2008

An integrated software environment for protein structure refinement

Rahul Ravindrudu
Iowa State University

An integrated software environment for protein structure refinement

by

Rahul Ravindrudu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Zhijun Wu, Co-major Professor
David Fernández-Baca, Co-major Professor
Srinivas Aluru
Guang Song
Kai-Ming Ho

Iowa State University

Ames, Iowa

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The development of an integrated software environment for protein structure refinement is reported. Energy minimization is combined with geometric embedding in the refinement program. The energy minimization procedure is used to sample the conformational space and find a group of low energy structures for further improvement. The distance geometry also known as geometric embedding is then applied to the structures with a set of statistical distances (distance derived statistically from known protein structures). The CHARMM potentials along with a set of recently developed statistical potentials are used in energy minimization. For distance geometry, in addition to the statistical distances, a set of distance bounds is also generated for each of the structures based on their normal mode fluctuations. The final output of the refinement program is an ensemble of plausible structures. The implementation of the algorithms, the organization of the software, and the parallelization of the computation is described. Some sample refinement results are also presented.

CHAPTER 1. INTRODUCTION

In order to fully understand biochemical systems and processes, the determination of three-dimensional protein structures is crucial. The accuracy and precision required of an experimentally determined model of a macromolecule, such as a protein or DNA, depends on the biological questions being asked of the structure. Questions involving say, the overall fold of a protein, or its topological similarity to other proteins, can be answered by structures of fairly low precision such as those obtained from very low resolution X-ray crystal diffraction data. Despite the low resolution, most of these structures are able to show the overall conformation of the protein in both its induced and repressed states and provide a framework for understanding the interactions it makes in performing its biological function. Questions involving reaction mechanisms, on the other hand, require much greater accuracy and precision as obtained from well-refined, high-resolution X-ray structures, including proper statistical analyses of the standard uncertainties of atomic structures and bond lengths. The most accurate and precise structures are those solved by X-ray crystallography to atomic resolution, which implies it should be better than 1.2 Å, and the number of such macromolecular structures is rapidly increasing. Structures at this level of accuracy can begin to address detailed functional biological questions.

The computer generated comparative models as well as the NMR models of protein structures do not have such high resolution structures. This was the basis for a 'between-CASP's refinement experiment, using some of the models submitted in Critical Assessment of Techniques for Protein Structure Prediction 6 (CASP 6). This experiment was called

Continuous CASP model refinement experiment (CASPR) with an aim to increase discussion and increase progress as well, using some of the models submitted in CASP6 as starting structures. The goal was to use any method to refine these approximate structures closer to experiment. These were not blind predictions. The work done in this was the starting point for this project.

The method used by us was to do energy minimization on a randomly generated coordinate structure from the given starting structure. The best result from this energy minimization was then used by distance geometry calculations to further refine the structure. Because the generation of structures used for minimization was totally random, only a small percentage of the result was useful. The goal of this research project is to improve the methods used to create the structures using energy minimization and also improve the distance constraints/restraints used by distance geometry modeling to hopefully give better results. It is hoped that these two software packages can be seamlessly combined to create a protein structure refinement software environment. The use of parallel processing is also hoped to improve the chances of finding a better refined structure than using a single processor.

The rest of the report is outlined as follows

Chapters 2 and 3 give the background information about the biological and computational aspects of the problem, including details about protein structures and various computational methods currently in existence for these protein structures. Chapter 4 presents the general structure refinement approach taken by us. The results are presented in Chapter 5, and the conclusions are drawn as well as the possible future work is presented in Chapter 6.

CHAPTER 2. BIOLOGY BACKGROUND

Computational biology is an interdisciplinary field that applies the techniques for computer science, applied mathematics, and statistics to address problems inspired by biology. Major fields in biology that use computational techniques include bioinformatics, computational genomics, molecular modeling, systems biology, protein structure prediction, structural genomics, computational biochemistry and biophysics.

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry. Its aim is the prediction of the three-dimensional structure of proteins from their amino acid sequences, sometimes including additional relevant information such as the structures of related proteins. In other words, it deals with the prediction of a protein's tertiary structure from its primary structure. Protein structure prediction is of high importance in medicine, for example in drug design, and biotechnology for example, in the design of novel enzymes. Every two years, the performance of current methods of protein structure prediction is assessed in the CASP experiment.

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein structures – typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy – is lagging far behind the output of protein sequences.

Proteins are the molecular workhorses of all known biological systems. Among other functions, they are the motors that cause muscle contraction, the catalysts that drive life-

sustaining chemical processes, and the molecules that hold cells together to form tissues and organs.

The following is a list of a few of the diverse biological processes mediated by proteins:

- Proteins called enzymes catalyze vital reactions, such as those involved in metabolism, cellular reproduction, and gene expression.
- Regulatory proteins control the location and timing of gene expression.
- Cytokines, hormones, and other signaling proteins transmit information between cells.
- Immune system proteins recognize and tag foreign material for attachment and removal (1).
- Structural proteins prevent cells from collapsing on themselves, as well as forming large structures such as hair, nails, and the protective, largely impermeable outer layer of skin. They also provide a framework along which molecules can be transported within cells.

The estimate of the number of genes in the human genome has been changing dramatically since it was annotated. Each gene encodes one or more distinct proteins. The total number of distinct proteins in the human body is larger than the number of genes due to alternate splicing. Of those, only a small fraction have been isolated and studied to the point that their purpose and mechanism of activity is well understood. If the functions and relationships between every protein were fully understood, there will most likely be a much better understanding of how our bodies work and what goes wrong in diseases such as cancer, amyotrophic lateral sclerosis, Parkinson's, heart disease and many others. As a result,

protein science is a very active field. As the field has progressed, computer-aided modeling and simulation of proteins have found their place among the methods available to researchers.

Protein Structure

An amino acid is a simple organic molecule consisting of a basic, amine group bound to an acidic carboxyl group via a single intermediate carbon atom:

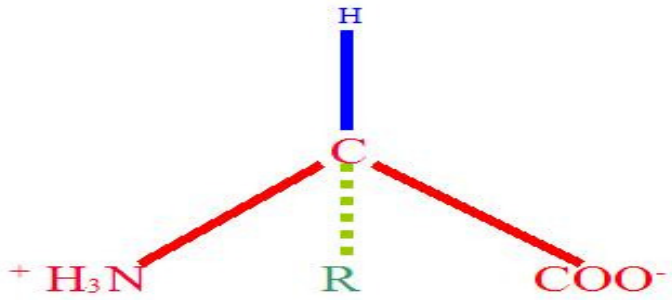


Figure 1. A generic α-amino acid.

Figure 1 shows a generic α-amino acid. The "R" group is variable, and is the only difference between the 20 common amino acids. This form is called a zwitterion, because it has both positive and negatively charged atoms. The zwitterionic state results from the amine group (NH_2) gaining a hydrogen atom from solution, and the acidic group (COO) losing one.

During the translation of a gene into a protein, the protein is formed by the sequential joining of amino acids end-to-end to form a long chain-like molecule also known as a polymer. A polymer of amino acids is often referred to as a polypeptide. The genome is capable of coding for 20 different amino acids whose chemical properties depend on the composition of their side chains represented by "R" in Figure 1. Thus, to a first

approximation, a protein is nothing more than a sequence of these amino acids. A more proper term to use would be amino acid residues, because both the amine and acid groups lose their acid and base properties respectively when they are part of a polypeptide. The sequence is called the primary structure of the protein.

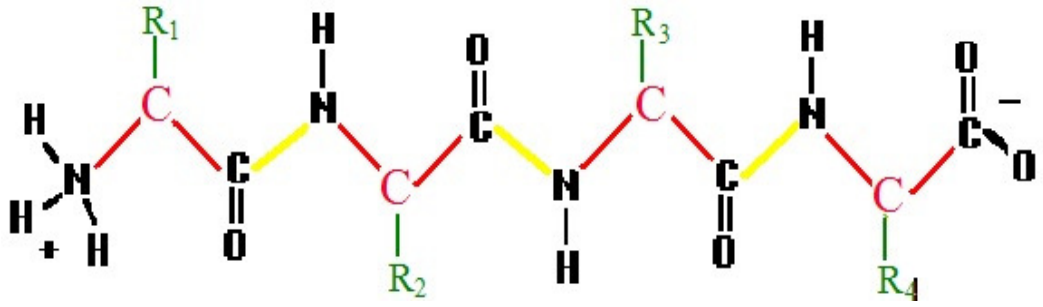


Figure 2. A generic polypeptide chain.

The primary structure of a protein is easily obtainable from its corresponding gene sequence, as well as by experimental manipulation. Unfortunately, the primary structure is only indirectly related to the protein's function. In order to work properly, a protein must fold to form a specific three-dimensional shape, called its native structure or native conformation. The three dimensional structure of a protein is usually understood in a hierarchical manner. A secondary structure refers to folding in a small part of the protein that forms a characteristic shape. The most common secondary structure elements are α-helices and β-sheets, one or both of which are present in almost all natural proteins.

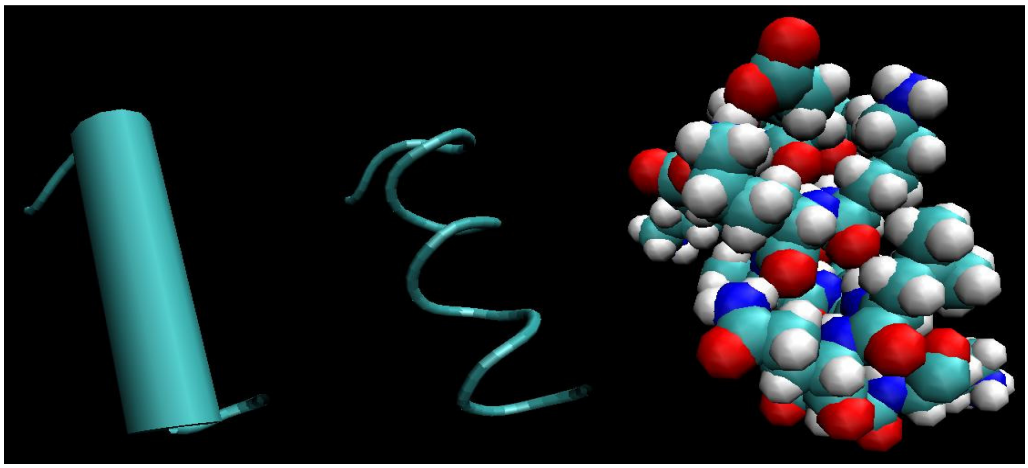
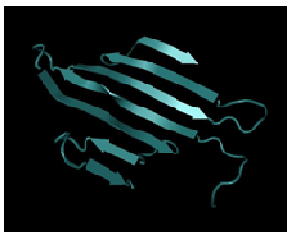


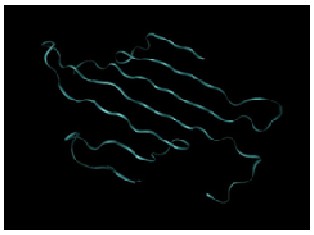
Figure 3. α-helices, rendered three different ways.

Figure 3 shows α-helices rendered three different ways. On the left is a typical cartoon rendering, in which the helix is depicted as a cylinder. The center shows a trace of the backbone of the protein. Right shows a space-filling model of the helix, and is the only rendering that shows all atoms including those on the side chains.

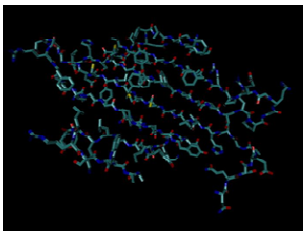
Table 1. Beta-sheets: cartoon, ribbon, and bond representations



Different parts of the polypeptide strand align with each other to form a β-sheet. This β-sheet is anti-parallel, because adjacent segments of the protein run in opposite directions.



β-sheets are sometimes referred to as β pleated sheets, because of the regular zigzag of the strands evident in this representation.



The alignment of oxygen atoms (red) toward nitrogen atoms(blue) is due to hydrogen bonding, the primary interaction involved in stabilizing secondary structure.

Tertiary structure refers to structural elements formed by bringing more distant parts of a chain together into structural domains. The spatial arrangement of these domains with respect to each other is also considered part of the tertiary structure. Finally, many proteins consist of more than one polypeptide folded together, and the spatial relationship between these separate polypeptide chains is called the quaternary structure. It is important to note that the native conformation of a protein is a direct consequence of its primary sequence and its chemical environment, which for most proteins is either aqueous solution with a biological pH which is roughly neutral, or the oily interior of a cell membrane. Nevertheless, no reliable computational method exists to predict the native structure from the amino acid sequence, and this is a topic of ongoing research. Thus, in order to find the native structure of a protein, experimental techniques are deployed. The most common approaches are outlined in the next section.

Protein 3D structures

The Protein Data Bank (PDB) is a repository for 3-D structural data of proteins and nucleic acids. Most of the three-dimensional macromolecular structure data in the Protein Data Bank were obtained by one of three methods: X-ray crystallography (2), solution nuclear magnetic resonance (NMR) or theoretical modeling. The first two are experimental methods.

X-ray crystallography

The most obvious way to determine the shape of an object is to look at it. If the object is small, a microscope can be used. But there is a limit to how small an object can be seen

under a light microscope. The limit is that it is not possible to image things that are much smaller than the wavelength of light that is being used. The wavelength for visible light is measured in hundreds of nanometers, while atoms are separated by distances of the order of 0.1 nanometer or 1 angstrom (Å). A look at the electromagnetic spectrum shows that X-rays have the right wavelength range to observe atoms. An X-ray microscope cannot just be built to look at molecules. There are a couple of reasons for this, one is there is no X-ray lens, the other is even if there was such a lens, it would have to be made with tolerances significantly less than the distance between two atoms. However, an X-ray lens can be simulated on a computer. The microscope can be thought of as working in two stages. First, light strikes the object and is diffracted in various directions. The diffracted rays are collected by the lens and reassembled to form an image. In the case of X-rays, the diffraction from the molecules can be detected, but a computer has to be used to reassemble the image. This is the essence of the method, even though it is not as simple as described above. Other types of waves with wavelengths in the correct range can be considered. One of the unexpected results of quantum mechanics is that particles have a wave nature. The faster the particles are moving the shorter the wavelength. Two types of particles can be accelerated to speeds sufficient to bring their wavelengths into the Angstrom range: neutrons and electrons. Neutron diffraction works more or less like X-ray diffraction.

Electrons can diffract too, but they can also be focused by magnetic fields, which allows the construction of electron microscopes. The very best electron microscopes have resolving power near atomic resolution. It turns out that electron microscopy tends to be most useful for very large assemblies, which is where crystallography tends to become very difficult, so the two techniques are quite complementary.

Often, X-rays emitted from copper targets bombarded with high energy electrons which emit at several characteristic wavelengths are used. The one that is used is called CuKα, which has a wavelength of 1.5418Å. This is well suited to the study of molecular structure as it is very similar to the distance between bonded carbon atoms. The result of a crystallographic experiment is not really a picture of the atoms, but a map of the distribution of electrons in the molecule, also called the electron density map. The electron density map gives a pretty good picture of the molecule, since the electrons are mostly tightly localized around the nuclei. X-ray scattering from a single molecule would be incredible weak and extremely difficult to detect above the noise level, which would indicate scattering from air and water. A crystal arranges a large number of molecules in the same orientation, so that scattered waves can add up in phase and raise the signal to a measurable level. In essence, the crystal acts as an amplifier.

There are a number of potential bottlenecks in determining a crystal structure, but growing a useful crystal can be the most serious one. Apart from growing useful crystals, the phase problem is often the most serious bottleneck in determining a new structure. Because the density map doesn't resolve individual atoms, fitting models to density is a bit of an art. It requires the use of computer programs. An atomic model can never be perfect, but it can be improved a great deal by a process called refinement, in which the atomic model is adjusted to improve the agreement with the measured diffraction data. The Ramachandran plot is a good indicator of the quality of a structure.

Solution nuclear magnetic resonance (NMR)

Nuclear magnetic resonance spectroscopy is a powerful and theoretically complex analytical tool. Protein NMR spectroscopy provides an important complement to X-ray crystallography for structural genomics, both for determining three-dimensional protein structures and in characterizing their biochemical and biophysical functions (3). The following subsections cover the theory behind the technique. The experiments are performed on the nuclei of atoms and not the electrons. The chemical environment of specific nuclei is deduced from information obtained about the nuclei.

Several features of solution state NMR make it particularly suitable for structure-function analysis and structural genomics. Structural analysis by NMR does not require protein crystals. Nearly 75% of the NMR structures in the Protein Data Bank (PDB) do not have corresponding crystal structures, and many of these simply do not provide diffraction quality crystals. Moreover, NMR studies can be carried out in aqueous solution under conditions quite similar to the physiological conditions under which the protein normally functions. This feature allows comparisons to be made between subtly different solution conditions that may modulate structure-function relationships. While most crystal structures are determined under physiologically relevant conditions, in many cases somewhat exotic solution conditions are required for crystallization.

The accuracy of protein structures determined by NMR is very dependent on the extent and quality of data that can be obtained. The highest quality NMR structures have accuracies comparable to 2.0-2.5 Å x-ray crystal structures (4; 5). Although atomic positions in high-resolution crystal structures are more precisely determined than in the corresponding NMR structures, the crystallization process may select a subset of conformers present under

solution conditions. NMR is particularly valuable in structural genomics for analyzing protein structures that are outside the scope of crystallographic studies. Included in the classes of proteins that do not form crystals suitable for crystallographic analysis are those that are partially unfolded in the absence of binding partners, as well as some membrane-associated proteins that can be studied in micelle environments using solution state NMR. Solid state NMR methods can also provide structural information for some integral membrane proteins that may not be accessible by crystallographic methods.

NMR spectroscopy is relatively insensitive, which severely limits experimental design. Typically samples around 1mM protein concentration are required, preventing studies of proteins with very low solubilities. Because of constraints on pulse sequence design arising from these sensitivity limitations, several different NMR spectra recorded over a four to six week period are necessary to obtain the information needed for a high-quality structure determination. These long data collection periods, in turn, put significant constraints on sample stability. Although multiple samples can be used in the structure determination process, each one must be stable for days to weeks with respect to precipitation, aggregation, and other forms of degradation. Manual analysis of these multiple NMR data sets is laborious and requires significant expertise. Another important limitation of NMR analysis is that the density of constraints is sometimes inadequate for accurate structural analysis. In particular, general methods for cross validation analogous to a free R-factor, a statistical measurement used in crystallographic studies to evaluate how well a structural model fits the diffraction data, are not yet available.

Recent technological advances

The reduction of the data collection time required for a structure determination is a major challenge for NMR-based structural genomics. Technological advances enhancing sensitivity, such as the construction of new high-field magnets are of keen interest. The sensitivity of the acquired NMR data depends critically on the performance of the NMR probe, a sophisticated electronic device used to detect NMR signals. In the near future, the introduction of cryogenic probes is expected to have a significant impact. Radiofrequency (RF) coils constitute the heart of these probes, and their sensitivity scales with the thermal noise associated with the coil's temperature. Cryogenic probes utilize RF-coils cooled to around 25 K, and the resulting sensitivity enhancement reduces instrument time requirements by factors that range from 4 to 16. Another key advance involves partial deuteration, providing samples that can be studied with improved signal-to-noise ratios that result from their sharper line widths and longer transverse relaxation times. The combination of partial deuteration and cryogenic probes can provide a factor of 10 or more reduction in the requisite data collection times. These technologies provide the basis for high throughput NMR, and are particularly valuable for samples exhibiting limited stabilities and/or low solubilities. NMR structure determinations rely on the nearly complete assignment of chemicals shifts (6), which are obtained using multidimensional ^{13}C, ^{15}N, ^{1}H-triple resonance NMR methods.

Another important area of development involves automated analysis of NMR data. It has been recognized for some time that many of the interactive tasks carried out by an expert in the process of spectral analysis could, in principle, be carried out more efficiently and rapidly by computational systems. Recent developments provide automated analysis of NMR assignments and three-dimensional structures of proteins ranging from around 50 to 200

amino acids. When good quality data are available, automated analysis of protein NMR data can be very rapid. Many of the available resonance assignment programs execute in tens of seconds, and automated structure refinements are being carried out in tens of minutes using arrays of processors for coarse-grain parallel calculations. However, while progress over the last few years is encouraging, more work is required, even for small proteins, before automated analysis of side chain resonance assignments are not yet well developed, and there are as yet no examples of completely automated protein structure determinations. Moreover, little work has focused on the specific problems associated with nucleic acid structure determinations. It is the intention of this work to address some of these issues.

