12-17-2010

A Web Service for Protein Refinement and Refinement of Membrane Proteins

Kapil Pothakanoori
University of New Orleans

A Web Service for Protein Refinement and Refinement of Membrane Proteins

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
In partial fulfillment of the
Requirements for the degree of

Master of Science
In
Computer Science

By

Kapil Pothakanoori

B.E (I.T) Osmania University, 2008

December, 2010

Acknowledgments

I acknowledge my advisor Dr. Christopher Summa. His problem solving approach accompanied with immense knowledge, expertise and commitment has been a great source of inspiration for me. Thank you for all the patience throughout the entire process. Under his guidance, I have not only completed my thesis work but have also learned so much that will help me in my future academic and professional life.

My study at UNO would not have been possible without the support of my family, Thanks Mom, Dad and my brothers. I take this opportunity to thank all the people who supported me throughout the 2 years of my study here.

I would also like to thank Mr. Austin Ada Orgah who has been a good friend. I thank him for his help and support at all times.

To all my friends who reside in 6239 and 6241 Wain Wright drive New Orleans during my time here, thank you all!!

Last but not the least; I would like to render my sincere gratitude to all those who have directly or indirectly helped in making this happen.

Table of Contents

List of Figures

List of Tables

Abstract

The structures obtained from homology modeling methods are of intermediate resolution 1-3Å from true structure. Energy minimization methods allow us to refine the proteins and obtain native like structures. Previous work shows that some of these methods performed well on soluble proteins. So we extended this work on membrane proteins. Prediction of membrane protein structures is a particularly important, since they are important biological drug targets, and since their number is vanishingly small, as a result of the inherent difficulties in working with these molecules experimentally. Hence there is a pressing need for alternative computational protein structure prediction methods. This work tests the ability of common molecular mechanics potential functions (AMBER99/03) and a hybrid knowledge-based potential function (KB_0.1) to refine near-native structures of membrane proteins *in vacuo*.

A web based utility for protein refinement has been developed and deployed based on the KB_0.1 potential to refine proteins.

Keywords: Protein refinement Web server, membrane proteins, protein refinement, amber03/99, Knowledge based mechanics model.

CHAPTER 1: INTRODUCTION

Proteins are the micro machines on which biological systems are based. Understanding protein's structure and functionality helps us design new drugs and understand the underlying mechanism of human disease. Protein structure prediction and protein folding are considered fundamental problems of modern computational/ molecular biology,

There has been research progress from the past several decades in understanding the underlying biophysical interactions in proteins either by simulating the proteins or by studying their behavior using experimental techniques.

1.1) Introduction to Proteins

Proteins are composed of individual units called amino acids. These amino acids are linked by peptide bonds. There are twenty different naturally occurring amino acids which differ in the chemical nature of their side chains. The nature of the side chain influences the properties and structure of the proteins.

Each amino acid consists of

- A central carbon atom usually referred as Alpha Carbon ($C\alpha$) atom.
- an amino group
- a carboxyl group
- a side chain

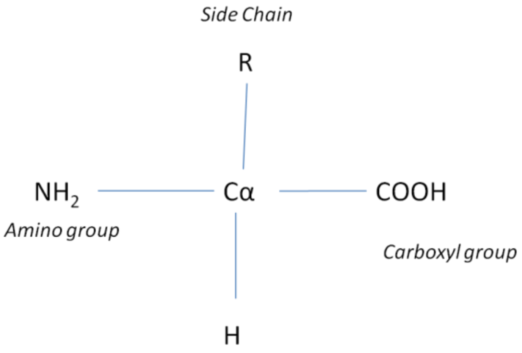


Figure 1: The general structure of an alpha amino acid

Proteins are intimately involved in nearly every cell activity, from replication of genetic code to transporting oxygen, and are generally responsible for regulating the cellular machinery and determining the phenotype of an organism. Diseases like transmissible spongiform encephalopathies [24], are usually caused due to improper folding of protein which may result from genetic and/or environmental influences. Examples of proteins include hemoglobin, thyroid hormone, insulin, and myosin.

There are 20 possibilities of amino acids based on the R group side chain associated with the amino acids

These amino acids are classified as acidic (negatively charged) , basic (Positively charged), polar(side chains with pure hydro carbon alkyl groups or aromatic), non-polar(Side chains with functional groups acids, amides, alcohols, and amines), Hydrophobic (un likely to be in contact with the aqueous environment), hydrophilic (likely to be in contact with the aqueous environment)based on chemical and structural behavior of their side chains.

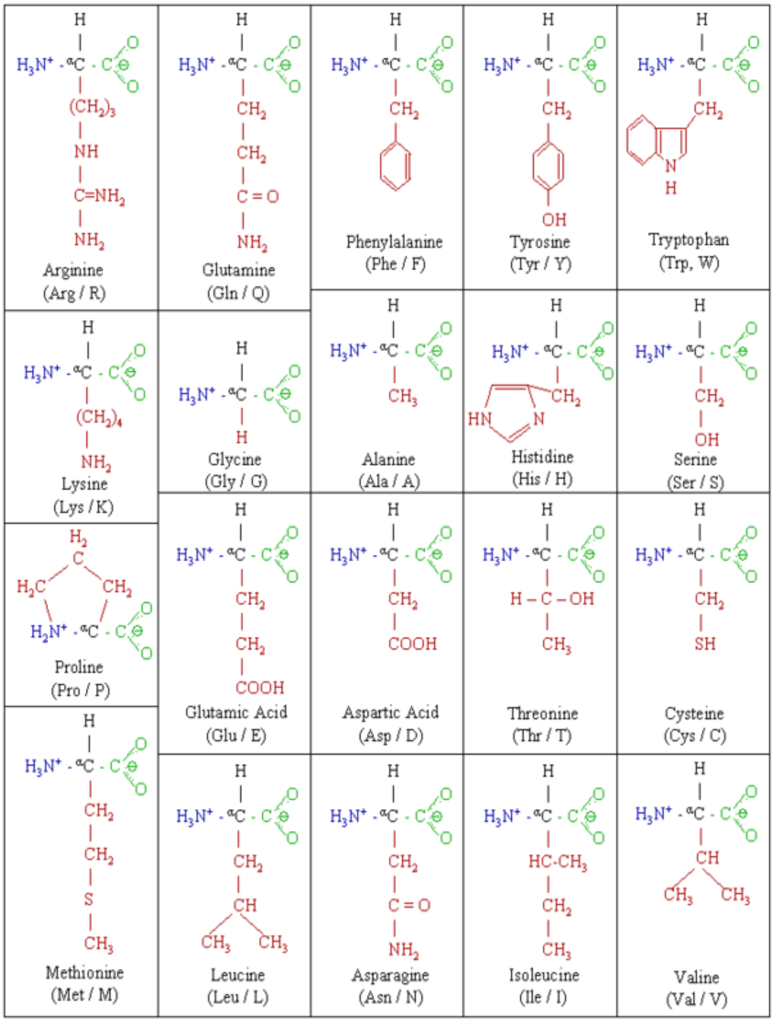


Figure 2: List of 20 different amino acids[36],Reprinted from [36]

1.2) Protein Structure

Primary Structure

The primary structure refers to the sequence of amino acid of the protein .The amino acids form covalent bonds (peptide bonds) between each other during the protein biosynthesis. The polypeptide chain thus formed has two ends a carboxyl terminus (C-terminus) and amino terminus (N- terminus).The counting of residues starts from the N-terminus. The sequence of the protein defines its structure and functions [41].

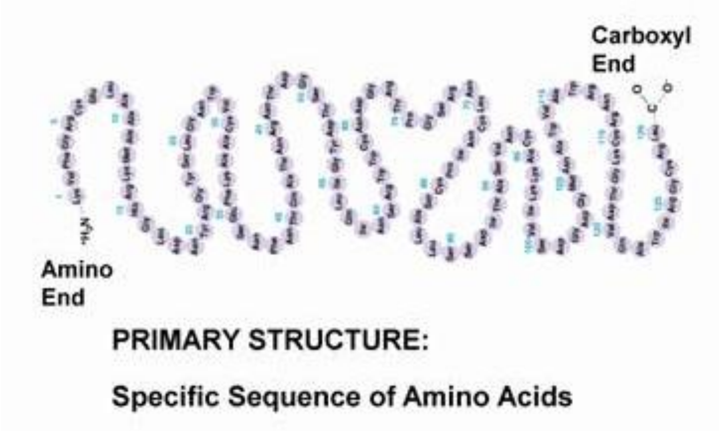


Figure 3: Primary structure of protein [35]; the figure shows Sequence of amino acids forming a protein

Secondary Structure in Proteins

The secondary structure of a protein is the local spatial arrangement of its Cα carbons (those atoms that are not part of the side chain, often referred to as the main chain) [IUPAC-IUB, 1970]. There are commonly three secondary structures in proteins, namely α-helices, β-sheets, and turns. The other structures are usually classified as random coil, or other.

Alpha helix

The **alpha helix** (α-helix) is a major part of secondary structure. The polypeptide chain turns about itself to form a spring like structure with each of the poly peptide bond, hydrogen bonded with other peptide bonds on the chain thus forming alpha-helix [34].

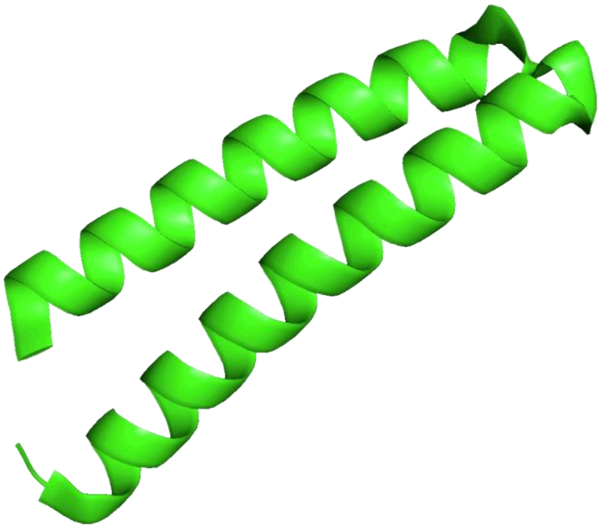


Figure 4: Two alpha helixes in protein 1ROP.

