Refinement and Improvement of Template Based Protein Modelling Algorithms

THESIS

Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

By

ASHISH RUNTHALA
2008PHXF406

Under the Supervision of
Prof. SHIBASISH CHOWDHURY

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)
INDIA

2015

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE PILANI (RAJASTHAN)

CERTIFICATE

This is to certify that the thesis entitled "**Refinement and Improvement of Template Based Protein Modelling Algorithms**" submitted by **Ashish Runthala, ID No. 2008PHXF406P** for award of Ph. D. degree of the Institute embodies the original work done by him under my supervision.

Signature in full of the supervisor	:	
Name in capital block letters	:	**SHIBASISH CHOWDHURY**
Designation	:	**Associate Professor**

Date:

ACKNOWLEDGEMENT

It’s really a great pleasure and immense satisfaction in expressing my deep gratitude towards my research supervisor, Prof. Shibasish Chowdhury Associate Professor, Department of Biological Sciences, BITS, Pilani, for his dexterous guidance, suggestions and support which he bestowed to me. Mere acknowledging with words is inadequate to express my gratitude to him. He was always an inspiration to me in the research. The work environment given to me under him, the experiences gained from him and his creative working culture are extremely treasured and will be remembered throughout my life.

I deeply acknowledge and earnestly thank Prof. AK Das and Dr. Rajesh Mehrotra (Head of the Department of Biological Sciences, BITS, Pilani) for their valuable suggestions, guidance and precious time and support which he offered me throughout my research. I also sincerely thank Dr. Debashree Bandyopadhyay for reviewing my thesis and providing the fruitful comments.

I am grateful to Vice-Chancellor (BITS, Pilani) and Prof. AK Sarkar, Director (Pilani campus), for allowing me to carry out my doctoral research work in the institute.

I am happy to express my sincere thanks to S.C Sivasubramanian, Dean Administration (Pilani Campus), Dean, Educational Development Division and Faculty Division III, for his uninterrupted support during my research work.

I am extremely thankful to Prof. S. K. Verma, Dean, Academic Research Division, BITS, Pilani, for his co-operation and for providing me with all the necessary academic facilities and helping me at various stages of my research work.

I sincerely acknowledge the help rendered by Dr. B.Vani, Prof. Jitendra Panwar, Prof. Lalita Gupta, Mr. Manoj Kannan, Dr. Pankaj K Sharma, Dr. Prabhat N Jha, Dr. Rajdeep Chowdhury, Dr. Sandhya Marathe, Dr. Sandhya Mehrotra, Prof. Sanjeev Kumar, Dr. Shilpi Garg, Dr. Sudeshna Mukherjee, Prof. Uma S Dubey and Prof. Vishal Saxena.

I am highly grateful to all the research scholars especially Amit Subudhi, Akanksha Pareek, Anyaa Mittal, Arpit Bhargava, Boopathi PA, Chetna Sangwan, E.Divya Niveditha, Gagan Deep, Gurpreet Kaur, Isha Pandey, Jyothi Nagraj, Kuldeep, Leena Fageria, Manohar Lal, Mithilesh Kajla, Monika M, Panchsheela Nogia, Parik Kakani, Parva Sharma, Poonam Singh, Rajnish, Ramandeep Kaur, Ramapuram Dilip, Ranita De, Rini Dhawan, Sandeep Poonia, Senthil, Shraddha Mishra, Subhra Das, Swarna Kanchan, Tania Pal Choudhury, Vandana, Vidushi Asati, Vikram Pareek, Zaiba Hasan Khan and Zarna Pala for the time they had spent for me.

I express my thanks to our office staff members Mr. Parmeshwar Nayak, Mr. Subhash Chander, Mr. Mukesh Saini, Mr. Kamlesh Kumar Soni and Mr. Naresh Kumar Saini for all of their help and support in one way or the other.

I deeply acknowledge the University Birla Institute of Technology and Science, Pilani for extending the financial assistance and providing the required resources and support.

I would like to dedicate this pious piece of research work to my father, mother, brothers, sisters, wife, grandparents and parents-in-law, whose dreams had come to life with me getting the highest degree in education. I specially owe my doctorate degree to my mother, who kept with her continuous care, support and encouragement. Thanks are due if I don't dedicate this thesis to all of my other family members whose continuous support, love and affection well supported me to reach this height.

Lastly, and above all, I would like to thank the God almighty; for all that he has given or always planned to give me.

Ashish Runthala

Abstract

Protein structure prediction algorithms are studied to construct accurate models of the protein sequences for bridging the ever-increasing gap between the available count of protein sequences and the experimentally determined protein structures. Comparative modelling is considered as most popular and accurate structure prediction algorithm to model protein structure. Template selection is considered as one of the most important steps of a comparative modelling algorithm. However, selection of the best set of templates is still a major challenge. An effective template ranking algorithm is developed to efficiently select only the reliable hits for predicting the accurate protein structures. The algorithm employs the pairwise as well as multiple sequence alignments of template hits to respectively capture their key sequence and structural information based scores for effectively ranking them, selecting their best possible set and constructing an accurate target model. Modelling accuracy of the algorithm is tested and evaluated on TBM-HA domain containing CASP8, CASP9 and CASP10 targets. In-house C, Python and PERL scripts are used to select the functionally similar and structurally complimentary template hits to model the protein sequences. Protein models sampled through MODELLER are evaluated through different assessment scores viz. MOLPDF, GA341, DOPE Score, Normalized DOPE Score, GDT-TS, GDT-HA and TM_Score. TM_Score along with Normalized DOPE score (Z_Score) is lastly selected as the best set of model assessment measures and is employed to evaluate the model sampling for selecting the accurate target model. The statistical ranking based template selection and combination algorithm, further integrated with TM_Score and Z_Score assessed iterative sampling strategy, significantly improves the modelling accuracy of the targets. The algorithm predicts accurate models with an average GDT-TS, GDT-HA and TM_Score improvement of 3.531, 4.814 and 0.022 along with the individual relevant standard deviation of 4.142, 6.353 and 0.037 over the best CASP models. The predicted models are found more accurate than the best CASP models not only for the individual domains but also for the overall target conformation. Our results suggest that the inclusion of structurally similar templates with ample conformational diversity is vital for the modelling algorithm to maximally as well as reliably span a target sequence and construct its accurate model. The optimal model sampling also holds the key to predict the best possible structure for a target.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACCPRO	Solvent accessibility prediction of a protein
AMBER	Assisted model building with energy refinement
ANOLEA	Atomic non-local environment assessment
BLOSUM	Blocks substitution matrix
CASP	Critical assessment of structure prediction
CATH	Class architecture topology homology
CE	Combinatorial extension
COG	Clusters of orthologous groups
CG	Conjugated gradient
CHARMM	Chemistry at harvard molecular mechanics
COMPASS	Comparison of multiple protein sequence alignments with assessment of statistical significance
CONGEN	Conformation generator
CSA	Conformational space annealing
CS-BLAST	Context specific basic local alignment search tool
DALI	Distance matrix alignment
DELTA-BLAST	Domain enhanced lookup time accelerated blast
DFIRE	Distance-scaled, finite ideal-gas reference
DISOPRED	Disorder prediction
DOMAC	Accurate, hybrid protein domain prediction
DOMPRO	Protein domain predictor
DOPE	Discrete optimized potential energy
DSSP	Dictionary of protein secondary structure

EVA	Evaluation of protein structure
EBI	European bioinformatics institute
FAMS	Fully automated modeling system
FASTA	Fast alignment
FM	Free modelling
GA	Genetic algorithm
GBSA	Generalized born surface area
GDT	Global displacement test
GDT-HA	Global displacement test – high accuracy
GDT-TS	Global displacement test – total score
GROMOS	Groningen molecular simulation
HMM	Hidden markov model
ICM	Internal coordinate modeling
INDEL	Insertion or deletion
ISS	Intermediate sequence search
LCS	Longest continuous segments
LGA	Local global alignment
LOMETS	Local meta threading server
MC	Monte carlo
MD	Molecular dynamics
MOLPDF	Molecular probability density function
MSA	Multiple sequence alignment
MQAP	Model quality assessment program
MULTICOM	Multi-template combination
MUSCLE	Multiple sequence comparison by log- expectation

MUSTER	Multi-sources threader
NMR	Nuclear magnetic reasonance
PAM	Point accepted mutation
PDB	Protein data bank
PDF	Probability density function
PFAM	Protein family
PHS	Parallel hyperbolic sampling
PIR	Protein information resource
PHYLIP	The phylogeny inference package
PHYRE	Protein homology recognition
PPA	Profile profile alignment
PROCHECK	Protein structure check
PROSA	Protein structure analysis
PSI-BLAST	Position specific iterative - basic local alignment search tool
PSIPRED	Psi-blast based secondary structure prediction
PSSM	Position specific scoring matrix
PULCHRA	Powerful chain restoration algorithm
RAPTOR	Rapid protein threading by operation research
REMD	Replica exchange molecular dynamics
REMO	Reconstruct atomic model
R-factor	Reliability factor
RMSD	Root mean square deviation
SA	Simulated annealing
SAGA	Simulated annealing genetic algorithm
SALIGN	Sequence/structure alignment server

SAM	Sequence alignment and modeling system
SCOP	Structural classification of proteins
SCWRL	Side-chains modelling with a rotamer library
SIB	Swiss institute of bioinformatics
SPICKER	Structure picker
SSPRO	Secondary structure prediction of a protein
STAMP	Structural alignment of multiple proteins
TASSER	Threading assembly refinement
TBM	Template based modelling
TBM-HA	Template based modelling high accuracy
TM_Score	Template Modelling Score
TOPITS	Threading one-dimensional predictions into three-dimensional structures
TrEMBL	Translated european molecular biology laboratory
UniprotKB	Universal protein knowledgebase
UNIRES	United residue
VDW	Van der waals

Chapter I

Introduction

1.1 Protein

Proteins are the major building blocks of cell machinery and are involved in almost every cellular function like genetic regulation, metabolism and the cell proliferation (Tomkins & Martin 1970). Cellular proteins are produced through ribosomes by the process of translation from a set of 20 naturally occurring amino acids with different chemical features and properties (Lucas-Lenard 1971). In this process, transcribed gene sequence or messenger RNA is translated into a linear chain of amino acids connected by the peptide bonds (Wilson 1971). The amino acids of a protein sequence or primary protein structure interact with each other and the surrounding environment to produce the stable three dimensional conformation, which is also known as the native state (Hanley et al. 1983). Both the protein sequence as well as its functionally-active three-dimensional structure are essential to study its cellular activity and are biologically important (Mihaesco et al. 1983).

1.2 Protein sequence

Protein sequence information or primary structure of a protein is essential to understand its function in a cell. Proteins are sequenced through different techniques like the Edman degradation (Edman 1949), Sanger sequencing (Sanger & Tuppy 1951) and the high throughput sequencing methodology (Metzker 2010). Protein sequence database was created as the Atlas of Protein sequence and structure (Dayhoff et al. 1965) in 1965 and was later linked with european bioinformatics institute (EBI), swiss institute of bioinformatics (SIB) and protein information resource (PIR) databases in 2002 to form a complete Uniprot consortium. Recently, secondary databases Swiss-prot and translated european molecular biology laboratory (TrEMBL) were also developed from the EBI and SIB databases.

Currently (as on January 19th, 2015), 547,357 annotated protein sequences exist in the swissprot section of universal protein knowledgebase (UniProtKB) database. It is interesting to note here that this number is quite small when compared to the 89,451,166 protein sequences available in the complete translated european molecular biology laboratory (TrEMBL) repository of UniProtKB (http://www.ebi.ac.uk/uniprot/TrEMBLstats). The TrEMBL database also includes the computationally translated copies of all coding segments present in the nucleotide sequences available in the UniProtKB and is substantially larger than UniProtKB database. Moreover, many of these database sequences are homologous to each other and share a substantial (30%) sequence similarity. Even after excluding these homologous sequences from the sequence databases, the remaining 67,265,680 protein sequences existing in the TrEMBL database also significantly exceed the number of experimentally solved protein structures.

Although the protein sequence information is important to understand its function, its three dimensional structure play important functional role in a cell. As a protein sequence can adopt different structures in different chemical environments due to altered folding kinetics, the three-dimensional conformation of a protein sequence is more informative than its primary sequence (Anfinsen 1973; Hellberg et al. 1987; Petrey et al. 2015).

1.3 Protein structure

Proteins are involved in every aspect of biological activity and their detailed structural study is important to understand the mechanism of their biological activity or to interfere with it, as for example in the case of drug design and folding related diseases. Structural details of a protein assist us to identify the key residues that are primarily responsible for its biological function, i.e. if a protein sequence is mutated to delete such key residues through a knockout and site-directed mutagenesis experiment, its biological function is completely lost in a cell system (Filippis et al. 1994; Iverson et al. 2002; Goyal et al. 2015). The structural properties of a protein sequence like superficial surface, conformational topology, loop flexibility and residue accessibility help us in understanding its biological function. Further, the conserved domain that is functionally important for a protein sequence is also studied through its structure (Matsui et al. 2004; Chapple et al. 2015). Moreover, the recombinant proteins cloned in a cell sometimes aggregate into misfolded conformations which lead to the formation of inclusion bodies or abnormal assemblies and loss of its functional role(s) in a cell (McCallus et al. 1992). Such functionally defective protein aggregations lead to several diseases like amyloidosis (Chiti & Dobson 2006), cystic fibrosis (Luheshi et al. 2008) and Alzheimer's disease (Gadad et al. 2011). Certain specific chemical interactions and structural constraints between the solvent-exposed hydrophobic residue stretches usually drive the *in-vitro* aggregation of partially folded recombinant proteins into functionally abnormal aggregates (Carrio et al. 2005). Studying the biophysical reasons for this abnormal aggregation of a recombinant protein allows us to identify and substitute the key residues responsible for the aggregation with the structurally tolerable residues, and it further empowers us to prevent its excruciating cellular aggregation and result in its increased *in-vitro* expression rate (Carrio et al. 2005). Detailed knowledge of protein structures further helps us to map their functionally active cross-talks in a cell (Wolf et al. 1998; Kevin et al. 2011). As protein structures are more conserved than sequences, the evolutionary relationship among two distantly related proteins is studied by assessing their structural similarity (Kim 1998; Montelione & Anderson 1999; Marti-Renom et al. 2000; Pentony et al. 2012).

Proteins contain a well-defined pattern of secondary structures and a complex three-dimensional conformation. The secondary structure of a protein is usually defined as the local hydrogen-bonded configuration which is formed while the protein is folding into its native conformation. Apart from this hydrogen bonding pattern, the key signature of the backbone of a protein secondary structure is its specific set of right-handed phi (Φ) and psi (ψ) torsion angles that describe the rotation of polypeptide backbone around the N-Cα and Cα-C bonds respectively. Sterically favourable combination of these torsion angles formed by a set of consecutive amino acids provides the flexibility required to fold this stretch into a specific secondary structure since the third possible torsion angle (ω) around the peptide bond within the protein backbone is planar and almost invariably fixed at 180° due to its double bond character. Thus the phi and psi angles form the conformational basis of a protein secondary structure that majorly contains α-helices, β-sheets and turns.

Depending on the set of encoded secondary structures, the protein conformations are categorized into four structural classes viz. all-α, all-β, α/β and α+β. The all-α and all-β proteins respectively contain only the α-helices and β-strands. However, the α/β and α+β protein structures encode the mixed sets of α-helices and β-strands, where the former consists of alternating α-helices and β-strands (mostly parallel) along its main-chain and the latter consists of α-helices and β-strands (mostly anti-parallel) that occur separately across the main-chain (Murzin et al. 1995). Several such super-secondary structures interact together to form a compact tertiary structure for a protein and the further alliance of these tertiary structures through non-covalent and disulphide bonds results in the formation of a multi-subunit or quaternary structure (Rossmann & Argos 1981).

The protein structures including the ones that form a complex with nucleic acids (NA) are solved through several experimental techniques like X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy techniques, as shown in Table 1.1. Protein structures are also experimentally determined with some other techniques viz. Solid-state NMR, Electron crystallography, Neutron diffraction, Fiber diffraction, Solution scattering, Infrared spectroscopy, Powder diffraction and their hybrid set (shown as Other in Table 1.1).

Myoglobin was the first protein that was experimentally solved through X-ray crystallography in 1962. This 153 residue protein structure of the *Physeter catodon* (sperm whale) was solved at 2Å resolution. Currently all the experimentally solved protein structures are deposited in a single database as protein data bank (PDB, Berman et al. 2000) which was founded in 1971. Presently (*as on January 19th, 2015*), 102994 experimentally solved protein structures have been released by the PDB, as listed in Table 1.1.

Table 1.1 Current PDB Holdings (January 19th, 2015, http://www.rcsb.org).

		←Molecule Type→			
		Proteins	**Protein/NA Complexes**	**Other**	**Total**
Experimental Techniques	X-RAY	88026	4332	4	92362
	NMR	9466	222	7	9695
	ELECTRON MICROSCOPY	522	164	0	686
	HYBRID	68	2	1	71
	Other	161	6	13	180
	Total	98243	4726	25	102994

As it is well noticed (Number of protein sequences is 89,451,166 and number of experimentally solved protein structures is 102994), the protein sequencing rates are significantly higher than the rate at which their structures are getting experimentally solved, and this gap between the number of experimentally solved protein structures and the count of protein sequences is constantly increasing. Although the structure determination methodologies have been developed to a great extent through the high throughput experimental approaches, the sequence-structure gap is constantly increasing due to several technical and resource limitations. It is because the X-Ray analysis requires an extremely pure protein crystal and many proteins do not crystallize. NMR analysis on the other hand is just limited to small, soluble proteins only, with slightly lower accuracy of approximately 2.5Å. Cryo-electron microscopy also suffers from the problems that the electron-microscopy maps are not unambiguous for a protein and its resolution is also much lower than that of the X-ray and NMR techniques (Baker & Johnson 1996; Rossman MG. 2000). Moreover, all these experimentally determined structures require conformational refinement through costly and time consuming experimental steps. Protein structure prediction methods or protein modelling algorithms are thus being developed to build a protein structure simply from its primary sequence information (Lushington 2015). Currently, it seems to be a complete realistic objective as it promises to quickly construct near-native protein models (Battey et al. 2007; Das et al. 2007; Kryshtafovych et al. 2007; Kryshtafovych et al. 2009; Li et al. 2015).

1.4 Protein structure prediction

The count of experimentally determined protein structures is substantially smaller than the number of known protein sequences, as shown in Fig. 1.1. Due to the development of high throughput sequencing methodologies, the number of protein sequences is exponentially increasing, as shown in Figure 1.1 (a). The yearly growth and an overall increase in the number of experimentally solved protein structures, existing in the PDB database, are diagrammatically represented in Fig. 1.1 (b). Hence, an experimentally solved structure is unavailable for majority of the protein sequences and this gap has widened over the last decade despite the development of high-throughput dedicated crystallography pipelines (Berman et al. 2000; Metzker 2010). Protein structure prediction is thus a major research objective to bridge this sequence-structure gap.

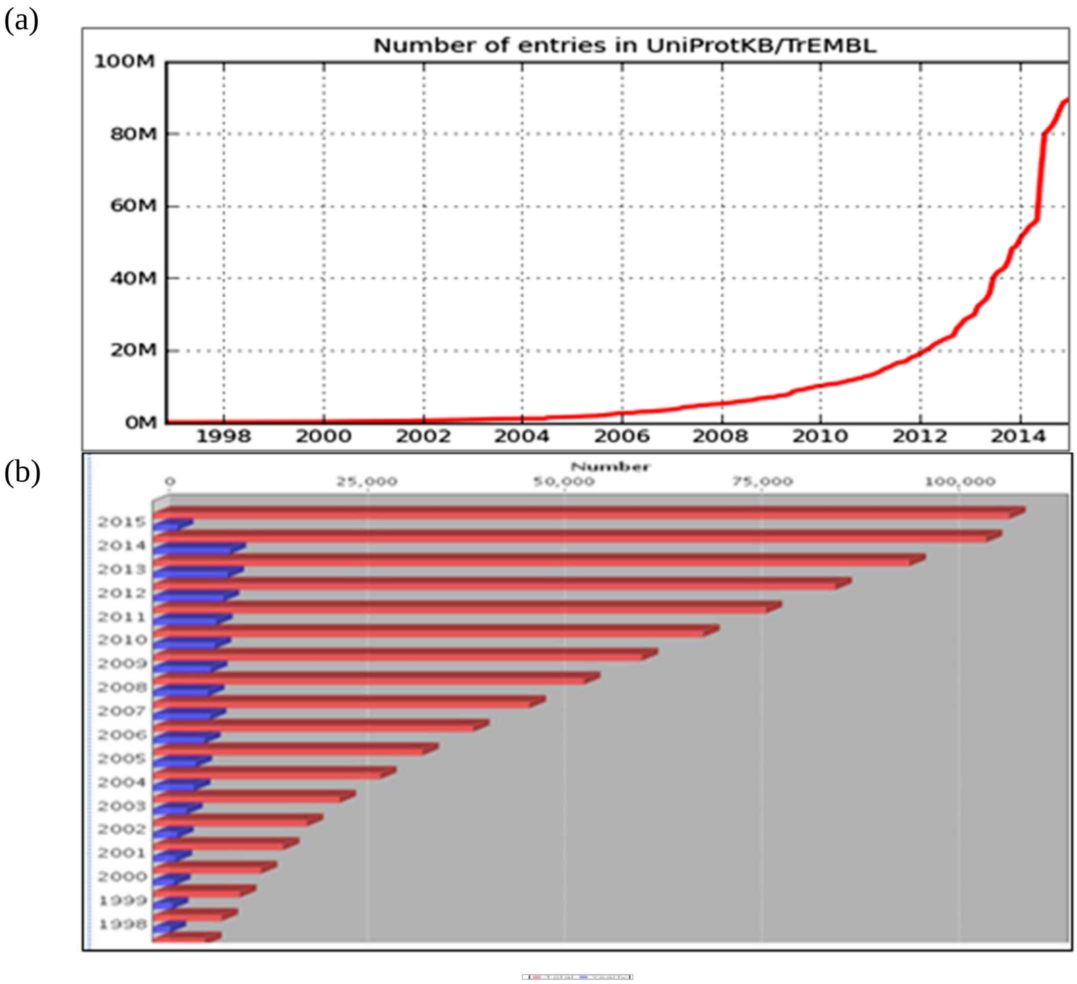


Fig. 1.1 Yearly growth of (a) total number of protein sequences in the UniprotKB/ TrEMBL database (http://www.ebi.ac.uk/uniprot/TrEMBLstats) (b) total number of protein structures in the PDB database (http://www.rcsb.org/pdb/statistics/holdings.do).

1.5 Protein structure prediction algorithms

The protein structure prediction methodologies have been categorized in two broad categories namely free modelling (FM) and the template based modelling (TBM) (Kryshtafovych et al. 2005; Cheng 2007). TBM encompasses the comparative modelling and threading algorithms that predict a protein structure on the basis of its similarity with the available set of experimentally solved structures (templates). Conversely, FM includes the *ab-initio* modelling methodologies which predict a protein structure solely from its sequence information, as further detailed below. As TBM algorithm employs the best homologous set of templates to reliably cover the target sequence, it is able to escape the extensive conformational search problem of a FM algorithm (Tress et al. 2005; Zhang 2007). By employing an optimally scoring target-template alignment, a TBM algorithm is thus able to quickly construct the accurate target models (Tress et al. 2007; Tong et al. 2015) than the FM methodologies which involve several computationally complicated steps (Tress et al. 2009). The TBM and FM algorithms have different degree of modelling accuracy, computational and algorithmic complexity, as shown below with a flowchart in the Fig. 1.2.

