DISSERTATION PROJECT ON

A. Design of Novel Tubulin Binding Molecules as Anti-Cancer Drugs and their Molecular Modeling Evaluation

B. A Computational Protocol for Protein Tertiary Structure Refinement

Submitted in partial fulfillment of the requirements for the award of degree of

MASTER OF TECHNOLOGY

In

BIOINFORMATICS

by

Ms. Komal Soni
(05/BIN/2K10)

Under the guidance of

Dr. Yasha Hasija
DTU (formerly DCE)

Prof. B Jayaram
SCFBio, IIT Delhi

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)

2010-2012

CERTIFICATE

This is to certify that the work entitled, "(A) Design of Novel Tubulin Binding Molecules as Anti-Cancer Drugs and their Molecular Modeling Evaluation and (B) A Computational Protocol for Protein Tertiary Structure Refinement" submitted by Ms. Komal Soni in partial fulfillment for the award of degree of Master of Technology in Bioinformatics from Delhi Technological University, (formerly Delhi College of Engineering), Delhi has been carried out under my supervision.

Dr. Yasha Hasija

Associate Head & Assistant Prof.
Department of Biotechnology
Delhi Technological University (formerly DCE)

INDIAN INSTITUTE OF TECHNOLOGY, DELHI
Hauz Khas, New Delhi -110016, INDIA

Dr. B. Jayaram
Ph.D. City University, NY, USA
Professor & Coordinator,
Supercomputing Facility for Bioinformatics &
Computational Biology, IIT Delhi.

Ph: +91-11-2659 1505
+91-11-2659 6786
Fax: +91-11-2658 2037
E-mail: bjayaram@chemistry.iitd.ernet.in
Website: www.scfbio-iitd.res.in

TO WHOM IT MAY CONCERN

This is to certify that the dissertation project entitled "*A Computational Protocol for Protein Tertiary Structure Refinement*" submitted by Ms. Komal Soni, a student of M.Tech Bioinformatics (Roll no - 05/BIN/2K10), Delhi Technological University (formerly DCE), in partial fulfillment of the award of the degree of M.Tech in Bioinformatics embodies the record of the original investigation carried out under my guidance at SCFBio, IIT Delhi.

It is certified that this work or part thereof has not been submitted to any other university/institution for the award of any degree.

June 27, 2012

Prof. B Jayaram

Dr. B. Jayaram
Prof. of Chemistry &
Coordinator, School of Biological Sciences &
Coordinator, Supercomputing Facility for
Bioinformatics & Computational Biology, IIT Delhi
Hauz Khas, New Delhi-110 016

DECLARATION

I hereby declare that the dissertation entitled "(A) Design of Novel Tubulin Binding Molecules as Anti-Cancer Drugs and their Molecular Modeling Evaluation and (B) A Computational Protocol for Protein Tertiary Structure Refinement" submitted for the partial fulfillment of the Master of Technology Degree in Bioinformatics is a record work done by me, during the period August 2011- June 2012 and has not formed the basis for the award of any Degree or other similar titles under my University in India or Abroad.

KOMAL SONI

(05/BIN/2K10)

ACKNOWLEGEMENT

I express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my Masters degree. I am highly indebted to my supervisor, Dr. Yasha Hasija, who has guided me through thick and thin.

I am grateful to Dr. B Jayaram for giving me the opportunity to work in his prestigious lab. I sincerely thank Ms. Priyanka Dhingra for helping and supporting me in every possible manner. The timely completion of this project is attributed to their positive attitude towards research and zest for high quality research.

Dr. P. K. Naik (Associate Professor of Biotechnology and Bioinformatics, JUIT) has been a pillar throughout.

I am thankful to Prof .P .B. Sharma (Vice Chancellor, DTU), for according me permission to undertake the project work at DTU, Delhi for partial fulfillment for the degree of Master of Technology.

I express gratitude to Prof. S. Maji (Head Of Department), Department of Biotechnology, for his moral support.

KOMAL SONI

(05/BIN/2K10)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Insight

The first section of the project deals with the designing of novel anti-cancer molecules as inhibitors of *β*-tubulin, the first step of which was modeling the structures of different tubulin isoforms. Further, I went on to design noscapine derivatives which show better binding affinities to the target and thus, are promising substitutes to noscapine (which is in clinical trials). After this part of the project was completed, it was realized that protein structure prediction and protein structure refinement are the two main steps which should be precisely carried out as the downstream processing is entirely dependent upon them. Therefore, the second section deals with the protein structure refinement problem. To this end, two main protocols have been developed. Firstly, *ab initio* modeling was executed on the longer loops, end loops and the missing secondary structural elements that exist in the protein as they are the regions which are poorly modeled and increase the divergence from the native structure. Secondly, a molecular dynamics simulation protocol has been developed and used to further refine the structures.

A. DESIGN OF NOVEL TUBULIN BINDING MOLECULES AS ANTI-CANCER DRUGS AND THEIR MOLECULAR MODELING EVALUATION

ABSTRACT

Tubulin is the target for numerous small molecule ligands which alter microtubule dynamics leading to cell cycle arrest and apoptosis. Many of these ligands are currently used clinically for the treatment of several types of cancer, and they bind to one of three distinct binding sites within *β*-tubulin (paclitaxel, vinca, and colchicine), all of which have been identified crystallographically. Unfortunately, serious side effects always accompany chemotherapy since these drugs bind to tubulin indiscriminately, leading to the death of both cancerous and healthy cells. However, the existence and distribution of divergent tubulin isoforms provide a platform upon which we may build novel chemotherapeutic drugs that can differentiate between different cell types and therefore reduce undesirable side effects. We report results of computational analysis that aims at predicting differences between the binding energies of a family of noscapine derivatives against 8 human α/*β*-tubulin isoforms. Docking and binding energy calculated using molecular mechanics generalized born surface area (MM/GB-SA) method has been used in our calculations and the results provide a proof of principle by indicating significant differences both among the derivatives and between tubulin isoforms.

Key words: tubulin isoforms, noscapine, noscapine derivatives, docking, binding energy

Chapter 1

INTRODUCTION

Cancer is a term for a large group of different diseases, all involving unregulated cell growth. In cancer, cells divide and grow uncontrollably, forming malignant tumors, and invade nearby parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or bloodstream. Not all tumors are cancerous. Benign tumors do not grow uncontrollably, do not invade neighboring tissues, and do not spread throughout the body. Healthy cells control their own growth and will destroy themselves if they become unhealthy. Cell division is a complex process that is normally tightly regulated. Cancer occurs when problems in the genes of a cell prevent these controls from functioning properly. These problems may come from damage to the gene or may be inherited, and can be caused by various sources inside or outside of the cell. Faults in two types of genes are especially important: oncogenes, which drive the growth of cancer cells, and tumor suppressor genes, which prevent cancer from developing.

Determining what causes cancer is complex and it is often impossible to assign a specific cause for a specific cancer. Many things are known to increase the risk of cancer, including tobacco use, infection, radiation, lack of physical activity, poor diet and obesity, and environmental pollutants [1]. These can directly damage genes or combine with existing genetic faults within cells to cause the disease [2]. A small percentage of cancers, approximately five to ten percent, are entirely hereditary. Cancer can be detected in a number of ways, including the presence of certain signs and symptoms, screening tests, or medical imaging. Once a possible cancer is detected it is diagnosed by microscopic examination of a tissue sample. Cancer is usually treated with chemotherapy, radiation therapy and surgery. The chances of surviving the disease vary greatly by the type and location of the cancer and the extent of disease at the start of treatment.

A mitotic inhibitor is a drug that inhibits mitosis, or cell division. These drugs disrupt microtubules, which are structures that pull the cell apart when it divides. Mitotic inhibitors are used in cancer treatment, because cancer cells are able to grow and eventually spread through the body (metastasize) through continuous mitotic division and so are more sensitive to inhibition of mitosis than normal cells.

Microtubules (MTs) are filamentous intracellular structures in all eukaryotic cells that are responsible for moving vesicles, granules, organelles like mitochondria, and chromosomes. They are important in mitosis or nuclear cell division, organization of intracellular structure, and intracellular transport, as well as ciliary and flagellar motility. A microtubule is a hollow filament of about 24 nm in diameter and is formed via polymerizations of the α/*β*-tubulin heterodimer as shown in Figure 1.

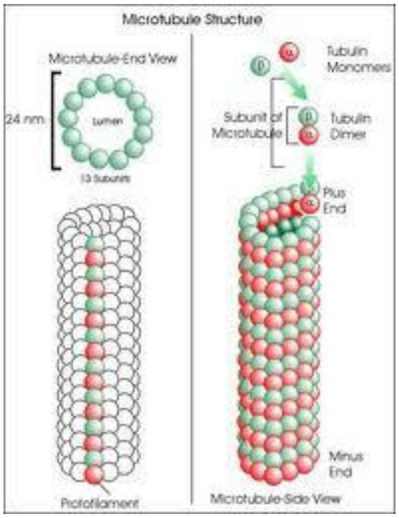


Figure 1. Microtubule structure showing tubulin polymerization

Tubulin is constituted of two 50kDa monomers: an α-subunit and a related *β*- subunit, with 40% sequence identity between them. During polymerization, α*β* tubulin heterodimers arrange head-to-tail to form straight protofilaments, which are therefore polar structures. Protofilaments interact laterally to constitute the wall of the MT. The polar nature of protofilaments gives polarity to MTs, with a fast growing (+) end exposing β subunits and a slower growing (−) end exposing α subunits [3]**.** Under conditions of steady state, at the plus end more dimers are added than lost and at the minus end more dimers are released than new ones re-associate.

Polarity is a very important feature for microtubule functioning. It is the basic property for direction-dependent cellular events, e.g., vesicle transport. Since cancer cells divide much more rapidly than normal cells, a search for inhibitor drugs that would prevent cell division, is of great importance. During mitosis, microtubules construct the mitotic spindle, which are responsible for the segregation of aligned chromosomes prior to cell division. Microtubules of mammalian cells disintegrate at temperatures below 10 °C. On the other hand, they reconstitute from tubulin *in vitro* at physiological temperature in the presence of GTP (at equimolar concentrations to tubulin) and magnesium ions. Microtubules have become the target for a large number of antimitotic agents including antitumor drugs such as the taxanes, epothilones, colchicine, and vinca alkaloids. The antimitotic and cytotoxic activity of these drugs is believed to primarily arise from suppression of dynamics in the mitotic spindle, the inhibition of spindle assembly and/or the disruption of spindle checkpoint functions [4]. The method of action of these drugs is to promote or inhibit microtubule polymerization by binding at specific sites on the interface of α/β-tubulin heterodimers. For example, the vinca alkaloid, vinblastine, binds at the intertubulin dimer interface, ultimately resulting in microtubule depolymerization [5]. Colchicine also inhibits the microtubule polymerization by binding to tubulin. On the other hand, binding of the taxanes results in an overall increase in the spindle microtubule mass with a concurrent reduction in microtubule dynamics [6,7]. They stabilize GDP-bound tubulin in the microtubule, thereby inhibiting the process of cell division - a "frozen mitosis". Epothilones also act in a manner similar to taxanes, although early trials have shown that they have better efficacy than the latter. The resulting cellular phenotype of these drugs is the induction of mitotic arrest leading to apoptosis, making them effective chemotherapeutic agents for targeting rapidly dividing cells. Nevertheless, even the most successful chemotherapy drugs have undesirable side effects that limit their utility. Their drawback is that when these drugs are given systemically, they bind tubulin indiscriminately, leading to the destruction of both cancerous and healthy cells, the consequence of which is the presence of serious side effects in all known cancer chemotherapy applications.

Noscapine (also known as Narcotine) is a benzylisoquinoline alkaloid from plants of the Papaveraceae family. It is a non addictive derivative of opium. This agent is primarily used for its antitussive (cough-suppressing) effects [8,9]. It has also been shown to have anticancer activity. Noscapine is currently under investigation for use in the treatment of several cancers and hypoxic ischemia in stroke patients.

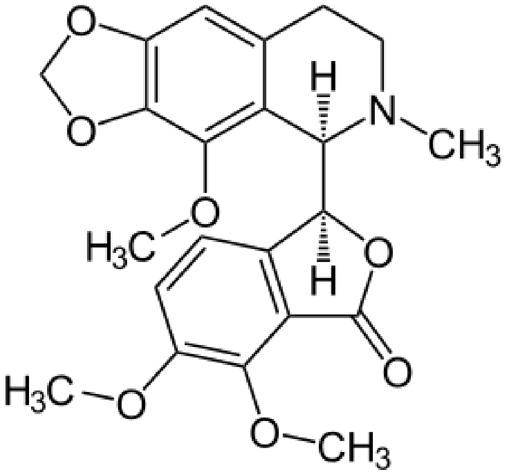


Figure 2. Scaffold structure of Noscapine

In cancer treatment, noscapine appears to interfere with microtubule function, and thus the division of cancer cells in a way similar to the taxanes. Antimicrotubule drugs disturb the assembly of the microtubules, thus preventing cell division. In normal cell growth, microtubules are formed when a cell starts to divide. Once the cell stops dividing, the microtubules are either broken down or destroyed. Anti-microtubule drugs stop the microtubules from breaking down, thus causing cancer cells to become so clogged with microtubules, so that they cannot further grow and divide. Early studies in treatment of prostate cancer are very promising. In stroke patients, noscapine blocks the bradykinine b-2 receptors. In this piece of research, we present a brief description of the tubulin isoforms, their 3-dimensional models, and the noscapine binding site. Our purpose in this study is to employ computational modeling in order to find derivatives of noscapine as potential anticancer drug candidates which would fit in the colchicines binding site of the β-tubulin. To this end, we performed both docking and free energy of binding [10,11,12] on the noscapine bound to several β-tubulin isoforms as well as on a number of derivatives of noscapine also bound to the same set of β- tubulin isoforms to elucidate which of these molecules may be considered to be potential substitute(s) for noscapine as an anticancer drug.

Chapter 2
OBJECTIVES

Cancer has long been an elusive field of research but we are yet to discover extremely efficient drugs with minimal side-effects. In this study, we plan to design novel tubulin binding drugs by derivatization from noscapine (a drug, which is already in the testing phase as an anti-cancer molecule) as potent anti-cancer molecules.

We wish to employ computational modeling in order to find derivatives of noscapine as potential anticancer drug candidates which would fit in the colchicines binding site of the β-tubulin.

Chapter 3
MATERIALS AND METHODS

3.1. Tubulin isoforms and their 3D models

Homologous protein families, such as tubulin, are collectively known as isoforms. They have amino acid sequences that diverged as a result of accumulated mutations since their separation by speciation events [13]. The resulting variations in sequence can be neutral (when they are irrelevant to the process of natural selection) or essential (when they adapt the function of a protein to a given selective pressure).The role of tubulin at the molecular level is extremely complex and seems to be related to structural variations observed between α- and *β*- isoforms [14]. The existence and distribution of tubulin isoforms provide a link to their structure and their role in the polymerization and stability of microtubules. It is clear that much of the tubulin's surface is invariant; however, those substitutions that do occur are clustered at positions that comprise the longitudinal interface between protofilaments [15]. This observation implies that there must be a contribution from the interdimer interface, between protofilaments, that is a key to our understanding of the properties that each isoform contributes to microtubule stability. Isoform composition has previously been recognized as having a demonstrable effect on microtubule assembly kinetics [16,17] whereby small differences in the binding energies and chemical affinities of different tubulin isoforms surprisingly translate into significant deviations in the growth rates and catastrophe frequencies. Short-range interactions have been studied by calculating the energy of protofilament - protofilament interactions [18]. Cells, especially cancer cells, are capable of altering the expression of each tubulin isoform (encoded by different genes) in response to external conditions that affect microtubule stability. There are several examples of this response; the most recent is the overexpression of *β*- tubulin isoform III (*β* III) following exposure to microtubule stabilizing agents such as paclitaxel [19-23]. Table 1 identifies the distribution of *β*-tubulin isoforms in normal human cells. Current antitubulin drugs bind to all of these isoforms, having only slight preference for one over another [24,25]. For example, the vinca alkaloids bind best to β II [24] providing an explanation as to their efficacy in leukemia and Hodgkin's lymphoma, since these cancerous cells express β II while normal lymphocytes do not [26]. It is to be noted that cancerous cells seem to express a variety of tubulin isoforms and are not limited to those expressed in the noncancerous cells from which they are derived [27]. Therefore, a drug that is highly specific for an isoform that is found within a cancerous cell could preferentially affect only those cells, while not harming significantly noncancerous cells. To examine the molecular properties of tubulin isoforms and the effect that they have on microtubule dynamics and drug interactions, we earlier performed a search of both the SWISS-PROT and Entrez protein databases [28]. We identified a total of eight β-tubulin isoforms. Following the identification of 83 individual protein sequences, corresponding only to β-tubulin, a ClustalW alignment was performed [29]. The alignments between the isoforms of β-tubulin were unambiguous due to the highly conserved amino acid sequences between these proteins. This alignment resulted in the filtering of both duplicates and fragmentary sequences and produced a final set of unique sequences, from which eight distinct subtypes were classified, generally based on their overall amino acid sequence and specifically their carboxy terminal tail sequence [28-32]. Of the eight subtypes, βI(GI:338695) contained two additional isoforms (gi: 18088719, 338695). The βII isoform contained additional two sequences (gi: 4507729, 29788768) that carried differences in their carboxy terminal tail, however, class βIIa may actually correspond to a pseudogene [30]. In addition to the two βII tubulin genes, we also identified three additional proteins (gi: 27227551, 49456871, 7441369) that carry minor substitutions within the coding sequence. The βIII tubulin isoform (gi: 50592996) also contained two additional proteins (gi: 62897639, 1297274). Two-class βIV tubulin genes (gi:21361322, 135470) differ only in their carboxy terminal tail sequences. The class *β*V (gi: 14210536) and *β*VI (gi: 62903515) are unique in their sequences. The class *β*VII tubulin gene family (gi: 55770868) contains additional two proteins (gi: 1857526, 12643363) that show slightly greater sequence variability than any of the other *β*-tubulin isoforms. Finally, the *β*VIII tubulin gene (gi: 42558279) is also unique in its sequence and has yet to be officially classified [16].