It can be either a right-handed or left-handed coiled conformation. This secondary structure is also sometimes called a classic **Pauling-Corey-Branson alpha helix** [19].

Beta sheet

The secondary structure has another major element called beta-sheet; it has several individual beta strands. A beta strand is chain of 3-10 amino acids connected by polypeptide bonds. These beta strands are connected to one another by two or more hydrogen bonds forming a pleated sheet [33].

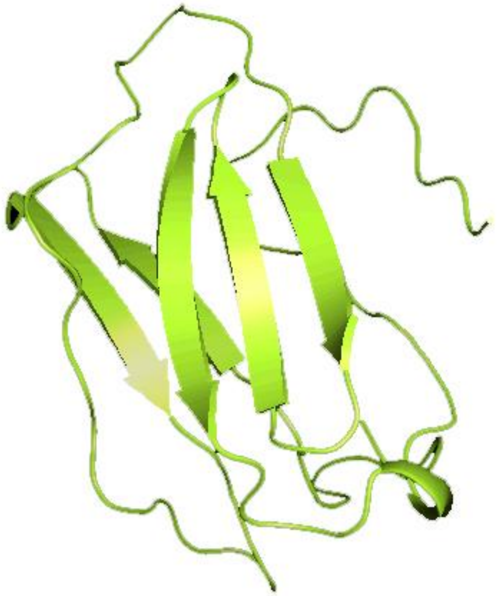


Figure 5: A beta sheet in protein 1BKZ

Tertiary Structure of Proteins

Tertiary structure is the final geometric shape a protein assumes. The bonding interactions between the side chains of each amino acids will result in several folds and bends in the protein chain [20]. These interactions finally stabilize to form tertiary structure. The tertiary structure consists of several secondary structure elements i.e. α-helices and beta sheets.

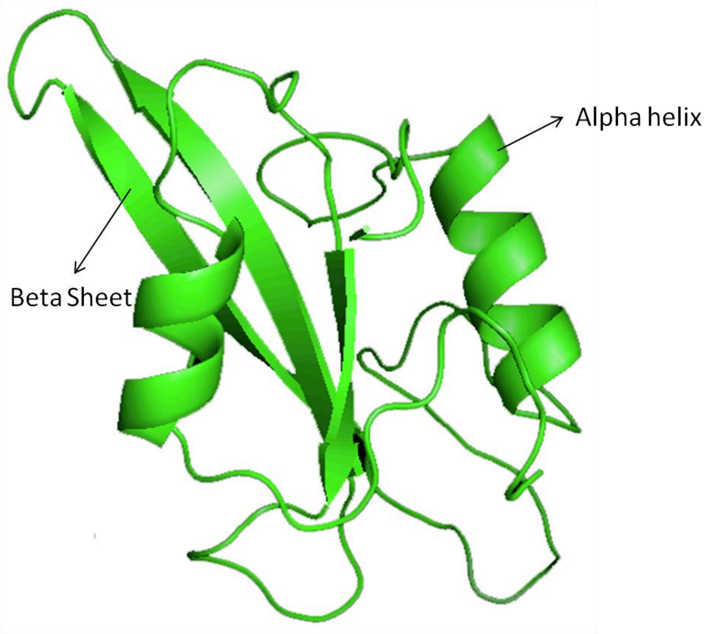


Figure 6: Tertiary structure showing alpha helix and beta sheet

Quaternary Structure

Quaternary structure is the assembly of two or more protein chains – if the protein chains have the same sequence, then this is known as a homooligomers – if they have different sequences, then the protein is a heterooligomer. The quaternary structure can be assumed as a structure consisting of stable tertiary structures as subunits which are stabilized by variety of interactions like non-covalent forces, disulfide bonds, hydrogen bonding and salt bridges [27, 28].

1.3) Protein Folding Theory

Protein folding is a physical process where a random coil (sequence of amino acids) undergoes hydrophobic collapse and eventually comes to a stable 3-dimensional structure (native structure). It is puzzling that in the protein folding process a protein undergoes a spontaneous self assembly of amino acids and forms a unique three dimensional structure despite the possibility of enormous conformational spaces [29][43]. It is assumed that if we understand the underlying physical mechanisms involved in protein folding, we can model them on computers and use algorithms to predict native structures from their amino acid sequences. Science magazine in 2005 listed the protein folding problem as one of the 125 biggest unsolved problems in science [15].

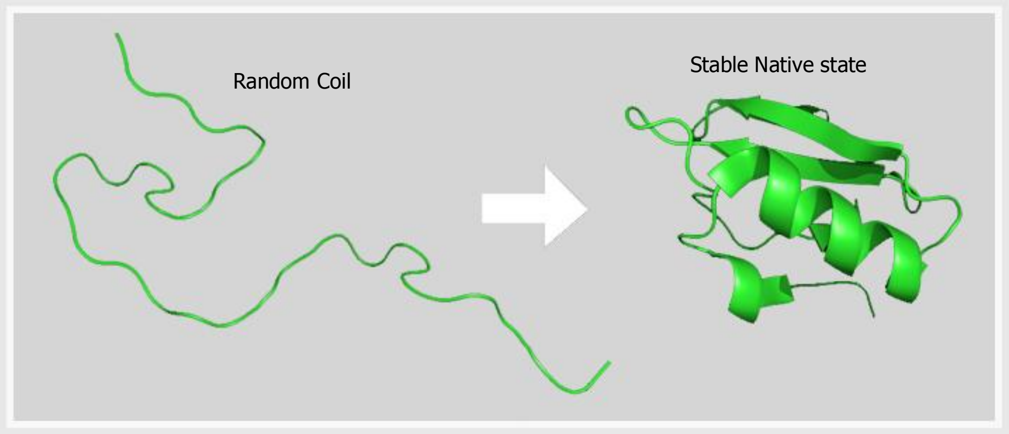


Figure 7: Protein Folding [initial state (left) → Final folded state (Right)][37] reprinted from Wikipedia

1.4) Protein Structure Prediction

One of the most challenging problems of computational molecular biology is the prediction of the proteins spatial conformation from its primary structure. Scientists and researchers have been trying to find the protein's 3-dimensional structure which helps us to understand its functionality and provides means for planning experiments and drug design.

Why is predicting 3D structure so important?

The gene is the basic unit of hereditary. It is comprised of DNA (genetic information), and its gene products, which through an extremely complex series of interactions with the products of other genes, play a large role in determining the appearance and behavior of an organism. The DNA in the gene is transcribed into messenger RNA which is translated in to sequence of amino acids. This sequence of amino acids in turn folds up in to a three dimensional structure to form a protein. This protein now interacts with other proteins (lock and key arrangements, etc.) and this interaction mediates the functions of the organism [16].

"*In fact, the 3D interactions between proteins and substrates are essentially the organism. We cannot completely understand (any predictions about) the phenotype of the organism without knowing the 3D structure of the proteins in a genome.*" -Ram Samudrala [39].

The classical method of solving the protein structure is done by using X-ray Crystallography and Nuclear magnetic resonance (NMR). These methods give good accuracy rate with resolution of 0.1A-0.3A in many cases [40]. The disadvantage with these techniques is that these techniques are very expensive and take quite a long time to determine the structure.

In 1973 Prof B. Christian Anfinsen has proposed that the information determining the tertiary structure of a protein resides in the chemistry of its amino acid sequence [41]. His research sets a new challenge for many researchers to start predicting tertiary structure from the amino acid sequence.

After the discovery of a protein's propensity to fold into its unique native state without any additional genetic mechanisms, there has been a great deal of research over the past 25 years on the prediction of 3D structure from sequence alone, without further experimental data. Sequencing of proteins is relatively fast, simple and inexpensive. Despite significant efforts, the protein folding and protein structure prediction problem remains as an unsolved problem. Due to several genomic projects increasing over time around the world, it is evident that there is a large gap between the number of known sequences and number of known three-dimensional structures. There are a few contemporary approaches toward protein structure prediction that can be roughly divided into three categories of increasing difficulty.

Homology/Comparative Modeling

The homology Modeling is based on the fact that evolutionarily related proteins with similar sequences usually exhibit similar structures. For example, two sequences that have just 35% sequence identity usually have the same overall fold. Proteins which are evolutionarily related have similar sequences and the proteins which are naturally homologous have similar protein structure [17]. By evolution three dimensional structure of protein is more conserved than the sequence itself [17].

Threading Methods

The threading methods compare the unknown protein sequence against a library of structural templates thus producing a list of scores for each template in the library [30]. These scores are sorted, the template structure with the best score is assumed to be adopted by the unknown protein. The threading method, along with the comparative modeling method, use the structures of already solved proteins as templates.

Ab initio *Method*

The *ab initio* method is based on thermodynamic hypothesis which states that "The native state of a protein is the one for which free energy achieves the global minimum" [30]. This approach does not depend on prior information from any other proteins. It is clearly the most difficult approach and arguably the most useful approach. But there could be some unresolved issues with this approach.

This method requires two things, a search algorithm to explore the protein conformational space and an energy function which evaluates whether a state is a native or not. It is extremely complex to find a useful energy function and a useful search algorithm to traverse the conformational space [30].

It can be understood that under normal physiological conditions there is a possibility of one and only one conformation which has low energy than any other conformation. So, now the search can be done in all possible conformations in random fashion until the conformation with lowest energy is found. The time complexity involved searching all the conformations is extremely high. Decreasing the time complexity of the folding process is an extremely complex task, as our algorithm should be capable of calculating the heuristic methods of finding a kinetic pathway by escaping the irrelevant conformations and finally lead us to one conformation which has lower energy than any other conformation [23]. We will further discuss these methods in next chapter.