CHAPTER 3. COMPUTATIONAL BACKGROUND

Theoretical modeling

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry. Its aim is the prediction of the three-dimensional structure of proteins from their amino acid sequences, sometimes including additional relevant information such as the structures of related proteins. In other words, it deals with the prediction of a protein's tertiary structure from its primary structure. Protein structure prediction is of high importance in medicine and biotechnology. Every two years, the performance of current methods is assessed in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment.

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structure genomics, the output of experimentally determined protein structures – typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy – is lagging far behind the output of protein sequences.

A number of factors exist, that make protein structure prediction a very difficult task. The two main problems are, the number of possible protein structures is extremely large, and that the physical basis of protein structural stability is not fully understood. As a result, any protein structure prediction method needs a way to explore the space of possible structures efficiently, which can be a search strategy, and a way to identify the most plausible structure.

In comparative structure prediction, the search space is pruned by the assumption that the protein in question adopts a structure that is reasonably close to the structure of at least one known protein. In de novo or ab-initio structure prediction, no such assumption is made, which results in a much harder search problem. In both cases, an energy function is needed to recognize the native structure, and to guide the search for the native structure. Unfortunately, the construction of such an energy function is to a great extent an open problem.

Direct simulation of protein folding in atomic detail, via methods such as molecular dynamics with a suitable energy function, is typically not tractable due to the high computational cost, despite the efforts of distributed computing projects such as Folding@home. Therefore, most de novo structure prediction methods rely on simplified representations of the atomic structure of proteins.

The above mentioned issues apply to all proteins, including well-behaving, small, monomeric proteins. In addition, for specific proteins such as multimeric proteins and disordered proteins, the following issues also arise:

1. Some proteins require stabilization by additional domains or binding partners to adopt their native structure. This requirement is typically unknown in advance and difficult to handle by a prediction method.

2. The tertiary structure of a native protein may not be readily formed without the aid of additional agents. For example, proteins known as chaperones are required for some proteins to properly fold. Other proteins cannot fold properly without modifications such as glycosylation.

3. A particular protein may be able to assume multiple conformations depending on its chemical environment.

4. The biologically active conformation may not be the most thermodynamically favorable.

Due to the increase in computer power, and especially new algorithms, much progress is being made to overcome these problems. However, routine de novo prediction of protein structures, even for small proteins, is still not achieved.

Ab initio protein modeling

Ab initio- or de novo- protein modeling methods (7; 8) seek to build three-dimensional protein models "from scratch", that is based on physical principles rather than on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search for possible solutions, for example, by applying global optimization of a suitable energy function. These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. Prediction of protein structure de novo for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers or distributed computing. Although these computational barriers are vast, the potential benefit of structural genomics makes ab initio structure prediction an active research field.

Comparative protein modeling

Comparative protein modeling (9) uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has

been suggested that there are only around 2000 distinct protein folds in nature, though there are many millions in different proteins. These methods may also be split into two groups.

Homology modeling is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved that its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modeling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment. Unsurprisingly, homology modeling is most accurate when the target and template have similar sequences.

Protein threading scans the amino acid sequence of an unknown structure against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. This type of method is also known as 3D-1D fold recognition due to its compatibility analysis between three-dimensional structures and linear protein sequences. This method has also given rise to methods performing an inverse folding search by evaluating the compatibility of a given structure with a large database of sequences, thus predicting which sequences have the potential to produce a given fold.

Side chain geometry prediction

Even structure prediction methods that are reasonable accurate for the peptide backbone often get the orientation and packing of the amino acid side chains wrong. Methods that specifically address the problem of predicting side chain geometry include dead-end

elimination and the self-consistent mean field method. Both discretize the continuously varying dihedral angles that determine a side chain's orientation relative to the backbone into a set of rotamers with fixed dihedral angles. The methods then attempt to identify the set of rotamers that minimize the model's overall energy. Rotamers are the side chain conformations with low energy. Such methods are most useful for analyzing the protein's hydrophobic core, where side chains are more closely packed. They have more difficulty addressing the looser constraints and higher flexibility of surface residues.

Molecular modeling

Molecular modeling is a collective term that refers to theoretical methods and computational techniques to model or mimic the behavior of molecules. The techniques are used in the fields of computational chemistry, computational biology and materials science for studying molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The simplest calculations can be performed by hand, but inevitably computers are required to perform molecular modeling of any reasonably sized system. The common feature of molecular modeling techniques is the atomistic level description of the molecular systems. The lowest level of information is individual atoms or a small group of atoms. This is in contrast to quantum chemistry which is also known as electronic structure calculations, where electrons are considered explicitly. The benefit of molecular modeling is that it reduces the complexity of the system, allowing many more particles (atoms) to be considered during simulations.

Molecular mechanics is one aspect of molecular modeling, as it is refers to the use of classical mechanics/Newtonian mechanics to describe the physical basis behind the models. Molecular models typically describe atoms (nucleus and electrons collectively) as point charges with an associated mass. The interactions between neighboring atoms are described by spring-like interactions (representing chemical bonds) and van der Waals forces. The Lennard-Jones potential is commonly used to describe van der Waals forces. The electrostatic interactions are computed based on Coulomb's Law. Atoms are assigned coordinates in Cartesian space or in internal coordinates, and can also be assigned velocities in dynamical simulations. The atomic velocities are related to the temperature of the system, a macroscopic quantity. The collective mathematical expression is known as a potential function and is related to the system internal energy (U), a thermodynamic quantity equal to the sum of potential and kinetic energies. Methods which minimize the potential energy are known as energy minimization techniques (e.g., steepest descent and conjugate gradient), while methods that model the behaviour of the system with propagation of time are known as molecular dynamics.

$$E = E_{bonds} + E_{angle} + E_{dihedral} + E_{non-bonded}$$

$$E_{non-bonded} = E_{electrostatic} + E_{vanderWaals}$$

This function, referred to as a potential function, computes the molecular potential energy as a sum of energy terms that describe the deviation of bond lengths, bond angles and torsion angles away from equilibrium values, plus terms for non-bonded pairs of atoms describing van der Waals and electrostatic interactions. The set of parameters consisting of

equilibrium bond lengths, bond angles, partial charge values, force constants and van der Waals parameters are collectively known as a force field. Different implementations of molecular mechanics use slightly different mathematical expressions, and therefore, different constants for the potential function. The common force fields in use today have been developed by using high level quantum calculations and/or fitting to experimental data. The technique known as energy minimization is used to find positions of zero gradient for all atoms, in other words, a local energy minimum. Lower energy states are more stable and are commonly investigated because of their role in chemical and biological processes. A molecular dynamics simulation, on the other hand, computes the behaviour of a system as a function of time. It involves solving Newton's laws of motion, principally the second law,

F = ma , where F is force, m is mass and a is acceleration

Integration of Newton's laws of motion, using different integration algorithms, leads to atomic trajectories in space and time. The force on an atom is defined as the negative gradient of the potential energy function. The energy minimization technique is useful for obtaining a static picture for comparing between states of similar systems, while molecular dynamics provides information about the dynamic processes with the intrinsic inclusion of temperature effects.

Molecules can be modelled either in vacuum or in the presence of a solvent such as water. Simulations of systems in vacuum are referred to as *gas-phase* simulations, while those that include the presence of solvent molecules are referred to as *explicit solvent*

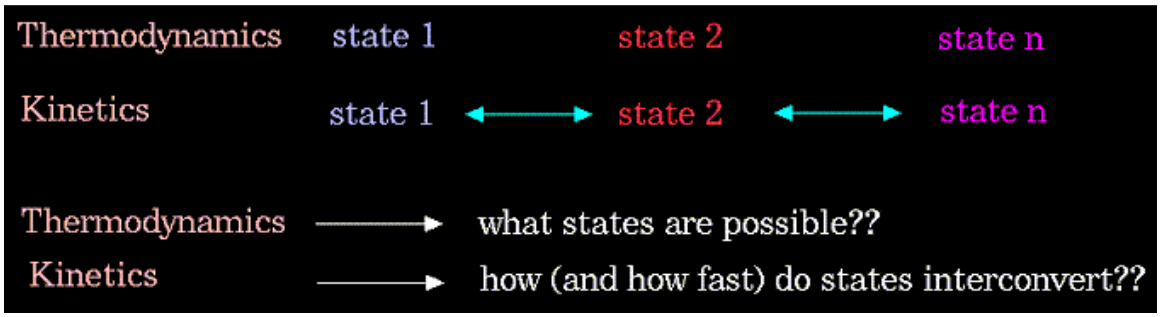
simulations (10). In another type of simulation, the effect of solvent is estimated using an empirical mathematical expression; these are known as *implicit solvation* simulations.

Molecular modelling methods are now routinely used to investigate the structure, dynamics and thermodynamics of inorganic, biological, and polymeric systems. The types of biological activity that have been investigated using molecular modelling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Molecular dynamics

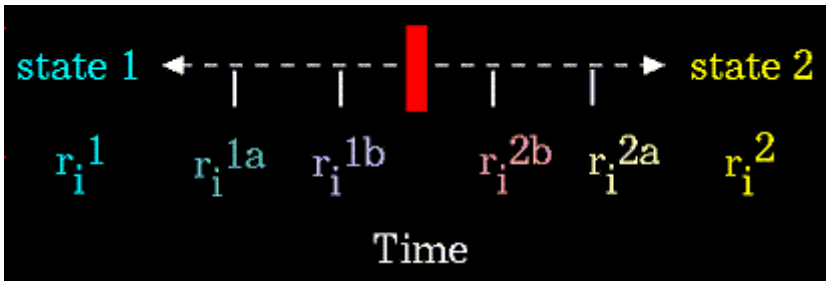
In the broadest sense, molecular dynamics (11; 12) is concerned with molecular motion. Motion is inherent to all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to infrared spectra. Chemical reactions, hormone-receptor binding, and other complex processes are associated with many kinds of intra- and intermolecular motions.

The driving force for chemical processes is described by thermodynamics. The mechanism by which chemical processes occur is described by kinetics. Thermodynamics dictates the energetic relationships between different chemical states, whereas the sequence or rate of events that occur as molecules transform between their various possible states is described by kinetics:



Conformational transitions and local vibrations are the usual subjects of molecular dynamics studies . Molecular dynamics alters the intramolecular degrees of freedom in a step-wise fashion, analogous to energy minimization. The individual steps in energy minimization are merely directed at establishing a down-hill direction to a minimum. The steps in molecular dynamics, on the other hand, meaningfully represent the changes in atomic position r_i over time, that is velocity.

For the "i" atoms of the system:



Newton's equation is used in the molecular dynamics formalism to simulate the atomic motion:

$$\text{force} = \text{mass} \times \text{acceleration} \quad (F_i = m_i a_i)$$

The rate and direction of motion (velocity) are governed by the forces that the atoms of the system exert on each other as described by Newton's equation. In practice, the atoms

are assigned initial velocities that conform to the total kinetic energy of the system, which in turn, is dictated by the desired simulation temperature. This is carried out by slowly "heating" the system, which is initially at absolute zero and then allowing the energy to equilibrate among the constituent atoms. The basic ingredients of molecular dynamics are the calculation of the force on each atom, and from that information, the position of each atom throughout a specified period of time, which is typically on the order of picoseconds.

The force on an atom can be calculated from the change in energy between its current position and its position a small distance away. This can be recognized as the derivative of the energy with respect to the change in the atom's position:

$$-\frac{dE}{dr_i} = F_i$$

Energies can be calculated using either molecular mechanics or quantum mechanics methods. Molecular mechanics energies are limited to applications that do not involve drastic changes in electronic structure such as bond making/breaking. Quantum mechanical energies can be used to study dynamic processes involving chemical changes. The latter technique is extremely novel, and of limited availability. CHARMM is an example of such a program.

Knowledge of the atomic forces and masses can then be used to solve for the positions of each atom along a series of extremely small time steps on the order of femtoseconds. The resulting series of snapshots of structural changes over time is called a trajectory. The use of this method to compute trajectories can be more easily seen when Newton's equation is expressed in the following from:

$$-\frac{dE}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

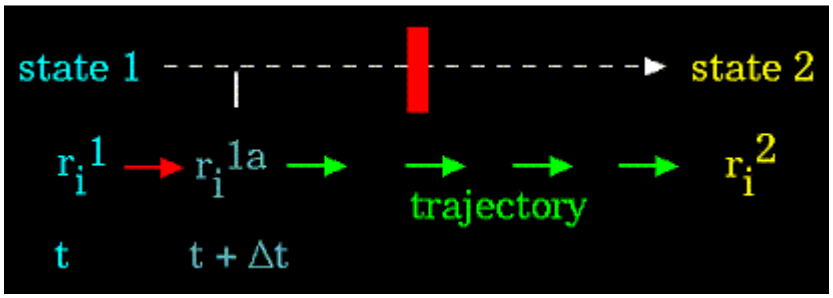
In practice, trajectories are not directly obtained from Newton's equation due to lack of an analytical solution. First, the atomic accelerations are computed from the forces and masses. The velocities are next calculated from the accelerations based on the following relationship:

$$a_i = \frac{dv_i}{dt}$$

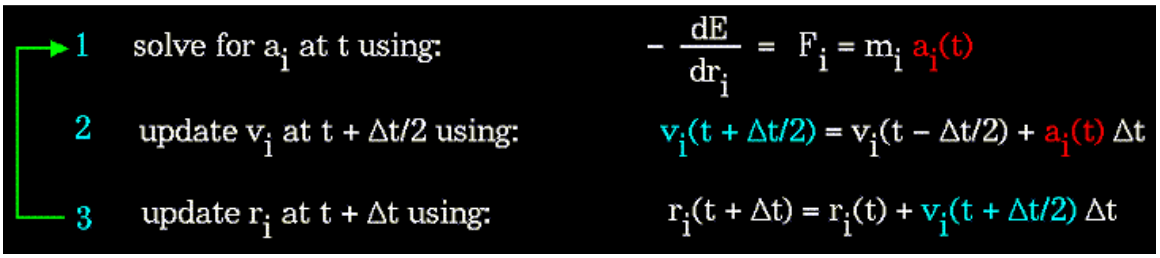
Lastly, the positions are calculated from the velocities:

$$v_i = \frac{dr_i}{dt}$$

A trajectory between two states can be subdivided into a series of sub-states separated by a small time step, "Δ t".



The initial atomic positions at time "t" are used to predict the atomic positions at time "t+ Δ t". The positions at "t+ Δ t" are used to predict the positions at "t+2*Δ t", and so on. The "leapfrog" method is a common numerical approach to calculating trajectories based on Newton's equation. The steps can be summarized as follows:



The method derives its name from the fact that the velocity and position information successively alternate at ½ time step intervals.

Molecular dynamics has no defined point of termination other than the amount of time that can be practically covered. Unfortunately, the current picoseconds order of magnitude limit is often not long enough to follow many kinds of state to state transformations, such as large conformational transitions in proteins. Molecular dynamics calculations can be performed using software tools like CHARMM or GROMACS.

Energy minimization

Energy minimization can repair distorted geometries by moving atoms to release internal constraints. Energy minimization is good to release local constraints, “make room” for a residue, but it will not pass through high energy barriers and stops in a local minima.

Energy minimization methods are common techniques to compute the equilibrium configuration of molecules. The basic idea is that a stable state of a molecular system should correspond to a local minimum of their potential energy. This kind of calculation generally starts from an arbitrary state of molecules, and then the mathematical procedure of optimization allows for the movement of atoms in such a way so as to reduce the net forces, which are the gradients of potential energy, to nearly zero. Like molecular dynamics and Monte-Carlo approaches, periodic boundary conditions have been allowed in energy

minimization methods, to make small systems. A well established algorithm of energy minimization can be an efficient tool for molecular structure optimization.

Unlike molecular dynamics simulations, which are based on Newtonian dynamic laws and allow calculating atomic trajectory with kinetic energy, molecular energy minimization does not include the effect of temperature, and hence the trajectories of atoms during the calculation do not really make any physical sense. That is, only a final state of system that corresponds to a local minimum of potential energy can be obtained. From physical point of view, this final state of the system corresponds to the configuration of atoms when the temperature of system infinitely approximates to zero.

The algorithms of gradient are the most popular methods for energy minimization. The basic idea of gradient methods is to move atoms by the total net forces acting on them. The force on atoms is calculated as the negative gradient of total potential energy of system, as follows:

$$F(r_i) = \overrightarrow{\nabla}_{r_i} U^{\mathrm{tot}}, \quad i = 1, \ldots, N,$$

Where r_i is the position of atom i and U^{tot} is the total potential energy of the system.

An analytical formula of the gradient of potential energy is preferentially required by the gradient methods. If not, one needs to calculate numerically the derivatives of the energy function. In this case, the Powell's direction set method or the downhill simplex method can generally be more efficient than the gradient methods.

Simple gradient method or steepest descent

Here a single function of the potential energy is to be minimized with 3N independent variables, which are the 3 components of the coordinates of N atoms in the system. The net

force on each atom **F** is calculated at each iteration step *t,* and the atoms are moved in the direction of **F** with a multiple factor *k. k* can be smaller at the beginning of calculation if the minimization was started with a very high potential energy. Note that similar strategy can be used in molecular dynamics for reducing the probability of divergence problems at the beginning of simulations.

$$r_i^t = r_i^{t-1} + \kappa \cdot F(r_i), \quad i = 1, \ldots, N.$$

This step in the above equation t = 1,2… is repeated until **F** reaches zero for every atom. The potential energy of the system goes down in a long narrow valley of energy in the procedure. Even though it is also called the "steepest descent", the simple gradient algorithm is in fact very time consuming if it is compared to the conjugated gradient algorithm. It is therefore known as a not very good algorithm. However, its advantage is its numerical stability, that is, the potential energy can never increase if a reasonable *k* is chosen. Thus, it can be combined with a conjugated gradient algorithm for solving the numerical divergence problem when two atoms are too close to each other.

Conjugate gradient method

The conjugate gradient algorithm (13) includes two basic steps: adding an orthogonal vector to the direction of research, and then move them in another direction nearly perpendicular to this vector. These two steps are also well known as: step on the valley floor and then jump down. The following figure shows a highly simplified comparison between the conjugated and the simple gradient on a one dimensional energy curve.

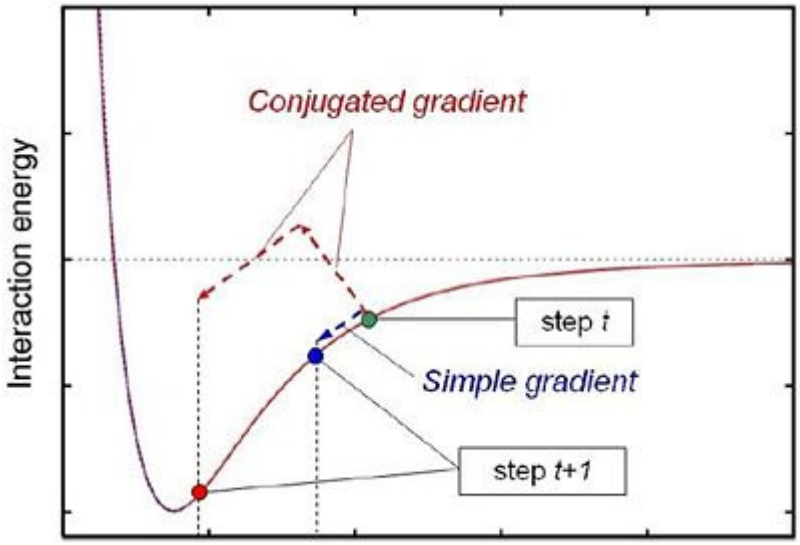


Figure 4. Comparison between two gradient algorithms

In this algorithm, the energy function is minimized by moving the atoms as follows,

$$r_i^t = r_i^{t-1} + \kappa \cdot h_i^t, \quad i = 1, \ldots, N,$$

where

$$h_i^t = F\left(r_i^t\right) + \gamma_i^{t-1} h_i^{t-1}$$

and *gamma* is updated using the Fletcher-Reeves formula as:

$$\gamma_i^{t-1} = \frac{F\left(r_i^t\right) \cdot F\left(r_i^t\right)}{F\left(r_i^{t-1}\right) \cdot F\left(r_i^{t-1}\right)}$$

Here it is to be noted that *gamma* can also be calculated by using the Polak-Ribiere (14) formula, however, it is less efficient than the Fletcher-Reeves (15) one for certain energy functions. At the beginning of calculation (when t = 1), we can make the search direction vector h0 is set as 0.

