc. They are widely used for macromolecular structure determination and refinement, molecular dynamics simulation, and structural analysis. They usually have very complicated infrastructure and include numerous built-in modules that require a lot of parameters. Therefore, modifying some modules or replacing them with newly developed algorithms is extremely difficult, if possible, within such complex software package suites. On the other hand, many scientific computing environments such as Mathematica$^{74}$ or Matlab$^{75}$ do not have direct support for the computation often required in macromolecular modeling, for example, small modules for manipulating structure files in PDB format, for comparing two structures in terms of similarity, etc. As a result, mathematicians or structural biologists have to repeatedly write their own codes to meet their daily modeling needs. To facilitate those modeling professionals, we have developed a Matlab toolbox MTMM as a special computational environment for macromolecular modeling (Chapter 6). Another reason we implement such a toolbox in Matlab is to take advantage of the numerical computing capability of Matlab so that advanced numerical computing tools can be directly accessible to macromolecular modeling.

1.6 Dissertation organization

This dissertation is organized as follows. Chapter 1 is General Introduction, which includes some theoretical and experimental background. Literature review on NMR structure determination approaches and knowledge-based NMR structure refinement is also provided.

Chapter 2 is a manuscript, entitled, “A knowledge-based, structural bioinformatics approach to the refinement of NMR-determined protein structures”. This manuscript has been
submitted to the Journal of Bioinformatics and Computational Biology and is under the second revision.

Chapter 3 is a manuscript, entitled, “Enhancement of torsion angle constraints in NMR structure refinement via database-derived distance constraints”. This manuscript has been submitted to the Journal of Biomolecular NMR and is currently in revision.

Chapter 4 is a manuscript, entitled, “Improvement of under-determined loop regions of human prion protein by database-derived distance constraints”. This manuscript has been submitted to Protein: Structure, Function, and Bioinformatics and is currently under review. Among the co-authors, Feng Cui did most work on method development, some analysis and manuscript preparation. Kriti Mukhopadhyay worked on some data analysis. Dr. Won-Bin Young provided generous help on understanding the materials and preparing the manuscript.

Chapter 5 is a manuscript, entitled, ”A geometric build-up algorithm for solving the molecular distance geometry problem with sparse and inexact distance data”. This manuscript is to submit to Journal of Global Optimization. Among the co-authors, Feng Cui did most work on algorithm implementation, data analysis and manuscript preparation. Qunfeng Dong and Peter Vedell did some related work on algorithm development.

Chapter 6 is a paper, entitled, “MTMM - A Matlab toolbox for macromolecular modeling”, to appear in the Proceedings of 2004 International Conference on Bioinformatics Application, December 2004, Fort Lauderdale, FL. Among the co-authors, Feng Cui did most work on coding and programming and part of manuscript preparation. Tauqir Bibi worked on some related analysis.

Chapter 7 is General Conclusion, which summarizes the results.
1.7 References


40. Havel, T. F. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance.


CHAPTER 2. A Knowledge-Based, Structural Bio-informatics Approach to the Refinement of NMR-Determined Protein Structures

A paper submitted to the Journal of Bioinformatics and Computational Biology

Feng Cui, Robert Jernigan, and Zhijun Wu

2.1 Abstract

It is well known that protein structures determined by NMR (Nuclear Magnetic Resonance Spectroscopy) are not as detailed or as high in quality as those derived from X-ray crystallography and are underdetermined due to the limited distance data available from NMR experiments. With increased numbers of structures being determined, a computational enhancement to the long-standing problem now is possible by utilizing additional distance constraints based on the distributions of the distances derived from a set of high quality protein structures. Here we show that in fact, many inter-atomic distances in the NMR-determined structures can deviate significantly from their average distributions, and while the distances in a particular NMR-determined structure do not have to completely agree with their prior distributions, the structure can indeed be refined when a selected set of distances are confined to their most probable ranges.

2.2 Introduction

The structures determined by NMR (Nuclear Magnetic Resonance Spectroscopy) are not as detailed and accurate as those by X-ray crystallography due to the inadequate distance data available from NMR experiments\(^1\). The uses of NMR-determined structures in such important applications as homology modelling and rational drug design have thus been severely limited.
The distance data can only be obtained from NMR for specific atoms and in most cases, hydrogen atoms and be estimated approximately with a set of lower and upper bounds. As a result, an ensemble of structures, instead of a single unique one, usually is determined for a protein. While the variation of the structures in the ensemble is often considered as a reflection of the flexibility of the structures in solution, it could be misleading since the variation can also occur from structural under-determination.

In order to increase the accuracy of NMR structures, more distance data has been sought by using various techniques. Experimental approaches such as dipolar coupling \(^2\)\(^-\)\(^4\) have been developed, but they are costly in general. Theoretical approaches include techniques to obtain additional conformational constraints from databases of known protein structures such as to derive constraints on dihedral angles based on their distributions in known X-ray structures in structural databases\(^5\).

With the increased number of high-resolution protein structures being determined, many structural properties such as secondary structure motifs, contact patterns, and hydrophobic core formations, have been revealed from their statistical distributions in known protein structures\(^6\). The inter-atomic distances are also subject to certain statistical distributions, depending on the types of the distances. Such distributions have been employed for constructing various statistical potentials for contact determination, inverse folding, structure alignment, and X-ray structure refinement\(^7\)\(^-\)\(^12\).

In this work, the distributions of inter-atomic distances in known protein structures and in particular, in known X-ray structures, are studied and used to extract additional distance constraints for NMR structure refinement. In order to estimate the distributions, a large set of high-resolution protein structures from PDB Data Bank\(^13\) have been utilized. The distances for selected pairs of atoms across one or two residues along the protein backbones (called cross-residue, inter-atomic distances) are sampled to obtain the probability distributions of the distances.

The distribution functions are then used to evaluate a set of NMR structures. The cross-residue inter-atomic distances in each of the structures are compared with their corresponding
distribution functions, and the deviations of the distances from their average distributions (means) are recorded. The results show that many cross-residue inter-atomic distances in the structures deviate significantly from their average distributions. More specifically, in each structure, on average, about 22% of the residue pairs that are separated by at most one residue along the protein backbone are found to have cross-residue inter-atomic distances deviating from their means by more than two standard deviations. While the inter-atomic distances in a particular NMR structure do not have to agree with their distributions in known protein structures completely, the large number of cases having large deviations of the distances from the means suggest that many of the distances may be incorrectly formed due to the lack of proper constraints for the distances in the NMR data.

In order to reduce the errors in the distances and hence improve the NMR structures, the distribution functions for selected cross-residue inter-atomic distances are used to extract probable ranges for the distances. The obtained distance constraints (called database distance constraints) are then applied to refining a set of NMR structures, using the modeling software CNS (Crystallography and NMR System) developed by Brnger and co-workers\textsuperscript{14}. The structures are refined through combining the original NMR distance constraints with additional database distance constraints. The refined structures are compared in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, the RMSD (root-mean-square-deviation) values of the ensembles of structures, the RMSD values of the structures compared with their X-ray crystal structures (for available ones), as well as the remaining distance errors in the structures. The results show that with additional database distance constraints, the numbers of improperly formed inter-atomic distances in the refined structures are significantly reduced, while the RMSD values of the ensembles of structures are reduced and the acceptance rates of the structures are more than doubled, suggesting that protein structures can indeed be determined more accurately and efficiently by combining the distance constraints obtained from NMR experiments with additional distance constraints extracted from known protein structures in structural databases.
2.3 Results

2.3.1 Distributions of cross-residue inter-atomic distances

To estimate the distributions of cross-residue inter-atomic distances of proteins in known protein structures, 2150 X-ray crystal structures with resolution of 2.0 or higher and sequence similarity of 90% or less were downloaded from PDB Data Bank. The distances are specified together with the types of the atom pairs, the types of the residue pairs, and the sequential separations. More specifically, let \( D \) be the distance between two atoms, \( A_1 \) and \( A_2 \) the types of the two atoms, \( R_1 \) and \( R_2 \) the types of the two residues the two atoms are associated with, respectively, and \( S \) the number of residues separating \( R_1 \) and \( R_2 \). Then, the distribution of the distance \( D \) between atoms \( A_1 \) in \( R_1 \) and \( A_2 \) in \( R_2 \) where \( R_1 \) and \( R_2 \) are separated by \( S \) residues can be represented by using a probability distribution function \( P[A_1, A_2, R_1, R_2, S](D) \). In this study, only five different types of atoms are considered: the amide N, \( C_\alpha \), and the carbonyl C and O along the backbone and the carbon \( C_\beta \) in the side chain. Residue types include all twenty different amino acid types. The separation \( S \) is either one or zero. So in total there are \( 5 \times 5 \times 20 \times 20 \times 2 = 20,000 \) possible distance distributions considered. For each set of \( A_1, A_2, R_1, R_2, \) and \( S \), all corresponding distances in the downloaded crystal structures are computed. The distances are collected into bins of uniform distance intervals \([D_i, D_{i+1}]\), where \( D_i = 0.1 \times i \), \( i = 0, 1, \ldots, n \). The distribution function \( P[A_1, A_2, R_1, R_2, S](D) \) for any \( D \) in \([D_i, D_{i+1}]\) is then defined to be the number of distances in \([D_i, D_{i+1}]\) normalized by the total occurrences of distances in all intervals. Two sample graphs for \( P[A_1, A_2, R_1, R_2, S](D) \) are illustrated in Figure 1, one with \( A_1 = C_\beta, A_2 = C_\beta, R_1 = ALA, R_2 = ALA, \) and \( S = 1 \), and another with \( A_1 = N, A_2 = C, R_1 = ALA, R_2 = ALA, \) and \( S = 0 \). These graphs show clearly non-uniform distributions of distances.

2.3.2 Cross-residue inter-atomic distances in NMR structures

The inter-atomic distances for 462 averaged and energy-minimized NMR structures downloaded from PDB Data Bank are examined and compared with their distribution functions as defined and calculated above. The results show that many of these distances have deviations
larger than two standard deviations. For example, the distribution of the distance between \( \text{C}_\beta \)
in ALA and the carbonyl C in ASP separated by one residue is found to have a mean around 7.1 Å and standard deviation equal to 1.05 Å, while the distance between such a pair of atoms across the 20th and 22nd residues in the NMR structure 2GB1 is 4.6293 Å, which is 0.3707 Å smaller than the mean minus two standard deviations. More example cases of distance deviations in 2GB1 are given in Table 1. In fact, in each of 462 NMR structures, similar deviations are found in 2% to 44%, or in an average of 21.98% of the residue pairs that are separated by one and zero residue along the protein backbone. The deviations are not only found among backbone atoms (N, O, C, \( \text{C}_\alpha \)), but also between backbone (N, O, C, \( \text{C}_\alpha \)) and side-chain atoms (C\( \beta \)). In most cases, the residues having such distance deviations are located on exposed parts of the proteins, which is consistent with the fact that the surface residues are usually of high mobility and more difficult to determine by NMR\(^{15} \).

### 2.3.3 Refining NMR structures with database distance constraints

The large deviations of inter-atomic distances in NMR structures from their average distributions in known protein structures are clear indications of modeling errors in NMR structures that are probably due to the lack of proper constraints on the corresponding distances in the NMR data. One possible way to reduce the errors is to confine the distances to their most probable ranges according to their distributions in known protein structures. To test such an approach, the distribution functions for selected cross-residue inter-atomic distances are used to generate a set of bound constraints for the distances, with the lower and upper bounds equal to the mean values of the distances minus and plus twice the standard deviations, respectively. The generated distance bounds are then taken as additional distance constraints to refine a set of NMR structures, including five structures for 1EPH, 1GB1, 1IGL, 2IGG, 2SOB and five for 1CEY, 1CRP, 1E8L, 1ITL, 1PFL. The last five are selected because they have X-ray structures available. The original NMR experimental constraints for the structures are downloaded from BioMagResBank\(^{16} \). The structures are refined using the standard torsion angle dynamic simulated annealing protocol implemented in CNS\(^{14} \). The results obtained with and
without additional statistically derived distance constraints (database distance constraints) are examined on the deviations of selected inter-atomic distances from their average distributions, and compared and assessed in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, the RMSD values of the ensembles of structures, and the RMSD values of the structures compared with their X-ray structures. As summarized in Table 2, after being refined with additional database distance constraints, the numbers of incorrectly distributed cross-residue inter-atomic distances in the structures are clearly reduced. For example, in structure 1GB1, there are 15 residue pairs with 28 cross-residue inter-atomic distances deviating from their average distributions by more than twice the standard deviations, but after being refined with additional database distance constraints, the numbers drop to 11 residue pairs and 14 cross-residue inter-atomic distances.

2.3.3.1 The acceptance rates

Given an ensemble of accepted NMR structures, the acceptance rate for the ensemble of structures is defined as the number of accepted structures divided by the total number of trial structures including the "rejected" ones. Here, the default acceptance criteria in CNS are used, including the bond lengths, bond angles, NOE distances, and dihedral angles restraints. A trial structure is accepted if all these requirements are satisfied. With additional database distance constraints, the acceptance rates of the refined NMR structures become much higher than those of the structures obtained with only original NMR distance constraints. As shown in Figure 2, for protein 1E8L, only 97 structures need to be determined to obtain 50 accepted structures when additional database distance constraints are used, while 223 structures are required if without them. The acceptance rate for protein 1E8L is increased from about 0.25 to more than 0.50. For protein 1IGL, only 29 structures need to be determined to obtain 17 accepted structures if additional database distance constraints are used, while 67 structures are required otherwise. The acceptance rate is increased from about 0.30 to more than 0.60. These increases in efficiency indicate that additional database distance constraints not only help to correct the distance errors in the NMR structures but also improve the performance of
the modeling program for obtaining acceptable ensembles of structures.

**2.3.3.2 The structure ensembles**

The precision of an ensemble of structures determined by NMR usually is measured by the RMSD values of the structures in the ensemble compared with the average structure of the ensemble, and in particular, by the mean and standard deviation of these values\(^{14}\). The precision may be overestimated since the ensemble of structures determined by current modeling software may not necessarily contain the whole range of structures determined by the given distance constraints\(^{17}\). Nevertheless, as shown in Table 3, the means and standard deviations of the RMSD values for the listed ensembles of structures all become smaller after the structures are refined with additional database distance constraints. Note that the RMSD values are calculated in terms of either just backbone atoms or all non-hydrogen atoms. The results are consistent in both calculations.

**2.3.3.3 Comparing NMR structures with corresponding crystal structures**

The refined NMR structures for five proteins (1CEY, 1CRP, 1E8L, 1ITL, 1PFL) are compared with the corresponding X-ray structures for the RMSD values of the pairs of NMR and X-ray structures. Since each protein has an ensemble of NMR structures, the mean and standard deviation of the RMSD values of the member structures are calculated and used as an assessment for the whole ensemble of structures. As shown in Table 4, both means and standard deviations of the RMSD values for the ensembles of structures refined with additional database distance constraints are clearly smaller than those refined without them, indicating strongly that the structures agree more closely with the X-ray structures after being refined with the additional database distance constraints. More detailed residue-residue comparisons for a particular protein 2IGG\(^{18}\) are also demonstrated in Figure 3, where the RMSD values for all corresponding pairs of residues of NMR and X-ray structures are plotted. The two curves in each graph show the residue RMSD values for two NMR structures of 2IGG, one refined with original NMR distance constraints and another with additional database distance constraints,
when they are compared with the corresponding X-ray structure. The graph on the left is for two accepted structures randomly chosen from their corresponding ensembles of structures. The graph on the right is for two averaged and minimized structures. Both graphs demonstrate the differences between NMR structures refined with and without additional database distance constraints, although the differences in the latter graph are not as large as the former. Figure 4 further displays in 3D graphics the differences among three structures determined for 2IGG, one refined with NMR distance constraints only, one with additional database distance constraints, both without being averaged and energy minimized, and one determined by X-ray crystallography. The picture shows clearly that the NMR structure determined by using additional database distance constraints agree with the X-ray structure better in many regions than the one without using additional database distance constraints, especially in loops and tails, where the structure is not well defined by the NMR experimental data\textsuperscript{18}.

2.4 Discussion

The analysis of NMR-determined protein structures by comparing selected cross-residue inter-atomic distances with the distributions of the distances in known protein structures can always provide a statistical estimate of the accuracy of the NMR structures. While some of the deviations of the inter-atomic distances in NMR structures may be attributable to the additional flexibilities of the NMR structures in solution beyond the crystalline state, many of them must originate in modeling errors, as justified indirectly by the higher acceptance rates and smaller RMSD values of the ensembles of structures when selected distances are confined in high probable regions of their distributions. However, how to distinguish the variations of the distances due to the flexibilities of the NMR structures from those caused by modeling errors is not so clear and remains a question to pursue in future studies. Several approaches may be taken to determine the fluctuations of NMR structures, such as based on the order parameters or temperature factors that can be obtained from NMR or X-ray diffraction data, respectively, or using the Gaussian Network Model\textsuperscript{19} or the Normal Mode Analysis\textsuperscript{20}. If the fluctuations of NMR structures can be determined, the structural variations inconsistent with
the fluctuations may be better targeted for refinement.

While a distance constraint can be derived for every selected pair of cross-residue atoms based on the distribution of the distance in known protein structures, not all the constraints are necessary for the refinement of a given NMR structure since some distances may not necessarily be incorrect even if they deviate significantly from their average distributions. In this work, the distance constraints for all pairs of atoms N, Cα, C, O, Cβ in nearby residues along the protein backbone are derived based on their distribution functions. However, only four such constraints (Cβ-Cβ, C-Cβ, N-Cα, O-Cβ) are selected for pairs of neighboring residues and one (Cβ-R-Cβ) for every two separated residues, where R represents the separating residue. In general, the constraints may be most effective for distances or interactions in regions that are under-determined by NMR experimental data.

On the other hand, the atom types used can certainly be extended to include more side-chain atoms and longer-range interactions. In general, the backbone and other non-hydrogen atoms are perhaps most likely to have distances among them disagreeing with their distributions in known protein structures, since the non-hydrogen atoms usually do not have as much distance data available as hydrogen atoms and therefore cannot be determined as directly and accurately. Indeed, some initial test results show that for many NMR structures, the RMSD values of the ensembles of structures compared with the corresponding X-ray structures in terms of all non-hydrogen atoms are much larger than the RMSD values of the ensembles in terms of hydrogen atoms, while the RMSD values of the ensembles in terms of only backbone atoms are in between the two cases (data not shown).