CHAPTER 2: WEB SERVICE FOR PROTEIN STRUCTURE REFINEMENT

Protein structure refinement is the process of improving the structures of protein models to make them more like the native structures. The models obtained from comparative modeling and threading have a resolution within 1-to 3Å root means square deviation range with true structure [1]. The root mean square deviation (RMSD) is the measure of the average distance between the backbones of superimposed proteins. It is extremely challenging problem to minimize this rmsd from near native structure to true structure. The energy minimization methods help us refine the proteins models at such low resolution [1]. Recently, there has been some very encouraging progress toward solving this problem by using techniques that involve optimization of new potential functions [1]. Dr. Christopher Summa and Dr. Levitt have tested whether Potential energy minimization (PEM) could be applied for refinement and they proved it worked. The web server is uses their method for refinement of proteins over internet. This allows any user to upload his model (either pdb, ent) and obtain a refined structure.

2.1) Introduction to KB/MM Structure Refinement Method

The quest for proteins modeling has given rise to modeling of protein based on its energetics. There has been some progress in developing a molecular mechanics potential energy function, which model a protein as a collection of atoms connected by springs that hold bold lengths and angles [16]. These molecular mechanics potential energy functions (MMPEF) have two types of terms: “bonded” and “non-bonded” [16]. These functions treat atoms as spheres, and bonds as springs. Thus bond stretching, bends, twists are modeled based on spring deformation (Hooks Law) [31].

The bonded terms also include a torsional potential contributed by torsional angle rotation between atoms that are adjacent. The non bonded interactions between the atoms are due to van der Waals attractions, steric repulsion and electrostatic attraction/repulsion depending on their distance from each other[16, 31]. These van der Waals attractions and electrostatic forces are modeled based on the Lennard-Jones function and Coulomb's Law respectively. All these bonded and non-bonded interactions of the Molecular mechanics potential energy functions are derived from several sources either based on quantum mechanics, thermodynamic data, or some combination of the two [4, 5].

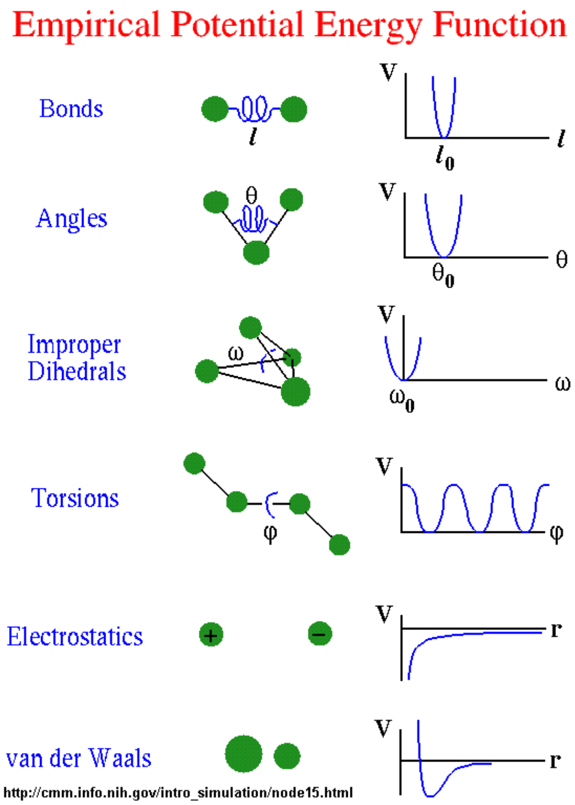

Empirical Potential Energy Function
Bonds
l
V
l0
l
Angles
θ
V
θ0
θ
Improper
Dihedrals
ω
V
ω0
ω
Torsions
φ
V
φ
Electrostatics
+
−
V
r
van der Waals
V
r
http://cmm.info.nih.gov/intro_simulation/node15.html

Figure 8: Interactions included in representative potential energy function for Molecular Dynamic simulation. Reprinted from [26]

These functions are largely used in protein folding simulation, template free modeling methods, and are also used to refine X-ray crystal structures [32]. The molecular dynamics simulation is a computer simulation where models of atoms and molecules are allowed to interact with each other and the forces between them are approximated by solving Newton's classical laws of motion. So during the refinement process if a force is applied on the protein, the effect of force on one atom is calculated and additional effect of displacement of this atom due to the force affects other atoms (either bonded or non bonded) is calculated, thus resulting in displacement of atoms resulting in new positions and velocities of the atoms**.**

The energy functions which are derived based on the statistics of the known native proteins are known as "knowledge based". These energy functions derive statistics based on the probabilities of pair wise appearance of residues in a specific geometry [16]. These probabilities are converted into potential energy using the Boltzmann equation:

$$\Delta G = -RT \ln\, (Pobserved\ /Pexpected),$$ [6, 11, 16]

Where *Pobserved* is the probability of finding a particular structural element [6, 11, 16],

Pexpected is the expected probability of finding that structural element based on chance [6, 11, 16].

The advantage of these energy functions is that they can model behavior seen in the protein database .The disadvantage is that they can't predict new behavior out of scope of database. [16]. Dr. Christopher Summa and Dr. Levitt introduced Knowledge Based /Molecular Mechanics (KB/MM). The equation representing the potential energy of the protein is given as follows

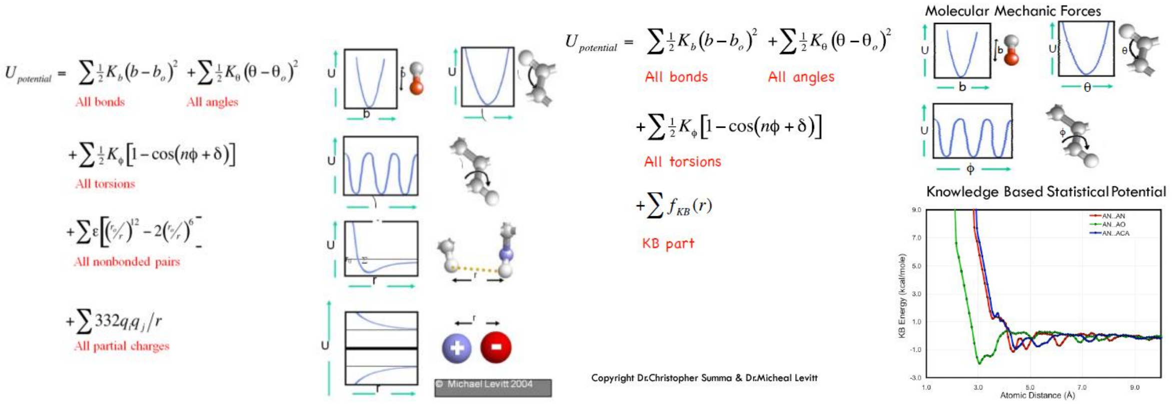
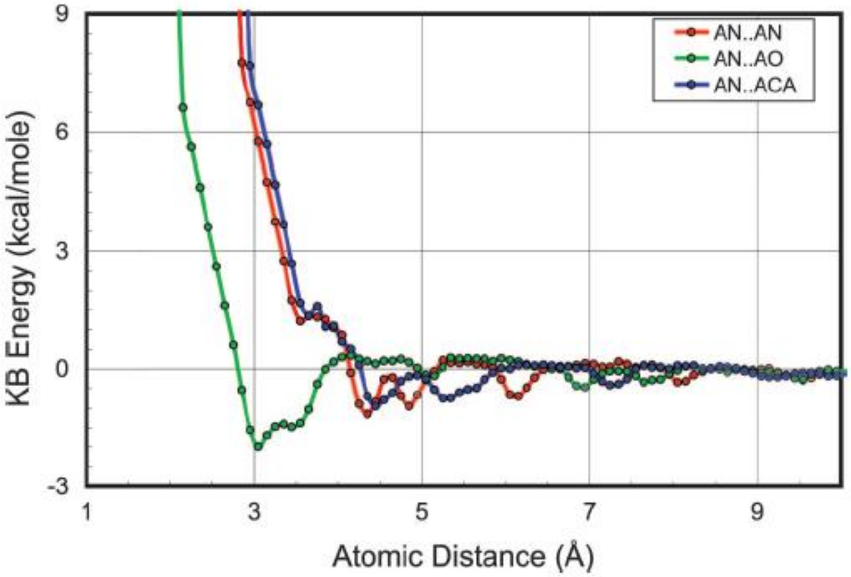


Figure 9: Equation for calculating potential energy of a protein (left); Modified energy function having Knowledge based term, Reprinted from [1].

The equation has energy terms from both the bonded and non bonded interactions .The bond angle bends, bond stretches, bond torsion angle twists contribute to the bonded interactions. The van der Walls forces and columbic forces contribute to the non bonded interactions. The energy equation was transformed in to a new equation by replacing the non-bonded interactions and all partial charges with a knowledge based statistically calculated term f_{KB} .The term f_{KB} was calculated based on the atomic pairwise Potential Mean Force (PMF) .For all the 167 atoms types which were used a pairwise PMF was derieved to describe energy interaction with every other atom type[1][9][10].



Copyright Dr. Christopher Summa & Dr. Michael Levitt Reprinted from [1]

Figure 10: Energy profiles for : AN (alanine backbone nitrogen), AO (alanine backbone carbonyl oxygen) and ACA is the alanine α carbon. The symbols shown are the energies from the PMF derivation, and the fit shown is a simple smooth curve fit in Excel, not the quintic spline as generated in ENCAD [1].

They used three different bins (0.5Å, 0.2Å, 0.1Å) .For generating the energies they have used Lu and Skolnick derivation [9]. Depending on the bins they made a histogram and when the counts go to zero they have added a repulsive part to account for steric over lap [1] These curves are smoothened by fitting a quintic spline function [1][9][10]. Using this probabilities of occurances of atoms they finally derieved an energy term which contribute non-bonded interactions(f_{KB}).

The KB/MM Hybrid potential is used to perform the refinement of proteins on the webserver

2.2) Implementation of Web Services

Summa Protein Refinement Server

The main idea of the web server is to provide a utility where users can refine their proteins. The web server takes a protein model as an input and returns the user with a refined protein.