This algorithm is very efficient. However, it is not quiet stable with certain potential functions, that is, it sometimes can step so far into a very strong repulsive energy range,

when two atoms are too close to each other, where the gradient on this point is almost infinite. It can directly result in a typical data-overrun error during the calculation. For resolving this problem, the conjugated gradient algorithm can be combined with the simple one. The following figure shows the schematics of this combined predicting algorithm. It is to be noted that for implementation, the steps 2 and 5 can be combined to one single step.

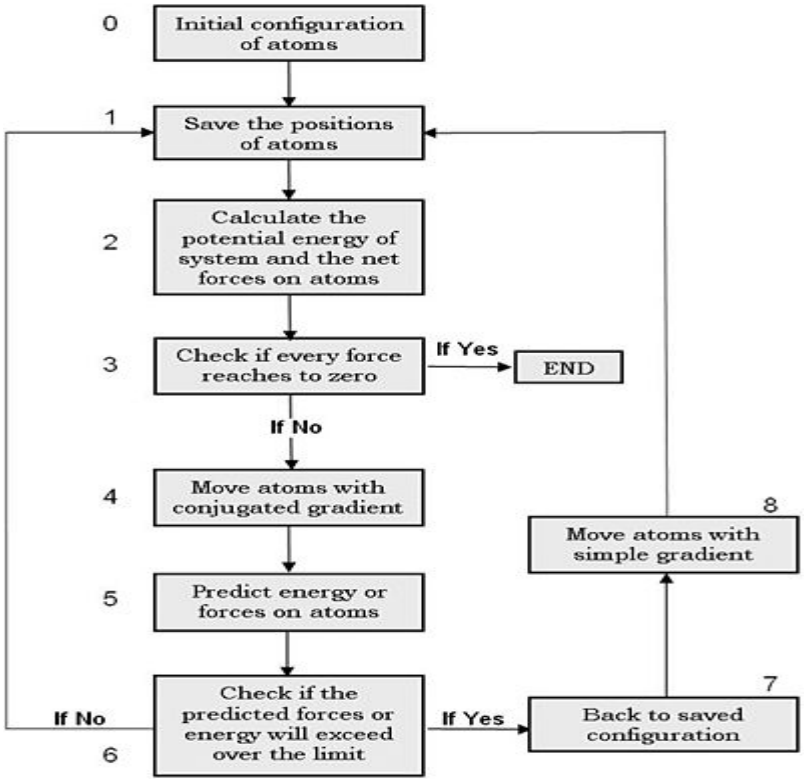


Figure 5. Schematics of a computational energy minimization procedure

Boundary conditions

The atoms in this system can have different degrees of freedom. Moreover, one can equally add other boundary conditions to the minimized energy function, such as adding

external forces or external electric fields to the system. In these cases, the terms in potential energy function will be changed but the number of variables remains constant.

Protein Normal Modes

In the analysis of protein dynamics, an important goal is the description of slow large-amplitude motions (16; 17; 18). These motions, while strongly damped, typically describe conformational changes which are essential for the function of proteins. Only global collective motions can significantly change the exposed surface of the protein and hence influence interactions with its environment. Such structural rearrangements in the protein can occur on a local level within a single domain or can involve large movements of protein domains in a multi-domain protein. Protein dynamics thus cover a broad timescale ranging from 10^{-14}seconds to 10 seconds. However, many large-amplitude conformation changes are not on a timescale accessible by most time dependent theoretical methods, such as phase space sampling techniques or molecular dynamics for example. Therefore, in order to gain insight into the mechanism of slow, large amplitude motions, one must resort to the use of a time independent approach, such as normal mode analysis.

Normal modes of vibration are simple harmonic oscillations (19) about a local energy minimum, characteristic of a system's structure $\vec{R}$ and its energy function, $V(\vec{R})$. For a purely harmonic $V(\vec{R})$, any motion can be exactly expressed as a superposition of normal modes. For an anharmonic $V(\vec{R})$, the potential near the minimum will still be well approximated by a harmonic potential, and any small-amplitude motion can still be well described by a sum of normal modes. In other words, at sufficiently low temperatures, any classical system behaves

harmonically. In a typical normal mode analysis, the characteristic vibrations of an energy minimized system ($T = 0^+ K$) and the corresponding frequencies are determined assuming $V(\vec{R})$ is harmonic in all degrees of freedom. Normal mode analysis is less expensive than MD (molecular dynamics) simulation, but requires much more memory.

As a globular protein is heated from very low temperature, the fluctuations (20) of its atoms begin to deviate measurably from harmonic behavior around 200K. The motion at 300K is considerably an-harmonic. This must be kept in mind when attempting to interpret physiological behavior in terms of normal modes. Still, calculation of the normal mode spectrum is less expensive than a typical MD simulation, and the spectrum may provide qualitative, if not quantitative, insight. The normal mode spectrum of a 3-dimensional system of N atoms contains 3N-6 normal modes (3N-5 for linear molecules in 3D). In general, the number of modes is the system's total number of degrees of freedom minus the number of degrees of freedom that correspond to pure rigid body motion (rotation or translation). Each mode is defined by an eigen vector and its corresponding eigen frequency, $\boldsymbol{\omega}$. The eigen vector contains the amplitude and direction of motion for each atom. In mode *i*, all *N* atoms oscillate at the same frequency, $\boldsymbol{\omega_i}$.

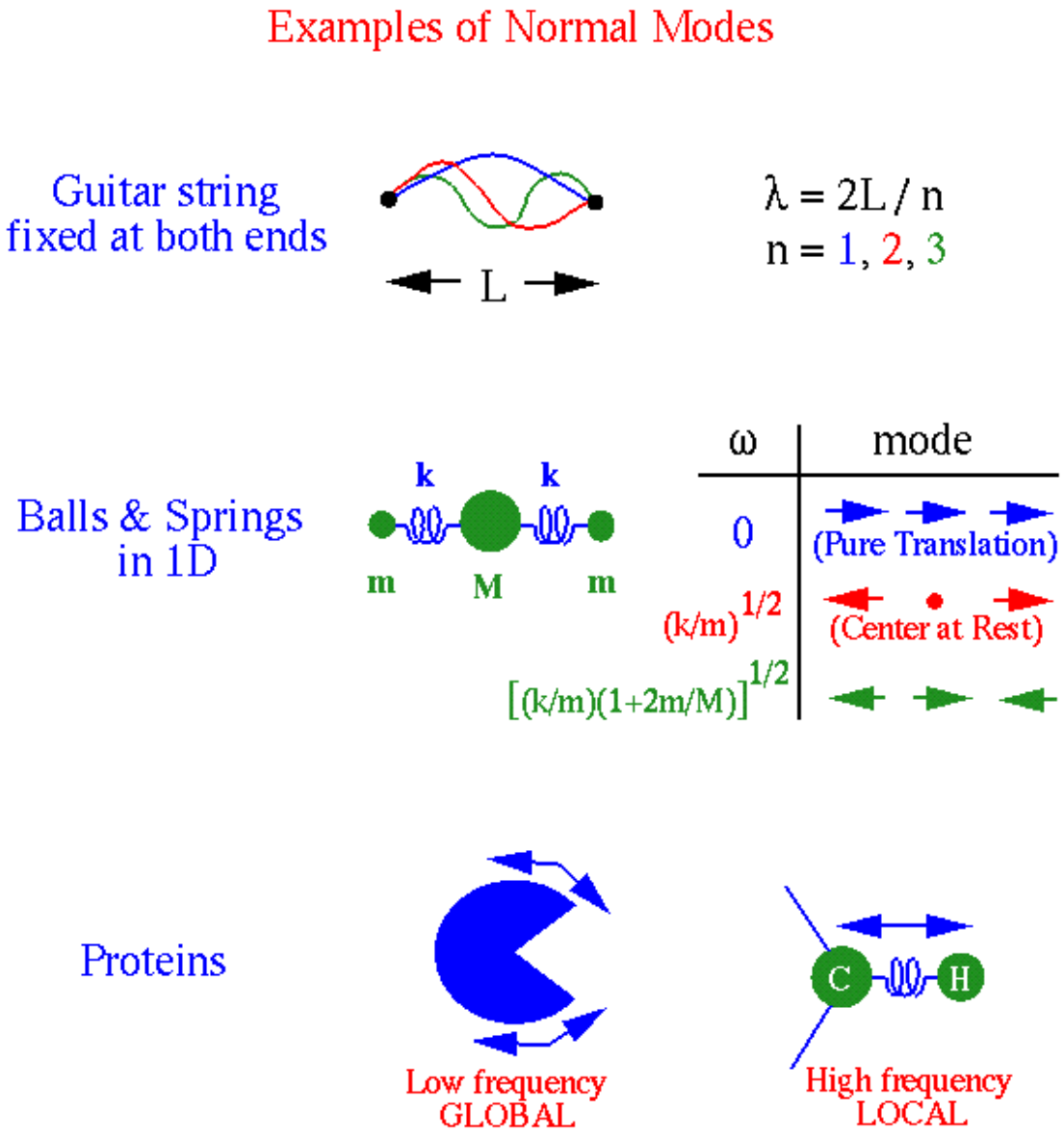


Figure 6. Small oscillations about an equilibrium position

In macromolecules, the lowest frequency modes correspond to delocalized motions, in which a large number of atoms oscillate with considerable amplitude (21). The highest

frequency motions are more localized, with appreciable amplitudes for fewer atoms, for example the stretching of bonds between carbon and hydrogen atoms.

Scoring functions

A key ingredient for the development of a solution to protein folding, protein design, and docking of ligands to protein structures entails the development of a scoring function (22; 23; 24) that can identify the native-like fold of a given sequence of amino acids from a pool of decoy conformations. There exist at least two different types of approaches to this problem: one that uses a scoring scheme based on statistical considerations (25) applied to a database of sequences and structures, and another that uses only energetic considerations to extract the quantities that make up the scoring function (26; 27; 28). In the first approach, to find the most likely structure for a given protein sequence, one first determines the distribution of amino acids in various environments (29) and/or the distribution of the contacts between the 20 types of amino acids in proteins with known tertiary structure (30).Then based on the quasi-chemical approximation and Boltzmann statistics or on Bayes theorem, one converts these distributions into a scoring function. For a given sequence, the structure that corresponds to the best score is considered to be the most native-like conformation. This method has been used in a wide range of problems which include identification of structures from a pool of decoys that can house a sequence of amino acids whose tertiary structure is previously unknown, judging the quality of protein structure models, predicting docking of ligands to protein structures, simulating the folding of a protein, and identifying the native fold of a protein sequence among many incorrect alternatives. In (31) a thorough analysis through a lattice model study of the degree of

accuracy of statistical potentials extracted from protein structures based on Boltzmann statistics and on the quasi-chemical approximation was presented. It was concluded that these potentials are not accurate enough to lead to good predictions regarding the folding of a sequence of amino acids because the method neglects the excluded volume in proteins and the use of the Boltzmann distribution to convert frequencies of contacts between various amino acids into energies of interaction is not firmly grounded.

The second method (32) starts from the idea that the interaction energies between amino acids parametrizing a coarse grained free energy must be such that the energy of a sequence in its own native state is lower than in any other alternative conformation. For each sequence in a data bank, assuming a simple free energy, one obtains a set of linear inequalities involving the unknown interaction parameters. These inequalities can then be solved to obtain the interaction potentials that give an energetic measure of the goodness of the fit between a sequence and a structure. This method is extremely powerful on lattice models. When applied to real proteins, there are difficulties in generating viable alternative conformations that compete significantly with the native structure in housing each of the sequences in the training set. However, using decoy structures obtained by simple gapless threading, the performance of the method is slightly superior to those of previously proposed strategies despite the fact that gapless threading does not produce sufficiently competitive alternatives.

Short-range potentials

Short-range interactions, also termed local interactions, refer to those taking place between near neighbor amino acids along the main chain; they determine the conformational distributions of bond angles and bond torsional states of the backbone. The paper (33) explores the short-range interactions observed in globular proteins. This is a one-dimensional problem, which is suitably analyzed by the tools of linear Ising or Markov chain models, as well as the classical rotational isomeric state approximation of polymer statistics. A set of residue-specific empirical energy parameters is extracted here and used for interpreting experiments and recognizing correct sequence-structure pairs.

4-body contacts

Two-body inter-residue contact potentials for proteins have often been extracted and extensively used for threading. In (34) a new scheme was developed to derive four-body contact potentials as a way to consider protein interactions in a more cooperative model. Several datasets of protein native structures were used to demonstrate that around 500 chains are sufficient to provide a good estimate of these four-body contact potentials by obtaining convergent threading results. Also two sets of protein native structures differing in resolution were deliberately chosen, one with all chains resolution better than 1.5 Å and the other with 94.2% of the structures having a resolution worse than 1.5 Å to investigate whether potentials from well-refined protein datasets perform better in threading. However, potentials from well-refined proteins did not generate statistically significant better threading results. The four-body contact potentials can discriminate well between native structures and partially unfolded or deliberately misfolded structures. Compared with another set of four-body

contact potentials derived by using a Delaunay tessellation algorithm, the four-body contact potentials appear to offer a better characterization of the interactions between backbones and side chains and provide better threading results, somewhat complementary to those found using other potentials.

CHAPTER 4. GENERAL STRUCTURE REFINEMENT

Approach

Significant progress has been made toward the longstanding goal of predicting the structure of proteins from their amino acid sequence with computational methods. In particular, the use of templates from known structures of homologous proteins can routinely generate reliable models of at least the overall fold topology of unknown protein structures. At the same time, the rapid increase in available experimental protein structures, especially from structural genomics efforts, has led to a near complete coverage of protein fold space. Consequently, it is possible to predict the structure of most genes at least to some degree through comparative modeling. Nonetheless, even the best available methods often remain unable to predict structures at a sufficiently high level of accuracy to fully appreciate biological function and to serve as a reliable starting point for rational drug design efforts. Further progress in protein structure prediction therefore depends crucially on improved methods for refining template-based predictions towards experimental accuracy. However, only limited progress has been made in this direction (35; 36; 37; 38).

The general idea was that native structures have the lowest potential energy. The best molecular modeling programs can get the structure close to, and approximate very well the experimental structural values. The near native structure models generated are quite accurate in the general shape of the protein. These methods were hoped to augment or even replace, the experimental determination of a protein structure in cases where the protein is a close

relative of a known structure or experimentally difficult to obtain like the integral membrane proteins.

Typically the models generated by these modeling applications are within the 3-6 Å Cα root mean square deviation (rmsd) range of the true structure. Cα is usually the choice for the root mean square alignment as it reduces the chances of errors and the side chains are quite hard to model. Also a number of the x-ray or other experimental structures do not have coordinates for all the atoms, mostly just the main backbone atoms.

Traversing this seemingly tiny distance between the near native structure of the protein to the native structure has been extremely challenging. This problem of protein structure refinement has turned out to be a major bottleneck in the overall improvement of protein structure prediction.

The most common and popular assumption is that the native structure of the protein is the most energetically favorable structure. This is usually the global minimum of the potential energy function of the protein structure. Potential functions used in structure predictions and refinement can typically be grouped into two general classes: traditional molecular mechanics (MM) potentials and statistically derived knowledge-based (KB) potentials. In both cases, the energy of the system is defined as the sum over energetic terms that are themselves functions of the 3D coordinates of the atoms.

The energy minimization methods described in the previous chapters are suitable to find the local minima. Local minima indicate a preferred state relative to neighboring states on a 3-dimensional plot of energy and space. Using just the minimization protocol it is not possible to tell if the local minima reached is in-fact a global minimum. This means different

approaches have to be taken to be able to cross the local energy barrier if, there exists a lower energy state than the current local minima.

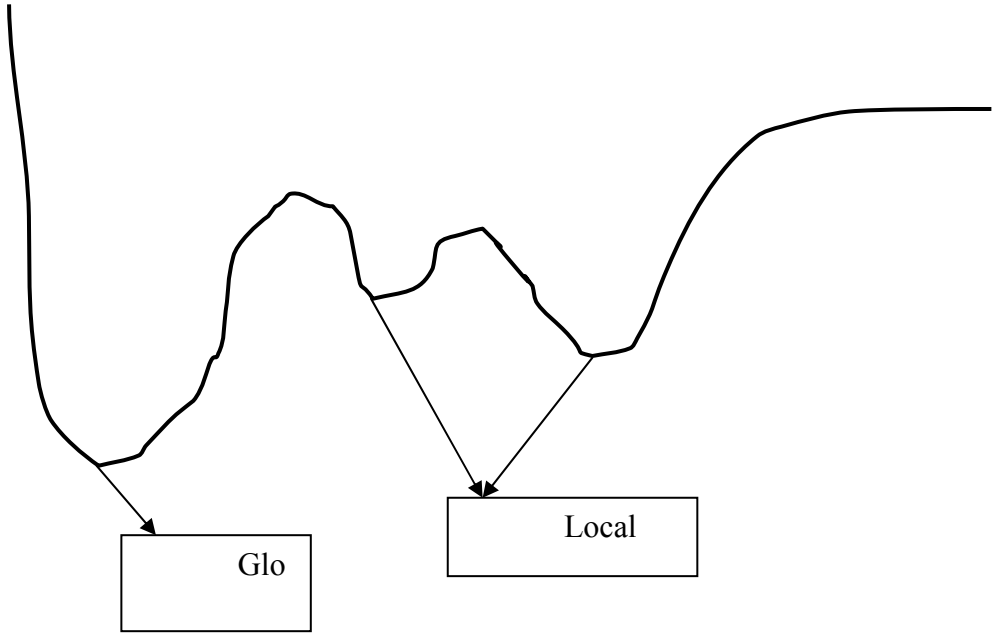


Figure 7. Local and global energy minima

In the actual biological process of protein folding there are various parameters influencing the folding of a protein to its native state. Some include the surrounding chaperone proteins, water molecules in solution. It is however widely believed that the native state of a protein is intrinsic to the protein sequence, since the protein always folds to the same shape for a given sequence. So to reduce the computational costs and computing times, the common practice is to do the simulation in vacuum.

The approach taken by us, in the hope of crossing this energy barrier is to make small modifications to the structure of the protein and then proceed to do energy minimization on

the modified structure. The key to the structural modifications is that it should be large enough that the energy barrier is crossed and also not too large that the basic protein structure itself is modified as this is essentially a protein structure refinement. Since this is a blind structure refinement, we cannot really know for sure if there is an energy barrier and if it exists how large is the energy barrier? So the simplest strategy would be to randomly perturb the structure and hope it gives rise to a structure with a lower potential energy value. Since the new protein structures generated are random, the more structures that are generated the better the probability that a structure which crosses the potential energy barrier is found. This process can be repeated as many times as needed. An important point to note here is that the new structures obtained might actually be worse than the original structure. So for the process to be useful steps must be taken to ensure that this sequence of structure manipulation followed by energy minimization is improving the structure.

However the potential energy functions are not very accurate, even though there have been considerable improvements since they were first used in protein structure determination.

Once it has been determined that further improvements cannot be made based on potential energy minimization, the next approach we decided to use was distance geometry based molecular modeling.

Database Derived Mean-Force Potentials

Wu et al (39) have investigated an alternative, generalized, and in certain sense, improved approach of utilizing the distributions of the protein inter-atomic distances in databases of known protein structures for structure refinement as proposed in Cui et al (40). Instead of extracting the distance ranges from the distributions of the distances, a distribution

function was used to define a mean-force potential for the distance so that the potential is maximized when the probability of the distance in the distribution is maximized. For a selected set of distances, a set of mean-force potentials can be obtained. The sum of the potentials can then be used to define an energy function, and a structure can be refined through energy minimization.

The distances of a specific type are typically distributed in certain range. A range constraint for the distances may be derived by restricting the distances in the most populated range, say in between mean minus and plus two standard deviations. Or, a mean-force potential may be defined for the distances based on the distribution of the distances, e.g., $E = -kT \ln P$, where P is the distribution function, E the potential, T the temperature, and k the Boltzmann constant.