2.5 References


3. Clore, G. M. and Gronenborn, A. M. New methods of structure refinement for macro-


Figure 2.1 Sample distance distribution functions. The graph on the left is for the distance between the two $C_\beta$ atoms in two ALA residues separated by one residue, and the one on the right is for the distance between the two backbone atoms $N$ and $C$ in two adjacent ALA residues.
Figure 2.2  Example acceptance rates of refined NMR structures. The graphs show the acceptance rates for two ensembles of NMR structures, 1E8L on the left and 1IGL on the right, refined with (green line) and without (blue line) using database distance constraints.
Figure 2.3  Example residue-residue comparisons between refined NMR and X-ray structures. The graphs show the residue RMSD values for an accepted structure (left) and an averaged and energy minimized structure (right) of 2IGG refined with (red line) and without database distance constraints (blue line).
Figure 2.4  **NMR and X-ray crystal structures of 2IGG.** The NMR structures are refined with (green line) and without (red line) using additional database distance constraints. They are compared against the structure determined by X-ray crystallography (blue line).
Table 2.1  Deviations of distances in NMR-determined structures*

<table>
<thead>
<tr>
<th>Res No</th>
<th>Res 1 Atom 1</th>
<th>Res No</th>
<th>Res 2 Atom 2</th>
<th>Mean</th>
<th>$2 \times $ STD</th>
<th>$D$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>GLU N</td>
<td>20</td>
<td>ALA C</td>
<td>5.0</td>
<td>0.8</td>
<td>5.94</td>
</tr>
<tr>
<td>20</td>
<td>ALA CB</td>
<td>22</td>
<td>ASP C</td>
<td>7.1</td>
<td>2.1</td>
<td>4.63</td>
</tr>
<tr>
<td>20</td>
<td>ALA CB</td>
<td>22</td>
<td>ASP CA</td>
<td>6.7</td>
<td>1.5</td>
<td>5.09</td>
</tr>
<tr>
<td>20</td>
<td>ALA CB</td>
<td>22</td>
<td>ASP N</td>
<td>5.6</td>
<td>1.3</td>
<td>4.24</td>
</tr>
<tr>
<td>20</td>
<td>ALA CB</td>
<td>22</td>
<td>ASP O</td>
<td>7.8</td>
<td>2.5</td>
<td>3.51</td>
</tr>
<tr>
<td>21</td>
<td>VAL N</td>
<td>22</td>
<td>ASP O</td>
<td>5.9</td>
<td>1.0</td>
<td>4.28</td>
</tr>
<tr>
<td>21</td>
<td>VAL CB</td>
<td>23</td>
<td>ALA CB</td>
<td>7.2</td>
<td>1.6</td>
<td>9.37</td>
</tr>
<tr>
<td>21</td>
<td>VAL CB</td>
<td>23</td>
<td>ALA CA</td>
<td>6.7</td>
<td>1.1</td>
<td>8.19</td>
</tr>
<tr>
<td>21</td>
<td>VAL CB</td>
<td>23</td>
<td>ALA N</td>
<td>5.7</td>
<td>0.9</td>
<td>6.95</td>
</tr>
<tr>
<td>22</td>
<td>ASP CB</td>
<td>23</td>
<td>ALA C</td>
<td>5.4</td>
<td>0.6</td>
<td>4.69</td>
</tr>
</tbody>
</table>

*Shown in the table are sample atomic pairs (Atom 1 and Atom 2) across some of the residues (Res 1 and Res 2) in NMR structure 2GB1 with distances ($D$) deviating more than twice their standard deviations (STD) from their average distributions (Mean) in known protein structures.

Table 2.2  Incorrect cross-residue inter-atomic distances*

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue</th>
<th>DA§</th>
<th>NOE¶</th>
<th>Incorrect Distance / Affect Residue (without Database Distance Constraints)</th>
<th>Incorrect Distance / Affect Residue (with Database Distance Constraints)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1EPH</td>
<td>53</td>
<td>24</td>
<td>6.7</td>
<td>58/25</td>
<td>48/24</td>
</tr>
<tr>
<td>1GB1</td>
<td>56</td>
<td>93</td>
<td>16.5</td>
<td>28/15</td>
<td>14/11</td>
</tr>
<tr>
<td>1IGL</td>
<td>67</td>
<td>11</td>
<td>7.8</td>
<td>83/30</td>
<td>65/28</td>
</tr>
<tr>
<td>2IGG</td>
<td>64</td>
<td>39</td>
<td>7.4</td>
<td>75/31</td>
<td>29/20</td>
</tr>
<tr>
<td>2SOB</td>
<td>103</td>
<td>49</td>
<td>8.0</td>
<td>143/57</td>
<td>74/41</td>
</tr>
</tbody>
</table>

*Numbers of incorrect distances (outside mean ± 2 x standard deviations) versus numbers of affected residue pairs for structures refined with and without database distance constraints; §DA - dihedral angle constraints; ¶NOE - NOE distance constraints per residue.
Table 2.3  RMSD values of the ensembles of refined NMR structures

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue</th>
<th>Distance Data</th>
<th>Means ± Standard Deviations*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Backbone Atoms§</td>
</tr>
<tr>
<td>1EPH</td>
<td>53</td>
<td>NMR + Database</td>
<td>2.04 ± 0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR</td>
<td>1.78 ± 0.40</td>
</tr>
<tr>
<td>1GB1</td>
<td>56</td>
<td>NMR + Database</td>
<td>0.38 ± 0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR</td>
<td>4.50 ± 1.52</td>
</tr>
<tr>
<td>1IGL</td>
<td>67</td>
<td>NMR + Database</td>
<td>3.81 ± 1.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR</td>
<td>2.62 ± 0.85</td>
</tr>
<tr>
<td>2IGG</td>
<td>64</td>
<td>NMR + Database</td>
<td>2.16 ± 0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR</td>
<td>7.25 ± 1.60</td>
</tr>
<tr>
<td>2SOB</td>
<td>103</td>
<td>NMR + Database</td>
<td>5.54 ± 1.77</td>
</tr>
</tbody>
</table>

*The means and standard deviations of the RMSD values of the structure ensembles refined with and without database distance constraints;§RMSD values in terms of backbone atoms;¶RMSD values in terms of all non-hydrogen atoms.

Table 2.4  RMSD values of refined NMR structures compared to X-ray structures

<table>
<thead>
<tr>
<th>NMR ID</th>
<th>X-ray ID</th>
<th>Residue</th>
<th>Means ± Standard Deviations*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Backbone Atoms§</td>
</tr>
<tr>
<td>1CEY</td>
<td>3CHY</td>
<td>128</td>
<td>1.85 ± 0.19</td>
</tr>
<tr>
<td>1CRP</td>
<td>1HAQ.A</td>
<td>166</td>
<td>1.77 ± 0.29</td>
</tr>
<tr>
<td>1E8L</td>
<td>193L</td>
<td>129</td>
<td>2.05 ± 0.22</td>
</tr>
<tr>
<td>1ITL</td>
<td>1RCB</td>
<td>129</td>
<td>2.88 ± 0.76</td>
</tr>
<tr>
<td>1PFL</td>
<td>1FIK</td>
<td>139</td>
<td>1.66 ± 0.07</td>
</tr>
</tbody>
</table>

*The means and standard deviations of the RMSD values for the ensembles of NMR structures compared with their X-ray structures;§Refined with only NMR distance constraints;¶Refined with NMR and database distance constraints.
CHAPTER 3. Enhancement of Torsion Angle Constraints in NMR Structure Refinement via Database Derived Distance Constraints

A paper submitted to the Journal of Biomolecular NMR

Feng Cui, Robert Jernigan, and Zhijun Wu

3.1 Abstract

Protein structures determined by Nuclear Magnetic Resonance (NMR) are not as accurate and detailed as those determined by X-ray Crystallography because of the inadequate distance data available from NMR experiments. The uses of NMR-determined structures in such important applications as homology modeling and rational drug design have thus been severely limited. Here we show that with the increased number of high-quality protein structures determined, additional distance data can actually be derived just based on the distributions of the distances in databases of known protein structures, and the derived distance data can then be used to improve the quality of the NMR-determined structures. We present the results for the refinement of a set of NMR-determined structures by using just a set of short-range distance constraints derived from structural databases and show in particular that the derived distance constraints can be used to enhance and even replace some of the experimental constraints such as the torsion angle constraints obtained from J-coupling experiments.

3.2 Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy has been used as one of main experimental techniques for protein structure determination. The latest Protein Data Bank (PDB)
(Bernstein et al., 1977; Berman et al., 2000) statistics has shown that approximately 3700 out of 25100 protein structures (15% of total entries) deposited in PDB are determined by NMR. One of the advantages of using NMR is that it does not require the protein to be crystallized as X-ray crystallography does. NMR can therefore be applied to modeling a protein in more realistic environments such as in solution. Because of its ability of revealing both structural and dynamical properties of a bio-molecule, NMR has also been used in probing protein-ligand interaction (Shuker et al. 1996; Hajduk et al. 1999) and protein folding (Jonas 2002).

However, compared to the X-ray structures of the same molecules, NMR-determined structures are not as accurate and detailed (Creighton, 1993), and hence their uses in some critical applications such as structure-based drug design are limited. What causes the discrepancies between NMR and X-ray structures and how to improve the NMR-determined structures have been long-standing questions in protein modeling yet to be answered.

Previous analysis showed that considerable numbers of violations in geometrical indicators such as bond length, bond angle, dihedral angle, and planarity existed widely in many NMR-determined structures (Doreleijers et al. 1998). These violations were believed to be associated with the limitation of the refinement protocols and the lack of information content of the NMR experimental data (Doreleijers et al. 1999a; Spronk et al. 2002; Nabuurs et al. 2003; Linge et al. 1999, 2003; Kuszewski et al. 1996). The latter is referred to the fact that the NMR experimental data is not complete and accurate enough to determine the structures exclusively: a considerable number of inter-proton distances within 5 Å are not obtainable (Doreleijers et al. 1999b) and the distances can only be estimated with rough ranges, say <2.5, <3.5, or <5.0 Å (Creighton 1993). Moreover, up to 50% NMR experimental data is redundant and therefore they cannot provide meaningful information (Nabuurs et al. 2003). The information content of the NMR experimental data may be increased through either introducing other types of restraints from, for example, residual dipolar coupling experiments (Tjandra and Bax 1997), or applying some prior knowledge of local conformational properties such as the distributions of ψ/φ angle combinations (Kuszewski et al. 1996). With the increased number of high-quality protein structures determined, additional distance data can also be derived just based on the
distributions of the distances in databases of known protein structures. Cui et al. 2004 showed that NMR-determined protein structures could be improved significantly with the database derived distance constraints, in terms of various NMR modeling criteria (the acceptance rate, the RMSD of the structural ensemble, the comparison with the X-ray crystalline structure, etc.).

In this paper, we investigate more specific functions of the database derived distance constraints for NMR structure determination and in particular, why they can improve NMR structures and what types of experimental constraints they can specifically enhance. We consider the constraints only for short-range distances among a few heavy atoms mostly in the backbone of the protein (as used in Cui et al. 2004). The reason that such constraints can improve the local conformations of NMR-determined structures is clear since they provide additional distance information that may be missing from NMR experiments. However, it is not clear how critical the constraints are for the overall structure determination. For example, the constraints are only for short-range distances and therefore, may help to enhance directly or indirectly the NOE constraints, but not be able to completely replace them, since the NOE constraints do contain distance information for long-range interactions. On the other hand, the J-coupling constraints are for short-range distances and may be enhanced completely by the database derived distance constraints. Indeed, we show that NMR-determined structures can be refined by database derived distance constraints without using original torsion angle constraints (derived from J-coupling), while the same quality of structures can be obtained. More specifically, we first refine a set of NMR structures using original NMR distance data (NMR.NOE) with or without original torsion angle constraints (NMR.TOR) or database-derived distance constraints (DB.DIST), and evaluate the refined structures in terms of their accuracy (agreement with "true" structures) and precision (consistence of the ensemble of the structures) and examine the constraint violations, acceptance rates, as well as the Ramachandran plots of the structures. We then show that the structures refined using the NMR experimental constraints plus the database derived distance constraints can increase accuracy and precision of the structures with fewer distance violations, higher acceptance rates, and sig-
significantly improved Ramachandran plots even when the experimental torsion angle constraints are removed.

3.3 Results

Proteins used in this study are listed in Table 1 with additional information on the sizes of the proteins, the numbers of the NOE constraints per residue, and the numbers of the torsion angle constraints. All the experimental constraints were downloaded from BioMegResBank (BMRB) (Markley and Ulrich 2002; Doreleijers et al. 2003). The proteins whose NMR experimental constraints are CNS or XPLOR compatible (Brünger 1992; Brünger et al. 1998) were selected since we only used CNS in our structure refinement. However, the results and their implications should not be restricted to only such cases.

The distributions of inter-atomic distances between selected heavy atoms (N, O, C, Cα, and Cβ) across two residues either next to each other or separated by one residue were generated from a large set of known high-resolution protein structures, in particular, X-ray structures, downloaded from PDB. All the distributions appeared to have a high probability region around two standard deviations (2σ) from the mean (μ). A set of distance constraints were generated with -2σ and +2σ as the lower and upper bounds for the distances and used for structure refinement.

NMR structures for proteins in Table 1 were refined with original NMR experimental data (NMR_NOE and NMR.TOR constraints), original NMR experimental data without torsion angle constraints (NMR_NOE constraints only), or original NMR experimental data without torsion angle constraints but with additional database derived distance constraints (NMR_NOE and DB_DIST constraints). The ensembles of refined structures were examined, at both atomic and ensemble levels. At the atomic level, large deviations of distances (more than two standard deviations) in the structures were checked. At the ensemble level, the structures were evaluated in terms of their acceptance rates, precision, accuracy, and Ramachandran plots.
3.3.1 Inter-atomic distance deviations

Inter-atomic distances with large deviations (more than two standard deviations) exist widely in the NMR structures in PDB as shown in Cui et al. 2004. While the inter-atomic distances in a particular NMR structure do not have to agree with their distributions in known protein structures completely, many of them appear to be improperly formed due to the lack of further information on the distances from NMR experiments. The number of improperly formed distances can be greatly reduced using database derived distance constraints (DB.DIST) as additional distance constraints to the original NMR experimental constraints (NMR.NOE and NMR.TOR) (Cui et al. 2004).

Table 2 shows the numbers of improperly formed distances and affected residue pairs in the NMR structures for four of the proteins listed in Table 1 refined using NMR.NOE distance constraints with or without NMR.TOR and DB.DIST constraints. The number of improperly formed distances was large when the structure was refined without using the database derived distance constraints, but was reduced significantly afterwards. For example, for protein 2IGG, there were 75 improperly formed distances in 31 residue pairs in the structure refined with both NMR.NOE and NMR.TOR constraints and 78 improperly formed distances in 31 residue pairs when the NMR.TOR constraints were removed. However, the numbers drop to 27 distances in 18 residue pairs when additional DB.DIST constraints were used. The protein 1GB1 is another interesting case. When the structure was refined with NMR.NOE but without NMR.TOR constraints, it became so poor in quality that no trial structure was accepted. However, when DB.DIST constraints were used, the refined structures became acceptable. Both examples suggest that database derived distance constraints can be used to correct the “improperly formed distances” in NMR-determined structures, and they may even replace the function of the original NMR torsion angle constraints.

3.3.2 Acceptance rates

Given an ensemble of accepted NMR structures, the acceptance rate for an ensemble of structures is defined as the number of “accepted” structures divided by the total number of
trial structures including the "rejected" ones. Structures are "accepted" only if they satisfy certain experimental restraints and stereo-chemical criteria built in the refinement software. Therefore, the acceptance rate is determined by the quality of the trial structures. The better the quality of the trial structures, the fewer rejected structures and the higher acceptance rate of the ensemble of structures. Figure 1 shows the numbers of trial structures versus the numbers of accepted structures of the ensembles of structures for proteins 1E8L (Figure 1 (a)) and 1IGL (Figure 1 (b)), refined with NMR.NOE and NMR.TOR constraints (black), NMR.NOE constraints only (blue), and NMR.NOE and DB.DIST constraints (green). For protein 1E8L, only 84 trial structures were required to obtain 50 accepted structures when NMR.NOE and DB.DIST constraints were used, while 223 trial structures were required when NMR.NOE and NMR.TOR constraints were used. The acceptance rate was increased from about 0.25 to more than 0.50 when the torsion angle constraints were replaced by the database derived distance constraints. For protein 1IGL, 54 trial structures were required to obtain 17 accepted structures when NMR.NOE and DB.DIST constraints were used, while 22 trial structures were required when NMR.NOE and NMR.TOR constraints were used. The acceptance rate was increased from about 0.3 to more than 0.8 when the torsion angle constraints were replaced by the database-derived distance constraints.

3.3.3 Precision of refined structures

The precision of an ensemble of NMR structures, measured by the average RMSD value of the member structures compared with the mean of the structures in the ensemble, is one of the quality indicators for NMR structures (Spronk et al., 2003). The smaller the average RMSD value is, the more precisely the structures are determined. Table 3 shows the average RMSD values for the ensembles of NMR structures for four of the proteins listed in Table 1 refined using NMR.NOE distance constraints with or without NMR.TOR and DB.DIST constraints. For example, for protein 1E8L, the average RMSD values in terms of backbone atoms and non-hydrogen atoms of the structures refined with NMR.NOE and NMR.TOR constraints are 2.302Å and 2.971Å respectively, which are comparable to the corresponding values (2.392Å...
and 3.314 Å) for the structures refined with NMR.NOE constraints only. However, when the torsion angle constraints were replaced by the database derived distance constraints, the values were reduced to 1.921 Å and 2.673 Å, respectively. Together with the results for 1EPH, 1GB1, and 2IGG, it is clear that the database derived distance constraints appear to be able to improve the precision of the structures even when the torsion angle constraints are removed.

3.3.4 Accuracy of refined structures

The accuracy of an ensemble of NMR structures is measured by the closeness of the structures to a reference X-ray structure of the same molecule (Spronk et al., 2003), i.e., the average RMSD value of the structures in the ensemble compared with the reference X-ray structure. Table 4 shows the average RMSD values for the ensembles of NMR structures for four of the proteins (1CEY, 1CRP, 1PFL, and 2IGG) listed in Table 1 refined using NMR.NOE distance constraints with or without NMR.TOR and DB.DIST constraints. The results show that in general, both torsion angle restraints and database derived distance constraints can help to increase the accuracy of NMR structures, while the former can actually be replaced by the latter without compromising the accuracy of the structures. For example, for protein 1CRP (Kraulis et al. 1994), the average RMSD value of the ensemble of its NMR structures refined with NMR.NOE and NMR.TOR constraints is 1.774 Å (when compared against the reference X-ray structure 1IAQ Chain A). The value is smaller than that (1.8178 Å) refined with NMR.NOE constraints only, but larger than that (1.7226 Å) refined with DB.DIST constraints added while NMR.TOR constraints removed.

A more detailed residue-residue comparison for the structures for 1CRP in terms of the RMSD values of the main-chain atoms (Cα, C, and N) is plotted in Figure 2. Here, the averaged and energy-minimized structures, obtained by using NMR.NOE and NMR.TOR constraints, NMR.NOE constraints only, and NMR.NOE and DB.DIST constraints, are compared with the reference X-ray structure 1IAQ Chain A. The secondary structures are marked on the top of the figures with 'h' for helices, 's' for β-sheets, and 't' for turns. For example, the X-ray structure shows that 1CRP has 5 β-sheets (Residue 2 ~ 9 (s1); 38 ~ 46
(s2); 49 ~ 57 (s3); 77 ~ 84 (s4); 110 ~ 118 (s5)), 5 helices (Residue 15 ~ 25 (h1); 66 ~ 75 (h2);
87 ~ 104 (h3); 127 ~ 138 (h4); 152 ~ 166 (h5)), and 8 turns (Residue 10 ~ 14 (t1); 26 ~ 37
(t2); 47 ~ 48 (t3); 58 ~ 76 (t4); 85 ~ 86 (t5); 105 ~ 109 (t6); 119 ~ 126 (t7); 139 ~ 151 (t8)).

Among the three structures, the one refined with NMR.NOE and DB.DIST constraints has
the smaller RMSD values at residue segments 7 ~ 10, 97 ~ 98, 100 ~ 112, and 142 ~ 150,
which are located around turns t1, t6, and t8, but relatively larger RMSD values at segments
20 ~ 26, 49 ~ 50, 81 ~ 88, and 138 ~ 141 (The values at segment 30 ~ 38 and 56 ~ 66 are
equal to zero because of the absence of the coordinate data in the reference X-ray structure
(Spoerner et al. 2001)). Some of the differences can be easily seen in a 3-D picture (Figure 3),
where the structures refined with NMR.NOE and NMR.TOR constraints (blue) and with
the NMR.TOR constraints replaced with the DB.DIST constraints (red) are superimposed
to the reference X-ray structure (green), with region 97 ~ 98 in h3 marked in blue and region
48 ~ 49 between s2 and s3 marked in orange. The region 97 ~ 98 of the structure refined with
NMR.NOE and DB.DIST constraints appears closer to the X-ray structure than that of
the structure refined with NMR.NOE and NMR.TOR constraints. Interestingly, in Figure
2, most of the regions with differential RMSD values are around one or more glycine residues
(in positions 10, 12, 13, 15, 48, 60, 75, 77, 115, 138, and 151). Because there is no side chain,
glycine is extremely flexible and may therefore be changed more easily by external constraints
like database derived distance constraints. The changes in glycine conformation may in turn
affect the conformations of the neighboring residues, and so forth.