Isoform	Organ of expression	Cellular expression
*β*_I	constitutive	most cells
*β*_II	brain, nerves, muscle; rare elsewhere	restricted to particular cell types
*β*_III	Brain,testis, colon (very slight amounts)	neurons only sertoli cells epithelial cells only
*β*_IVa	brain only	neurons and glia
*β*_IVb	constitutive (not as widespread as β_I)	high in ciliated cells, lower in others
*β*_V	unknown	Unknown
*β*_VI	blood, bone marrow, spleen	erythroid cells, platelets
β_VII	brain	Unknown

Table 1. Tissue Distribution of β-tubulin Isoforms in Normal Cells

3.2. Modeling of structure of *β*-tubulin isoforms

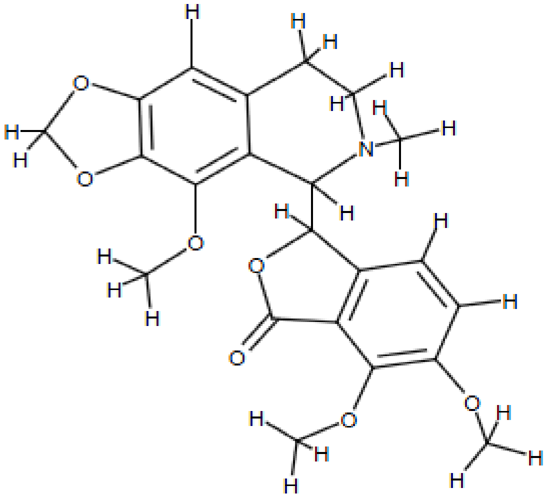
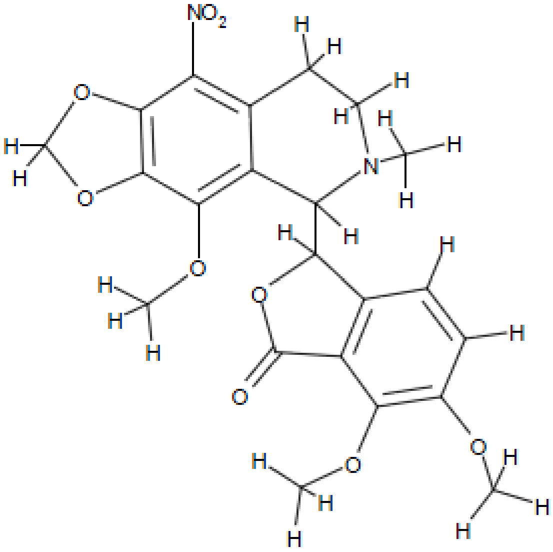
The presence of both numerous tubulin structures and multiple tubulin isoform sequences offers a unique opportunity to apply homology modeling and create a library of human *β*-tubulin isoforms, from which we can determine their key biochemical characteristics. Following the solution of the three-dimensional structure of a protein, it becomes possible to use homology modeling to predict the structure of a protein that has a similar sequence. Homology modeling utilizes several structural motifs from template proteins and pieces them together to form a final model. A scoring function assesses both the sequence identity between the target sequence and template and the overall quality of the template that is being considered. The scores are ranked and the fold with the best score is assumed to be the one adopted by the target sequence. The tubulin structure co-crystallized with colchicine (PDB ID:1SA0; resolution 3.5 A) was used as template for homology modeling of the *β*-tubulin isoforms using PRIME (version 4.5) (Schrodinger Inc.). The quality of the resulting models was then investigated, using two software packages: WHAT_CHECK and PROCHECK.

3.3. Noscapine and its binding site

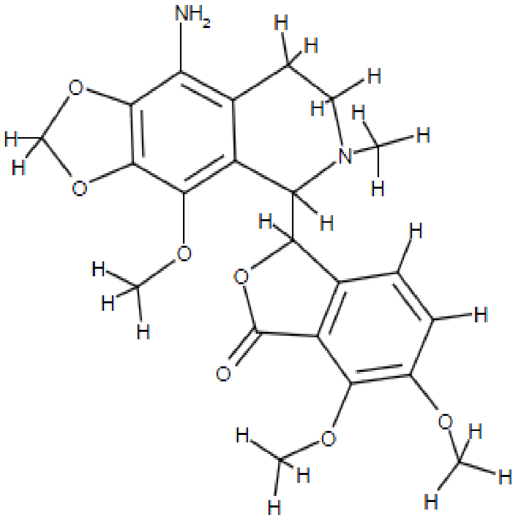
Noscapine is a non-narcotic, phthalide isoquinoline alkaloid derived from the opium poppy *Papaver somniferum*. It arrests mammalian cell cycle with intact bipolar microtubules spindles in mitosis even at high concentrations [35,36,37]. Noscapine binds tubulin dimer with a 1:1 stoichiometry [35] and alters the auto-fluorescence and circular dichroism spectrum of tubulin suggesting an alteration of the secondary structure of tubulin upon binding and arrests the mammalian cells at mitosis [35]. During the onset of mitosis, tubulin subunits assemble and disassemble vigorously to make the attachment between kinetochores of chromosomes and the plus ends of microtubules [38]. Physical tension is generated across kinetochore pairs following attachment to kinetochores and is probably regulated by the combined action of MT dynamics and MT motors within the vicinity of kinetochores [39,40,41]. The careful real time observation of individual polymerizing MTs *in vitro* and tracking the plus end growth over time revealed that noscapine affected MT-dynamics primarily by increasing the amount of time MTs spent in an attenuated pause state rather than engaging into active depolymerization and repolymerization. As a result, noscapine treatment reduced the tension generated across the kinetochore pairs as well as reduced the number of MTs attached to each pair of kinetochore [37]. During mitosis, spindle assembly checkpoint (SAC) prevents the onset of anaphase until all the chromosomes are correctly attached with MTs and proper tension is applied to the chromosomes [39]. Owing to its effect on MT dynamics, noscapine reduces tension as well as the number, therefore fails to deactivate the spindle assembly checkpoint, and the active checkpoint is responsible for the sustained mitotic arrest [37,42]. Noscapine docks onto *β*-tubulin near the interface with its dimerization partner, α- tubulin [43]. This study showed the presence of an empty space around position-9 of noscapine and that can accommodate small chemical moieties such as electronegative halogen atom (e.g. Cl-, Br-). The study suggested that addition of chemical moiety in the empty space might confer additional electrostatic interactions and hence enhanced biological activity [44]. In this study, we have taken the co-crystal structure of tubulin complexed with colchicine (PDB ID:1SA0) from *Bos taurus* for docking studies. It has been identified that colchicine also binds between α- and *β*-tubulin molecules within the heterodimer itself [45]. The binding site for colchicine was shown to be at the interface between α- and *β*-tubulin.

3.4. Ligand preparation

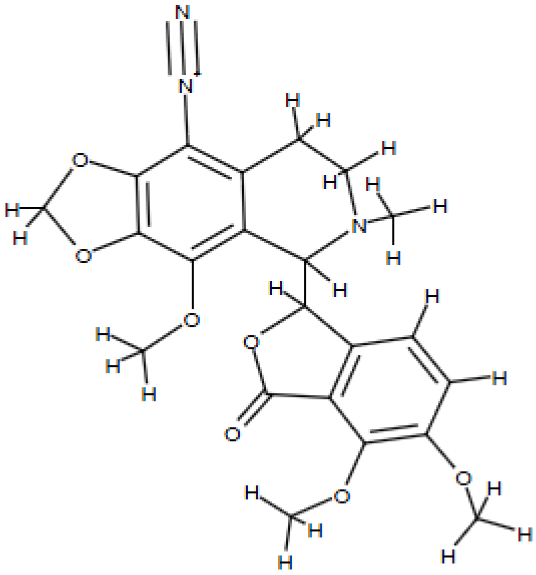
The structural derivatives of noscapine as shown in Table 2, were built from the scaffold structure of noscapine (Figure 2) and substitution of functional groups as mentioned in Table 1. We used Maestro-molecular builder for building the scaffold and structural derivatives. LigPrep [46] was used for final preparation of ligands. LigPrep is a utility of Schrodinger software suit that combines tools for generating 3D structures from 1D (Smiles) and 2D (SDF) representation, searching for tautomers and steric isomers and performing a geometry minimization of ligands. The ligands were energy minimized using Macromodel module of Schrodinger with default parameters and applying molecular mechanics force fields (MMFFs). Truncated Newton Conjugate Gradient (TNCG) minimization method was used with 500 iterations and convergence threshold of 0.05 kJ/mol.

Sl. No.	Ligand	Structure
1.	Noscapine	
2.	Nitro-noscapine	

3. Amino-noscapine



4. Azido-noscapine



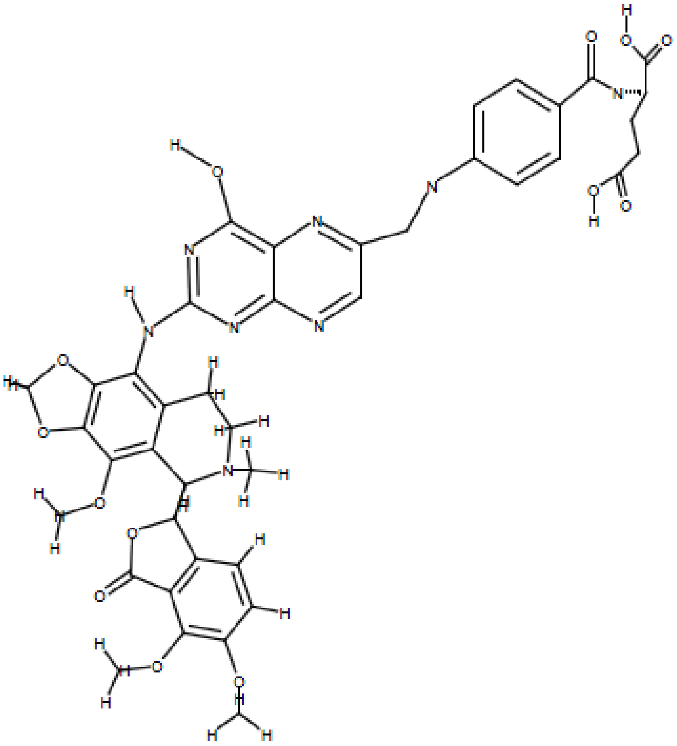
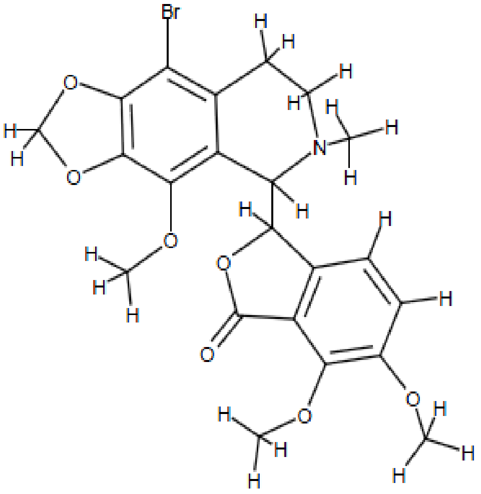
5.	Folate-noscapine	
6.	Bromo-noscapine	

Table 2. Noscapine and its structural derivatives used in the study

3.5. Docking of the ligands

The Glide program [47] was used for docking study. The Glide docking algorithm performs a series of hierarchical searches for locations of possible ligand affinity within the binding site of a receptor. A rough positioning and scoring algorithm is applied during the initial search step, followed by torsional energy optimization on an OPLS-AA non-bonded potential energy grid for enduring candidate poses. The pose conformations of the very best candidates are of MTs attached to each pair of kinetochore. It is further refined by using Monte Carlo sampling. Selection of the final docked pose is accomplished using a Glide score, which is a model energy function that combines empirical and force field based terms. The Glide score is a modified and extended version of the ChemScore function [46]. All the ligands were docked to the tubulin isoforms using Glide 4.0. After ensuring that protein and ligands are in correct form for docking, the receptor-grid files were generated using grid-receptor generation program by selecting the colchicine binding site, using van der Waals scaling of the receptor at 0.4. The default size was used for the bounding and enclosing boxes. The ligands were docked initially using the “standard precision” method and further refined using “xtra precision” Glide algorithm. For the ligand docking stage, van der Waals scaling of the ligand was set at 0.5. Out of the 50,000 poses that were sampled, 4,000 were taken through minimization (conjugate gradients 1,000) and the 30 structures having the lowest energy conformations were further evaluated for the favorable Glide docking score. A single best conformation for each ligand was considered for further analysis.

3.6. Molecular Mechanics and Free Energies of Binding

After obtaining preferable binding structure from docking simulation, the complex was partially minimized by relaxing ligand and atoms of side chains that are within 7A^{o} away from the ligand while all other atoms were fixed. Bimolecular Association with Energetics (eMBrAcE) developed by Schrodinger was used for physics based rescoring procedure [48]. The eMBrAcE (*MacroModel v9.1*) program calculates binding energies between ligands and receptors using molecular mechanics energy minimization for docked conformations. eMBrAcE applies multiple minimizations, during which each of the specified pre-positioned ligand is minimized with the receptor. For each ligand, the protein-ligand complex (*E*lig-prot), the free protein (*E*prot), and the free ligand (*E*lig) were all subjected to energy minimization in implicit solvent (generalized Born) [48,50]. It uses traditional molecular mechanics (MM) methods to calculate ligand-receptor interaction energies (*G*ele and *G*vdW), with a Gaussian smooth dielectric constant function method [51] for electrostatic part of solvation energy and solvent-accessible surface for the nonpolar part of solvation energy. A conjugate gradient minimization protocol was used in all minimization.

The eMBrAcE calculation was performed using the Ligand & Structure-Based Descriptors (LSBD) application of the Schrodinger software package. This calculation was applied to the ligand-receptor complex structures obtained from Glide docking.

Chapter 4
RESULTS AND DISCUSSION

In forming the initial model for the α/*β*-tubulin heterodimer, only one type of α –tubulin subunit was used throughout the simulations and was taken from the RCSB Protein Data Bank with the PDB identifier 1SA0[46]. The 8 human *β* -tubulin isoforms used in the calculations were obtained from Huzil et al. [49,52]. The α- and *β*-monomers were then joined together to form the 8 α/*β*-tubulin dimers. The coordinates of all the missing hydrogen atoms in the PDB structures were added using the PRIME program. The geometry of each system was optimized by energy minimization to refine the structure as well as to relieve any bad contacts among atoms due to the creation of hydrogen atom coordinates.

The original crystal structure of tubuline-colchicine complex (PDB ID: 1SA0) was used to validate the Glide-XP docking protocol. This was done by moving the co-crystallized colchicines ligand outside of active site and then docking it back into the active site. The RMSD was calculated for each configuration in comparison to the co-crystallized colchicine and the value was found to be in between 0.02–0.85 A°. This revealed that the docked configurations have similar binding positions and orientations within the binding site and the docking protocol successfully reproduces the crystal tubulin-colchicine complex. After validation of docking protocol the noscapine and its structural derivatives were docked into the colchicines binding site of tubulin isoforms.

	I	II	III	IV	V	VI	VII	VIII
Azido	-2.782	-1.684	-1.790	-2.808	-2.425	-1.092	-2.431	-1.092
Nitro	-2.671	-1.390	-1.712	-1.918	-1.942	-1.030	-2.391	-1.030
Amino	-1.618	-0.697	-1.785	-0.820	-0.629	0.149	-0.810	0.149
Bromo	-1.380	-0.351	-0.479	-1.28	-0.131	-0.152	-0.213	-0.152
Folate	-1.734	0.346	-3.599	0.139	-3.438	-0.833	-1.022	-0.695

Table 3. Calculated ΔΔGGlide score for Noscapine and its structural derivatives in 8 α/*β*-Tubulin Dimers.

	I	II	III	IV	V	VI	VII	VIII
Azido	-23.64	-58.82	-24.55	-0.9	-22.35	-36.49	-90.18	-36.49
Nitro	-34.78	-42.42	-13.86	2.58	23.85	-28.7	-45.88	-28.7
Amino	-11.06	-23.11	-24.59	-34.35	-2.91	-29.87	-31.05	-29.87
Bromo	-8.04	-30.16	1.33	-23.52	-18.63	-45	-58.36	-45
Folate	-174.95	-112.43	-136.15	-149.86	-118.28	-117.47	-187.51	-203.34

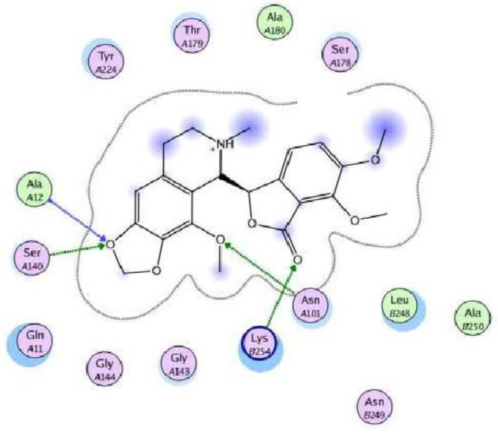
Table 4. Calculated ΔΔGGvdw for Noscapine and its structural derivatives in 8 α/*β*-Tubulin Dimers.

	I	II	III	IV	V	VI	VII	VIII
Azido	-390.71	-502.79	-487.76	-405.94	-8.45	-45.28	-560.96	-45.28
Nitro	-98.48	-196.18	-92.4	-65.94	155.1	-62.74	-69.79	-62.74
Amino	-126.78	-224.17	-119.52	-39.95	-198.47	299.91	-61.48	299.91
Bromo	-130.88	-135.21	-152.73	-40.81	-237.38	179.34	-161.75	179.34
Folate	-171.25	-643.21	-617.92	-612.13	-302.66	-193.09	-623.77	-480.09

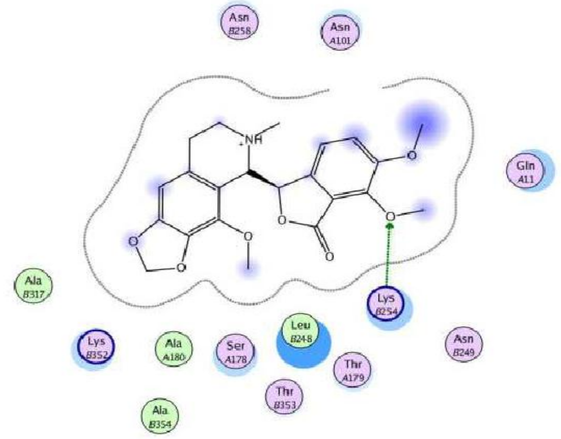
Table 5. Calculated ΔΔGGele for Noscapine and its structural derivatives in 8 α/*β*-Tubulin Dimers.

Tables 3, 4 and 5 show the calculated changes in the binding free energies with respect to noscapine which was used as a reference point for all the molecules when used as a replacement for noscapine for the 8 α/*β*-tubulin isoforms that were used in the study. The ΔΔGide score, ΔΔ*Gvdw* and ΔΔ*Gele* with negative values correspond to analogues that are more strongly bound to the α/*β*-tubulin dimers than noscapine, while those with positive values correspond to those that are less strongly bound to them. The binding mode of all these analogues with all 8 α/*β*-tubulin isoforms is shown in Figure 3. The results of the simulations indicate that there are compounds that potentially could be superior substitutes for noscapine. It has been found that, the compound azido-noscapine is the best among the compounds studied as it shows that it is more strongly bound to all isoforms with respect to noscapine followed by nitro-noscapine, amino-noscapine, bromo-noscapine and folate-noscapine on the basis of Glide score. However, the analogue folate-noscapine shows better electrostatic interaction with tubulin isoforms followed by azido-noscapine, bromo-noscapine, amino-noscapine and nitro-noscapine. Similarly the analogue folate-noscapine shows better van der Waal interaction with tubulin isoforms followed by azido-noscapine, bromo-noscapine, amino-noscapine and nitro-noscapine. This revealed that all the noscapine analogues bind efficiently with all 8 types of tubulin isoforms and could be having better therapeutic indices.