Work Flow:

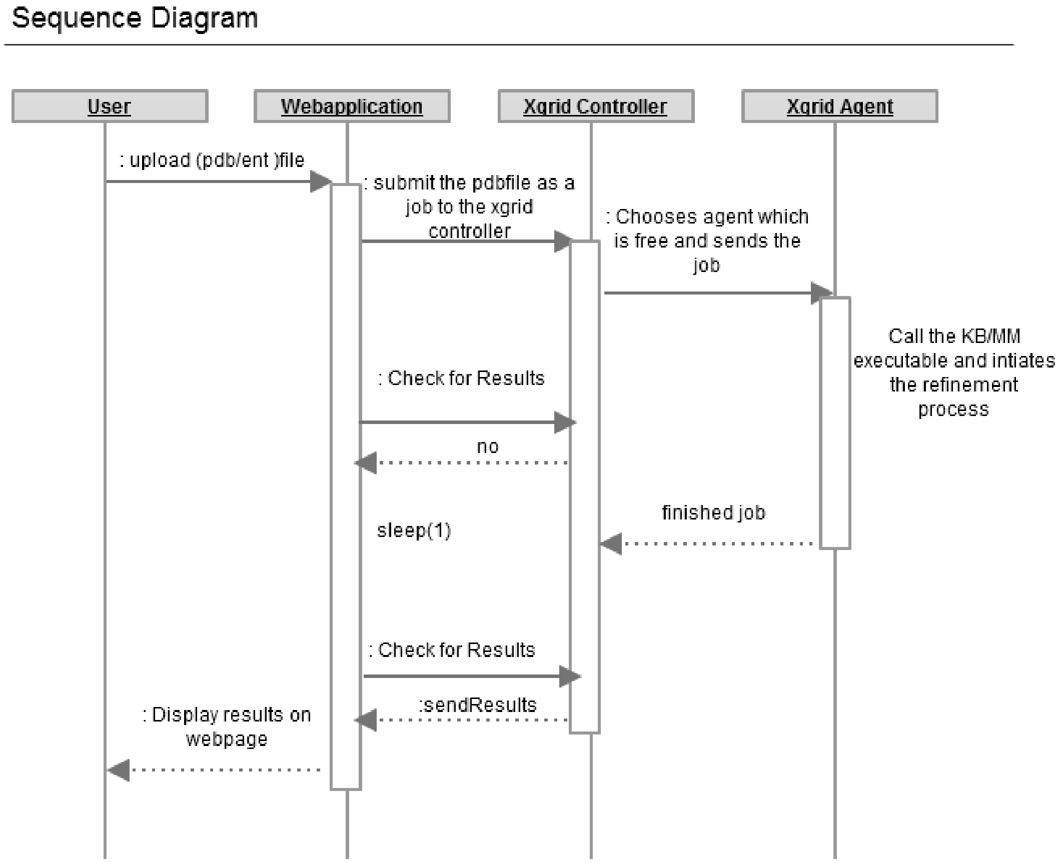

Sequence Diagram
User
Webapplication
Xgrid Controller
Xgrid Agent
: upload (pdb/ent)file
: submit the pdbfile as a job to the xgrid controller
: Chooses agent which is free and sends the job
Call the KB/MM executable and intiates the refinement process
: Check for Results
no
sleep(1)
finished job
: Check for Results
:sendResults
: Display results on webpage

Figure 11: Work Flow of Web Server

The web application work flow is as follows

- The user uploads a protein database file (pdb/ent). The web application cautions if the file uploaded is a pdb/ent file for security reasons.

- The web application creates a temporary directory and stores the uploaded file in to that folder.
- The web application gives response to the user that the file has been uploaded and then calls the script file which creates the job on the grid.
- The job is created and submitted to the Xgrid controller by the web application , then the controller checks for the agent which is idle and sends the job to that agent.
- The agent now starts to execute the job by calling encadv6lg.exe (is the main executable which handles the refinement process) under the environment variables of software ENCAD[44] (Dr. Michael Levitt).
- The XGgrid agent executes the job and generates the results.
- The web application keeps on querying the controller for results. The controller finally fetches the results from the agent and stores them in the temporary directory that it generates in the first place.
- The results thus generated are displayed as links on the webpage.
- The links for the output folder are generated on the webpage.

2.3) Tools & Methods

The basic idea of the web application is to use the executable (encadv6lg.exe) to minimize the protein file. The web application is built on PHP (Pre Hyper text Processor), Apple Xgrid Technology, Unix shell scripts.

PHP (Hypertext Preprocessor)

PHP is widely used general purpose scripting language which was designed for developing dynamic web pages. The PHP code between the php tags in php page with html content gets interpreted by php web server. PHP is available as an open source and it supports mostly all widely used operating systems.

Apple XGrid Technology

The field of bioinformatics and computational biology are emerging as an important discipline for research and industrial applications. The grid computing techniques are very useful to reduce time complexity involved in huge genomic data projects and large scale distributed applications. Apple has introduced the XGrid software (a proprietary software which implements distributed computing protocol) which was developed by Advanced Computation Group under supervision of Apple. XGrid was used for the jobs as each of them were time consuming. The XGrid architecture is outlined in **Figure 12**.

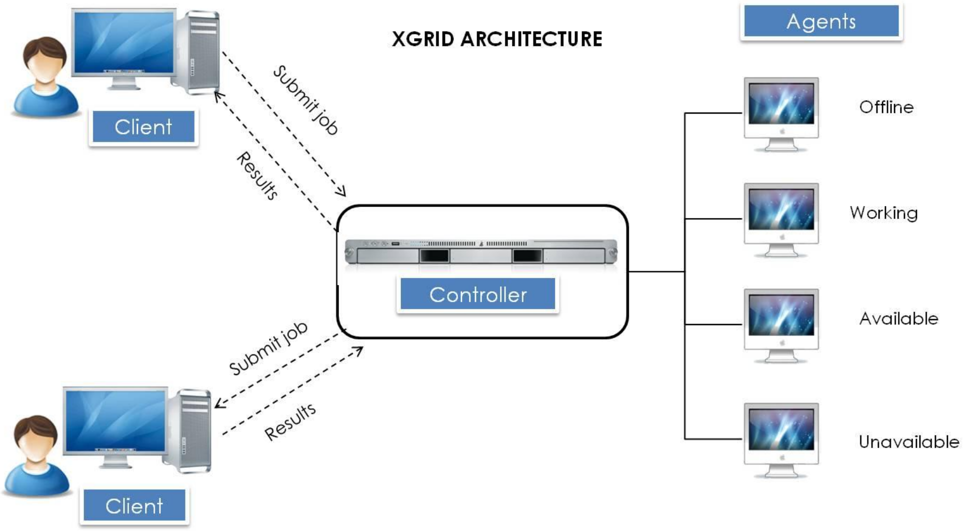


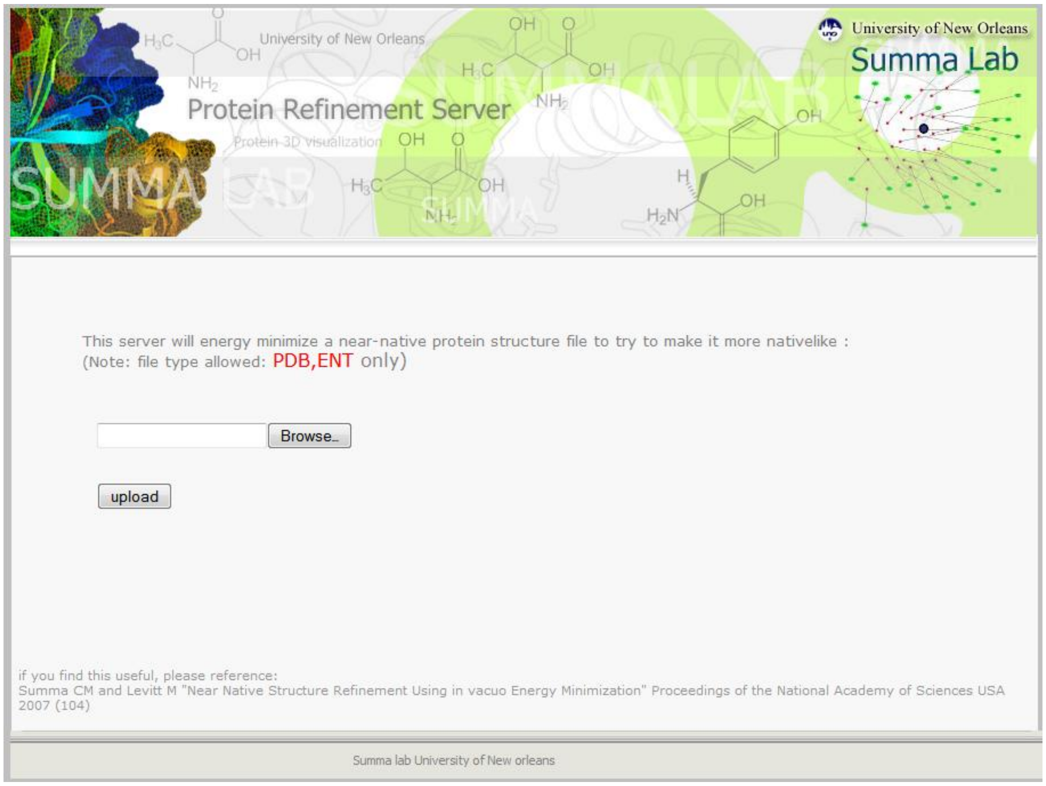
figure 12: Xgrid architecture

The XGrid works as follows,

- A client submits a job to the controller.
- The controller looks for the agent with status “available” .The “available” status indicates that the agent has free processors and is ready to work on the job.
- The job is then sent to that agent by the controller.
- The agent completes the job and stores the results in the XGrid folder, which can be obtained from XGrid command line utility.