Ramachandron plots of the structures obtained for 1CRP using PROCHECK (Laskowski
et al. 1993; Morris et al. 1992) further show that the improvements in local residue RMSD
values actually lead to better formations of ψ and φ angles along the backbone of the protein,
as shown in Table 4 and Figure 4, 5. For example, residues THR87 and TYR64 appear in
disallowed regions of the Ramachandron plot of the structure refined with NMR.NOE and
NMR.TOR constraints. They appear in a most favored region and a generously allowed
region, respectively, in the plot of the structure refined with NMR.NOE and DB.DIST
constraints. On the other hand, residues VAL7, ILE24, ASN26, GLU49, ALA59, ASN85,
and ASN86 appear in additional allowed regions in the plot of the structure refined with NMR.NOE and NMR.TOR constraints, but they appear in most favored regions of the plot of the structure refined with NMR.NOE and DB.DIST constraints. Furthermore, segments 30 ~ 38 and 56 ~ 66 are poorly determined experimentally due to large-scale motions in both regions (Kraulis et al. 1994). However, with database derived distance constraints, one of the under-determined residues, ASP33, moves to a most favored region of the Ramachandran plot from an additional allowed region of the plot of the structure refined with only NMR.NOE and NMR.TOR constraints.

### 3.4 Discussion

With the increased number of high-resolution protein structures being determined, many structural properties such as secondary structure motifs, contact patterns, and hydrophobic core formations have been revealed from their statistical distributions in known protein structures (Bourne & Weissig 2003). The inter-atomic distances are also subject to certain statistical distributions, depending on the types of the distances. Such distributions have been employed for constructing various statistical potential functions for contact determination, inverse folding, structure alignment, and X-ray structure refinement (Miyazawa & Jernigan 1985, 1996; Rojnuckarin & Subramaniam 1999; Sippl 1990; Sippl & Weitckus 1992; Wall et al. 1999). Cui et al. 2004 proposed to extract additional distance constraints from the distributions of the distances in known protein structures and applied the extracted distance constraints successfully to NMR structure refinement. In this paper, we showed that the database derived distance constraints can in particular enhance the experimental constraints and even replace some of them without affecting the increased precision and accuracy of the refined structures.

The database derived distance constraints we generated are just for short distances (3 ~ 7 Å) among some selected heavy atoms, but they are already strong enough to improve NMR structures significantly, as shown in the study. This is because the experimental distance constraints usually act on hydrogen atoms and can only help determine the positions of the heavy atoms indirectly. The additional database derived distance constraints, acting directly on
selected heavy atoms, are able to determine the positions of the heavy atoms more accurately and hence improve the quality of the overall structures. Indeed, as shown in Table 5, the experimental NMR structures are more accurate in terms of the RMSD values of HN and HA (the hydrogen atoms attached to N and C\textsubscript{\textalpha} atoms) than those of backbone atoms N, C\textsubscript{\textalpha}, and C as well as those of all non-hydrogen atoms, when compared with the reference X-ray structures. For example, for the ensemble of NMR structures for protein 1E8L, the mean RMSD value of HN and HA is 1.475\AA, which is much smaller than that of backbone atoms N, C\textsubscript{\textalpha}, and C (2.302\AA) and that of all non-hydrogen atoms (2.971\AA), although the HN and HA atoms are only one chemical bond away from the backbone atoms N and C\textsubscript{\textalpha}.

Several quality indicators such as the precision and accuracy of the ensemble of structures have been used in NMR modeling (Spronk et al. 2003). In this study, an additional indicator, the acceptance rate of the ensemble of structures, defined as the ratio of the accepted structures and the trial structures including “rejected” structures, has also been used. The acceptance rate of an ensemble of structures reflects not only the efficiency of a refinement process but also the quality of the trial structures since the better quality a trial structure has, the less likely the structure will be rejected. Our work showed that the database derived distance constraints were able to “correct” a large portion of improperly formed or in other words, large deviated distances in the tested NMR structures and therefore increased the acceptance rates of the structures dramatically.

3.5 Methods

3.5.1 Collection of structures and distances

Up to 2150 X-ray structures with resolution of 2.0\AA or higher and sequence similarity of 90% or less were downloaded from PDB for the computation of the distributions of the inter-atomic distances in known proteins. A structure was separated into different segments by alternative locations (symbolized as ALT in PDB files), unknown residue types (UNK), or heterogen atoms (HETATM). The coordinates of the atoms in all the segments were extracted and the distances between selected heavy atoms were calculated.
3.5.2 Calculation of distance distributions

Let D be the distance between two atoms, A1 and A2 the types of the two atoms, R1 and R2 the types of the two residues the two atoms are associated with, respectively, and S the number of residues in between R1 and R2. Then, the distribution of the distance D between atoms A1 in R1 and A2 in R2 with R1 and R2 being separated by S residues can be represented by using a distribution function \( P[A_1,A_2,R_1,R_2,S](D) \). For each set of A1, A2, R1, R2, and S, the corresponding distances in the downloaded structures were collected and grouped into a set of uniformly divided distance intervals \([D_i, D_{i+1}]\), where \( D_i = 0.1 \cdot i \, \text{Å}, \) \( i = 0, 1, \ldots, n \). The function value, \( P[A_1,A_2,R_1,R_2,S](D) \), for any D in \([D_i, D_{i+1}]\), was then defined to be the number of distances in \([D_i, D_{i+1}]\) divided by the number of distances in all the intervals. For each distribution function P, the mean \( \mu \) and standard deviation \( \sigma \) were also calculated and stored. Note also that in this study, only five different types of atoms were considered. They were the amide N, the carbon C\(_\alpha\), and the carbonyl C and O along the backbone and the carbon C\(_\beta\) in the side chain. The residue types included all twenty different amino acid types. The separation S was either one or zero. So there are total 5 * 5 * 20 * 20 * 2 = 20,000 possible distance types.

Two example distance distribution functions are plotted in Figure 6. One of the graphs is for the distance between the atom C in arginine on any position i and the atom O in the isoleucine on position i+1 (Figure 1a), while another one is for the distance between the atom C\(_\beta\) in alanine on any position i and the atom N in leucine on position i+2 (Figure 1b). Both functions appear to be Gaussian-like.

3.5.3 Distance constraints and refinement protocols

For each distribution function P, the mean plus and minus 2 standard deviations were used as the upper and lower bounds for the corresponding distance D. For a protein to be refined, a selected set of distance bounds was generated and stored in the same format as the NOE distance constraints. A standard torsion angle dynamic simulated annealing protocol implemented in CNS was used for structure refinement with default settings for all the parameters.
including the number of simulation steps, the annealing temperatures, the acceptance criteria, the tolerances for the NOE distance constraints, etc.

3.6 References


Figure 3.1 Example acceptance rates of refined NMR structures. The graphs show the acceptance rates for the ensembles of structures of (a) 1E8L and (b) 1IGL, refined with NMR.NOE and NMR.TOR constraints (black), NMR.NOE constraints only (blue), and NMR.NOE and DB.DIST constraints (green) of proteins.
Figure 3.2 Example residue-residue comparisons between refined NMR and X-ray structures. The graphs show the residue-residue RMSD values for the backbone atoms (N, C, and Ca) of the averaged and energy-minimized structures of 1CRP, obtained using NMR.NOE and NMR.TOR constraints (green), NMR-NOE constraints only (magenta), and NMR.NOE and DB.DIST constraints (blue).
Figure 3.3 Superimposition of NMR and X-ray crystal structures of 1CRP. The averaged and energy-minimized structures of 1CRP, obtained using NMR.NOE and NMR.TOR constraints (blue) and NMR.NOE and DB.DIST constraints (red), are superimposed to the X-ray structure 1IAQ Chain A (green).
Figure 3.4 Ramachandran plot of the averaged and energy-minimized structure of 1CRP, obtained using NMR.NOE and NMR.TOR constraints. The plot was generated using PROCHECK.
Figure 3.5  Ramachandran plot of the averaged and energy-minimized structure of 1CRP, obtained using NMR_NOE and DB_DIST constraints. The plot was generated using PROCHECK.
Figure 3.6 Example distance distributions. (a) Distribution of distances between atom C in residue ARG and atom O in residue ILE. The residues are adjacent. (b) Distribution of distances between atom Cβ in ALA and atom N in LEU. The residues are separated by one residue.
Table 3.1  **Data Sets**

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residues</th>
<th>NOE</th>
<th>DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CEY</td>
<td>128</td>
<td>7.2</td>
<td>112</td>
</tr>
<tr>
<td>1CRP</td>
<td>166</td>
<td>19.7</td>
<td>69</td>
</tr>
<tr>
<td>1E8L</td>
<td>129</td>
<td>13.1</td>
<td>110</td>
</tr>
<tr>
<td>1EPH</td>
<td>53</td>
<td>6.7</td>
<td>24</td>
</tr>
<tr>
<td>1GB1</td>
<td>56</td>
<td>16.5</td>
<td>93</td>
</tr>
<tr>
<td>1IGL</td>
<td>67</td>
<td>7.8</td>
<td>11</td>
</tr>
<tr>
<td>1PFL</td>
<td>139</td>
<td>12.9</td>
<td>200</td>
</tr>
<tr>
<td>2IGG</td>
<td>64</td>
<td>7.4</td>
<td>39</td>
</tr>
</tbody>
</table>

*NOE - NOE distance constraints per residue;
 *DA - dihedral angle constraints.

Table 3.2  **Numbers of incorrectly formed inter-atomic distances verses affected residue pairs in ensembles of NMR structures**

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>NMR_NOE+NMR.TOR (Distance/Residue Pair)</th>
<th>NMR_NOE (Distance/Residue Pair)</th>
<th>NMR_NOE+DB_DIST (Distance/Residue Pair)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1EPH</td>
<td>58/25</td>
<td>244/48</td>
<td>55/26</td>
</tr>
<tr>
<td>1GB1</td>
<td>28/15</td>
<td>N/A</td>
<td>16/12</td>
</tr>
<tr>
<td>1IGL</td>
<td>83/30</td>
<td>78/34</td>
<td>68/27</td>
</tr>
<tr>
<td>2IGG</td>
<td>75/31</td>
<td>78/31</td>
<td>27/18</td>
</tr>
</tbody>
</table>

*The averaged and energy-minimized structures, refined with different combinations of distance and torsion angle constraints.
Table 3.3  **RMSD values of ensembles of refined NMR structures against mean structures**

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Distance Data</th>
<th>Means ± Standard Deviations*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Backbone Atoms§</td>
</tr>
<tr>
<td>1E8L</td>
<td>NMR_NOE+NMR_TOR</td>
<td>2.30 ± 0.50</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE</td>
<td>2.39 ± 0.49</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE+DB_DIST</td>
<td>1.92 ± 0.52</td>
</tr>
<tr>
<td>1EPH</td>
<td>NMR_NOE+NMR_TOR</td>
<td>2.04 ± 0.61</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE</td>
<td>2.65 ± 0.64</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE+DB_DIST</td>
<td>2.08 ± 0.51</td>
</tr>
<tr>
<td>1GB1</td>
<td>NMR_NOE+NMR_TOR</td>
<td>0.45 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE+DB_DIST</td>
<td>0.54 ± 0.19</td>
</tr>
<tr>
<td>2IGG</td>
<td>NMR_NOE+NMR_TOR</td>
<td>2.62 ± 0.85</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE</td>
<td>2.54 ± 0.77</td>
</tr>
<tr>
<td></td>
<td>NMR_NOE+DB_DIST</td>
<td>2.32 ± 0.77</td>
</tr>
</tbody>
</table>

*The means and standard deviations of the RMSD values of the NMR structures in each ensemble against the mean of the structures in the ensemble. Each ensemble contains 50 accepted structures. §Shown are RMSD values for backbone atoms; ¶Shown are RMSD values for non-hydrogen atoms.

Table 3.4  **RMSD values of refined NMR structures against their X-ray structures**

<table>
<thead>
<tr>
<th>NMR Structure ID</th>
<th>X-ray Structure ID</th>
<th>NMR_NOE+NMR_TOR (μ ± σ)</th>
<th>NMR_NOE (μ ± σ)</th>
<th>NMR_NOE+DB_DIST (μ ± σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CEY</td>
<td>3CHY</td>
<td>1.85 ± 0.19</td>
<td>1.98 ± 0.29</td>
<td>1.84 ± 0.14</td>
</tr>
<tr>
<td>1CRP</td>
<td>1IAQ_A</td>
<td>1.77 ± 0.29</td>
<td>1.82 ± 0.40</td>
<td>1.72 ± 0.27</td>
</tr>
<tr>
<td>1PFL</td>
<td>1FIK</td>
<td>1.66 ± 0.07</td>
<td>1.67 ± 0.08</td>
<td>1.64 ± 0.09</td>
</tr>
<tr>
<td>2IGG</td>
<td>1FCC.C</td>
<td>1.97 ± 0.79</td>
<td>1.93 ± 0.67</td>
<td>1.83 ± 0.51</td>
</tr>
</tbody>
</table>

*The means (μ) and standard deviations (σ) of RMSD values of NMR structures in each ensemble against their reference X-ray structures. Each ensemble contains 50 accepted structures.
### Table 3.5 Ramachandran plots of NMR structures of 1CRP*

<table>
<thead>
<tr>
<th></th>
<th>NMR.NOE + NMR.TOR</th>
<th>NMR.NOE</th>
<th>NMR_NOE + DB.DIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in Most Favored Regions</td>
<td>111 (74.0%)</td>
<td>121 (80.7%)</td>
<td>118 (78.7%)</td>
</tr>
<tr>
<td>Residues in Additionally Allowed Regions</td>
<td>36 (24.0%)</td>
<td>23 (15.3%)</td>
<td>29 (19.3%)</td>
</tr>
<tr>
<td>Residues in Generously Allowed Regions</td>
<td>1 (0.7%)</td>
<td>2 (1.3%)</td>
<td>3 (2.0%)</td>
</tr>
<tr>
<td>Residues in Disallowed Regions</td>
<td>2 (1.3%)</td>
<td>4 (2.7%)</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Non-Glycine and Non-Proline Residues</td>
<td>150 (100%)</td>
<td>150 (100%)</td>
<td>150 (100%)</td>
</tr>
</tbody>
</table>

*The Ramachandran plots of averaged and energy-minimized structures of 1CRP, refined with different combinations of distance and torsion angle constraints. Each ensemble contains 50 accepted structures.

### Table 3.6 RMSD values of hydrogen atoms, backbone atoms, and non-hydrogen atoms of NMR structures compared with the mean structures*

<table>
<thead>
<tr>
<th>NMR Structure ID</th>
<th>HN and HA</th>
<th>Backbone Atoms</th>
<th>Non-hydrogen Atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E8L</td>
<td>1.48 ± 0.40</td>
<td>2.30 ± 0.50</td>
<td>3.01 ± 0.60</td>
</tr>
<tr>
<td>1EPH</td>
<td>1.76 ± 0.38</td>
<td>2.04 ± 0.61</td>
<td>2.94 ± 0.70</td>
</tr>
<tr>
<td>1GB1</td>
<td>0.33 ± 0.05</td>
<td>0.45 ± 0.12</td>
<td>1.04 ± 0.18</td>
</tr>
<tr>
<td>2IGG</td>
<td>2.35 ± 0.97</td>
<td>2.62 ± 0.85</td>
<td>3.29 ± 0.83</td>
</tr>
</tbody>
</table>

*The means and standard deviations of the RMSD values for hydrogen atoms, backbone atoms and non-hydrogen atoms of NMR structures compared with their mean structures. Each ensemble contains 50 accepted structures.
CHAPTER 4. Improvement of Under-Determined Loop Regions of Human Prion Protein by Database-Derived Distance Constraints

A paper submitted to Protein: Structure, Function, and Bioinformatics

Feng Cui, Kriti Mukhopadhyay, Won-Bin Young, Robert Jernigan, Zhijun Wu

4.1 Abstract

Computational simulations of the conversion from the normal cellular prion (PrPc) to the scrapie prion (PrPSc) are usually based on the structures determined by NMR because of the difficulties in crystallizing prion protein. Due to insufficient experimental restraints, a biologically critical loop region in PrPc (residues 167-171), which is a potential binding site for Protein X, is under-determined. Here, we show that by adding information about distance constraints derived from a database of high-resolution protein structures, this under-determined loop and some other secondary structural elements of the E200K variant of human PrPc can be refined into more realistic structures within an ensemble having improved quality and increased accuracy. We show, in particular, that the ensemble becomes more compact after this refinement and the percentage of residues in the most favourable region of the Ramachandran diagram is increased from the 80 to 85% range in the previously reported structures to about 90% in the refined structures. Our results not only provide significantly better structures of the prion protein and hence would facilitate insights into its conversion in the spongiform encephalopathies, but also demonstrate the strong potential for using databases of known protein structures for structure determination and refinement.
4.2 Introduction

Spongiform encephalopathies, or prion diseases, are a group of neuro-degenerative diseases in mammalian species characterized by a progressive vacuolation of brain tissue, amyloid protein deposits, and astrogliosis\(^1\). Specific examples of the diseases include scrapie in sheep, transmissible mink encephalopathy in mink (TME), chronic wasting disease in mule deer (CWD), bovine spongiform encephalopathy (BSE) in cows, Gerstmann-Sträussler-Scheinker disease (GSS), fatal familial insomnia (FFI), kuru, and Alpers syndrome in humans\(^2\text{-}^3\). The pathogenesis of the diseases is associated with accumulation of the infectious "scrapie" form of prion protein (\(PrP^{sc}\)) in brain tissue, which is transformed from its normal cellular form (\(PrP^c\)) with no detectable covalent modification of the molecule\(^2\text{-}^5\). The two forms of prion protein are distinct in many aspects. Bio-chemically, \(PrP^c\) is soluble and sensitive to proteases K digestion, whereas \(PrP^{sc}\) is highly insoluble and resistant to proteases K digestion. Structurally, \(PrP^c\) is \(-\)helices-rich and soluble, whereas \(PrP^{sc}\) is in \(\beta\)-sheet-rich conformation and insoluble\(^6\text{-}^{11}\). The striking difference in secondary structures implies a major conformational transition from \(PrP^c\) to \(PrP^{sc}\), which has been considered as a key process involved in the pathogenesis of prion diseases. However, this prion-protein-only hypothesis is still not fully validated, and the mechanism of the conformation conversion is still unclear. One of the obstacles for understanding the details of prion conformational conversion is that the \(PrP^{sc}\) sample is hard to purify for biochemical and structural characterization. The cellular and scrapie isoforms of \(PrP\) have also proven difficult for high-resolution spectroscopic or crystallographic study\(^12\). Therefore, high-quality structures of prion protein are urgently needed to provide better insights into its transition process.

As are other membrane glycoproteins, prion protein (\(PrP^c\)) is extremely difficult to crystallize when glycosylated\(^13\). Thus far, only two X-ray structures of \(PrP^c\) were reported\(^14\text{-}^{15}\). Most normal and disease-related variants (e.g. E200K) of the human prion protein (\(hPrP^c\)) have been determined by NMR spectroscopy\(^16\text{-}^{18}\). Due to the lack of NOE (Nuclear Overhauser Effects) restraints\(^16\), one particular loop region that comprises residues 167-171 is under-determined in these NMR structures. This region is believed to be a species-specific
binding site for “Protein X”, which might function as a mediator for the transition from $PrP^c$ to $PrP^Sc^{19-22}$. A structure quality analysis by PROCHECK$^{23-24}$ shows that none of the residues in the critical region fall within the most favored regions of the Ramachandran plot$^{25}$ (see below), suggesting that enhancing the structure in this critical region might be vitally important to elucidate the interaction between prion protein and “Protein X”. An alternative explanation may be that this loop is flexible, and the NMR structure is built from data reflecting some average of these forms that might not actually be a feasible form.