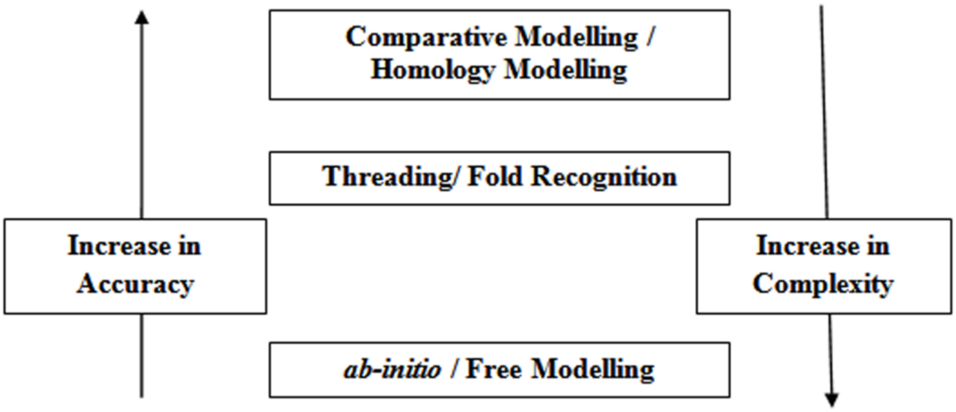

Comparative Modelling / Homology Modelling
Threading/ Fold Recognition
Increase in Accuracy
Increase in Complexity
ab-initio / Free Modelling

Fig. 1.2 Schematic representation of different protein modelling categories and their accuracy as well as complexity orders.

1.5.1 Free Modelling

The *ab-initio* or *de-novo* prediction algorithms attempt to build a protein model simply from its sequence information without employing the available set of experimentally solved structures and are extremely helpful in the construction of novel structural folds. The FM methodology is considered as the “Holy Grail” of modelling algorithms as its solution would ultimately solve all the modelling problems (Zhang 2008). This FM methodology is based on the core physical principles of energy and geometry and it assumes that the actual native state of a protein sequence exists at the lowest free energy conformation. Mathematically, it means that the native state conformation of a target sequence is a model existing at the global minima of its energy landscape. The FM algorithm thus searches the entire possible conformational space of a target protein sequence to identify and construct its native conformation. The FM algorithms basically include the molecular dynamics (MD, Robson & Platt 1987), molecular mechanics (MM, Schiffer et al. 1990) and monte carlo (MC, Higo et al. 1992) methodologies. However, the FM methodology has been developed through several other improved algorithms like genetic algorithms (GA, Ring et al. 1993), multiple copy simultaneous search (Miranker & Karplus 1991), neural networks (Holley & Karplus 1991), field optimization (Koehl & Delarue 1995) and graph theory (Samudrala & Moult 1998) to predict the accurate protein model (Moult & Melamud 2000).

Considering that the conformational space of an amino acid is reasonably approximated with only three discrete torsion angles (Φ and Ψ for the protein backbone and χ_1 for the side-chain), a conformational space of 3^n structures is theoretically possible for a target protein sequence with n amino acids (Chakrabarti & Pal 2001). The FM algorithm is therefore expected to maximally search these 3^n structures for constructing the accurate target conformation. The FM algorithm considers protein structure prediction problem as two sub-problems, i.e. (a) development of an extremely accurate energy function to score and select the accurate conformation among all the generated model decoys and (b) development of a very efficient search protocol to quickly screen the energy landscape possible for a target sequence (Bonneau & Baker 2001) and find its global minima with minimum number of steps. Theoretically, an *ab-initio* approach can model any protein sequence. However, as the possible count of conformations increases exponentially with the total number of atoms in the protein sequence, the computation process becomes prohibitively expensive (Lu & Skolnick 2003) and usually results into relatively less accurate protein model. The ultimate objective of such an algorithm is to reach the global minima conformation for a target sequence (Fiser 2004; Kolinski 2004). *Ab-initio* approaches mainly use geometric optimization algorithms, like Newton-Raphson, Steepest Descent, Conjugated Gradient and Adopted basis Conjugated Gradient to explore the energy landscape.

A protein structure can be represented by two ways, a) an all atom representation of amino acids or b) a reduced representation of amino acids. An all-atom representation considers all the atoms of an amino acid to represent a protein structure and it greatly increases the conformational space of a target sequence. Reduced representation of protein structure only includes Cα atom, Cβ atom, C and N atoms of the peptide bond and the center of mass of the side-chain (Herzyk & Hubbard 1993). However, a different reduced representation of a protein model symbolizes every amino acid with its Cα coordinates and as a single center of interaction (Rotkiewicz & Skolnick 2008) which is sometimes also positioned at the side-chain center of mass (Zhou et al. 2007). Construction of the protein models by these two different representations of an amino acid uses two different energy functions (Sun 1993). The reduced amino-acid representation greatly decreases the computational complexity required to efficiently sample the conformational space of a target sequence. It consequently allows us to easily construct the fine-detailed all-atom target protein model after sampling the target conformational space and constructing its best possible minimal energy structure.

The reduced representation approach has even been considered to chemically represent the behavior of solvent condition that is physically present in a cell system. It is termed as a solvent model and is traditionally categorized as explicit and implicit model (Cramer & Truhlar 1999). The implicit solvent model replaces physical solvent molecules surrounding a protein model with a continuous medium (Roux & Simonson 1999; Tomasi et al. 2005) to equivalently represent the solvent molecules with a reasonable accuracy through a few specific physical parameters like the surface tension, pH and the dielectric constant (García-Moreno & Fitch 2004; Marenich et al. 2009). However, the explicit solvent model overtly considers the solvent molecules to understand their interaction with a protein model and it also spontaneously considers the interactions among the solvent molecules (Borjesson & Hunenberger 2001; Borjesson & Hunenberger 2004; Goh et al. 2014). These solvent models are usually employed for the MM (Still et al. 1990), MC (Burgi et al. 2002) and MD (Ferrara et al. 2002) simulations of a protein structure to accurately predict, simulate and analyze its folding mechanism (Zhou 2003; Xu et al. 2012; Masuda 2015). The MD simulation trajectory has even been employed to assess the conformational stability of a protein model (Cunha et al. 2015).

Recently, Beveridge protocol has combined united residue (UNIRES) empirical energy function of assisted model building with energy refinement (AMBER; Cornell et al. 1995) with an implicit solvent model i.e. generalized born surface area (GBSA; Weiser et al. 1999) to construct a more realistic protein model in an appropriate solvent continuum (Liu & Beveridge 2002). This method includes dielectric polarization of solvent, van der waals (VDW) interaction and cavitation effects (evacuation of the solvent molecules from the intervening space between the hydrophobic residues of a protein structure), mostly observed at the protein structural pockets. MD forms the basis of all the energy functions that are employed to energetically refine the protein models and is the core of *ab-initio* modelling as well as sampling algorithms (van Vlijmen & Karplus 1997; Das et al. 2007). Yet another methodology i.e. replica exchange molecular dynamics (REMD) iteratively includes the predicted models with each energy minimization step to efficiently sample the conformational landscape of a target sequence and further improve the modelling accuracy (Sugita & Okamoto 1999). It energetically relaxes the protein model in a set of multiple non-interacting replicas at different temperatures and at definite step, it exchanges the considered conformation of the initial model with the most stable conformation constructed among the sampled replica decoys. REMD has been found to be extremely successful with the cooperative formation of correct secondary structures when the folding transitions are successfully employed (English & Garcia 2014). Even the implicit solvent based energy minimization protocol has been employed to construct the complete protein model from its reduced-representation structure with only the Cα backbone and the side-chain centre-of-mass (Feig et al. 2000).

The implicit solvent methodology has long been employed to construct and analyze a protein structure that is subjected to a specific solvent (Baysal & Meirovitch 2000). It has even been supplemented with the force field energy equations to evaluate the structural flexibility of a model for selecting the lowest energy as well as topologically correct conformation among the sampled set of target protein structures (Lazaridis & Karplus 2000; Hsieh & Luo 2004; Rodriguez et al. 2011). The implicit solvent models have even been tested to explore the energetic landscape of a protein model (Steinbach 2004; Yelena et al. 2009) and to add the side-chains to a protein backbone model (Lopes et al. 2007) for constructing an accurate protein model.

Altogether, both the explicit and implicit solvent models have been really handy to predict an accurate protein structure and to investigate the stability of a protein conformation (Razzokov et al. 2014). However, even after considering the appropriate solvent continuum for energetic refinements, the target protein model is not always structurally improved with a topology that is closer to the actual global minima of the target sequence. Hence, several optimization techniques are employed to efficiently sample the conformational landscape of the target sequence.

MC sampling guided lattice-based parallel hyperbolic sampling (PHS) algorithm (Zhang et al. 2002) has even been tried in this category and it considers logarithmic flattening of the local high energy barriers prevailing in the conformational landscape of a target sequence by an inverse hyperbolic sine function to quickly bypass the local minima. It spans a considerable conformational space of a target sequence and successfully models the lower energy target models (Zhang et al. 2002). A multiple-copy conformational space annealing (CSA) approach that represents a protein structure with an array of atomic interaction potentials has even been used (Hardin et al. 2002). While progressing towards global minima, it considers interactions between the conformational sites located at Cα atom, Cβ atom and the peptide bond or at the center of mass of the side-chains. A global optimization algorithm has even been attempted through a modified UNRES force field with CSA global optimization approach (Liu & Beveridge 2002). This UNIRES force field has also been employed along with TBM methodologies to improve the modelling accuracy (Krupa et al. 2015). Meanwhile, MC program employing the correct topology of hydrogen bonds and hydrophobic interactions has also been developed to estimate the atomic violations of a protein model for energetically refining it (Srinivasan & Rose 2002). Some other algorithms first construct only the alpha carbon model of all the target residues before adding the backbone and side-chain atoms to build the complete target model (Iwata et al. 2002; Pokarowski et al. 2003; Kolinski 2004; Gront et al. 2007; Zhang et al. 2010; Lyons et al. 2014). Even the modelling methodologies considering the protein model on a cubic or tetrahedral lattice have been developed and this approach has also considered the interactions between the hydrophobic residues and the orientation dependent repulsive interactions existing between the polar and non-polar charged moieties (Jacobsen 2008).

Stochastic and minimum perturbation (Fine et al. 1986) and accurate, hybrid protein domain prediction (DOMAC) methodology (Ginalski 2006) have even been developed in this category. DOMAC aligned the target sequence with the selected set of top-ranked templates to construct the target model through the TBM algorithm. Moreover, it also employed the FM methodology of the protein domain predictor (DOMPRO; Tress et al. 2007) to construct the target segments that were uncovered by the selected templates for building the complete target structure. As the DOMAC algorithm employed both FM as well as TBM methodologies to construct a target model, it was termed as a hybrid modelling algorithm and subsequently the *ab-initio* methodology is termed as FM modelling approach (Cheng 2007). DOMAC also respectively predicted secondary structure and relative solvent accessibility through secondary structure prediction of a protein (SSPRO) and solvent accessibility prediction of a protein (ACCpro) module of its complete SCRATCH suite. But despite these many efforts, domain boundary specificity and sensitivity was found to be just 27% and 14% respectively for the *ab-initio* prediction of the unaligned target segments, as compared to 50% and 76.5% respectively for the aligned target segments modelled with MODELLER through the selected templates (Jauch et al. 2007). The modelling accuracy of this method is too low and still an improved algorithm is needed for practical usage (Jauch et al. 2007).

Although the FM based protein structure prediction algorithms have been developed to a great extent, the predicted structures are still not closer to their actual experimental conformations. Several different algorithms categorized as TBM methodologies employing the already available set of templates are being developed to construct accurate protein models, as discussed below.

1.5.2 Template Based Modelling

TBM modelling methodology includes the comparative modelling and the threading algorithms. Comparative modelling algorithm aligns a target sequence with the template sequences and employs the experimentally solved templates that share a statistically significant sequence similarity with the target sequence to predict its structure (Sali & Blundell 1993; Schoonman et al. 1998).

In between the *ab-initio* and comparative modelling methodologies, there is threading or the fold recognition method that attempts to construct a protein model from several known protein structures which may not share a significant sequence similarity with target. It is based on the observation that nature reuses existing folds for accommodating new protein sequences and functions during evolution (Bray et al. 2000; Moult & Melamud 2000). Very few novel protein folds with unique topologies are experimentally found every year, as graphically represented in the Fig. 1.3 for the data released by class architecture topology homology (CATH; http://www.cathdb.info/wiki/doku/?id=release_notes) database (Orengo et al. 1997). As represented, the total number of available protein folds slowly increase as the nature reuses the existing protein folds in different combinations and topological orientations across all the protein structures. Numerous protein sequences are structurally encoded through a unique and limited set of folds (Pearl et al. 2003; Andreeva et al. 2004).

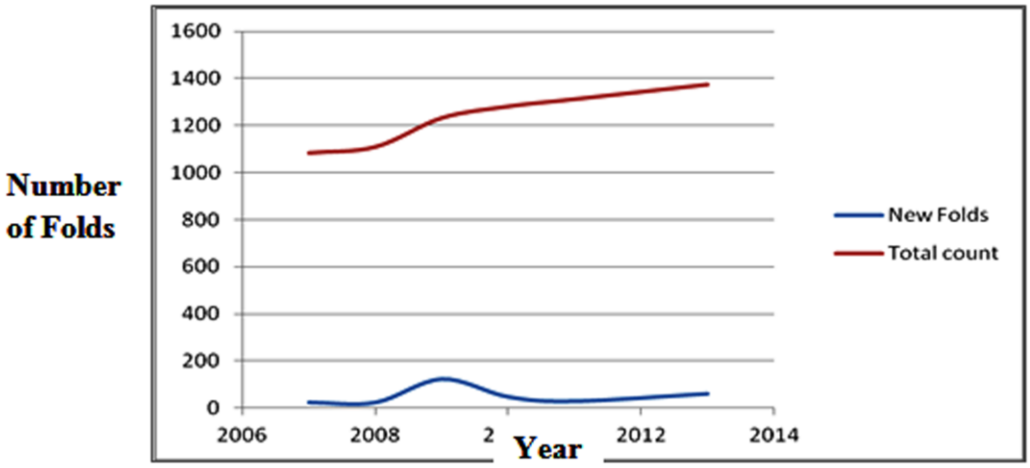
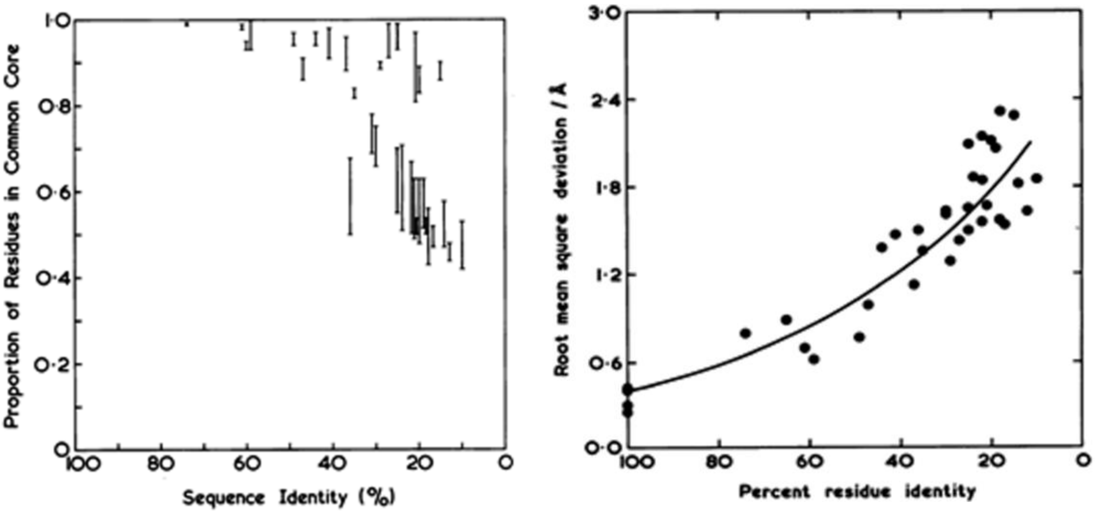


Fig. 1.3 Limited set of novel protein folds exists in the nature (http://www.cathdb.info/wiki/doku/?id=release_notes).

1.6 TBM algorithms

A single point mutation in a gene sequence has been shown to have a deleterious effect on the structure as well as function of its encoded protein(s). It has been shown that a missense mutation altering the sixth codon GAG to GUG in the β-hemoglobin gene results in a substitution of glutamic acid to valine in its protein and it deforms the red blood cells, makes them sickle-shaped and decreases their oxygen carrying capacity (Moo-Penn et al. 1977). Further, the point mutation in the Neurofibromin1 gene has also been shown to cause the Neurofibromatosis disease (Serra et al. 2001). A few missense mutations in the tumor suppressor gene APC have even been shown to cause a cancer (Minde et al. 2011) and the hereditary disorders (Sarig et al. 2012; George et al. 2014). However, it has also been shown that the proteins are structurally too robust over most of these mutations and it has been shown that a minor residue substitution in a protein sequence normally results in a negligible structural change in its protein (Chothia & Lesk 1986; Martinez & Serrano 1999; Sinha & Nussinov 2001; Berrondo & Gray 2011). It has been also observed that evolutionarily related proteins sequences share similar three dimensional structures and the protein structures belonging to the same family are much more conserved than their sequences (Lesk & Chothia 1980). Hence, if an appreciable sequence similarity among two proteins is detected, their structural similarity can be assumed and this relationship forms the basis of TBM algorithm that includes both the comparative or homology modelling, and the threading methodologies.

While the comparative modelling is helpful to construct the model of a target protein sequence that is maximally spanned by the reliable template(s), the threading algorithm is extremely handy for constructing the model when the target sequence shares a distant homology with the templates and can be spanned with the set of conserved structural folds. The availability of statistically significant templates maximally covering the target sequence and also having a substantially higher sequence similarity with it is the key constraint to differentiate between these two modelling algorithms (Tress et al. 2005). The comparative modelling methodology constructs accurate models and is routinely employed to build protein structures (Zhang & Skolnick 2005). However, estimating the accuracy of a predicted model is of prime importance for a biologist intending to employ it for further analysis. A direct correlation between the sequence identity of a pair of proteins and the topological similarity of their common core has already been observed (Chothia & Lesk 1986, Rost 1999), as shown in the Fig. 1.4 (a). The target-template sequence identity has thus usually been considered as a first indicator for the expected accuracy of a protein model (Chothia & Lesk 1986; Kopp and Schwede 2004), as shown in the Fig. 1.4 (b). The Fig. 1.4 (a) highlights the sequence identity percentage of the 8 sets of homologous proteins with the proportion of residues structurally conserved in their hydrophobic core and Fig. 1.4 (b) represents RMSD scores of the mutually compared structures for each of these 8 sets (Chothia & Lesk 1986).



(A) (B)

Fig. 1.4 Sequence identity percentage between a pair of proteins indicates (a) an equivalent proportion of residues conserved in their structural core (b) their minimal structural variation computed in terms of RMSD scores (adapted from Chothia & Lesk 1986).

It has been shown that the templates with 30% identity to a target approximate the structural core of its model to around 1.5-2Ǻ against its native structure and the modelling accuracy may improve to less than 1Ǻ RMSD when the target-template sequence identity is more than 50% (Chothia & Lesk 1986; Cozzetto and Tramontano, 2005; Wishart et al. 2008). However, when the target sequence shares only a statistically insignificant similarity against several templates, the threading algorithm is helpful to predict its structure. If the boundary, location and conformation of each of the folds in a target sequence are precisely marked, then its structure can be very well predicted by the threading algorithm. For its simplicity and reliability, TBM is currently the most widely employed methodology to predict accurate protein structures. A TBM algorithm usually involves several steps, namely, template search, target-template alignment, model building, loop modelling, modelling of side-chains, model assessment and model refinement, as shown in the Fig. 1.5 and the tools and servers normally employed for each of these steps are enlisted in the Table 1.2.

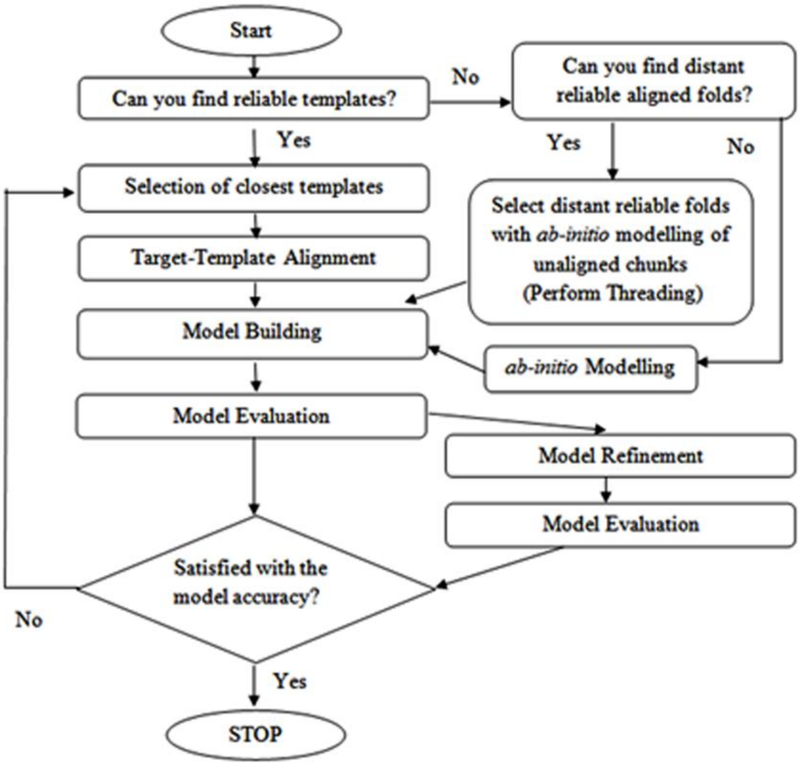

Start
Can you find reliable templates?
No
Can you find distant reliable aligned folds?
Yes
Yes
No
Selection of closest templates
Select distant reliable folds with *ab-initio* modelling of unaligned chunks (Perform Threading)
Target-Template Alignment
Model Building
ab-initio Modelling
Model Evaluation
Model Refinement
Model Evaluation
Satisfied with the model accuracy?
No
Yes
STOP

Fig. 1.5 Schematic representation of protein structure prediction algorithms.

1.6.1 Identification of suitable templates

The reliable templates are normally screened through their sequence identity and sequence similarity scores against a target sequence. Although the sequence identity score computes the fraction of target residues that are identical and aligned to the corresponding residues of the template sequence, it is not beneficial to optimally align the protein sequences. It is because a protein structure robustly tolerates a mutation in its sequence when its amino acid is substituted with a chemically equivalent residue sharing the similar charge or hydrophobicity and the sequence similarity score comes handy to align the protein sequences. Thus the algorithms screening the correct templates for a target sequence routinely employ several amino acid substitution matrices like PAM (point accepted mutation, Dayhoff et al. 1978), BLOSUM (blocks substituion matrix, Henikoff & Henikoff 1992) and Gonnet (Gonnet et al. 1994), gap penalty schemes like the one employed by PSI-BLAST (Altschul & Erickson 1986), Bayesian penalty (Lathrop et al. 1998) and the variable gap penalty scheme (Madhusdhan et al. 2006) and sequence identity. The best set of these high-scoring templates maximally spanning the target sequence is used to construct its model. It has already been shown that a target protein sequence sharing atleast 40% sequence identity with the available template(s) can predict model structure with accuracy comparable to a medium accuracy NMR structure or a low resolution X-ray structure (Tramontano & Morea 2003).

Moreover, it is also observed that some templates are promiscuous enough to show false positive sequence similarity with several targets and it becomes a major problem especially when the target contains structurally conserved segments (Jauch et al. 2007). A mutated template residue can be aligned with a target residue to show a biologically irrelevant sequence similarity score which is termed as false positive sequence similarity. However, a single template that shares a significantly high sequence similarity with several target sequences and can be easily employed to construct several different target models is not always available (Battey et al. 2007; Kryshtafovych et al. 2007; Kryshtafovych et al. 2009; Mariani et al. 2011). Hence, it is always required to search the best possible template for a target sequence and construct its accurate structure.

The template search algorithms are categorized in two groups. The first category includes the pairwise sequence comparison methods like basic local alignment search tool (BLAST, Altschul et al. 1990) and fast alignment (FASTA, Pearson 1990) which construct pairwise alignment of a target against the templates (Brenner et al. 1998). Pairwise sequence alignment slithers the two considered protein sequences on one-another to identify the segments that may share an evolutionary relationship with substantial functional, structural and conformational similarity. The second category includes the methods that evaluate the target-template sequence profiles to search the correct templates for a target sequence. Sequence profile is a powerful sequence comparison methodology that is competent enough to find even the distantly related templates. For a target sequence profile, computed as per the considered template scoring parameters through the sequence database, this methodology scans the PDB database to find the reliable hits and further iteratively employs each of these resultant hits to search all the significant templates by employing the position specific scoring matrix (PSSM; Henikoff & Henikoff 1997). The PSSM represents the amino acid frequency at each position in the target-template sequence profile and it estimates the conservation probability of every single target residue at every single profile position (Rychlewski et al.