2.4) Results

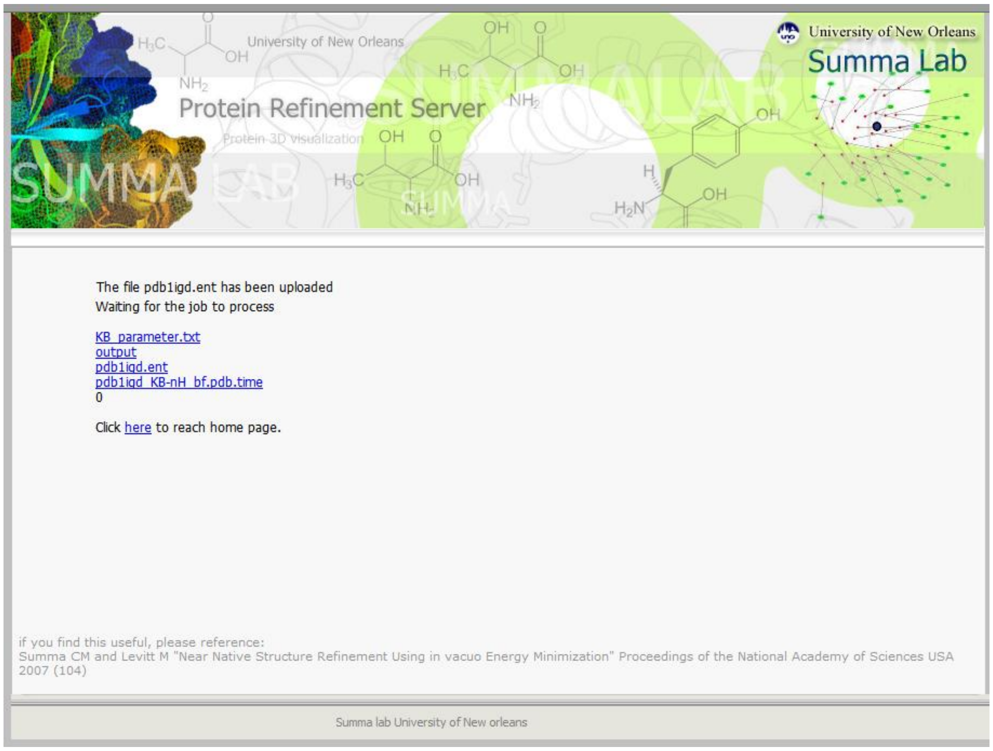
The web application was able to take the input file pbd/ent files. There is a validation system that only accepts pdb/ent files while rejecting other extensions. The input pdb file is sent to XGrid to perform minimization. The screen shot for the web server is as follows.


University of New Orleans
Summa Lab
University of New Orleans
Protein Refinement Server
Protein 3D visualization
SUMMA LAB
Browse…
upload
Summa lab University of New orleans

This server will energy minimize a near-native protein structure file to try to make it more nativelike :
(Note: file type allowed: PDB,ENT only)

if you find this useful, please reference:
Summa CM and Levitt M "Near Native Structure Refinement Using in vacuo Energy Minimization" Proceedings of the National Academy of Sciences USA 2007 (104)

Figure 13: The Home page for Protein refinement server



University of New Orleans
Protein Refinement Server
Protein 3D visualization
SUMMA LAB
University of New Orleans
Summa Lab

The file pdb1igd.ent has been uploaded
Waiting for the job to process

KB_parameter.txt
output
pdb1igd.ent
pdb1igd_KB-nH_bf.pdb.time
0

Click here to reach home page.

if you find this useful, please reference:
Summa CM and Levitt M "Near Native Structure Refinement Using in vacuo Energy Minimization" Proceedings of the National Academy of Sciences USA 2007 (104)

Summa lab University of New orleans

Figure 14: The webserver showing the links for the refined protein

2.5) Future Work

The web application can be extended with some user friendly features. The application is now capable of generating results on the webpage. The user must wait for the results for some time (until the minimization is done). In few cases if the grid is filled the job has to wait for a substantial amount of time. The other problems are connection problems, if the client waiting for execution of the minimization process has some network problems then the client fails to get results, though the job has successfully completed. The solution for both of these problems is to have an email system implemented on server side where the user inputs his pdb to the web server and provides his email id to the server. Later after the job gets completed the server emails the results to the user.

CHAPTER 3: REFINEMENT OF NEAR-NATIVE MODELS OF MEMBRANE PROTEINS

3.1) Introduction

The computational prediction of the structures of proteins that span the cellular membrane is still in its infancy relative to prediction of the structures of water-soluble proteins. Prediction of membrane protein structures is a particularly important endeavor, since they are important biological drug targets, and since the number of experimental structures of membrane proteins relative to water soluble proteins is vanishingly small, as a result of the inherent difficulties in working with these molecules experimentally. Computational prediction methods represent an alternative to expensive, time consuming, and often difficult experimental methods. In this work we test the ability of common molecular mechanics potential functions (AMBER99 and AMBER03) and a hybrid knowledge-based potential function (KB_0.1) to refine near-native structures of membrane proteins *in vacuo*. We employ the technique of potential energy minimization to a set of 88 native membrane protein structures to determine the extent to which they are perturbed away from their native, experimental structures and show that, for the majority of the proteins in our dataset, this technique does not significantly alter the structure, even in the absence of treatment of explicit or implicit membrane. As a more stringent test, large sets of near-native decoys were generated for each of the 88 membrane proteins using the technique of normal-mode perturbation. This technique was employed to sample the configuration space around the native state as evenly as possible.

The mean percentage improvement in Ca-rmsd (root mean square distance of the coordinates of the backbone a-Carbons) of the decoy sets (averaged over all 88 proteins) was 4.50% for KB_0.1, 3.97% for AMBER99 and 4.75% for AMBER03. We conclude that, while all three potentials are able to generate a modest improvement of the decoys, they clearly are able to draw near-native structures toward the native state rather than away from it for most examples we tested even in *in vacuo* simulations. More robust search methods can be used to greater improvement values, but also represent a significant increase in computational cost.

3.1.1) Membrane Proteins – Importance and Difficulties

Integral membrane proteins are defined by their ability to associate with, and span the plasma membrane of cells. They differ from water soluble proteins in the nature of the amino acid sidechains on their exterior surface – water soluble proteins have a marked tendency to display polar, hydrophilic amino acids which interact favorably with water, whereas, in the transmembrane regions of integral membrane proteins, the side chains displayed are non-polar and hydrophobic, in order to interact with the non-polar hydrocarbon chains at the interior of the phospholipid bilayer. This propensity to be associated with the bilayer makes it very difficult to work with these proteins experimentally, and the relative dearth of structural information for integral membrane proteins is the result.

There are currently ~63,000 experimental structures of water-soluble proteins in the Protein Data Bank (PDB), but only 246 (unique) structures of membrane proteins.

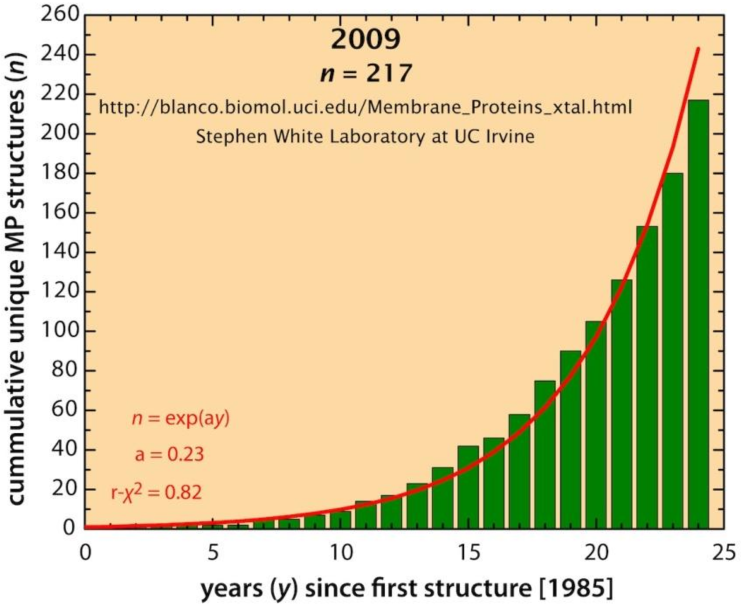


Figure 15: Cumulative growth of membrane protein structural data (reprinted from website of Stephen White's Lab, UC Irvine 2010)

Figure 16 shows the count of membrane proteins structures over the last 25 years. The graph indicates that there has not been a considerable growth in number of structures signifying the fact that these structures are hard to work with and there is a great need of research to be done in this domain.

3.1.2) Generation of Membrane Protein Dataset (Data Collection)

The Membrane protein dataset was generated by collecting membrane protein files from the Stephen Whites Lab. Stephen Whites lab website has several membrane protein structures solved either by diffraction or NMR methods. These protein files are in the form of PDB file format. The PDB file stores the 3-dimensional structural information of the protein (The pdb data is derived from X-ray crystallography, Nuclear Magnetic Resonance, and theoretical simulation. The pdb stores all the 3-dimesional co-ordinates of atoms in the proteins). The proteins initially collected were 301 membrane proteins. We have selected the proteins having single chain .The final set generated contained 88 proteins.

1AP9.pdb
1AT9.pdb
1BRX.pdb
1BXW.pdb
1BY3.pdb
1C3W.pdb
1C8R.pdb
1E12.pdb
1FI1.pdb
1FQY.pdb
1FX8.pdb
1G90.pdb
1H68.pdb
1IH5.pdb
1J4N.pdb
1JGJ.pdb
1K24.pdb
1KMO.pdb
1LKF.pdb
1MM4.pdb
1N9P.pdb
1NQE.pdb
1OKC.pdb
1ORM.pdb
1P49.pdb
1P4T.pdb
1PNZ.pdb
1PRN.pdb
1PW4.pdb
1Q9F.pdb
1QD5.pdb
1QFG.pdb
1QHJ.pdb
1QJ8.pdb
1QJP.pdb
1QKP.pdb
1SOR.pdb
1SU4.pdb
1T5S.pdb
1THQ.pdb
1XIO.pdb
1XQF.pdb
1YC9.pdb
1YGM.pdb
1YMG.pdb
2A65.pdb
2B2F.pdb
2B6O.pdb
2BRD.pdb
2C3E.pdb
2CFQ.pdb
2D1U.pdb
2D57.pdb
2F1C.pdb
2F2B.pdb
2GUF.pdb
2H8A.pdb
2IC8.pdb
2JK4.pdb
2JLN.pdb
2JMM.pdb
2JO1.pdb
2JQY.pdb
2K4T.pdb
2K73.pdb
2NR9.pdb
2O7L.pdb
2O9J.pdb
2OMF.pdb
2OQO.pdb
2POR.pdb
2QDZ.pdb
2QEI.pdb
2QJU.pdb
2UUH.pdb
2ZFG.pdb
2ZIY.pdb
3B9W.pdb
3B9Y.pdb
3C02.pdb
3DWO.pdb
3EFC.pdb
3EFM.pdb
3EMN.pdb
3F3A.pdb
3FWM.pdb
3GD8.pdb
3GJD.pdb

Table 1: Final set of pdb files used for minimization

3.1.3) Generation of Near-Native Decoys

What are decoys?