The enhancement of NMR structural conformations can be achieved by adding information such as dihedral angles$^{26}$ and inter-atomic distances$^{27}$ based on statistical analysis of databases of high-resolution protein structures. In particular, it has been shown that the inter-atomic distance constraints could improve NMR structures with increased precision and accuracy$^{27}$, and replace some experimental NMR restraints such as torsion angle restraints without compromising the quality of NMR structures$^{28}$. Moreover, these distance constraints impose literally no extra cost on NMR structure refinement$^{27-28}$. In this work, we use a selected set of inter-atomic distance constraints between heavy atoms derived from databases of high-resolution protein structures as additional constraints to refine the E200K variant of human prion protein$^{16}$. Our results show that the critical loop region between residue 167-171 as well as the whole protein can be significantly improved in terms of the precision and accuracy as well as the Ramachandran plots of the structures by using database-derived distance constraints. It is the first evidence that the distance constraints derived from structural databases can be used to optimize the under-determined regions of a protein, in this case, a prion protein. The results provide significantly better structural information about the prion protein and hence could ultimately provide a better starting point for studies on its conversion in the spongiform encephalopathies. And, generally it can be expected that this approach will be important for improving under-defined NMR structures.
4.3 Results

4.3.1 Deriving distance constraints from structural databases

Inter-atomic distances can be categorized into different types according to their related atom pairs, residue pairs, and sequential residual separations. Different types of distances are subject to different statistical distributions in structural databases, which have been employed to construct various statistical potentials for contact determination, inverse folding, structure alignment, X-ray structure refinement, etc.29–34 A set of X-ray structures with similarity of 90% or less and resolution of 2.0Å or higher was downloaded from Protein Data Bank33 (PDB) and used to obtain the statistical distributions of distances of different types. Only the distances between two heavy atoms (N, C\(\alpha\), C, C\(\beta\), and O) in two residues separated by no more than one residue were considered. The distribution of each type of distances was defined by the occurrences of the distances in a set of distance intervals. Two example distance distributions are illustrated in Figure 1, one for the distances between C\(\beta\) of GLU on position i and C\(\beta\) of ASP on position i + 1 (Figure 1a) and another for the distances between C\(\beta\) of ALA on position i and C\(\alpha\) of GLU on position i + 2 (Figure 1b). Two hundred distance intervals are specified in the horizontal direction. The length of each interval is equal to 0.1 Å. Ordinate values show the frequencies of the distances in the corresponding distance intervals. The means \(\mu\) and standard deviations \(\sigma\) of the distributions have been used to specify the range constraints to be between \(\mu - 2\sigma\) and \(\mu + 2\sigma\) for the corresponding distances.

4.3.2 Refining NMR structures of the hPrP\(^{c}\) E200K variant

The NMR structure of the hPrP\(^{c}\) E200K variant of human prion protein was then refined using database derived distance constraints. Two biologically critical but under-determined loop regions (residues 167-171 and 195-199) were targeted particularly for improvement. NMR experimental data for the hPrP\(^{c}\) C-terminal globular domain (residues 125-228) was downloaded from BioMagResBank34 (BMRB). The structure of the domain was then refined using NMR experimental constraints plus additional database derived distance constraints. The standard torsion angle dynamic simulated annealing protocol implemented in CNS35 was used.
for the refinement. Two ensembles of 50 accepted structures have been collected, one that has used the additional database-derived distance constraints (denoted as (E200K)\textsuperscript{NMR+D}) and one that has not (denoted as (E200K)\textsuperscript{NMR}). The structures and in particular, the lowest energy structures of the obtained ensembles have been evaluated in terms of their agreement with the experimental constraints and optimal covalent geometry, their local and global potential energy, and the overall structure quality. As summarized in Tables 2-5, the average root-mean-square deviations of the structures in the ensembles as well as the root-mean-square deviations of the lowest energy structures in the ensembles from the experimentally specified constraints and optimal covalent geometry are comparable for both (E200K)\textsuperscript{NMR} and (E200K)\textsuperscript{NMR+D} ensembles. The total energy and the energy of improper angle restraints and dihedral angle restraints of (E200K)\textsuperscript{NMR+D} structures are lower than those of (E200K)\textsuperscript{NMR} structures, although not so much. However, the results from PROCHECK on the average and energy-minimized structures and the lowest energy structures of both ensembles show a significantly higher percentage (89.6%) of residues in most favorable regions of the Ramachandran plot of the structures in (E200K)\textsuperscript{NMR+D} than the 85.4% of such residues found in (E200K)\textsuperscript{NMR}. Note that the latter percentage (85.4%) is consistent with what was reported by Zhang et al. for their experimental structures (85.7%) 16. The increase in the percentage of residues in the most favorable regions indicates a clear improvement of the structures coming from the use of the database-derived distance constraints. Since the increases are observable from both the average and energy-minimized structure and the lowest energy structure, the improvements have occurred overall throughout the ensemble of structures. However, previously, such a high percentage of residues in most favorable regions of the Ramachandran plot were observed only in the lowest energy structure of a structural ensemble but not in both the average and energy-minimized structure and the lowest energy structure\textsuperscript{16}.

4.3.3 Comparisons with NMR and X-Ray structures of \textit{PrP}\textsuperscript{C} wild types

The C-terminal domain (residues 125-231) of prion protein (\textit{PrP}\textsuperscript{C}), which consists of three \(\alpha\)-helices, helix 1 (residues 144-153), helix 2 (residues 172-194), and helix 3 (residues 200-
and a short β sheet (residues 129-131 and 161-163), has been determined by NMR under neutral (1HJM) and mildly acidic (1QM0) conditions and also by X-ray Crystallography (1I4M for a dimeric form of hPrPc and 1UW3 for a monomeric form of shPrPc)14-15, 17-18. Residue-residue comparisons of averaged and minimized structures of (E200K)\textsuperscript{NMR+D} and (E200K)\textsuperscript{NMR} with the NMR and X-ray structures (also averaged and energy-minimized) of the PrPc wild types were conducted to justify the improvement of the accuracy of the structures, especially the under-determined loop regions, after the database derived distance constraints were employed.

The residue RMSD values for the average and energy-minimized structures (calculated for the backbone atoms, N, Ca, C, and O) of (E200K)\textsuperscript{NMR+D} and (E200K)\textsuperscript{NMR} compared with 1QM0, 1HJM, 1UW3, and 1I4M are plotted in Figure 2, with magenta for (E200K)\textsuperscript{NMR+D} and green for (E200K)\textsuperscript{NMR} in each of the plots. The secondary structures are indicated along the top of each part of the figure with h representing alpha helix and s beta sheet. The residue RMSD values in the loop regions (residues 167-171 and 195-199) in all these plots are all significantly higher than in the remainder of the structure (> 4 Å), which suggests that the loop regions are relatively more flexible. The helix regions seem more rigid with residue RMSD values around 2 Å. However, in the loop region between the helix 2 and helix 3 (residues 195-199), the residue RMSD values for (E200K)\textsuperscript{NMR+D} are consistently smaller than those for (E200K)\textsuperscript{NMR} (in Figure 2a-d), showing that the database-derived distance constraints modify this loop to be more consistent with the NMR and X-ray structures of other prion variants. Indeed, the loop was poorly determined in the (E200K)\textsuperscript{NMR} case, especially around GLY95, mainly due to not having sufficient NMR data in the region (Table 1), but was improved by the introduction of additional database-derived distance constraints (A similar situation was also previously observed by us in refinement of the NMR structure of Streptococcal protein G (2IGG)\textsuperscript{27}). In the other loop between the β-sheet and α-helix 2 (residues 167-171), (E200K)\textsuperscript{NMR+D} appeared closer to the hPrPc X-ray structure (1I4M) with smaller residue RMSD values than (E200K)\textsuperscript{NMR} (in Figure 2d), although not so obvious to other wide types.

In addition to the under-determined loop regions, differences between (E200K)\textsuperscript{NMR+D}}
and (E200K)\textsuperscript{NMR} were observed in well-defined helix regions (helix 2 and helix 3) as well, as shown in Figure 3. A monomer of the dimeric hPr\textsuperscript{PC} X-ray structure is used as a reference structure because it can be superimposed to the NMR structures refined, especially along the residues 125-190 and 197-225\textsuperscript{12}. The RMSD values of (E200K)\textsuperscript{NMR+D} are slightly smaller than those of (E200K)\textsuperscript{NMR} in the N terminal of helix 2 (residues 172-190) and helix 3 (residues 201-228), showing that the helix regions of (E200K)\textsuperscript{NMR+D} are nearer to the corresponding X-ray structure than those of (E200K)\textsuperscript{NMR} (Figure 3a, 3c). In the loop region between the \(\beta\)-sheet and helix 2 (residues 191-199), (E200K)\textsuperscript{NMR+D} is closer to the sheep Pr\textsuperscript{PC} X-ray structure than (E200K)\textsuperscript{NMR} (Figure 3b) with smaller residue RMSD values. Here, the human Pr\textsuperscript{PC} X-ray structure was not used as the reference structure for this region because it is a switch region connecting the helix 2 to the swapped helix 3 in the hPr\textsuperscript{PC} X-ray structure\textsuperscript{12}. Overall, (E200K)\textsuperscript{NMR+D} has slightly better agreement than (E200K)\textsuperscript{NMR} in both the under-determined loop regions and in well-defined helix regions, when compared against the NMR and X-ray structures of other Pr\textsuperscript{PC} variants.

### 4.3.4 Ramachandran plots of residues of loop regions

To further elucidate the improvement of the under-determined loop regions of E200K, angles \(\psi\) and \(\phi\) of each residue of (E200K)\textsuperscript{NMR} and (E200K)\textsuperscript{NMR+D} were evaluated (Figure 4) and plotted in Ramachandran plots (Figure 5). As shown in Figure 4, the angles \(\psi\) and \(\phi\) of (E200K)\textsuperscript{NMR+D} (magenta square) and (E200K)\textsuperscript{NMR} (green square) are very close to each other, except for the two loop regions (residues 167-171 and 195-199). In the loop region between helix 2 and helix 3 (residues 195-199), the angles \(\psi\) and \(\phi\) of (E200K)\textsuperscript{NMR+D} are closer to those of the mean structure of the 30 best structures (1F07) reported by Zhang et al.\textsuperscript{14} than the \(\psi\) and \(\phi\) angles of (E200K)\textsuperscript{NMR}, where the 30 best structures were selected from 60 calculated structures, which were believed to be the most accurate\textsuperscript{16}. In contrast, the residues of (E200K)\textsuperscript{NMR} in the loop (residues 167-171) lie far outside the most favorable regions of the Ramachandran plot (Figure 5a), and so do the residues of the same loop in the mean structure of the 30 best (1F07) structures as well as in the average and energy-minimized
structure (1FKC) of the ensemble generated by Zhang et al.\(^\text{16}\) However, after the structures were refined with additional database-derived distance constraints, i.e., in the average and energy-minimized structure of (E200K)\(^{NMR+D}\), most of these residues have moved into the most favorable regions of Ramachandran plot (Figure 5b). These are small but important changes.

4.3.5 Backbone structure superimposition and side-chain packing of loop regions

Differences between (E200K)\(^{NMR+D}\) and (E200K)\(^{NMR}\) in \(\psi/\phi\) angles of the loops reflect the different loop conformations between the two structures (Figure 4, 5). The backbones of these two structures can be superimposed, as shown in green and magenta cylinders for (E200K)\(^{NMR}\) and (E200K)\(^{NMR+D}\), respectively in Figure 6a. Figure 6b and 6c shows in greater detail the backbones of the loops (residues 167-171 and 195-199). The loop (residues 167-171) in (E200K)\(^{NMR+D}\) appears quite different from the corresponding region in (E200K)\(^{NMR}\) (Figure 6b). This implies that the database derived distance constraints can actually affect the backbone conformations in regions where experimental restraints are insufficient (Table 1), and in other words, these new constraints applied here do not exert their influence uniformly through the structure, but rather in a localized way to improve the most under-determined parts. The conformations of another loop (residues 195-199) in both structures appear quite similar except at residue GLY95 (Figure 6c). It has been shown that the conformations of glycine and neighboring residues can indeed be improved by using database-derived distance constraints\(^\text{27-28}\). Examination of side-chain packing in the two loop regions show that overall conformations of side-chains are quite similar between (E200K)\(^{NMR+D}\) and (E200K)\(^{NMR}\) (Figure 6d, 6e). No change in either hydrogen bonds or salt bridges was observed. It suggests that the database-derived distance constraints do not affect particularly the side-chain packing in general, which is not so surprising since only the constraints between backbone atoms have been utilized in this study. The impact of introducing further distance constraints in the refinement process could be substantial and is currently under investigation.
4.4 Discussion

Prion diseases can be sporadic (spontaneous), inherited, or transmitted by infectious agents. Some prion diseases in humans, such as the familiar CJD, FFI, and GSS, are inherited and linked to mutations in the PrP<sup>c</sup>-coding gene, PRNP. More than 20 mutations in this gene have been associated with prion diseases<sup>38</sup>. The mutation E200K, where the glutamic acid is substituted by the lysine at residue 200, is the major cause of familiar CJD<sup>39</sup>. The tertiary structure of the variant E200K of human PrP<sup>c</sup> is almost identical to the wild-type prion<sup>16</sup>. This mutation only changes the surface potential of the protein, which might affect the interaction with Protein X or other cellular components and the conversion from PrP<sup>c</sup> to PrP<sup>sc</sup>. On the other hand, the mutation itself cannot lead to the conversion that might require additional modification of the protein<sup>40</sup>. Due to the possible involvement of Protein X in the pathogenic process of the familiar CJD, the binding site between the mutated prion (E200K variant) and Protein X, which is the loop region encompassing residues 167-171, requires more accurate determination. However, the lack of experimental restraints makes this task hard to accomplish<sup>16–18</sup>. To refine the conformation of this critical region, here we employed distance constraints that are derived from databases of high-resolution protein structures. The results showed that the loop regions as well as the overall structure of E200K was all significantly improved by several comparisons.

With the database-derived distance constraints, both loop regions (residues 167-171 and 195-199) in (E200K)<sub>NMR+D</sub> showed more reasonable conformations. Although it is difficult to determine whether the calculated conformation reflects the 'true' conformation of the protein in solution, the comparisons with various NMR and X-ray structures of the wild type prion protein confirmed a convergence between the structures, implying an increase in the accuracy of the structure and in particular of the loop regions where experimental restraints tended to be insufficient (Table 1). The improved structure will afford a better structural understanding of prion proteins and hence facilitate insights into its conversion in the spongiform encephalopathies.
4.5 References


corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR* **26**, 139-146 (2003).


Figure 4.1 Sample distance distribution functions. (1a) (left) Distances between the $C_\beta$ atom of GLU at position $i$ and the $C_\beta$ atom of ASP at position $i + 1$. (1b) (right) Distances between the $C_\beta$ atom of ALA at position $i$ and the $C_\alpha$ atom of GLU at position $i + 2$. 
Figure 4.2 Residue-residue comparisons between the refined mutant and wild-type structures. The graphs show the residue RMSD values (with the backbone atoms, N, Cα, C, and O) for the average and energy-minimized structures of (E200K)^{NMR+D} (magenta line) and (E200K)^{NMR} (green line) when compared against the same types of structures of (a) the wild-type hPrP^c NMR structure (1QM0) at mildly acidic condition (pH 4.5); (b) the wild-type hPrP^c NMR structure (1HJM) at neutral condition (pH 7.0); (c) the wild-type shPrP^c X-ray structure (1UW3) in a monomeric form; (d) the wild-type hPrP^c X-ray structure (1H4M) in a dimeric form.
Figure 4.3 Detailed residue RMSD plots for helix and loop regions. The graphs show the detailed residue RMSD values (of the backbone atoms, $N$, $C_{\alpha}$, $C$, and $O$) for the average and energy-minimized structures of (E200K)$^{NMR+D}$ (magenta line) and (E200K)$^{NMR}$ (green line) at (a) N-terminal of helix 2 (residues 172-190) when compared with the $hPrP^c$ X-ray structure (1I4M); (b) C-terminal of helix 2 and the loop between helix 2 and helix 3 (residues 191-199) when compared with the shPrP$^c$ X-ray structure (1UW3); and (c) helix 3 (residues 201-228) when compared with the hPrP$^c$ X-ray structure (1I4M).
Figure 4.4  The $\psi$ and $\phi$ angles of the E200K residues. Those for $(E200K)^{NMR+D}$ are represented by magenta squares and for $(E200K)^{NMR}$ by green squares.
Figure 4.5  **Ramachandran plots.** (a) Ramachandran plot showing the values of $(\psi, \phi)$ angles of the average and energy-minimized structure of (E200K)$^{NMR}$ (green squares); (b) Ramachandran plot of (E200K)$^{NMR+D}$ (magenta squares). The residues in the loop between the $\beta$-sheet and helix 2 (residues 167-171) are shown. In (a), a number of the residues are found outside the most favorable (red) regions, while in (b), most of the residues lie in most favorable regions.
Figure 4.6 Superimposition of tertiary structures. Tertiary structures of the average and energy-minimized structures of (E200K)\textsuperscript{NMR+D} (small magenta cylinders) and (E200K)\textsuperscript{NMR} (small green cylinders) are superimposed in (a) the backbone of the whole protein; (b) the backbone of the loop, residues 167-171, between \( \beta \)-sheet and helix 2; (c) the backbone of the loop, residues 195-199, between helix 2 and helix 3; (d) the backbone and sidechain of the loop between \( \beta \)-sheet and helix 2 (residues 167-171); (e) the backbone and sidechain of the loop between helix 2 and helix 3 (residues 195-199).
Table 4.1  **Experimental restraints**

<table>
<thead>
<tr>
<th>Residue</th>
<th>NOE</th>
<th>Torsion</th>
<th>H-bond</th>
<th>J-coupling</th>
</tr>
</thead>
<tbody>
<tr>
<td>166</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>167</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>168</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>169</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>170</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>171</td>
<td>21</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>172</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Experimental Restraints in Loop2 (residue 195-199)

<table>
<thead>
<tr>
<th>Residue</th>
<th>NOE</th>
<th>Torsion</th>
<th>H-bond</th>
<th>J-coupling</th>
</tr>
</thead>
<tbody>
<tr>
<td>194</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>195</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>196</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>197</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>198</td>
<td>76</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>199</td>
<td>32</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>27</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

NOE  | Torsion | H-bond | J-coupling |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3157</td>
<td>177</td>
<td>96</td>
<td>44</td>
</tr>
<tr>
<td>Per Res.</td>
<td>29.8</td>
<td>1.7</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Note: Total number of Residue is 106.

*Total distance restraints; † J HNHA-coupling constants.

Table 4.2  **Average RMSD from experimental restraints**

<table>
<thead>
<tr>
<th></th>
<th>$\langle E_{200K} \rangle^{NMR}$</th>
<th>$\langle E_{200K} \rangle^{NMR}$</th>
<th>$\langle E_{200K} \rangle^{NMR+D}$</th>
<th>$\langle E_{200K} \rangle^{NMR+D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distances (Å)</td>
<td>$0.0046 \pm 0.0018$</td>
<td>$0.0030$</td>
<td>$0.0047 \pm 0.0016$</td>
<td>$0.0040$</td>
</tr>
<tr>
<td>Angles (degrees)</td>
<td>$0.1664 \pm 0.0368$</td>
<td>$0.1540$</td>
<td>$0.1589 \pm 0.0340$</td>
<td>$0.1380$</td>
</tr>
<tr>
<td>J-couplings (Hz)</td>
<td>$0.3787 \pm 0.0951$</td>
<td>$0.2470$</td>
<td>$0.2105 \pm 0.0186$</td>
<td>$0.2550$</td>
</tr>
</tbody>
</table>

†Average RMSD ± standard deviations for the ensemble of structures, ‡RMSD for the lowest energy structure in the ensemble.
Table 4.3  Average RMSD from idealized geometries

<table>
<thead>
<tr>
<th></th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR+D}$</th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR+D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond lengths (Å)</td>
<td>0.0014 ± 0.0002</td>
<td>0.0012</td>
<td>0.0014 ± 0.0002</td>
<td>0.0012</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.3128 ± 0.0212</td>
<td>0.2990</td>
<td>0.3108 ± 0.0153</td>
<td>0.3040</td>
</tr>
<tr>
<td>Improper angles (°)</td>
<td>0.2148 ± 0.0236</td>
<td>0.2000</td>
<td>0.2105 ± 0.0186</td>
<td>0.2120</td>
</tr>
</tbody>
</table>

†Average RMSD ± standard deviations for the ensemble of structures, †RMSD for the lowest energy structure in the ensemble.