2000). PSSM significantly improves the sensitivity of such sequence profile based template search algorithms (Teichmann et al. 2000) like position specific iterative BLAST (PSI-BLAST, (Altschul & Koonin 1998), HMMER (Eddy 1998), intermediate sequence search (ISS, Teichmann et al. 2000), HHPred or HHSearch (Söding 2005), comparison of multiple alignments (COMA, Margelevicius and Venclovas 2010) and domain enhanced lookup time accelerated BLAST (DELTA-BLAST, Boratyn et al. 2012).

Although the sequence profile based methodologies like PSI-BLAST are more accurate than BLAST, the remotely related homologues are not effectively screened (Gonzalez & Pearson 2010) for a target sequence. In contrary, the profile-profile alignment (PPA) method compares the sequence profile of target with that of template to compute the PSSM or position-specific degree of conservation for each of the target residue against the templates. As this method compares the target-template sequence profiles, it is more powerful than BLAST or PSI-BLAST (Jaroszewski et al. 2000; Roland 2006). HHPred further increases the efficiency of these profile based template search algorithms (Söding 2005).

To construct an optimal target-template alignment (Knudsen & Miyamoto 2003; Marko et al. 2007; Wohlers et al. 2010; Kuziemko et al. 2011; Barbato et al. 2012; Yoon 2014), protein secondary structure prediction tools like PSI-BLAST based secondary structure prediction (PSIPRED; Jones 1999) are employed (Pirovano et al. 2007). However, assignment of an incorrect secondary structure to a target segment results in the erroneous template screening and it further leads to incorrect placement of gaps in the target-template alignment which can even dissect the secondary structure segments of a target sequence and result in a biologically futile target-template alignment.

1.6.2 Template selection

Among the templates screened for a target, the reliable structures are selected to construct accurate target models (Greer 1980; Murphy et al. 1988; Ruan et al. 1994; Lewis et al. 2002; Tress et al. 2005; Tress et al. 2007; Tress et al. 2009; Taylor et al. 2014) through many parameters like the presence of a specific ligand, specific microenvironment of source organism of the template, phylogeny relationship among the target and template sequences, resolution and reliability factor (R-factor), E-value and sequence identity (Srinivasan et al. 1993; Sanchez & Sali 1997; Navaratnam et al. 1998; Reva et al. 2002; Nguyen et al. 2011).

Template resolution is a criterion to preferentially select the high resolution templates among all the available hits for constructing the target models (Srinivasan et al. 1993). R factor or the Reliability factor is another such criterion to define the accuracy of a template and it estimates the structural agreement between the template and its experimentally solved diffraction data. As it quantifies the structural deviation between the template and its ideal conformation expected for its crystallographic diffraction data, its lower value ideally implies the better quality of a protein structure solved through the X-ray crystallography. However, for the NMR structural ensemble of a protein, the degree of topological convergence or precision is computed through their RMSD after superimposition and a NMR structure does not have any direct measure of the resolution (Snyder et al. 2005; Montelione et al. 2013). It is often suggested that the accuracy of a NMR structure corresponds to a 2 - 3Ǻ resolution X-ray crystal structure (Mao et al. 2011; Mao et al. 2014).

E-value is a factor to estimate the number of hits one can expect to occur only by chance for a target sequence (Sanchez & Sali 1997). It may also be considered as the random background noise existing between sequence matches while searching the database of a particular size. Although the template with the lowest E-value is mathematically the best possible match for a target, consideration of similar analogy to select other good templates on the orderly basis of their E-value scores is purely a coincidence.

Presence of a specific ligand or a microenvironment in the source organism of a template is an important criterion to be considered for predicting the target protein structure (Navaratnam et al. 1998). If the objective of model prediction is to study the active sites responsible for its biological function, it is often advised to consider the templates that have similar pH, solvent environment, ligand(s), and quaternary interactions as probably present in the source organism of a considered target sequence. For example, if we want to study a sodium ion receptor in a cell, then it is advised to consider the templates that perform similar function in similar solvent conditions for constructing its best possible model. However, if the predicted target model is required to study its docking with a specific ligand, then such a constraint need not be considered.

Phylogeny relationship of the templates is also screened against the target sequence for selecting the best structures as it is always advised that the evolutionarily related templates justify the biological significance and evolutionarily conserved nature of every single target residue in its predicted structure (Reva et al. 2002).

Sequence identity is an unreliable measure to select the correct templates for a target sequence (Nguyen et al. 2011), as its score is correct only when the considered target-template alignment is biologically meaningful. A considerably higher target-template sequence identity is an indicator of a good template. However, when two templates have almost equivalent as well as lower sequence identity scores for a target sequence, selecting the best template becomes difficult. Such seemingly equivalent sequence identity among different templates is normally the result of homoplasy (The case of evolutionarily distinct sequences that have different random mutations, although it appears to be parallel or convergent in evolution on the basis of their sequence identity).

1.6.3 Constructing a target-template alignment

An alignment making the best use of complete biological information encrypted in templates is mandatory to build the best model for a target sequence. Many alignment algorithms based on Smith-Waterman (Local alignment) and Needleman-Wunsch (Global alignment) are currently available and employing the best of these methodologies is essential to accurately align the target-template sequences for constructing the best possible target model. All these alignments are normally evaluated for the number, length and location of gaps to improve its accuracy as the incorrectly placed gaps dissecting the conserved core of templates can sometimes be manually corrected by employing the conserved structural information of templates available in the dictionary of protein secondary structure (DSSP, Kabsch & Sander 1983). Constructing the target-template alignment is easy when their sequence identity is above 40% (Moult 2005), although alternative alignments constructed with different scoring schemes (Topf et al. 2006) are also employed to compute the best-scoring and biologically significant target-template alignment (Cozzetto et al. 2008). However if pairwise target-template identity is comparatively lower, an optimal alignment can result in a better model topology. A misalignment of just a single residue can result in an almost 4Ǻ structural deviation in the predicted model. Several algorithms like CLUSTALW (Jeanmougin et al. 1998), PRALINE (Heringa 1999), multiple sequence comparison by log-expectation (MUSCLE, Edgar 2004) and TCOFFEE (O'Sullivan et al. 2004) have been traditionally employed to construct the best possible target-template alignment.

1.6.4 Model building

TBM algorithm employs the target-template alignment file to extract the structural information of the template residues for the corresponding target residues and to construct the target protein model. The model building algorithms have been grouped into several categories viz. Rigid-body assembly that constructs the target model from the structural framework of aligned segments of the selected templates through the alignment file (Blundell et al. 1987; Sutcliffe et al. 1987; Topham et al. 1993; Guex & Peitsch 1997; Schwede et al. 2003; Kopp & Schwede 2004), Segment matching that considers a target sequence as continuous sequence of hexapeptides and constructs its Cα backbone model by concatenating the best set of hexapeptide structural segments of the templates (Jones & Thirup 1986; Unger et al. 1989; Bruccoleri & Karplus 1990; Levitt 1992) and the optimal satisfaction of the structural restraints that models the target through the distance map of template(s) on the basis of their alignment (Sali & Blundell 1993).

Protein modelling methodologies are also categorized as Cα backbone, loop and side-chain construction algorithms. The Cα backbone construction methodology copies backbone topology of the template residues for the equivalent target residues as per their alignment to further add the backbone atoms through algorithms like Maxsprout (Holm & Sander 1991), powerful chain restoration algorithm (PULCHRA, Rotkiewicz & Skolnick 2008) and reconstruct atomic model (REMO, Li & Zhang 2009), and add side-chain atoms by using the algorithms like side-chains with a rotamer library (SCWRL, Canutescu et al. 2003). Loop modelling algorithms consider the loop segments as insertion or deletion (INDEL, Fiser & Sali 2003) and model them through the *ab-initio* methods (Moult & James 1986; Bruccoleri & Karplus 1987; Fidelis et al. 1994; Fiser et al. 2002; Xiang et al. 2002; Fiser & Sali 2003; Zhang et al. 2003; Park et al. 2011; Park & Seok, 2012; Tyka et al. 2012) and the database search methods (Lee et al. 2010; Subramani & Floudas 2012; Bonet et al. 2014).

Side-chain modelling algorithms add side-chains to a protein backbone model while keeping it devoid of any atomic clash. Side-chains are added by simply replacing the target residue in its structure with the corresponding residue of the selected templates (Chothia & Lesk 1986; Sutcliffe et al. 1987) or by substituting the side-chain conformers (rotamers) to structurally satisfy the stereo-chemical and energetic constraints (Smith et al. 2007) through the VDW exclusion test (Ponder & Richards 1987; Xiang & Honig 2001). This algorithm is especially helpful for adding side-chains to the Cα backbone target model that is constructed by integrating the structural information of several different templates (Bower et al. 1997).

1.6.5 Model evaluation

It is usually observed that the models predicted through any of the modelling algorithms have several atomic clashes. This step assesses all the predicted models for the stereo-chemical and topological errors to select the accurate structure. Two types of model evaluation schemes are commonly employed. While the first scheme includes the assessment measures like RMSD (Martin et al. 1997), longest continuous segments (LCS, Zemla 2003), global displacement test (GDT, Zemla 2003) and TM_Score (Xu & Zhang 2010) to screen the predicted model that is structurally closest to the considered template(s), the second methodology evaluates a protein model on the basis of its alignment with the considered template(s) through the local global alignment (LGA) analysis (Zemla 2003). This second category also includes several other assessment measures like Main-chain reality score (Engh & Huber 1991), VERIFY3D (Luthy et al. 1992), ERRAT (Colovos & Yeates 1993), protein structure check (PROCHECK, Laskowski et al. 1993), protein structure analysis (PROSA, Sippl 1993), atomic non-local environment assessment (ANOLEA, Melo & Feytmans 1998), evaluation of protein structure (EVA, Koh et al. 2003), correct structural localization of hydrogen bonds and side-chain rotamers (Chakrabarti & Pal 2001), MolProbity (Davis et al. 2007), distance-scaled, finite ideal-gas reference (DFIRE) and disorder prediction (DISOPRED, Jonathan et al. 2004) to evaluate the core biophysical properties and structural topology of a protein model (Kryshtafovych & Fidelis 2008). Even the assessment scores viz. molecular probability density function (MOLPDF), discrete optimized potential energy (DOPE) score, GA341 and normalized DOPE score (Z_Score) employed by MODELLER also fall in this second category (Shen & Sali 2006). All these model evaluation schemes have been frequently employed by several protein modelling algorithms like LEE, rapid protein threading by operation research (RAPTOR), ZICO, FAIS@HGC, FIEG, DOMFOLD, DISOCLUST, GS-META model quality assessment program (MQAP), threading assembly refinement (TASSER), ZHOU-SPARX-, 3D-JIGSAW, MUFOLD, PLATO and PRECORS. Even the model clustering algorithms like structure picker (SPICKER) and CIRCLE, mutually comparing all the predicted models, have been employed to select the representative model that is structurally most similar to all the other constructed models (Zhang & Skolnick 2004). Although these measures are designed to evaluate different structural features of the generated set of protein models (Xu & Zhang 2010; Cao et al. 2015), the accurate model is not consistently selected (Manavalan et al. 2014). Moreover, majority of these measures do not unanimously select a single target model as the accurate structure for any of its individual domains or for its overall correct topology. An increased sampling becomes yet another confusing step especially when the target sequence contains multiple structural domains wherein the linker segments connecting these domains are even independently constructed (Shatnawi & Zaki 2015).

1.6.6 Model refinement

Several unfavorable and local steric clashes are often found even in the top-scoring model predicted for a target sequence (Arendall et al. 2005). These atomic clashes are strongly correlated to the incorrect local topology of the protein folds and are normally absent in an accurate model. To improve the structural topology, the top-scoring target model is energetically refined or relaxed through the increased model sampling (Zhang 2008) to remove all the energetically unfavorable atomic contacts (Hao et al. 2012). As the native conformation of a protein sequence exists at its lowest energy conformation, the energetic refinement methods attempt to search the global minima of the energetic landscape for a target sequence. Though being equivalent to the *ab-initio* modelling methodologies, the energetic refinement step is used by the TBM algorithms to extensively sample the target model for further improving its structural topology.

Although MC based energetic refinement algorithms have been improved several times, the simulation is not steered towards the global minima and is perplexed in the sampling landscape (Liang & Grishin 2002; Qian et al. 2004). It allows complete model to relax in a physically realistic all-atom force field and improves the target conformation both in terms of backbone topology and side-chain placement (Qian et al. 2007). However, it usually does not drive the target model away from the selected template(s) towards its native structure and fails to consistently improve its topology (Chen & Brooks 2007; Moult et al. 2007). Hence, several different algorithms viz. simulated annealing (SA, Holak et al. 1988), genetic algorithms (Tuffery et al. 1991), a combination of MC and SA (Holm & Sanders 1992), MC simulation (Eisenmenger et al. 1993), mean field optimization (Koehl and Delarue 1994), neural network with SA (Hwang & Liao 1995), combinatorial search algorithms (Subbiah & Harrison 1989), a combination of SA and GA as SAGA (Bayley et al. 1998; Standley et al. 1998) and dead-end elimination theorem (Looger & Hellinga 2001) have been developed for efficiently searching the conformational space of a target sequence and constructing its accurate model. All these refinement algorithms are found to have diligent conformational search protocol and their modelling accuracy is limited only due to their inaccurate energy function (Jacobson et al. 2002).

The model refinement step has been improved to a great extent although it is still obstructed by some logical problems. It is usually observed that an energetically refined model is not structurally closer to its native conformation or the refinement step is unable to consistently improve topological accuracy of the predicted target models. Though many refinement methodologies have been developed including the ones that only sample the side-chains of a protein model (Wallner 2014), the cellular *in-vivo* protein folding process is still not correctly implemented computationally in our structural refinement algorithms (Huang et al. 2015). The single long model sampling perturbs a target model for the specified number of steps. However, if a structurally incorrect model is constructed during this sampling, the further sampling path becomes completely erroneous and does not lead to structurally improved target model in comparison to the initial target structure.

Table 1.2 Mostly used protein modelling servers and tools (Brooks et al. 1983; Russell & Barton 1992; Sali & Blundell 1993; Pearlman et al. 1995; Altschul et al. 1997; Eddy 1998; MacKerell et al. 1998; Shindyalov & Bourne 1998; Scott et al. 1999; Notredame et al. 2000; Melo et al. 2002; Fiser & Sali 2003; Kosinski et al. 2003; Schwede et al. 2003; Bateman et al. 2004; Edgar 2004; Edgar & Sjolander 2004; Jonathan et al. 2004; Söding 2005; Biegert & Soding 2009; Lawrence & Michael 2009).

Usage	S. No.	TOOL	Website Link
Template Search	1.	PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
	2.	TOPITS	http://www.embl-heidelberg.de/ predictprotein/submit_adv.html
	3.	HMMER	http://bio.ifom-firc.it/HMMSEARCH/
	4.	CS-BLAST	http://toolkit.tuebingen.mpg.de/cs_blast
	5.	HHPred /	http://toolkit.tuebingen.mpg.de/hhpred
	6.	FUGUE	http://www-cryst.bioc.cam.ac.uk/ ~fugue/prfsearch.html
	7.	Threader	http://bioinf.cs.ucl.ac.uk/threader/
	8.	3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/
	9.	PFAM	http://www.sanger.ac.uk/Software/Pfam/
	10.	PHYLIP	http://evolution.genetics.washington.edu/phylip.htm
	11.	DALI	http://www2.ebi.ac.uk/dali/
Target-Template Alignment	1.	CLUSTALW	http://www.ebi.ac.uk/clustalw/
	2.	HMMER	http://bio.ifom-firc.it/HMMSEARCH/
	3.	STAMP	http://bioinfo.ucr.edu/pise/stamp.html
	4.	CE	http://cl.sdsc.edu
	5.	DSSP	http://bioweb.pasteur.fr/seqanal/ interfaces/dssp-simple.html
	6.	COMPASS	ftp://iole.swmed.edu/pub/compass/
	7.	MUSCLE	http://www.drive5.com/muscle
	8.	SALIGN	http://www.salilab.org/modeller
	9.	TCOFFEE	http://www.ch.embnet.org/software/TCoffee.html

Model Building	1.	COMPOSER	http://www-cryst.bioc.cam.ac.uk
	2.	SwissModel	http://swissmodel.expasy.org/
	3.	3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djigsaw/
	4.	MODELLER	http://salilab.org/modeller/
	5.	ICM	http://www.molsoft.com/bioinfomatics/
	6.	CONGEN	http://www.congenomics.com/congen/ congen_toc.html
Loop Modelling	1.	MODLOOP	http://alto.compbio.ucsf.edu/modloop//modloop.html
	2.	ARCHDB	http://sbi.imim.es/cgi-bin/archdb/loops.pl
	3.	Sloop	http://www-cryst.bioc.cam.ac.uk/ ~sloop/Browse.html
Modelling of Side-chains	1.	WHAT IF	http://swift.cmbi.kun.nl/whatif/
	2.	SCWRL	http://dunbrack.fccc.edu/SCWRL3.php
Model Assessment	1.	PROCHECK	http://www.biochem.ucl.ac.uk/~roman/ procheck/procheck.html
	2.	PROSA II	http://www.came.sbg.ac.at/
	3.	ANOLEA	http://protein.bio.puc.cl/cardex/servers/
	4.	AQUA	http://nmr.chem.uu.nl/users/jurgen/Aqua/server
	5.	BIOTECH	http://biotech.embl-heidelberg.de:8400
	6.	ERRAT	http://www.doe-mbi.ucla.edu/Services/ERRAT/
	7.	VERIFY3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/
	8.	EVA	http://cubic.bioc.columbia.edu/eva/
	9.	DFIRE	http://sparks.informatics.iupui.edu/ yueyang/server/dDFIRE/
	10.	DISOPRED	http://bioinf.cs.ucl.ac.uk/disopred/
Model Refinement	1.	AMBER	www.amber.scripps.edu
	2.	GROMOS	http://www.igc.ethz.ch/gromos/
	3.	CHARMM	http://www.charmm.org/

1.7 CASP

For assessing the significant progress and accuracy of protein structure prediction algorithms, a community wide blind test entitled critical assessment of structure prediction

(CASP) is organized every two years since 1994 (Dunbrack et al. 1997; Mariani et al. 2011). CASP assesses the modelling accuracy of all the structures predicted by the participants for the target protein sequences whose structures are experimentally solved and kept frozen till the end of the test. Target protein sequences with cloning artifacts, protein chains with significant structural deviations (>3.5A°) or the structures majorly influenced by the physically unreasonable crystal packing are not considered as the CASP targets (Clarke et al. 2007). Further, the target sequence segments are considered as sequence or structural domains only if they are covered by the existing templates with a significantly high alignment score (Tress et al. 2009). Moreover, the target sequences are also categorized as FM, FM/TBM overlap, TBM and High Accuracy (Tress et al. 2009; Kinch et al. 2011; Moult et al. 2014). While the FM category includes only the target domains that are not homologous or share a minimal homology to any existing template that is not considered as the current CASP target, the TBM group solely includes the targets that share a significant homology with the existing templates which are also not the current CASP targets. Further in the intermediate difficulty group FM/TBM group comprises the targets that have sequence segments individually assigned to each of the FM and TBM assessment categories (Li et al. 2015). Lastly, the High Accuracy category includes only the target domains for which the top-scoring and reliable templates with significantly high coverage are easily available and which can be readily modelled with GDT-TS score higher than 80 through the available templates (Tress et al. 2009). Since its first round, CASP has been considering a diverse set of target sequences (www.predictioncenter.org), as enlisted below in Table 1.3.

Table 1.3 Evolution Statistics of CASP, Targets and its Domains. NF: New Fold (*ab-initio* prediction), CM: Comparative Modeling, FR: Fold Recognition (Murzin & Hubbard 2001; Lisa et al. 2003; Moult et al. 2005; Moult et al. 2007; Moshe et al. 2009; Leaver-Fay et al. 2011; Taylor et al. 2014).

CASP# (YEAR)	NUMBER OF TARGETS (DOMAINS)	*AB-INITIO*/ THREADING PREDICTIONS	COMPARATIVE MODELING
1 (1994)	31	-	-
2 (1996)	42	22	15
3 (1998)	43	15	28
4 (2000)	40 (56)	16 NF and 36 FR Domains	18 Domains
5 (2002)	55 (80)	5 NF and 24 FR Domains	29 CM and 22 CM/FR Domains
6 (2004)	64 (90)	37 FR and 10 NF Domains	46 Domains
7 (2006)	95 (123)	15 Domains	108 Domains
8 (2008)	121 (165)	11 FR and 3 NF Domains	151 Domains
9 (2010)	116 (142)	27 Domains	115 Domains
10 (2012)	97 (128)	16 Domains	112 Domains
11 (2014)	93 (126)	45 Domains	81 Domains

As per the Table 1.3, various CASP rounds have successively evolved a lot in comparison to CASP1 in terms of the nature of the considered protein sequences. The number of CASP target proteins has constantly increased over the years, although the number of novel fold protein targets (supposed to be modelled through the *ab-initio* algorithm in CASP) has not greatly increased. However, as it is well understood that the accuracy of TBM algorithm is significantly higher than that of FM methodology, the proportion of TBM target domains has significantly increased in the number of CASP targets. As clear from the table 1.3, more and more targets are now being considered as sequential or structural TBM domains. The exquisite CASP development over the years is actually an indicator of the successful development of protein structure prediction algorithms (Moult 2005; Guo et al. 2008). Recently, the blind protein structure prediction tests CASP8, CASP9 and CASP10 were held in the years 2008, 2010 and 2012 respectively and here also the target domains have been mostly assigned under the TBM category.

Although the researchers have developed several TBM algorithms, most of the predicted models are still too divergent from their native structures. Among all the discussed protein modelling problems, the ones caused due to incorrect selection and combination of templates, and inaccurate assessment guided increased model sampling are the major hurdles to develop robust modelling methodologies. Considering the fact that protein structure prediction algorithms can successfully bridge the sequence-structure gap, the quest for an improved protein structure prediction methodology is further strengthened.

1.8 Objectives of the current research

Although every step of a TBM algorithm has some logical limitations, the protein modelling errors are mainly caused due to the consideration of inaccurate templates and the employment of inefficient sampling algorithm. Thus the objective of the present work was:

- To study the template selection measures and to identify the scoring schemes for selecting the best set of templates for a target sequence.
- To develop the model sampling strategy through consideration of best set of model assessment measures for constructing the optimal target structure.

Chapter II

Materials and Methods

Chapter II

2.1 Overview

This chapter describes the methods used in this study. The objective of the study is to develop a protein structure prediction algorithm that selects the best set of templates for a target sequence and optimally searches the conformational space of a target sequence to construct its accurate structure. We develop template selection and ranking algorithm, and apply it to construct the protein models for the recent CASP targets. HHPred is employed to screen the template structures available for a target sequence.

To estimate and evaluate the plausible modelling accuracy of a template for a target sequence, the pairwise alignments and multiple sequence alignment (MSA) of the screened templates are constructed. Robust set of diverse template scoring parameters is developed and then employed to assess all these alignments for selecting the best combination of templates. Through the selected template set, the target conformation is constructed with MODELLER9.9 (Sali & Blundell 1993). Model sampling is further employed to minimize the number of atomic clashes that are normally present in a predicted structure (Fiser & Sali 2003; Topf et al. 2006). All the sampled decoy structures are then assessed through the best set of model assessment measures and the top-scoring target conformation is selected as the best predicted structure. The generated model is further evaluated against the actual native target conformation in comparison to the best predicted CASP structure.

2.2 Target selection

To develop and test the accuracy of a modelling algorithm, the CASP TBM targets with at least one high accuracy (TBM-HA) domain are selected for the study. CASP, a community wide blind test, assesses the accuracy of these predicted TBM-HA domain structures to evaluate the modelling accuracy of a TBM algorithm. CASP considers a target as a TBM-HA domain if it can be successfully modelled with a GDT-TS accuracy of 80.00 through the available set of templates. Altogether, 21 CASP8 targets with 33 domains, 35 CASP9 targets with 52 domains and 22 CASP10 targets with 31 domains are considered, as enlisted in the following Tables 2.1, 2.2 and 2.3 respectively.