These distances, are the distances between atoms in separated residues in sequence, also called cross-residue inter-atomic distances. Such a distance can be specified by using the types of the two atoms it connects to, the types of the residues the two atoms are associated with, and the types of the residues separating the two end residues in sequence (see Figure). Since the distributions of the distances are non-uniform in general, constraints on the distances can immediately be extracted based on these distributions. As mentioned earlier Cui et all , have derived bound constraints on the distances by using the means minus and plus two standard deviations of the distances as the lower and upper bounds, and applied the constraints to the refinement of NMR-determined protein structures. The advantage of this approach is that the constraints are easy to generate and straightforward to implement with

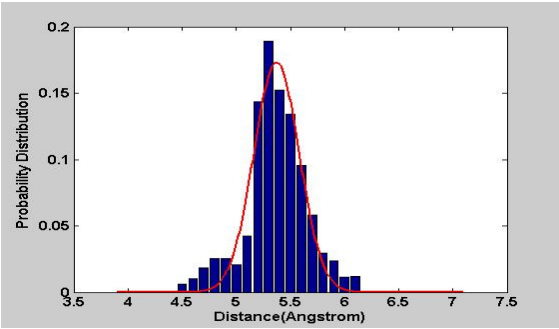


Figure 8. Typical distribution of the distance

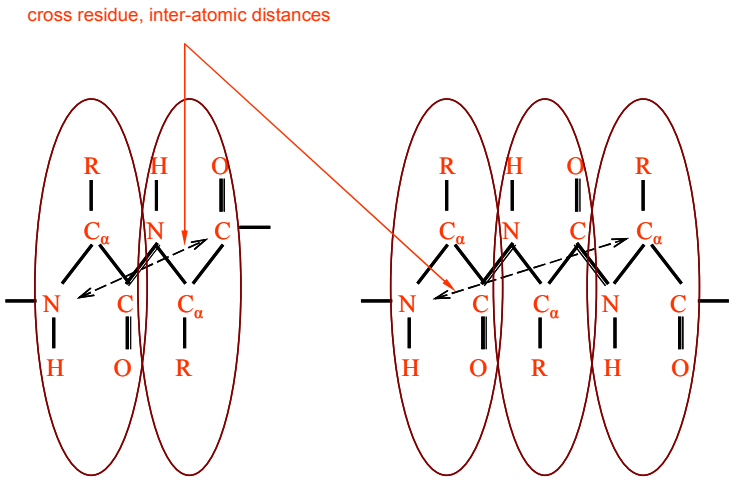


Figure 9. Cross residue, inter-atomic distances

current NMR modeling software such as CNS (41) because they can be applied for structure refinement in the same way as the NOE distance constraints. However, by using simple bounds, the information on the distances demonstrated in the distributions of the distances is not completely exploited, since the constraints exclude the possible distance values outside the bounds and also treat the distance values inside the bounds equally. In fact, the distances outside the bounds are still likely although with only small chances. Also, the distances inside the bounds are obviously distributed non-uniformly and the more probable ones should be considered with higher priorities. A relatively more complete approach is to incorporate

the information in the distribution functions as much as possible to restrict the use of the distances. To this end, for each type of distance, a potential function can be defined by using the distribution function for the distance so that the potential energy is minimized when the distance maximizes the probability distribution. One of such potential function can be defined with the idea of mean-force potentials in the statistical physics (42).

The potentials were then inserted into the energy function of CNS, used in simulated

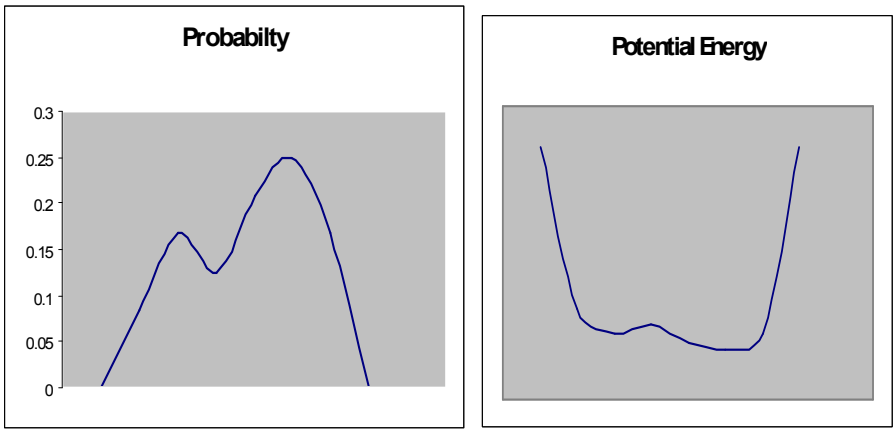


Figure 10. Mean-force potential vs. probability distribution

annealing above. This modified version of CNS was used in refining a selected set of test structures. Using the original NMR data downloaded from the PDB Databank [34], a total of 70 NMR-determined structures were refined. In refining these NMR structures, both the original and the extended energy functions were used. The results were compared to evaluate the effectiveness of the mean-force potentials for the refinement of the structures. Several standard methods were adopted in the comparison of the energy functions, these included the energy values in various different categories such as the bond length energy, the bond angle energy, the NOE energy, etc., the ensemble RMSD of the structures, the RMSD of the

structures against the X-ray reference structures (for available ones), and the Ramachandran plots of the structures. Using these terms, it was found that there was significant improvement of the structures after the refinement with the database derived mean-force potentials. The decreases in the overall energy, NOE and dihedral angle energies indicated that the mean-force potentials helped not only forming more energetically favorable structures but also forcing the structures to fit the experimental constraints even better, which was of great importance to NMR modeling [35].

As described in [25], the distance restraints normally used by CNS for the distance geometry simulated annealing are obtained from experimental data of NMR Spectroscopy. Since the PDB structures obtained in this step of the refinement process are just models and do not have any experimental data associated with them. This means, that the distance restraints can be generated in the format specified in the previous section on distance restraints, and used in place of the experimental NOE restraints normally used. This leads to a common question in NMR structure refinement, what types of distances must be used and what should the distance values be, to obtain a consistent improvement of protein structures across multiple sample structures.

The combination of using potential energy minimization techniques and distance geometry simulated annealing is hoped to give an ensemble of structures which improve upon the given template structure as well as provide further details about the characteristics of the molecule.

Implementation of energy minimization

From various research papers (43) (44) (45) it has been inferred that the native protein structure has one of the lowest energies for a given sequence, when compared with the generated comparative models. But the lowest energy is not necessarily the best structure or the structure obtained from experimental methods, which is also known as the native state. It is hoped that we can reach an energy minima close to the experimental structure using energy minimization.

Two molecular modeling software tools are used to perform energy minimization. One is CHARMM (46; 47) and the other is GROMACS. CHARMM (Chemistry at HARvard Macromolecular Mechanics) is a program for macromolecular simulations, including energy minimization, molecular dynamics and Monte Carlo simulations. GROMACS is a versatile package to simulate the Newtonian equations of motion for systems with a large number of particles. It is primarily designed for biochemical molecules like proteins and lipids that have a lot of complicated bonded interactions.

However, this energy minimization cannot be used just by itself to refine the protein structure. Additional methods are being investigated to help in obtaining a refined protein structure closer to the native state. This minimized protein structure can then be used by distance geometry modeling to refine it further and get it closer in alignment to the experimental values. During the initial design of the algorithm CHARMM was chosen to be the software tool that is going to be used for energy minimization. The input and output data formats for CHARMM were quite specific and different from the starting input data as well the input data format for distance geometry modeling software tools. This required the use of

perl scripts as well as other third party tools for the data conversion. The following sections explain how the parameters for CHARMM were calibrated for optimum results.

Input parameter initialization

Before the energy is calculated and minimization is done, the various force fields and parameters between atoms should be set. The interactions between atoms can be broadly divided into two categories bonded interactions and non-bonded interactions. The information about the bonds is given in the protein structure file (PSF) generated by CHARMM from the input PDB file. The PSF holds lists giving every bond, bond angle, torsion angle, and improper torsion angle as well as information needed to generate the hydrogen bonds and the non-bonded list. It is essential for the calculations of the energy of the system. The non-bonded interactions refer to van der Waals terms and the electrostatic terms between all atom pairs that are not specifically excluded from non-bonded calculations, for example the directly bonded atoms. A few examples are given below:

NBONd CDIE CUTNb 14.5 CTONnb 12.0 CTOFnb 13.5 SWITch – VSWITch EPSilon 1.0

NBONd GROUP RDIE CUTNb 14.5 CTONnb 12.0 CTOFnb 13.5 SWITch – VSWITch EPSilon 1.0

There are two basic methods for electrostatics, ATOM and GROUP. Atom electrostatics indicates that, interactions are computed on an atom by atom pair basis. This is the default. The GROUP based method performs electrostatics based on chemical groups instead of atom pairs. There are two options that specify the radial energy functional form.

The keywords CDIE and RDIE select the basic function form. The SHIFted and SWITched keywords determine the long-range truncations option.

CDIE – Constant dielectric. Energy is proportional to 1/R.

RDIE – Distance dielectric. Energy is proportional to 1/(R-squared).

SWITch – Switching function used from CTONnb to CTOFnb values.

SHIFt – Shifted potential acting to CTOFnb and zero beyond.

Initialization

1. The method to be used.
2. Distance cutoff in generating the list of pairs. CUTNb value.
3. Distance cut at which the switching function eliminates all contributions from a pair in calculating energies. CTOFnb value.
4. Distance cut at which the smoothing function begins to reduce a pairs contribution. This value is not used with SHIFting.

Various options have been tried for the electrostatics terms to investigate which values and parameters give the best results for the protein structure. Some of the results are given below.

The first experiment was choosing a good value for CUTNb, which was the cutoff distance used in generating the atom by atom list of pairs. At the time of testing these values energy was the only criteria available for comparing different protein structures. These values are obtained after two series of minimization, as only one iteration of energy minimization would not produce any significant change in energy or structure. As the table below shows, there was not any improvement in the results by using a distance greater than 15.5 Å. The columns where the RMSD with the experimental structure was calculated

indicated that the protein structure obtained was better if more atom pairs were used, but there was not a significant change in the energy, which was the criteria, for choosing good structures at the time.

Table 2. Energy calculation dependence on atom pairs cutoff distance (CUTNb)

	Comparing RMSD				Comparing Energy			
CUTNB	Min RMS	Energy1	Max RMS	Energy2	Min Ener	RMS3	Max Ener	RMS4
14.5	1.7644	-1197.81	2.3254	-1158.77	-1252.3	2.0881	-1121.82	2.01502
15.5	1.8161	-1199.35	2.3252	-1158.75	-1241.53	1.9168	-1121.79	2.01509
20.5	1.7147	-1193.28	2.3253	-1158.77	-1241.53	1.9168	-1121.79	2.01509

The next experiment was testing to see whether CDIE or RDIE was better at producing good structures by energy minimization. As mentioned previously, for constant dielectric or CDIE as the option indicates, energy is proportional to 1/R. For distance dielectric or RDIE, energy is proportional to 1/(R-squared). The program was run for the two electrostatics options, as well as for two different cutoff distances. The goal was to see if, the total number of atom pairs involved made a difference to the performance. Once again, energy was the only criteria used for comparing two protein structures.

Table 3. Comparison of electrostatic methods

	Distance*	Min Ener	RMS	Energy1*	Min RMS	Energy2*	Max RMS
RDIE	14.5	-1226.17	2.1506	-1207.71	1.7920	-1215.19	2.3545
	9.5	-1204.85	1.9388	-1184.64	1.7996	-1098.37	2.4070
CDIE	14.5	-3253.42	2.5940	-2940.5	1.9801	-3192.88	2.6993
	9.5	-3058.49	2.3853	-2802.77	1.8130	-2796.88	2.4382

* The distances indicated are the CUTNb which is the cutoff distance for calculating atom pairs used in energy calculations.

The above table shows that using RDIE gave better protein structures, even if we used just energy to compare different structures. RDIE option also performed better in

generating the overall structures, when we look at the minimum and maximum rmsd values with the native structure.

The next step then involved investigating the effect of switching and smoothing functions. The following table shows the different cutoff distances used to vary the energy calculations. The greater the cutoff distances the greater the number of atom pairs needed to be taken into energy calculations. From the table we can clearly see that the other structures do not differ much regardless of the cutoff distances used as the variation in either the minimum or maximum rmsd is about 0.3 Å. But when the energy is used to compare different structures there is a variation of almost 0.6 Å. This indicates that the cutoff distances make a different in calculating energies of a protein structure but not much impact on the protein structure itself.

Choosing and setting a set of cutoff distances and a proper electrostatic method initializes the system for structure refinement. An energy minimization is performed for the structure to relax the protein.

Normal modes with energy minimization

As described in the previous chapter, protein normal modes (48) are used to describe the conformational changes in a protein. The normal mode vibrations help in observing the motions in a protein. The goal is to generate new structures which are better than the existing structures. The energy minimization ensures that the protein structure is in an area of local

Table 4. Cutoff distances and structure variation

CUTNB	CTONNB	CTOFNB	Lowest Energy		Highest Energy		Minimum RMSD		Maximum RMSD	
			Energy	RMS	Energy	RMS	Energy	RMS	Energy	RMS
40.5	38.00	39.5	-1244.16	2.141	-1100.98	2.165	-1186.88	1.985	-1116.68	2.409
20.5	18.00	19.5	-1245.94	2.420	-1117.04	2.208	-1124.98	2.004	-1210.24	2.470
14.5	12.00	13.5	-1226.17	2.151	-1121.82	2.015	-1207.71	1.792	-1215.19	2.354
13.5	11.00	12.5	-1254.51	2.326	-1105.4	2.142	-1135.08	2.036	-1141.31	2.546
12.5	10.00	11.5	-1251.14	2.504	-1073.11	2.406	-1142.16	2.188	-1205.65	2.534
11.5	9.00	10.5	-1253.14	1.893	-1074.33	2.342	-1235.73	1.875	-1088.73	2.464
10.5	8.00	9.5	-1214.71	1.995	-1065.71	2.237	-1122.31	1.809	-1108.63	2.365
9.5	7.00	8.5	-1204.85	1.939	-1038.63	2.268	-1184.64	1.799	-1098.37	2.407
7.5	5.00	6.5	-1098.74	2.140	-966.914	2.183	-1051.57	1.923	-1001.17	2.533
6.5	4.00	5.5	-818.812	2.311	-649.775	2.392	-708.956	2.124	-775.849	2.718

The above results were obtained by running CHARMM on 4 processors. A total of 64 structures were produced and compared for each set of cutoff distances.

energy minimum and a reasonably stable state. Once the local minimum is attained, normal mode analysis is performed using the molecular dynamics software to calculate the normal modes (49) and the resulting vibrational amplitudes. These vibrational amplitudes will be referred to as the fluctuations for easier understanding. The fluctuations are totaled over a series of normal modes to obtain an overall fluctuation of the atoms. This gives the extent to which the atoms move away from the equilibrium position without the actual direction. Some atoms have large fluctuations whereas others show hardly any movement. Since the final outcome is supposed to be protein structure refinement, the three-dimensional structure itself does not change much as the target folded structure already approximates the experimental structure quite well.

The fluctuations are available for each atom in the protein, however to reduce the computational costs, each residue is treated as a single unit. To generate the new structures the residues are translated as a whole in the three-dimensional space. This step tweaks the structure a little bit yet retaining the similar overall structure. As the residues have been moved as a single unit they retain their three-dimensional structure. This process is illustrated in Figure 12. This step raises the obvious question of direction and distance of coordinate transformation. Since the residues are treated as a single unit, the Cα atoms of the residues are considered to be the center of each residue. The fluctuation of each Cα atom in the molecule is used as the distance by which each residue is translated in the coordinate space. In order to generate multiple structures a pseudo random number generator is used to obtain the direction in which a residue can be translated.

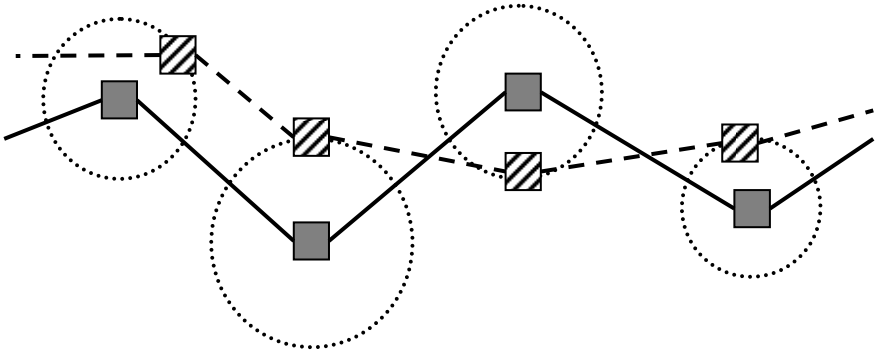


Figure 12. Normal mode perturbation

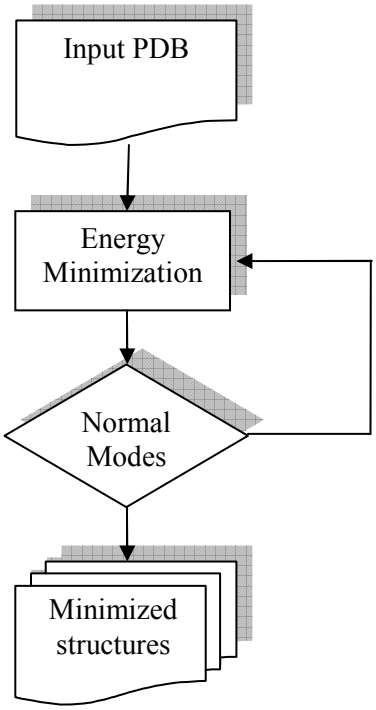


Figure 11. Energy minimization using normal modes

Implementation of distance geometry modeling

Crystallography and NMR System (CNS) is a flexible multi-level hierarchical approach for the most commonly used algorithms in macromolecular structure determination by X-ray crystallography or solution nuclear magnetic resonance (NMR) spectroscopy. The goals of CNS (50) (51) (52) (53) were to create a flexible computational framework for

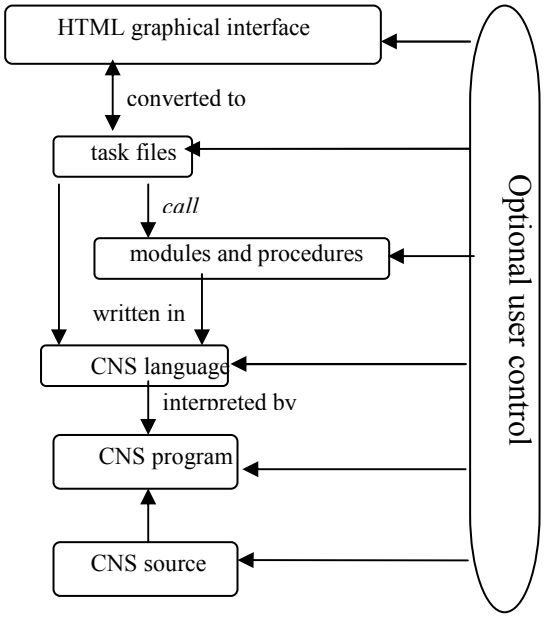
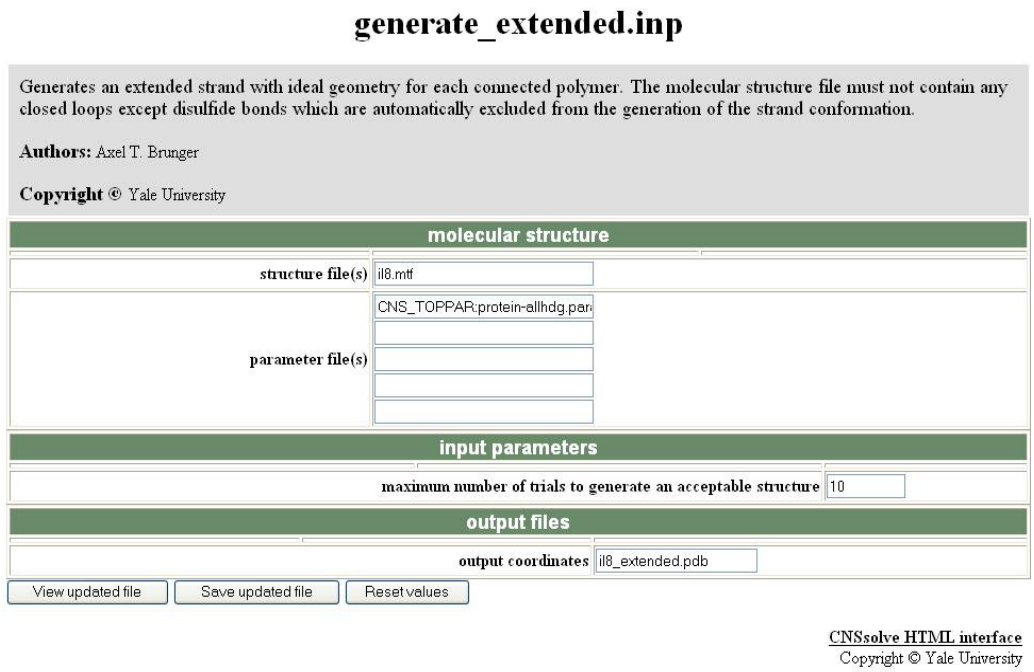


Figure 13. Overview of CNS

exploration of new approaches to structure determination, to provide tools for structure solution of difficult or large structures, to develop models for analyzing structural and dynamical properties of macromolecules, and to integrate all sources of information into all stages of the structure determination process.