Table 4.4  Potential energy of different types

<table>
<thead>
<tr>
<th></th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR+D}$</th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR+D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Energy</td>
<td>104.16 ± 24.80</td>
<td>82.08</td>
<td>102.31 ± 23.09</td>
<td>86.30</td>
</tr>
<tr>
<td>Bonds</td>
<td>3.30 ± 1.11</td>
<td>2.45</td>
<td>3.54 ± 1.58</td>
<td>2.70</td>
</tr>
<tr>
<td>Bond angles</td>
<td>46.78 ± 6.67</td>
<td>42.53</td>
<td>46.11 ± 4.92</td>
<td>44.11</td>
</tr>
<tr>
<td>Improper angles</td>
<td>6.78 ± 1.54</td>
<td>5.80</td>
<td>6.49 ± 1.21</td>
<td>6.53</td>
</tr>
<tr>
<td>Van der Waals</td>
<td>34.44 ± 9.44</td>
<td>26.23</td>
<td>31.81 ± 6.97</td>
<td>26.29</td>
</tr>
<tr>
<td>NOE</td>
<td>5.85 ± 4.80</td>
<td>2.11</td>
<td>6.97 ± 6.57</td>
<td>3.60</td>
</tr>
<tr>
<td>Dihedral angles</td>
<td>0.31 ± 0.14</td>
<td>0.26</td>
<td>0.28 ± 0.14</td>
<td>0.21</td>
</tr>
</tbody>
</table>

†Average energy ± standard deviations for the ensemble of structures, †energy for the lowest energy structure in the ensemble (kcal/mol).

Table 4.5  Percentage of residues in different Ramachandran plot regions

<table>
<thead>
<tr>
<th></th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR}$</th>
<th>$t(E200K)^{NMR+D}$</th>
<th>$t(E200K)^{NMR+D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most favorable</td>
<td>85.40%</td>
<td>84.40%</td>
<td>89.60%</td>
<td>88.50%</td>
</tr>
<tr>
<td>Additional allowed</td>
<td>14.60%</td>
<td>14.60%</td>
<td>10.40%</td>
<td>11.50%</td>
</tr>
<tr>
<td>Generously allowed</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Disallowed</td>
<td>0.00%</td>
<td>1.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

†Average percentage of residues in different $(\psi,\phi)$-regions for the average and energy-minimized structure in the ensemble, †the percentage of residues in different $(\psi,\phi)$-regions for the lowest energy structure in the ensemble.
CHAPTER 5. A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse and Inexact Distance Data

A paper to be submitted to the Journal of Global Optimization

Feng Cui, Qunfeng Dong, Peter Vedell and Zhijun Wu

5.1 Abstract

Nuclear magnetic resonance (NMR) experimental data are used to derive a sparse set of distance restraints with upper and lower bounds to model NMR structures. The structure modeling requires structures to be generated not only within a reasonable time period and but also in good accuracy. Current approaches such as EMBED algorithm are considered computationally too expensive to calculate the structure of a protein in median size (about 2000 atoms) and with relative low accuracy. To tackle the efficiency and accuracy issues, here, we report a new geometric build-up algorithm for solving a molecular structure by using sparse and inexact distance data. Our algorithm is based on concepts of previously reported geometric build-up methods, combining the partial metrization method with optimization techniques. The algorithm has been implemented in Fortran 77, integrated into Crystallography and NMR System (CNS), and tested on a set of poly-alanine chains in size up to 2,000 atoms. The results showed that our algorithm successfully generate these structures in less time and better accuracy than EMBED algorithm. It suggests that our algorithm could be used as an efficient approach for NMR structure modeling.
5.2 Introduction

One of the fundamental issues in modern biology is to understand functions of proteins in a cell. The functions of a protein are dictated by its tertiary structure, for example, what molecule the protein could interact with, which reaction it can catalyze, how it binds to its substrate, etc [Branden and Tooze 1991]. Therefore, structural information of a protein is indispensable for fully elucidation of the functions of the protein.

The structures of a protein can be determined by experimental methods such as nuclear magnetic resonance (NMR) spectroscopy and X-ray Crystallography or theoretical calculation such as homology modeling [Bourne and Weissig 2003]. Among these methods, NMR plays an increasingly important role in macromolecular structure modeling. The latest Protein Data Bank (PDB) [Berman et al. 2000] statistics show that about 15% of all the protein structures are determined by the NMR technique. This technique provides not only atomic-level structural information but also dynamical properties of a protein in solution.

The most valuable information that could be extracted from NMR data is distance restraints. These restraints may be derived from nuclear Overhauser enhancements (NOE), or with known knowledge of covalent bond geometry such as bond distance and bond angle. Because of uncertainty of atom position in solution, the distance between two protons is not an exact value. Instead, it can be any value within certain range. On the other hand, the set of distances is not complete because that experimentally, if two protons are more than 5Å apart in space, the distance between them is undetectable. Therefore, only a sparse set of inexact distance restraints can be obtained from NMR experimental data. Based on them, a (or a set of) molecular structure could be generated to satisfy these restraints. Mathematically, the problem of determining the structure(s) of a protein with given distance restraints is called a molecular distance geometry problem [Crippen and Havel 1988; Havel and Snow 1991].

The molecular distance geometry problem was initially introduced and studied by Crippen and Havel [Crippen and Havel 1988]. They proposed a so-called EMBED algorithm to solve this problem with NMR data. A couple of updated versions of EMBED algorithm to improve sampling efficiency and overall running time are reported [Kuszewski, Niles, and Brünger 1992].
Besides the EMBED algorithm, several other algorithmic approaches are developed including the graph reduction algorithm [Hendrickson 1991], the alternating-projection algorithm [Glunt, Hayden, and Raydan 1993], and the global smoothing and continuation algorithm [More and Wu 1996, 1997, 1999].

A novel algorithm called geometric build-up algorithm has been proposed and applied to a special class of the molecular distance geometry problem, the problem with a full set of exact distances, meaning that exact distances between all pair of atoms of a protein are provided [Dong and Wu 2002]. This algorithm is based on a principle in distance geometry, stating that in three-dimensional Euclidean space, any four points that are not on a plane form a metric basis for the space. Any point in this space can be determined by using the distances between the point and the four basis-forming points [Blumenthal 1953]. Applying this principle to protein structure modeling, we first determine four atoms that are not on the same plane and use them as a metric basis. Then we determine every other atom by solving a simple system of algebraic equations, which only needs fixed number of floating point operations. The total computation time is proportional to the number of atoms. In other words, given a full set of exact distances, we can solve the structure in $O(n)$ floating point operations, where $n$ is the number of atoms of the protein.

A similar geometric build-up method is developed to tackle another special class of the problem, the problem with a sparse set of exact distances [Dong and Wu 2003], meaning that only part of all exact distances between atoms are provided. This algorithm is based on the same idea as the previously reported method [Dong and Wu 2002]. In principle, to be determined uniquely, every atom is supposed to have distances with at least four atoms that form a metric basis. Since an atom is not guaranteed to have at least four distances in a sparse distance matrix, a complete structure is therefore not guaranteed to be generated. Commonly, a partial protein structure is reported.

The above two special cases do not reflect the molecular distance geometry problem in reality, the problem with a sparse set of inexact distances. In this problem, to ensure that any atom has enough distances (or distance ranges) and hence could be determined, the "missing"
distances between the atom and basis-forming atoms need to be estimated. A process named partial metrization in EMBED algorithm is developed for this purpose and used to estimate the distances on the level of triangle inequality [Kuszewski, Niles, and Brünger 1992]. As a result, the distances from the selected atoms to all other atom can be obtained. Nevertheless, current EMBED algorithm needs two expensive steps, bound smoothing and embedding, after partial metrization to solve the whole structure, which severely affects the computational performance of the algorithm. They make EMBED algorithm computationally too expensive to determine a protein in a median size (about 2000 atoms). Here, we describe a novel geometric build-up algorithm to tackle this efficiency issue. In short, we replace the expensive bound smoothing and embedding processes with a cheap geometric build-up process. More specifically, we first utilize partial metrization to obtain distance ranges between selected atoms and every other atom on the level of triangle inequality, and then choose four atoms from the selected atoms that are not on the same plane as base atoms. The coordinates of the base atoms and the distances between the base atoms and a new atom are used to determine the new atom. Instead of determining this atom uniquely, an approximate space that the atom might be located is estimated by determining some boundary points surrounding the space. The position of the atom is then refined by solving an optimization problem to make sure that all the restraints on this atom have been satisfied. This process is repeated until all the atoms have been fixed.

We applied our algorithm and EMBED algorithm to a set of poly-alanine chains (10mer ~ 200mer). The results showed that our algorithm outperformed EMBED algorithm with the partial metrization process. Besides, the generated structures have better accuracy than those generated by the EMBED algorithm in terms of their fits to the given distance restraints.

The paper is organized as follows. We first describe the distance geometry problem with a sparse set of exact distances in greater detail in Section 2. We then describe our algorithm in Section 3 and the implementation of the algorithm in Section 4. We then discuss some of our computational results in Section 5. In Section 6, we conclude the paper and discuss some related issues about this algorithm.
5.3 The algorithm for full sets of exact distances and metrization

Our geometric build-up algorithm for is based on the same idea of a linear-time algorithm we previously developed for solving the molecular distance geometry problem with a full set of exact distances [Dong and Wu 2002]. We describe the idea of the algorithm briefly in this section. For more details, readers are referred to [Dong and Wu 2002].

Suppose a molecule has \( n \) atoms and the exact distances between all pairs of atoms are given. We need to find the coordinates of atoms \( x_1, x_2, \ldots, x_n \) such that the distance between atom \( i \) and atom \( j \) is equal to the given value \( d_{ij} \) for all \( i \) and \( j \). This distance constraint between atoms \( i \) and \( j \) can be written in a mathematical formula as follows.

\[
\|x_i - x_j\| = d_{ij}, \quad i, j = 1, \ldots, n
\] (5.1)

Mathematically, the coordinates of atom \( i \), denoted as \( (u_i, v_i, w_i) \), can be obtained by solving a simple set of linear equations. Assume that the coordinates of the base atoms \( x_1, x_2, x_3, \) and \( x_4 \) are

\[
x_1 = (u_1, v_1, w_1)^T
\]
\[
x_2 = (u_2, v_2, w_2)^T
\]
\[
x_3 = (u_3, v_3, w_3)^T
\]
\[
x_4 = (u_4, v_4, w_4)^T
\] (5.2)

The distances between the base atoms and atom \( i \) are denoted as \( d_{i,j} \), where \( j = 1, 2, 3, \) and \( 4 \). Then we have the following equations for \( x_i \)

\[
\|x_i - x_1\| = d_{i,1}
\]
\[
\|x_i - x_2\| = d_{i,2}
\]
\[
\|x_i - x_3\| = d_{i,3}
\]
\[
\|x_i - x_4\| = d_{i,4}
\] (5.3)

The above equations can be reduced to
\[ Ax_i = b_i \]  

(5.4)

where

\[
A = 2 \begin{pmatrix} u_1 - u_2 & v_1 - v_2 & w_1 - w_2 \\ u_1 - u_3 & v_1 - v_3 & w_1 - w_3 \\ u_1 - u_4 & v_1 - v_4 & w_1 - w_4 \end{pmatrix}
\]  

(5.5)

and

\[
b_i = \begin{pmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{pmatrix}
\]

(5.6)

The variable \( x_i \) that includes coordinates \((u_i, v_i, w_i)\) for atom \( i \) can be obtained by solving the above linear system, which needs fixed number of floating point operations. Therefore, the whole molecule with \( n \) atoms can be solved in \( O(n) \) floating point operations. The detailed derivation could be referred to [Dong and Wu 2002].

However, a full set of exact distances is not able to be obtained from NMR experimental data. Instead, only a sparse set of inexact distances is available. These distance ranges are put into an \( n \) by \( n \) matrix, where \( n \) is the number of atoms. Lower bounds are placed into the lower triangle of the matrix, while upper bounds are placed into the upper triangle of the matrix. Missing distance ranges could be estimated via triangle inequalities. A process called metrization ensures any distance being picked up from its range is consistent with those previously picked up on the level of triangle inequality [Crippen and Havel 1988]. Metrization includes full metrization and partial metrization. Full metrization makes every chosen distance in the matrix consistent among one another, while partial metrization only take care the distances from a selected set of atoms to all other atoms and makes sure that these distances
are self-consistent. In this study, we use the partial metrization process to generate distances from a selected set of atom, base atoms, to all other atoms in a protein. A detailed description of the metrization process as well as EMBED algorithm can be found in [Crippen and Havel 1998, Kuszewski et al. 1992].

5.4 The algorithm for sparse sets of inexact distances

As we just describe that, if a full set of exact distances is given, all the atoms of a protein can be uniquely determined in $O(n)$ floating point operations, where $n$ is the number of atoms of the protein. However, this efficient algorithm cannot be directly applied to the molecular distance geometry problem with a sparse set of inexact distances since some distances from a set of base atoms to an unfixed atom may not be provided. Even if missing distance ranges from the base atoms to the atom are estimated via partial metrization on the level of triangle inequality, the atom cannot be uniquely determined because given distance data are not exact. Instead, certain space that the atom might be located in could be estimated by determining some boundary points of the space.

Figure 2 illustrates how to determine the boundary points and to estimate the position of the atom in 2-D space. Let atom 1 be put on the original point. Atom 2 is placed on the x-axis $(l_{12} + u_{12})$ away from the original point. Given the coordinates of atoms 1 and 2, and distance bounds $l_{13}$, $u_{13}$, $l_{23}$, and $u_{23}$, the area that atom 3 might be located can be estimated by determining four boundary points, $3_1$, $3_2$, $3_3$, and $3_4$. The four boundary points, $3_1$, $3_2$, $3_3$, and $3_4$ produced by intersecting lower bound lines and upper bound lines can be determined analytically by solving a small system of algebraic equations (see below). Then, the position of atom 3 can be estimated by taking the geometric center of the above four boundary points. In 3-D space, similarly, 8 boundary points need to be determined for estimating the space that the new atom might be located in. In this case, the atom position cannot be estimated by simply taking the geometric center of the 8 boundary points because it has to satisfy not only distance constraints from the base atoms, but also constraints from other previously determined atoms. Instead, we estimate it by solving a local optimization
problem so that it satisfies all the distance constraints between this atom and all previously determined atoms as much as possible. The boundary points are used as starting points for solving the optimization problem. The geometric center of the solutions of the problem is used to estimate the new atom. This process is repeated for determining every atom of the structure. At certain time, the error might reach a predetermined value, e.g. $10^{-4}$, meaning that the partial structure may not good enough to accommodate more new atoms and needs to be modified. All the determined atoms are relaxed and re-optimized with given distance constraints. An outline of our algorithm for sparse sets of inexact distances are given is Figure 3.

Our algorithm shares some components with the EMBED algorithm such as distance data processing, initial bound smoothing, and partial metrization. The time-consuming components of the EMBED algorithm such as bound smoothing and embedding are replaced with a geometric build-up process outlined in Figure 3. A comparison between both algorithms is outlined in Figure 4.

5.5 Computational issues

We can use any atom, e.g. the first atom of the protein, as the first base atoms. Let $u_1, v_1, w_1$ be the coordinates of atom $1$ (denoted as $x_1$). We put this atom on the origin by setting $u_1 = 0, v_1 = 0, \text{and } w_1 = 0$. Then the second atom $x_2$ can be used as the second base atom and fixed on an axis, e.g. $x$-axis, by setting $u_2 = (l_{12} + u_{12}), v_2 = 0, \text{and } w_2 = 0$, where $l_{12}$ and $u_{12}$ are the lower and upper bounds between atom 1 and atom 2. The third base atom is put into a plane formed by two axes, e.g. the one by $x$-axis and $y$-axis. Therefore, the third coordinate for the atom $w_3$ is set to zero. The other two coordinates, $u_3$ and $v_3$, can be obtained by determining four boundary points around atom 3, which are $x_{31}, x_{32}, x_{33}, \text{and } x_{34}$ (Figure 2). The boundary points can be determined by solving the following linear system,

$$\|x_{31} - x_1\| = l_{13}$$

$$\|x_{31} - x_2\| = u_{23}$$
Specifically, let us use point $31$ as an example. The coordinates $u_{31}$ and $v_{31}$ of the point $x_{31}$ can be obtained by solving

\[
\begin{align*}
    u_{31}^2 + v_{31}^2 &= l_{13}^2 \\
    (u_{31}^2 - u_{23}^2 + v_{21}^2)^2 + v_{31}^2 &= u_{23}^2
\end{align*}
\]  

(5.11)

and therefore,

\[
\begin{align*}
    u_{31} &= (l_{31}^2 - u_{23}^2 + v_{21}^2)^2 / (2u_2) \\
    v_{31} &= \pm (l_{13}^2 - u_{31}^2)^{1/2}
\end{align*}
\]  

(5.12)

Here, $v_{31}$ cannot be zero in order to avoid being on the same line determined by the first two atoms. Since $v_{31}$ can either be positive or negative without affecting the final structure, we always choose $v_{31}$ to be positive. All other three boundary points can be determined in the same way. Thus, the third base atom can be estimated as the geometric center of the four boundary atoms with $u_3 = 1/4 (u_{31} + u_{32} + u_{33} + u_{34})$, $v_3 = 1/4 (v_{31} + v_{32} + v_{33} + v_{34})$, and $w_3 = 0$. Finally, the fourth base atom can be estimated by fixing eight boundary points, $x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{46}, x_{47},$ and $x_{48}$, by solving the following linear system,

\[
\begin{align*}
    \|x_{41} - x_1\| &= l_{14} \\
    \|x_{41} - x_2\| &= l_{24} \\
    \|x_{41} - x_3\| &= l_{34}
\end{align*}
\]  

(5.13)
\[ \| x_{42} - x_1 \| = l_{14} \]
\[ \| x_{42} - x_2 \| = l_{24} \]
\[ \| x_{42} - x_3 \| = u_{34} \]  \hspace{1cm} (5.14)

\[ \| x_{43} - x_1 \| = l_{14} \]
\[ \| x_{43} - x_2 \| = u_{24} \]
\[ \| x_{43} - x_3 \| = u_{34} \]  \hspace{1cm} (5.15)

\[ \| x_{44} - x_1 \| = u_{14} \]
\[ \| x_{44} - x_2 \| = l_{24} \]
\[ \| x_{44} - x_3 \| = l_{34} \]  \hspace{1cm} (5.16)

\[ \| x_{45} - x_1 \| = u_{14} \]
\[ \| x_{45} - x_2 \| = u_{24} \]
\[ \| x_{45} - x_3 \| = l_{34} \]  \hspace{1cm} (5.17)

\[ \| x_{46} - x_1 \| = u_{14} \]
\[ \| x_{46} - x_2 \| = l_{24} \]
\[ \| x_{46} - x_3 \| = u_{34} \]  \hspace{1cm} (5.18)

\[ \| x_{47} - x_1 \| = l_{14} \]
\[ \| x_{47} - x_2 \| = u_{24} \]
\[ \| x_{47} - x_3 \| = u_{34} \]  \hspace{1cm} (5.19)

\[ \| x_{48} - x_1 \| = u_{14} \]
\[ \| x_{48} - x_2 \| = u_{24} \]
\[ \| x_{48} - x_3 \| = u_{34} \]  \hspace{1cm} (5.20)
Specifically for atom $x_41$, its coordinates, $u_{41}$, $v_{41}$, and $w_{41}$, can be obtained by solving

$$u_{41}^2 + v_{41}^2 + w_{41}^2 = l_{14}^2$$

$$(u_{41} - u_2)^2 + v_{41}^2 + w_{41}^2 = l_{24}^2$$

$$(u_{41} - u_3)^2 + (v_{41} - v_3)^2 + w_{41}^2 = l_{34}^2$$

(5.21)

and

$$u_{41} = (l_{14}^2 - l_{24}^2 + u_2^2) / (2u_2)$$

$$v_{41} = (l_{14}^2 - l_{34}^2 - (u_{41} - u_2)^2 + (u_{41} - u_3)^2 + v_3^2) / (2v_3)$$

$$w_{41} = \pm (l_{14}^2 - u_{41}^2 - u_{41}^2)^{1/2}$$

(5.22)