Table 2.1 CASP8 target proteins along with the length and source organism.

Target	Length	Source Organism
T0388	174	*Homo sapiens*
T0390	126	*Homo sapiens*
T0396	102	African swine fever virus BA71V
T0398	292	*Bacillus Halodurans*
T0400	162	*Staphylococcus aureus*
T0402	139	*Listeria Innocua*
T0404	110	*Anabaena variabilis*
T0418	222	*Bacteroides fragilis*
T0422	357	*Homo sapiens*
T0423	110	*Agrobacterium tumefaciens C58*
T0426	283	*Homo sapiens*
T0428	267	*Cryptosporidium parvum*
T0432	130	*Homo sapiens*
T0435	151	*Homo sapiens*
T0438	439	*Porphyromonas Gingivalis*
T0442	269	*Agrobacterium tumefaciens*
T0444	326	*Homo sapiens*
T0447	542	*Thermotoga maritima*
T0458	107	*Streptomyces avermitilis*
T0470	223	*Bacillus thuringiensis*
T0499	56	*Escherichia coli*

Table 2.2 CASP9 target proteins along with the length and source organism.

Target	Length	Source Organism
T0521	179	*Plasmodium falciparum*
T0522	134	*Sinorhizobium meliloti 1021*
T0523	120	*Burkholderia thailandensis*
T0528	388	*Rhodopseudomonas palustris CGA009*
T0530	115	*Bacillus subtilis*
T0538	54	*Nostoc sp. PCC 7120*
T0541	106	*Methanosarcina acetivorans*
T0559	69	*Bacteroides vulgatus ATCC 8482*
T0560	74	*Bacteroides thetaiotaomicron*
T0563	279	*Shewanella oneidensis MR-1*
T0566	156	*Plasmodium falciparum*
T0567	145	*Escherichia coli CFT073*
T0570	258	*Parabacteroides distasonis atcc 8503*
T0580	105	*Streptococcus pneumoniae TIGR4*
T0586	125	*Listeria innocua Clip11262*
T0589	465	*Nostoc sp. PCC 7120*
T0594	140	*Plasmodium falciparum*
T0596	213	*Nitrosomonas europaea ATCC 19718*
T0599	399	*Bacillus anthracis str. Ames*
T0600	125	*Chromobacterium violaceum ATCC 12472*
T0601	449	*Pseudomonas aeruginosa*
T0602	123	*Yersinia enterocolitica*
T0605	72	*Homo sapiens*
T0611	227	*Marinobacter aquaeolei vt8*
T0613	287	*Rhodopseudomonas palustris Cga009*
T0614	135	*Homo sapiens*
T0619	111	*Haloarcula marismortui*
T0620	312	*Homo sapiens*
T0626	283	*Pseudomonas syringae pv. tomato str. DC3000*
T0629	216	*Enterobacteria phage T4*
T0632	168	*Bacillus halodurans*
T0634	140	*Pelobacter carbinolicus*
T0635	191	*Legionella pneumophila subsp. pneumophila str.*
T0636	336	*Burkholderia pseudomallei*
T0640	250	*Bacteroides thetaiotaomicron*

Table 2.3 CASP10 target proteins along with the length and source organism.

Target	Length	Source Organism
T0645	537	*Bacteroides vulgatus ATCC 8482*
T0650	346	*Listeria monocytogenes serotype 4b str. F2365*
T0657	154	*Homo sapiens*
T0659	85	*Ruminococcus gnavus*
T0662	79	*Pseudomonas aeruginosa 2192*
T0663	205	*Peptoclostridium difficile 630*
T0664	540	*Bacteroides ovatus ATCC 8483*
T0674	340	*Staphylococcus aureus subsp. aureus Mu50*
T0675	75	*Homo sapiens*
T0689	234	*Parabacteroides distasonis ATCC 8503*
T0692	473	*Anabaena variabilis ATCC 29413*
T0708	196	*Pseudomonas putida KT2440*
T0712	223	*Bacteroides fragilis NCTC 9343*
T0714	88	*Homo sapiens*
T0716	71	*Homo sapiens*
T0721	301	*Bacillus anthracis str. 'Ames Ancestor'*
T0726	597	*Idiomarina loihiensis L2TR*
T0731	79	*Homo sapiens*
T0747	121	*Bacteroides thetaiotaomicron VPI-5482*
T0749	449	*Bacteroides uniformis ATCC 8492*
T0752	156	*Kribbella flavida DSM 17836*
T0757	247	*Spirosoma linguale DSM 74*

2.3 Template search

HMM based template search algorithm HHPred is employed to screen the reliable hits for the selected target sequences. HHPred uses several sequence and structural databases like PDB, structural classification of proteins (SCOP), InterPro, clusters of orthologous groups (COG) and PFAM to construct the target and template Hidden Markov Model (HMM) profiles through PSI-BLAST or HHblits (Remmert et al. 2012). It constructs the HMM profile of the target sequence and evaluates it against the pre-computed HMM profile of the templates. By comparing the target-template HMM profiles, it estimates the probability of conserved nature of an amino acid at every profile position by considering the entire local alignment ensemble and not only the top-scoring alignment (Hildebrand et al. 2009). HHPred also probabilistically estimates the correctness of every single amino acid pair aligned in the target-template HMM profiles to screen all the biologically significant templates for a target sequence (Sadowski & Jones 2007; Hildebrand et al. 2009; Peng & Xu 2010; Meier & Söding 2014). It considers the mutation probability of all the amino acids for insertions or deletions at specific locations (Söding 2005) as per the structural context (Meier & Söding 2015) and is accurate at screening even the distantly related templates for a target (Smith et al. 1997; Jones 1999; Gough et al. 2001; Meier & Söding 2015).

2.4 Template selection

The templates screened through HHPred are evaluated through their pairwise alignment against the considered target sequence. The MODELLER (Sali & Blundell 1993) align2d module (Madhusudhan et al. 2006) is employed to construct the pairwise alignments of all the selected templates against the target sequence. Although it is based on the dynamic programming algorithm, it also uses structural information of the template to construct its alignment with the target sequence. It employs a variable gap penalty function (Madhusudhan et al. 2006) that tends to place the gaps in the solvent exposed segments, in between two spatially close residues, in the curved segments of the main-chain and avoid gaps within the secondary structure segments to optimally align the target sequence with the considered template through the BLOSUM62 scoring matrix. It employs the dynamic programming matrix along with a variable gap penalty option to inset the minimal possible gap length at any location on the basis of the structural context of an insertion or deletion to construct the optimal target-template alignment.

The BLOSUM matrices are tailored to detect the significant similarity among the protein sequences with different levels of evolutionary divergence. While, the BLOSUM80 matrix is derived from the alignments of sequences that have no more than 80% identity, the BLOSUM45 matrix is computed from the alignments of sequences with no more than 45% identity. These matrices are employed for estimating the similarity of closely related and evolutionarily divergent protein sequences respectively (Henikoff & Henikoff 1992). However, the BLOSUM62 ideally implies an average of these matrices and is constructed from the alignments of sequences with no more than 62% identity. It is reasonably efficient over a relatively broad range of evolutionary substitutions and is used in this study to sensitively detect the weakest as well as meaningful similarity among the target-template protein sequences (Henikoff & Henikoff 1993; Styczynski et al. 2008; Pearson 2013). This BLOSUM62 residue substitution score is therefore employed along with several other scoring parameters viz. average proportion of mismatched hydrophobic, hydrophilic and identical residues, sequence coverage span, an average gap length available in an alignment considering all of its insertions or deletions (INDELs), proportion of gaps existing in the alignment, affine gap penalty score and the proportion of mismatched residues existing in a target-template alignment to assess all the pairwise alignments constructed for a target sequence. All these sequence similarity based template scoring parameters assess the biological credibility of every single target residue with the corresponding template residue in the target-template alignment.

The MSA of the selected templates is constructed through Salign_multiple_struc module (Marti-Renom et al. 2004; Madhusudhan et al. 2009) of the MODELLER (Sali & Blundell 1993). This module mutually superimposes the templates under a defined RMSD threshold of 3.5Ǻ which is the constraint to decide the maximal absolute distance deviation between the corresponding residues of the two templates so that these residues can be considered as the correctly aligned set of amino acids. Minimal such average distance deviation between any two templates implies the extent of their evolutionary relatedness and the correctness of their alignment. Aligned blocks of template residues are iteratively aligned until all the template motifs are structurally aligned within a predefined RMSD threshold cutoff and it shows the presence of structurally conserved motifs (Yang & Honig 1999; Jaroszewski et al. 2000; Al-Lazikani et al. 2001; Reddy et al. 2001). MODELLER employs these structural segments (sharing a minimal RMSD deviation) of the selected templates to construct their dendrogram for further computing their progressive MSA alignment. The MODELLER progressive alignment strategy constructs the template MSA through a combination of pairwise alignments. It starts with the most similar template to the target sequence and then progressively adds the distant templates without altering the sub-alignments of the considered template folds sharing a minimal RMSD deviation. The MODELLER align2d_mult module (Madhusudhan et al. 2006) is then used to append the considered target sequence to the computed MSA alignment. It considers the structural topology of the aligned templates to optimally align the target sequence. As it also employs the structural constraints and the variable gap penalty function (Madhusudhan et al. 2006) employed by the Salign module (Marti-Renom et al. 2004; Madhusudhan et al. 2009) and the dynamic programming methodology used by the align2d module, it also tends to optimally incorporate the gaps on basis of the structural context of the aligned template segments.

To select the best set of top-ranked templates that maximally cover the target sequence, the MSA is assessed through several scoring parameters viz. unique residues that are available only in a single template and not encoded in the other ones, average proportion of the mismatched hydrophobic residues, mismatched hydrophilic and the sequence identity, BLOSUM62 score of a hit against the target sequence with the seed template, additional target coverage over the seed template and structural topology of a hit against the seed template assessed in terms of TM_Score, GDT-TS, Cα backbone RMSD and the count of residues that are confined within 8Ǻ distance deviation.

Pairwise alignments of all the selected templates are evaluated against the selected templates through our in-house computer scripts written in "C" and "PERL" programming languages. These automated scripts rank the templates on basis of their pairwise alignments and then further rank their MSA alignment to select their best set for maximally covering the target sequence. The detailed methodology of these automated template ranking, selection and a combination script is further discussed later in Chapter III.

2.5 Model building

The target model is constructed through the MODELLER9.9 software that is a python based comparative modelling tool (Sali & Blundell 1993; Marti-Renom et al. 2000; Fiser & Sali, 2003). MODELLER technique structurally satisfies the spatial restraints of a target sequence on the basis of its alignment information with the selected templates (Havel &

Snow 1991). MODELLER (Sali & Blundell 1993) algorithm derives all the mutual distance and dihedral angle restraints of all the atoms of the target sequence residues through their alignment positions against the corresponding residues of the template sequence. MODELLER assumes that the distance and angle based restraints extracted from the selected templates are similarly applicable for the corresponding target sequence residues. For the unaligned target residues, MODELLER algorithm employs the *ab-initio* modelling protocol to fit these loop segments the best possible way so that the topology extracted from the template for the corresponding aligned target residues is not disturbed. For assessing the constructed target model, it quantifies the correlations between the equivalent Cα-Cα distances and the equivalent main-chain dihedral angles of the predicted model with the employed template by integrating them together for all the aligned and unaligned target segments into a molecular probability density function (MOLPDF). MOLPDF compares the structural features of the predicted target model with the spatial restraints extracted from the alignment and the template structure to estimate the atomic violations incurred in the target model. MOLPDF quantifies the degree of violating atomic distance restraints (both bonded and non-bonded) and dihedral angle restraints by considering the standardized thresholds of each of these restraints and its lower score implies the accuracy of a predicted model.

2.6 Model refinement

The predicted target model is subjected to an optimization procedure to refine its geometry and stereo-chemistry by increased MODELLER sampling. MODELLER samples the potential energy surface defined by a molecular mechanics force field for a target sequence to construct its minimal energy conformation (Shen & Sali 2006). MODELLER sampling structurally optimizes the first constructed target model as per its MOLPDF function in a way that the model minimally defies the employed set of distance restraints extracted from the selected template(s) (Braun & Go 1985) and the target-template alignment. MODELLER employs MD and conjugated gradient (CG) methodology along with simulated annealing (SA; Clore et al. 1986) to improve the topology of the first constructed target model by optimizing the DOPE energy function to construct an optimal structure through an increased model sampling. DOPE is an atomic distance-dependent statistical potential that is computed from a standard set of native protein conformations (Shen & Sali, 2006). DOPE employs the standard MODELLER energy function to estimate potential energy of a predicted model and is designed to select the best model from the decoy structures sampled for a target sequence.

2.7 Model evaluation

The generated set of protein models is lastly evaluated to select the best predicted structure. For the considered TBM targets of CASP dataset, the sampled set of target models are assessed against the considered template(s) to select the top-scoring conformation that accurately retains the template topology intact, as per their considered alignment. Several different model assessment methodologies viz. RMSD, GDT, LCS and TM_Score are employed for the study to select the correct near-native target conformation, as explained below in this section.

2.7.1 Root Mean Square Deviation (RMSD)

It is the average score of the squared distance differences between X, Y and Z coordinates of the model and the native structure (Martin et al. 1997). Its optimal value is calculated by structural superimposition of the templates through Kabsch algorithm (Kabsch 1978). It can be calculated only for the Cα atoms (*for assessing the model accuracy only for the Cα backbone*) or for all the atoms (*for assessing the overall model accuracy of the complete model*). It is computed through the following formula.

$$RMSD = \sqrt{\left(\sum_{i=1}^{L} d_i^2\right) / L}$$

Where, *L* is the number of residues encoded in a protein model and d_i is distance between the corresponding i^{th} pair of Cα atoms of the two structures. As two functionally as well as structurally different proteins can be superimposed to yield a lower RMSD score, its biological reliability is always doubtful. It is usually insensitive to evaluate the global topology as it equally weighs and considers all the model atoms (Carugo & Pongor 2001). So the topological mis-orientation at some local segments or the incorrect topology of a few residue chunks like loop regions in a model can result in its abruptly higher RMSD score (Zhang & Skolnick 2005; Zhang et al. 2005; Xu & Zhang 2010) and here the RMSD score cannot identify the topologically correct substructures of a model (Zhang & Skolnick 2004).

2.7.2 Global Displacement Test (GDT)

GDT measures the average percentage of the Cα residues of a predicted protein model that are present within the maximum pre-defined distance cutoff from the corresponding residues of its actual native conformation or the employed template through sequence-dependent or sequence-independent optimal superimposition of these structures (Zemla 2003; Jauch et al. 2007). The consideration of this number of model Cα residues as well as the individual distance deviation between each corresponding residue pair of the model and its template or the native structure empowers the Maxsub and TM_Score measures to efficiently evaluate the local as well as global structural similarity of a protein model (Zhang & Skolnick 2004). The GDT-Total Score (GDT-TS) and GDT- High Accuracy (GDT-HA) scores along with the maxcluster tool (Siew et al. 2000) are based on these Maxsub and TM_Score measures. However, the Maxsub score does not penalize the over-prediction or it does not penalize the residue pairs that are incorrectly superimposed and is not considered in our study (Söding 2005). The GDT-TS and GDT-HA scores are defined as:

$$\text{GDT\_TS:} \qquad (C\alpha_1 + C\alpha_2 + C\alpha_4 + C\alpha_8)/4$$

$$\text{GDT\_HA:} \qquad (C\alpha_{0.5} + C\alpha_1 + C\alpha_2 + C\alpha_4)/4$$

Where, $C\alpha_x$ refers to the percentage of a model's Cα atoms that are structurally localized within the maximum distance cutoff of xǺ from its actual experimental structure or the employed template (*as per the considered target-template alignment*) in one-to-one equivalent residue correspondence.

GDT_TS allows maximum distance deviation of 8Ǻ and is useful to distinguish the best predicted structure from the set of models generated for a difficult target for which the maximally covering templates are not easily available. For a simple target sequence (*where the correct templates are easily available*), the predicted model is normally found to be more accurate and the 4Ǻ distance deviation is used as the maximal allowed distance deviation between the Cα atoms of corresponding residues. GDT_HA score is found to significantly differentiate between the close model structures that have almost equivalent GDT-TS scores.

2.7.3 Longest Continuous Segment (LCS)

It is used to compute the longest continuous segments of a model structure that fall under the specified Cα RMSD cutoff to its actual native structure. Through LGA (Zemla 2003), this methodology computes standard RMSD, superimposed RMSD and GDT scores. As RMSD score may not be a reliable scoring parameter, GDT scoring becomes vital as it does not penalize the score of the complete target model for a few topologically incorrect target residues.

2.7.4 TM_Score

TM_Score is employed to estimate the conformational similarity of two structures as per their alignment, for the aligned as well as the paired residues. Based on the alignment, TM_Score rotation matrix is calculated for optimal superimposition of the two structures. Here, the scoring matrix is based on the features extracted from Voronoi tessellation. The best possible target model shows a TM_Score of 1.00 against its actual experimental structure and a structure with TM_Score greater than 0.5 is considered as a good model (Xu & Zhang 2010). It is calculated as:

$$TM - Score = \frac{1}{LM} \sum_{i=1}^{LA} \frac{1}{1 + (\frac{dist^i}{dist^0})^2}$$

Where, LM is the target sequence length, LA is the alignment length, $dist^i$ is the Cartesian distance between the ith pair of aligned target-template residues and $dist^0$ is the normalizing parameter. The $dist^0$ parameter is calculated as $1.24\sqrt[3]{LM\ -15}-1.8$ and it makes the TM_Score assessment independent of the target sequence length. TM_Score parameter is considered to assess the modelling accuracy of two target protein models (*with different lengths*) constructed through two different templates with varying lengths. It is the correct mathematical score that efficiently distinguishes between the correct and the bad models, unlike the other discussed model assessment scores that unanimously do not justify a single model as the accurate structure.

Chapter III

Development of Template Selection and Combination algorithm

Chapter III

3.1 Introduction

Template selection is considered as one of the most important steps of TBM algorithm as it is noticed that a model constructed through incorrect templates cannot be refined later and most of the modelling errors occur because of incorrect selection of templates for a target sequence. No template selection algorithm consistently screens the reliable hits for the targets. Even the modelling algorithms use different threading and TBM techniques together to construct protein models. However, most of these algorithms employ the correct templates along with the incorrect and unrelated structures and it leads to an incorrect target model. To construct accurate models, we try to develop an improved template selection and combination technique. We have screened the scoring schemes that are usually used to evaluate a template and we have standardized the process of template selection by a ranking scheme to select the correct set for a target sequence.

3.2 Developing the template ranking methodology

The developed template-ranking algorithm to rank and select the best set of templates for a target sequence is diagrammatically represented as a flowchart in the Fig. 3.1. T0388 target of the CASP8 TBM-HA dataset is considered for developing this algorithm. The HMM based template search algorithm, HHPred, with 8 successive MSA construction iterations and all the other default parameters, is employed to screen the top 100 templates for the considered target sequence. We select the top-ranked hit with the lowest E-value (available during the CASP) as the seed template. Through the functional details of this template, harnessed through PDB and HMMPFAM (Finn et al. 2011) databases, the other hits sharing a functional similarity with it and spanning atleast 75% of the target are selected as the candidate additional templates. To assess the reliability, the selected hits are ranked through their pairwise alignment information with the target, as explained below.

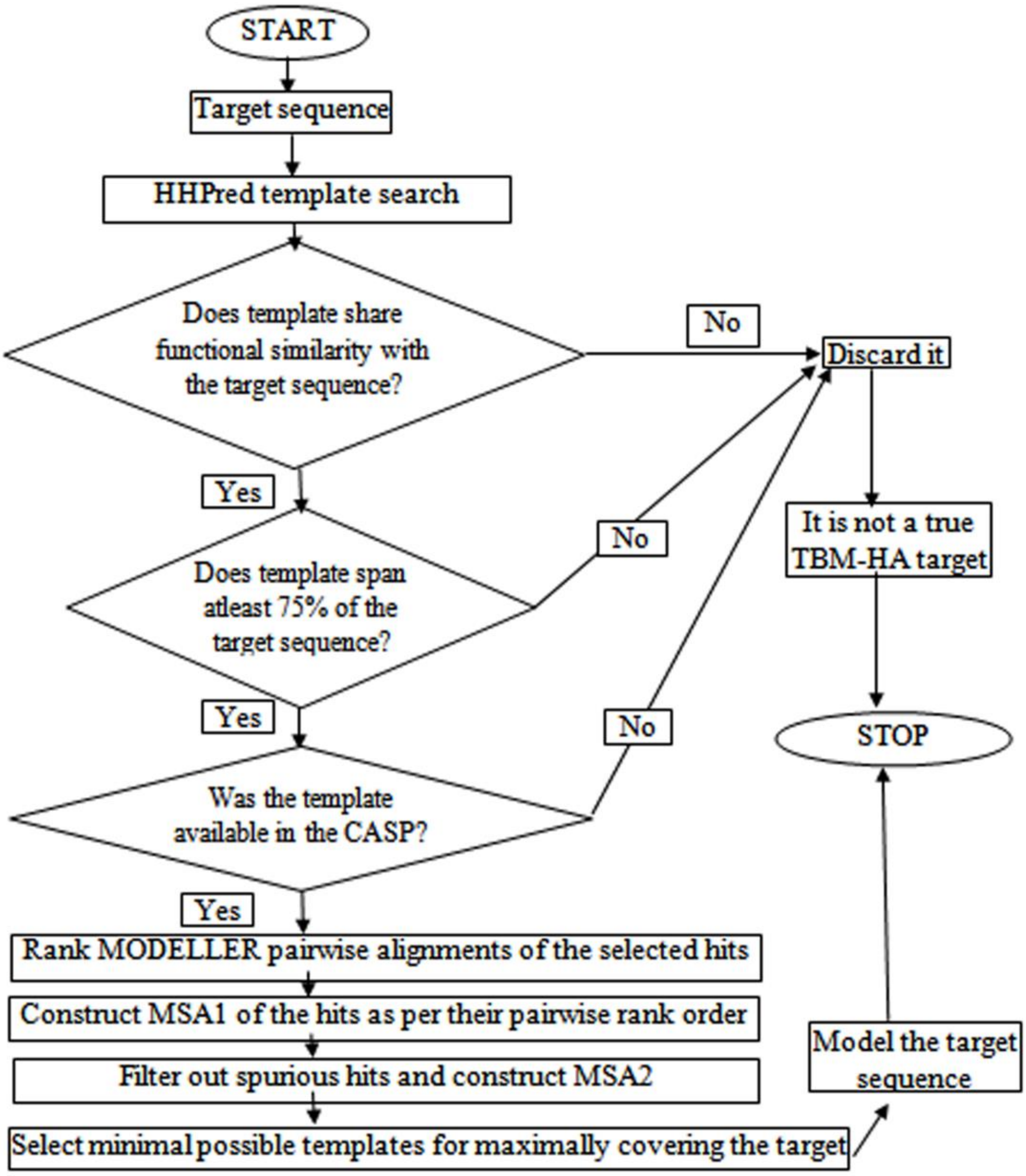

START
Target sequence
HHPred template search
Does template share functional similarity with the target sequence?
No
Discard it
Yes
Does template span atleast 75% of the target sequence?
No
It is not a true TBM-HA target
Yes
No
STOP
Was the template available in the CASP?
Yes
Rank MODELLER pairwise alignments of the selected hits
Construct MSA1 of the hits as per their pairwise rank order
Model the target sequence
Filter out spurious hits and construct MSA2
Select minimal possible templates for maximally covering the target

Fig. 3.1 Flowchart representing the template selection and combination algorithm.

3.2.1 Pairwise sequence alignment based template ranking

All the selected templates are individually aligned against the target sequence using MODELLER9v9 (Sali & Blundell 1993) to construct their pairwise alignments. Through these alignments, the selected templates are evaluated against the considered target sequence.