CNS consists of five layers which are under user control. The high-level HTML graphical interface interacts with the task oriented input files. The user can edit fields in the form, and then automatically generate the modified task file. The task files make use of a large variety of CNS modules for crystallographic and NMR structure determination. The task and module files all make use of the CNS language, which is plain ASCII text readable by the user. The CNS language is interpreted by the CNS program which is written in Fortran77. The program performs the data manipulations, data operations, and hard-wired algorithms.



generate_extended.inp

Generates an extended strand with ideal geometry for each connected polymer. The molecular structure file must not contain any closed loops except disulfide bonds which are automatically excluded from the generation of the strand conformation.

Authors: Axel T. Brunger

Copyright © Yale University

molecular structure

structure file(s): il8.mtf

parameter file(s): CNS_TOPPAR:protein-allhdg.par

input parameters

maximum number of trials to generate an acceptable structure: 10

output files

output coordinates: il8_extended.pdb

View updated file | Save updated file | Reset values

CNSsolve HTML interface

Copyright © Yale University

Figure 14. CNS HTML form page showing the graphical interface

```
{+ file: generate_extended.inp +}
{+ directory: nmr_calc +}
{+ description: Generates an extended strand with ideal geometry
                for each connected polymer.
                The molecular structure file must not contain any
                closed loops except disulfide bonds which are automatically
                excluded from the generation of the strand conformation.   +}
{+ authors: Axel T. Brunger +}
{+ copyright: Yale University +}

{- begin block parameter definition -} define(

{========================= molecular structure ==========================}

{* structure file(s) *}
{===>} structure_file="il8.mtf";

{* parameter file(s) *}
{===>} par_1="CNS_TOPPAR:protein-allhdg.param";
{===>} par_2="";
{===>} par_3="";
{===>} par_4="";
{===>} par_5="";

{========================= input parameters ============================}

{* maximum number of trials to generate an acceptable structure *}
{===>} max_trial=10;

{============================ output files ============================}

{* output coordinates *}
{===>} output_coor="il8_extended.pdb";

{=========================================================================}
{          things below this line do not normally need to be changed      }
{=========================================================================}

 ) {- end block parameter definition -}

 checkversion 1.1

 evaluate ($log_level=quiet)

 structure @&structure_file end

 parameter
    if (&par_1 # " ") then
       @@&par_1
```

Figure 15. The CNS task file

NMR structure calculation

The part of the CNS that is being used by us is the NMR structure calculation. The NMR structure calculation protocols in CNS consist of four main sections: data input, annealing protocols, acceptance tests and analysis of all NMR structures.

The starting points for the NMR structure calculation and refinement protocols are randomized extended strands corresponding to each disjoint molecular entity (polypeptide chain or oligonucleotide acid strand) or pre-folded structures. The first section of the protocol consists of reading the various data structures. This is followed by an initialization section for statistical analysis of average properties. A constant high-temperature Cartesian or torsion-angle annealing stage follows. This is followed by a slow-cooling stage with either torsion angle or Cartesian dynamics. Finally, an additional Cartesian dynamics cooling stage and a minimization stage follow. A number of trials are performed by starting the simulated-annealing calculation with different randomly selected initial atomic velocities.

Analysis of deviations and violations for the various experimental and chemical restraints is carried out and corresponding to the particular trial. The acceptability of the trial is tested and analysis of average properties carried out. The whole process begins again using different initial velocities (or coordinates) which in general produces a different result.

Initial template generation

This stage is divided into two steps, generating the molecular topology and generating the initial extended coordinates.

The molecular topology information must be first generated for the structure - this contains the information about molecular connectivity. This information is then to be used in

the next step to generate starting (extended conformation) coordinates. The molecular topology is generated from the sequence (not coordinates). This is done with the CNS task file `generate_seq.inp`.

```
cns_solve < generate_seq.inp > generate_seq.out
```

As an example, consider a structure which contains 2 separate chains, thus 2 sequence files are required. This will result in a molecular topology with 2 unconnected chains. In CNS there is no way to specify a break in a chain purely based on the sequence. The 2 sequence files have this format:

```
MET VAL LYS GLN ILE GLU SER LYS THR ALA
PHE GLN GLU ALA LEU ASP ALA ALA GLY ASP
LYS LEU VAL VAL VAL ASP PHE SER ALA THR
TRP CYS GLY PRO ALA LYS MET ILE LYS PRO
PHE PHE HIS SER LEU SER GLU LYS TYR SER
ASN VAL ILE PHE LEU GLU VAL ASP VAL ASP
ASP ALA GLN ASP VAL ALA SER GLU ALA GLU
VAL LYS ALA THR PRO THR PHE GLN PHE PHE
LYS LYS GLY GLN LYS VAL GLY GLU PHE SER
GLY ALA ASN LYS GLU LYS LEU GLU ALA THR
ILE ASN GLU LEU VAL
and
PRO ALA THR LEU LYS ILE CYS SER TRP ASN
VAL ASP GLY
```

The two chains are input as 2 different sequence files and given different segment identifiers. Also, the numbering for the second chain is begun at 106:

```
{* protein sequence file *}

{===>} prot_sequence_infile_1="trx_a.seq";

{* segid *}

{===>} prot_segid_1="A";

{* start residue numbering at *}

{===>} renumber_1=1;


{* protein sequence file *}

{===>} prot_sequence_infile_2="trx_b.seq";
```

```
{* segid *}
{===>} prot_segid_2="B";
{* start residue numbering at *}
{===>} renumber_2=106;
```

It is also important to include any disulphide bonds at this stage - as they require the addition of bond information to the molecular topology. Here there is a bond between the 2 chains (residue 32 to residue 112):

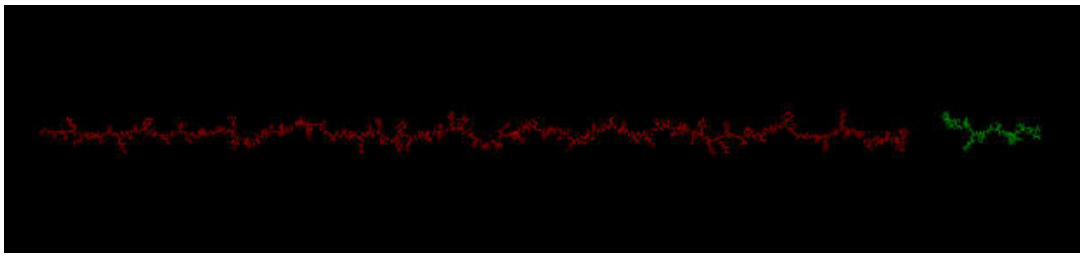
```
{============================ disulphide bonds ==============================}

{* Select pairs of cysteine residues that form disulphide bonds *}
{* First 2 entries are the segid and resid of the first cysteine (CYS A). *}
{* Second 2 entries are the segid and resid of the second cysteine (CYS B). *}
{+ table: rows=8 numbered
   cols=5 "use" "segid CYS A" "resid CYS A" "segid CYS B" "resid CYS B" +}

{+ choice: true false +}
{===>} ss_use_1=true;
{===>} ss_i_segid_1="A"; ss_i_resid_1=32;
{===>} ss_j_segid_1="B"; ss_j_resid_1=112;
```

There is one file generated: an MTF file (this contains the molecular topology information which describes the covalent topology of the molecule).

A starting model for structure calculation is needed. For the following calculations an extended conformation is generated. This provides good local geometry but contains no information about the fold of the structure – this will be generated in the structure calculation stage. The extended conformation is calculated with the CNS task file generate_extended.inp.

```
cns_solve < generate_extended.inp > generate_extended.out
```

The extended conformation is generated from the molecular topology information and initial random coordinates using an extensive series of minimization steps. The output coordinates form an extended conformation. The two separate chains are shown in red and green.



Structure calculation with distance geometry

There are two ways the structure can be calculated, one is simulated annealing, and the other is distance geometry simulated annealing.

In simulated annealing a structure is calculated using experimentally measured inter-proton distance estimates, hydrogen bonds and coupling-constant-derived dihedral angle restraints. This protocol uses ab initio simulated annealing starting from an extended template structure.

In distance geometry simulated annealing a structure is calculated similar to the simulated annealing method. The only difference is, the protocol uses ab initio simulated annealing starting from embedded substructures using distance geometry calculations (based on the experimental data). The experimental data is available for most NMR structures at the protein data bank (PDB). The structure calculation is performed with the CNS task file dg_sa.inp.

```
cns_solve < dg_sa.inp > dg_sa.out
```

In this protocol the extended coordinate template is used as a starting point for generation of an embedded structure. This embedded structure is generated using distance geometry calculations such that the coordinates satisfy the known geometric and experimental distance restraints. The resulting coordinates need to be further regularized with a simulated annealing protocol. The generated structures can be either trial structures or accepted structures. In general it takes a lot longer to generate accepted structures instead of trial structures, as these structures need to pass all the acceptance tests. This implies more trial structures need to be generated.

Distance restraints

NOE distance restraints (54) are specified with the following syntax:

```
ASSIgn atom-selection atom-selection real real real
```

The atom selections define the atoms (or groups of atoms) between which the distance restraint will be applied. The following real numbers determine the parameters of the distance restraint: d (distance), and dminus, and dplus (the extents either side of this distance) respectively.

Example:

assign (resid 112 and name n) (resid 74 and name o) 2.8 0.4 0.9

assign (resid 112 and name hn) (resid 74 and name o) 1.8 0.4 0.9

assign (resid 74 and name n) (resid 112 and name o) 2.8 0.4 0.9

assign (resid 74 and name hn) (resid 112 and name o) 1.8 0.4 0.9

In CNS, the setup of pseudoatoms is accomplished by the ASSIgn statement, with multiple protons in either atom selection. For the restraining functions, CNS computes either an R-6 averaged distance between the involved protons or the distance between the geometric centers of the two specified atom selections. For distance geometry, CNS automatically applies a pseudoatom correction to the specified distance ranges. Pseudoatoms (multiple atom selections) should be used primarily for unresolved NOE cross peaks, like those of methyl groups, prochiral centers, and aromatic rings. In the case of stereospecific assignments, the distances should be exact.

Example:

assign (resid 4 and name HG#) (resid 4 and name HE2#) 4.0 2.2 1.0

assign (resid 4 and name HG#) (resid 4 and name HE2#) 3.0 1.2 1.0

assign (resid 4 and name HA) (resid 4 and name HE2#) 4.0 2.2 1.0

Energy minimization in parallel

The energy minimization and normal mode analysis parts of the system involved some simplified assumptions to reduce the computation involved. Another goal of reducing the computational aspect was to make the algorithm scalable to multiple processors without affecting the processing time significantly.

When an algorithm is modified to work on multiple processors, the usual goal of such a process is to be able to get more processing done per a unit of time than when using a single processor (55). The increased performance of the new parallel algorithm is measured in terms of speedup of the algorithm compared to the original single standalone program. For

example, if the parallel algorithm was implemented on 4 processors, the performance should be 4 times faster than the original single processor algorithm. That is the ideal speedup expected from parallelization of a program.

Parallel computing, explained simply, is the simultaneous use of a number of compute resources to solve a computational problem. These parallel programs are designed to run on multiple processing units. A problem is broken into smaller parts that can be solved concurrently. Each part is further broken into a series of instructions which are executed simultaneously on different processing units. The computational resources used in parallel computing can include a single computer with multiple processors, an arbitrary number of computers connected by a network or a combination of both. The problems usually considered for parallel computing usually have characteristics such as the ability to be broken into discrete pieces that can be solved simultaneously, solved in less time with multiple compute resources than a single compute resource. Parallel computing is an evolution of serial computing that attempts to emulate what has always been the state of affairs in the natural world: many complex interrelated events happening at the same time, yet within a sequence. Traditionally, it has been considered that parallel computing is “the high end of computing” and has been motivated by numerical simulations of complex systems such as weather and climate, seismic activity or chemical and nuclear reactions. Presently, commercial applications are providing an equal or greater driving force in the development of faster computers. These applications require the processing of large amounts of data in sophisticated ways. Some examples include data mining, web search engines or computer

aided diagnosis in medicine. Ultimately, parallel computing is an attempt to maximize the infinite but seemingly scarce commodity called time.

There are different ways to classify parallel computers. One of the more widely used classifications is Flynn's Taxonomy. Flynn's taxonomy distinguishes multi-processor computer architecture according to how they can be classified along the two independent dimensions of instruction and data. Each of these dimensions can have only one of two possible states: single or multiple. The four possible classifications according to Flynn are single instruction single data (SISD), single instruction multiple data (SIMD), multiple instruction single data (MISD), multiple instruction multiple data (MIMD).

SISD is the serial non-parallel computer and MIMD is the traditional parallel computer system. The classification in the table above is a simple and basic differentiation scheme. When designing parallel programs, there are various factors to consider. The factors that were of importance to protein structure refinement using energy minimization and normal mode analysis will be discussed below.

With the wide array of parallel compute resources available as well as the different programming models that a parallel program can be designed for; there is more than one way to go about solving a problem. When the significant time consuming steps are considered for the energy minimization with normal mode analysis, two steps emerge. One of the limiting factors for many programs running on a sequential processing system is the amount of memory available for the computational requirements. But this also closely tied to another limiting factor for the sequential processing system, the processing time. Even though there has been a significant improvement in the processing power of computers the increase in the

problem complexities and problem sizes have been even greater. Higher memory capacities enable researchers to carry out ever increasing computational demands. However the programmers expect similar turnaround times as they did with the smaller less complex problems. The parallel computer memory architectures can be classified as shared memory, distributed memory and hybrid distributed-shared memory. The architectures differ in the way the memory is used and accessed by each processing unit of the parallel computer system. The programs themselves can also be designed to access memory different from the underlying architecture some of which include shared memory model, threads model or message passing model.

Since there is a large amount of data generated and written to files, having a good underlying file system was important. The input for the energy minimization was the starting target protein structure file and the output was a set of multiple protein structure files. The resulting energy minimized three-dimensional protein structures do not necessarily have better structures either structurally or in comparison with the experimental structure as the normal mode perturbations are randomly generated. The starting structure can be improved by either generating more energy minimized structures or have a better protein structure ranking system, or even a combination of both. Generation of more structures is simpler to achieve compared to an improved ranking system for the protein structures. The simplest way to accomplish this would be to use multiple processors doing the same series of energy minimization and normal mode analysis. The increase in the number of new structures will be proportional to the number of processors used. File input and output for the protein structures generated account for a significant portion of the total time taken. This requires a

good implementation of the underlying file system to achieve the benefits of using multiple processing units to generate more protein structures.

The shows the two main time consuming phases of the energy minimization and normal mode analysis steps. To achieve a proportional speedup of the performance for an increase in the number of the processing units, the file system that the generated protein structures are written to should also be independent.

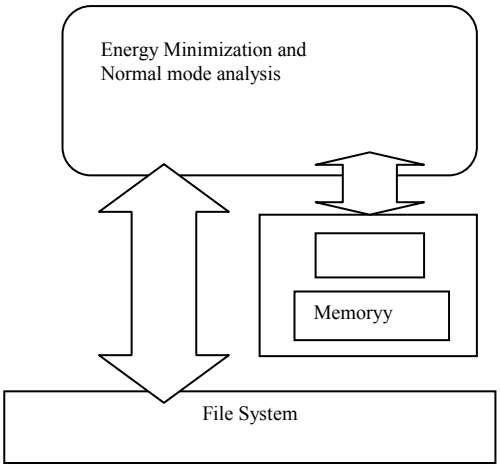


Figure 16. Single processing unit for energy minimization

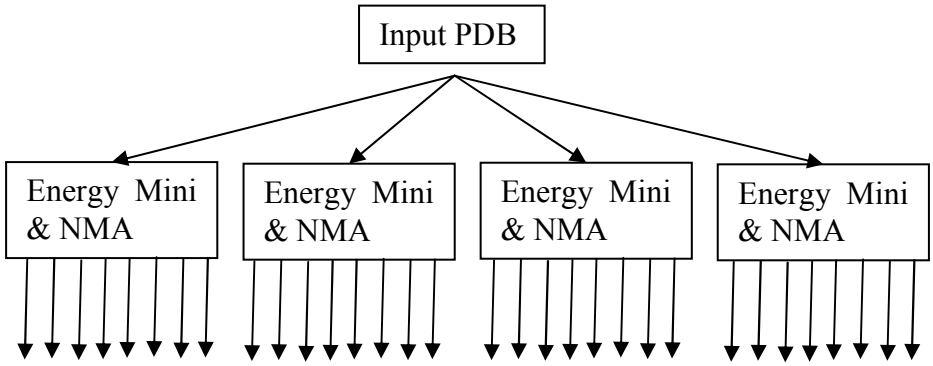


Figure 17. Simple parallelization

This avoids the data lost to resolve conflicts in data communication in an input and output bus. There are usually two ways to approach this problem. One would be to build a parallel processing system specifically designed to give the best performance for the existing algorithm. This is a more expensive process as the existing parallel systems do not usually satisfy these criteria, and a new parallel computing system must be built.

This works out better in the long run if there is a big class of problems that can be solved using this architecture and there is active research being conducted in the area. The other option is to design an algorithm that is going to best utilize the existing parallel processing systems available for access.

The parallel Linux cluster available for testing was a 20 node dual processor cluster with an underlying parallel file system. The first attempt was just to replicate the program and run it on multiple processors at the same time.

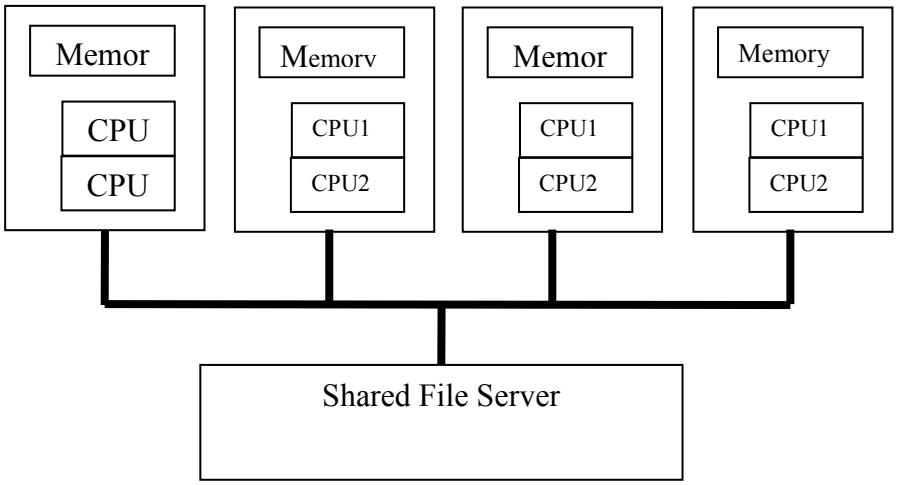


Figure 18. High performance computing architecture

In this model there is no communication between the different compute processors except at the beginning when the initial structure is distributed. If multiple iterations of energy minimization and normal mode analysis are performed the above design does not allow the program to select good structures and discard the bad ones. If all the processors worked on good structures at the beginning of each stage of energy minimization and normal mode analysis, instead of only some then better structures can be generated overall. Based on the existing architecture of the parallel linux cluster, there were two ways this communication of information could happen. One is to use the message passing interface across the network for the transfer of the actual files. The other is to use the underlying parallel file system and only transfer the relevant file information across the network. This reduces the overall amount of data sent over the network as well as the overhead involved in the transfer of the large amount of data.

The figure above shows the architecture of the high performance computing system used by the software environment. Each node of the cluster has a dual Intel processor with a shared memory. All the nodes are connected by a fast switch to enable high speed communications. The nodes are also connected to a shared file server as shown in the figure.

The file input output (I/O) has a significant overhead due to the large number of files begin generated from energy minimization and normal mode analysis. Since the files are shared among all the compute nodes using the underlying shared file server, only the file names and file path is necessary information to access the data.