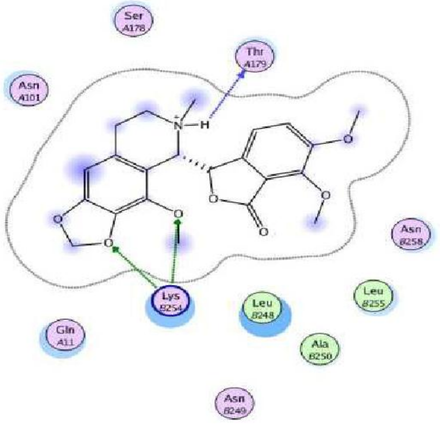
B1



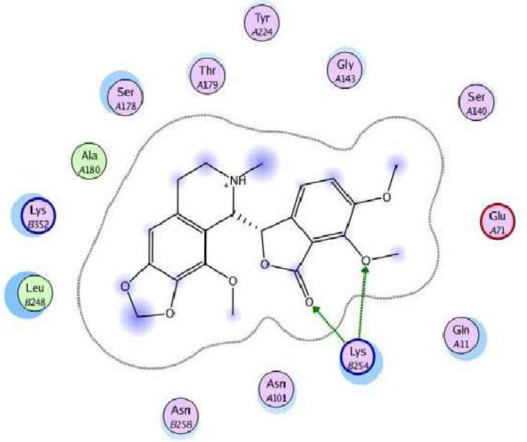
B2



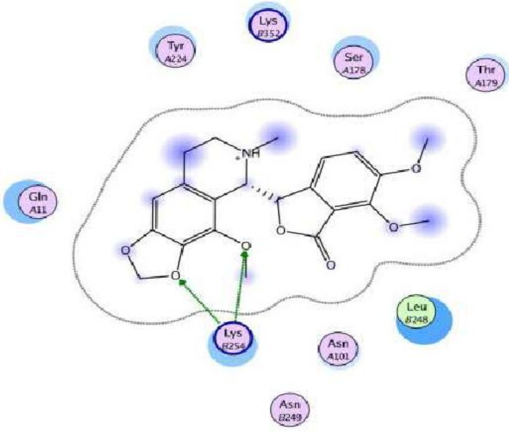
B3



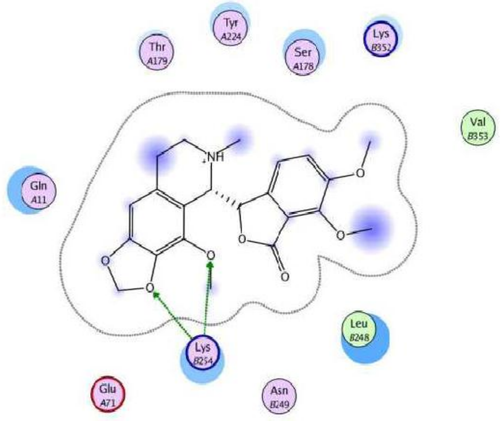
B4



B5



B6



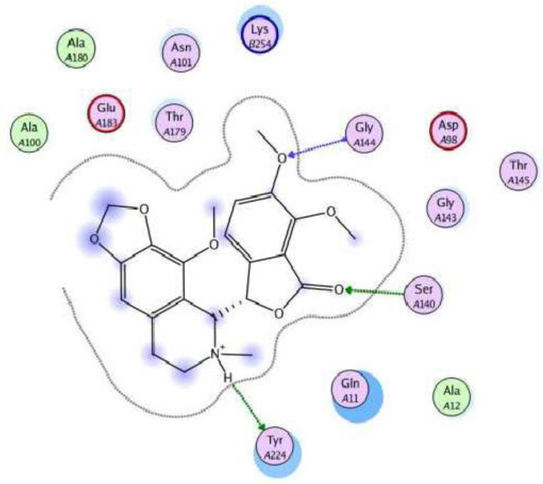
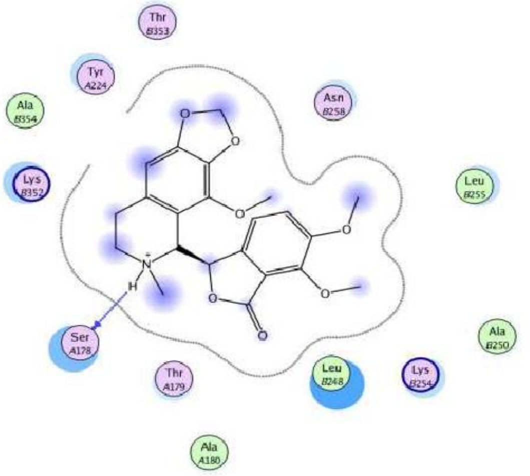
B7	B8
	

Figure 3. Ligplot showing interacting residues between the noscapine analogues and the different isoforms of *β*-tubulin.

Chapter 5
CONCLUSION

The estimated values of $\Delta\Delta G$Glide score and $\Delta\Delta G$bind presented in this study, can serve as a first of the many steps in designing, testing, and identifying specific chemotherapeutic compounds for targeting specific tubulin isoforms that are over expressed in cancer cells. The ultimate goal of cancer research is to develop a drug or treatment regimen that will target only cancer cells and will target them absolutely. The significance of microtubules as a target for chemotherapeutic treatments is outlined in a review by Jordan and Wilson [53]. They emphasize the importance of understanding the underlying mechanistic processes of these drugs when they bind to the target protein. While it is clear that a substantial amount of experimental work has to be done on obtaining kinetic data for drug binding to each β-tubulin isoform, the presence of minor variations within the structure of β-tubulin isoforms may provide us with an initial starting point for the development of novel drugs, or the derivatization of existing drugs that have increased specificity. This study provided an attempt in this direction by investigating diversity within a specific group of noscapine derivatives. We are encouraged by the results showing a degree of specificity for each tubulin isoform exhibited by the panel of the designed noscapine derivatives. We expect that these results will be supported by in vitro experiments. This type of approach, when brought to a successful completion, may eventually allow us to develop secondary treatments for cancer cell lines that have developed drug resistance, due to mutations or altered expression levels, as a result of standard chemotherapy treatments. We are aware of the existence of various mutations affecting the amino acid sequence of β-tubulin. Many of these mutations are somatic and develop within the tumor site. A number of common mutations have been identified to occur in the colchicine binding site and they may lead to the development of drug resistance over the course of chemotherapy. To overcome this complication, we intend to calculate the binding affinities of these tubulin mutants for the noscapine derivatives discussed in this study and select the optimized structures for the particular over expressed mutant which could eventually result in better cure outcomes.

REFERENCES

1. Anand P, Kunnumakkara AB, Kunnumakara AB, et al. Cancer is a Preventable Disease that Requires Major Lifestyle Changes. Pharm. Res. 2008;25 (9): 2097–116.
2. Kinzler, Kenneth W, Vogelstein B. Introduction: The genetic basis of human cancer (2nd, illustrated, revised ed.). New York: McGraw-Hill, Medical Pub. Division. p. 5, 2002.
3. Gigant B, Cormier A, Dorleans A, Ravelli R. B. G, Knossow M. Microtubule-Destabilizing Agents: Structural and Mechanistic Insights from the Interaction of Colchicine and Vinblastine with Tubulin. Top Curr Chem 2008;10:1007-1128.
4. Correia J. J, Lobert□S. Physiochemical Aspects of Tubulin-Interacting Antimitotic Drugs. Curr. Pharm. Des. 2001; 7:1213-1228.
5. Mitchison T, Kirschner M. Dynamic instability of microtubule growth. Nature 1984; 31:237–242.
6. Toso R, Jordan M, Farrell K, Matsumoto B, Wilson L. Kinetic stabilization of microtubule dynamic instability in vitro by vinblastine. Biochemistry 1993; 32:1285–1293.
7. Gigant B, Wang C, Ravelli R, Roussi F, Steinmetz M, Curmi P, Sobel A, Knossow M. Structural basis for the regulation of tubulin by vinblastine. Nature 2005; 435: 519–522.
8. Empey, D.W., Laitinen, L.A., Young, G.A., Bye, C.E. and Hughes, D.T. Comparison of the antitussive effects of codeine phosphate 20 mg, dextromethorphan 30 mg and noscapine 30 mg using citric acid-induced cough in normal subjects. European journal of clinical pharmacology 1979;16:393-397.
9. Karlsson, M.O., Dahlstrom, B., Eckernas, S.A., Johansson, M and Alm, A.T. Pharmacokinetics of oral noscapine. European journal of clinical pharmacology 1990;39:275-279.
10. Zagrovic B, Van Gunsteren, W. F. Computational analysis of the mechanism and thermodynamics of inhibition of phosphodiesterase 5A by synthetic ligands. J. Chem. Theory Comput. 2007; 3: 301–311.
11. Klauda, J. B, Brooks B. R. Sugar binding in lactose permease: Anomeric state of adisaccharide influences binding structure. J. Mol. Biol. 2007; 367: 1523–1534.

12. Bren M, Florian J, Mavri J, Bren U. Do all pieces make a whole? Thiele cumulants and the free energy decomposition. Theor. Chem. Acc. 2007; 117: 535–540.
13. Hamel E. Microtubule Proteins. CRC Press; Boca Raton, FL, 1990.
14. Wilson L, Jordan M. Microtubules. Wiley-Liss; New York, 1994.
15. Huzil J, Luduena R, Tuszynski J. Comparative modelling of human_ tubulin isotypes and implications for drug binding. Nanotechnology 2006; 17:S90–S100.
16. Panda D, Miller H, Banerjee A, Luduena R, Wilson L. Microtubule dynamics in vitro are regulated by the tubulin isotype composition. Proc. Natl. Acad. Sci. U.S.A.1994; 91:11358–11362.
17. Banerjee A, Kasmala L. Differential assembly kinetics of R-tubulin isoforms in the presence of paclitaxel. Biochem. Biophys. Res. Commun. 1998; 245:349–351.
18. Sept D, Baker N, McCammon, J. The physical basis of microtubule structure and stability. Protein Sci. 2003; 12: 2257–2261.
19. Mane Jonathan Y, Klobukowski Mariusz, Huzil J Torin, Tuszynski Jack. Free energy calculations of binding of colchicine and its derivatives with the α/β-tubulin isoforms. J. Chem. Inf. Model. 2008;48:1824-1832.
20. Ranganathan S, Dexter D, Benetatos C, Chapman A, Tew K, Hudes G. Increase of βIII and βIVa - tubulin isotopes in human prostate carcinoma cells as a result of estramustine resistance. Cancer Res. 1996; 56: 2584–2589.
21. Liu B, Staren E, Iwamura T, Appert H, Howard J. Mechanisms of taxotere-related drug resistance in pancreatic carcinoma. J. Surg. Res. 2001; 99: 179–186.
22. Hari M, Yang H, Zeng C, Canizales M, Cabral F. Expression of class III β –tubulin reduces microtubule assembly and confers resistance to paclitaxel. Cell Motil. Cytoskeleton 2003; 56: 45–56.
23. Mozzetti S, Ferlini C, Concolino P, Filippetti F, Raspaglio G, Prislei S, Gallo D, Martinelli E, Ranelletti F, Ferrandina G, Scambia G. Class III β-tubulin overexpression is a prominent mechanism of paclitaxel resistance in ovarian cancer patients. Clin. Cancer Res. 2005; 11: 298 305.
24. Khan I, Luduena R. Different effects of vinblastine on the polymerization of isotypically purified tubulins from bovine brain. InVest. New Drugs 2003; 21: 3–13.

25. Banerjee A, Luduena R. Kinetics of colchicine binding to purified β-tubulin isotypes from bovine brain. J. Biol. Chem. 1992; 267:13335– 13339.

26. Hardman J, Limbird L. Goodman & Gilman's The Pharmaco-logical basis of therapeutics, 9th ed; McGraw-Hill: New York, 1996.

27. Scott C, Walker C, Neal D, Harper C, Bloodgood R, Somers K, Mills S, Rebhun L, Levine P. β-tubulin epitope expression in normal and malignant epithelial cells. Arch. Otolaryngol. Head Neck. Surg. 1990; 116: 583–589.

28. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh L. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2004; 32: D115– D119.

29. Thompson J, Higgins D, Gibson T. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22: 4673–4680.

30. Lee M, Lewis S, Wilde C, Cowan N. Evolutionary history of a multigene family: An expressed human β-tubulin gene and three processed pseudogenes. Cell 1983; 33: 477–487.

31. Hall J, Dudley L, Dobner P, Lewis S, Cowan N. Identification of two human β-tubulin isotypes. Mol. Cell. Biol. 1983; 3:854–862.

32. Lewis S, Gilmartin M, Hall J, Cowan N. Three expressed sequences within the human β-tubulin multigene family each define a distinct isotype. J. Mol. Biol. 1985; 182:11–20.

33. Burgoyne R. D, Cambray-Deakin M, Lewis S, Sarkar S, Cowan N. Differential distribution of β-tubulin isotypes in cerebellum. EMBO J. 1988; 7:2311–2319.

34. Strausberg R. Mammalian Gene Collection Program Team. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. U.S.A. 2002; 99:16899–16903.

35. Ye K, Ke Y, Keshava N, Shanks J, Kapp J A, Tekmal R R, Petros J, Joshi H C. Opium alkaloid noscapine is an antitumor agent that arrests metaphase and induces apoptosis in dividing cells. Proceedings of the National Academy of Sciences of the United States of America 1998; 95:1601-1606.

36. Ke Y, Ye K, Grossniklaus H E, Archer D R, Joshi H C and Kapp J A. Noscapine inhibits tumor growth with little toxicity to normal tissues or inhibition of immune responses. Cancer immunology and immunotherapy. 2000; 49:217-225.

37. Zhou J, Gupta K, Yao J, Ye K, Panda D, Giannakakou P, Joshi H C. Paclitaxelresistant human ovarian cancer cells undergo c-Jun NH2-terminal kinase-mediated apoptosis in response to noscapine. J. Bio. Chem. 2002; 277: 39777-39785.

38. Mitchison T, Evans L, Schulze E, Kirschner, M. Sites of microtubule assembly and disassembly in the mitotic spindle. Cell. 1986; 45:515-527.

39. Joshi H C, Palacios M J, McNamara L, Cleveland D W. Gamma-tubulin is a centrosomal protein required for cell cycle-dependent microtubule nucleation. Nature. 1992; 356:80-83.

40. Cassimeris L, Rieder C L, Salmon E D. Microtubule assembly and kinetochore directional instability in vertebrate monopolar spindles: implications for the mechanism of chromosome congression. J. Cell Sci. 1994;107: 285-297.

41. Nicklas R B. How cells get the right chromosomes. Science. 1997; 275:632-637.

42. Landen J W, Lang R, McMahon S J, Rusan N M, Yvon A M, Adams A W, Sorcinelli M D, Campbell R, Bonaccorsi P, Ansel J C, Archer D R, Wadsworth P, Armstrong C A, Joshi H C. Noscapine alters microtubule dynamics in living cells and inhibits the progression of melanoma. Cancer res. 2002; 62:4109-4114.

43. Checchi P M, Nettles J H, Zhou J, Snyder J P, Joshi H C. Microtubule-interacting drugs for cancer treatment. Trends in pharmacological science. 2003; 24:361-365.

44. Aneja R, Lopus M, Zhou J, Vangapandu S N, Ghaleb A, Yao J, Nettles J H, Zhou B, Gupta M, Panda D, Chandra R, Joshi H C. Rational design of the microtubule-targeting anti-breast cancer drug EM015. Cancer Res. 2006; 66:3782-3791.

45. Ravelli R, Gigant B, Curmi P, Jourdain I, Lachkar S, Sobel A, Knossow M. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. Nature 2004; 428:198–202.

46. Eldridge M.D, Murray C.W, Auton T.R, Paolini G.V, Mee R.P, Empirical scoring functions: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J. Comput-Aided Mol .Desig. 1997; 11:425-445.

47. Friesner R.A, Banks J.L, Murphy R.B, Halgren T.A, Klicic J.J, Mainz D.T, Repasky M.P, Knoll E.H, Shelley M, Perry J.K, Shaw D.E, Francis P, Shenkin P.S. Glide: a new approach for rapid, accurate docking and scoring 1 Method and assessment of docking accuracy. J. Med. Chem. 2004;47:1739-1749.
48. Guvench O, Weiser J, Shenkin P.S, Kolossvary I, Still W.C. Application of the frozen atom approximation to the GB/SA continuum model for solvation free energy. J. Comput. Chem. 2002; 23:214-221.
49. Wu X, Milne J.L.S, Borgnia M.J, Rostapshov A.V, Subramaniam S, Brooks B.R. A coreweighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy. J. Struct. Bio. 2003;141:63-76.
50. Todorov N.P, Mancera R.L, Monthoux P.H. A new quantum stochastic tunneling optimisation method for protein-ligand docking. Chemical Physics Letters 2003;369:257-263.
51. Reynolds C.H. Estimating liphophilicity using GB/SA continuum solvation Model: a direct method for computing partition coefficients. J. Chem. Inf. Comput. Sci. 1995; 35:738-742.
52. Capranico G, Zagotto G, Palumbo M. Development of DNA topoisomerase-related therapeutics: a short perspective of new challenges. Curr. Med. Chem. Anti-Canc. Agents 2004; 4:335–345.
53. Jordan M, Wilson L. Microtubules and actin filaments: dynamic targets for cancer chemotherapy. Current Opinion in Cell Biology 1998; 10:123-130.

B. A COMPUTATIONAL PROTOCOL FOR PROTEIN TERTIARY STRUCTURE REFINEMENT

ABSTRACT

The field of protein structure prediction and modeling has long been researched and a number of softwares have been made available for the same. Although this field is promising, obtaining a model with the same accuracy as a crystal structure is still an unsolved problem and the refinement problem is holding back the performance of many softwares. Thus, the structure refinement of a rough model, to bring it closer to the native structure remains a major challenge. Work on this area has been ongoing for many decades and various methodologies have been used, but no method has emerged as a clear winner. Two major problems were identified in the modeled structures, one being that of loop optimization, as loops are the most variable regions of the protein which generally do not match the template and therefore, require special attention. The other one deals with the presence of steric clashes in the modeled structures, especially the ones obtained using multiple templates for modeling different regions of the protein in the form of patches. In this project, a computational protocol has been developed to tackle the above mentioned problems using *ab initio* modeling for loop optimization and a final molecular dynamics simulation on the modeled protein structures. This protocol helped decrease the RMSD and remove the steric clashes that are present in the predicted models thereby making them physically plausible.

Key words: structure refinement, loops, clashes, RMSD

Chapter 1
INTRODUCTION

Proteins are macromolecules that play a very important role in the functioning of living organisms. They are responsible for catalyzing and regulating biochemical reactions, transporting molecules, and they form the basis of structures such as skin, hair, and tendon. They are polymeric chains that are built from monomers called amino acids. Proteins are made up of 20 amino acids formed in different combinations [1].

Structural organization of proteins- They form four different levels of structural organization which can be classified into primary, secondary, tertiary and quaternary as represented in Figure 4. Primary structure is defined as the linear sequence of amino acids in a polypeptide chain. The secondary structure refers to certain regular geometric figures of the chain. Tertiary structure results from long range contacts within the chain and quaternary structure is the organization of protein subunits or 2 or more independent polypeptide chains [2]. The Secondary structure of a protein is characterized by regular elements such as alpha helices (α helices), Beta sheets (β sheets) and irregular elements such as Beta bulges, tight turns and random coils [3]. Alpha helices, Beta sheets and Turns are the three common secondary structures in proteins and the segment of polypeptide which cannot be classified into these three are grouped into the category of **Loops** [4].