For the target-template residues aligned in a pairwise alignment, an average BLOSUM_score (Sc1) is computed through the BLOSUM62 residue substitution matrix (Henikoff & Henikoff 1992). The Sc1 score is calculated as decimal logarithm of the total BLOSUM62 score (BLS) of all the aligned target-template residues divided with the total number of residues (TR) encoded in the target and template sequences, as shown in the equation 1 below. Here, if a template shows a negative score, decimal logarithm of the modulus value is employed to compute its overall negative Sc1 score.

$$\boxed{Sc1 = \log_{10}(\frac{BLS}{TR})} \qquad - (1)$$

Further the IDENT_score (Sc2), an average fraction of mismatched hydrophobic residues, hydrophilic residues and sequence identity, is computed for an alignment, as shown in equation 2 below. Among a total of TR residues and as per the alignment, fraction of non-identical or mismatched hydrophobic residues and mismatched hydrophilic residues are orderly computed as AHR and AHYR scores. The rightly aligned hydrophobic and hydrophilic residues are important to define the accuracy of an alignment (Jefferys et al. 2010; Radzicka & Wolfenden 1988). Phe, Ala, Gly, Leu, Ile, Met and Val, and Ser, Thr, Gln, Asn, Asp, Glu and Arg are respectively considered as hydrophobic and hydrophilic residues after excluding the Cys, Trp, His, Tyr and Lys residues common in both these residue sets

(Radzicka & Wolfenden 1988). Sequence identity (ASID) score is computed as fraction of the TR residues that are identical in the alignment.

$$\boxed{Sc2 = \frac{(AHR + AHYR + ASID)}{3}} \quad - (2)$$

Coverage_span score (Sc3), is computed as the proportion of continuous segment of target residues spanned by the considered template. Sc3 score is the maximal target sequence length, including all the gaps, spanned by a template.

Average gap length or INDEL_score (Sc4) is computed to score the reliability of a target-template alignment. It calculates the sum length of all the INDELs longer than 5 gaps (GAP5L) and the INDELs smaller than 5 gaps (GAP5S) to compute their average gap length score (Sc4) for the total of TR residues encoded in the target-template alignment, as shown in the equation 3 below. Here, the coefficients of 0.5 and 1 are respectively considered for the gaps shorter than 5 residues and atleast 5 residues. The INDELs in a target-template alignment are built through the *ab-initio* algorithm to construct an overall target conformation. The residue chunk, modelled through an *ab-initio* algorithm, is expected to be topologically more accurate for the shorter gap lengths and the gap length longer than 5 gaps is considered as a vital factor in defining the accuracy of an alignment.

$$\boxed{Sc4 = \frac{(0.5 * GAP5S + GAP5L)}{TR}} \quad - (3)$$

To evaluate the reliability of a template, the total number of gaps existing in an alignment is further considered. For a pairwise alignment, Average_gaps score (Sc5) of a template is computed as a fraction of the total number of gaps (TOTG) that are employed to optimally align the TR residues, as shown in equation 4 below.

$$\boxed{Sc5 = \frac{TOTG}{TR}} \qquad \text{- (4)}$$

Moreover, the gap segments incurred in the target-template alignment are expected to decrease the modelling accuracy of a target sequence and their evaluation through an affine gap penalty score becomes essential. As HHPred employs four different probabilities (opening as well as extension penalties for the residue insertions and similar two penalties for the residue deletions) on the basis of the structural context of a sequence segment to evaluate the frequency profile of an INDEL (Wang et al. 2011), the default gap penalties employed by the BLAST (11 for insertion and 1 for extension) are simply considered (Altschul et al. 1997; Schäffer et al. 2001) to compute an Affine_gap score (Sc6) for a total of TR residues.

As per the pairwise alignment, count of mismatched residues (TNMMR) is converted to a Sc7 score by dividing it with the total number of TR residues, as shown in the equation 5 below.

$$\boxed{Sc7 = \frac{TNMMR}{TR}} \qquad \text{- (5)}$$

The modelling reliability of a template for a target sequence is directly proportional to all these scoring schemes. A template with higher Sc1, Sc2 and Sc3 scores and lower Sc4, Sc5, Sc6 and Sc7 scores is expected to be a better template. The template similarity score (TLS) is computed by considering all these scores to assess the modelling reliability of a template for a target, as shown in the equation 6 below.

$$\boxed{TLS = Sc1 + Sc2 + Sc3 - Sc4 - Sc5 - Sc6 - Sc7} \quad - (6)$$

To estimate statistical reliability of this template score, its standard statistical Z_Score (TLS2) is further computed by using the mean (μ) and standard deviation (σ) of the similarity scores for all the selected templates, as shown in the equation 7 below.

$$\boxed{TLS2 = \frac{TLS - \mu}{\sigma}} \quad - (7)$$

Template ranking on the basis of overall similarity score and Z_Score is enlisted in table 3.1 for the CASP8 target. Here, the considered top-ranked HHPred resultant hit with the lowest E-value and also available during the CASP is mentioned as SEED and is sequentially employed in all the subsequent rankings of all the selected templates. All the considered templates are subsequently fed to the sequence/structure alignment (SALIGN) function of MODELLER9v9 (Sali & Blundell 1993), as per their pairwise rank order, to construct their structural topology guided MSA (MSA1) and to further statistically rank them by their structural similarity with the seed template.

Table 3.1 Pairwise alignment based scoring parameters for representative T0388 target of CASP8. Seven scoring parameters (Sc1-Sc7) are shown in different columns. The overall score and the statistical Z score are shown in the last two columns.

Target T0388 (174 residues)		**SCORE**								
TEMPLATE	**LENGTH**	**SC1**	**SC2**	**SC3**	**SC4**	**SC5**	**SC6**	**SC7**	**OVERALL SIMILARITY SCORE**	**Z SCORE** **(** $Z-Score=\frac{X-\mu}{\sigma}$ **)**
2P31A	155	0.482	0.247	0.943	0.014	0.027	0.356	0.407	0.868	SEED
2P5QA	161	0.306	0.217	0.972	0.040	0.054	0.403	0.531	0.467	0.767
2P5RA	164	0.299	0.219	1.000	0.041	0.059	0.462	0.550	0.405	0.540
2GS3A	171	0.326	0.207	0.995	0.048	0.064	0.499	0.528	0.389	0.482
2OBIA	165	0.291	0.199	1.000	0.050	0.068	0.490	0.561	0.321	0.233
2I3YA	188	0.177	0.175	1.000	0.087	0.099	0.514	0.569	0.083	-0.641
2R37A	191	0.172	0.186	0.995	0.081	0.093	0.521	0.597	0.061	-0.719
1GP1A	184	0.154	0.171	0.955	0.084	0.103	0.480	0.564	0.049	-0.766
2F8AA	186	0.101	0.176	0.974	0.040	0.064	0.525	0.611	0.011	-0.905
2HE3A	185	0.115	0.176	0.995	0.049	0.078	0.618	0.618	-0.077	-1.226

3.2.2 Multiple sequence alignment based template ranking

Through MSA1, all the selected hits are evaluated against the target and the seed template through several sequence as well as structural similarity based scoring schemes. For the length of a considered template (LT), the Unique_Residue (S1) score of a template counts the fraction of its residues that spans certain specific target residues (STR) which are uncovered by all the other hits, as represented in equation 8 below.

$$\boxed{S1 = \frac{STR}{LT}} \qquad - (8)$$

The proportion of mismatched hydrophobic residues (AHR), mismatched hydrophilic residues (AHYR) and the sequence identity score (ASID) are further employed to compute the Aligned_Charges (S2) score of a hit. As used for the pairwise alignment based template ranking, these 3 scores are computed to estimate the S2 score for the target and template sequences (Radzicka & Wolfenden 1988), as shown in equation 9 below.

$$\boxed{S2 = \frac{(AHR + AHYR + ASID)}{3}} \qquad - (9)$$

Moreover, the BLOSUM62 score is computed for all the hits against the target sequence (S3) and the seed template (S4). The S3 score computes the decimal logarithm of the total BLOSUM62 score (BLS1) of all the residues of a hit, aligned against the target sequence, divided with their total number of residues encoded in the hit and the target sequence (TR1), as shown in equation 10 below. Similarly, the S4 score computes the decimal logarithm of the total BLOSUM62 score (BLS2) of all the residues of a hit, aligned against the seed template, divided with their total number of residues encoded in the hit and the seed template (TR2), as shown in equation 11 below.

$$\boxed{S3 = \frac{BLS1}{TR1}} \qquad - (10)$$

$$\boxed{S4 = \frac{BLS2}{TR2}} \qquad - (11)$$

Despite all these sequence similarity based scores, several other structural similarity based scores of a template are considered. As per the considered MSA for the aligned residues of a hit and the seed template, structural similarity of a hit is assessed against the seed template in terms of TM_Score (S5) by employing the count of residues encoded in the seed template as the normalization factor. Similarly for all the aligned residues of the hit and the seed template, the GDT-TS (S6) score of a hit is computed against the seed template. S6 score is computed as the average of the residue fractions R1, R2, R3 and R4 of a hit, respectively fitting within 1, 2, 4 and 8Ǻ against the equivalent residues of the seed template.

Moreover, among the total number of residues encoded by a template, the fraction of its topological correct residues, fitting within 8Ǻ distance deviation against the equivalent residues of the seed template, is computed as S7 score to efficiently evaluate its structural relationship with the seed template, as shown in equation 12 below.

$$\boxed{S7 = \frac{R8}{LT}} \qquad - (12)$$

Individually all these seven scores are expected to be linearly proportional to the modelling accuracy and reliability of a template and are used to compute its structural similarity based credibility score for a target. This reliability score is additionally employed to statistically rank the templates as per their Z_Score. The MSA1 ranking of all the selected hits is shown in table 3.2, where similarity score and Z_Score are shown along with other scoring parameters. The MSA based similarity score (TMS) and Z_Score (TMS2) are calculated by the following equations 13 and 14.

$$\boxed{TMS = S1 + S2 + S3 + S4 + S5 + S6 + S7} \quad \text{- (13)}$$

$$\boxed{TMS2 = \frac{TMS - \mu}{\sigma}} \quad \text{- (14)}$$

On the basis of MSA1 ranking, the templates with TM_Score lesser than 0.5 or higher than 0.975 against the seed template (Zheng et al. 2010), an additional coverage span lesser than 5 residues over the target and statistically negative Z_Score are considered as unreliable hits. It has been observed that the hits having these unreliable scores (shown as DNC (Do Not Consider) by our MSA ranking) are not found to improve the modelling accuracy of the target over the seed template and are not considered as reliable structures. Moreover, the structurally redundant hits, clustered or culled in the HHPred results, with lower Z_scores are also discarded after keeping the top most Z_score hit. However, if a set of all the redundant hits fits with our considered DNC constraints, it is completely discarded. The high-scoring hits are then employed, as per their MSA1 rank order, to construct another MSA (MSA2) for re-ranking them and selecting their best set for a target sequence. The MSA2 alignment with the top-ranked representative templates is observed to be a good source to select templates for maximally covering the target sequence. This MSA2 script ranks the templates on basis of all the parameters that are employed in the MSA1 ranking. It also parses the MSA2 alignment to map the target residues that are aligned by the additional hits over the seed template. This MSA2 ranking is shown in table 3.3 for the selected CASP8 target T0388.

Table 3.2 MSA1 based scoring parameters for T0388 target of CASP8. 7 scoring parameters (S1-S7) are shown in different columns. Overall score and Z score are also shown for all the hits along with their additional coverage of the target sequence and the respective BLOSUM62 score. The templates those are not considered are also mentioned as “DNC”.

TEMP LATE	**LEN GTH**	**SCORE**							**CA RMSD**	**OVERALL SIMILARITY SCORE**	**Z SCORE**	**ADDITIONAL COVERAGE**		
		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**				**TARGET RESIDUES**	**BLOSUM62 SCORE**	
2P31A	155	SEED	0.37	0.48	SEED	1.00	1.00	1.00	0.00	SEED	SEED	SEED	SEED	
2P5QA	161	0.00	0.30	0.26	0.26	0.96	0.91	0.96	1.07	3.661	1.846	2.000	-0.301	DNC
2OBIA	165	0.00	0.27	0.23	0.30	0.94	0.92	0.92	1.52	3.584	1.635	6.000	-0.090	
2GS3A	171	0.00	0.27	0.21	0.27	0.94	0.90	0.89	2.07	3.478	1.344	6.000	-0.090	
2F8AA	186	0.00	0.25	0.15	0.24	0.92	0.88	0.83	1.70	3.278	0.796	3.000	0.544	DNC
2P5RA	164	0.00	0.29	0.22	0.21	0.85	0.83	0.87	2.35	3.270	0.775	5.000	-0.105	
1GP1A	184	0.00	0.25	0.14	0.21	0.92	0.86	0.84	1.63	3.225	0.653	2.000	0.204	DNC
2I3YA	188	0.02	0.25	0.13	0.21	0.90	0.83	0.81	2.58	3.154	0.457	10.000	0.176	
2HE3A	185	0.00	0.24	0.13	0.20	0.90	0.83	0.83	1.95	3.134	0.403	4.000	0.352	DNC
2R37A	191	0.04	0.24	0.09	0.16	0.91	0.84	0.81	1.74	3.091	0.284	11.000	0.019	
Redundant Hits														
2P5QA		2P5RA												
2HE3A		1GP1A		2F8AA										

Table 3.3 MSA2 based scoring parameters for T0388 target of CASP8. 7 scoring parameters (S1-S7) are shown in different columns. Overall score and Z score are also shown for all the hits along with their additional coverage of the target sequence and the respective BLOSUM62 score.

TEMP LATE	LEN GTH	SCORE							CA RMSD	OVERALL SIMILARITY SCORE	Z SCORE	ADDITIONAL COVERAGE	
		S1	S2	S3	S4	S5	S6	S7				TARGET RESIDUES	BLOSUM62 SCORE
2P31A	155	SEED	0.37	0.47	SEED	1.000	1.00	1.00	0.00	SEED	SEED	SEED	SEED
2OBIA	165	0.00	0.27	0.24	0.28	0.95	0.93	0.92	1.29	3.61	1.57	6.00	0.81
2GS3A	171	0.00	0.27	0.22	0.26	0.95	0.93	0.89	1.93	3.51	1.38	6.00	0.11
2P5RA	164	0.00	0.29	0.21	0.21	0.85	0.82	0.87	2.45	3.25	0.90	7.00	-0.17
2I3YA	188	0.02	0.25	0.12	0.21	0.89	0.83	0.80	2.68	3.13	0.68	11.00	0.11
2R37A	191	0.04	0.24	0.09	0.16	0.91	0.84	0.80	1.85	3.08	0.59	12.00	-0.02

Second round of MSA analysis is expected to eliminate the templates that are structurally dissimilar to the seed template. It is also able to select the other hits for additionally covering the target. Moreover, several hits have shown equal additional target coverage over the seed template. These templates have also shown equal BLOSUM62 score for the target and one representative hit is chosen among these hits. However, for a target that is maximally covered by the seed template itself, the seed template is solely employed to model the target sequence by curating its pairwise alignment through the PRALINE server (Heringa 1999). The selected MSA2 hit(s) maximally covering the target sequence are then employed to construct the most reliable conformation of the selected CASP8, CASP9 and CASP10 targets, as shown in the tables 3.4, 3.5 and 3.6 respectively. These tables enlist the CASP target domains along with their length, residues assessed during the CASP, the templates and accuracy scores of the best predicted CASP model in comparison to the GDT-TS score of our first constructed target model. These TBM-HA targets also encode FM and TBM domains and hence these domains are also considered as our modelling targets for this study. Moreover, all the target domains are evaluated against their native structure through maxcluster (Siew et al. 2000) and not with current scripts as these latter algorithmic tools are solely employed by CASP assessors and are not publically available. Therefore, some of these target domains show a GDT_TS score lesser than 80, a standard CASP threshold score for defining a TBM-HA target. Further, some of these domains (T0470_D2, T0674_D1 and T0726_D2) are threading targets and no high-scoring reliable templates are available for them. Hence, we have lesser modelling accuracy for them and this is the major reason for the large standard deviation of the modelling accuracy scores.

Table 3.4 Model assessment scores of the best predicted CASP8 model along with the target length, assessed segment, considered residues and its best predicting CASP8 team in comparison to GDT-TS assessment score of our first predicted conformation.

Target/ Domain	**Length**	**Assessed region**	**Assessed Residues**	**CASP8 Best Model**			**Our First Model**	
				Best Predictor CASP Group	**Templates**	**GDT-TS**	**Templates**	**GDT-TS**
T0388_ D1	174	11-174	164	pro-SP3-TASSER	2P31_A, 2GS3_A, 2I3Y_A, 2P5Q_A, 1GP1_A	91.616	2P31_A, 2I3Y_A, 2R37_A	91.006
T0390_ D1	182	29-123, 130-160	126	EB_AMU_ Physics	1SHW_A	90.726	1SHW_A	90.726
T0396_ D1	105	3-104	102	CpHModels	1JRA_D	89.706	1JR8_A	84.314
T0398	292	1-292	292	MUSTER Complete Model	2RIR_G, 2RIR_D, 2RIR_F, 2RIR_B, 2RIR_C, 2RIR_A, 2RIR_E	87.500	2RIR_A, 2CUK_A	94.010
T0398_D1		1-124, 272-290	143	MUSTER_ D1 Model	2RIR_G, 2RIR_D, 2RIR_F, 2RIR_B, 2RIR_C, 2RIR_A, 2RIR_E	96.786		96.071
T0398_D2		125-271	147	Zhang-Server_ D2 Model	2RIR_A, 1PJC_A, 1B0A_A, 1X13_A, 2RIR_H	98.980		99.490

T0400_D1	162	2-146, 152-161	155	PS2-Server	2Q7B_A	88.387	2Q7B_A	86.774
T0402_D1	139	4-100, 112-127	113	McGuffin	NA	79.630	2FHQ_A, 2HQ7_A	74.769
T0404_D1	110	2-98,	97	MULTICOM-REFINE	NA	89.815	2CZ4_A, 2J9C_A	89.506
T0418	222	1-222	222	MULTICOM-RANK	2HSZ_A, 2NYV_A 2HI0_A, 2AH5_A 2HDO_A, 2GO7_A 2HOQ_A, 2FI1_A	87.976	2HDO_A, 2HI0_A,	82.500
T0418_D1		2-16, 86-211	141	SAMUDRALA	NA	90.780		85.816
T0418_D2		17-85	69	3DShot1	1AAB, 1AAC, 1AAF, 1AAJ	86.957		85.145
T0422	357	51-354	304	3D-JIGSAW_AEP	NA	71.313	3B9P_A, 3CF0_A, 1IN4_A	70.683
T0422_D1		57-250, 342-357	210	3D-JIGSAW_AEP	NA	79.500		81.625
T0422_D2		251-340	90	Phyre_de_novo	NA	87.500		86.875
T0423_D1	110	2-98,	97	Zhang-Server	2OTM_A, 2B33_A, 1JD1_A, 1JD1_B, 2UYK_A	85.870	2OTM_A, 1QAH_A	89.402
T0426_D1	283	27-283	257	MULTICOM-RANK	1HCB_A, 1AZM_A, 1LUG_A, 1FLJ_A 1MOO_A, 1V9E_A 1Z93_A, 1RJ5_A	97.860	2FOY_A	94.942

T0428_D1	267	20-248	229	3Dpro	1XQ9_B	97.052	1XQ9_A, 1T8P_A, 1FZT_A	94.978
T0432_D1	130	1-130	130	MULTICOM	NA	87.885	2DKW_A, 1E6I_A. 2YQD_A	86.923
T0435_D1	151	15-58, 73-148	120	MULTICOM	NA	84.091	2QPW_A	84.091
T0438	439	1-439	439	MULTICOM	NA	79.716	2OAS_A, 2G39_A	78.101
T0438_D1		2-185	184	MULTICOM	NA	87.805		85.366
T0438_D2		186-430	245	METATASSE	2OAS_A	93.161		91.816
T0442	269	1-269	269	3D-JIGSAW_AEP	NA	82.872	2PIF_A	81.170
T0442_D1		11-124, 158-166, 233-266	157	LEE	2PIF_A,2PIF_B	92.994		93.471
T0442_D2		131-157, 167-213, 219-232	88	EB_AMU_ Physics	2PIF_A	97.603		96.575
T0444_D1	326	34-317	284	LEE	1JK0_A,2O1Z_A	94.669	1SMQ_A, 1H0O_A	94.485
T0447_D1	542	1-152	152	PSI	1EG7_A	88.330	1EG7_A	88.653
T0458_D1	107	12-88.	77	FFASstandard	2OKA_A	97.727	2OKA_A	97.728
T0470	223	1-223	223	FEIG	NA	82.447	2QGS_A	81.516
T0470_D1		2-125	124	HHpred5	2QGS_A, 3B57_A	85.586		82.883
T0470_D2		126-214	89	PS2-server	2QGS_A	41.329		81.358
T0499_D1	56	1-56	56	Xianmingpan	1IGD_A	85.714	2ONQ_A	84.375

Table 3.5 Model assessment scores of the best predicted CASP9 model along with the target length, assessed segment, considered residues and its best predicting CASP9 team in comparison to GDT-TS assessment score of our first predicted conformation.