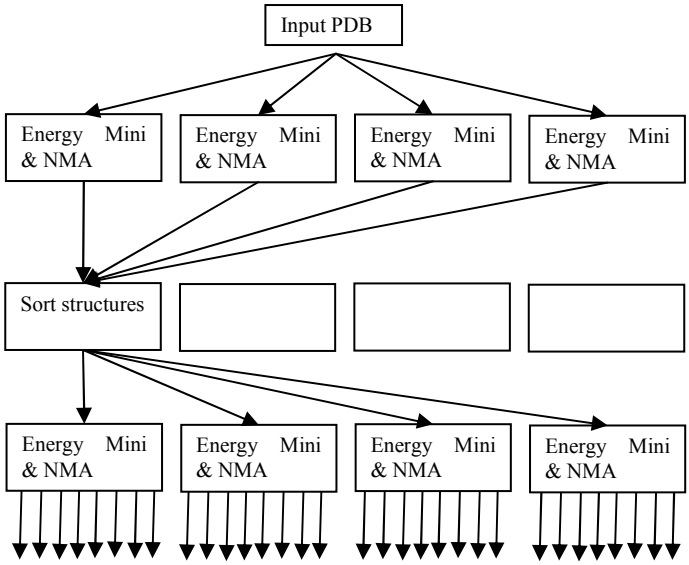


Figure 19. Interprocess communication

The figure above shows the design of energy minimization and normal mode analysis with inter-process communication. The figure shows the steps and data transfer that occurs for two iterations of energy minimization and normal mode analysis. This sequence of communication of protein structures information between parallel processors takes place at the end of each iteration of energy minimization and normal mode analysis. Since all the

generated protein structure files are written to a shared file server, there is no overhead of transferring the actual structure files. The final iteration of the energy minimization and normal mode analysis sequence results in a set of protein structures which is directly proportional to the number of processors used. These protein structures are sorted one more time based on the scoring method selected and required number of protein structure files are selected for distance geometry phase of the structure refinement which has been previously described.

Distance geometry in parallel

The distance geometry implementation for a single processor has been described in the earlier section. This section describes how this algorithm has been implemented for multiple processors. The input is a single protein sequence file with a set of distance restraints, and the out put is a set of structures satisfying the distance restraints. This makes the parallelization of the distance geometry simulated annealing stage quite straight forward. Each processor can work on one protein structure at a time and if there are more input structures, each has to be processed after the structure ensemble for the previous structure has been generated. The implementation is illustrated in the figure below.

Since the number of protein structure files used for distance geometry simulated annealing is a small percentage of the number of protein structure files generated during energy minimization and normal mode analysis, the protein structure selection and sorting

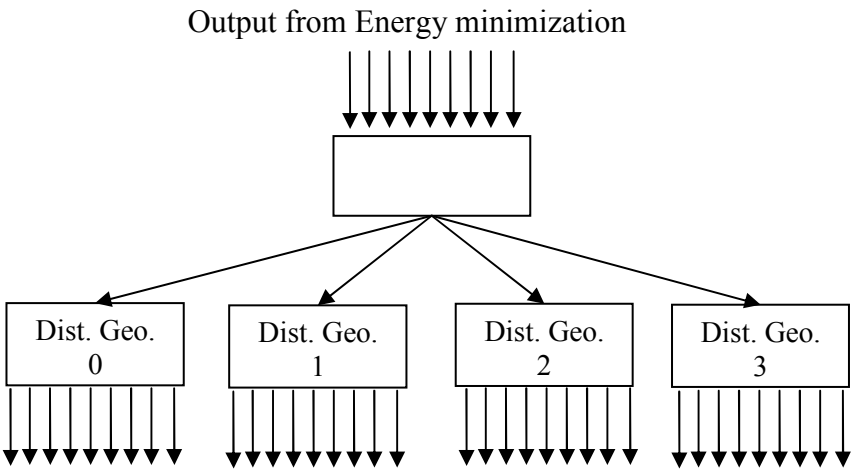


Figure 20. Distance geometry simulated annealing using 4 parallel processors

methods used should be good at picking the structures that have best chance of providing a refined structure. This process of selection of structures is an active area of research and it becomes increasingly harder to correctly identify structures as the structures get closer to and better at representing the actual experimental structures. This process is described in more detail in the section on protein scoring functions.

The structures generated at the end of distance geometry simulated annealing are analyzed and the final refined structures are selected based on the sorting methods specified.

Software System

The previous sections described how energy minimization, normal mode analysis and distance geometry based simulated annealing methods were implemented. This section describes how these methods work in conjunction with one another. Both energy

minimization and distance geometry based simulated annealing can independently refine protein structures to a certain extent.

Energy minimization is a coarser form of structure refinement with respect to the resolution of the structure, compared to distance geometry based simulated annealing. Energy minimization was chosen to be the first step followed by distance geometry based simulated annealing. During the initial stage of the project, CHARMM was used as the software tool to perform the energy minimization and normal mode analysis. One of the reasons for CHARMM as the software of choice was the popularity of the package among computational biologists, access to the program, as well as the expertise of the existing members of the research group. Distance geometry based simulated annealing was based on the work of Wu (39), and CNS was the software tool used there. The distance based mean force potentials were implemented for the potential energy functions of CNS. Hence the use of CNS would reduce the work involved in a fresh implementation of the mean force potentials. However, the primary input data formats and requirements for CHARMM and CNS were quite different. This required the use of Perl and UNIX shell scripts to perform the necessary data format correction and input generation. Furthermore, there was a need for external tools to evaluate structures for protein structure ranking by scoring functions, as well as comparison between structures. These tools have been described earlier in detail in the sections on scoring functions.

With this information in mind, an overall illustration of the software environment for protein structure refinement is shown in the figure below.

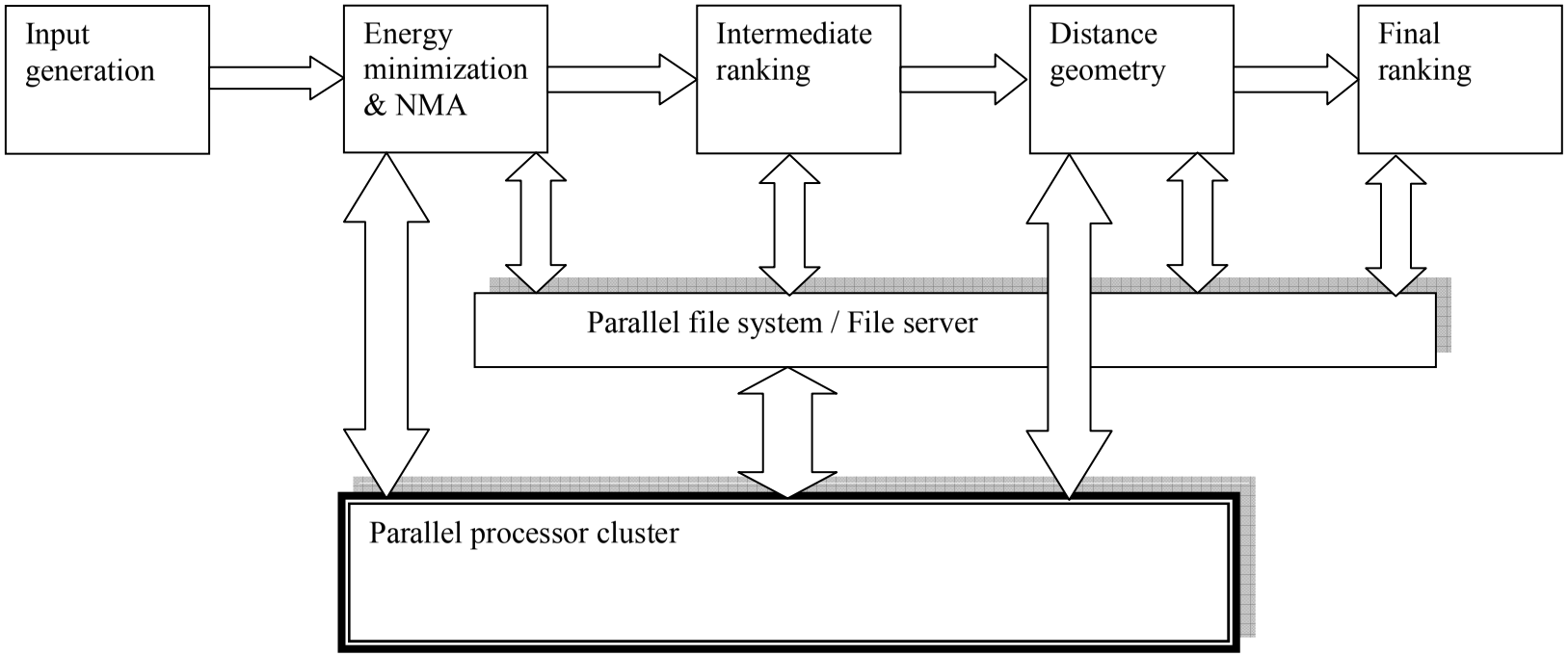
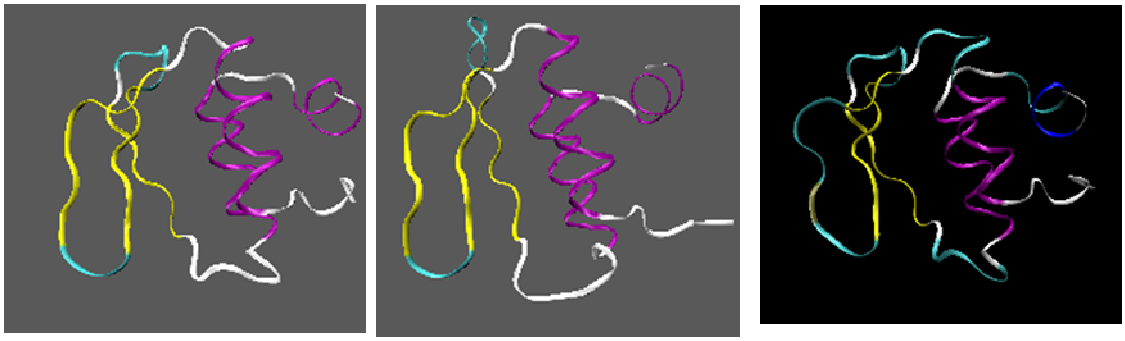


Figure 21. Software environment for protein structure refinement

CHAPTER 5. RESULTS

This chapter presents the results from energy minimization and distance geometry calculations carried out on an a few sample protein structures from the protein structure refinement experiment (CASPR). The parallel performance of the algorithm is also demonstrated.

For the protein structures shown below, Figure 22(a) is the initial modeled protein structure provided by the Baker group as one of the results from CASP 6 and Figure 22(b) is the experimental X-ray crystal structure (PDB id 1WHZ). The root mean square deviation between the two 70 residue structures is 3.1829Å for all atoms in the chain, and 2.1954Å for the Cα atoms in the backbone. The third image shows the protein structure after 1000 steps of potential energy minimization. Energy was used to rank all the structures generated during that run.



a) b) c)

Figure 22. a) Initial template b) X-ray target c) Minimization result

The ensemble of structures below shows the protein structures obtained after energy minimization.

Figure 23. Ensemble of energy minimized structures

It is easy to see that the alpha helices are quite well modeled, but the loop regions, chain ends as well as the beta sheets are the regions which have the most variability. These are the areas that have traditionally been quite difficult to model very well compared to the experimental crystal structures.

The Ramachandran plots in Figure 24 and Figure 25 show how the structures compare structurally. The X-ray crystal had 98.2% in the favorable region, and the starting template structure had 89.1% in the favorable region. These Ramachandran plots were obtained using Procheck (56). The structures obtained after energy minimization demonstrate quite a variation in their Ramachandran values, but they give a good indication on the overall chemical structure.

The Table 5 shows the results comparing different methods that were evaluated to score the structures obtained after energy minimization. The calculations were done on an 8 nodes, with each node having 2 SMP processors. Each processor generated 16 structures with a total of 256 structures generated. The RMSD column compares the generated structure with the X-ray structure 1WHZ. The structure with the best root mean square deviation does not

have the minimum energy or best values for Ramachandran scores, short range energy or even Four-body contact energy. The table indicates that no one value can be used to

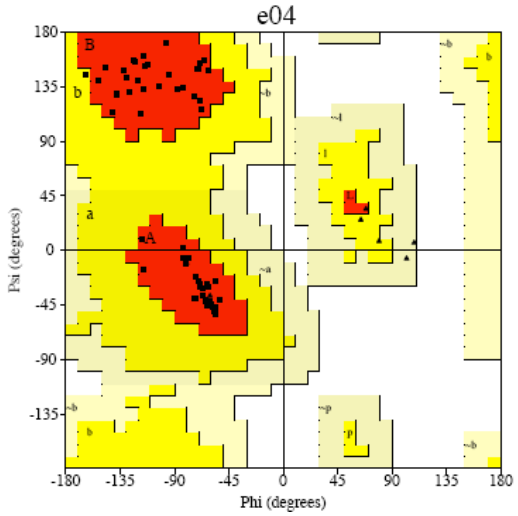


Figure 24. Ramachandran plots of X-ray crystal structure (1WHZ)

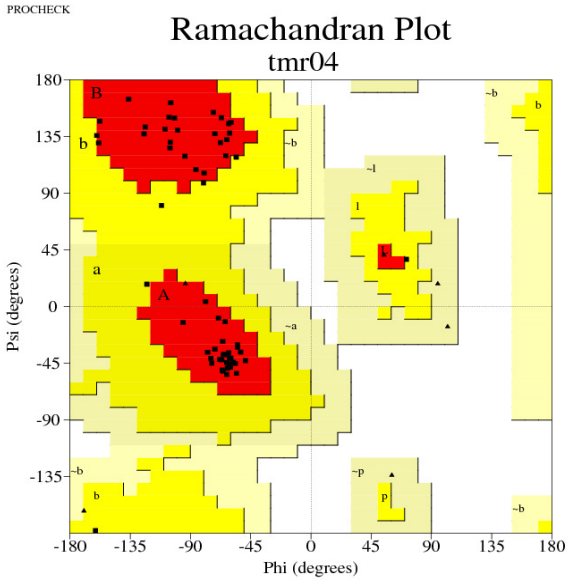


Figure 25. Ramachandran plot of template(TMR04) from Baker group

effectively rank the best structures. However they can be used to differentiate between the bad structures and structures which are reasonable close to the experimental structure.

Table 5. Comparisons after energy minimization

Type	Energy	RMSD (Ca) Å	Rama.	Short range	4 body
Min RMSD	-1226.88	1.81	89.1	-118.944	-11.604
Max Rama	-1181.21	2.01	98.2	-113.981	-14.861
Min Short	-1122.66	2.13	92.7	-160.57	-12.759
Min 4 body	-853.82	2.35	87.3	-113.63	-18.007
Min Energy	-1257.77	2.05	90.9	-107.33	-14.656

Table 6. Results after distance geometric calculations

Proc	Start RMS	Min RMS	Max RMS	Mean	SD
1	2.0467	1.9861	2.0188	2.0007	0.0066
2	2.0327	2.0007	2.0243	2.0143	0.0052
3	2.1518	2.1085	2.1293	2.1194	0.0038
4	2.0912	2.0457	2.071	2.0611	0.0045
5	2.0184	1.9864	2.0108	2.0002	0.0057
6	2.1217	2.0974	2.1199	2.1108	0.0050
7	2.0596	2.0185	2.0432	2.0353	0.0048
8	1.9673	1.9705	1.9818	1.9767	0.0026

Table 7. After distance geometric calculations

Type	Energy	RMSD	Rama.	Short range	4 body
Min RMSD	-4328.22	1.9705	87.3	-176.61	-16.514
Max Rama	-4112.26	1.9861	90.9	-160.67	-17.111
Min Short range	-4123.99	1.9818	87.3	-187.22	-16.6
Min 4body	-4045.99	2.0383	85.5	-157.54	-21.756
Min Energy	-4388.91	2.0034	89.1	-171.29	-19.913

Table 8. Comparison of proteins

Protein	Length	Initial RMS	Best RMS	Energy RMS
1XE1	91	2.9244	1.461	1.7894
1VM0	103	5.9511	5.5383	5.7782
1O13	107	4.1584	3.3129	3.536
TMR04	70	2.19	1.8098	2.0467

It can be seen from Figure 26, Figure 27,Figure 28, Figure 29 and Figure 30, the secondary structures in a protein can impact the refinement of protein structures. The alpha helices can be modeled very accurately compared to beta sheets and loops. So the potential for improvement is greatest when there are significantly less alpha helices. But this also makes it much more harder as there are no clear cut algorithms for predicting the structures of loops and beta sheets.

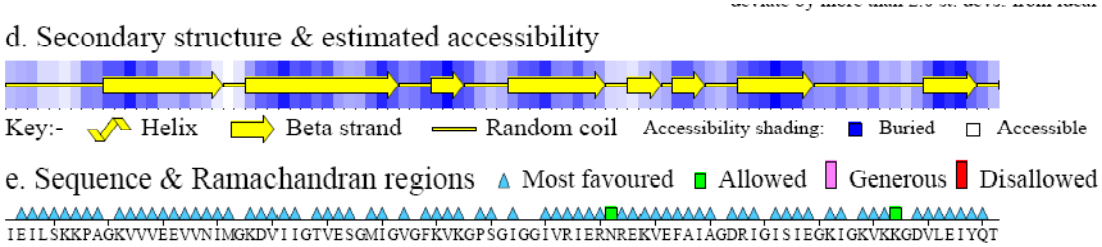


Figure 26. Secondary structures for 1XE1

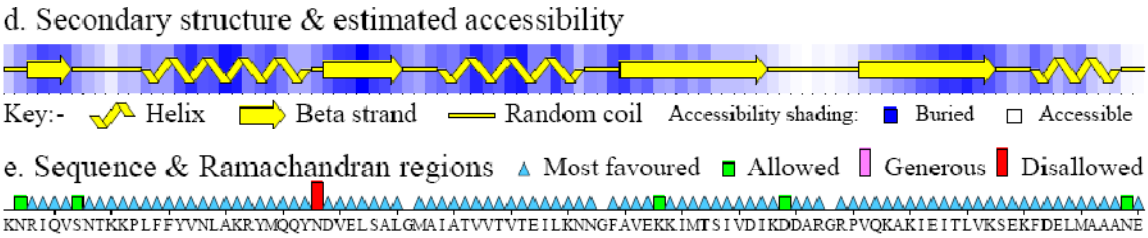


Figure 27. Secondary structures for 1VM0

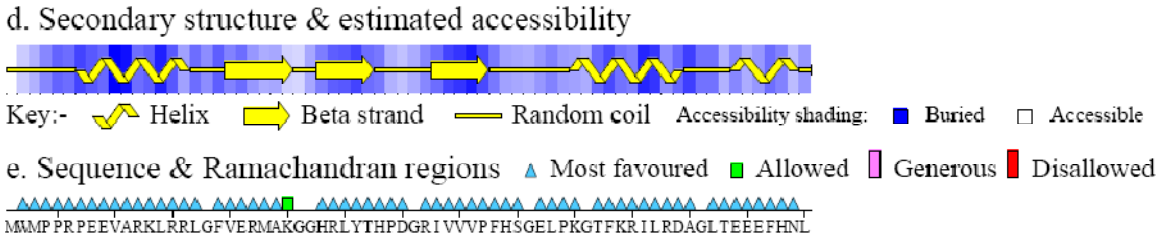


Figure 28. Secondary structures for TMR04

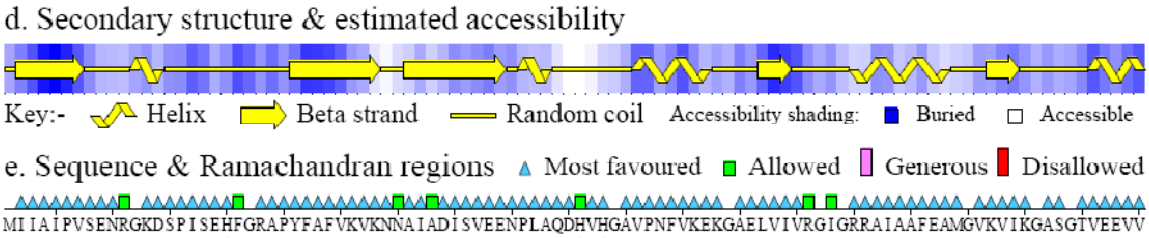


Figure 29. Secondary structures for 1O13

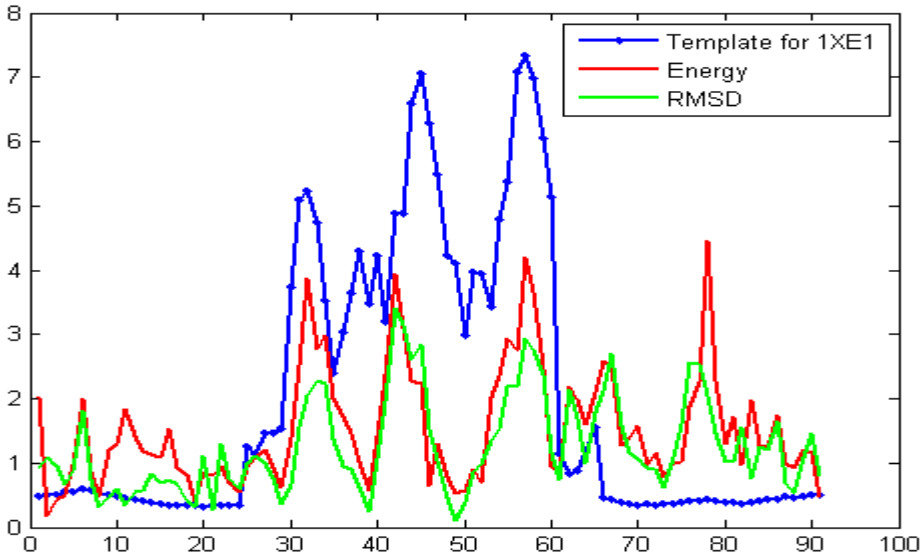


Figure 30. RMS deviation for each residue (CA) for 1XE1

Energy minimization results

Figure 31 shows the energy RMSD plot for CHARMM energy minimization for 2 iterations, 16 generation for each file. 16 processors.