Loops are the structural elements which connect two secondary structures comprising the core and are important for the overall three dimensional structure and function of proteins [5]. Despite their short length, loops are of major importance. Without loops many proteins cannot fold into compact structures. They are commonly located on the surface of the protein and therefore may be involved in binding or recognition of other molecules. They do not contribute much to protein stability but may be important for protein specific function and for interaction with other components of the cell. They vary widely in both sequence and size, even between two closely related homologous proteins [6]. Correct prediction of loops structure is considerably more difficult than the geometrically regular structures like alpha helices and beta strands.

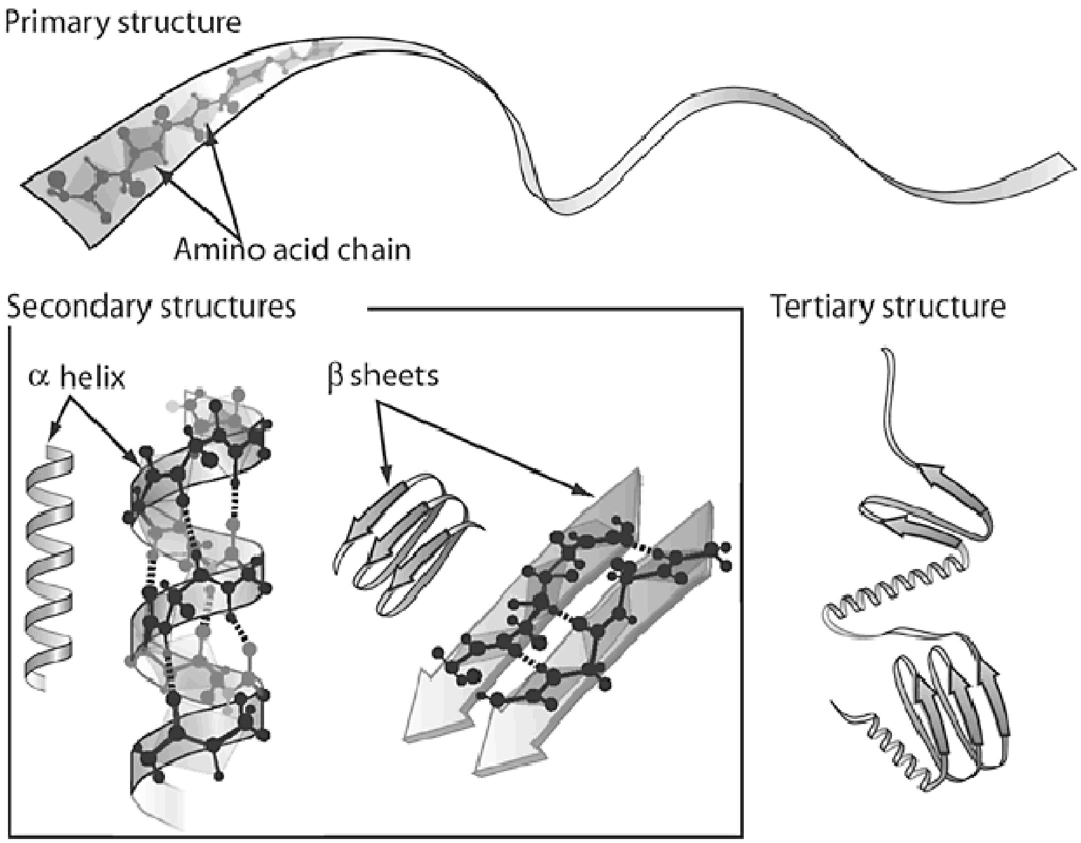


Figure 4. Basic architecture of proteins representing primary, secondary and tertiary structure (adapted from http://www.press.uillinois.edu/epub/books/brown/ch6.html)

Large scale study of proteins, their structures and functions is known as proteomics. In the 1970s, when researchers all over the world were creating databases of proteins based on the techniques like two-dimensional gel electrophoresis and relatively modern methods like mass spectrometry [7, 8], the term proteomics was coined. Today, it refers to a procedure that characterizes large sets of proteins which is only possible due to the large amount of human genome sequence available [9]. It basically deals with the protein structures and functions as well as protein-protein interactions. A subsidiary of proteomics which has emerged now and is of great interest to researchers is that of Structural Genomics [10]. It involves methodologies such as X-ray crystallograppy, NMR (Nuclear Magnetic Resonance), cryo-EM (electron microscopy) which are facilitating in acquiring angstrom-level knowledge of protein structure. The three dimensional structure of proteins is required for understanding the biological role, functional annotation and mechanism of molecular recognition. With the sharp escalation in the amount of genomic data available, there is an immediate need for elucidation of the three dimensional structure of proteins [11]. Sequence determines the protein structure which leads to knowledge. Insights into the protein structure give way to applications in many different areas of biology and medicine with the help of studies of protein-protein interactions.

The difference between the number of known protein sequences and the number of known protein 3-D structures is referred to as the sequence gap. The proliferation of genome sequencing projects is rapidly widening this gap [12]. The number of available structures for known protein sequences is limited to a meager 15%. Currently there are 36 complete eukaryotic genomes, 1695 complete prokaryotic genomes and 2683 complete viral genomes in the National Center for Biotechnology Information (NCBI) database [13, 14]. In comparison, the number of structures deposited in Protein Data Bank (RCSB) [15] as on 22 July 2011 was 74,601 as compared to the 530264 sequences available in the 2011_07 release of the UniProtKB/Swiss-Prot protein knowledgebase database [16]. Numbers of eukaryotic, prokaryotic and viral proteins with experimentally known structures, clustered based on <30% sequence similarity in PDB, are 8031, 8048 and 1013 respectively. This clearly demonstrates the high rise in the sequence-structure gap, which necessitates computational aids. It has been estimated that the generation of an experimental protein structure costs, on average, between U.S. $250,000 and $300,000. Improved methods in structure prediction, therefore, hold the promise of shifting some of the cost burden from experimentalists to (relatively) cheap computations, allowing experimentalists to focus on those structures of particular interest. The field of protein structure prediction has been revolutionized ever since the first homology model was predicted in 1969 by Browne et al. [17]. A lot of structural genomics initiatives have also taken place world wide, the biggest one being that by the National Institutes of Health at nine centers through the Protein Structure Initiative (PSI) in 2000 [18]. It started with the aim to determine the 3D structure of all proteins which could be achieved by, organizing known protein sequences into families, selecting family representatives as targets, solving the 3D structure of targets by X-ray crystallography or NMR spectroscopy and finally building models for other proteins by homology to solved 3D structures. In 2008, the PSI launched the Structural Genomics Knowledgebase to make the fruits of their work available to the society. Over the years, this initiative has greatly increased the number of submissions made to the PDB and has also helped in broadening the knowledge and understanding of protein structure and function [19]. Alongside this, developed another initiative in the form of a biannual competition, the Critical Assessment of Protein Structure Prediction (CASP) competition, which challenged the homology modeling community to validate their programs in truly blind tests [20]. These experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. The CASP competition has spurred a lot of research into homology modeling, and the ninth CASP meeting showed that this work is slowly but steadily leading to better models. Currently, the tenth CASP competition is going on, which promises to provide insight into the latest developments in the area of protein structure prediction.

1.1. Protein Structure Prediction

The field of protein structure prediction concerns itself with the generation of models of protein structures that approximate the true, native protein structure as accurately as possible. These methods are intended to augment, or even replace, the experimental determination of a protein structure in cases where the structure is either highly derivative (such as a protein with a close relative of known structure) or experimentally difficult to obtain (as with integral membrane proteins). With the help of the large PDB library as well as the number of unique SCOP [21, 22] defined folds, predicting the structure of single domain proteins has been made possible. Together PDB and sophisticated computer modeling approaches will enlarge the scope of modelable proteins [23]. Traditionally, computational protein structure prediction methods can be divided into three categories: comparative modeling or homology modeling (CM or HM), threading or fold recognition and *ab initio* methods. Although, now they are broadly classified into two basic categories: template based modeling and template free modeling.

(i). Template Free Modeling

The free modeling category was introduced in CASP7 (2006) as a replacement for the historic "*Ab initio*" category which does not use any information from the known structures. The *ab initio* methods are used to predict the structures of proteins which do not have any sequence or structural similarity with data in PDB, making them most difficult methods for protein structure prediction [24]. The *ab initio* methods have, in general, failed to predict reliable protein structures which calls for need for improvisation. Nevertheless, they still perform impressively for filling the gaps/loops or for unaligned regions in proteins with low template similarity [25] which has been seen in the CASP8 results. The new and improved free modeling category includes fragment based recombination, hybrid methods that combine multiple methods for sequence comparison and *ab initio* methods. In simper terms, the free modeling structures demonstrate a lack of sequence detectable templates and exhibit difficulties for structure prediction. Thus, the predictors are free to do whatever they can do to model structures of this category.

(ii). Template Based Modeling

Earlier attempts to compute the structure of an amino acid sequence using nothing else but a complete *ab intio* approach had very little success rate [26]. However, using information from the databases to model structures has made CM and fold recognition methods routine for protein structure prediction. In CM, an all-atom model for a target protein is predicted based on the alignment with a template of known 3D structure. The reliability of the prediction is based on the sequence similarity as well as evolutionary relationship. Threading goes slightly beyond the CM [27]. Threading is based on the principle that there are a limited number of folded protein structures in comparison to the number of sequences. It is used to identify distinctly related template sequences in the PDB, which are skipped in CM template search [28]. Threading method does a thorough scan of the databases to search for homologous as well as analogous proteins that adopt similar folds without any evolutionarily relationship [25].

The original concept of threading was introduced to calculate the potential of a structure to fit to a sequence, unlike the CM, which fits a sequence to a structure. Similarly, fold recognition term was used to compare a target sequence with a library of known folds and score based on energy as well as different scoring methods [29]. With the increasing popularity of the threading and fold recognition methods, 'threading' became a common name for these methods. Today the method of CM and threading are grouped together under the name of "Template based modeling" (TBM). The workflow of the steps involved in template based modeling has been represented in Figure 5 and the detailed description of the protocol has also been given in the next few pages.

(a). Homolog (Template) Detection and Target-Template Alignment

The initial step in TBM is based on the hypotheses that the amino acid sequence determines the native conformation of a protein, by Afinsen [30] and also that if the query sequence has high sequence identity (>30%) to the structure, the homology detection is quite straightforward which is usually done by comparing the query sequence with all the sequences of the structures in the PDB [31]. This can easily be achieved simply with dynamic programming method [32] and its derivatives [33–34]. The most popular software is BLAST that searches sequence databases for optimal local alignments to the query. The BLAST program functions very well for alignment of sequences with high similarities. But when the sequence identity is well below 30%, homology hits from BLAST are not reliable. A number of alternative strategies including template consensus sequences [35-36] and profile analysis [37–39] have been developed. All these approaches, based on either multiple sequence or structure alignments, are more sensitive because the consensus sequences are better representative of the sequence family, and the profile reflects the conserved structural or functional preferences.

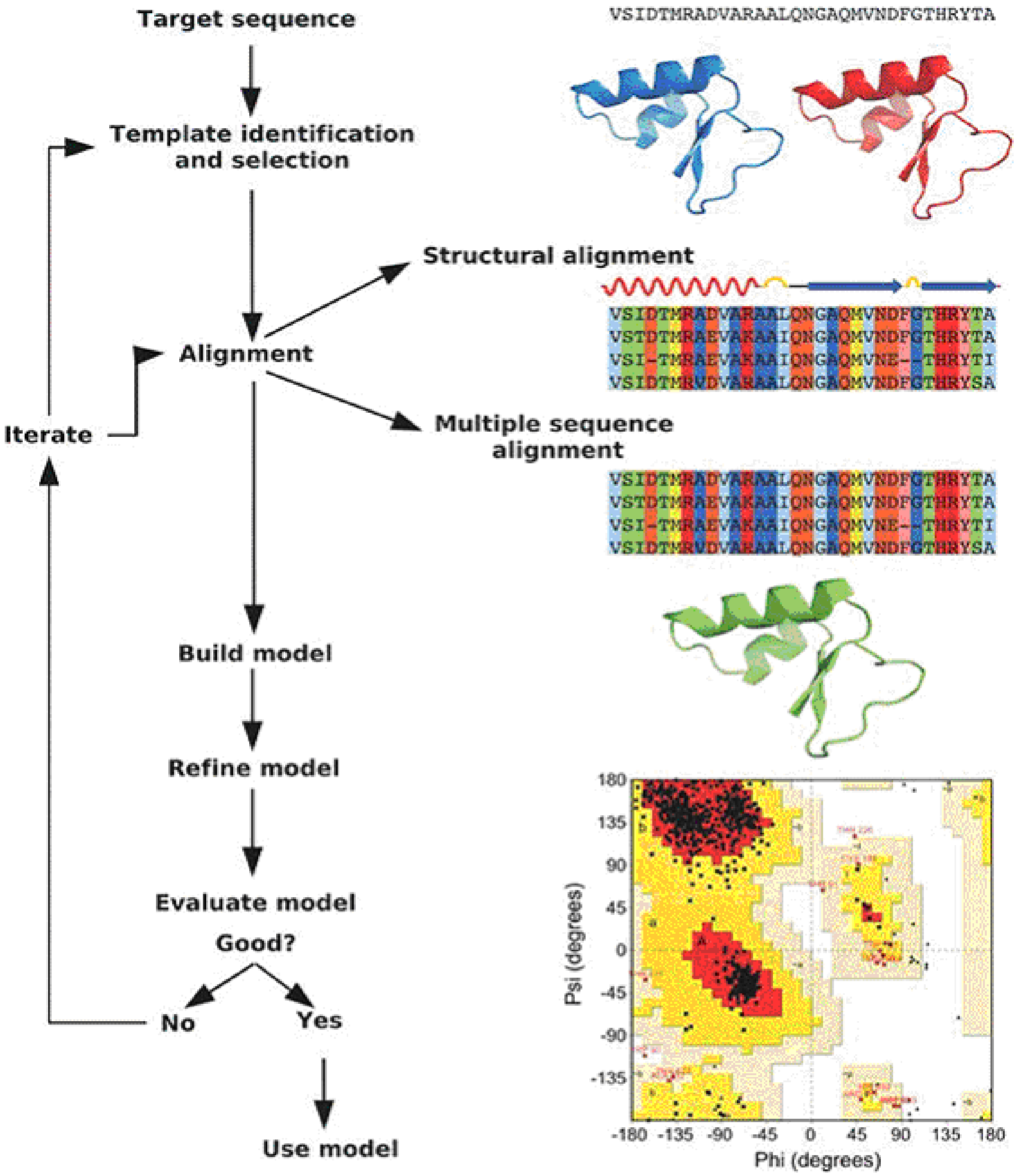

Target sequence
VSIDTMRADVARAALQNGAQMVNDFGTHRYTA
Template identification and selection
Structural alignment
VSIDTMRADVARAALQNGAQMVNDFGTHRYTA
VSTDTMRAEVAKAAIQNGAQMVNDFGTHRYTA
VSI-TMRAEVAKAAIQNGAQMVNE--THRYTI
VSIDTMRVDVARAALQNGAQMVNDFGTHRYSA
Alignment
Iterate
Multiple sequence alignment
VSIDTMRADVARAALQNGAQMVNDFGTHRYTA
VSTDTMRAEVAKAAIQNGAQMVNDFGTHRYTA
VSI-TMRAEVAKAAIQNGAQMVNE--THRYTI
VSIDTMRVDVARAALQNGAQMVNDFGTHRYSA
Build model
Refine model
Evaluate model
Good?
No
Yes
Use model
180
135
90
45
0
-45
-90
-135
Psi (degrees)
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)

Figure 5. Workflow of TBM adapted from-

http://www.scielo.org.za/scielo.php script=sci_arttext&pid=S0038-23532008000100001&lng=en

In the past several years, profile methods have emerged as the primary approach in distant homology detection. Position-specific profile search methods such as PSI-BLAST [40] and

Hidden Markov Models (HMMs) [41], as implemented in the SAM [42] and HMMER (http://hmmer.wustl.edu) packages, have vastly improved the accuracy of sequence alignments and have extended the boundaries of detectable sequence similarity. Sequence profiles methods, e.g. PSI-BLAST, start from performing a pair-wise search of the database. The significant alignments are then used by the program to construct a position specific score matrix (PSSM). This matrix replaces the query sequence in the next round of database searching. The procedure may be iterated until no new significant alignments are found. This method of iterative profile generation helps in detection of remote sequence homologs [43].

Although a major goal of the profile analysis has been remote homolog detection, an important side benefit has been significant improvement in alignment quality, even at levels of sequence identity for which pairwise alignment methods are known not to work. This, in turn, has had a positive impact on the starting alignments used in homology modeling, and thus has the potential to extend the applicability of homology modeling to increasingly lower levels of sequence similarity. Also, there are a number of publicly available tools in this area.

(b). Model Building

After the identification of the template (or templates), the next step is model generation. There are different methods which can be employed for model building, including rigid body assembly [44–47], segment matching [48], spatial restraint satisfaction method [49]. Rigid body assembly model, builds a model by assembling rigid bodies obtained from the target-template alignment. In this, all the templates are rotated and translated into a common frame of reference initially. Templates structures are then superimposed and a residue by residue correspondence is established. An initial framework is calculated by averaging coordinates of the structurally conserved regions in the templates. The main chain coordinates of the template showing maximum sequence similarity in the core region are used directly to generate core region in the target model. For identical side chains, coordinates of the template are used directly and for variable side chains rotamer libraries are used. If the side chains and loop regions differ, these are generated by scanning database of the known protein structures for loops which are compatible with the target core region. The final model is refined to remove steric clashes using energy minimization or molecular dynamics simulations [50-54]. Swiss Model [55] and 3D JIGSAW [56] use rigid modeling approach for model building. In the segment matching method, the target sequence is split into short segments and each segment is matched with the databases to generate independent models of the segments. These segments are fitted into the growing target chain until the atomic coordinates for all the residues are generated. The final model is refined using energy minimization. This method is implemented in the SEGMOD/ENCAD [57, 58]. In the last method, a 3D model is obtained by satisfying the spatial restraints derived from the alignment. Spatial restraints of the protein conformation are derived from the databases and are based on secondary-structure packing, distance geometry, hydrophobicity, stereo chemical properties, NMR experiments etc. A final model is generated by minimizing these restraints. MODELLER software uses this methodology [49, 50]. The accuracy of the generated model depends largely on the sequence similarity and the target-template alignment.

(c). Model Refinement

This step of TBM is extremely important yet difficult task as the accuracy of model prediction can be greatly improved at this step. The main focus at this stage is to improve the side chain and loop conformations. Loops are usually the most variable regions of a structure where insertion and deletion often occur. When the sequence identity is above 40%, errors in the homology structure mainly comes from side chains; when the sequence identity is between 30–40%, loops and side chains become most problematic.

The most common approach to solve the problem of side chain packing is to iteratively search the rotamer libraries for a combination of side chain conformations which are energetically favorable. These rotatmer libraries are compiled from the known structures and the resolution of the libraries has increased with the availability of more experimental structures [59-61]. Various approaches such as simulated annealing, monte carlo search, molecular dynamics refinement and mean field optimization have been used for conformational searching. Lately, improvements in side chain prediction are contributed mainly from better energy scoring functions and molecular dynamics simulations.