In the protein's conformational space there are lots of possible conformations a protein can assume. Based on the proteins conformational space and conformations we try to make similar conformations of proteins using computer which have some characteristics of native proteins called decoys [8].

Generating all or few of the possible protein conformations by using several algorithms is called as decoy generation (Samudrala *et al*, 1999a).The decoy data sets, consists of a solved protein structure and numerous alternative native-like structures. Decoys are used for the testing, development of scoring functions and refinement process in protein structure prediction [8, 14]. The possible protein conformations are infinite, using validation methods only few of the conformations are selected thus reducing the search space.

There are few packages available to generate decoys like Decoys R Urs (Ram Samudrala and Michael Levitt) [8], ENCAD[44] packages. We generated the decoy sets using ENCAD package.

We use the Tirion3 method to calculate the low frequency normal modes of motion for the native structures. We then perturb the native structure along those low frequency normal modes.

The total decoys sets were 97, one for each pdb file and a mean of ~504 near native structure decoys per set were generated.

3.2) Methods/Tools

3.2.1) Potential Energy Minimization / Refinement Process

The Potential energy minimization is one of the earliest methods used for refinement of protein structures [35], the structures obtained from several modeling applications are refined to obtain a native like structures. The PEM is based on thermodynamic hypothesis "The native state of a protein is the one where its free energy achieves the global minimum" [1].

Test Criteria:

We test with various kinds of force fields if they are applied on the native proteins they should not perturb them as if the force field was a perfect energy function of protein then its global minimum should match with proteins native state. The idea of comparison and test criteria is based on the idea of Dr.Christopher Summa and Dr. Michael Levitt's "Near-native structure refinement using in *vacuo* energy minimization". Their work was on refinement of soluble proteins. Similarly now we apply this refinement process on the membrane proteins. We compare and contrast the ability of different force fields to move the near native structure towards the native state [1]. To simplify this comparison we perform single refinement technique i.e., PEM in *vacuo* [1].

To setup comparison criteria between the force fields *in vacuo* we test

The refinement process should not significantly perturb the native structure [1].

The refinement process should result in movement of near native structures towards native [1].

Considering the test criteria 1, it is weak because when a force field is applied on the native structure, it may reach a local minimum and stop there, here we cannot strongly say if the force field really doing a good job. But if we consider the criteria 2 we can have an idea of the movement near the native state, whether it is towards the native or away thus we could at least analyze if our force field is trying to make structures more like native or deteriorate them. In criteria 2 we can get a global picture of shape of the curve at the native state [1].

3.2.2) Force fields Employed/Tools

GROMACS:

GROMACS stands for GROningen MAchine for Chemistry Simulation. Gromacs is a package which performs molecular dynamics by simulating the Newtonian equations of motion for systems with hundreds to millions of particles [42]. Gromacs is considered as optimized and fast software for molecular dynamics simulations. Gromacs is a command line utility for UNIX based systems, but there are few user interface implemented by developers [7].

AMBER 03/99 Force Field

Molecular dynamics is a computer simulation which calculates how a molecular system behaves over a time span [25]. Gromacs is a package which performs molecular dynamics simulation. These packages have built-in routines for energy calculations and minimization [19]. AMBER stands for Assisted Model Building with Energy Refinement.

The amber force field package is a set of molecular mechanics force fields which can be applied on the bio molecules. For implementing the refinement process the force fields amber03, amber99 are applied on the proteins and observe whether these force fields have moved these structures towards the native structure or away from native structure.

Work flow of the refinement Process

The data used in this project was collected from Stephen Whites Lab .There were 97 pdbs and for each pdb decoys were generated and a mean count of decoys was ~504.Now based on criteria 2 we perform refinement process on these decoys. The refinement process involved writing lot of scripts in perl and c-shell which are explained in detail in the following workflow.

1) For every pdb, decoys are generated and are stored in respective folder named with "pdbname_decoys".

2) The script files takes inputs as pdbfoldername and creates a job and submits to the XGrid

3) XGrid checks if there are any agents available in the grid and sends the job for execution to the agent.

4) The agent starts executing the job. The script file initializes environment and calls minimize.pl to collect all pdbs and submit each one to the actual gromacs minimization script.

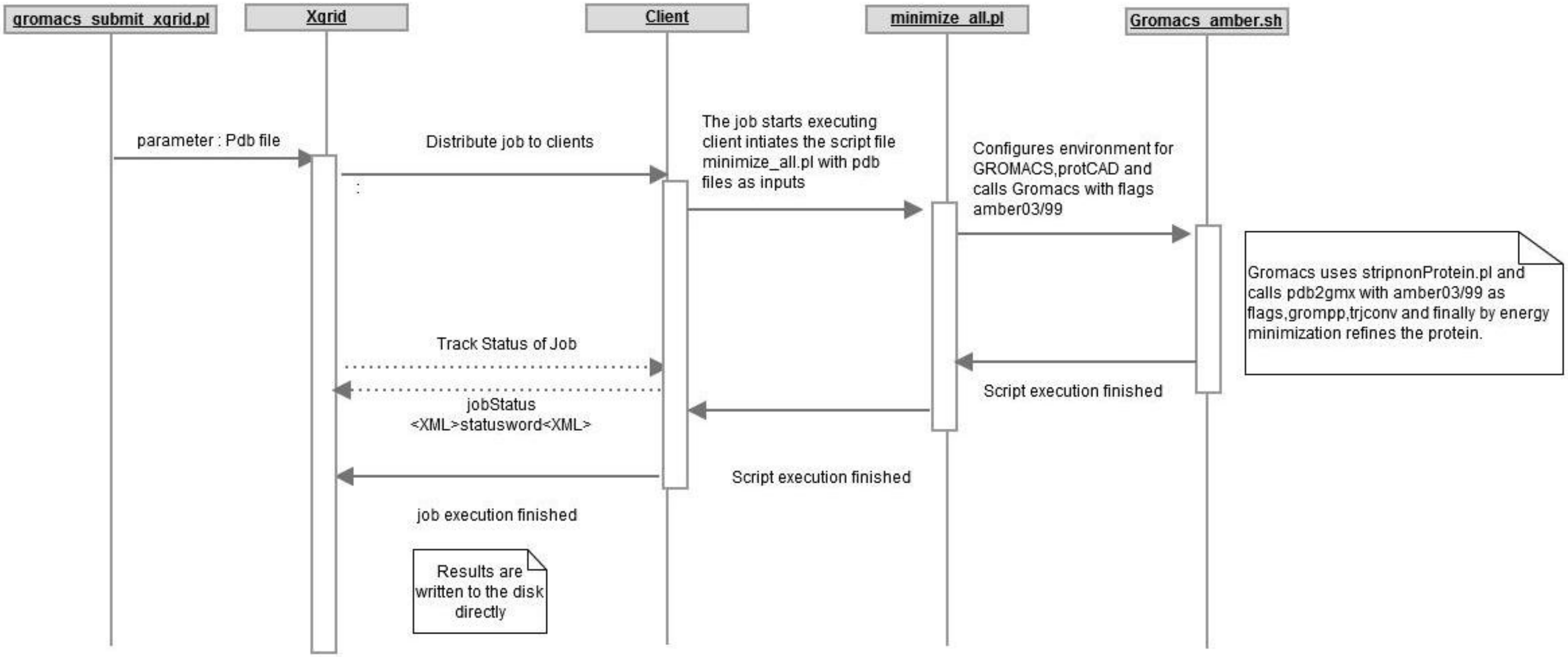
5) The gromacs minimization script calls stripNonProtein.pl script to remove non protein elements from the pdb file.

6) The gromacs script calls a set of executables to perform minimization .Initially pdb2gmx is executed with arguments amber03/99 and pdb decoy name, pdb2gmx outputs .top, .gro which serve as input grompp, the outputs from grompp serve as input to trjconv. Finally after the execution of all the three the minimized output is obtained.

7) After the minimization is done the minimized file is copied back to the decoys folder.

According to criterion 2 the force fields are applied on the decoys to see their improvement/deterioration .According to criterion 1 way the same force fields are applied on the native structures to see how much the force filed is deteriorating the native structure.

Work Flow of the Refinement Process



The job starts executing client intiates the script file minimize_all.pl with pdb files as inputs

Configures environment for GROMACS,protCAD and calls Gromacs with flags amber03/99

Gromacs uses stripnonProtein.pl and calls pdb2gmx with amber03/99 as flags,grompp,trjconv and finally by energy minimization refines the protein.

Results are written to the disk directly

Figure 16: Work flow sequence diagram for protein refinement process.

3.2.3)Knowledge Based /Molecular Mechanics Force Field

The Knowledge based molecular mechanics potential derived from the equations proposed by Dr. Christopher Summa and Dr. Levitt in **Figure 10** is used to refine proteins. The knowledge based molecular mechanics hybrid model was discussed in chapter 2 , we used the same method here to perform the energy minimization process.

The workflow of the refinement process is similar to the amber03/99 refinement process but here the executable encadv6lg.exe is used to minimize the proteins. The executable encadv6lg.exe implements the Knowledge Based /Molecular Mechanics hybrid Force Field.

3.3) Results

3.3.1) How to compare protein conformations?

There should be a method to compare our initial native conformation with refined conformation so that we know how much we have improved or deteriorated the structure. The measure used for such purpose is root mean square deviation [19].

“*The root mean square deviation (RMSD) is the measure of the average distance between the backbones of superimposed proteins*.” [38] The rmsd for a protein is calculated for the C-α atomic co-ordinate and is represented by the equation.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N} \delta_i^2}$$

Figure 17: Root means square deviation equation (Reprinted from Wikipedia)

Where δ is the distance between N pairs of equivalent atoms (usually *Cα* and sometimes *C, N, O, Cβ*).