Target/ Domain	**Length**	**Assessed region**	**Assessed Residues**	**CASP9 Best Model**			**Our First Model**	
				Best Predictor CASP Group	**Templates**	**GDT-TS**	**Templates**	**GDT-TS**
T0521	179	1-179	179	HHpredB	3K21_A, 3KHE_A, 2GGM_A, 3LIJ_A, 2AAO_A, 2MYS_C, 1EXR_A, 2EHB_A	45.238	3KHE_A	40.030
T0521_D1		1-34, 107-179	107	Jones-UCL	N/A	77.806		69.898
T0521_D2		35-104	60	fams-ace3	N/A	87.143		75.357
T0522_D1	134	4-134	131	BAKER-ROSETTASERVER	3I4S_A	94.656	3I4S_A	94.655
T0523_D1	120	4-114	111	ZHOU-SPARKS-M	3LYX_B	88.288	3LYX_A	79.505
T0528	388	1-388	388	LEE	1QO0_A, 3I09B 3I45_A, 3EAF_A, 3LKB_A, 3LOP_A, 3H5L_B	48.315	3I45_A	44.340

T0528_D1		18-138, 269-351	204	MULTICOM-NOVEL	3I45_A, 1PEA_A, 1USG_A, 3I09_A, 3HUT_A, 2LIV_A, 3H5L_A, 3EAF_A	86.275		80.270
T0528_D2		139-268, 352-381	160	United3D	N/A	60.000		55.625
T0530_D1	115	36-115	80	YASARA	2K5Q_A	82.500	2K5Q_A	75.313
T0538_D1	54	2-54	53	PconsR	N/A	91.509	1J7K_A	81.132
T0541_D1	106	1-17, 19-71, 75-106	102	YASARA	3IDU_B	80.637	3IDU_A	72.059
T0559_D1	69	3-69	67	BAKER-ROSETTASERVER	N/A	92.910	2VXZ_A	73.883
T0560_D1	74	3-66	64	Splicer	N/A	92.188	1R1U_A	81.641
T0563_D1	279	1-60, 66-70, 86-156, 169-260	228	HHpredC	1GP6_A, 1DCS_A, 1ODM_A, 1W9Y_A	76.549	1OC1_A, 1UNB_A, 1GP4_A	65.819
T0566_D1	156	10-152	143	fams-ace3	N/A	76.538	1USV_B	71.346
T0567_D1	145	10-144	135	BAKER-ROSETTA SERVER	1NY6_A	79.444	1NY5_A	74.630
T0570_D1	258	24-256	233	LEEcon	N/A	80.687	2PZ0_A, 3L12_A	78.433

T0580_D1	105	2-105	104	BAKER	1IIB_A, 1H9C_A, 1TVM_A, 2R48_A, 2Q9U_A	89.663	2WY2D	78.606
T0586	125	1-125	125	Zhang	3BY6_E, 3BY6_A, 3BY6_B, 2EK5_C, 3IC7_A	75.841	2DU9_A, 3BY6_A	77.941
T0586_D1		5-84	80	LTB	N/A	92.813		88.438
T0586_D2		85-123	39	Distill	N/A	88.462		83.3335
T0589	465	1-465	465	gws	N/A	60.035	1WU7_A	49.884
T0589_D1		24-65, 96-188, 271-369	234	United3D	N/A	74.232		71.820
T0589_D2		189-270	82	Seok-server	1WU7_A, 3LC0_A	77.744		63.110
T0589_D3		370-464	95	MUFOLD-MD	2EL9_A	93.085		82.714
T0594_D1	140	1-140	140	LTB	N/A	84.643	1X53_A	79.107
T0596	213	1-213	213	FEIG	N/A	57.759	3C07_A	58.908
T0596_D1		6-58	53	FEIG	N/A	95.755		90.565
T0596_D2		59-188	130	PconsR	N/A	61.364		54.339
T0599_D1	399	10-99, 113-182, 186-392	367	LEE	2FN1_B, 3HWO_B, 2G5F_D, 2G5F_B, 1I1Q_A, 3BZN_A	81.557	3HWO_A, 2FN0_A, 3H9M_A	81.079
T0600	125	1-125	125	ProQ2	N/A	43.396	3LYX_A, 3EEH_A	42.925
T0600_D1		17-75	59	PconsD	N/A	78.814		75.847
T0600_D2		76-122	47	pro-sp3-TASSER	3EEH_A, 3H9W_A, 3LYX_A, 2GJ3_A, 3ICY_A	89.894		86.170

T0601_D1	449	7-448	442	McGuffin	N/A	84.070	1VPB_A, 1VL4_A	80.045
T0602_D1	123	1-55	55	BAKER	3A7M_A, 3H3M_A	89.545	3H3M_A, 3A7M_A,	84.091
T0605_D1	72	18-66	49	BAKER	1ZXA_A, 1ZXA_B	93.230	2NPS_A, 1H89_A	93.229
T0611	227	1-227	227	YASARA	3G1L_A	51.131	1UI5_A	50.377
T0611_D1		3-55	53	Seok-server	2G7S_A, 2HKU_A, 2ID3_A, 3DEW_A, 3HIM_A, 2HYJ_A, 3G1L_A, 1T56_A, 1RKT_A, 3DCF_A, 1PB6_A, 3F0C_A, 3KNW_A, 3LHQ_A, 2EH3_A, 2NX4_A, 1VI0_A, 3LWJ_A, 2ZCM_A, 2ID6_A, 2QTQ_A	98.078		96.153
T0611_D2		56-169, 179-213	149	LEEcon	N/A	48.630		52.055
T0613_D1	287	5-130, 137-285	273	ProfileCRF	3LOU_A	93.937	3LOU_A	92.350
T0614_D1	135	2-12, 23-34, 51-66, 74-79, 86-111	71	LEEcon	N/A	86.972	2CY5_A, 1EAZ_A	84.858

T0619_D1	111	1-101	101	FEIG	N/A	88.614	3LYS_A	71.535
T0620_D1	312	3-164, 179-312	296	HHpredC	3MTC_A, 1I9Z_A, 2IMQ_X	87.456	3MTC_A	85.976
T0626_D1	283	2-283	282	Jones-UCL	N/A	89.628	3LOU_A	83.156
T0629	216	1-216	216	bujnicki-kolinski	N/A	27.546	1OCY_A, 2FKK_A	25.579
T0629_D1		1-49, 209-216	57	PconsM	N/A	83.333		75.877
T0629_D2		50-208	159	GSmetaserver	1OCY_A	8.176		10.377
T0632_D1	168	44-57, 65-164	114	Phyre2	N/A	92.544	3DKZ_A, 1Q4T_A.	90.132
T0634_D1	140	3-83, 95-126	113	BAKER-ROSETTA SERVER	1YS6_A	90.421	3JTE_A, 3GT7_A	85.982
T0635_D1	191	10-170	161	HHpredA	3MN1_A, 1K1E_A, 2P9J_A, 2R8E_A, 3EWI_A, 3MMZ_A, 3IJ5_A	99.068	3MN1_A, 3IJ5_A	97.826
T0636_D1	336	17-334	318	HHpredB	3GET_A, 3EUC_A, 1UU1_A, 3LY1_A, 3FFH_A, 3CQ5_A, 3FTB_A	81.682	3HDO_A, 1FG7_A, 3EZ1_A, 3CQ5_A	76.651
T0640_D1	250	9-145, 157-196, 208-237	207	Jones-UCL	N/A	82.486	1IY8_A, 3KVO_A, 3IOY_A	77.401

Table 3.6 Model assessment scores of the best predicted CASP10 model along with the target length, assessed segment, considered residues and its best predicting CASP10 team in comparison to GDT-TS assessment score of our first predicted conformation.

Target/ Domain	Length	Assessed region	Assessed Residues	CASP10 Best Model			Our First Model	
				Best Predictor CASP Group	Templates	GDT-TS	Templates	GDT-TS
T0645_D1	537	40-537	498	Phyre2_A	N/A	78.514	3GZS_A, 3EHN_A	72.892
T0650_D1	346	4-342	339	HHpredAQ	1O6V_A, 1H6U_A, 3O6N_A, 1M9S_A	92.478	2OMU_A	81.932
T0657_D1	154	4-17, 23-44, 51-68, 71-149	133	Distill	N/A	82.519	1BWN_A, 2DHI_A	83.647
T0659_D1	85	1-74	74	Phyre2_A	N/A	93.581	3LD7_A	90.878
T0662_D1	79	4-79	76	RBO-MBS	N/A	82.896	3GZL_A	76.975
T0663	205	53-204	152	LEEcon	N/A	31.926		35.642
T0663_D1		53-138	86	LEE	N/A	59.821		47.917
T0663_D2		139-204	66	Baker	N/A	87.891		81.253
T0664_D1	540	43-540	498	PMS	N/A	83.368	3CGH_A	79.782

T0674	340	1-340	340	MULTICOM-CONSTRUCT	N/A	36.293	3H41_A, 2K1G_A	36.379
T0674_D1		46-206	161	BAKER	N/A	63.365		7.862
T0674_D2		207-340	134	Mufold	N/A	82.443		80.534
T0675	75	1-75	75	Bilab-ENABLE	2ELR_A	34.333	2CT1_A	38.000
T0675_D1		17-43	27	QUARK	2YT9_A, 1F2I_G, 2COT_A	81.481		83.335
T0675_D2		44-73	30	Phyre2_A	N/A	85.000		79.165
T0689_D1	234	23-130, 132-234	211	PconsM	N/A	86.178	3FZX_A	85.216
T0692_D1	473	1-470	470	MULTICOM-CLUSTER	1KY8_A, 1WND_A, 1UZB_A, 1BXS_A, 2JG7_A, 1O9J_A, 1EUH_A, 2J6L_A	79.185	3N83_A, 3IWK_A	75.109
T0708_D1	196	1-196	196	Seok-server	N/A	83.995	3IRV_A, 3OT4_A	79.101
T0712_D1	223	38-223	186	Mufold-MD	3U22_B, 3H8T_A	92.319	3U22_A	91.566
T0714_D1	88	1-88	88	BAKER-ROSETTA SERVER	1WAAA_201	89.489	2E6P_A	85.511
T0716_D1	71	10-60	51	BAKER-ROSETTA SERVER	2DMQA_201	93.627	2CRA_A, 1SAN_A	93.137

T0721_D1	301	3-301	299	HHpredA	3FBS_A, 1HYU_A	78.912	3FBSA, 3F8PA	76.020
T0726	597	1-597	597	LEE	N/A	32.742	1Z1W_A	30.536
T0726_D1		1-447	447	MUFOLD-Server	3SE6_B 3Q7J_A 1Z5H_A	43.580		39.773
T0726_D2		484-564	81	Jones-UCL	N/A	87.500		20.938
T0726_D3		448-483, 565-587	59	Zhang-IRU	N/A	21.552		22.414
T0731_D1	79	8-62	55	MATRIX	2KZ5_A, 1SKN_P, 2WT7_B, 3A5T_A	80.909	1SKN_P	80.000
T0747_D9	121	24-34, 43-121	90	Mufold-MD	3SD2_A 3D33_A	69.382	3D33_A	67.416
T0749_D1	449	35-446	412	Pcons-net	N/A	91.479	3EU8_A, 3QWT_A	88.409
T0752_D1	156	2-149	148	PconsM	N/A	88.699	4STD_A, 3EF8_A	78.253
T0757_D1	247	1-247	247	BAKER-ROSETTA SERVER	2OWNA_201	82.563	2ESS_A, 2OWN_A	78.676

3.3 Results and Discussion

Conventional template selection algorithms employ either the pairwise alignment or the MSA information of templates to select their logical set for a target sequence. However, both these pairwise and MSA alignments provide diverse information about a template. A template ranking scheme employing both the pairwise and MSA alignment information together into a single algorithm is expected to select the best set of templates.

Compared to the traditional template scoring measures that screen the credibility of a template on the basis of a single scoring measure, our pairwise ranking methodology employs the diverse set of seven different measures to rank the templates as per their pairwise alignments for a target sequence. By statistically ranking the templates through all these different reliable attributes, our scoring methodology is expected to sturdily assess the modelling credibility of a template for a target sequence. Sc1 score computes the BLOSUM62 score of the aligned target-template residues. As it employs the total number of residue encoded in the target and template sequences, it makes the BLOSUM62 score dependent on the length of the target and template sequences. Higher Sc1 value of a template implies its higher BLOSUM62 score for each of its residue against the target and is expected to increase its credibility. As sequence identity or similarity scores normally prove to be confusing measures for evaluating the credibility of a template, employing an average of the sequence identity, mismatched hydrophobic residues and mismatched hydrophilic residues for the aligned target-template residues makes the Sc2 score really credible. Consideration of coverage span of a template against the target sequence as Sc3 score further increases its credibility. Its higher score should certainly affirm that higher scoring value of all the other employed measures is uniformly applicable for the major chunk of the target against a template and should confirm its credibility. Besides considering these measures for the aligned target-template residues, the gaps are also equally important to define the credibility of a template. However, a higher gap length decreases the modelling accuracy of a template and evaluating the Sc4 score by considering the coefficients of 0.5 and 1 for all the gap lengths, respectively shorter and longer than 5 gaps, is anticipated to affirm the credibility of a pairwise alignment. Moreover, to confirm that a lower number of gaps exist in a pairwise alignment, the Sc5 proportion score of gaps required to align a total of target and template residues should also be computed. Consideration of the benchmarked affine gap penalty scheme (Sc6 score) employed by the BLAST algorithm (11 for insertion and 1 for extension) and the average proportion of mismatched residues existing in a pairwise alignment (Sc7 score) should further strengthen the template ranking algorithm to efficiently screen the best template(s) for a target sequence. As our statistical ranking algorithm specifically employs GDT-TS, TM_Score and the fraction of topologically correct residues of a hit against the seed template along with several other sequence and structural scoring parameters, it reliably evaluates the structural topology of a hit against the target sequence and the seed template.

Traditional sequence or structural MSA profile based template selection and combination algorithms simply attempt to maximally cover the target sequence through minimal possible structures. Contrarily, our MSA based template ranking algorithm assesses the sequence as well as structural similarity of the screened set of templates to select the statistically top-ranked set of templates. Our algorithm employs S1 score for a template to compute the proportion of its unique residues that spans certain specific target residues that are uncovered by all the other hits. Quite similar to pairwise ranking methodology, the S2 score of the MSA ranking algorithm employs the proportion of mismatched hydrophobic residues, mismatched hydrophilic residues and the sequence identity score to compute the credibility score of a template. The reliability score of a template is further strengthened by the assessment of per residue BLOSUM62 score of a template both against the target sequence (S3) and the seed template (S4). In contrast to the normal algorithms that compute the BLOSUM62 score of a template against the target, BLOSUM62 based evaluation of a template against the seed template efficiently screens their conformational relationship and is expected to strengthen our algorithm for selecting the best set of topologically similar templates. However, as the functionally similar proteins evolving in different microenvironments might still retain the topologically equivalent conformation, all these sequence based similarity scores could have failed to efficiently select the best possible templates. Evaluation of the structural similarity of these templates circumvents this problem. As per the considered MSA, our ranking algorithm evaluating the structural similarity scores, in terms of TM_Score (S5), GDT-TS (S6) and the average proportion of topologically correct residues fitting within 8Ǻ distance deviation, of all the templates against the top-ranked and credible seed template (S7) is expected to efficiently rank the templates. Statistically ranking the MSA templates through all these diverse scoring schemes is expected to efficiently discriminate their credibility.

For the single domain target T0388 having 174 residues, the best predicted CASP8 model constructed by the Pro-SP3-TASSER group employed 5 templates (2P31_A, 2GS3_A, 2I3Y_A, 2P5Q_A and 1GP1_A). The GDT-TS, GDT-HA and TM_Score of this model is found to be 91.616, 77.896 and 0.951 in comparison to the respective scores of 95.732, 82.165 and 0.971 for our model predicted through three templates (2P31_A, 2I3Y_A and 2R37_A). The best predicting CASP8 group PRO-SP3-TASSER unnecessarily employed 5 templates for threading this target sequence and it did not assess the structural similarity of these templates for selecting their best set. However, our algorithm employed only the top-ranked set of templates with ample structural similarity to construct the target model and its modelling accuracy as well as applicability for the selected CASP TBM-HA targets is discussed later in chapter V. Our algorithm is expected to correctly rank the templates for selecting their best set to maximally cover the target sequence and construct its accurate models consistently.

Chapter IV

Development of protein model sampling methodology

4.1 Introduction

A single long model sampling step is traditionally employed to energetically refine the predicted target structure for relieving the energetically unstable and erroneous atomic contacts that are normally present in a modelled conformation. However, this sampling step does not improve the accuracy of a protein model consistently. This sampling inefficiency is not only due to inaccuracy of the energy function and inefficient step parameter employed by the sampling algorithm, but is also due to the inaccuracy of our assessment measures to select the accurate model. The correct evaluation of accuracy of the model sampling step is impractical if all the sampled structures are not correctly evaluated to select the best model (Siew et al. 2000; Melo et al. 2002; John & Sali 2003; Zhang & Skolnick 2004; Joo et al. 2010; Trojanowski et al. 2010). The assessment measures employed by the MODELLER (MOLPDF, DOPE, Z_Score and GA341) and the other scoring parameters like GDT, TM_Score (Melo et al. 2002; John & Sali 2003), DFIRE (Yang & Zhou 2008) and RMSD are mutually non-linear and these scoring parameters do not unanimously screen a single predicted model as the accurate structure. Selecting the correct structure among the models sampled for a target sequence is still a major problem and the model assessment step should be improved to consistently predict accurate structures. We have screened the reliable assessment measures to guide and assess the optimal sampling of a target sequence.

The interplay between model sampling and model assessment steps is analyzed for increasing the accuracy of protein modelling algorithms. The reliable assessment scores are sorted out and streamlined in an iterative algorithm to exploit their consensus scoring criteria for selecting a reliable model among the ones sampled for a target sequence. Modelling accuracy results of the traditional single long sampling is further evaluated in comparison to our developed iterative sampling strategy.

4. 2 Methodology

A 182 residue CASP8 TBM-HA target (Moult 2005) T0390 is selected and the EphB2 / EphrinA5 complex structure (PDB ID: 1SHW_A) is used as the template. The target model is constructed and sampled with the MODELLER9.9 package (Sali & Blundell 1993).

4.2.1 Model Assessment and Selection Strategy

The following model assessment and selection criteria are developed to choose the best protein model. The target sequence is used to model 100 decoy structures through MODELLER. The constructed model with the highest TM_Score is then employed to further build 100 target models. The top 10 models with the highest TM_Scores are then scored through the other usually employed assessment measures, viz. MOLPDF, DOPE, GA341 and Z_Score, to study the correlation of these different measures with the TM_Score in reliably discriminating two close models. The consensus scoring set of assessment measures is subsequently employed to score and select the highly accurate models.

4.2.2 Model Sampling Strategy

In contrary to traditional single long sampling, we employ an iterative model sampling strategy. Initially 1000 models are generated for the target sequence. The top 10 models with the highest TM_Score are then sorted out and the one with the lowest Z_Score amongst these structures is selected. The combination of TM_Score and Z_Score criteria for selecting the best model is adopted after evaluating the performance of other selected scoring schemes in reliably discriminating the two pretty close models. This selected model is then used as a template to construct another set of 1000 models. Such 1000 model sampling runs are iteratively employed, each time starting with the best model of the current sampling.

These iterative runs are employed until the convergence is attained for the selected assessment measures for atleast 3 sampling runs. The consensus scoring, optimally sampled model with the highest TM_Score and lowest Z_Score is then selected as the best predicted structure. To compare our results with the conventional single sampling run, an equivalent number of models are further manually constructed with a single sampling run.

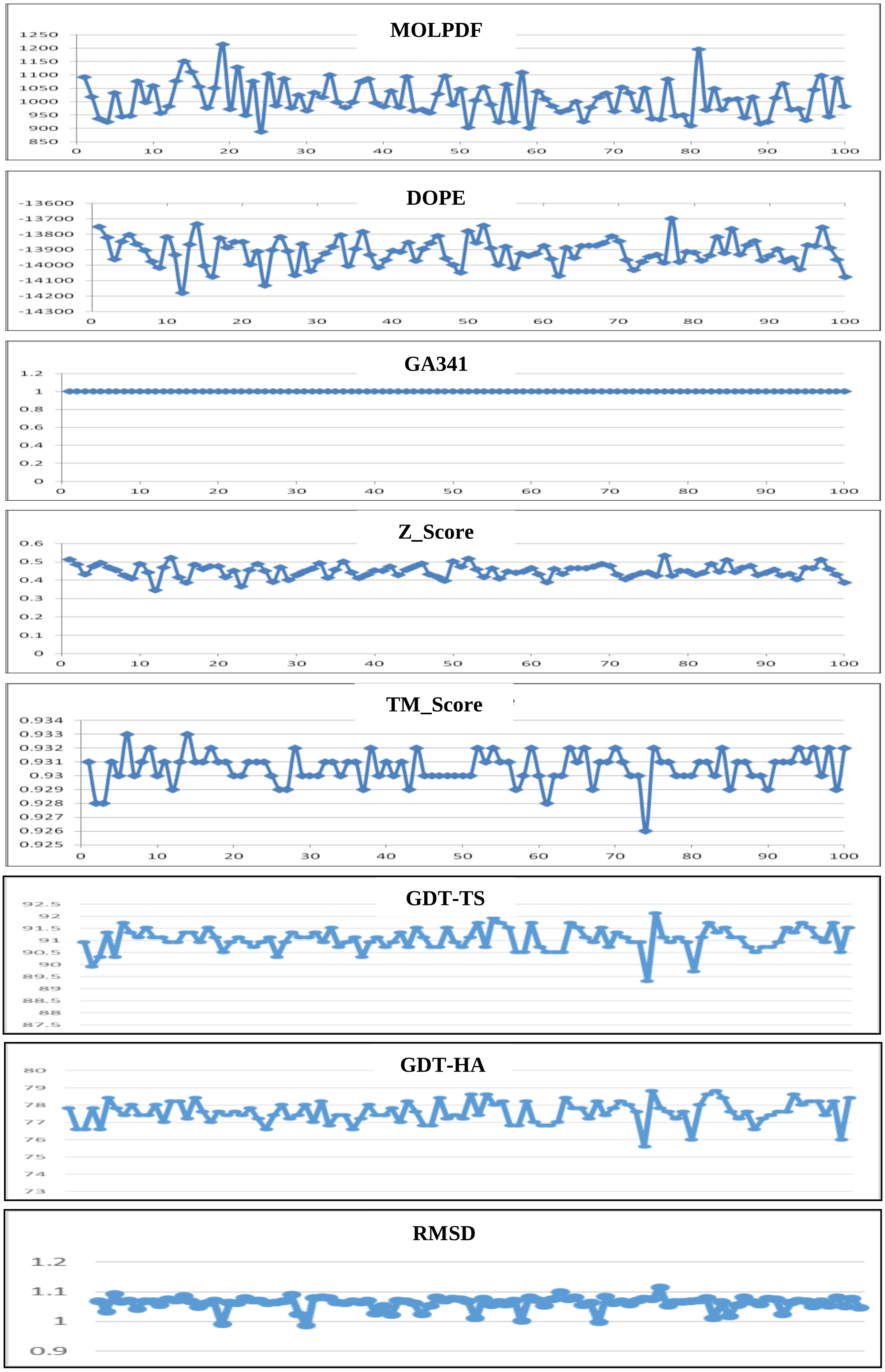
4.3 Results and Discussion

Conventional sampling algorithms keep perturbing the model structure in an attempt to minimize the energy. A single assessment measure is not found to consistently select the accurate predicted model from the sampled set of decoys. Hence, if an incorrect model is wrongly selected, the sampling accuracy seems to be further decreased. However, both these steps are mutually interrelated and iteratively sampling the model through the best set of assessment measures is expected to construct and select the best possible sampled model.

4.3.1 Model Assessment and Selection

It is observed that scores of different model assessment measures keep fluctuating during the model sampling and it becomes difficult to select the accurate model from the generated decoys. Assessment scores do not justify a single model as the single most reliable structure. Evaluating all the selected assessment measures viz. TM_Score, MOLPDF, DOPE, GA341, Z_Score, GDT-TS, GDT-HA and RMSD for the short 100 model sampling run, we observe significant deviations in all the other scoring measures except the TM_Score, as shown in the following Fig. 4.1.

Fig. 4.1 Assessment scoring undulations of the 100 models sampled for the target T0390.



Selecting a reliable model from the generated decoys is a complicated step and to make it simpler, we screen the scoring measures which consistently rank and select the most accurate models. We observe that TM_Score measure (Sadreyev et al. 2009) selects the utmost reliable and accurate model. It effectively evaluates both the Cα distance deviations between the equivalent residues of the two protein structures and also considers the count of such residues fitting within a minimal distance of 3.5Ǻ to calculate an overall structural similarity score for a protein model (Siew et al. 2000). Hence, the TM_Score could reliably discriminate an accurate model topology from the generated decoy structures and it should thus be considered as the initial evaluation criterion. However, when models show fairly equivalent TM_Score, selecting an accurate one becomes difficult. Hence all other selected measures viz. MOLPDF, DOPE, GA341, Z_Score, GDT-TS, GDT-HA and RMSD are used to evaluate the top 10 models with the highest TM_Score, as enlisted in the Table 4.1.

Table 4.1 Assessment details MOLPDF, DOPE, GA341, Z_Score, GDT-TS, GDT-HA and RMSD for the top 10 models with the highest TM_Score.

Model	TM_ Score	MOL PDF	DOPE	GA 341	Z_ Score	GDT-TS	GDT-HA	RMSD
Model6	0.933	889.00	-14187.18	1	0.3441	91.734	78.427	1.041
Model14	0.933	1171.95	-13712.27	1	0.5289	91.331	78.226	1.049
Model9	0.932	918.82	-14143.62	1	0.3610	91.532	78.024	1.055
Model17	0.932	1014.07	-13953.83	1	0.4349	91.532	78.427	0.99
Model28	0.932	942.70	-14080.99	1	0.3854	91.331	78.024	0.986
Model38	0.932	957.79	-14170.31	1	0.3507	90.927	77.218	1.052
Model44	0.932	1012.26	-14171.64	1	0.3501	91.532	78.226	1.052
Model52	0.932	965.55	-14068.48	1	0.3903	91.734	78.629	1.055
Model54	0.932	992.71	-14204.79	1	0.3372	91.935	78.629	1.057
Model64	0.932	888.58	-14138.81	1	0.3629	91.734	78.427	1.055

MOLPDF and GA341 scoring seems to be ineffective as MOLPDF shows too much deviations across these models. As per Table 4.1, GA341 score stands equal for all the models and it thus does not differentiate among the close and structurally correct, near-native models. Models with high GA341score are good structures but GA341 fails to discriminate between the good models that are structurally too close. It does not unanimously rank the best model and is not correlated with the TM_Score ranking. Similarly from Table 4.1, MOLPDF, GDT-TS, GDT-HA and RMSD do not correlate with TM_Score and these scores are only found to be effective in simply discriminating between a good and a bad model.