The chart in Figure 32 shows the energy vs. all atom rms for 15000 minimization steps. The results were obtained by executing 16 processors using GROMACS. The worst rms was 3.5653, and the best rms was 3.0612, and the median rms was 3.3174.

Performance of Gromacs by increasing the number of steps used in potential energy minimization using steepest descent is shown in Figure 33.

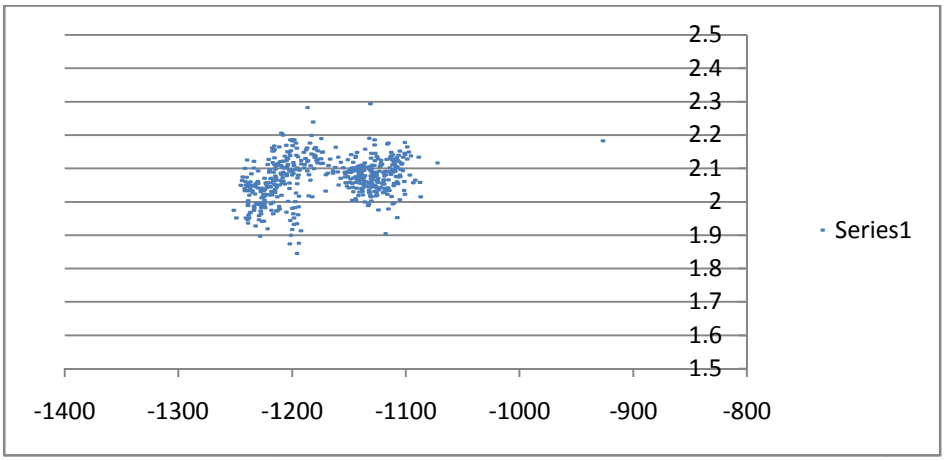


Figure 31. RMSD(Y-axis) vs. Energy (X-axis)

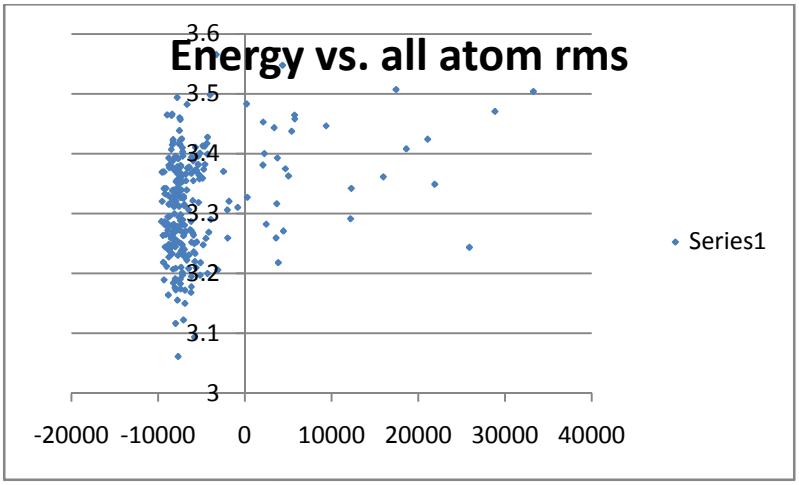


Figure 32. Energy vs. RMSD (Y-axis)

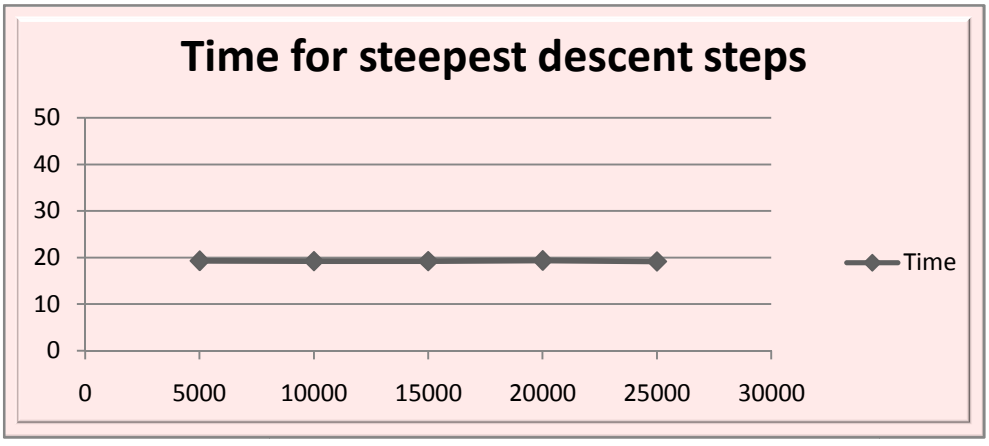


Figure 33. Steepest descent

Performance of Gromacs energy minimization for varying number of processors

The following charts show the performance of energy minimization part based on the number of processors used. There are two plots here, one shows the total time taken for a specified number of processors and the other shows the number of protein model pdb files generated at the end of the energy minimization stage. As the two plots show the time taken increases proportionately to the number of model files generated, both of which increase linearly with the number of processors used.

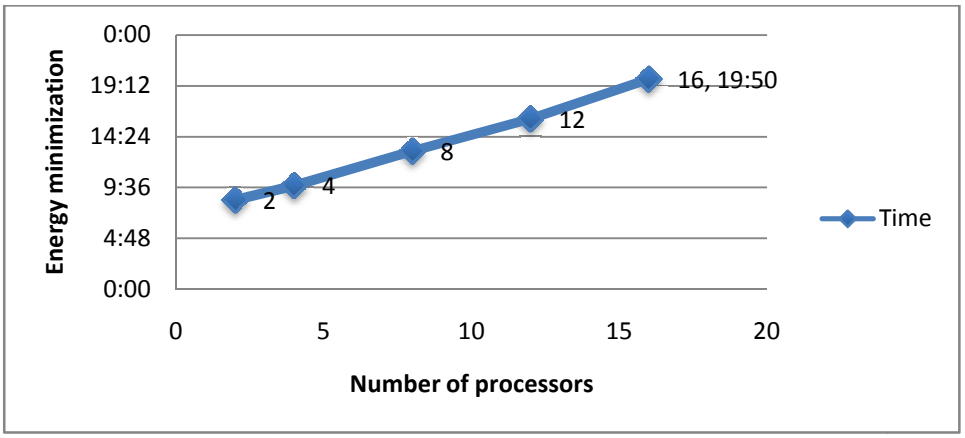


Figure 34. Time vs Num. of processors

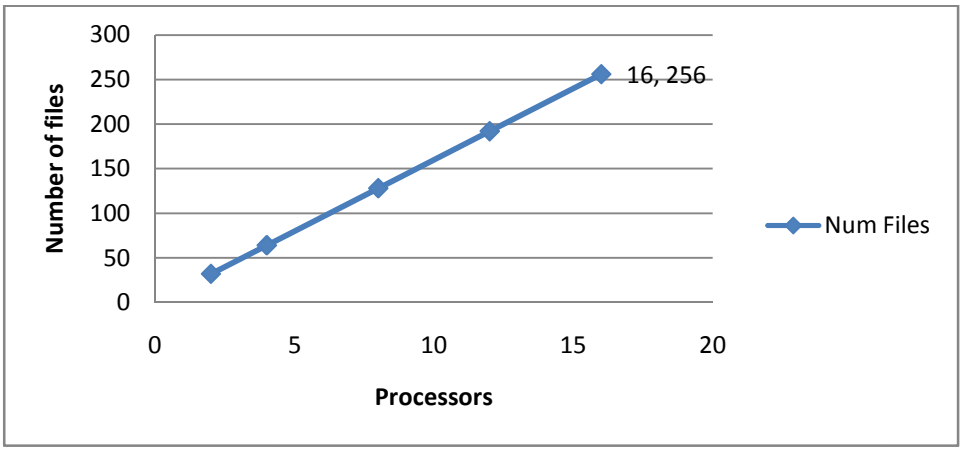


Figure 35. Files generated vs. Num. of processors

Performance based on the number of iterations of energy minimization

The following two plots show the time taken by 16 processors, with each processor generating 8 protein structures after each iteration. The structures from earlier iterations are not discarded but compared with the new ones that are generated. It can be seen that the time taken is proportional to the total number of files generated at the end of each iteration.

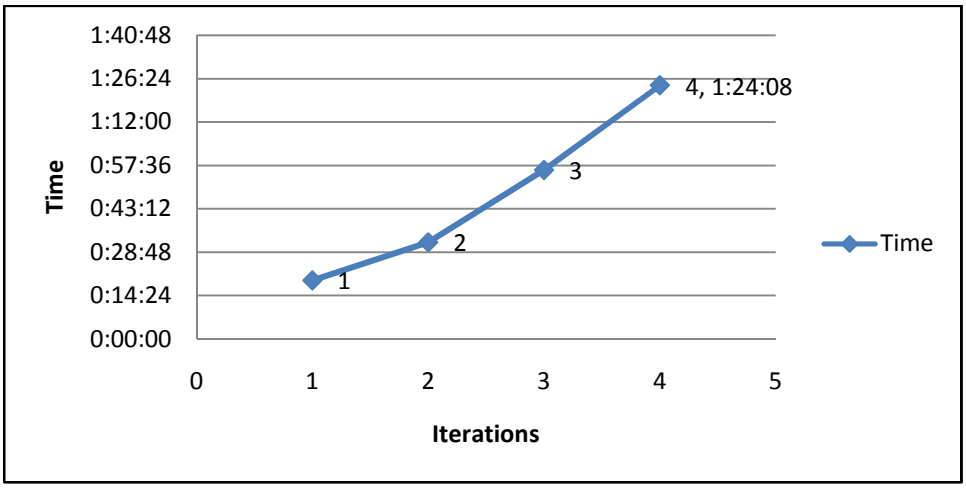


Figure 36. Time taken for energy minimization iterations

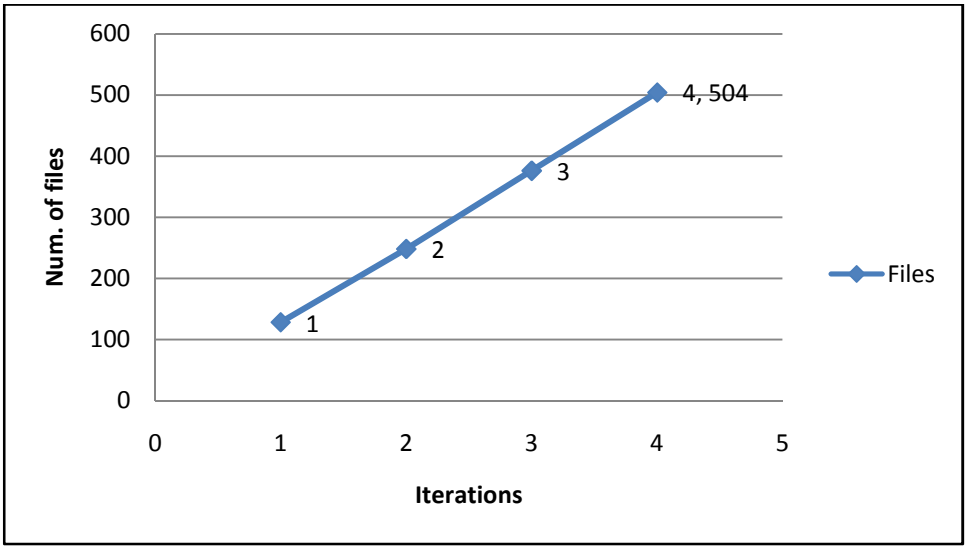


Figure 37. Number of files generated with increasing iterations

Performance of energy based scoring methods

The plot below are results of using potential energy as means of comparing the final protein structures obtained. The plot shows the majority of the structures are around the 3Å to 2Å.

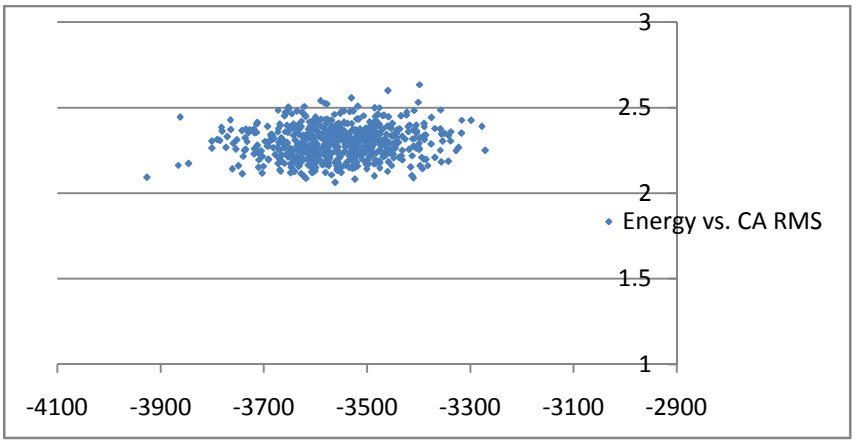


Figure 38. Evergy vs. RMS

The Figure 39 shows the structure with the lowest energy from the above plot, and aligned with the X-ray structure. The Cα root mean square is about 2.0925Å. The X-ray structure is shown as blue and the modeled structure is shown as red. It can be clearly seen that the refined structure very closely resembles the X-ray structure except the beta-sheet region on the left side of the picture.

The difference of the final energy minimized structure obtained after distance geometry calculations with the X-ray structure and in comparison to the starting template structure is shown in the plot below generated using Matlab, and using the Cα coordinates do calculate the root mean square deviation. It is can be clearly noted that the segment between

residues 30 and 40 is the most significant deviation from the X-ray structure, which is in fact the location of the beta-sheet and the associated loops.

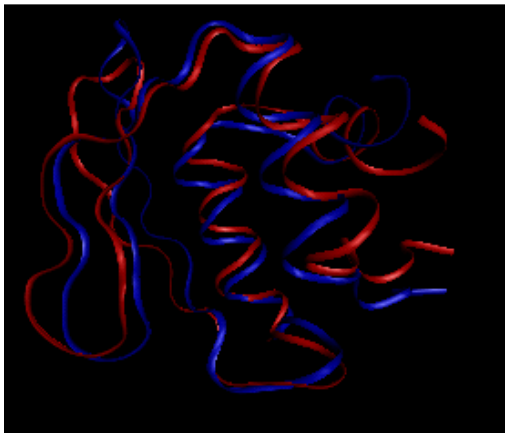


Figure 39. Alignment of minimum energy structure with X-ray structure

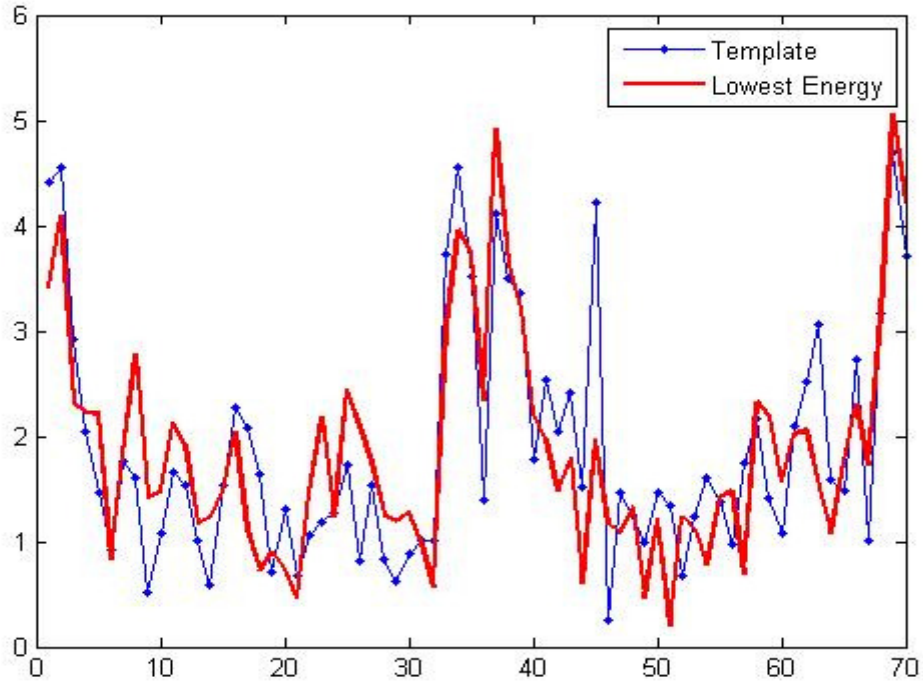


Figure 40. RMS deviation for each residue with X-ray structure

Performance of Ramachandran plots

This section shows the comparison of ramachandran values to the structures of the protein refinement. The plot shown below is very similar to the plots using energy as the X-axis shown in the previous page. Likewise there is no clear pattern between ramachandran values and the root mean square of the refined structures compared to the X-ray structure. However it is easily noted that the structures with the worst root mean square values are the ones with ramachandran values of 50% or less.

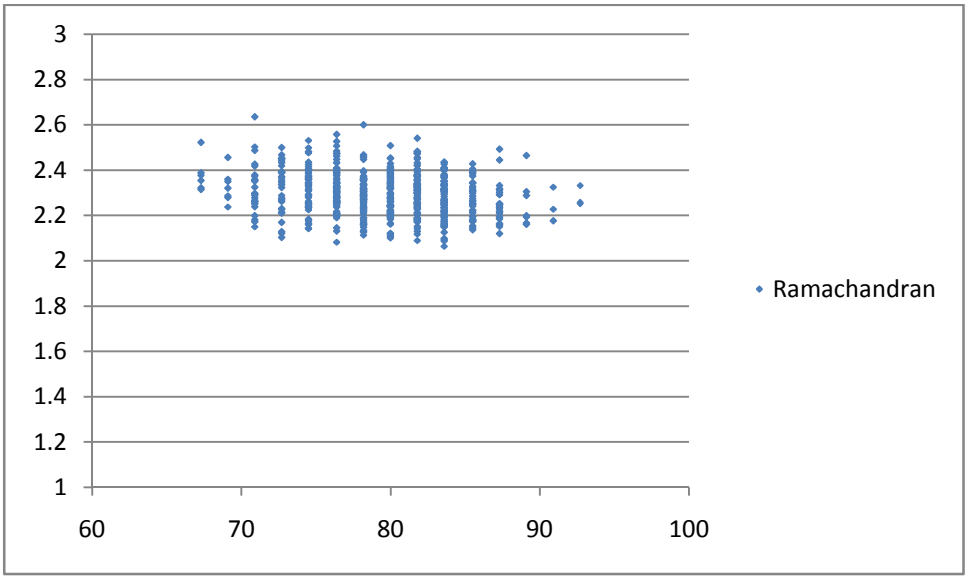


Figure 41. Ramachandran values vs. RMSD

The figure below shows the structure comparison of the X-ray structure of the protein with the refined structure having the night ramachandran value. The Cα root mean square deviation of this structure is 2.256Å. It can be seen that the beta-sheet region on the left side is the segment that is not closely modeled.

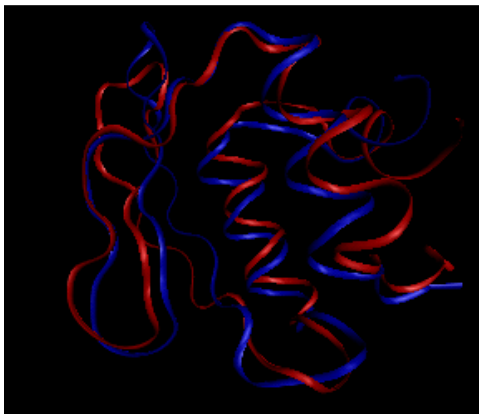


Figure 42. Alignment of best ramachandran structure with X-ray structure

Performance of short range scoring function

The plots below show the performance of the short range potentials used as a scoring function. It is easy to note that the short range potentials clearly differentiate between the good and the bad structures.