3.3.2) Criterion 1: Energy Minimization of the Native Structures

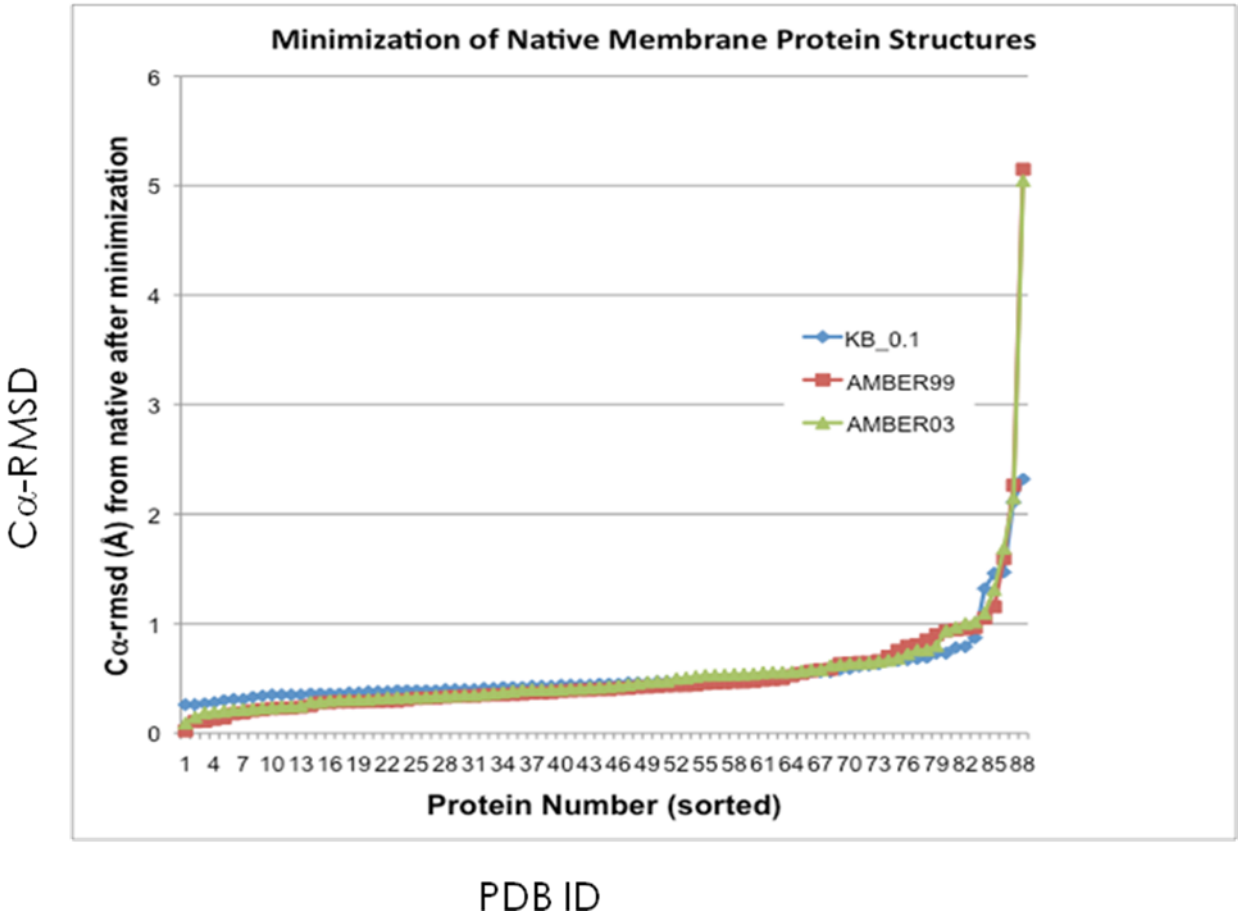


Figure 18: Minimization of Native Membrane Protein Structures

For testing criterion 1 the energy minimization is applied on the native structures as starting point and find how much the native structure is affected. Each point in the graph (**Figure 18**) represents the native structure of one of the proteins in our dataset, and the Ca-RMSD after minimization using a particular potential function is shown. A value of 0.0 for Ca-RMSD indicates that the minimized structure is exactly the same as the native, experimental structure, which is the ideal case. Lower values indicate better performance than higher values. The AMBER99, AMBER03, KB_0.1 potentials applied on the native structures show following results.

The mean deviation in rmsd for AMBER03 is 0.55 Å rmsd, AMBER99 is 0.53 Å rmsd and KB_0.1 is 0.54Å rmsd. This indicates that these force fields have not significantly deteriorated the structures.

3.3.3) Criterion 2: Energy Minimization of the Near Native Structures

For testing criterion 2 the energy minimization is applied on the near native structures as starting point and sees how much they have improved/ deteriorated when compared to the native structure.

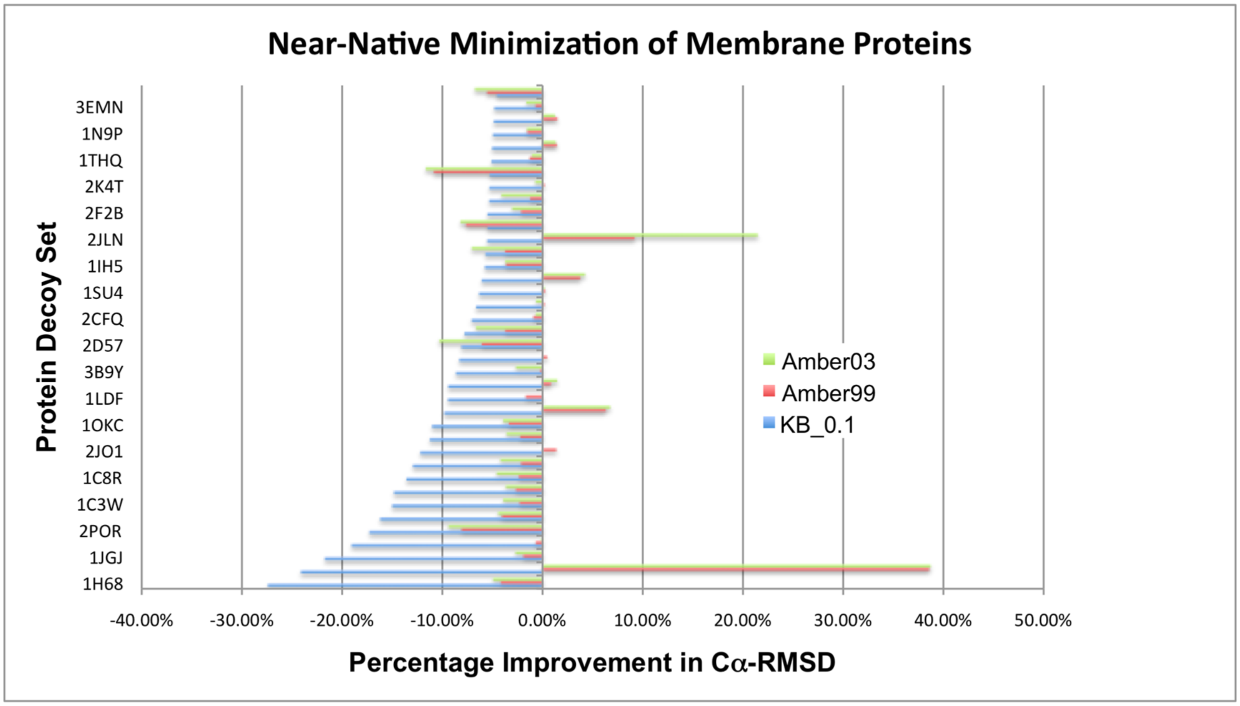


Figure 19: Near native minimization of Membrane Proteins

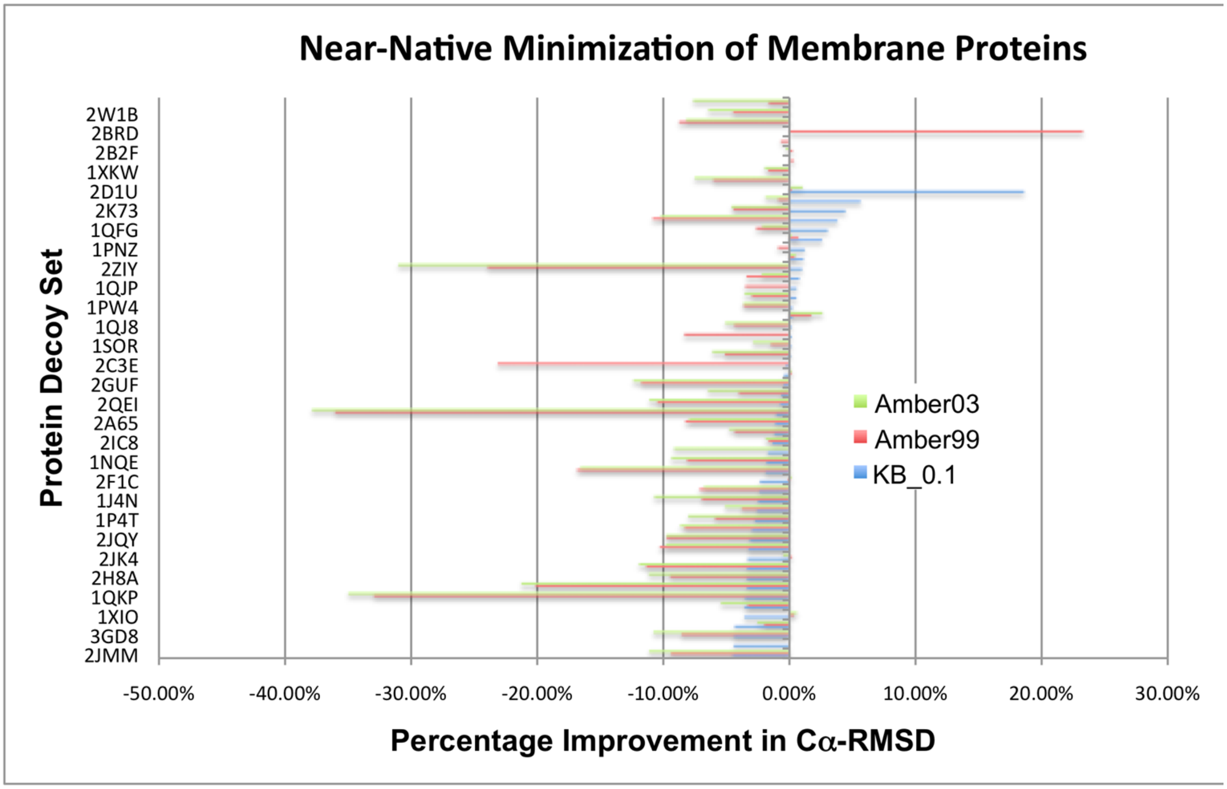


Figure 20: Near native minimization of Membrane Proteins

The graph (**Figure 20**) represents the mean percentage improvement in Cα-RMSD for the near native structure model sets .If the graph shows the bar in the left side from 0.0% for a protein it means that the structure has been improved, if it's the right side then the structure has been degraded.