Quite interestingly, compared to all the other selected scoring measures, Z_Score is found to be much more minimally deviant with the TM_Score measure across these 10 models. As the energy of a good biologically meaningful as well as topologically correct protein model should be statistically as low as possible and, as also quite clear from the scoring values of these parameters for the models 9 and 64, the normal DOPE score or energetic assessment is found to be quite unreliable in contrast to the statistically significant Z_Score. Proportional alteration of DOPE score for these models is hereby found to be negligible in comparison to the significant alteration of their energetic Z_Score. Further as per Table 4.1, Model54 has the best energetic score, but its TM_Score is not the highest among the selected models. Hence, the usual DOPE or Z_Score itself does not select the accurate predicted model. To make the assessment measure more robust and effective at selecting the best model conformations, we therefore employ the TM_Score along with the best of the other selected scoring measures i.e. Z_Score together as a single model assessment measure. Here for each model sampling, we have sorted the top 5 models with the highest TM_Score and then finally selected the model with the lowest Z_Score as the best conformation. We further find this sampling methodology very effective. It is important to understand here that we have used only these measures to score our model predictions because when the experimental structure of the considered target sequence is not available, we cannot use the other scoring measures like GDT-TS, GDT-HA and RMSD.

4.3.2 Model Sampling

During model sampling, we iteratively generate one thousand models with the help of MODELLER and select the best model on the basis of TM_Score and Z_Score. This iterative sampling is continued until the employed scoring measures get saturated for a minimum of 3 sampling runs.

It took altogether 12 sampling runs in our iterative sampling strategy and the assessment scores of each of the chosen high scoring and intermittent sampling conformation is enlisted in Table 4.2. Here, the TM_Score and Z_Score guided best model predicted in each iteration cycle along with the first model and the finally predicted models are evaluated as per the CASP defined domain boundary information (Tress et al. 2009) and assessed both against the actual native structure of the target sequence and the employed template. The TM_Score of model computed against its native structure is normalized by 126 residues (*Assessed domain length*) and is termed as TM_Score_answer. The TM_Score normalized by the length of the selected template (*138 residues*) for the same domain boundary information is referred as TM_Score.

The iterative sampling strategy yields a model with GDT_TS, GDT_HA, TM_Score and RMSD score of 93.548, 82.863, 0.945 and 0.919 respectively, as highlighted with bold characters in Table 4.2. This model is more accurate than the first target structure predicted through the templates. Hence, in contrast to the single long model sampling step, the iterative sampling strategy optimally samples the conformational space of the target sequence through the best set of model assessment measures to construct a more accurate structure.

Table 4.2 Iterative sampling results of the considered assessment measures for top model of each iterative run along with the best CASP8 and the best single long sampling model.

Model	**Assessment against the template (*TM_Score*)**	**Z_ Score**	**Assessment against the solved structure**			
			GDT-TS	**GDT-HA**	**TM_Score _Answer**	**RMSD_ Answer**
First Model	0.98439	0.370	90.726	77.621	0.931	1.075
1	0.98032	0.504	90.726	76.613	0.937	1.014
2	0.97501	0.571	91.935	78.427	0.937	0.990
3	0.97155	0.509	92.137	79.234	0.939	0.979
4	0.96655	0.493	92.137	80.040	0.940	0.961
5	0.96518	0.505	92.137	80.847	0.942	0.949
6	0.96482	0.496	92.742	81.250	0.943	0.933
7	0.9594	0.481	92.540	81.452	0.944	0.918
8	0.95139	0.477	92.742	82.056	0.945	0.902
9	**0.94876**	**0.467**	**93.548**	**82.863**	**0.945**	**0.919**
10	0.9478	0.546	92.944	81.250	0.943	0.943
11	0.951	0.568	93.347	80.645	0.942	0.942
12	0.94936	0.566	92.742	79.839	0.942	0.952
Best single long sampling model	0.98159	0.376	91.129	78.226	0.933	0.988
Best CASP8 Model	0.90066	-1.012	90.726	78.831	0.929	0.990

In comparison to iterative sampling result, a single long increased sampling of 12000 models (*equivalent to our set of iterative sampling models*) has produced interesting results. TM_Score of the top model is marginally improved over the TM_Score of first constructed model (0.933 vs 0.931). Similarly, the GDT-TS scores of these two models remain same and it indicates that the single long sampling technique does not improve the structure quality. Increased sampling however improves the model quality and accuracy compared to the first built conformation for a target sequence, only when it is carefully employed and assessed.

In comparison with iterative sampling, we obtained best TM_Score of 0.933 from single long run sampling. Upon long single sampling, the initial model GDT-TS score of

90.726 marginally improves to 91.129. Even if the best model is correctly selected, the GDT-TS score is improved by 1.106 after sampling of 12000 models. On the other hand, our proposed iterative sampling reliably predicts near-native target conformation with GDT-TS, GDT-HA, TM_Score and RMSD score of 93.548, 82.863, 0.945 and 0.919 respectively, as shown in Table 4.2. As per the closeness of these assessment score values, it is really difficult to blindly select only this particular model and it is even more difficult to select it for every single sampling step for each of the considered target sequences consistently. So probably we might have also missed a more accurate target model that was correctly constructed by our sampling strategy and which could not be selected for every single iterative sampling step. However, in contrast to this target model, the GDT-TS, GDT-HA, TM_Score and RMSD scores of the best CASP8 Model are 90.726, 78.831, 0.929 and 0.990 respectively.

The usually employed sampling measures do not keep a track of the sampling path to improve the accuracy of model prediction towards its native conformation. Despite this lacuna, it often gets stuck in some wrong local minima prevailing in an energetic landscape. Here if a wrong topology is incurred during sampling, it is sequentially maintained till the end and if we could track this sampling path through the reliable assessment measures, we would be logically sampling the protein conformation on the correct path. We realize that employing a correct set of assessment measures together significantly and consistently predicts more accurate conformation for a protein sequence.

Our sampling methodology improves the model accuracy by 2.822 GDT_TS score in comparison to the first constructed model. The TM and Z score guided iterative sampling methodology yields a better model topology than the conventional single sampling run for the considered target sequence. This iterative sampling methodology pushes the topology of the predicted model structure towards its native conformation. Quite interestingly, GDT-TS improvement rate per model is almost negligible in the conventional sampling run. However, in our methodology the model accuracy is significantly improved.

Further as shown in Fig. 4.2, TM_Score of the model against the selected template marginally decreases in the sampling iterations despite the fact that it still retains the structural topology harnessed from the template and the alignment file. It indicates that all these models still have the same count of the topologically correct Cα residues localized within 5Å distance deviation against the equivalent template residues, as assessed by the TM-align tool (Zhang & Skolnick 2005). Attaining almost the similar structural decoys with almost equivalent TM_Score for at-least 3 successive and iterative sampling runs implies the saturation of TM_Score measure during this iterative sampling methodology. It further implies that our sampling strategy optimally constructs the target model with exactly the similar topology harnessed from the template(s), but simultaneously the model is also not exactly like the employed template(s) as the TM_Score against the template decreases in each of the successive iterations, as represented in Fig. 4.2. Quite interestingly, as represented in Fig. 4.3(a), our iterative sampling predicts an accurate near-native conformation. It is because TM_Score evaluates the correctness of the overall model topology with the considered templates as per the considered target-template alignment and the normalized DOPE score further allows us to select the model with the statistically minimal energy. Hence, the TM_Score_Answer increases in each of the successive iterations. Here in Fig. 4.3(b), the RMSD_answer decreases in each of the successive iteration, i.e. the modelling strategy slowly leaps towards the actual native target conformation and is not biased towards the template. The model, where the TM_Score and Z_Score parameters scores almost equivalently during the iterative sampling strategy with no further significant TM_Score alterations, seems to be the best sampled model and should be finally selected. It is observed that Z_Score along with TM_Score properly discriminates between the correct and bad models. Increased single sampling run with no switching and selection of the better intermediary structures does not improve the model quality, as shown by the Table 4.2 data.

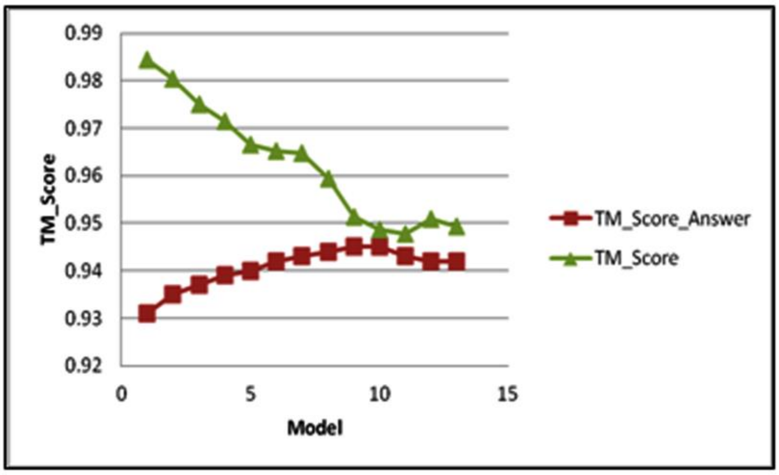


Fig. 4.2 TM_Score and TM_Score_Answer delineating the correct converging nature of our model sampling methodology.

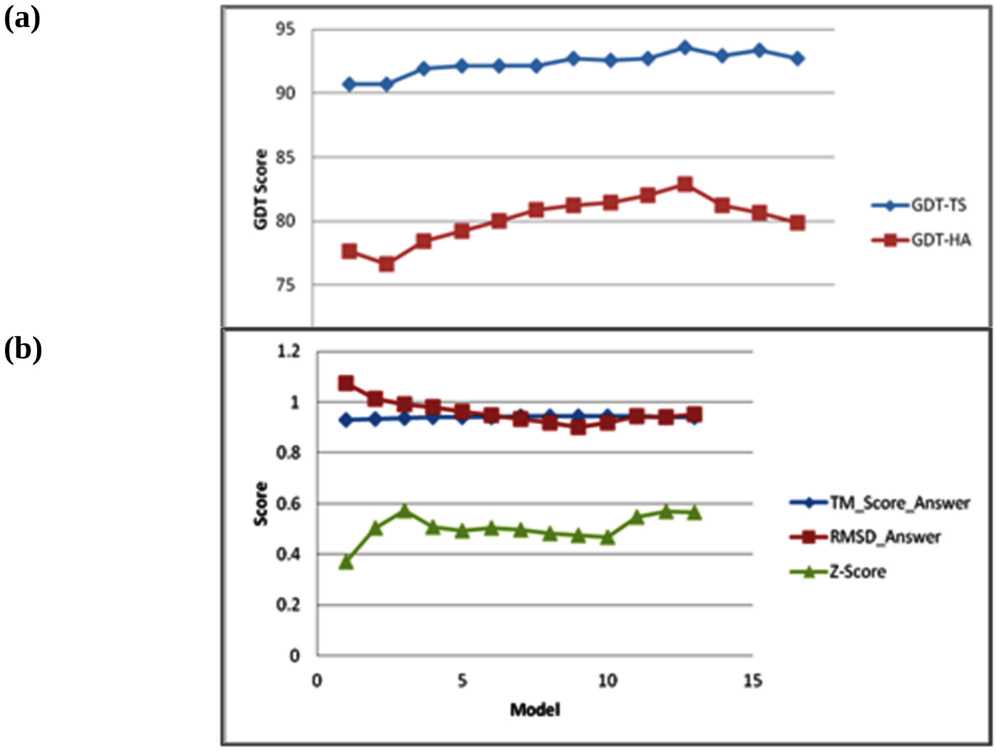


Fig. 4.3 Assessment results of **(a)** GDT-TS and GDT-HA & **(b)** RMSD, TM_Score and Z_Score in iterative optimal model sampling run.

Our sampling strategy optimally relaxes the predicted model to relieve its atomic clashes significantly faster than the conventional sampling algorithms. It is considerably better than the normal single long sampling to predict accurate models. This sampling strategy efficiently bypasses several intermittent saddle points to predict an improved model conformation (Onuchic et al. 2000). However, we still need more efficient energy functions to reliably screen the correctly modelled structures and for their sampling, we also need an improved algorithm that is extremely competent at tackling the extremely large conformational space of the target sequence. The development of an efficient model clustering method may further improve the accuracy of predicted protein models.

With the results obtained we conclude that the TM_Score and Z_Score guided sampling measure significantly improves the sampling accuracy of predicted protein models. The predicted model conformation is found to be accurate for its individual structural domains and for its overall structure with the correct mutual orientation of these domains. Our sampling-cum-assessment strategy substantially improves the accuracy of predicted protein models in comparison to their first constructed target model and is employed to build the accurate structures for all the selected CASP targets. The details of these results are discussed in Chapter V.

Chapter V

Accuracy and robustness of the template ranking, selection and iterative sampling algorithm

Chapter V

5.1 Introduction

The protein modelling problems are mainly caused due to the consideration of incorrect templates or inaccurate sampling algorithm. We have developed the template selection as well as combination algorithm and the iterative sampling strategy, as respectively reported in the chapters III and IV. The developed algorithms are employed to model the selected CASP targets. Altogether 33 domains of 21 CASP8 targets, 52 domains of 35 CASP9 targets and 31 domains of 22 CASP10 targets are considered to test our algorithm. We compare the accuracy of our models with the best predicted CASP models by assessing them against their actual native conformations. We assess the accuracy of our sampling methodology and evaluate the reliability of our employed set of assessment measures.

5.2 Comparison between our predicted models and the CASP results

The constructed target models are assessed against the employed template(s) for selecting the most accurate model which is finally evaluated for all of its encoded domains through GDT-TS, GDT-HA, RMSD and TM score against its experimentally solved native structure. However, for targets encoding multiple domains, the domains are individually assessed along with the evaluation of their topological orientation in the overall target model.

Among all the modelled target domains, we have excluded the ones that are not enlisted in the TBM-HA dataset of the CASP. Amongst the selected CASP8 target domains we could model 31, 26 and 27 target domains with improved GDT-TS score, GDT-HA score and TM_Score respectively. Out of these 31 target domains, the best predicted CASP models have shown the employed templates for 21 domains. For predicting these 21 domain structures, we have employed a different set of templates for 15 domains. All the predicted CASP8 models show an average GDT-TS, GDT-HA and TM_Score improvement of 2.958, 3.936 and 0.017 respectively along with the individual standard deviation of 2.917, 4.917 and 0.020. For the selected CASP9 target domains, we have successfully modelled 43, 42 and 43 domains with improved GDT-TS, GDT-HA and TM_Score respectively. For these target domains, the best predicted CASP models have shown the employed templates for 25 domains, out of which we have employed a different set of templates for 21 domains. All the predicted CASP9 models show an average GDT-TS, GDT-HA and TM_Score improvement of 3.641, 4.938 and 0.020 respectively along with the individual standard deviation of 5.126, 6.832 and 0.041. Similarly, for CASP10 targets, we have modelled 25, 22 and 22 target domains with improved GDT-TS, GDT-HA and TM_Score respectively. Among these target domains, the best predicted CASP models have shown the employed templates for 12 domains, out of which we have employed a different set of templates for all the 12 domains. All the predicted CASP10 models show an average GDT-TS, GDT-HA and TM_Score improvement of 3.995, 5.568 and 0.029 respectively along with the individual standard deviation of 4.382, 7.309 and 0.051. Overall for all the selected targets, our models have shown an average respective GDT-TS, GDT-HA and TM_Score improvement of 3.531, 4.814 and 0.022 along with the individual respective standard deviation of 4.142, 6.353 and 0.037. Tables 5.1, 5.2 and 5.3 respectively enlist the modelling accuracy of our predicted protein models against their experimentally solved native structures in comparison to their most accurate CASP models, as discussed further. As RMSD score does not evaluate the topologically correct substructures of a protein model (Zhang & Skolnick 2004), it is not officially employed by the CASP assessors to rank the target protein models predicted by the CASP participants (www.predictioncenter.org/CASP10/results.cgi) and is thus not enlisted here in these tables. For some of the models,

Moreover, we have not evaluated our template-ranking methodology with similar strategies employed by the I-TASSER, SWISSMODEL and MODELLER algorithms for the following two reasons. Firstly, as the objective of our template-ranking methodology is simply to choose the best set of templates for a target sequence, we have not compared the modelling accuracy of the selected templates with their computed rank-order. Secondly, as enlisted in the Tables 4, 5 and 6, we have simply restricted our discussion to only the evaluation of modelling accuracy of our predicted target structures with the most accurate CASP algorithms for all the selected target domains. Hence we have not compared the template-ranking scores of our method with the similar scores employed by all the other modelling methodologies used in the CASP test.

Table 5.1 Model assessment results in terms of GDT-TS, GDT-HA and TM_Score of the best predicted models against the best CASP8 models. The last three columns show differences in GDT-TS, GDT-HA, TM_Score of best CASP8 model and our best predicted model. A negative value indicates superior CASP model.

	Our Best Model				**Modelling accuracy better than the best predicted CASP model**		
Target/ Domain	**Templates**	**GDT-TS**	**GDT-HA**	**TM_Score**	**GDT-TS**	**GDT-HA**	**TM_Score**
T0388_D1	2P31_A, 2I3Y_A, 2R37_A	95.732	82.165	0.971	4.116	4.269	0.020
T0390_D1	1SHW_A	93.548	82.863	0.945	2.822	4.032	0.016
T0396_D1	1JR8_A	91.667	76.225	0.925	1.961	5.147	0.026
T0398	2RIR_A, 2CUK_A	96.788	81.076	0.985	9.288	15.364	0.019
T0398_D1		97.500	85.536	0.976	0.714	0.715	0.001
T0398_D2		100.000	90.476	0.986	1.020	2.381	0.000
T0400_D1	2Q7B_A	92.742	73.387	0.955	4.355	6.774	0.018
T0402_D1	2FHQ_A, 2HQ7_A	84.028	64.815	0.864	4.398	2.546	0.027
T0404_D1	2CZ4_A, 2J9C_A	97.840	85.185	0.951	8.025	16.975	0.061
T0418	2HDO_A, 2HI0_A,	88.333	67.143	0.953	0.357	-2.500	0.008
T0418_D1		91.312	71.986	0.945	0.532	-2.482	-0.006
T0418_D2		88.043	67.029	0.870	1.086	-0.725	0.011
T0422	3B9P_A, 3CF0_A, 1IN4_A	79.856	58.094	0.924	8.543	9.982	0.035
T0422_D1		82.125	61.375	0.906	2.625	1.875	0.024
T0422_D2		91.563	70.938	0.890	4.063	5.000	0.014

T0423_D1	2OTM_A, 1QAH_A	93.207	77.717	0.933	7.337	9.510	0.060
T0426_D1	2FOY_A	97.957	85.895	0.988	0.097	-2.821	-0.002
T0428_D1	1XQ9_A, 1T8P_A, 1FZT_A	98.035	87.118	0.988	0.983	4.258	0.005
T0432_D1	2DKW_A, 1E6I_A, 2YQD_A	94.231	79.038	0.953	6.346	6.730	0.033
T0435_D1	2QPW_A	88.864	72.273	0.919	4.773	5.909	0.035
T0438	2OAS_A, 2G39_A	80.879	58.204	0.952	1.163	1.873	0.000
T0438_D1		85.671	66.768	0.923	-2.134	-2.287	-0.017
T0438_D2		93.834	77.803	0.970	0.673	1.121	0.001
T0442	2PIF_A	83.085	61.915	0.940	0.213	2.128	0.007
T0442_D1		94.268	75.478	0.955	1.274	-2.547	-0.003
T0442_D2		97.603	81.507	0.944	0.000	1.028	-0.001
T0444_D1	1SMQ_A, 1H0O_A	95.404	83.180	0.982	0.735	1.562	0.001
T0447_D1	1EG7_A	91.697	76.245	0.980	3.367	5.212	0.009
T0458_D1	2OKA_A	99.675	93.182	0.976	1.948	8.442	0.022
T0470	2QGS_A	88.165	68.617	0.932	5.718	6.649	0.022
T0470_D1		86.712	67.793	0.900	1.126	0.000	0.013
T0470_D2		88.584	69.364	0.927	47.255	32.514	0.508
T0499_D1	2ONQ_A	92.857	74.107	0.887	7.143	9.821	0.070
				Average	**2.958**	**3.936**	**0.017**
				Standard Deviation	**2.917**	**4.917**	**0.020**

Table 5.2 Model assessment results in terms of GDT-TS, GDT-HA and TM_Score of the best predicted models against the best CASP9 models. The last three columns show differences in GDT-TS, GDT-HA, TM_Score of best CASP9 model and our best predicted model. A negative value indicates superior CASP model.

	Our Best Model				Modelling accuracy better than the best predicted CASP model		
Target / Domain	**Templates**	**GDT-TS**	**GDT-HA**	**TM_Score**	**GDT-TS**	**GDT-HA**	**TM_Score**
T0521	3KHE_A	47.321	35.417	0.533	2.083	3.423	0.000
T0521_D1		81.888	61.224	0.842	4.082	3.061	0.016
T0521_D2		86.428	63.929	0.862	-0.716	-1.785	-0.014
T0522_D1	3I4S_A	98.282	88.740	0.981	3.625	6.106	0.012
T0523_D1	3LYX_A	93.019	77.027	0.933	4.731	7.658	0.033
T0528	3I45_A	67.588	45.216	0.867	19.273	19.677	0.081
T0528_D1		82.108	59.314	0.927	-4.167	-6.617	-0.011
T0528_D2		58.281	35.000	0.705	-1.719	-0.781	-0.052
T0530_D1	2K5Q_A	88.125	69.375	0.876	5.625	5.625	0.030
T0538_D1	1J7K_A	98.585	88.208	0.946	7.076	12.736	0.068
T0541_D1	3IDU_A	87.255	65.686	0.914	6.618	6.862	0.038
T0559_D1	2VXZ_A	94.403	73.507	0.915	1.493	-0.374	0.009
T0560_D1	1R1U_A	94.533	83.203	0.917	2.345	9.375	0.013
T0563_D1	1OC1_A, 1UNB_A, 1GP4_A	76.438	55.420	0.878	-0.111	-0.332	0.003

T0566_D1	1USV_B	79.808	57.885	0.879	3.270	3.462	0.019
T0567_D1	1NY5_A	84.072	64.630	0.900	4.628	6.667	0.026
T0570_D1	2PZ0_A, 3L12_A	86.588	66.202	0.942	5.900	9.120	0.011
T0580_D1	2WY2D	89.904	72.356	0.928	0.241	-0.961	0.009
T0586	2DU9_A, 3BY6_A	90.336	76.261	0.917	14.495	23.530	0.051
T0586_D1		87.188	75.000	0.869	-5.625	-3.438	-0.061
T0586_D2		97.437	85.897	0.903	8.974	18.589	0.125
T0589	1WU7_A	69.026	46.520	0.898	8.991	9.513	0.042
T0589_D1		74.781	51.096	0.889	0.549	0.328	-0.005
T0589_D2		64.939	41.768	0.664	-12.805	-14.939	-0.132
T0589_D3		87.766	68.085	0.894	-5.319	-7.713	-0.032
T0594_D1	1X53_A	91.964	75.357	0.953	7.321	11.607	0.051
T0596	3C07_A	70.259	45.977	0.847	12.500	12.500	0.079
T0596_D1		98.585	82.075	0.925	2.830	1.886	0.010
T0596_D2		66.322	42.355	0.770	4.958	4.752	0.040
T0599_D1	3HWO_A, 2FN0_A, 3H9M_A	87.158	66.189	0.958	5.601	5.123	0.037
T0600	3LYX_A, 3EEH_A	48.821	34.198	0.507	5.425	4.953	0.016
T0600_D1		84.746	61.017	0.786	5.932	2.119	0.025
T0600_D2		93.085	77.128	0.870	3.191	6.915	0.051
T0601_D1	1VPB_A, 1VL4_A	86.961	66.100	0.966	2.891	4.705	-0.001
T0602_D1	3H3M_A, 3A7M_A, 2FUP_A	90.000	78.182	0.864	0.455	4.091	0.001

T0605_D1	2NPS_A, 1H89_A	98.957	91.665	0.954	5.727	15.623	0.074
T0611	1UI5_A	61.558	38.442	0.778	10.427	9.296	0.087
T0611_D1		98.558	83.654	0.935	0.480	0.000	0.004
T0611_D2		56.164	32.877	0.684	7.534	9.247	0.009
T0613_D1	3LOU_A	95.709	80.504	0.979	1.772	2.426	0.007
T0614_D1	2CY5_A, 1EAZ_A	93.660	77.817	0.915	6.688	9.155	0.040
T0619_D1	3LYS_A	90.594	74.010	0.913	1.980	3.465	-0.004
T0620_D1	3MTC_A	86.760	72.038	0.940	-0.696	-2.614	0.010
T0626_D1	3LOU_A	90.426	73.138	0.967	0.798	1.329	0.002
T0629	1OCY_A, 2FKK_A	33.218	25.231	0.380	5.672	6.134	0.036
T0629_D1		92.544	74.561	0.898	9.211	12.719	0.097
T0629_D2		12.421	7.547	0.157	4.245	3.459	0.013
T0632_D1	3DKZ_A, 1Q4T_A, 1SC0_A, 2FS2_A	96.711	81.578	0.961	4.167	5.262	0.017
T0634_D1	3JTE_A, 3GT7_A	94.626	83.178	0.951	4.205	10.514	0.019
T0635_D1	3MN1_A, 3IJ5_A	98.913	89.441	0.987	-0.155	0.311	0.000
T0636_D1	3HDO_A, 1FG7_A, 3EZ1_A, 3CQ5_A	81.918	61.164	0.943	0.236	0.315	0.002
T0640_D1	1IY8_A, 3KVO_A, 3IOY_A	84.887	65.113	0.925	2.401	2.684	0.018
				Average	**3.641**	**4.938**	**0.020**
				Standard Deviation	**5.126**	**6.832**	**0.041**

Table 5.3 Model assessment results in terms of GDT-TS, GDT-HA and TM_Score of the best predicted models against the best CASP10 models. The last three columns show differences in GDT-TS, GDT-HA, TM_Score of best CASP10 model and our best predicted model. A negative value indicates superior CASP model.