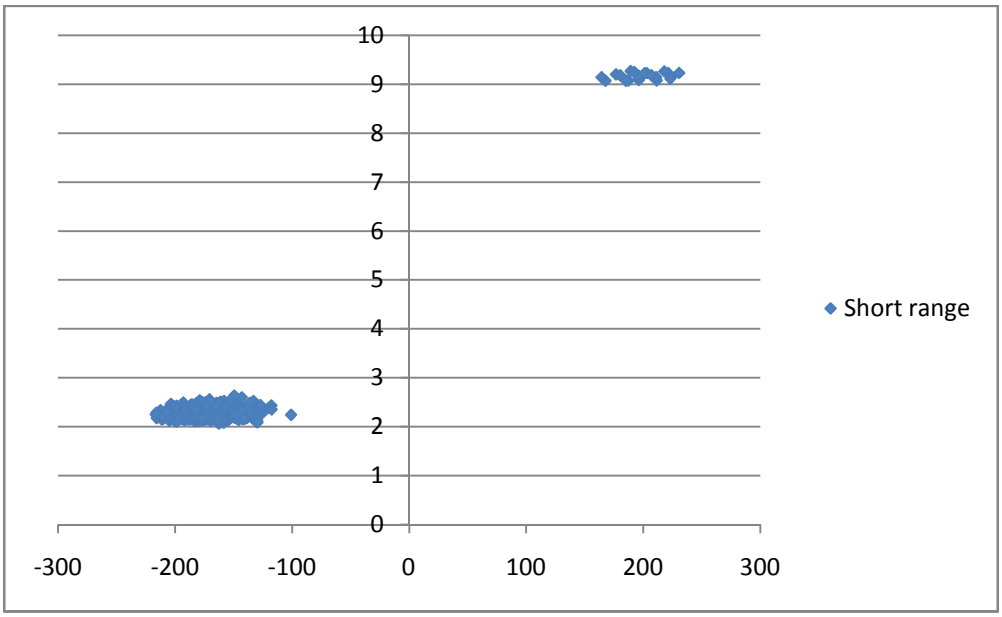


Figure 43. Short range scoring function vs. RMSD

The figure below shows the three-dimensional comparison of the structure with the best short range potential score with the X-ray structure. The Cα root mean square deviation is 2.2465Å.

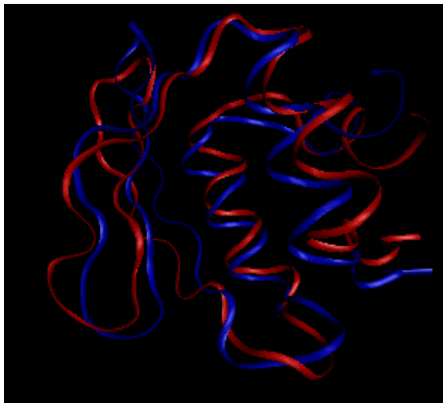


Figure 44. Alighment of short range structure with X-ray structure

Performance of 4 body function

The plots below show the results using another scoring function called the 4 body contact potential. From the first plot is clearly seen that the potential cannot distinguish from good structures and bad structures. It also does not perform much better at a close up of the plot area of the good structures.

This three-dimensional structure comparison with the X-ray structure clearly shows the difference from the experimental values. The best scoring structure is infact one of the bad structures with an Cα root mean square deviation of 9Å.

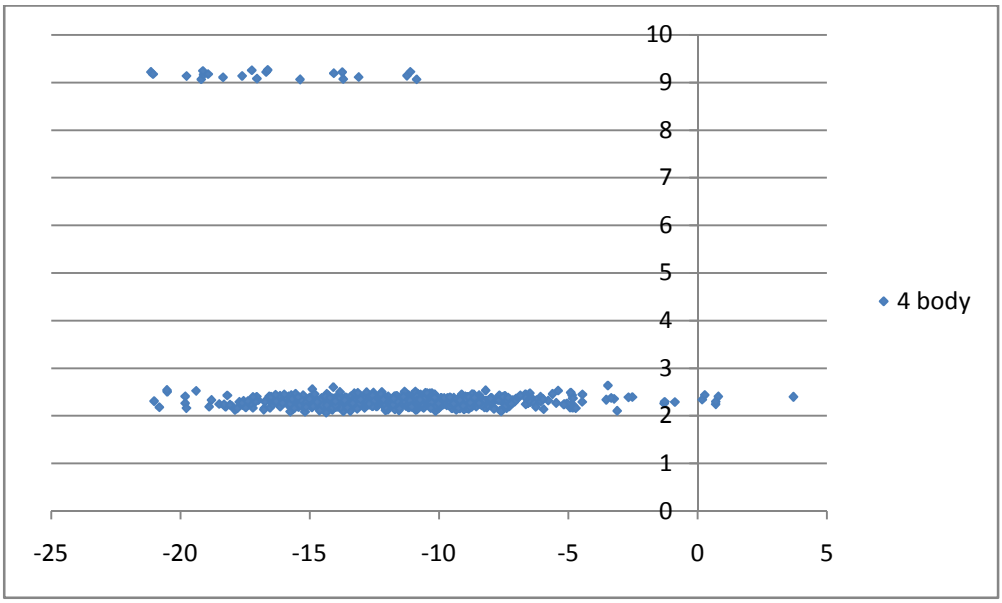


Figure 45. Four-body function vs. RMSD

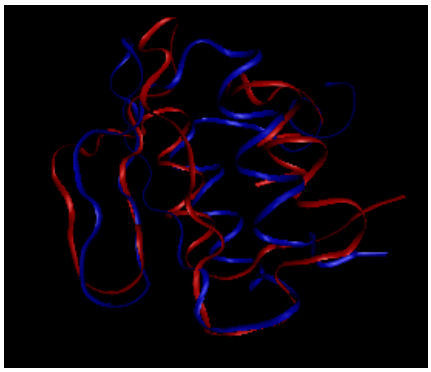


Figure 46. Alignment of four body structure with X-ray structure

The plot in Figure 47 compares all the previously described functions with the highest scoring structures of each function. The plot in blue is the starting template structure. It is quite easy to notice that all the methods had difficulty in improving the beta-sheet region between residues 30 and 40. A combination of these functions is also being investigated to see if better results can be obtained.

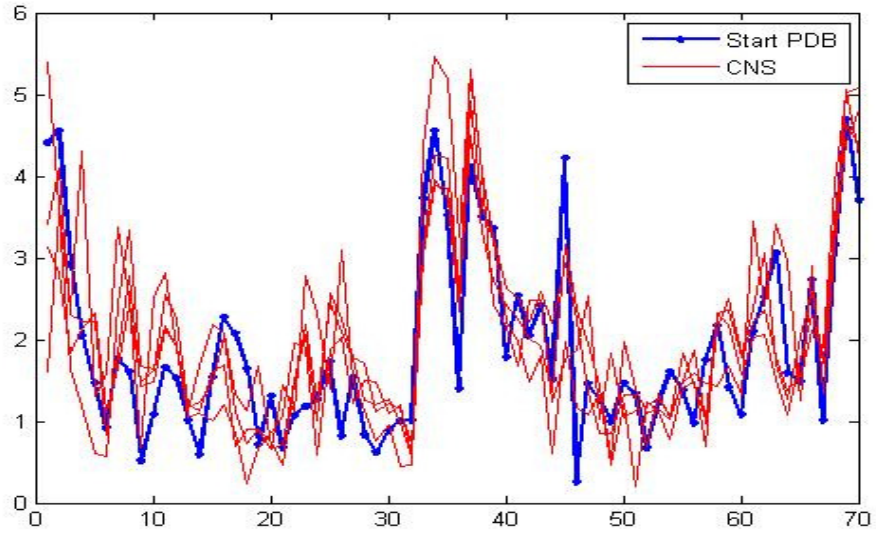


Figure 47. RMS deviation for each residue with X-ray structure

Show below is a different representation of the three-dimensional structures of the X-ray structure with the structure having the lowest root mean square deviation. It is clearly evident that the loop region of the structure is the area that is difficult to refine. This is an area of considerable active research and the hardest part of structure prediction and refinement.

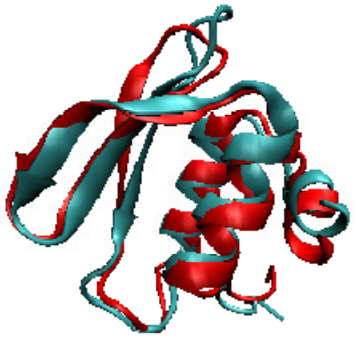


Figure 48. Alignment of best RMSD structure with X-ray

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

Conclusions

In the previous chapter the results from the different approaches to protein structure refinement were presented. The goal of this project was to arrive at an algorithm or criteria for selecting good structures relative to the experimentally derived results like X-ray crystallography and NMR spectroscopy.

It can be inferred that using the potential energy of a protein structure can be a good idea to differentiate a structure from a bad stereo chemical structure. Normal mode fluctuations are used to indentify the flexible regions of a protein structure and could certainly aid in structure refinement if a good three dimensional search algorithm is used. Use of distance geometric calculations with the aide of normal mode fluctuations also can provide valuable information in distinguishing the good structures from the bad.

However, all of the previously described methods cannot with a high certainty, distinguish the structures which are near the native protein structures with the structures that are good stereo chemically but not close to the native structures. This leaves room for considerable scope of future work.

Future work

There are various alternative methods that can be investigated. Energy functions other than the one provided by CHARMM can be used. Knowledge based energy functions have also been known to provide very good results. The scoring functions used here only

considered short range interactions. Other scoring functions as well as long range and intermediate range interactions can also be considered. The distance constraints can also be modified to use other parameters than just normal mode fluctuations. The geometric embedding parameters can also be investigated to see how they compare with just distance restraints based on normal mode fluctuations.

BIBLIOGRAPHY

1. *RDOCK: Refinement of rigid-body protein docking predictions.* **L. Li, R. Chen, Z. Weng.** 2003, PROTEINS: Structure, Functions, and Genetics, Vol. 53, pp. 693-707.

2. **Reed, Randy J.** Overview of macromolecular crystallography. *Protein crystallography course.* [Online] 2008. [Cited: June 7, 2008.] http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html.

3. *Protein NMR spectroscopy in structural genomics.* **Gaetano T. Montellone, Deyou Zheng, Yuanpeng J. Huang, Kristin C. Gunsalus, Thomas Szyperski.** 2000, Nature Structural Biology, Vol. 7, pp. 982-985.

4. *Comparison of protein structures determined by NMR in solution and by x-ray diffraction in single crystals.* **Billeter, Martin.** 3, 1992, Quarterly Reviews of Biophysics, Vol. 25, pp. 325-377.

5. *Improving the quality of protein structures derived by NMR spectroscopy.* **C.A.E.M Spronk, J.P. Linge, C.W. Hilbers, G.W. Vuister.** 2002, Journal of Biomolecular NMR, Vol. 22, pp. 281-289.

6. *Protein structure refinement based on paramagnetic NMR shifts: Applications to wild-type and mutant forms of cytochrome c.* **M. Gochin, H. Roder.** 1995, Protein Sci., Vol. 4, pp. 296-305.

7. *Ab Initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement.* **Skolnick, J., Kolinski, A., Kihara, D.,**

Betancourt, M., Rotkiewicz, P., Boniecki, M. 5, 2001, PROTEINS: Structure, Function, and Genetics Suppl., pp. 149-156.

8. *Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction.* **Z. Xiang, C. Soto, B. Honig.** USA : s.n., 2002. Proceedings of Natl. Acad. Sci. Vol. 99, pp. 7432-7437.

9. *Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation.* **B. Qian, A.R. Ortiz, D. Baker.** 43, 2004, Proceedings of the National Academy of Science, Vol. 101, pp. 15346-15351.

10. *Refinement of protein structures in explicit solvent.* **J.P. Linge, M.A. Williams, C.A.E.M. Spronk, A.M.J.J. Bonvin, M. Nilges.** 2003, PROTEINS: Structure, Function, and Genetics, Vol. 50, pp. 496-506.

11. *Gaussian dynamics of folded proteins.* **T. Haliloglu, I. Bahar, B. Erman.** 1997, Physical Review Letters, Vol. 79, pp. 3090-3093.

12. *Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water.* **B. Xia, V. Tsui, D.A. Case, H.J. Dyson, P.E. Wright.** 2002, Journal of Biomolecular NMR, Vol. 22, pp. 317-331.

13. *Fnction minimization by conjugate gradients.* **R. Fletcher, C.M. Reeves.** 1964, The Computer Journal, Vol. 7, pp. 149-154.

14. *Generalized Polak-Ribiere algorithm.* **K.M. Koda, Y. Liu, C. Storey.** 2, s.l. : Springer Netherlands, November 1992, Journal of Optimization Theoy and Applications, Vol. 75, pp. 345-354.

15. *Global convergence of the fletcher-reeves algorithm with inexact linesearch.* **Liu Guanghui, Han Jiye, Yin Hongxia.** 1, March 1995, Applied Mathematics - A Journal of Chinese Universities, Vol. 10, pp. 75-82.

16. *ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement.* **Sanejouand, Karsten Suhre and Yves-Henri.** July 2004, Nucleic Acids Research, Vol. 32, pp. W610-W614.

17. *A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca2+-ATPase.* **G. Li, Q. Cui.** November 2002, Biophysical Journal, Vol. 83, pp. 2457-2474.

18. *Normal modes for predicting protein motions: A comprehensive database assessment and associated Web tool.* **Vadim Alexandrov, Ursula Lehnert, Nathaniel Echols, Duncan Milburn, Donald Engelman, Mark Gerstein.** 2005, Protein Science, Vol. 14, pp. 633-643.

19. **Steinbach, Peter J.** Normal mode (harmonic) analysis. *Introduction to Macromolecular Simulation.* [Online] [Cited: May 1, 2008.] http://cmm.cit.nih.gov/intro_simulation/intro_simulation.

20. *Direct evaluation of thermal fluctuations in roteins using a single-parameter harmonic potential.* **I. Bahar, A.R. Atilgan, B. Erman.** 2, May 1997, Folding and Design, pp. 173-181.

21. **Tirion, M. M.** Large amplitude elastic motions in roteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 1996, Vol. 77, pp. 1905-1908.

22. *Scoring functions in protein folding and design.* **Dima, Ruxandra I., Banavar, Jayanth R. and Maritan, Amos.** s.l. : Cambridge University Press, 2000, Protein Science, Vol. 9, pp. 812-819.

23. *Processing and evaluation of predictions in CASP4.* **A. Zemla, C. Venclovas, J. Moult, K. Fidelis.** 5, 2001, PROTEINS: Structure, Function, and Genetics Suppl., pp. 13-21.

24. *Scoring function for automated assessment of protein structure template quality.* **Y. Zhang, J. Skolnick.** 2004, PROTEINS: Structure, Function, and Bioinformatics, Vol. 57, pp. 702-710.

25. *Empirical potentials and functions for protein folding and binding.* **S. Vajda, M. Sippl, J. Novotny.** 2, April 1997, Current Opinion in Structural Biology, Vol. 7, pp. 222-228.

26. *MaxSub: an automated measure for the assement of protein structure prediction quality.* **Naomi Siew, Arne Elofsson, Leszek Rychlewski, Daniel Fischer.** 9, 2000, Bioinformatics, Vol. 16, pp. 776-785.

27. *Evaluating CASP4 predictions with physical energy functions.* **M. Feig, C.L. Brooks III.** 49, 2002, PROTEINS: Structure, Function, and Genetics, pp. 232-245.

28. *Improved protein structure selection using decoy dependent discriminatory functions.* **K. Wang, B. Fain, M. Levitt.** 8, 2004, BMC Structural Biology, Vol. 4.

29. *Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.* **KT, Simons, et al.** 1999, Proteins, Vol. 34, pp. 82-95.

30. *Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues.* **Miyazawa, S and Jernigan, RL.** 1999, Proteins, Vol. 34, pp. 49-68.

31. *An iterative method for extracting energy-like quantities from protein structures.* **D.P., Thomas and K.A., Dill.** 257, 1996, Journal of Molecular Biology, pp. 457-469.

32. *Determination of interaction potentials of amino acids from native protein structures: Test on simple lattice models.* **Mourick, J. von, et al.** 110, 1999, J Chem Phys, pp. 10123-10133.

33. *Short-range conformational energies, secondary structure propensities, and .* **Bahar, I., Kaplan, M. and Jernigan, R.L.** 1997, PROTEINS: Structure, Function, and Genetics, Vol. 29, pp. 292-308.

34. *Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys.* **Feng, Yaping, Kloczkowski, Andrzej and Jernigan, Robert L.** 1, 2007, Proteins: Structure, Function, and Bioinformatics, Vol. 68, pp. 57-66.

35. *Progress and challenges in high-resolution refinement of protein structure models.* **K.M.S. Misura, D. Baker.** 59, 2005, PROTEINS: Structure, Function, and Bioinformatics, pp. 15-29.

36. *Screening and refinement of protein structures from fold recognition.* **R. Zhou, B.D. Silverman, G. Dent, A. Royyuru, A. Curioni, A. Logen.** s.l. : MIT, 2005. RECOMB2005.

37. *Developing a move-set for protein model refinement.* **M.N. Offman, P.W. Fitzjohn, P.A. Bates.** 15, 2006, Bioinformatics, Vol. 22, pp. 1838-1845.

38. *Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations.* **Andrew W. Stumpff-Kane, Katarzyna Maksimiak, Michael S. Lee, Michael Feig.** 4, 2007, Proteins, Vol. 70, pp. 1345-1356.

39. *Refinement of NMR-determined protein structures with database derived mean-force potentials.* **D. Wu, R. Jernigan, Z. Wu.** s.l. : PROTEINS: Structure, Function, and Bioinformatics, 2006.

40. *Refinement of NMR-determined protein structures with database derived distance constraints.* **F. Cui, R. Jernigan, Z. Wu.** 6, 2005, Journal of Bioinformatics and Computational Biology, Vol. 3, pp. 1315-1329.

41. *Crystallography & NMR system: A new software suite for macromolecular structure determination.* **A.T. Brunger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren.** 1998, Acta Crystallographica, Vol. D54, pp. 905-921.

42. *Calculation of conformational ensembles from potentials of mean force.* **Sippl, M.J.** 4, 1990, Journal Molecular Biology, Vol. 213, pp. 859-883.

43.

44. *Progress and challenges in high-resolution refinement of protein structure models.* **Misura, K.M.S and Baker, D.** 59, 2005, PROTEINS: Structure, Function, and Bioinformatics, pp. 15-29.

45. *Developing a move-set for protein model refinement.* **Offman, M.N, Fitzjohn, P.W and Bates, P.A.** 15, 2006, Bioinformatics, Vol. 22, pp. 1838-1845.

46. *CHARMM: a program for macromolecular energy, minimization and dynamics calculation.* **B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus.** 1983, J Comp Chem, Vol. 4, pp. 187-217.

47. CHARMM. *Chemistry at HARvard Macromolecular Mechanics.* [Online] 2006. www.charmm.org.

48. *Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme.* **M. Levitt, C. Sander, P.S. Stern.** 1985, Journal of Molecular Biology, Vol. 181, pp. 423-447.

49. *Computation of low-frequency normal modes in macromolecules: Improvements to the method of diagonalization in a mixed basis and application to hemoglobin.* **D. Perahia, L. Mouawad.** 3, 1995, Computers Chem., Vol. 19, pp. 241-246.

50. *Crystallography and NMR System (CNS).* [Online] cns.csb.yale.edu.

51. *Multiscale Modeling Tools for Structural Biology (MMTSB).* [Online] mmtsb.scrips.edu.

52. *Visual Molecular Dynamics (VMD).* [Online] University of Illinois, Urbana-Champaign. www.ks.uiuc.edu/Research/vmd.

53. RCSB Protein Data Bank. *An Information Portal to Biological Macromolecular Structures.* [Online] www.rcsb.org.

54. *Concepts and tools for NMR restraint analysis and validation.* **S.B. Nabuurs, C.A.E.M. Spronk, G. Vriend, G.W. Vuister.** 2, 2004, Concepts in Magnetic Resonance, Vol. 22A, pp. 90-105.

55. **Barney, Blaise.** *Introduction to parallel computing.* [Online] Livermore computing. https://computing.llnl.gov/tutorials/parallel_comp/.

56. *PROCHECK: a program to check the stereochemical quality of protein structures.* **R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton.** 1993, J. Appl. Cryst., Vol. 26, pp. 283-291.