The overall percentage improvement in Cα (c-alpha) rsmd with KB/MM was -4.50% Amber99 had a -3.97% while Amber03 had -4.75%.

The top and worse performers with respect to the force fields are listed in the following tables

The top performers with respect to force fields

KB	Amber99	Amber03
1H68 -27.41834 2UUH -24.11730 1JGJ -21.73361 1P49 -19.06765 2POR -17.28706	1BRX -36.0340 1QKP -32.9541 2ZIY -23.9361 2C3E -23.1384 1XQF-20.2731	1BRX -37.8618 1QKP -34.9888 2ZIY -31.0247 2C3E -23.2400 1XQF-21.2627

Table 2: Top performers with each of the force field

The worst performers with respect to force fields

KB	Amber99	Amber03
1QFG 3.039096 1YGM 3.810338 2K73 4.442954 1K24 5.646301 2D1U 18.607957	3FWM 3.75568 2QDZ 6.30944 2JLN 9.18768 2BRD 23.28000 2UUH 38.59200	3FWM 4.23802 2QDZ 6.75380 2JLN 21.47155 2BRD 23.27045 2UUH 38.72682

Table 3: Worst performers with each of the force field

Structures improved:

- KB has improved 72 of 88 protein structures
- Amber99 has improved 67 of 88 protein structures
- Amber03 has improved 70 of 88 protein structures

3.4) Discussion

The results show that three potential functions tested were able to refine the membrane proteins in our dataset using potential energy minimization. Interestingly, the KB_0.1 potential worked as well as, an in many cases better than, the traditional molecular mechanics force fields, despite having been derived using interatomic distance statistics from a dataset of water soluble proteins only. The study suggests that if a low resolution membrane protein fold has been found then we can use either traditional or knowledge based techniques to refine the membrane proteins.

3.5) Future Work

In addition to our work in this project we intended to test other energy functions (CHARMM, ENCAD, GROMACS, implicit solvent models) and test other algorithms for searching (local backbone moves, simulated annealing). Improvements in decoys set generation such that each decoy is diverse to every other in a decoy set in a range of specific resolution and perform analysis on how the force fields behave.

REFERENCES

[1] Summa CM and Levitt M. Near Native Structure Refinement Using *in vacuo* Energy Minimization. Proceedings of the National Academy of Sciences USA 2007 Feb 27;104(9):3177-82.

[2] Yu Xiaa and Michael Levitt ,Extracting knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model (*Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305)*

[3]Lazaridis, T. "Effective energy function for proteins in lipid membranes", *Proteins*, 52:176-192 (2003)

[4]. Mackerell, A. D., Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* 25, 1584-604.

[5]. Jorgensen, W. L. & Tirado-Rives, J.Potential energy functions for atomic-level simulations of water and organic and bio-molecular systems. (2005). *Proc. Natl. Acad. Sci. USA* **102**, 6665-70.

[6] Thomas PD, Dill KA.Statistical potentials extracted from protein structures: how accurate are they?
J Mol Biol. 1996 Mar 29; 257(2):457-69.

[7] Filip Jagodzinski. Guest Lecture, GROMACS, MD Tutorial, Smith College, CS 334, Bioinformatics .16 October 2008

[8] RAM SAMUDRALA and MICHAEL LEVITT, Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction.
Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305

[9] Lu H, Skolnick J (2001) *Proteins* 44:223–232.

[10] Samudrala R, Moult J (1998), *J Mol Biol* 275:895–916.

[11] Dehouck, Y, Gilis, D. & Rooman, M. (2006). A new generation of statistical Potentials for proteins. *Biophys J.* **90**, 4010-7.

[12] Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* **11**, 371-9.

[13] Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker,

D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8.

[14] **Kai Wang, Boris Fain,** Michael **Levitt** and **Ram Samudrala**
Improved protein structure selection using decoy-dependent discriminatory functions
BMC Structural Biology 2004, **4:**8doi:10.1186/1472-6807-4-8

[15] Editorial: So much more to know. Science 2005, 309:78-102.

[16] F. Edward Boas, Physics-based design of protein ligand binding [Doctor of philosophy Thesis] Stanford University May 2008.Chapter 2, P.16, 17

[17] Kaczanowski S and Zielenkiewicz P (2010). Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts* 125:543-50

[18]Ab Initio Protein Structure Prediction Using a Combined Hierarchical Approach [Ram Samudrala, Yu Xia, Enoch Huang and Michael Levitt]

[19] Zhijun Wu, Lecture notes on computational structural biology.[Internet]
(Iowa State University, USA) p.88.

[20] Elmhurst College: Elmhurst, Illinois [Internet][18th Oct 2010]
Available from: http://www.elmhurst.edu/~chm/vchembook/567tertprotein.html

[21] Birk Beck Crystallography, The Ramachandran Plot [Internet] [4th Feb 1996,18th Oct 2010] Available from : http://www.cryst.bbk.ac.uk/PPS2/course/section3/rama.html

[22] College of Saint Benedict Saint John's University [Internet]
Available from:
http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olunderstandconfo.html

[23] http://bioinsilico.blogspot.com/2008/11/secondary-structure-prediction_25.html

[24] Protein Folding, Structure and Function
Heinrich Roder, Ph.D.,Hong Cheng, Ph.D.,Harvey H. Hensley, Ph.D.,Dharmaraj Samuel, Ph.D. Paul W. Riley, B.S., Jayme Staub,* B.S.,Colin M. Hayden,*

[25] Swiss EMBnet node server,Theory of Molecular Dynamcis Simulation
Avialble from : http://www.ch.embnet.org/MD_tutorial/pages/MD.Part1.html

[26] Center for Molecular Modeling [Internet] The Empirical Potential Energy Function Steinbach [2005-08-12, 18-10-2010]
Available from: http://cmm.cit.nih.gov/intro_simulation/node15.html

[27] Protein Structure –Wikipedia the free encyclopedia [Internet] [Oct 15 2010: Oct 16 2010] Available from: http://en.wikipedia.org/wiki/Protein_structure

[28] Elmhurst College: Elmhurst, Illinois [Internet][18th Oct 2010]
Available from : http://www.elmhurst.edu/~chm/vchembook/567quatprotein.html

[29] Ken A Dill, S Banu Ozkan, Thomas R Weikl, John D Chodera and Vincent A Voelz
The protein folding problem: when will it be solved?

[30] Anna Bernasconi and Alberto M. Segre, Ab Initio Methods for Protein Structure Prediction: A New Technique based on Ramachandran Plots.

[31] Suzanne W. Slayden, Molecular Mechanics Theory in Brief
Available from: http://classweb.gmu.edu/sslayden/Chem350/manual/docs/MM.pdf

[32] F Edward Boas, Pehr B Harbury, Potential energy functions for protein design. Current Opinion in Structural Biology 2007, 17:199–204

[33] Beta sheet –Wikipedia the free encyclopedia [Internet] [Oct 15 2010: Oct 16 2010] Available from: http://en.wikipedia.org/wiki/Beta_sheet

[34]Bruce Alberts, Dennies Bray, Julian Lewis, Martin Raff, Keith Roberts, James D.Watson. Molecular Biology of the Cell. Newyork and London: Garland Publishing Inc; 1983. P.114-115.

[35] Refinement of Protein Conformations using a Macromolecular Energy Minimization Procedure , Micheal Levitt and Shneior Lifson, *Weixmann Institute of Science Rehovot, Israel (29 July, 1969) J. Mol. Biol.* (1969) 46, 269-279

[36] Luong, P.2009. Basic Principles of Genetics. Connexions,[Internet] [July 2, 2009: Oct 19 2010] Available from: http://cnx.org/content/m26565/1.1/.

[37]Protein Folding –Wikipedia the free encyclopedia [Internet] [Oct 15 2010: Oct 16 2010] Available from: http://en.wikipedia.org/wiki/proteinfolding

[38] Zhang Y. 2009. Protein Structure Prediction: Is It Useful? PubMed Central [Internet]. 19(2): 1-17[cited 2010 Oct 19]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2673339/ doi: 10.1016/j.sbi.2009.02.005

[39] Ram Samudrala [Internet] University of Washington in Seattle. Primer on protein folding problem. [Internet] Available from : http://www.ram.org/research/pfp.html

[40] Krieger E, Koraimann G, Vriend G (2002) *Proteins* 47:393–402.

[41] Anfinsen, C., Principles that govern the folding of protein chains. Science, 1973. **181**: p. 223-30.

[42]Gromacs .About Gromacs[Internet][Oct 19 2010]
Available from : http://www.gromacs.org/About_Gromacs

[43] Robert Zwanzig, Attila Szabo, and Biman Bagchi, Levinthal's paradox
Laboratory of Chemical Physics, National Institutes of Health, Bethesda, MD 20892.

[44] Levitt M, Hirshberg M, Sharon R, Daggett V (1995) *Comput Phys Commun* 91:215–231.

VITA

Kapil Pothakanoori was born in Medchal, Ranga Reddy Dist, AP, India. He received his Bachelor of Engineering in Information Technology degree from Osmania University, Hyderabad in May 2008.