Loop modeling is another major problem which various researchers worldwide are trying to address for many years now. Loops are the most variable regions in a protein structure. They are the regions which are often different in the template and the target sequence and therefore require special attention. The two main approaches used for loop sampling are *ab inito* modeling and database search. Database driven methods involve searching known protein structure for segments having similar topological constraints. On the other hand, *ab initio* method involves sampling of the loop conformational space and selecting the best conformations based on energy scoring functions.

A popular approach for model refinement is to perform molecular dynamics simulations. Several attempts have been made in the past to refine models. For example Lee et al. were able to refine and select a near native model (1.8 Å Cα RMSD) starting from a 2.8 Å RMSD model among

Rosetta *de novo* models using molecular dynamics simulations with an explicit solvent and MM-PBSA free energy function [62]. Lu et al. was able to refine *ab intio* models of 30 proteins using a combination of short MD simulations and scoring using knowledge based potentials [63]. Various groups used methods such as knowledge based functions for model selection along with MD [64], all atom force field optimization for model refinement [65], replica exchange molecular dynamics simulations (REMD) with Born solvent model [66-68] and temperature dependent REMD [69]. Very recently Zhang et al. used fragment guided MD for model refinement in CASP9 [70]. An alternative to MD simulations for refinement are the Monte Carlo simulations. Misura and Baker devised a refinement protocol consisting of low resolution step followed by high resolution step. In addition to these methods for model refinement, various other servers and protocols are available for model refinement.

(d). Model Assessment

Once a model is built and refined using various strategies the next step is model quality assessment. Errors are inevitable especially when a structure is built using a template which may or may not completely resemble the target. Most of the TBM methods do not provide much information on the quality of the predictions. The choice of selecting the best model and deciding the suitability of the model is always left to the user. Further, it depends on the application. Models of high accuracy (models with >70% sequence identity) are generally used in drug design projects while low accuracy models (<50% sequence identity) have greater alignment errors and have limited utility. Several approaches have been developed in the recent past to verify the model quality in terms of global and local accuracy and the residue specific local qualities when experimental structure is not available. The Protein Structure Analysis ProSA is a widely used tool for evaluating structures. It calculates a z score indicating the quality of the model. The z-score measures the deviation of the total energy of the model with respect to energy distribution of random conformations [71-72]. Due to the bias of the z-scores on the protein size, various other size independent methods have been developed. These include methods based on stereochemistry, energy, statistical potentials and machine learning approaches. The major and commonly used estimator for protein model quality is the sequence similarity of the predicted model with the known structure. Other methods include scoring based on (a) fragment comparison in combination with a statistical potential [73], (b) distance constraints extracted from alignments of known structures [74], (c) all atom energy based scoring [75] and (d) template based scoring.

Template based scoring refers to template modeling scoring function TM-score. TM-Score extends the MaxSub and Global distance test (GDT) approaches to calculate a score for accessing model quality. Global Distance test score counts the number of Cα pairs with distance of 1, 2, 4 and 8Å after superimposition. MaxSub identifies maximum substructure within a distance limit of < 3.5 Å. Unlike MaxSub and GDT score TM-score counts residue pairs using Leviit-Gerstein weights [76]. Protein pairs with TM score >0.5 are mostly in same fold whereas pairs with TM score < 0.5 do not have the same fold [77]. Another method for model quality assessment is QMEAN. QMEAN is a composite scoring function i.e. a linear combination of six structural descriptors torsion angle potential, solvation potential, distance dependent interaction based potentials, terms describing agreement of predicted and calculated secondary structure and solvent accessibility. The QMEAN score ranges from 0 to 1 with higher values referring to more reliable models [78-81]. Other methods for quality assessment include Verify-3D [82-83], WHAT-IF [84] and PROCHECK [85]. A plethora of algorithms for model quality assessment are available ever since CASP7 when for the first time quality assessment was treated as a separate category [86]. Together, these algorithms allow a more detailed view of the models for its potential applications. Although a number of quality assessment parameters are available as shown above, Global Distance Test Total Score (GDT_TS) remains the predominant one, and is used in determining the best predictions in the CASP scenario. Alongside this, Root Mean Square Deviation (RMSD) is also considered to be an important parameter.

1.2. Bhageerath and Bhageerath-H for Protein Structure Prediction

Bhageerath [87-91] is an energy based software suite for predicting tertiary structures of small globular proteins. The protocol comprises eight different modules which uses physicochemical properties of proteins and *ab initio* methodology to predict five candidates for the native from the input query sequence. The methodology has been validated on 80 small globular proteins with < 100 amino acids. For each of these proteins a structure within 3-7Å RMSD (root mean square deviation) from the native is predicted within few minutes to hours on a 280 processor cluster (~2 Teraflops of computing capacity). In this project, this software suite has been used for the *ab initio* loop optimization in the modeled protein structures.

Bhageerath-H [92] is a homology *ab initio* hybrid server for protein tertiary structure prediction. The protocol identifies regions having local sequence similarity with database to generate 3D fragments which are patched with *ab initio* modeled fragments to put together complete structure of proteins. The Bhageerath-H methodology has been validated on 115 CASP 9 targets. For each of these cases, structures were predicted excluding native and close native homologs as templates. In each case, a structure within 7Å rmsd from native has been obtained. In this project, the modeled protein structures which have been used for protein structure refinement, are obtained using the Bhageerath-H software suite. Thus, it serves as the starting point for the project.

Other notable softwares for TBM which participate in CASP are provided in Table 6.

Software	**Method**	**URL**
3D-JIGSAW (version 3.0) [56]	Comparative modeling	http://bmm.cancerresearchuk.org/~populus/
HHPred [93]	Comparative modeling	http://toolkit.tuebingen.mpg.de/hhpred
Phyre2[94]	*De novo* and template based	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
Pcons.net[95]	Meta Server	http://pcons.net/
Robetta [96]	*Ab initio* and comparative modeling	http://robetta.bakerlab.org/
MUFOLD [97]	Hybrid	http://gene.rnet.missouri.edu/dbselect/prediction.php
YASARA [98]	Comparative modeling	http://www.yasara.org/homologymodeling.htm
RAPTOR [99]	Threading	http://www.bioinformaticssolutions.com/raptor-overview

SWISS-MODEL [55]	Comparative modeling	http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1
I-TASSER [100,70,99]	Threading	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
QUARK [101]	*Ab initio*	http://zhanglab.ccmb.med.umich.edu/QUARK/
Modeller [101]	Comparative Modeling	http://salilab.org/modeller/
TASSER [102, 103]	Meta Server	http://cssb.biology.gatech.edu/skolnick/webservice/TASSER/index.html

Table 6. Table listing various softwares for TBM, participating in CASP.

Chapter 2

OBJECTIVES

Protein structure refinement is a very challenging problem and we wish to tackle it by developing a novel protocol which helps in refining the tertiary structure of proteins predicted by the Bhageerath-H software.

The main objectives are to obtain clash free structures with the side chains and loops correctly modeled, making the structure feasible in the biomolecular environment and therefore achieving protein structure refinement.

For accomplishing the aforementioned goals, the employment of *ab initio* loop optimaization and molecular dynamics simulations is considered.

Chapter 3

MATERIALS AND METHODS

3.1. *Ab initio* modeling for loop optimization

In the process of protein structure refinement, *ab initio* modeling was carried out on the longer loops, end loops and missing secondary structural regions using the Bhageerath Software Suite. The *ab initio* protocol helped in conformational sampling of these problematic regions in the modeled protein structures which are responsible for the drift of the model from the native.

Bhageerath Software Suite

Bhageerath [87-91] is an energy based software suite for narrowing down the search space of tertiary structures of small globular proteins. It is a web-enabled tool and is freely accessible at http://www.scfbio-iitd.res.in/bhageerath.40 . Bhageerath protocol comprises of eight different computational modules that form an automated pipeline as represented in Figure 6.

- The amino acid sequence is taken as an input. The software first predicts the secondary structure (helix/strand/loop) of the input amino acid sequence.
- In the second module, an atomic-level extended structure using the secondary structure information is created.
- In the third module, Bhageerath does a systematic sampling of the conformational space of loop dihedrals and generates a large number of trial structures. The number of trial structures generated is $128^{(n-1)}$ where 'n' is the number of secondary structural elements and '$n-1$' is the number of loops/junctions between the secondary structural units.
- These structures are generated by choosing seven dihedrals from each of the loops (three at both ends and one dihedral from the middle of the loop) and sampling two conformational states for each dihedral.

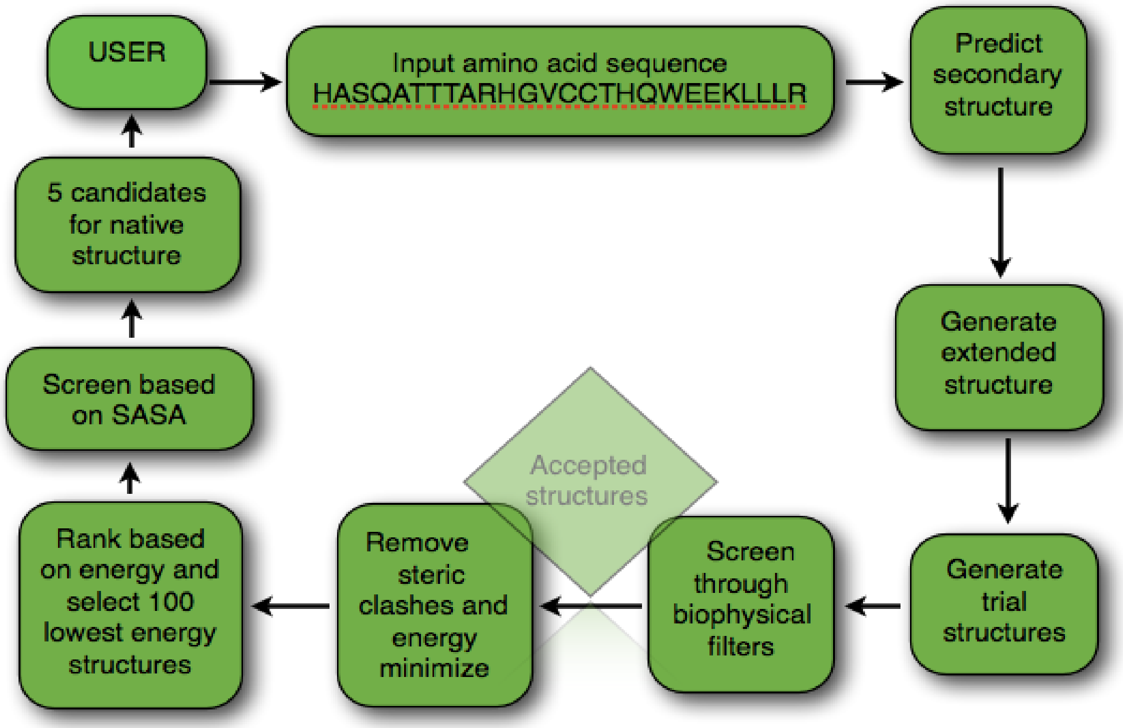
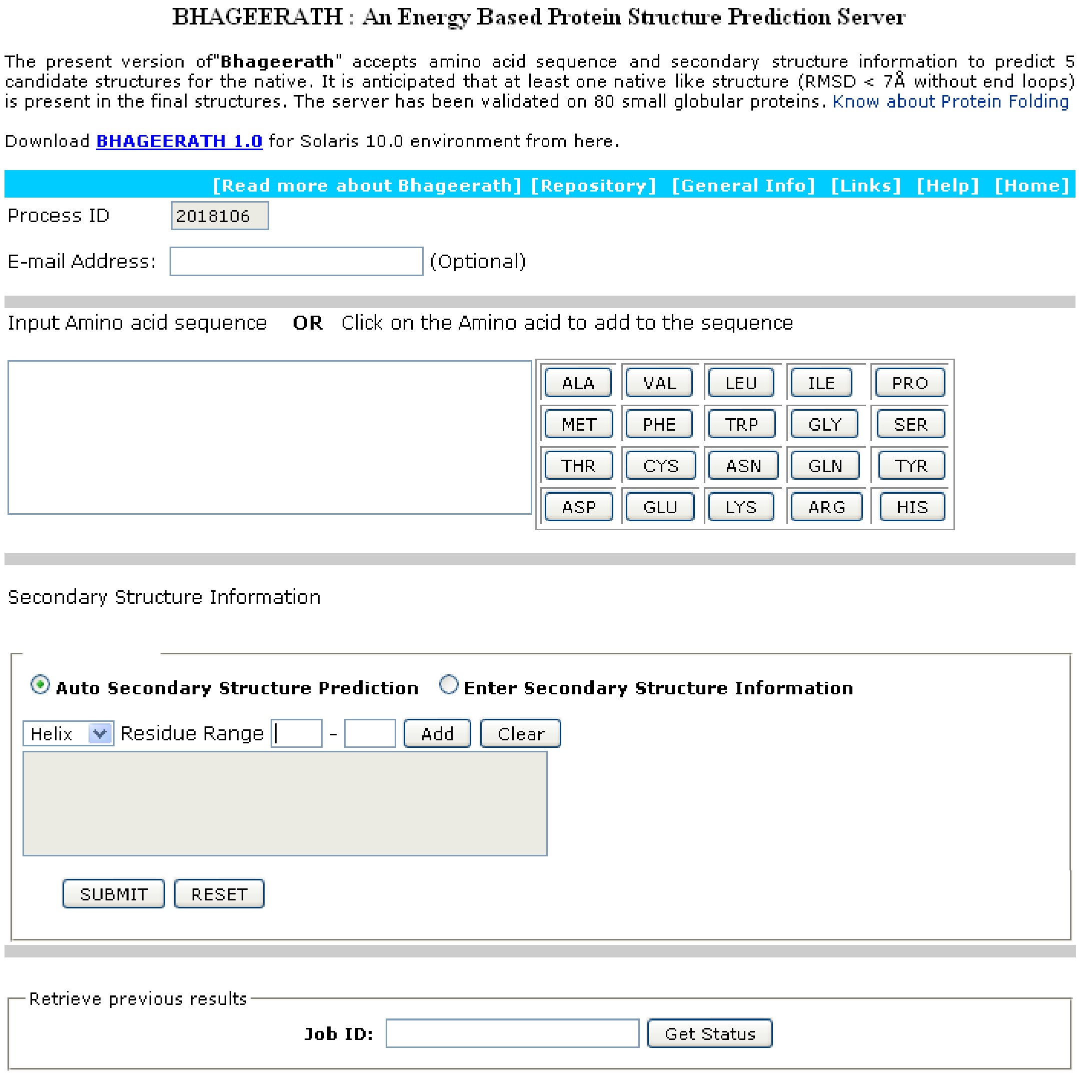

USER
Input amino acid sequence
HASQATTTARHGVCCTHQWEEKLLLR
Predict secondary structure
Generate extended structure
Generate trial structures
Screen through biophysical filters
Accepted structures
Remove steric clashes and energy minimize
Rank based on energy and select 100 lowest energy structures
Screen based on SASA
5 candidates for native structure

Figure 6. Workflow of the protocol followed by Bhageerath software suite (adapted from http://nar.oxfordjournals.org/content/34/21/6195.full.pdf)

- The generated trial structures are screened in the fourth module through persistence length, radius of gyration, topological distinctness of generated structures, inter-atomic distance and Cα loop distance filters [88], developed for the purpose of reducing the number of improbable candidates.
- The resultant structures are refined by the fifth module by a Monte Carlo sampling in dihedral space to remove steric clashes and overlaps involving atoms of main chain and side chains.
- In module six, energy of the structures is minimized, for further optimization of the side chains.
- In module seven, ranking of the structures using all atom energy based empirical scoring function [104] and subsequent selection of the 100 lowest energy structures is done.

- In module eight, structures selected in the previous module are reduced to five using solvent accessible surface areas (SASA) [105].

The software gives five candidate structures for the input amino acid sequence as the output. Figure 7 shows the screenshots of the Bhageerath software suite.



BHAGEERATH : An Energy Based Protein Structure Prediction Server

The present version of"**Bhageerath**" accepts amino acid sequence and secondary structure information to predict 5 candidate structures for the native. It is anticipated that at least one native like structure (RMSD < 7Å without end loops) is present in the final structures. The server has been validated on 80 small globular proteins. Know about Protein Folding

Download **BHAGEERATH 1.0** for Solaris 10.0 environment from here.

Process ID 2018106

E-mail Address: (Optional)

Input Amino acid sequence **OR** Click on the Amino acid to add to the sequence

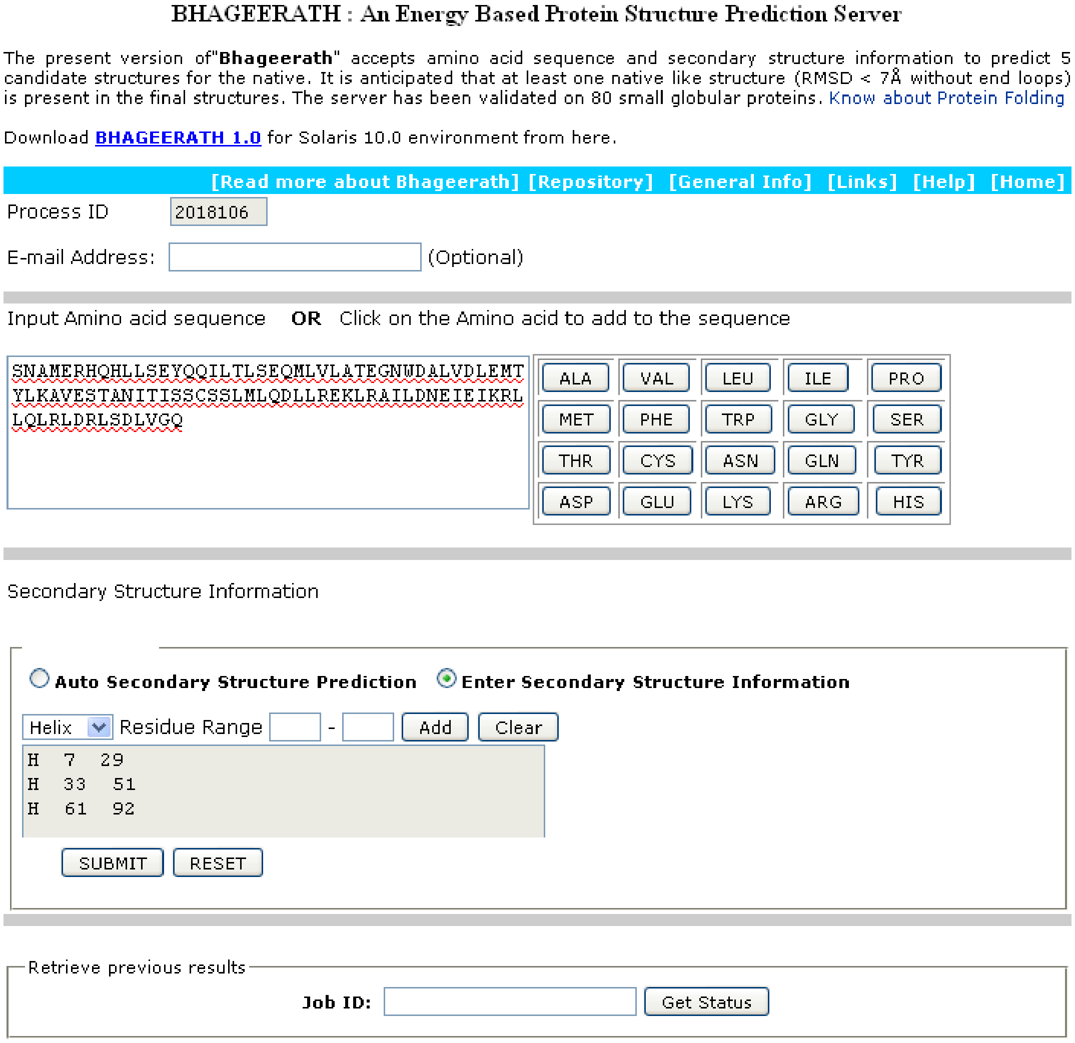
Secondary Structure Information

● **Auto Secondary Structure Prediction** ○ **Enter Secondary Structure Information**

Helix Residue Range -

Retrieve previous results

Job ID:



BHAGEERATH : An Energy Based Protein Structure Prediction Server

The present version of"**Bhageerath**" accepts amino acid sequence and secondary structure information to predict 5 candidate structures for the native. It is anticipated that at least one native like structure (RMSD < 7Å without end loops) is present in the final structures. The server has been validated on 80 small globular proteins. Know about Protein Folding

Download **BHAGEERATH 1.0** for Solaris 10.0 environment from here.