Obviously, $w_{41}$ cannot be zero in order to avoid being in the same plane determined by the first three base atoms. Here $w_{41}$ can either be positive or negative, corresponding to two mirror symmetric structures. We choose one of the structures with $w_{41}$ positive. If one of above points, $j = 1, \ldots, 8$, cannot be determined in real space, the point can be estimated by solving a least square problem. For example, let $x_j$ be the coordinates $(u_j, v_j, w_j)$ of the point $j$ and $x_i$ be the coordinate vector of the $i$th determined atom, $i = 1, 2, 3$, with which the point $j$ has a distance $d_{ij}$ bound by $l_{ij}$ and $u_{ij}$. Then, $x_j$ needs to be determined to satisfy the inequalities,

$$l_{ij} \leq \|x_i - x_j\| \leq u_{ij} \quad i = 1, 2 \text{ or } 3$$

(5.23)

The solution of the inequalities can be easily found by solving the following least square problem,

$$\min \sum_{i=1}^{3} (l_{ij}^2 - \|x_i - x_j\|^2)^+ + (\|x_i - x_j\|^2 - u_{ij}^2)^+ \quad x_j \in R^3$$

(5.24)

where $(\cdot)^+ = (\cdot) \text{ if } (\cdot) \geq 0 \text{ and } (\cdot)^+ = 0 \text{ otherwise.}$ Consequently, two sets of solutions, one above the x-y plane and one below the x-y plane, are obtained, which corresponds to two mirror symmetric structures. Each set includes the coordinates of eight points. The geometric center of the set of eight points above the x-y plane is usually chosen to estimate atom 4.
Given the coordinates of four base atoms and the distance bounds between the base atoms and all other atoms obtained through partial metrization, in principle, these undetermined atoms can somehow be fixed one at a time in a similar way as atom 4. More specifically, suppose we want to determine atom $i$, where $i = 5, \ldots, n$. Let $x_1, x_2, x_3$ and $x_4$ be the coordinates of four base atoms, and $x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}$ and $x_{18}$ be the coordinates of eight boundary points around the atom $i$. The lower bounds and upper bounds between the base atoms and atom $i$ are $l_{1i}, l_{2i}, l_{3i}, u_{1i}, u_{2i}, u_{3i}$ and $u_{4i}$. The eight boundary points can be determined by solving

\begin{align*}
\|x_1 - x_{11}\| &= l_{1i} \\
\|x_1 - x_{12}\| &= l_{2i} \\
\|x_1 - x_{13}\| &= l_{3i} \\
\|x_2 - x_{11}\| &= l_{1i} \\
\|x_2 - x_{12}\| &= l_{2i} \\
\|x_2 - x_{13}\| &= u_{3i} \\
\|x_3 - x_{11}\| &= l_{1i} \\
\|x_3 - x_{12}\| &= u_{2i} \\
\|x_3 - x_{13}\| &= u_{3i} \\
\|x_4 - x_{11}\| &= u_{1i} \\
\|x_4 - x_{12}\| &= l_{2i} \\
\|x_4 - x_{13}\| &= l_{3i} \\
\|x_5 - x_{11}\| &= u_{1i} \\
\|x_5 - x_{12}\| &= u_{2i} \\
\|x_5 - x_{13}\| &= l_{3i}
\end{align*}

(5.25) (5.26) (5.27) (5.28) (5.29)
\[ \begin{align*}
\| x_{16} - x_1 \| &= u_{1i} \\
\| x_{16} - x_2 \| &= l_{2i} \\
\| x_{16} - x_3 \| &= u_{3i} \\
\| x_{17} - x_1 \| &= l_{1i} \\
\| x_{17} - x_2 \| &= u_{2i} \\
\| x_{17} - x_3 \| &= u_{3i} \\
\| x_{18} - x_1 \| &= u_{1i} \\
\| x_{18} - x_2 \| &= u_{2i} \\
\| x_{18} - x_3 \| &= u_{3i}
\end{align*} \] (5.30)

Similarly, two set of solutions with eight points in each, one is above the x-y plane and another is below the x-y plane, are obtained. Atom 4 is used to determine which set would be chosen by comparing their fit to the distance constraints,

\[
\min \left( \sum_{p=1}^{g} \left( \left( l_{4p}^2 - \| x_p - x_4 \|_2^2 \right)_+ + \left( u_{4p}^2 - \| x_p - x_4 \|_2^2 \right)_+ \right), \sum_{q=1}^{s} \left( \left( l_{4q}^2 - \| x_q - x_4 \|_2^2 \right)_+ + \left( u_{4q}^2 - \| x_q - x_4 \|_2^2 \right)_+ \right) \right) \] (5.33)

where \( p \) represents the \( p^{th} \) points in the set above the x-y plane, while \( q \) represents the \( q^{th} \) points in the set below the x-y plane. Once the desired set is chosen, the geometric center of the set is used to estimate the atom \( i \). This position of atom \( i \) only satisfy constraints between atom \( i \) and the base atoms, which may not necessarily satisfy constraints between atom \( i \) and determined atoms other than base atoms. To make sure atom \( i \) satisfy all the constraints, the position of atom \( i \) might be adjusted by solving

\[
\min \sum_{i=1}^{k} \left( \left( l_{ij}^2 - \| x_i - x_j \|_2^2 \right)_+ + \left( \| x_i - x_j \|_2^2 - u_{ij}^2 \right)_+ \right) \quad x_j \in \mathbb{R}^3
\] (5.34)

where \( i \) is the \( i^{th} \) atom in a set of \( k \) determined atoms having distance constraints with atom \( j \). In this system, only \( x_j \) is treated as a variable and can be solved very efficiently. Therefore,
we call this problem the local least-squares problem. A cutoff value to stop the optimization process is pre-determined, in our case, $10^{-4}$. At certain time, the error of the system cannot be lower than the cutoff value, meaning that the partial structure we have may not good enough to accept new atoms and needs to be modified. We have to relax and adjust the structure such that the new atoms become well accommodated. The problem (5.34) needs to be reformulated such that all coordinate vector $x_i$, $i = 1, \ldots, k$, and $x_j$ are treated as variables. We call this problem the global least-squares problem. Mathematically, this problem is hard to be solved. However, in our case, it is relatively easy because we have a good starting structure, the previous partial structure. Moreover, this problem only needs to be solved occasionally, which would not affect the overall runtime efficiency too much.

5.6 Computational results

We implemented our algorithm in Fortran 77 and integrated into Crystallography and NMR System (CNS) [Brünger et al. 1998]. A set of poly-alanine chains (10 ~ 200mer) have been used as test cases for both EMBED algorithm and our algorithm. The covalent bond information of the poly-peptide chain is automatically provided by CNS. The missing distance constraints are estimated via the partial metrization process. EMBED algorithms with the full metrization process or partial metrization process is compared with the geometric build-up algorithm in runtime performance (Figure 5). The results show that the EMBED algorithm with the partial metrization process is computationally cheaper than that with the full metrization process, which is consistent with the previous study (Kuszewski, Nilges, and Brünger 1992). Our algorithm is more efficient than the EMBED algorithms with either metrization process. In particular, in case of 200mer poly-alanine chain, our algorithm is more than six times faster than the EMBED algorithm with the partial metrization process and more than thirteen times faster than the EMBED algorithm with the full metrization process.

Detailed analysis of CPU total usage time and time for different processes of EMBED algorithm (Figure 6) and of the geometric build-up algorithm (Figure 7) has been shown. For the EMBED algorithm, the CPU time for bound smoothing increases dramatically with the
number of atoms. In other words, the bound smoothing process dominates the whole algorithm. We avoid this time-consuming process by replacing it with a cheap geometric build-up process. It explains why we have performance gain in the geometric build-up algorithm (Figure 5).

For the geometric build-up algorithm, the dominant process is the metrization process. The geometric build-up process is actually very cheap (Figure 7).

Besides runtime performance, EMBED algorithm and the geometric build-up algorithm are compared in terms of accuracy of structures they generate. Accuracy is defined as fitness of structures to the given distance constraints. Let $d_{i,j}$ be the distance between atoms $i$ and $j$ in a solved structure and $l_{i,j}$ and $u_{i,j}$ be the lower bound and upper bound between atoms $i$ and $j$ in the given distance constraints. The total distance error ($E$) is defined as follows:

$$E = \sum_{i,j \in S} \max(l_{i,j} - d_{i,j}, 0, d_{i,j} - u(i,j))$$

(5.35)

where $S$ denotes the set of $i$-$j$ atom pair in the given distance constraints data. The unit of $E$ is angstrom (Å). For all test structures determined by EMBED algorithm (partial metrization method) and the geometric build-up algorithm, total error values have been computed and shown in Figure 8. The total error values of test structures generated by using EMBED algorithm increase as the number of atoms in the structures increases, while the errors of structures generated by the geometric build-up algorithm are all small than 1Å. It indicates that the structures obtained by the geometric build-up approach have better accuracy in terms of their fits to given distance constraints. To visualize the accuracy discrepancy between structures generated by both algorithms, we super-impose and compare the 3-D models of one of the test cases (poly-alanine 50mer) generated by using EMBED algorithm (4-atom metrization method) (blue) and the geometric build-up (green) algorithm (Figure 9). The structure generated by EMBED appears more straight and the chain is not connected at one end on the bottom, which suggests a lot of errors exist in the EMBED-generated structures.
5.7 Concluding remarks

The distance data derived from NMR experimental data can be used to calculate the 3-D structures of a protein. If the distances between all pairs of atoms are available and exact, the structure of the protein can be determined in linear time. The linear-time algorithm called geometric build-up algorithm we developed previously is based on a simple geometric relationship between coordinates and distances [Dong and Wu 2002, 2003]. This concept has been applied to sparse sets of exact distance data. We find that a protein structure can still be determined in a reasonable accuracy with only 3% of all distances in a molecule. Numerical errors might be accumulated in calculating the structure and can be avoided by using an updated method [Wu and Wu 2003]. However, in reality, distance data derived from NMR experimental data are sparse and inexact, which makes the molecular distance geometry problem NP-complete. Heuristic algorithms, such as the EMBED algorithm, have been applied to solve the problem [Crippen and Havel 1988]. Some processes in this algorithm, especially bound smoothing, make it unaffordable. As a result, the EMBED algorithm is computational too expensive for solving the structure of a protein in median size. Modifications on the EMBED algorithm to expedite the determination process have been made [Kuszewski et al. 1992].

In this paper, we reported a novel geometric build-up algorithm for protein structure determination using sparse sets of inexact distances. Our algorithm is based on the same idea in previously reported geometric build-up algorithms for full sets of exact distances and sparse sets of exact distances. The difference is that we utilize partial metrization to estimate missing distances between the base atoms and all other atoms on the level of triangle inequality. Moreover, instead of determining every atom directly, we determine boundary points of a space suitable for each atom and estimate the position of the atom by solving a least squares problem so that all constraints between the atom and all other determined atoms are satisfied. Atoms can be solved and built up repeatedly in this way, one at a time. The atom build-up process is actually used to replace the time-consuming bound smoothing process and embedding process of EMBED algorithm (Figure 4). The result on a set of poly-alanine chain (10 ~ 200mer) shows
that our algorithm is more efficient than EMBED algorithm (Figure 5). Moreover, structures generated by our algorithm have less error than those generated by EMBED algorithm (Figure 8). It indicates that the structures obtained by our approach have better accuracy in terms of their fits to the distance constraints.

Our algorithm has been integrated into CNS and therefore is available for practical use. It works in conjunction with a simulated annealing method as an alternative approach to EMBED algorithm. Applications of our approach on some test proteins are underway and will be reported elsewhere.

5.8 References


P. Bourne and H. Weissig, Structural Bioinformatics, Wiley-Liss, Inc., 2003


G. M. Crippen and T. F., Havel, Distance Geometry and Molecular Conformation, John Wiley & Sons, 1998


D. Wu and Z. Wu, An Updated Geometric Build-Up Algorithm for Molecular Distance Geometry Problems with Sparse Distance Data, submitted, 2003

(http://www.math.iastate.edu/wu)
Figure 5.1  The atom $i$ can be determined with its distances to four base atoms (atoms 1, 2, 3, and 4).

Figure 5.2  The coordinates of the atom 3 can be estimated by the four points surrounding it.
A Geometric Build-Up Algorithm for Problems with Sparse Sets of Inexact Distances

1. Find four base atoms that are not in the same plane; determine the coordinates of the base atoms by solving a set of linear system.

2. Repeat:
   For each of the remaining atoms,
   If there are at least four determined atoms having distance constraints with the atom, determine the atom by solving a local least-squares problem, using its distance constraints with the determined atoms.
   If the sum of the squares is not equal to a predetermined value, solve a least-squares problem to refine the coordinates of the atoms.
   If no atom is determined in the whole loop, stop.

3. All atoms are determined

Figure 5.3 A geometric build-up algorithm for problems with sparse sets of inexact distances

EMBED Algorithm     Geometric Build-Up Algorithm

Distance Bounds Input
(NOEs and Covalent Bonds)

Initial Bound Smoothing

Metrization

Bound Smoothing    Geometric Build-Up

Embedding

Pre-folded Structures

Figure 5.4 Flowchart of EMBED algorithm and Geometric Build-Up algorithm for distance data with bounds
Figure 5.5  Performance comparison between EMBED algorithms (full metrization and partial metrization) and the geometric build-up algorithm.

Figure 5.6  CPU total time and part time for difference processes of EMBED algorithm
Figure 5.7  CPU total time and part time for different processes of the geometric build-up algorithm

Figure 5.8  Comparison of error values of structures generated by EMBED algorithm (4-atom metrization) and the geometric build-up algorithm. The error value is defined in Equation (5.35)
Figure 5.9 Poly-alanine chains (50mer) generated by EMBED algorithm with 4-atom metrization (cyan) and the geometric build-up algorithm (green).
CHAPTER 6. MTMM – A MATLAB Toolbox for Macromolecular Modeling

A paper is to appear on the Proceeding of the International Conference of Bioinformatics Applications, December 2004, Fort Lauderdale, Florida

Feng Cui, Tauqir Bibi, and Zhijun Wu

6.1 Abstract

A MATLAB toolbox for macromolecular modeling is developed. The toolbox consists of routines for extracting and updating PDB data files, for calculating DME (Distance Matrix Error) and RMSD (Root-Mean-Square Deviation) of given structures, for building structural models with known inter-atomic distances, etc. Two algorithms, a singular-value decomposition algorithm and a geometric build-up algorithm, are used for distance-based structure modeling. A simulated annealing algorithm is implemented for energy minimization for structure refinement or determination. Functions facilitating structural conversion among three coordinate systems defined in terms of Cartesian coordinates, inter-atomic distances, and dihedral angles are also built in the toolbox.

6.2 Introduction

Macromolecular modeling requires intensive numerical computation such as matrix-vector calculations, optimization, and solving differential equations. The required computation is currently supported through several commercial codes including Amber (Pearlman 1995), Charmmm (Brook et al. 1983), Xplor (Brünger 1992), Insight (Molecular Simulations Inc., CA 1995,
Most of these systems are written in Fortran or C and can only be used for some general modeling purposes such as dynamics simulation and energy minimization. Many specific modeling functions such as structural comparison and transformation or distance-based structure modeling are either not available or not easy to use. On the other hand, many scientific computing environments such as Mathematica (Wolfram 1991) and Matlab (Gilbert 1992) do not have the direct support for the computation often required in macromolecular modeling. As a result, biologists often have to repeatedly write their own codes, especially numerical computing codes, to meet their daily modeling needs. For this reason, we have developed a Matlab toolbox \textit{MTMM} as a special computational environment for macromolecular modeling. The toolbox consists of a set of Matlab routines with a variety of molecular modeling functions. The reason we implement such a toolbox in Matlab is to take advantage of the numerical computing capability of Matlab so that advanced numerical computing tools become directly accessible to macromolecular modeling.

The current version of \textit{MTMM} has five basic modules including one for reading and updating PDB files, one for comparing structures, one for transforming structures from one coordinate system to another, one for modeling structures with distance data, and one for refining structure via energy minimization. Each module involves certain type of numerical computation, but most of them can be implemented in a relatively straightforward manner by using Matlab built-in functions.

DNA or protein structures are usually stored in certain type of database. One of such database is the PDB Data Bank, which hosts several tens of thousands of protein structures. In this database, each structure is stored in a separate file. Every line in the file contains 80 characters and is used as a record. A record is divided into a list of fields. In macromolecular modeling, it is always important to be able to get access to the PDB Data Bank and to process the data in the PDB files. The \textit{MTMM} module for reading and updating PDB files contains routines to read a PDB file into a character matrix and to extract or update certain structural data for the molecule such as the coordinates of the atoms.

\textit{MTMM} has a special module for computing the root-mean-square deviation (RMSD) and
the distance matrix error (DME) of two structures. This is not only necessary for comparing two structures of interest but also useful for other structure modeling purposes such as structure alignment. RMSD and DME calculations require many matrix-vector operations and in particular, for structure translation and rotation. However, the operations can be carried out easily in Matlab.

A DNA or protein structure can be represented in different coordinate systems. A coordinate system can be defined in terms of Cartesian coordinates of the atoms, or the distances between pairs of atoms, or the dihedral angles around flexible bonds (called internal coordinates along with certain bond lengths and angles). Different coordinate systems may be used in different contexts, and transformation among them is often needed in practice. The MTMM module for structural transformation contains a set of routines that can be used to transform a given structure among its Cartesian, distance, and dihedral angle representations.

A molecular structure may be determined when a set of inter-atomic distances is given. It can be done by mathematically solving a so-called molecular distance geometry problem. The method has been used in NMR structure determination as well as homology modeling. A special module for distance-based structure modeling is developed in MTMM. However, in the current version of MTMM, only the routines for solving exact molecular distance geometry problems are included. Routines for more general and practical cases will be added in future development. Two algorithms have been implemented for solving the molecular distance geometry problem. One is based on singular-value decomposition and another on geometric build-up. Both can solve the problem in polynomial time if all inter-atomic distances for the molecule are given.

A protein structure is assumed to correspond to the global energy minimum of the protein potential energy. Therefore, the structure can be determined in principle by minimizing an energy function of the protein. An energy minimization routine has been implemented in MTMM using a simulated annealing algorithm combined with a quasi-Newton algorithm for local minimization. The global energy minimum is hard to locate in general, but the routine can be used in applications such as structure refinement, where a good initial structure is
provided and the minimum energy structure can be found without exhausted searches.

The current version of MTMM has implemented only a small and basic set of structure computing functions. More general functions will be implemented in future versions. However, even in the current version, the routines have already been very useful for many routine calculations in structure modeling. In the following sections of the paper, we give a more detailed description on the implementation of MTMM. Sections 2 to 4 present the algorithms used in MTMM for distance-based structure modeling, energy minimization, and structural comparison and transformation. Section 5 describes the implementation details for all the modules and routines. Section 6 presents the performance results of selected MTMM routines.

6.3 Distance-based structural modeling

A molecular structure can be determined if a sufficient set of distances or their ranges between pairs of atoms in the molecule is given. This approach to molecular structure determination has been widely used in macromolecular modeling and in particular, in NMR-based structure modeling (Crippen and Havel 1988; Kuntz, Thomason and Oshiro 1993; Brünger and Niges, 1993; Havel 1995; More and Wu 1999).