[Read more about Bhageerath] [Repository] [General Info] [Links] [Help] [Home]

Process ID 2018106

E-mail Address: (Optional)

Input Amino acid sequence **OR** Click on the Amino acid to add to the sequence

```
SNAMERHQHLLSEYQQILTLSEQMLVLATEGNWDALVDLEMT
YLKAVESTANITISSCSSLMLQDLLREKLRAILDNEIEIKRL
LQLRLDRLSDLVGQ
```

Secondary Structure Information

○ **Auto Secondary Structure Prediction** ● **Enter Secondary Structure Information**

Helix Residue Range -

```
H  7   29
H  33   51
H  61   92
```

Retrieve previous results

Job ID:

Figure 7. Screenshots of Bhageerath (URL: http://www.scfbio-iitd.res.in/bhageerath/index.jsp)

Bhageerath methodology was used for the *ab initio* sampling of the loop regions of the modeled structures. The protocol was used for refinement of the longer loops, end loops and missing secondary structural regions. For our analysis, we took the CASP9 dataset. The CASP9 dataset comprises of the target systems which were used in the CASP9 competition which took place in the year 2010. A CASP9 like scenario was created where in the structures for each of the target systems were modeled using proteins which were available before the start of CASP9 competition. This was primarily carried out to exclude the natives and any other structures which bearing homology with the target. This ensured that the structures modeled using Bhageerath-H (which were taken as an input for our *ab initio* protocol) were unbiased.

The CASP9 dataset consists of a total of 115 targets out of which we have considered 60 target proteins of varying length and complexity for our analysis. The detailed methodology for *ab initio* modeling using Bhageerath software suite has been described below.

1. 60 proteins from the CASP9 dataset were considered for structure refinement so as to derive structures with better RMSD out of the 115 target dataset.
2. The models for these 60 CASP9 targets were obtained using Bhageerath-H software suite.
3. For each of the modeled protein structure, secondary structure information was obtained using Stride [106]. Stride is a program that extracts the secondary structural elements in proteins from their atomic coordinates based on the secondary structural definitions given by X-ray crystallography and protein NMR methods.
4. The problematic regions of the proteins, which form the basis for deviation of the modeled structure from that of the native, were identified by manual visualization using PyMOL.
5. These regions were then classified into four categories-
    a. Longer loops: Loops with length of more than 5 residues were considered as long loops.
    b. End loops: Loops present either at the beginning of the modeled protein structure or at the end, containing more than 2 residues, were classified as end loops.

c. Missed Strand: Comparison of the secondary structures for protein models predicted by Bhageerath-H (using PSIPRED* on back end) and Stride* was done. The regions which were predicted to be strands by Bhageerath-H alone or Stride alone, showing mismatch between their predictions were classified as missed strands. Strands missed in the 3D modeled protein structure were considered as long loops and were modeled accordingly.

d. Missed Helix: Comparison of the secondary structures for protein models predicted by Bhageerath-H (using PSIPRED* on back end) and Stride* was done. The regions which were predicted to be helices by Bhageerath-H alone or Stride alone, showing mismatch between their predictions were classified into the missed helix category. A missed helix in the modeled structure, with length more than 4 residues, was forced to form a helix using a helix formation algorithm.

   Contrastingly, forcing the formation of helix might have an adverse effect on the structure. So, in another step, the missing helices were considered as loops and modeled as long loops.

6. Each of these loop regions were modeled using Bhageerath methodology. For each loop 128 structures were generated although, for very long loops (>10 residues) the number of structures generated for each loop were increased from 128 to 1024 (2^{10}).

*[NOTE: PSIPRED – It is a protein secondary structure prediction server which predicts the protein secondary structures from the input amino acid sequence. STRIDE – It is a secondary structure assignment program which gives the protein secondary structure information from the input PDB data]

For a final refinement of the structures obtained, a simulation protocol was developed using Molecular Dynamics. This step is extremely important as the 3D modeled structures of proteins often contain numerous clashes in the backbone as well as side chain atoms. In order to refine the structure and model a naturally feasible structure, we need to get rid of such clashes that make the modeled protein structure highly unstable.

MD simulations have provided extremely high resolution spatial and temporal data, enhancing knowledge and understanding of the protein folding mechanism. Now-a-days, simulations of biologically relevant processes, with atomistic accuracy on timescales beyond microsecond are possible due to advances in software and hardware. In this project, we have exploited ability of compute power for carrying out computationally expensive MD simulations using the AMBER software package (version 10).

3.2. Assisted Model Building with Energy Refinement (AMBER) for MD simulations

AMBER is a family of force fields for molecular dynamics of biomolecules originally developed by the late Peter Kollman's group at the University of California, San Francisco. AMBER is also the name for the molecular dynamics software package that simulates these force fields.

The functional form of AMBER force field is:

$$V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$
$$+ \sum_{\text{torsions}} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{i,j} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

- The first term represents the energy between covalently bonded atoms or the bond stretch energy.
- The second term represents the energy due to the geometry of electron orbitals involved in covalent bonding or the bond angle energy.
- The third term represents the energy for twisting a bond due to bond order (e.g. double bonds) and neighboring bonds or lone pairs of electrons or the torsional energy.
- The fourth term represents the non-bonded energy between all atom pairs, which can be decomposed into Van der Waals and electrostatic energies respectively.

The AMBER software suite consists of a number of programs out of which the following were of particular importance to this project:

1. **LEaP:** LEaP is a program that provides for basic model building and Amber coordinate and parameter/topology input file creation.

2. **SANDER:** Simulated Annealing with NMR-Derived Energy Restraints

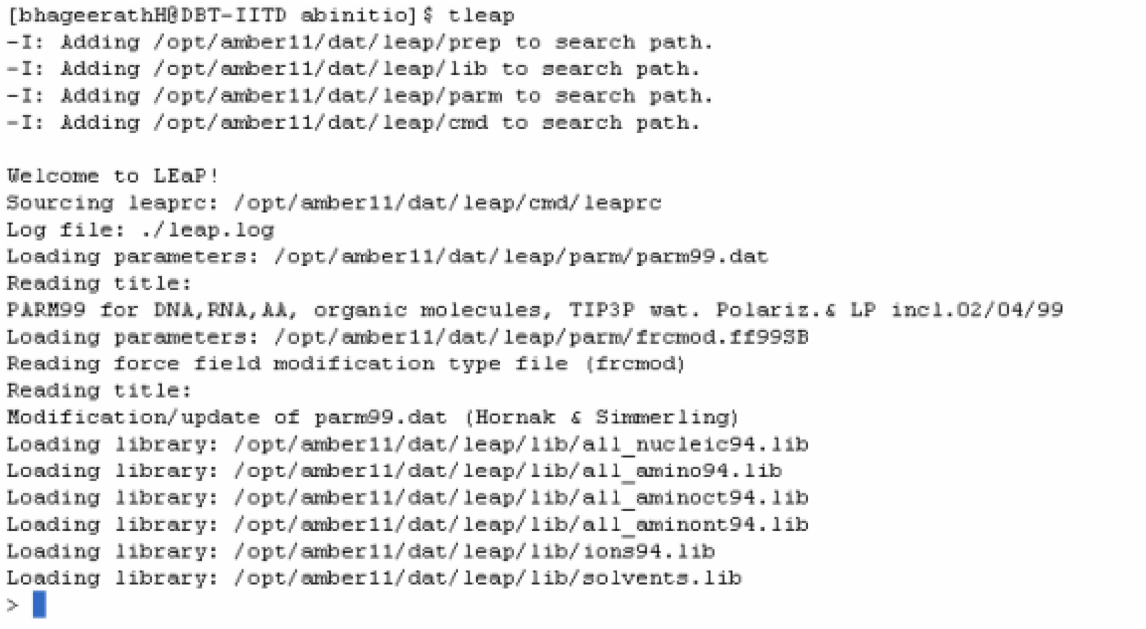
   This program allows for NMR refinement based on NOE-derived distance restraints, torsion angle restraints, and penalty functions based on chemical shifts and NOESY volumes.

3. **pmemd**: This program is an extensively-modified version of the sander program. It was designed with parallel processing in mind and has significantly better performance than sander when running on more than 8–16 processors.

Molecular dynamics simulations have been used for the final refinement of the modeled protein structures. The entire protocol has been described below.

1. **Generation of input files:**

The coordinate and topology files of the modeled protein structure which are to be used as input files for the downstream MD simulations are generated using tleap as shown in Figure 8.



```
[bhageerathH@DBT-IITD abinitio]$ tleap
-I: Adding /opt/amber11/dat/leap/prep to search path.
-I: Adding /opt/amber11/dat/leap/lib to search path.
-I: Adding /opt/amber11/dat/leap/parm to search path.
-I: Adding /opt/amber11/dat/leap/cmd to search path.

Welcome to LEaP!
Sourcing leaprc: /opt/amber11/dat/leap/cmd/leaprc
Log file: ./leap.log
Loading parameters: /opt/amber11/dat/leap/parm/parm99.dat
Reading title:
PARM99 for DNA,RNA,AA, organic molecules, TIP3P wat. Polariz.& LP incl.02/04/99
Loading parameters: /opt/amber11/dat/leap/parm/frcmod.ff99SB
Reading force field modification type file (frcmod)
Reading title:
Modification/update of parm99.dat (Hornak & Simmerling)
Loading library: /opt/amber11/dat/leap/lib/all_nucleic94.lib
Loading library: /opt/amber11/dat/leap/lib/all_amino94.lib
Loading library: /opt/amber11/dat/leap/lib/all_aminoct94.lib
Loading library: /opt/amber11/dat/leap/lib/all_aminont94.lib
Loading library: /opt/amber11/dat/leap/lib/ions94.lib
Loading library: /opt/amber11/dat/leap/lib/solvents.lib
> 
```

Figure 8. Screenshot showing tleap startup window

prmtop: This is the parameter/topology file. This defines the connectivity and parameters for our current model. This information does not change during the entire simulation as the topology of the model remains the same.

inpcrd: This is the coordinates file. This data is not static and changes during the simulations as the coordinates change after each step.

2. Neutralization of structure:

This step involves neutralization of the structure explicitly by calculating the charge and subsequently, adding the requisite ions.

3. Solvation in water:

This step involves solvation of the protein in water as shown in Figure 9. Solvation is done with the command "solvateoct" where an octahedral water box is created. For protein molecules, we used 8Å buffer of TIP3P water around the protein. The TIP3P water model is the simplest water model that treats the water molecule as rigid and relies only on non-bonded interactions. The electrostatic interaction is modeled using Coulomb's law and the dispersion and repulsion forces using the Lennard-Jones potential.

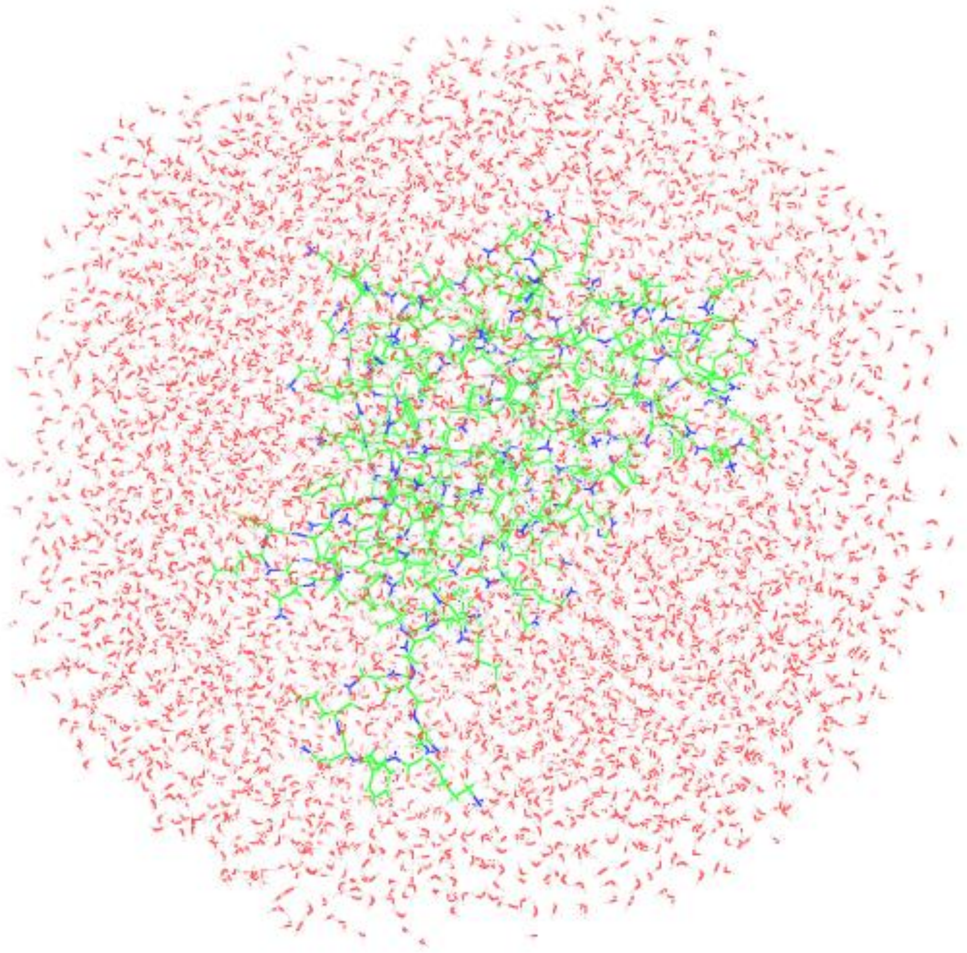


Figure 9. Screenshot of protein molecule surrounded by water molecules

4. Running Minimization and MD in explicit solvent

Molecular dynamics in explicit solvent is carried out to mimic the real world situation in which the protein molecules are surrounded by water molecules. Running MD in vacuum or in implicit solvent would not be able to accurately represent the biomolecular systems which exist in a solvated environment. Despite the computationally expensive nature of the explicit solvent molecular dynamics, it helps in mimicking the real world situation appropriately.

Our minimization procedure for solvated protein consists of a two stage approach. In the first stage we keep the protein fixed and just minimize the positions of the water and ions. Then in the second stage we minimize the entire system. This two stage approach works well when the modeled protein structure is far from the equilibrium, and requires multiple steps of minimizations.

Step1: Minimization 1:

Positional restraints on each of the protein atoms to keep them essentially fixed in the same position were applied. Such restraints work by specifying a reference structure which is the starting structure, thus, keeping the entire protein (solute) fixed using restraint (RES) command. A total of 1000 minimization steps were conducted. We have used two different algorithms to carry out the minimization. For the first 500 steps, steepest descent algorithm was used which is good for quickly removing the largest strains in the system but converges slowly when close to a minima. Thus, for the next 500 steps, the conjugate gradient method was used which is more efficient.

Step2: Minimization 2:

Now we have minimized the water and ions the next stage of our minimization is to minimize the entire system (modeled protein structure along with the water molecules). In this case we run 2,500 steps of minimization without any restraints. The first 500 steps were done using the steepest descent algorithm and the conjugate gradient algorithm for the rest 2000 steps.

Step3: Molecular dynamics heating with restraints on solute:

After successful minimization of the system, the next stage in the equilibration protocol is to allow the system to heat up from 0K to 300K. At this stage, we restrained the entire protein as otherwise the heating would unfold the whole protein. A short MD was run for 20ps.

Step4: Molecular Dynamics Equilibration over entire system:

The equilibration phase is the final step of our molecular dynamics protocol. The temperature is maintained at 300K and the simulation is run for 100ps, 200ps, 500ps, 1ns and 5ns successively.

Cα-Cα Secondary Structural Restraints

To avoid disruption of the modeled protein structure, we restrained the secondary structures in order to maintain the topology.

So, another set of simulations were carried out where the Cα-Cα distances between the secondary structures were calculated. The residues which fall within the 5Å cut-off were kept fixed in the backbone using NMR distance restraints at the molecular dynamics equilibration stage.

To obtain an input distance restraint file for imposing the NMR restraints, we need to put the restraints in a 7-column file. This file contains 7 columns with the following information for each restraint:

1st_res# 1st_res_name 1st_atom_name 2nd_res# 2nd_res_name 2nd_atom_name upper_bound

The upper bound distance for each restraint is taken as the actual Cα-Cα distance (which falls within the cut-off for example, 5Å in this case) in the modeled protein structure. The program used to convert the 7-column file into the AMBER restraint file is makeDIST_RST.

```
makeDIST_RST -upb <7 column distance file> -pdb <PDB structure> -rst RST.dist
```

The RST.dist is the final restraint file which is read by AMBER during the molecular dynamics equilibration phase as an argument in the MD input file. It helps in restraining the backbone of the modeled protein structure, so that side chain refinement can take place. Similarly, restraints were also imposed with 7Å and 10Å distance cut-offs respectively, in order to restrain the structure more aggressively during the simulations.

Chapter 4

RESULTS AND DISCUSSION

Protein structure refinement problem has proven to be a major bottleneck to further improvements in structure prediction. Researchers for long have tried solving this problem by using techniques that involve either optimization of new potential functions or inclusion of contact restraints from homologous proteins. We decided to test whether *ab initio* modeling of the problematic regions (mainly loops) could help improve the structure and bring it closer to the native. On the other hand, we also tried to improve the structures using molecular dynamics simulations. We were, in part, successful in our attempts.

The *ab initio* modeling protocol was applied to the set of models obtained from Bhageerath-H software suite for 60 CASP9 targets. We observed significant improvements in the RMSD of the structures from their natives in the systems T0516, T0531, T0538, T0544, T0558, T0559, T0564, T0571, T0576, T0581, T0600, T0603, T0606, T0612, T0616, T0618, T0622, T0625, T0635 and T0643. Table 7 summarizes the comparison of the RMSD from the native between the Bhageerath-H predicted models before *ab initio* modeling and that after *ab initio* modeling on the respective problematic regions, where the systems showing substantial improvements have been highlighted. System T0600, in particular, has shown notable improvement after application of the loop optimization protocol. Bhageerath-H along with all the other leading servers (Zhang et al, Baker et al), were only able to model structures with approximately 19.4Å RMSD from the native. On the other hand after applying the *ab initio* modeling protocol for loop optimization, we were able to build a structure with 4.73Å RMSD from the native. Upon tinkering of the loop region present in the Bhageerath-H modeled structure for T0600, the RMSD shows considerable improvement. Due to wrong direction of this problematic loop region, the rest of the structure was modeled in the exact opposite direction as can be easily seen from Figure 10.

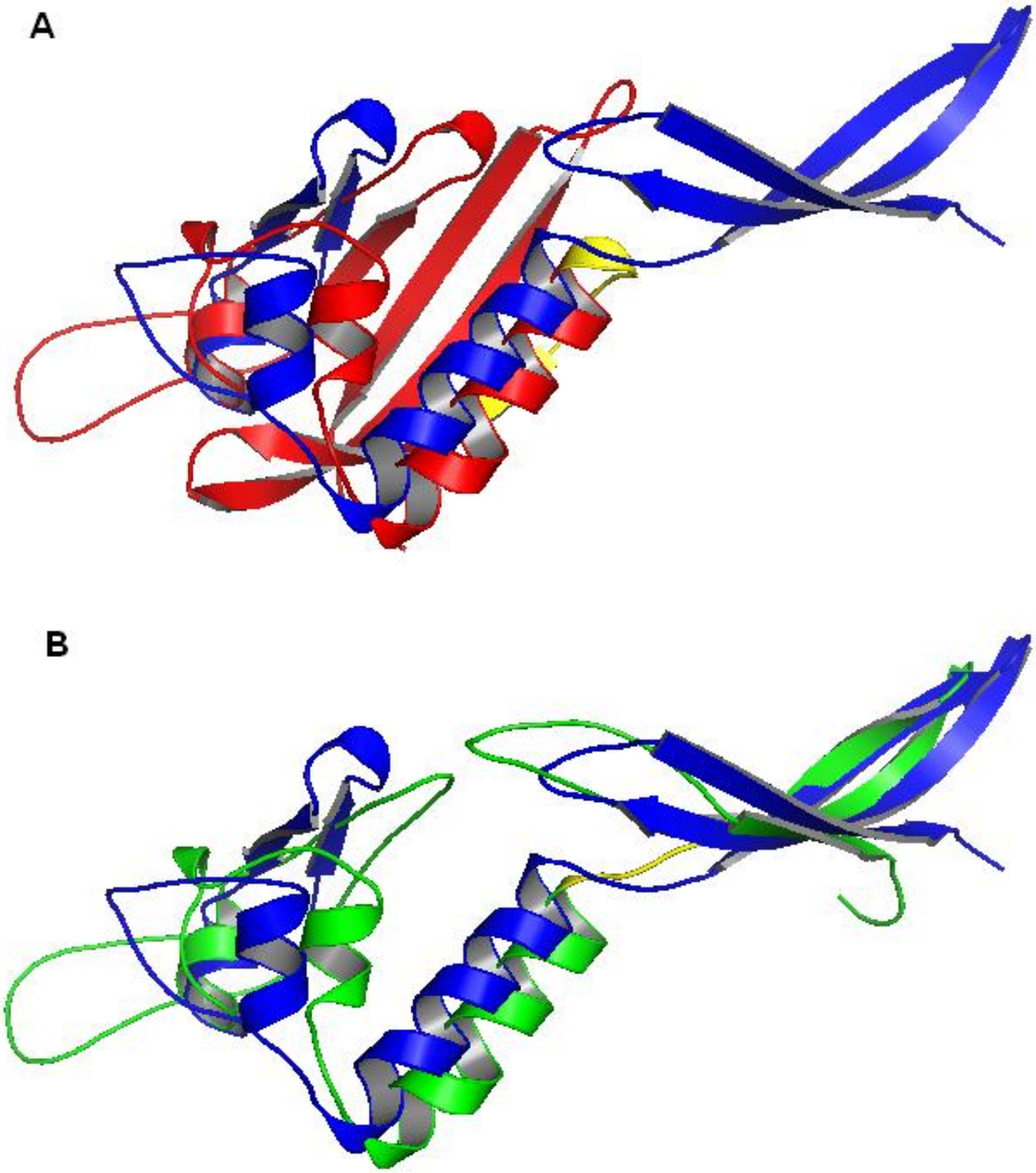


Figure 10. Colour scheme: Blue-Native T0600, Red-Bhageerath-H model, Green-Structure after ab inito modeling on loop (residues 75-82), Yellow-Loop on which *ab initio* is done.

A. Bhageerath-H model for T0600 native superimposed on the native. **B**. Model obtained after *ab initio* modeling on the loop region superimposed on the native.

Therefore, this result clearly demonstrates the high applicability of the *ab initio* modeling protocol in protein structure refinement. In order to avoid the loss the good starting structures (in cases where the RMSD from the native increases after *ab initio* modeling), the initial starting structure was retained.

An important observation is that the *ab initio* modeling protocol has been helpful in reducing the RMSD in cases where the models are far from the native whereas in other cases, where the RSMD is 5Å or lower from the native, deriving a lower resolution structure to a further low resolution becomes difficult.

After the entire structure has been modeled, it is extremely important to carry out final refinement of the structures so as to remove the steric clashes, optimization of side-chains, bond angles, bond lengths and reduce the Ramachandran outliers. We do not expect short MD simulations (upto 1-5 ns) to help reduce the RMSD from the native, but we do expect to obtain a structure with minimal clashes. We have carried out molecular dynamics simulations restraining residues for which the Cα-Cα distances of the secondary structure elements are within 5Å, 7Å and 10Å, during the equilibration phase. The restrained simulations have been implemented in order to prevent MD to open up the correctly folded regions and therefore, maintain the topology of the modeled protein structure. Thus, our motive of using MD simulations on the modeled protein structures is to retain their basic topology while reducing the clashes in the structure, making it physically plausible. The MD simulations were run for 100ps, 200ps, 500ps, 1ns and, for some cases, 5ns.

The structures then obtained were subsequently validated using Molprobity [107-108], a protein structure validation software, the results of which have been summarized in Table 8. After the

MD simulations, the structures are free of clashes, with lesser Ramachandran outliers, bad angles and bad bonds and greater Ramachandran favored residues. Thus, we were able to generate geometrically optimized structures using MD simulations.

Chapter 5

CONCLUSION AND FUTURE DIRECTIONS

Homology or template based modeling has been the most successful method for protein structure prediction in the critical assessment of protein structure prediction (CASP) experiments. The power of this technique progressively increases as more and more structures are solved by world-wide structural genomics initiatives. Although efforts in this field are promising, obtaining a model with the same accuracy as a crystal structure is still an unsolved problem. Thus, the structure refinement of a rough model, to bring it closer to the native structure remains a major challenge. Work on structure refinement has been ongoing for many decades, including those using Molecular Mechanics (MM) energy minimization, knowledge-based (KB) statistically derived potentials among others. During this period many different potentials and a variety of simulation methodologies such as energy minimization, molecular dynamics, and replica exchange Monte Carlo have been used for structure refinement, but no method has emerged as a clear winner.

In this project, two methodologies have been developed to tackle the protein structure refinement problem. The *ab initio* modeling using Bhageerath software suite on the regions of the protein structures that are modeled improperly (end loops, longer loops and missing secondary structural regions). This protocol helped decrease the RMSD of the predicted structures with that of their corresponding native structures, thereby bringing them closer to their natives. It has been demonstrated that in certain cases it has benefited the task tremendously, as in the case of system T0600. Secondly, restrained molecular dynamics simulations were applied on the proteins so as to improve the overall quality of the modeled structures. Restraints with different distance cut-offs (5Å, 7Å and 10Å) were used on the proteins in the production phase and encouraging results were obtained.

Protein structure refinement is a major problem that is holding back the performance of a number of structure prediction tools. Although a lot still has to be explored in this field, here we have presented here, two protocols for the same cause. We have been successful in part in our efforts. Further on, it is suggested that the molecular dynamics simulations should be run for longer durations so that better configurations (with lower RMSD) of the proteins can be generated.

REFERENCES

1. Dayalan S, Gooneratne N.D, Bevinakoppa S, Schroder. Dihedral angle and secondary structure database of short amino acid fragments. Bioinformation 2006; 1:78-80.

2. Creighton T.E. Proteins: Structures and Molecular Properties (Freeman W.H & Company) pp. 3, New York, 1983.

3. Kaur H, Raghava G.P.S. Prediction of α-Turns in Proteins Using PSI-BLAST Profiles and Secondary Structure Information. Proteins: Struct. Funct. Bioinfo. 2003; 55: 83-90.

4. Creighton T.E. Proteins: Structures and Molecular Properties (2nd edit) (Freeman W.H & Company), New York, 1996.

5. Monnigmann M, Floudas C.A. The role of flexible stem geometries in Protein loop structure prediction. NIC workshop 2006; 34: 115-116.

6. Rosenbach D, Rosenfeld R.F. Simultaneous modeling of multiple loops in proteins. Protein Sci. 1995; 4: 496-505.

7. Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000; 405:837-46.

8. O’Farrell P.H. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 1975; 250: 4007–21.

9. Fields S. Proteomics. Proteomics in genomeland. Science 2001; 291:1221-4.

10. Burley S.K, Almo S.C, Bonanno J.B, Capel M, Chance M.R, Gaasterland T, Lin D, Sali A, Studier F.W, Swaminathan S. Structural Genomics: beyond the Human Genome Project. Nat. Genet. 1999; 23:151-57.

11. Venter J.C et. al. The Sequence of the Human Genome. Science 2001; 291:1304-51.

12. Stoesser G, Sterk P, Tuli M.A, Stoehr P.J, Cameron G.N. The EMBL nucleotide sequence database. Nucl. Acids Res. 1997; 25:7–13.

13. McEntyre J.O, Ostell J. The NCBI handbook National Library of Medicine (US), National Center for

Biotechnology Information. Internet: Bethesda (MD) National Center for Biotechnology Information US 2002. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

14. Pruitt K.D, Tatusova T, Klimke W, Maglott D.R. NCBI Reference Sequence: current status, policy and new initiatives. Nucl. Acids Res. 2009; v.39:D52-57.

15. Berman H.M, Westbrook J, Feng Z, Gilliland G, Bhat T.N, Weissig H, Shindyalov I.N, Bourne P.E. The Protein Data Bank. Nucl. Acids Res. 2000; 28:235-42. www.pdb.org

16. Boeckmann B, Bairoch A, Apweiler R, Blatter M.C, Estreicher A, Gasteiger E, Martin M.J, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL. Nucl. Acids Res. 2003; 31:365-70.

17. Browne W.J, North A.C, Phillips D.C, Brew K, Vanaman T.C, Hill R.L. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. 1969; 42:65-86.

18. Chandonia J.M, Brenner S.E. The impact of structural genomics: expectations and outcomes. Science 2006; 311:347-51.

19. Lee D, de Beer T.A, Laskowski R.A, Thornton J.M, Orengo C.A. 1,000 structures and more from MCSG. BMC Struct. Biol. 2011; 11:2.

20. Moult J, Pedersen J.T, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995; 23:II–IV.

21. Murzin A.G, Brenner S.E, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995; 247:536-40.

22. Andreeva A, Howorth D, Chandonia J.M, Brenner S.E, Hubbard T.J, Chothia C, Murzin A.G. Data growth and its impact on the SCOP database: new developments. Nucl. Acids Res. 2008; 36:D419-25.

23. Zhang Y. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 2008; 18:342-8.

24. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein

structure prediction - Round VIII. Proteins 2009; 77:1-4.

25. Skolnick J. Protein Structure Prediction. Ency. of Life Sci. 2007; 1-7.

26. Murzin A.G. Progress in protein structure prediction. Nat. Struct. Biol. 2001; 8: 110-12.

27. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. 2004; 101:7594-99.

28. Petrey D, Honig B. Protein structure prediction: Inroads to Biology. Mol. Cell 2005; 20:811-9.

29. McGuffin L.J. In: Schwede T, Peitsch M.C, Ed. Protein fold recognition and threading. World Scientific Co. 2008; 37-60.

30. Anfinsen C.B. Principles that govern the folding of protein chains. Science 1973; 181:223-30.

31. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991; 9:56-68.

32. Needleman S.B, Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970; 48:443–453.

33. Smith T.F, Waterman M.S. Identification of common molecular subsequences. J. Mol. Biol. 1981; 147:195–197.

34. Gotoh O. An improved algorithm for matching biological sequences. J. Mol. Biol. 1982; 162:705–708.

35. Taylor W.R. Identification of protein sequence homology by consensus template alignment. J. Mol. Biol. 1986; 188:233–258.

36. Chappey C, Danckaert A, Dessen P, Hazout S. MASH: an interactive program for multiple alignment and consensus sequence construction for biological sequences. Comput. Appl. Biosci. 1991; 2:195-202.

37. Suyama M, Matsuo Y, Nishikawa K. Comparison of protein structures using 3D profile alignment. J. Mol. Evol. 1997; 44:S163–173.

38. Lolkema J.S, Slotboom D.J. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. Mol. Membr. Biol. 1998; 15:33–42.

39. Barton G.J, Sternberg M.J. Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. J. Mol. Biol. 1990; 212:389-402.

40. Altschul S, Madden T.L, Schaffer A.A, Zhang J, Zhang Z, Miller W, Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 1997; 25:3389-3402.

41. Krogh A, Brown M, Mian I, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol.1994; 235:1501-1531.

42. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998; 14:846–856.

43. Altschul S.F, Madden T.L, Schäffer A.A, Zhang J, Zhang Z, Miller W, Lipman D.J. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucl. Acids Res. 1997; 25:3389-402.

44. Greer J. Model for haptoglobin heavy chain based upon structural homology. Proc. Natl. Acad. Sci. 1980; 77:3393-3397.

45. Greer J. Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 1981; 153:1027.

46. Sutcliffe M.J, Haneef I, Carney D, Blundell T.L. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng. 1987; 1:377-384.

47. Sutcliffe M.J, Hayes F.R, Blundell T.L. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. Protein Eng. 1987; 1:385-392.

48. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 1992; 226:507-533.

49. Sali A, Blundell T.L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 1993; 234:779-815.

50. Martí-Renom M.A, Stuart A.C, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 2000; 29:291-325.

51. Greer J. Model for haptoglobin heavy chain based upon structural homology. PROC Natl. Acad. Sci. 1980; 77:3393-97.

52. Srinivasan N, Blundell T.L. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein Eng. 1993; 6:501-12.

53. Lee R.H. Protein model building using structural homology. Nature 1992; 356:543-44.

54. Greer J. Comparative model-building of the mammalian serine proteases.J. Mol. Biol. 1981; 153:1027-42.

55. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. Bioinformatics 2006; 22: 195-201.

56. Bates P.A, Kelley L.A, MacCallum R.M, Sternberg M.J. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2002; Suppl 5:39-46.

57. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 1992; 226:507-33.

58. Jones T.A, Thirup S. Using known substructures in protein model building and crystallography. EMBO J. 1986; 5:819-822.

59. Dunbrack R.L Jr. Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. 2002; 12:431-40.

60. Dunbrack R.L Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. 1993; 230:543-74.

61. Ponder J.W, Richards F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 1987; 193:775-91.

62. Lee M.R, Baker D, Kollman P.A. 2.1 and 1.8 Å Average Cα RMSD Structure Prediction on Two Small Proteins, HP-36 and S15. J. Am. Chem. Soc. 2001; 123:1040-46.

63. Lu H, Skolnick J. Application of statistical potentials to protein structure refinement from low resolution *ab initio* models. Biopolymers 2003; 70:575-84.

64. Flohil J.A, Vriend G, Berendsen H.J. Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. Proteins 2002; 48:593-604.

65. Krieger E, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA-a self-parameterizing force field. Proteins 2002; 47:393-402.

66. Chen J, Brooks C.L 3rd. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 2007; 67:922-30.

67. Im W, Lee M.S, Brooks C.L 3rd. Generalized born model with a simple smoothing function. J. Comput. Chem. 2003; 24:1691-702.

68. Chen J, Im W, Brooks C.L 3rd. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. J. Am. Chem. Soc. 2006; 128:3728-36.

69. Zhu J, Fan H, Periole X, Honig B, Mark A.E. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins 2008; 72:1171-88.

70. Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based *ab initio* folding and FG-MD-based structure refinement. Proteins: Structure, Function and Bioinformatics 2011; 79:147–160.

71. Wiederstein M, Sippl M.J. ProSA-web: interactive web service for the recognition of errors in three dimensional structures of proteins. Nucl. Acids Res. 2007; 35:W407-10.

72. Sippl M.J. Recognition of errors in three-dimensional structures of proteins. Proteins 1993; 17:355-62.

73. Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins 2008; 71:1211-18.

74. Paluszewski M, Karplus K. Model quality assessment using distance constraints from alignments. Proteins 2009; 75:540-9.

75. Narang P, Bhushan K, Bose S, Jayaram B. Protein structure evaluation using an all-atom energy based empirical scoring function. J. Biomol. Str. Dyn. 2006; 23:385-406.

76. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004; 57:702-10.

77. Xu J, Zhang Y. How significant is protein structure similarity with TM-score=0.5. Bioinformatics 2010; 26:889-895.

78. Benkert P, Tosatto S.C, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008; 71:261-77.

79. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2011; 27:343-50.

80. Benkert P, Schwede T, Tosatto S.C. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct. Biol. 2009; 9:35.

81. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucl. Acids Res. 2009; 37:W510-4.

82. Lüthy R, Bowie J.U, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992; 356:83-5.

83. Bowie J.U, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three dimensional structure. Science 1991; 253:164-170.

84. Vriend G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 1990; 8:52-6.

85. Laskowski R.A, MacArthur M.W., Moss D.S, Thornton J.M. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 1993; 26:283-91.

86. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins. 2007; 69:175-83.

87. Jayaram B, Bhushan K, Shenoy S.R, Narang P, Bose S, Agarwal P, Sahu D, Pandey V. Bhageerath : An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucl. Acids Res. 2006; 34:6195-6204.

88. Narang P, Bhushan K, Bose S, Jayaram B. A computational pathway for bracketing native-like structures fo small alpha helical globular proteins. Phys. Chem. Chem. Phys. 2005; 7:2364-75.

89. Thukral L, Shenoy S.R, Bhushan K, Jayaram B. ProRegIn: a regularity index for the selection of native-like tertiary structures of proteins. J. Biosci. 2007; 32:71-81.

90. Jayaram B, Dhingra P, Lakhani B, Shekhar S. Bhageerath - Targeting the Near Impossible: Pushing the Frontiers of Atomic Models for Protein Tertiary Structure Prediction. Journal of Chemical Sciences 2011.

91. Shenoy S.R, Jayaram B. Proteins: sequence to structure and function--current status. Curr. Protein Pept. Sci. 2010; 11:498-514.