A distance geometry problem can be stated as to find the coordinates for the atoms of a molecule, given a set of distances between pairs of atoms. In general, the problem can be stated as to find the coordinates for a set of points in $k$-dimensional Euclidean space $R^k$ for any $k$, given a set of distances between pairs of points in $R^k$. If the distances between all pairs of atoms are given, the problem is relatively easy to solve. Algorithms have been developed to solve the problem in polynomial time for example using a singular value decomposition algorithm (see for example Crippen and Havel 1988). However, if only a subset of all distances is available, the problem has been proved to be NP-hard (Saxe 1979; Moré and Wu 1996a).

The singular-value decomposition algorithm is based on the following idea: given distances $d_{i,j}$ between atoms $i$ and $j$ for all $i, j = 0, 1, \ldots, n$, we have

$$
\|x_i - x_j\| = d_{i,j} \quad i, j = 0, 1, \ldots, n
$$

(6.1)
or equivalently,

\[ \|x_i - x_j\|^2 = d_{i,j}^2 \quad i, j = 0, 1, \ldots, n \] (6.2)

Let \( x_0 \) be located at the origin. Then

\[ \|x_i\|^2 = d_{i,0}^2, \quad i = 1, 2, \ldots, n \]
\[ \|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, 2, \ldots, n \] (6.3)

We then obtain

\[ d_{i,0}^2 - d_{i,j}^2 + d_{j,0}^2 = 2x_i^T x_j, \quad i, j = 1, 2, \ldots, n \] (6.4)

Let \( D = (d_{i,0}^2 - d_{i,j}^2 + d_{j,0}^2) / 2 \). We can then define a matrix \( D = [D_{i,j}] \). Let \( X \) be an \( n \times 3 \) matrix and \( X = [x_1, x_2, \ldots, x_n]^T \). We then have

\[ D = XX^T. \] (6.5)

If a solution exists for this equation, matrix \( D \) must be of \( \text{rank} \leq 3 \). Therefore, we can make a singular-value-decomposition for \( D \) to obtain

\[ X = U\Sigma U^T, \] (6.6)

where \( U \) is an \( n \times 3 \) orthogonal matrix and \( \Sigma \) is a \( 3 \times 3 \) diagonal matrix with the diagonal elements \( \sigma_1, \sigma_2, \) and \( \sigma_3 \) being the three largest singular values of \( D \). A solution for \( D = XX^T \) can then be obtained with

\[ X = U\Sigma^{1/2}. \] (6.7)

Note that the singular value decomposition can be done in at most \( O(n^3) \) floating point operations.

A more straightforward approach to the distance geometry problem is using a pure geometric build-up method (Dong and Wu 2002, 2003), which can also be considered as a matrix completion method (Huang, Liang and Pardalos 2002). The method is based on a simple relationship between distances and coordinates. For example, whenever the coordinates of four
atoms and four distances between the four fixed atoms and an unfixed atom are given, the coordinates of the unfixed atom can immediately be determined by solving a small system of algebraic equations formed by the four distance constraints (see Figure 1). More specifically, let \( x_1, x_2, x_3 \) and \( x_4 \) be the coordinate vectors of the first four atoms, say atoms 1, 2, 3, and 4, and let \( x_i \) be the coordinate vector of the fifth atom, say atom \( i \). Assume that the distances between the first four atoms are not in a plane. The coordinate-vector \( x_i \) of atom \( i \) can then be determined uniquely by the following system of equations,

\[
\begin{align*}
\| x_i - x_1 \| &= d_{i,1} \\
\| x_i - x_2 \| &= d_{i,2} \\
\| x_i - x_3 \| &= d_{i,3} \\
\| x_i - x_4 \| &= d_{i,4}
\end{align*}
\]

which can further be reduced to a linear system of equations and solved efficiently in constant time (Dong and Wu 2002).

An overall geometric build-up procedure can be constructed as follows. Given a molecule of \( n \) atoms and a set of distances between pairs of atoms, the coordinates of four selected atoms can always be determined if the distances among them are available. Then, each of the remaining unfixed atoms is examined to find the distances between the unfixed atom and four previously fixed atoms. If such four atoms are found, the coordinates for the unfixed atom can be determined immediately by using the four distances between the unfixed atom and the four fixed atoms. The process continues until all the atoms are fixed (see Figure 2 for an example). If the distances between all pairs of atoms in the molecule are given, the build-up algorithm can determine the coordinates of \( n \) atoms in order of \( n \) floating-point operations. If a sparse set of distances is given, the algorithm can be implemented to solve the problem either completely or partially, depending on given distances. In case the problem is solved partially, a partial structure is determined.
6.4 Potential energy minimization

The potential energy minimization approach for structure determination is based on the hypothesis that the structure of a molecule corresponds to the global energy minimum of the molecule. Therefore, the structure can be determined by minimizing an energy function for the molecule in its conformational space.

The potential energy minimization problem is hard to solve since the energy function usually is highly non-convex and has an enormous number of local minima. In many cases, the number of local minima may grow as an exponential function of the number of variables, which makes the search for the global minimum an extremely difficult job. It is possible only if a reasonable starting point is provided. But efficient optimization algorithms with adequate computing power are still critical for a successful search.

A simulated annealing algorithm can be used to search for the global minimum of a function. The algorithm is expected to be able to find the global minimum of the function since physical annealing can often bring a physical system successfully to its ground state (Kirkpatrick, Gelatt, Jr. and Vecchi 1983).

A physical annealing process starts from a high temperature, and then cools down gradually to the zero temperature where the system reaches its ground state. The process usually proceeds slowly so that at each cooling stage the system has enough time to reach equilibrium. Otherwise, it is trapped in a local state.

A simulated annealing algorithm mimics this process by considering the function for a global minimization problem as the energy function of a simulated system. A temperature parameter is introduced and decreased stage by stage. At each stage, the function values are randomly sampled. When a point within the function domain with a lower function value is found, it is accepted as the current point. Otherwise the point is accepted and rejected randomly using the Metropolis criterion, which depends on the temperature: If the temperature is higher, the probability of accepting the points is also higher. This property allows the algorithm to accept more points at high temperature and gradually settle down at lower temperature to small regions where the point with the lowest function value may be located. A pseudo code for the
simulated annealing algorithm is illustrated in Figure 3. It has been proved that the sequence of the points sampled by the simulated annealing algorithm forms a Boltzmann distribution and converges to the global minimum of the function with probability one as the temperature goes to zero (Aarts and Korst 1989).

A successful simulated algorithm requires fine-tuning several parameters such as cooling schedule, step size, and starting state. A slower cooling schedule gives the algorithm more time to sample the conformation space, but of course, it takes longer computing time. The step size and temperature can be controlled so that the algorithm can overcome large energy barriers without being trapped in local energy minima. At each cooling stage, we can either use the initial conformation (usually it is randomly selected) as the starting point or choose one with the lowest energy value at the previous cooling stage.

6.5 Structural comparison and transformation

In order to compare two DNA or protein structures, we often need to compute the DME or RMSD of the two structures. If two distance matrices $C$ and $D$ for two structures are given, the DME for the two structures can be computed as the Frobenius norm of the difference matrix of $C$ and $D$. More accurately, $DME(C, D) = ||C - D||_F/n$. The computation of RMSD is more involved since it requires structure translation, rotation, as well as norm calculation. Assume that the structure is defined in terms of the Cartesian coordinates of the atoms and represented by an $n \times 3$ coordinate matrix, where $n$ is the number of atoms in the molecule. Structure translation can then be done via adding a translation vector to each row of the coordinate matrix, and structure rotation via multiplying the coordinate matrix by a rotation matrix. Let $X$ and $Y$ be the coordinate matrices of two structures after they are translated so that their centers of geometry coincide. The RMSD of the two structures is then defined as

$$RMSD(X, Y) = \min_Q \|X - YQ\|_F/\sqrt{n},$$

where $Q$ is a rotation matrix and $QQ^T = I$. Let $C = Y^TX$, and let $C = U\Sigma V^T$ be the singular-value decomposition of $C$. Then it can be verified that $Q = UV^T$ solves the
minimization problem in the definition of RMSD (Golub and van Loan 1989). The computation of RMSD can therefore be carried out through a series of matrix-vector operations.

A molecular structure may be represented in different coordinate systems, or in other words, in terms of different sets of parameters or variables. Transformation of a structure from one coordinate system to another is routinely needed in practice, but often requires tedious algebraic calculations. Three coordinate systems are commonly used including the Cartesian coordinates, the inter-atomic distances, and the dihedral angles.

Let $x_i$, $x_j$, $x_k$ and $x_l$ be the Cartesian coordinates of four atoms. Let $a$, $b$, and $c$ be three vectors, $a = x_j - x_i$, $b = x_k - x_j$, and $c = x_l - x_k$, and $\alpha_{ijk}$, $\alpha_{jkl}$, and $\alpha_{ikl}$ be the angles between these vectors as shown in Figure 4. Define the dihedral angle $\alpha_{ijkl}$ to be the angle between the plane formed by $a$ and $b$ and the plane formed by $b$ and $c$. Then the following formulas define the relationships among the Cartesian coordinates, the inter-atomic distances, and the dihedral angles. They can be used as a basic set of rules for the transformation among the three different coordinate systems.

Coordinates vs. Dihedral Angles:

\begin{align*}
    a \cdot b &= -\|a\|\|b\| \cos \alpha_{ijk} \\
    b \cdot c &= -\|b\|\|c\| \cos \alpha_{jkl} \\
    a \cdot c &= -\|a\|\|c\| \cos \alpha_{ikl}
\end{align*} 

(6.10)

Inter-Atomic Distance vs. Dihedral Angles:

\begin{align*}
    d_{ik}^2 &= \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \cos \alpha_{ijk} \\
    d_{jl}^2 &= \|b\|^2 + \|c\|^2 - 2\|b\|\|c\| \cos \alpha_{jkl} \\
    d_{il}^2 &= \|a\|^2 + \|b\|^2 + \|c\|^2 - 2\|a\|\|b\| \cos \alpha_{ijk} - 2\|b\|\|c\| \cos \alpha_{jkl} - 2\|a\|\|c\| \cos \alpha_{ikl}
\end{align*}

(6.11)

Coordinates vs. Inter-Atomic Distances:

\[ \|x_i - x_j\| = d_{ij}, \quad i, j = 1, 2, \ldots, n \]

(6.12)
6.6 Implementation

6.6.1 Reading and updating PDB files

The MTMM module for reading and updating PDB files has two routines, one to read the coordinate data from a PDB file and save it to a local matrix and another to rewrite the coordinate data in the PDB file with a given coordinate matrix.

The coordinate reading routine is defined as a Matlab function, \( \text{ret.info} = \text{read.PDB} \) \((PDB\_file, \text{coord.file})\), where \( PDB\_file \) is the name of the PDB file and \( \text{coord.file} \) is the name of the file to store the output coordinates. The function reads the coordinate data from \( PDB\_file \) and saves it in \( \text{coord.file} \) as an \( n \times 3 \) matrix, where \( n \) is the number of atoms in \( PDB\_file \). The function returns \( 1 \) if it succeeds or \( 0 \) otherwise.

The coordinate updating routine is defined as a Matlab function, \( \text{ret.info} = \text{update.PDB} \) \((PDB\_file, \text{coord.file})\), where \( \text{coord.file} \) is the name of the file with given coordinate data and \( PDB\_file \) is the name of the PDB file to be updated. The function reads the coordinate data from \( \text{coord.file} \) and uses it to overwrite the coordinate data in \( PDB\_file \). The coordinate data in \( \text{coord.file} \) must be stored as an \( n \times 3 \) matrix, where \( n \) is the number of atoms in \( PDB\_file \). The function returns \( 1 \) if it succeeds or \( 0 \) otherwise.

6.6.2 DME computation

Distance Matrix Error (DME) is one of the commonly used methods to compare two molecular structures. It can be computed as the Frobenius norm of the difference matrix of two distance matrices for the two structures to be compared.

Let \( X \) and \( Y \) be the coordinate matrices for two structures to be compared. The distance matrices \( C \) and \( D \) for the two structures can be obtained easily as follows,

\[
C(i,j) = \text{norm}[X(i,:) - X(j,:)]
\]

\[
D(i,j) = \text{norm}[Y(i,:) - Y(j,:)]
\]  

(6.13)

where \( i,j = 1, \ldots, n \) and \( n \) is the number of the atoms. Once the distance matrices are
obtained, the DME of the two structures can be computed with the formula,

\[
DME(C, D) = \frac{\|C - D\|_F}{\sqrt{n}}\frac{1}{n} = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} [C(i,j) - D(i,j)]^2}{n}}
\]

In MTMM, the DME calculation is implemented as a Matlab function \(dme = DME(coord\_file1, coord\_file2)\), where \(coord\_file1\) and \(coord\_file2\) are two files containing two \(n \times 3\) coordinate matrices for the two structures to be compared. The function returns the DME value of the two structures.

6.6.3 RMSD computation

The computation of Root-Mean-Square-Deviation (RMSD) of two structures requires structural translation and rotation and matrix norm calculation. Let \(X\) and \(Y\) be the coordinate matrices for the two structures after they are translated to the same location. The RMSD of the two structures is then defined as

\[
RMSD(X, Y) = \min_{Q} \|X - YQ\|_F/\sqrt{n},
\]

(6.15)

where \(Q\) is a rotation matrix and \(QQ^T = I\). Based on previous discussion, we can compute the RMSD as follows. First, compute

\[
xc = \frac{1}{n} \sum_{i=1}^{n} X(i,:) \quad yc = \frac{1}{n} \sum_{i=1}^{n} Y(i,:)
\]

(6.16)

where \(xc\) and \(yc\) are the mean vectors of matrices \(X\) and \(Y\). Then, set the difference matrix \(XX\) between \(X\) and its mean vector \(xc\) to be

\[
XX(:,1) = X(:,1) - xc(1)
XX(:,2) = X(:,2) - xc(2)
XX(:,3) = X(:,3) - xc(3)
\]

(6.17)
and the difference matrix $YY$ between $Y$ and its mean vector $yc$ to be

$$
YY(:,1) = Y(:,1) - yc(1)
$$
$$
YY(:,2) = Y(:,2) - yc(2)
$$
$$
YY(:,3) = Y(:,3) - yc(3)
$$

(6.18)

Let $C = YY^T XX$. Let $U\Sigma V^T = C$ be the singular-value decomposition of $C$. Then $Q = UV^T$ is the rotation matrix required in the calculation of RMSD and

$$
RMSD(X,Y) = \min_Q \| X - YQ \|_F / \sqrt{n}
$$

$$
= \sqrt{\frac{\sum_{i=1}^{n} \| XX(i,:) - YY(i,:)Q \|^2}{n}}
$$

(6.19)

In MTMM, the computation of RMSD is implemented as a Matlab function $rmsd = \text{RMSD(coord\_file1, coord\_file2)}$, where $coord\_file1$ and $coord\_file2$ are two data files containing two $n \times 3$ coordinate matrices for the two structures to be compared.

### 6.6.4 Structural transformation

The MTMM module for structural transformation contains a set of Matlab routines to transform a structure from one coordinate system to another among three coordinate systems defined in terms of Cartesian coordinates, inter-atomic distances, and dihedral angles. The names of the routines are $\text{coord\_to\_distance}$, $\text{distance\_to\_coord}$, $\text{coord\_to\_angle}$, $\text{angle\_to\_coord}$, $\text{distance\_to\_angle}$, and $\text{angle\_to\_distance}$. The functions of the routines are demonstrated in Figure 5. We describe them in greater detail in the following.

#### 6.6.4.1 Cartesian coordinates vs. angles

Given Cartesian coordinates of arbitrarily four atoms $i$, $j$, $k$, and $l$, the bond length, bond angles and dihedral angle formed by these four atoms can be easily determined (see Figure 6).
Let $x_i$, $x_j$, $x_k$ and $x_l$ be the coordinate vectors for the atoms. Then

\[
\begin{align*}
    a &= x_i - x_j \\
    b &= x_k - x_j \\
    c &= x_l - x_k
\end{align*}
\]  

(6.20)

and the bond angles $\alpha_{ijk}$, $\alpha_{jkl}$, $\alpha_{ikl}$, and the dihedral angle $\alpha_{ijkl}$ can be computed using the formulas,

\[
\begin{align*}
    a \cdot b &= -\frac{\|a\|\|b\| \cos \alpha_{ijk}}{\|a\|\|b\|} \\
    b \cdot c &= -\frac{\|b\|\|c\| \cos \alpha_{jkl}}{\|b\|\|c\|} \\
    a \cdot c &= -\frac{\|a\|\|c\| \cos \alpha_{ikl}}{\|a\|\|c\|}
\end{align*}
\]  

(6.21)

and

\[
\cos \alpha_{ijkl} = \frac{(a \cdot b)(b \cdot c) - (a \cdot c)(b \cdot b)}{\|a\|\|b\|\|c\| \sin \alpha_{ijk} \sin \alpha_{jkl}.}
\]  

(6.22)

In MTMM, these calculations are done by a Matlab function, $[\alpha_{ijk}, \alpha_{jkl}, \alpha_{ikl}] = \text{coord.to.angle}(x_i, x_j, x_k, x_l)$, where $x_i$, $x_j$, $x_k$, $x_l$ are the coordinate vectors for four given atoms $i$, $j$, $k$, and $l$, and $\alpha_{ijk}$, $\alpha_{jkl}$, $\alpha_{ikl}$ are the corresponding bond and dihedral angles.

### 6.6.4.2 Angles vs. cartesian coordinates

Given the Cartesian coordinates of three atoms $i$, $j$, and $k$, the bond angles and the dihedral angle among them, the Cartesian coordinate of the fourth atom $l$ can be determined as shown in Figure 7.

Let $n_a = a/\|a\|$, $n_b = b/\|b\|$ and $n_{ab} = a \times b/(\|a\|\|b\| \sin \alpha_{ijk})$, the components of vector $c$ are

\[
\begin{align*}
    c_1 &= -\|c\| \cos \alpha_{ikl} \cdot n_a \\
    c_2 &= -\|c\| \cos \alpha_{jkl} \cdot n_b \\
    c_3 &= -\|c\| \sin \alpha_{jkl} \sin \alpha_{ijkl} \cdot n_a
\end{align*}
\]  

(6.23)
Then \( c = c_1 + c_2 + c_3 \) and \( x_l = x_k + c \). This process can be repeated to obtain the Cartesian coordinates for a set of atoms when the bond angles and dihedral angles associated with the atoms are all given. In MTMM, the above calculation is implemented as a Matlab function, 
\[
\text{angle_to_coord}(x_i, x_j, x_k, \alpha_{ijk}, \alpha_{jkl}, \alpha_{ijkl})
\]
where \( x_i, x_j, x_k \) are the coordinate vectors for atoms \( i, j, k \), and \( \alpha_{ijk}, \alpha_{jkl}, \alpha_{ijkl} \) are the bond and dihedral angles associated with atoms \( i, j, k, l \).

### 6.6.4.3 Inter-Atomic distances vs. angles

Given all inter-atomic distances among four atoms \( i, j, k, \) and \( l \), we can find the bond angles and the dihedral angle associated with them using the formulas (6.11) and

\[
\cos \alpha_{ijkl} = \frac{\cos \alpha_{ijk} \cos \alpha_{jkl} + \cos \alpha_{ikl}}{\sin \alpha_{ijk} \sin \alpha_{jkl}} \quad (6.24)
\]

where \( d_{xy} \) is the distance between atoms \( x \) and \( y \). In MTMM, the calculation of the angles from a given set of distances is done by a Matlab function, \([\alpha_{ijk}, \alpha_{jkl}, \alpha_{ijkl}] = \text{distance_to_angle}(d_{ij}, d_{jk}, d_{kl}, d_{ki}, d_{jl}, d_{il})\) where \( d_{ij}, d_{jk}, d_{kl}, d_{ki}, d_{jl}, d_{il} \) are the distances among four atoms \( i, j, k, l \), and \( \alpha_{ijk}, \alpha_{jkl}, \) and \( \alpha_{ijkl} \) are the corresponding bond and dihedral angles.

### 6.6.4.4 Angles vs. inter-atomic distances

Given the bond angles and dihedral angle for four atoms \( i, j, k, \) and \( l \) and inter-atomic distances for adjacent atoms, we can compute all other distances among the atoms. First, from the formula (6.24), we can compute the angle \( \alpha_{ikl} \). Then, from the formulas

\[
\begin{align*}
d_{ik}^2 &= d_{ij}^2 + d_{jk}^2 - 2d_{ij}d_{jk}\cos \alpha_{ijk} \\
d_{jl}^2 &= d_{jk}^2 + d_{kl}^2 - 2d_{jk}d_{kl}\cos \alpha_{jkl} \\
d_{il}^2 &= d_{ij}^2 + d_{jk}^2 + d_{kl}^2 - 2d_{ij}d_{jk}\cos \alpha_{ijk} - 2d_{jk}d_{kl}\cos \alpha_{jkl} - 2d_{ij}d_{kl}\cos \alpha_{ikl} \quad (6.25)
\end{align*}
\]

we can get the distances \( d_{ik}, d_{kl}, d_{il} \). In MTMM, this calculation is done by a Matlab function, \([d_{ik}, d_{kl}, d_{il}] = \text{angle_to_distance}(d_{ij}, d_{jk}, d_{kl}, \alpha_{ijk}, \alpha_{jkl}, \alpha_{ijkl})\) where \( \alpha_{ijk}, \alpha_{jkl}, \) and \( \alpha_{ijkl} \) are the bond and dihedral angles associated with atoms \( i, j, k, l \), and \( d_{ij}, d_{jk}, d_{kl} \).
are the distances between adjacent atoms. The function returns other distances, \( d_{ik}, d_{kl}, d_{il} \), among the atoms.

6.6.4.5 Cartesian coordinates vs. inter-atomic distances

Given the Cartesian coordinate of atoms, the distances among atoms can be computed straightforwardly. MTMM has a routine, \( D = \text{coord.to.distance}(X) \), to compute the distances among all atoms of a molecule given the Cartesian coordinates of the atoms in \( X \), where \( X \) is an \( n \times 3 \) coordinate matrix. The routine returns an \( n \times n \) matrix \( D \) containing all the distances.

6.6.4.6 Inter-atomic distances vs. cartesian coordinates

The conversion from distances to Cartesian coordinates is equivalent to solving a molecular distance geometry problem when given all distances. We use a traditional singular-value decomposition algorithm for this conversion. The MTMM routine is \( X = \text{distance.to.coord}(D) \), where \( D \) is an \( n \times n \) matrix containing all the distances. The routine returns an \( n \times 3 \) coordinate matrix \( X \) for the molecule.

6.6.5 Solving the molecular distance geometry problem

Two algorithms have been implemented in the current version of MTMM for solving the molecular distance geometry problem when all distances are given. One is the singular-value decomposition algorithm, and another the geometric build-up algorithm. The former solves a molecular distance geometry problem in \( O(n^2) \) floating-point operations, while the latter in \( O(n) \), provided all distances are available, where \( n \) is the number of atoms in the given molecule.

6.6.5.1 The singular-value decomposition algorithm

The routine in MTMM for solving a molecular distance geometry problem using the singular-value decomposition algorithm is defined as a Matlab function \( X = \text{dg.svd}(D) \), where \( D \) is an \( n \times n \) matrix containing the distances between all pairs of atoms. The function returns
an $n \times 3$ coordinate matrix $X$ for the molecule. The implementation of the algorithm in Matlab is straightforward. An outline of the Matlab code is given in Figure 8.

### 6.6.5.2 The geometric build-Up algorithm

The routine in MTMM for solving a molecular distance geometry problem using the geometric build-up algorithm is defined as a Matlab function $X = dg.gbu(D)$, where $D$ is an $n \times n$ matrix containing the distances between all pairs of atoms. The function returns an $n \times 3$ coordinate matrix $X$ for the molecule. The implementation of the algorithm is outlined in Figure 9. Note that if all distances are available, the atoms $b_1, b_2, b_3, b_4 \in F$ can always be found, and the for-loop will run for $n-4$ times and the while-loop will only run once. Every for-loop requires solving a system of 3 linear equations to obtain the coordinates of $a$, which takes a constant time. Therefore, the total time to obtain the coordinates of the atoms when all distances are given is in order of $n$, where $n$ is the number of atoms to be determined in the molecule.

### 6.6.6 Energy minimization functions

A simulated annealing algorithm is implemented in the current version of MTMM for small-scale energy minimization. The routine is defined as a Matlab function $[x, f] = simulated.annealing(fname, x_0)$, where $fname$ is the name of the energy function and $x_0$ is the starting point for the solution. The function returns an approximate solution $x$ and its energy value $f$. Note that $x_0$ and $x$ are vectors of coordinates for the atoms in the molecule. The coordinates are stored in the order that the first three elements of the vector are the coordinates for the first atom, and the second three for the second one, and so on and so forth.

The current implementation of the simulated annealing algorithm for energy minimization is preliminary. The parameters are not optimized. The routine has currently only a simple cooling schedule and a fixed number of cooling steps, and can probably find only an approximate solution to a given problem. It should be used along with a local optimization routine so that a further improved solution may be obtained. For example, the following calling sequence may
be used in minimizing a function,

\[ [x, f] = \text{simulated\_annealing('energy', } x_0) \]

\[ \text{options} = \text{optimset('GradObj', 'on')} \]

\[ [x, f] = \text{fminunc('energy', } x) \]

where \textit{energy} is the name of the function to be minimized. Note that since \textit{GradObj} is set to on, the function \textit{energy} must be defined to return not only the function value but also the gradient vector. Figure 10 contains an outline of the implemented simulated annealing algorithm.

### 6.7 Performance results

Here we present the performance results of some selected routines in MTMM, including \textit{read\_PDB}, \textit{update\_PDB}, \textit{RMSD}, \textit{DME}, and \textit{dg\_gbu}. The results were obtained from running the routines in Matlab Version 5.3 on a PC Pentium III with 512MB memory. They do not represent a complete and optimal test on the routines, but do provide a general time frame the routines may run in a regular platform. In order to see the performance of the routines on a problem of relatively practical size, we chose the protein 1HMV as the test problem. This protein has 4200 atoms. Some of the routines will need to run a significant amount of time for it. Table 1 contains the wall clock times and total numbers of floating point operations required by the routines for the protein.

The simulated annealing routine is designed to run along with a local minimization routine. We tested this routine on a simple energy function in cluster simulation. More specifically, we considered a cluster of 12 argon molecules interacting with each other through van der Waals forces. The potential energy of the system can then be calculated using a Lennard-Jones function,

\[ E = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left( \frac{1}{\|x_i - x_j\|^12} - \frac{2}{\|x_i - x_j\|^6} \right) \tag{6.26} \]

where \( n \) is the number of molecules in the cluster, and \( x_i \) and \( x_j \) are coordinates of molecules \( i \) and \( j \). We wanted to find the lowest energy state of the cluster using the simulated annealing
algorithm. Similar to finding the lowest energy state of protein, although much simpler, finding
the lowest energy state of a molecular cluster tends to be a difficult global optimization problem
as well, and therefore, it can be used as a reasonable test case for the simulated annealing
algorithm. We first generated a random vector of coordinates for the molecules in the cluster.
We then ran simulated annealing followed by local optimization using the Matlab function
\textit{fminunc} with a \textit{GradObj} option. The routines took about 181 seconds and 70 M flops to find
the global energy minimum of the cluster. Of course, this only shows the average computing
time required for running the routines, but does not imply that the routines can always find
the global minimum of an energy function. In fact, in many cases, they may fail to, since
global optimization problems are intractable in general. However, the routines can certainly
be used for solving relatively small problems or providing approximate solutions.

6.8 References

Ferguson, G. Seibel and P. Kollman. AMBER, a computer program for applying molecular
mechanics, normal mode analysis, molecular dynamics and free energy calculations to

B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M.
Karplus, CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics

A. T. Brünger, X-PLOR (version 3.1): A system for X-ray crystallography and NMR, Yale
University Press, New Haven, CT (1992)

Insight II System Guide, by MSI, CA 1997

Insight II, by MSI, CA 1995

S. Wolfram, Mathematica: A System for Doing Mathematics by Computer (2nd Edition),


Figure 6.1  The fifth atom is determined with the distances between the first four atoms and the fifth atom.

Figure 6.2  The coordinates of the atoms can be determined by a build-up procedure: First, determine the coordinates of atoms 1, 2, 3 and 4 with the distances among them. Then, determine the coordinates of atom 5 (and 6) and the first four atoms. Finally, determine the coordinates of atom 7 with the distances between atom 7 and atoms 1, 3, 5 and 6.
input initial $x_0$, $y_0 = f(x_0)$;
set $x = x_0$, $y = y_0$.
for $T = T_n, T_{n-1}, \ldots, T_1$ (decreasing)
for $k = 1, \ldots, n$
    $x_i = \text{perturb} (x_0)$;
    $y_i = f(x_i)$;
    $\Delta y = y_i - y_0$;
    $e = \exp (-\Delta y / T)$;
    if $(\text{rand} < e)$
        $x_0 = x_i, y_0 = y_i$;
    end
end
update $x, y$;
end

Figure 6.3 The function is evaluated on a set of points at each temperature of the simulated annealing algorithm. A point is accepted if it satisfies the Metropolis criterion and rejected otherwise. The algorithm converges to the global minimum of the function with probability one as $n \to \infty$ and $T \to 0$.

Figure 6.4 The Cartesian coordinates, inter-atomic distances, bond-angles, and dihedral angles can be defined in terms of each other.
Figure 6.5  A structure can be transformed from one coordinate system to another by using an appropriate transformation routine.

Figure 6.6  Bond angle $\alpha_{ij}$, $\alpha_{jkl}$, $\alpha_{ikl}$ and dihedral angle $\alpha_{ijkl}$ can be derived from the Cartesian coordinates of the atoms.
The coordinates of the fourth atom \( x_l \) can be determined, given the coordinates of \( x_i \), \( x_j \) and \( x_k \), bond angles \( \alpha_{ijk} \), \( \alpha_{ikl} \) and \( \alpha_{jkl} \), and the dihedral angle \( \alpha_{ijkl} \).

\[
\text{for } i = 1 : n-1 \\
\text{for } j = 1 : n-1 \\
\quad D(i,j) = [D(i,n)\cdot D(i,n) - D(i,j)\cdot D(i,j) + D(j,n)\cdot D(j,n)] / 2; \\
\text{end} \\
\text{end} \\
[U, S, V] = \text{svds}(D, 3); \text{ singular-value decomposition} \\
X = U \cdot S \cdot V; \\
X(n,:) = [0, 0, 0], \text{ the last atom is allocated at the origin.}
\]

The Matlab pseudo code for \( X = \text{dg.svd}(D) \). A Matlab function \text{svds} is used for the singular-value decomposition.
input $D$;
$F = \{\text{four initial atoms}\}; \text{fix first four atoms}$
$U = \{\text{unfixed atoms}\};$
while $U \neq 0$
for $a \in U$
do
find $b_1, b_2, b_3, b_4 \in F$ with distances to $a$ available;
fix $a$ with $b_1, b_2, b_3, b_4$;
$F = F / a; U = U \setminus a; \text{coordinates (a)} \rightarrow X;$
end
if no atom is fixed, stop; structure partially determined
end
structure completely determined

Figure 6.9 A pseudo code for the geometric build-up algorithm for solving the molecular distance geometry problem with exact distances, a small linear system needs to be solved in every for-loop.

\begin{verbatim}
Input initial $x_0$;
$y_0 = f(x_0)$;
set $x = x_0, y = y_0, T = 2, s = 0.2, alpha = 0.9$;
for $i = 1: 20$
$T = alpha * T;$
for $j = 1: 10$
accept = 0;
for $k = 1: 100$
$x_1 = x_0 + (0.5 - \text{rand(size (x)))} * s;$
$y_1 = f(x_1);$ y = $y_1 - y_0,$
$e = \exp(- dy / T);$ 
if ($\text{rand} < a$)
x0 = x1, y0 = y1; accept = accept + 1;
if $y_0 < y, x = x_0, y = y_0; end$
end
if accept < 25, s = s / 2; end
if accept > 75, s = s * 2; end
x0 = x, y0 = y;
end
\end{verbatim}

Figure 6.10 A pseudo code for simulated annealing energy minimization; the temperature is lowered in 20 steps.
Table 6.1  Wall clock times and floating-point operations of selected MTMM routines for protein 1HMV (with 4200 atoms) on a PC Pentium III

<table>
<thead>
<tr>
<th>routine name</th>
<th>Read_PDB</th>
<th>update_PDB</th>
<th>RMSD</th>
<th>DME</th>
<th>dg_gbu</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>32 secs</td>
<td>29 secs</td>
<td>&lt; 1 secs</td>
<td>1552 secs</td>
<td>&lt; 1 secs</td>
</tr>
<tr>
<td>flops</td>
<td>8779</td>
<td>8779</td>
<td>240 K</td>
<td>407 M</td>
<td>202 K</td>
</tr>
</tbody>
</table>
CHAPTER 7. General Conclusion

7.1 Text

X-ray crystallography\(^1\) and nuclear magnetic resonance (NMR) spectroscopy\(^2\) are two primary experimental techniques for determination of protein structures. Between them, NMR spectroscopy provides not only structure conformations but also dynamics information of a protein in solution. The latter is not obtainable from X-ray crystallography. However, limited distance data (e.g., NOEs\(^3\)) can be derived from NMR experiments due to the fact that NOEs are unrecognizable if two protons are separated by more than 5 Å in space. This leads to an under-determined system: the number of constraints is much fewer than the degrees of freedom of the system. Therefore, this system has multiple solutions. In other words, multiple structures (ensemble) with different conformations are satisfied stereo-chemistry criteria and given experimental data. It is believed that diverse conformations in an ensemble reflect protein flexibility in solution. Further analysis suggests that part of the disagreement among structures is due to imperfection of current refinement protocols and lack of long-range restraints. Once the protocols are improved (e.g., appropriate solvent models\(^4–8\)) or long-range restraints (e.g., residual coupling data\(^9–10\)) are applied, the disagreement becomes smaller, and the resulting structures appear closer to corresponding X-ray structures. It is not clear that how much disagreement among structures truly reflects the protein flexibility and how close should be between a NMR structure and its corresponding X-ray structure.

Force field improvement leads to more realistic treatment of non-bonded interactions in NMR structure refinement. As a result, resulting structures have been improved in terms of backbone dihedral angle distribution and hydrogen bond pattern\(^6,8\). However, the impact of force fields on final structures is limited, only when experimental restraints are not sufficient.
Further improvement on well-determined structures may have to rely on conformational database potentials derived from high-resolution X-ray structures.

Indeed, several different structural properties such as dihedral angle and hydrogen bonding pattern have been analyzed and used to refine NMR structures. Coordinate accuracy as well as backbone and side-chain dihedral angle distribution are significantly improved in refined structures. On the other hand, inter-atomic distances have never been used for refining NMR structures. However, their usefulness has been fully demonstrated through successful refinement of low-resolution X-ray structures.

Here, we show, for the first time, that inter-atomic distances derived from databases of high-resolution X-ray structures can be used to refine NMR structures. More specifically, we derived inter-atomic distances between atoms N, C, Cα, Cβ and O across two residues neighboring or separated by one residue. Totally, 20,000 distance types are considered. Distribution of distances in each distance type is calculated, as well as the most probable range (e.g. mean ± 2 × standard deviation) for the type. The obtained ranges are used as additional distance constraints to refine NMR structures. The refined structures are compared in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, RMSD (root-mean-square-deviation) values of the ensembles of structures, RMSD values of the structures compared with their X-ray crystal structures (for available ones), as well as the remaining distance errors in the structures. The results show that with additional database distance constraints, both the RMSD values are reduced and the acceptance rates of the structures are more than doubled, suggesting that protein structures can indeed be determined more accurately and efficiently by combining the distance constraints obtained from NMR experiments with additional distance constraints extracted from known protein structures in structural databases.

The distance constraints derived from databases are essentially short-range constraints, comparable to sequential NOEs in NMR experiments. It would be interesting to see whether these constraints can enhance or even replace some experimentally derived short-range restraints, for example, short-range NOEs (intra-residue or sequential) or dihedral angle re-
straints. If yes, the cost for experiments and labor would be reduced by using these computationally derived constraints. We refined NMR structures under different conditions: with an original set of NMR data; with an original set of NMR but without dihedral angle restraints; with an original set of NMR and additional constraints from databases but without dihedral angle restraints. We show that the structures refined using the NMR experimental constraints plus the database derived distance constraints can increase accuracy and precision of the structures with fewer distance violations, higher acceptance rates, and significantly improved Ramachandran plots even when the experimental torsion angle constraints are removed. On-going study shows that the database-derived distance constraints can replace sequential NOEs without compromising the precision and accuracy of NMR structures.

The database derived distance constraints are applied to prion, a biologically important protein responsible for the Mad Cow Disease. This protein heavily relies on NMR spectroscopy for structure determination because of its difficulty in crystallization. Due to insufficient experimental data, a critical loop on this protein has been constantly under-determined. We show that by adding distance constraints derived from databases of high-resolution protein structures, this under-determined loop can be refined into more realistic structures having improved quality and increased accuracy. We show, in particular, that the percentage of residues in the most favorable region of the Ramachandran diagram is increased from the 80 to 85% range in the previously reported structures to about 90% in the refined structures. It is the first evidence that the distance constraints derived from structural databases can be used to optimize the under-determined regions of a protein, in this case, prion protein.

In current molecular modeling software packages such as CNS or X-PLOR, EMBED algorithm is used to generate pre-folded structures that are in turn used as starting structures for refinement. Mathematically, the EMBED algorithm generates pre-folded structures by solving a so-called molecular distance geometry problem. However, from the algorithmic perspective, the EMBED algorithm is not very efficient due to some very time-consuming steps, bound smoothing and embedding. Here, we replace these two steps by a geometric build-up module, which is based on the simple relationship between coordinates and distances in
space. The resulting hybrid method significantly reduces computing time by more than six times.

A new Matlab\textsuperscript{22} toolbox has been developed for macromolecular modeling. It works in the Matlab environment in order to take advantage of the powerful computing capability of Matlab. It is a handy tool for mathematicians and structural biologists to test new algorithms and conduct trivial yet necessary structural calculation, \textit{e.g.} structural manipulation, compute RMSD between two structures, \textit{etc.}
7.2 References


ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deepest gratitude to all those who have helped me throughout my Ph.D. program. I especially want to thank my major professor Dr. Zhijun Wu for his extensive efforts in all aspects from selecting research topics, developing algorithms/approaches, interpreting data, to editing manuscripts. His instruction on presentation skills really helps me in my teaching practice as well as research presentations. His efforts have contributed and will continue to contribute, to my success as a scientist. I would also like to thank my co-major advisor, Dr. Robert Jernigan for his generous support and valuable advice in my research. My deepest appreciation also goes to Dr. Mark Gordon for his advice and guidance in adding an interdisciplinary component to my graduate education. I also extend my thanks to my committee members: Dr. Amy Andreotti and Dr. Xueyu Song for reading my dissertation and for the suggestion and comments provided.

I would also like to thank my fellow graduate students Ajith Gunaratne, Di Wu and Peter Vedell for helpful questions and discussions. Special thanks go to Ajith Gunaratne for his generous help on many different software packages including Latex. I want to thank Kriti Mukhopadhyay and Dr. Won-bin Young for their contributions on Prion structure refinement, and Tauqir Bibi’s work on MTMM.

I would also like to thank financial support from the Graduate Program on Bioinformatics and Computational Biology, Lawrence Baker Center for Bioinformatics and Biological Statistics, Department of Mathematics, of Iowa State University, and Pioneer Hi-Bred International, Inc.