92. Mohanty P., Lakhani B. and Jayram B. Bhageerath-H: a homology *ab initio* hybrid server for protein tertiary structure prediction. (manuscript in preparation)

93. Söding J, Biegert A, Lupas A.N. The HHpred interactive server for protein homology detection and structure prediction. Nucl. Acids Res. 2005; 33:W244-8.

94. Kelley L.A, Sternberg M.J. Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. 2009; 4:363-71.

95. Wallner B, Larsson P, Elofsson A. Pcons.net: protein structure prediction meta server. Nucl. Acids Res. 2007; 35:W369-73.

96. Kim D.E, Chivian D, Baker D. Protein structure prediction and analysis using Rosetta server. Nucl. Acids Res. 2004; 32:W526-31.

97. Kreiger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry and side-chain accuracy in homology modeling: four approaches that performed well in CASP9. Proteins 2009; 77:114-122.

98. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J. Bioinform. Comput. Biol. 2003; 1:95-117.

99. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008; 40:1-8.

100. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 2010; 5: 725-38.

101. Eswar N, Webb B, Marti-Renom M.A, Madhusudhan M.S, Eramian D, Shen M.Y, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci. 2007; Unit 2.9.

102. Zhou H, Skolnick J. *Ab initio* protein structure prediction using chunk-TASSER. Biophys. J. 2007; 93:1510-8.

103. Zhou H, Skolnick J. Protein structure prediction by pro-Sp3-TASSER. Biophys. J. 2009 ;96:2119-27.

104. Narang,P., Bhushan,K., Bose,S. and Jayaram,B. Protein structure evaluation using an all-atom energy based empirical scoring function. J. Biomol. Struct. Dyn. 2006; 23:385-406.

105. Hubbard S.J, Thornton J.M. ‘NACCESS’, Computer Program, Department of Biochemistry and Molecular Biology, University College London, UK, 1993.

106. Heinig M, Frishman D. STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. Nucl. Acids Res. 2004; 32, W500-2.

107. Chen et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr 2010; D66:12-21.

108. Davis et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucl. Acids Res. 2007; 35:W375-W383.

APPENDIX I

Table 7. Table showing the results of *ab initio* modeling

System	No. of res	Model name	RMSD (before ab_initio)	RMSD (after ab_initio)
T0515	345	TT5565.pdb	1.492	1.583
T0516	226	TT4413.pdb	2.961	2.434
T0517	159	model2822_2qs7_Profile0.pdb	6.404	7.395
T0518	256	model12781_3hbk_ffas2.pdb	8.597	8.633
T0520	173	model5140_1fx2_ffas6.pdb	2.871	2.809
T0522	132	model7505_3i4s_ffas2.pdb	0.817	0.817
T0523	112	model3263_3lyx_ffas2.pdb	2.182	2.868
T0524	322	model11092_1snz_Profile0.pdb	3.4	4.061
T0525	205	N23.pdb	3.459	3.427
T0526	290	model21846_1snz_ffas10.pdb	3.154	5.239
T0528	371	model8547_3n0w_ffas2.pdb	3.31	5.041
T0531	65	model577_2dhi_Profile25.pdb	9.005	7.249
T0532	469	model3985_3ejn_ffas2.pdb	3.406	3.79
T0538	54	model653_2c9o_Profile93.pdb	1.99	1.577
T0540	90	model10462_2kd2_ffas2.pdb	5.4	5.436
T0541	106	model791_3idu_DIR395.pdb	2.18	2.5
T0544	135	TT746.pdb	10.498	9.464
T0558	274	N449.pdb	6.222	5.555
T0559	67	model811_1ku9_Profile18.pdb	2.11	1.898
T0560	64	model510_2l01_DIR255.pdb	1.557	1.795
T0564	80	model1008_1wjj.pdb	8.255	7.659
T0565	287	model11380_3h41_ffas1.pdb	7.428	7.891
T0566	130	model4703_1usu_Profile0.pdb	2.283	3.578
T0567	135	model314_1ny6_DIR157.pdb	2.581	2.769
T0570	237	TT1671.pdb	2.932	2.922
T0571	315	TT813.pdb	28.308	21.248
T0576	133	TT2517.pdb	8.254	4.993
T0580	104	model4000_1iib_Profile0.pdb	1.958	1.958
T0581	112	TT4302.pdb	9.887	8.305
T0582	221	TT13202.pdb	5.79	5.68
T0594	140	model4756_1x53_Profile0.pdb	2.545	5.801
T0596	174	model5049_3c07_ffas1.pdb	3.551	3.551
T0600	106	TT5603.pdb	19.402	4.727
T0601	442	model737_1vpb_Profile1.pdb	1.934	1.833
T0602	55	N1.pdb	3.834	5.308
T0603	262	TT3230.pdb	5.791	4.908
T0605	48	TT1800.pdb	2.052	2.415
T0606	123	N5.pdb	6.918	5.155

System	No. of res	Model name	RMSD (before ab_initio)	RMSD (after ab_initio)
T0610	179	TT1057.pdb	3.275	4.009
T0611	203	TT2374.pdb	1.387	5.927
T0612	105	TT1680.pdb	6.533	5.913
T0613	285	model929_3lou_DIR464.pdb	1.512	2.314
T0615	178	model954_1vj7_Profile0.pdb	5.212	5.131
T0616	93	TT2110.pdb	7.676	6.335
T0617	123	TT9998.pdb	2.612	3.096
T0618	158	N2345.pdb	10.543	7.757
T0619	101	model4940_1z1b_ffas14.pdb	2.411	2.443
T0620	299	model5927_3mtc_ffas1.pdb	3.296	4.201
T0622	121	model5688_2vt3_Profile1.pdb	8.252	5.007
T0625	231	N485.pdb	9.262	8.284
T0626	283	model889_3lou_DIR445.pdb	1.772	1.772
T0632	114	TT2146.pdb	1.332	5.385
T0634	116	model677_3crn_Profile7.pdb	2.139	2.139
T0635	182	model125_2r8x_DIR62.pdb	3.056	2.181
T0636	322	model3180_3euc_ffas8.pdb	3.239	8.089
T0637	135	TT523.pdb	6.671	6.671
T0638	218	TT0.pdb	2.575	2.575
T0640	232	TT2730.pdb	2.722	2.618
T0641	293	TT3350.pdb	3.07	3.047
T0643	73	N1274.pdb	6.897	5.486

Table 8. Table listing the results of MD simulations after 100ps, 200ps, 500ps, 1ns, (in some cases 5 ns)

Note: The higher the percentile, the better is the accuracy of the structure.

System	No. of res	GDT_TS	RMSD (Å)	Clash score (percentile)	Poor rotamers Goal: <1%	Ramachandran outliers Goal: <0.2%	Ramachandran favored Goal: >98%	MolProbity score (percentile)	Residues with bad bonds: Goal: 0%	Residues with bad angles: Goal: <0.1%
T0544	**135**	**25.556**	**10.498**							
Initial (w/o MD)		25.556	10.498	1225.23 (0th per)	-	-	-	4.68 (0th per)	-	-
100ps-7A		24.63	10.926	19.86 (33rd per)	0.0571	0.0236	0.8661	3 (23rd per)	0.0465	0.062
200ps-7A		22.778	11.319	18.44 (36th per)	0.0286	0.0236	0.8583	2.75 (34th per)	0.0233	0.0543
500ps-7A		25.185	11.167	14.66 (51st per)	0.0571	0.0079	0.8661	2.87 (28th per)	0.0155	0.0775
1ns-7A		25.185	10.907	16.55 (43rd per)	0.0952	0.0079	0.8425	3.14 (18th per)	0.031	0.062
5ns-7A		22.593	11.199	15.13 (49th per)	0.0571	0.0157	0.8425	2.93 (26th per)	0.0465	0.0698
T0616	**93**	**40.86**	**7.676**							
Initial (w/o MD)		40.86	7.676	1214.54 (0th per)	-	-	-	4.67 (0th per)	-	-
100ps-10A		37.903	8.422	4.12 (96th per)	0.0761	0.0306	0.8265	2.56 (43rd per)	0.03	0.06
200ps-10A		37.097	8.768	4.12 (96th per)	0.0543	0.0306	0.8367	2.44 (51st per)	0.07	0.03
500ps-10A		36.29	8.491	2.94 (98th per)	0.0435	0.0102	0.8265	2.27 (61st per)	0.04	0.06
1ns-10A		35.484	8.233	2.35 (99th per)	0.0217	0.0408	0.8776	1.88 (82nd per)	0.01	0.03
T0578	**156**	**29.006**	**11.996**							
Initial (w/o MD)		29.006	11.996	1240.5 (0th per)	-	-	-	4.68 (1st per)	-	-
100ps-7A		25.962	12.457	34.64 (11th per)	0.0211	0.0696	0.7785	3.03 (22nd per)	0.0375	0.1375
200ps-7A		26.763	12.265	29.74 (16th per)	0.0352	0.0633	0.8165	3.09 (20th per)	0.025	0.0562
500ps-7A		26.923	11.944	25.98 (20th per)	0.007	0.0506	0.8101	2.62 (40th per)	0.0312	0.0688
1ns-7A		26.442	12.182	24.85 (22nd per)	0.0352	0.0696	0.7785	3.06 (21st per)	0.0312	0.0938

System	No. of res	GDT_TS	RMSD (Å)	Clash score (percentile)	Poor rotamers Goal: <1%	Ramachandran outliers Goal: <0.2%	Ramachandran favored Goal: >98%	MolProbity score (percentile)	Residues with bad bonds: Goal: 0%	Residues with bad angles: Goal: <0.1%
T0618	**158**	**29.905**	**10.543**							
Initial (w/o MD)		29.905	10.543	1161.26 (0th per)	-	-	-	4.65 (0th per)	-	-
100ps-7A		27.848	10.719	23.57 (23rd per)	0.0629	0.0115	0.8506	3.13 (18th per)	0.0171	0.0686
200ps-7A		29.747	10.592	18.86 (36th per)	0.0629	0.0172	0.8506	3.04 (21st per)	0.0229	0.0571
500ps-7A		29.43	10.683	17.51 (40th per)	0.0503	0.0287	0.8448	2.94 (25th per)	0.0514	0.0914
1ns-7A		27.057	10.759	15.15 (49th per)	0.0629	0.0115	0.8391	2.97 (24th per)	0.04	0.0457
5ns-7A		25.475	11.514	15.82 (46th per)	0.044	0.0287	0.8448	2.86 (29th per)	0.04	0.0686
T0564	158	**48.75**	**8.004**							
Initial (w/o MD)	80	44.944	8.255	76.38 (0th per)	0.1169	0.0115	0.9425	3.55 (8th per)	0	0.0112
100ps-7A		44.375	7.89	0.7 (99th per)	0.1053	0.0116	0.907	2.03 (74th per)	0.0682	0.0341
200ps-7A		42.5	8.151	3.52 (97th per)	0.0658	0.0116	0.8372	2.44 (50th per)	0.0227	0.0682
500ps-7A		45.312	8.892	1.41 (99th per)	0.0658	0.0349	0.8953	2.06 (73rd per)	0	0.0455
1ns-7A		46.562	8.469	0.7 (99th per)	0.0526	0.0116	0.8488	1.94 (79th per)	0.0227	0.0455
5ns-7A		42.812	8.526	0.7 (99th per)	0.0789	0.0116	0.8488	2.07 (72nd per)	0	0.0114
T0622	**121**	**64.256**	**8.252**							
Initial (w/o MD)		64.256	8.252	83.64 (0th per)	0.0333	0	0.9412	3.18 (17th per)	0	0.0072
100ps-5A		60.95	8.718	6.24 (90th per)	0.069	0.0152	0.9015	2.52 (46th per)	0.0672	0.0299
200ps-5A		63.017	8.637	2.68 (98th per)	0.0259	0	0.8712	1.99 (76th per)	0.0373	0.0746
500ps-5A		61.364	8.6	3.12 (98th per)	0.0431	0.0076	0.9015	2.13 (69th per)	0.0149	0.0597
1ns-5A		62.19	8.598	0.45 (99th per)	0.0517	0	0.9318	1.64 (91st per)	0.0373	0.0224
T0580	**104**	**80.288**	**1.958**							
Initial (w/o MD)		80.288	1.958	83.85 (0th per)	0.0706	0.0097	0.9515	3.37 (11th per)	0	0.0286

System	No. of res	GDT_TS	RMSD (Å)	Clash score (percentile)	Poor rotamers Goal: <1%	Ramachandran outliers Goal: <0.2%	Ramachandran favored Goal: >98%	MolProbity score (percentile)	Residues with bad bonds: Goal: 0%	Residues with bad angles: Goal: <0.1%
100ps-5A		73.798	2.343	4.89 (94th per)	0.0976	0.01	0.94	2.41 (52nd per)	0.0392	0.0784
200ps-5A		71.394	2.571	3.05 (98th per)	0.0488	0	0.93	2.07 (72nd per)	0.0294	0.0686
500ps-5A		75	2.155	4.28 (96th per)	0.0244	0	0.95	1.85 (83rd per)	0.0196	0.0294
1ns-5A		79.327	1.977	0 (100th per)	0.0488	0.02	0.94	1.42 (97th per)	0.0392	0.0686
5ns-5A		76.923	2.554	1.83 (99th per)	0.0122	0.02	0.91	1.53 (94th per)	0.049	0.0784
T0541	**106**	**75**	**2.181**							
Initial (w/o MD)		75	2.181	58.45 (2nd per)	0.0103	0	0.9519	2.58 (42nd per)	0	0
100ps-5A		67.689	2.636	2.59 (98th per)	0.0625	0	0.9029	2.19 (65th per)	0.0095	0.1238
200ps-5A		73.349	2.411	1.29 (99th per)	0.0417	0.0194	0.9417	1.72 (88th per)	0.0762	0.0952
500ps-5A		75.236	2.294	1.94 (99th per)	0.0521	0.0097	0.9223	1.98 (77th per)	0.0095	0.0381
1ns-5A		69.104	2.466	2.59 (98th per)	0.0625	0	0.9515	1.99 (76th per)	0.0381	0.1143
T0538	**54**	**86.574**	**1.994**							
Initial (w/o MD)		86.574	1.994	54.3 (3rd per)	0.0217	0	0.9808	2.46 (49th per)	0	0
100ps-5A		76.852	2.29	2.23 (99th per)	0.0435	0.0192	0.9423	1.87 (82nd per)	0.037	0.0556
200ps-5A		76.389	2.229	0 (100th per)	0	0	0.9615	0.76 (100th per)	0.037	0.0185
500ps-5A		78.704	2.051	0 (100th per)	0.0652	0.0192	0.9615	1.38 (97th per)	0.0556	0.037
1ns-5A		81.019	1.982	0 (100th per)	0.087	0	1	1.21 (99th per)	0	0.0556
5ns-5A		84.722	2.452	0 (100th per)	0.0217	0	0.9808	0.76 (100th per)	0.037	0.0556
T0588	**381**	**40.682**	**7.707**							
Initial (w/o MD)		40.682	7.707	100.6 (0th per)	0.0478	0.0352	0.8794	3.58 (7th per)	0	0.0125
100ps-5A		37.139	7.974	3.93 (96th per)	0.1	0.0407	0.8142	2.65 (39th per)	0.0228	0.0835
200ps-5A		36.089	8.131	3.3 (97th per)	0.0788	0.028	0.827	2.50 (47th per)	0.0405	0.0608

System	No. of res	GDT_TS	RMSD (Å)	Clash score (percentile)	Poor rotamers Goal: <1%	Ramachandran outliers Goal: <0.2%	Ramachandran favored Goal: >98%	MolProbity score (percentile)	Residues with bad bonds: Goal: 0%	Residues with bad angles: Goal: <0.1%
500ps-5A		34.186	8.413	3.62 (97th per)	0.0667	0.0204	0.8295	2.47 (49th per)	0.0253	0.0684
1ns-5A		33.661	8.451	2.2 (99th per)	0.0758	0.0254	0.8422	2.34 (56th per)	0.043	0.0633
T0594	**140**	**81.25**	**2.545**							
Initial (w/o MD)		81.25	2.545	61.51 (2nd per)	0.0556	0.029	0.9203	3.31 (13th per)	0	0
100ps-5A		79.821	2.959	1.40 (99th per)	0.041	0.0299	0.8806	1.92 (80th per)	0.0221	0.0441
200ps-5A		77.321	2.943	5.76 (91st per)	0.041	0.0149	0.9403	2.18 (66th per)	0.0294	0.0809
500ps-5A		77.679	3.358	2.21 (99th per)	0.0246	0.0224	0.9179	1.79 (86th per)	0.0147	0.0588
1ns-5A		74.107	3.176	3.1 (98th per)	0.0328	0.0075	0.9179	1.99 (76th per)	0.0294	0.0588
5ns-5A		72.679	3.705	2.21 (99th per)	0.041	0.0224	0.8731	2.08 (72nd per)	0.0221	0.0294
T0619	**101**	**75**	**2.411**							
Initial (w/o MD)		75	2.411	51.92 (3rd per)	0.0319	0.0192	0.9615	2.83 (30th per)	0	0
100ps-5A		68.812	2.989	1.15 (99th per)	0.0667	0.0198	0.9406	1.85 (83rd per)	0.0196	0.0588
200ps-5A		69.802	2.999	1.15 (99th per)	0.0111	0.0297	0.9406	1.26 (99th per)	0.0196	0.049
500ps-5A		66.337	3.035	1.72 (99th per)	0.0111	0.0396	0.9109	1.48 (96th per)	0.0392	0.0588
1ns-5A		65.099	3.516	1.15 (99th per)	0.0444	0.0198	0.9307	1.76 (87th per)	0.049	0.1078
5ns-5A		66.337	3.398	1.72 (99th per)	0.0222	0.0198	0.901	1.74 (88th per)	0.0098	0.0784

APPENDIX II

RMSD: The **root-mean-square deviation** (**RMSD**) is the measure of the average distance between the atoms (backbone atoms) of superimposed proteins. In the study of globular protein conformations, it is used to measure the similarity in three-dimensional structure by the RMSD of the Cα atomic coordinates after optimal rigid body superposition.

GDT_TS: The **global distance test total score** or **GDT** is a measure of similarity between two protein structures with identical amino acid sequences but different tertiary structures. It is most commonly used to compare the results of protein structure prediction to the experimentally determined structure as measured by X-ray crystallography or protein NMR. The metric is intended as a more accurate measurement than the more common RMSD metric, which is sensitive to outlier regions created by poor modeling of individual loop regions in a structure that is otherwise reasonably accurate.

GDT_TS measurements are used as major assessment criteria in the production of results from the Critical Assessment of Structure Prediction (CASP), a large-scale experiment in the structure prediction community dedicated to assessing current modeling techniques and identifying their primary deficiencies

CASP: **Critical Assessment of Techniques for Protein Structure Prediction** is a worldwide experiment for protein structure prediction taking place every two years since 1994. It provides research groups with an opportunity to test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users.

LGA: **Local-Global Alignment** method is designed to facilitate the comparison of protein structures or fragments of protein structures in sequence dependent and sequence independent modes. The LGA structure alignment program is available as an online service at http://PredictionCenter.llnl.gov/local/lga and can also be locally installed. Data generated by LGA can be successfully used in a scoring function to rank the level of similarity between two structures and to allow structure classification when many proteins are being analyzed. The GDT_TS and RMSD for all the structures was calculated using LGA program.