	Our Best Model				Modelling accuracy better than the best predicted CASP model		
Target / Domain	**Templates**	**GDT-TS**	**GDT-HA**	**TM_Score**	**GDT-TS**	**GDT-HA**	**TM_Score**
T0645_D1	3GZS_A, 3EHN_A	76.958	56.426	0.931	-1.556	-3.464	-0.003
T0650_D1	2OMU_A	92.848	76.106	0.974	0.370	1.548	0.000
T0657_D1	1BWN_A, 2DHI_A	89.662	72.744	0.927	7.143	11.842	0.039
T0659_D1	3LD7_A	99.324	90.203	0.968	5.743	11.487	0.045
T0662_D1	3GZL_A	91.118	75.000	0.893	8.222	15.461	0.068
T0663	2GU3_A	39.358	29.223	0.450	7.432	8.108	0.042
T0663_D1		59.524	41.964	0.618	-0.297	2.381	-0.027
T0663_D2		89.455	70.313	0.860	1.564	-2.734	0.001
T0664_D1	3CGH_A	82.848	62.526	0.946	-0.520	-3.586	0.007
T0674	3H41_A, 2K1G_A	36.466	29.052	0.411	0.173	0.949	-0.047
T0674_D1		10.377	3.931	0.174	-52.988	-39.780	-0.556
T0674_D2		81.298	64.313	0.860	-1.145	-1.718	-0.002
T0675	2CT1_A	42.667	30.333	0.436	8.334	7.000	0.069
T0675_D1		92.595	76.852	0.693	11.114	13.889	0.168
T0675_D2		88.333	67.500	0.633	3.333	6.667	0.041

T0689_D1	3FZX_A	88.943	70.192	0.951	2.765	4.447	0.011
T0692_D1	3N83_A, 3IWK_A	81.848	59.457	0.955	2.663	2.174	0.003
T0708_D1	3IRV_A, 3OT4_A	86.640	67.063	0.931	2.645	0.529	0.036
T0712_D1	3U22_A	94.127	80.572	0.956	1.808	4.066	0.008
T0714_D1	2E6P_A	96.307	80.682	0.951	6.818	11.080	0.039
T0716_D1	2CRA_A, 1SAN_A	99.510	92.647	0.960	5.883	19.608	0.084
T0721_D1	3FBSA, 3F8PA	80.782	60.374	0.934	1.870	3.486	0.005
T0726	1Z1W_A	36.981	20.329	0.645	4.239	4.455	0.020
T0726_D1		48.182	26.648	0.780	4.602	3.353	0.042
T0726_D2		20.625	12.500	0.224	-66.875	-56.563	-0.663
T0726_D3		22.414	15.517	0.180	0.862	-0.431	-0.037
T0731_D1	1SKN_P	99.545	84.545	0.941	18.636	27.272	0.187
T0747_D9	3D33_A	77.528	59.551	0.784	8.146	10.955	0.045
T0749_D1	3EU8_A, 3QWT_A	91.730	76.441	0.971	0.251	-1.754	-0.002
T0752_D1	4STD_A, 3EF8_A	89.041	70.719	0.920	0.342	-0.856	-0.008
T0757_D1	2ESS_A, 2OWN_A	86.975	68.277	0.926	4.412	5.252	0.010
				Average	**3.995**	**5.568**	**0.029**
				Standard Deviation	**4.382**	**7.309**	**0.051**

5.3 Targets with improved accuracy

Unlike conventional template ranking and selection measures, our algorithm employs the template information from both the pairwise as well as MSA alignments through multiple scoring parameters to robustly rank the templates. On the basis of pairwise alignment information for a target sequence, we assessed all the hits on the basis of their BLOSUM62 score, coverage span, affine gap penalty score, average proportion of mismatched hydrophobic residues, hydrophilic residues and sequence identity, average fraction of total number of gaps and the residues, average gap length longer than 5 gaps and the proportion of mismatched residues. However, the functionally similar protein sequences keep evolving in their diverse cellular micro-environments along with the conservation of their overall structural topology. While also knowing that incorporation of additional templates could actually improve the modelling accuracy over the single best available template, we assessed the structural similarity of the selected hits with the top ranked seed template through their MSA by using several scoring measures to screen their best complementary set for a target sequence. The employed scoring schemes assessed the conformational diversity of the hits against the seed template and the target sequence by separately computing the BLOSUM62 score against the target and the seed template sequence, coverage span, affine gap penalty, average proportion of mismatched hydrophobic residues, hydrophilic residues and sequence identity and the proportion of unique residues encoded in a template and not encoded in the other selected hits. Moreover, MSA based assessment of hits evaluated structural similarity of a template in terms of TM_Score, GDT- TS and fraction of the topologically correct residues fitting within 8Ǻ distance deviation against the equivalent residues of the seed template.

The developed algorithm attempted to solve the modelling problems normally incurred due to the consideration of culled PDB dataset. Among structurally similar templates, the most alike conformation is normally employed as the representative hit by conventional modelling algorithms. However, our algorithm considered all these structurally similar HHPred hits to respectively assess their sequence and structural similarity with the target sequence and the seed template. It allowed us to properly evaluate the set of redundant hits and select the best representative template that is biologically related to a target sequence. Our template ranking algorithm consequently allowed us to select the best set of structurally similar templates for maximally spanning the target sequence and constructing its biologically meaningful model. The algorithm consistently constructed trustworthy protein structures and improved the modelling accuracy because it reliably spanned the target segment(s) with best set of templates sharing a significant structural as well as sequence similarity with the seed template. It has already been observed that consideration of the low-ranked templates or the template fragments along with the seed-template never improves the modelling accuracy over the seed template always (Zheng et al. 2010) and our algorithm efficiently resolves such modelling errors. Our template selection algorithm eliminates the modelling errors that are usually caused due to incorrect template selection or unreliable assembly of structural fragments. For example, our algorithm employed 3 templates (2P31_A, 2I3Y_A and 2R37_A) in comparison to the best CASP8 model (pro-SP3-TASSER) which employed 5 templates (2P31_A, 2GS3_A, 2I3Y_A, 2P5Q_A and 1GP1_A) to construct the domain1 model of the target T0388. Our algorithm efficiently employed conformational diversity of three templates to construct a target model. Assessing against its native conformation, our model shows GDT-TS, GDT-HA, TM_Score and RMSD scores of 91.032, 77.730, 0.936 and 0.889 in comparison to the respective scores of 86.351, 73.420, 0.898 and 0.970 of the best predicted CASP8 structure. For this T0388 domain1 shown in the Figure 5.1, a comparison of the best CASP8 model, our predicted model and the actual structure of this target reveals that our template set constructs a better model of the 5 residue C-terminal loop segment (KKEDL).

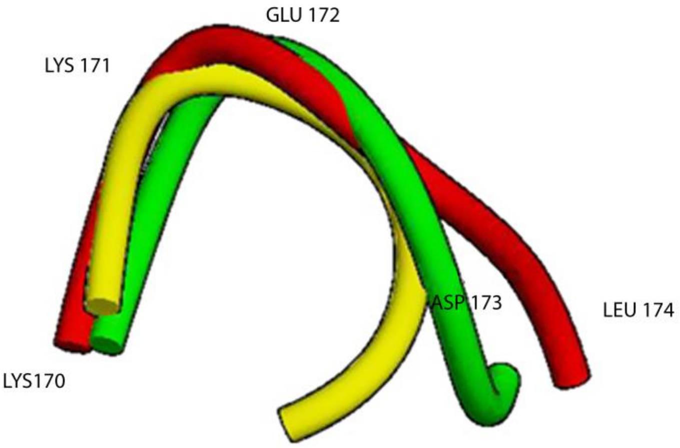


Fig. 5.1 Superimposed 170-174 residue segment of constructed model (Green), best predicted CASP8 model (Yellow) and experimental native conformation (Red). The main chain backbone conformation is shown by tube representation. TM_Score assessment of this 170-174 residue segment of the best predicted CASP model against the experimental structure is 0.36381 and that of our model against the experimental structure is 0.76125.

For the CASP9 target T0528, our modelling algorithm employed a single template 3I45_A in comparison to the best predicted CASP9 model (LEE) which used 7 templates (1QO0_A, 3I09_B, 3I45_A, 3EAF_A, 3LKB_A, 3LOP_A and 3H5L_B). Our algorithm could maximally cover the target with only the seed template and it did not employ additional hits for the reason that these hits did not share a significant structural similarity with the seed template. Assessing against the native structure, our model shows GDT-TS, GDT-HA, TM_Score and RMSD scores of 67.588, 45.216, 0.867 and 2.459 in comparison to the respective scores of 48.315, 25.539, 0.786 and 3.284 of the best predicted CASP9 model. A comparison of the best CASP9 model, our model and the native structure reveals that our selected set of templates constructs better topology for a 16 residue helical segment (GGDPLTKLQDMDPKRY), as shown in the Figure 5.2.

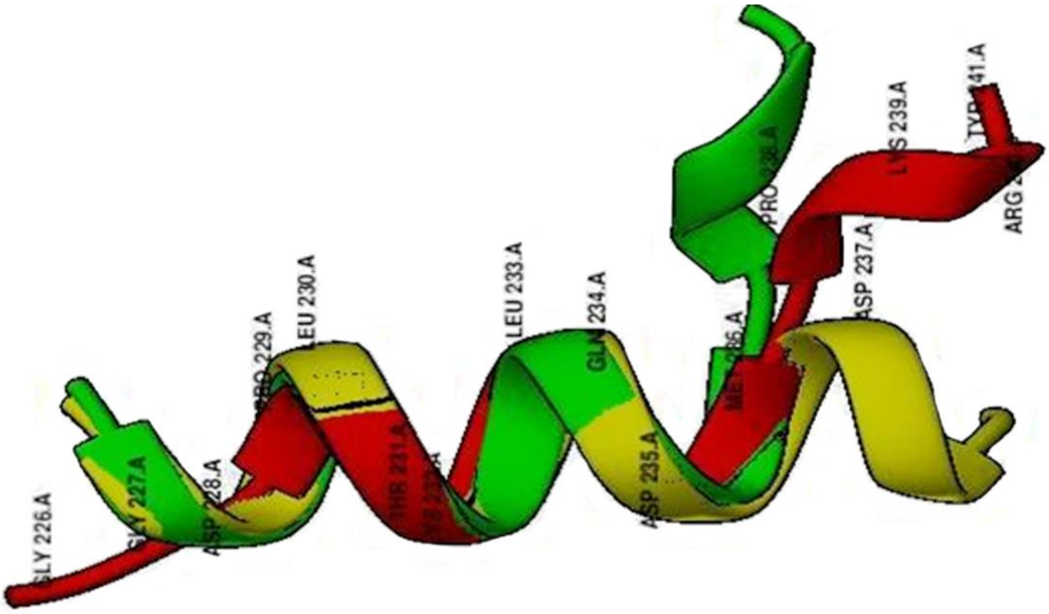


Fig. 5.2 Superimposed 226-241 residue segment of constructed model (Green), best predicted CASP9 model (Yellow) and experimental native conformation (Red). The main chain backbone conformation is shown by tube representation. TM_Score assessment of this 226-241 residue segment of the best predicted CASP model against the experimental structure is 0.36512, and that of our model against the experimental structure is 0.41268.

For the CASP10 T0731 Domain1 target, our algorithm used only the seed template 1SKN_P to maximally cover the target in comparison to the best CASP10 model (MATRIX) which used 4 templates (2KZ5_A, 1SKN_P, 2WT7_B and 3A5T_A). Assessing against the native structure, our model shows GDT-TS, GDT-HA, TM_Score and RMSD scores of 99.545, 84.545, 0.941 and 0.629 in contrast to the respective scores of 80.909, 57.273, 0.754 and 1.238 of the best predicted CASP10 model. A comparison of the best CASP10 model, our model and the native structure reveals that our algorithm constructs a better topology for the 10 residue segment (NEMMSKEQFN) which is a 7 residue helix segment (NEMMSKE) linked to a 3 residue loop segment (QFN), as shown in the Figure 5.3.

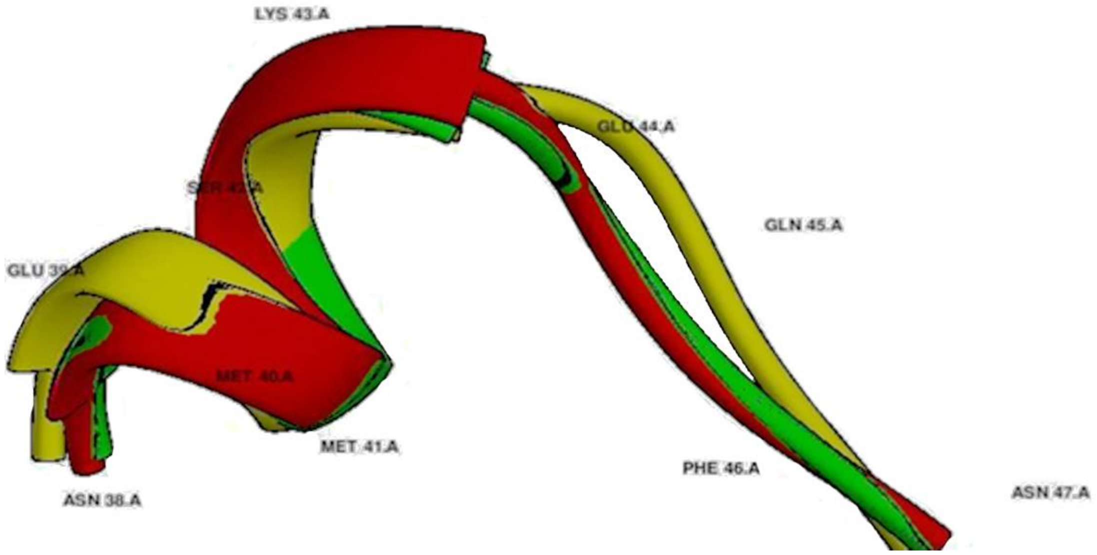


Fig. 5.3 Superimposed 38-47 residue segment of constructed model (Green), best predicted CASP10 model (Yellow) and experimental native conformation (Red). The main chain backbone conformation is shown by tube representation. TM_Score assessment of this 38-47 residue segment of the best predicted CASP model against the experimental structure is 0.48265, and that of our model against the experimental structure is 0.76429.

The combination of minimal number of correct, mutually complementary and the top-ranked templates in a structural topology guided sequence alignment significantly improves the modelling accuracy. The notably high modelling accuracy of almost all the CASP TBM-HA targets has shown that our algorithm is successful in consistently predicting the accurate near-native conformations for most of the target sequences, as shown for the CASP8 target T0398, CASP9 target T0586 and CASP10 target T0747 in the Figures 5.4, 5.5 and 5.6 respectively. These snapshots represent an overall topology of the predicted target models in comparison to their best predicted CASP structures against their native conformations.

For the CASP8 target T0398 encoding 292 residues, our algorithm used 2 templates (2RIR_A and 2CUK_A) in comparison to the best predicted CASP8 model (MUSTER) which employed 7 templates (2RIR_G, 2RIR_D, 2RIR_F, 2RIR_B, 2RIR_C, 2RIR_A and 2RIR_E). Assessing these predicted models against the native target structure for the structural domain (Residue segment 1-292), our predicted model shows the GDT-TS, GDT-HA, TM_Score and RMSD scores of 96.788, 81.076, 0.985 and 0.774 in comparison to the respective scores of 87.500, 65.712, 0.966 and 1.198 for the best predicted CASP8 model. A comparison of the best CASP8 model, our model and the experimental structure of this target sequence reveals that our set of templates constructs topologically more accurate T0398 model, as best represented by the violet colored encircled region in Figure 5.4. This region shows structural closeness of our model to its native structure in comparison to the best predicted CASP8 model.

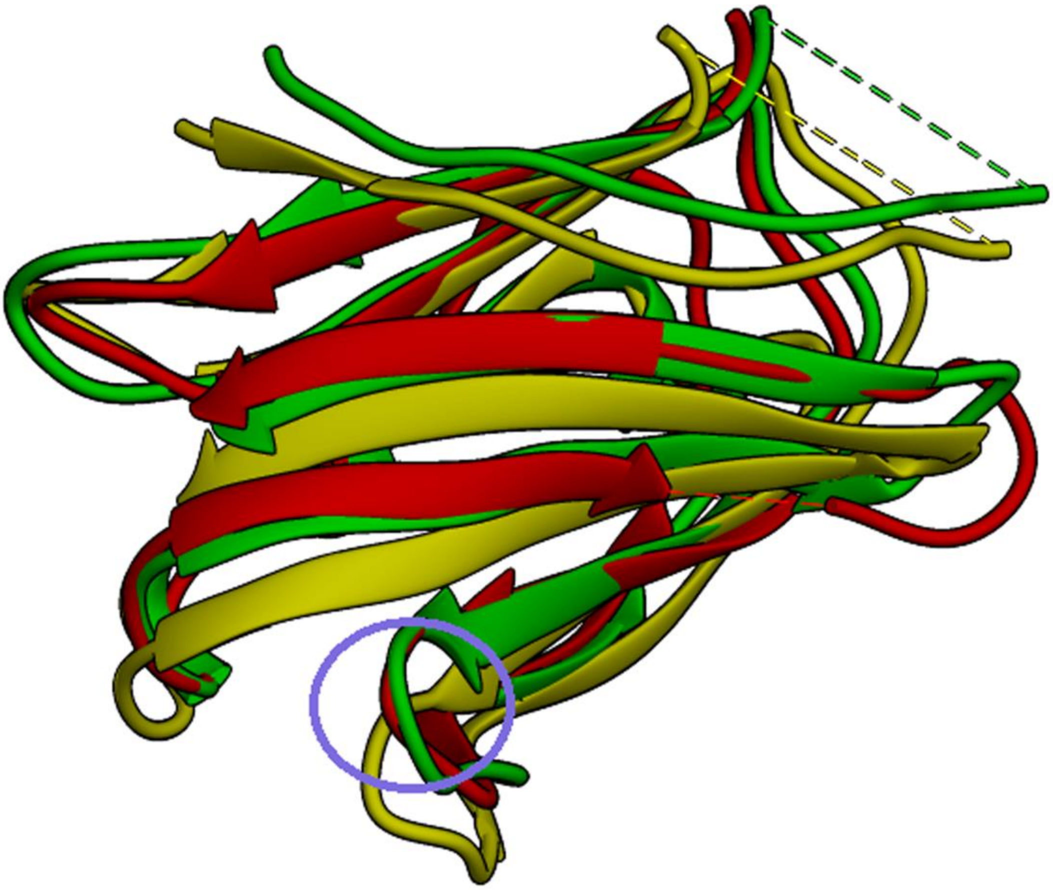


Fig. 5.4 Superimposed constructed model (Green), best CASP8 model (Yellow) and the experimental native conformation (Red) for the CASP8 target T0398. The main chain backbone conformation is shown by tube representation.

For the CASP9 target T0586 encoding 125 residues, our algorithm employed only 2 templates (2DU9_A and 3BY6_A) in comparison to the best predicted CASP9 model (Zhang) which employed 5 templates (3BY6_E, 3BY6_A, 3BY6_B, 2EK5_C and 3IC7_A). Assessing these predicted models against the native target structure for the structural domain (Residue segment 1-125), our predicted model shows the GDT-TS, GDT-HA, TM_Score and RMSD scores of 90.336, 76.261, 0.917 and 0.852 in comparison to the respective scores of

75.841, 52.731, 0.866 and 1.671 for the best predicted CASP model. A comparison of best CASP model, our model and the native target structure reveals that our set of templates constructs topologically more accurate T0586 model, as best represented by the violet colored encircle region in Figure 5.5. This region shows structural closeness of our model to its experimental structure in comparison to the best CASP model.

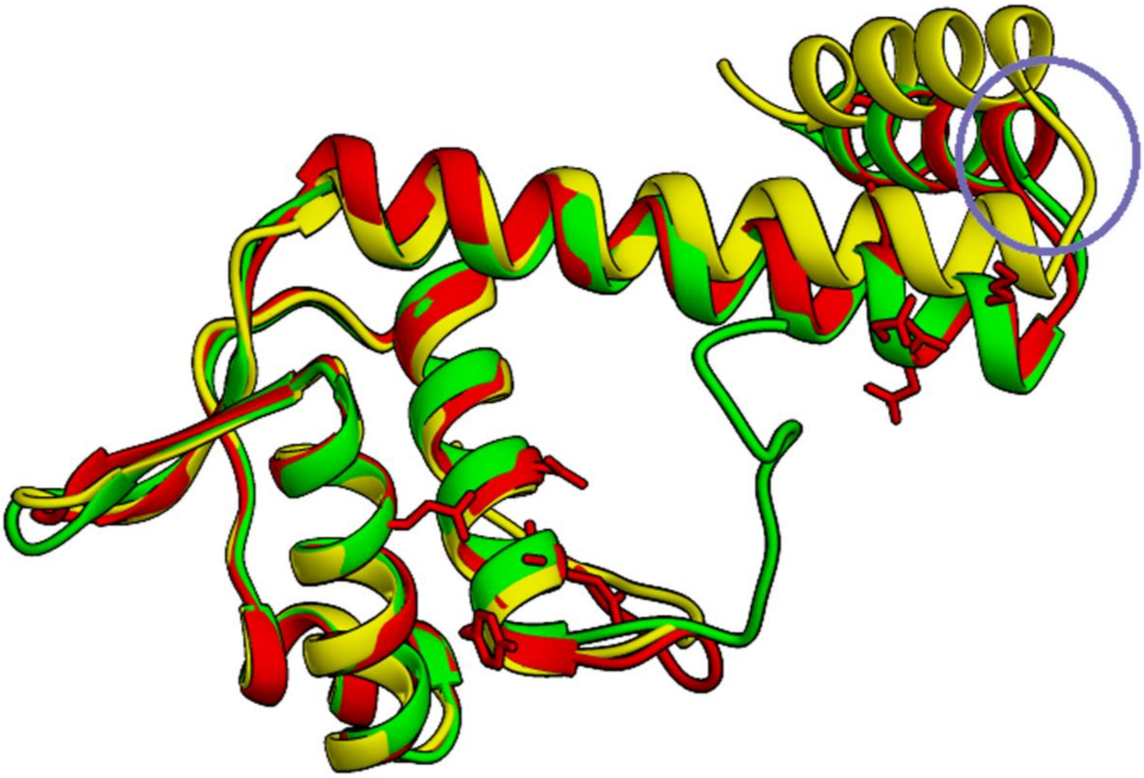


Fig. 5.5 Superimposed constructed model (Green), best CASP9 model (Yellow) and the experimental native conformation (Red) for the CASP9 target T0586. The main chain backbone conformation is shown by tube representation.

For the CASP10 target T0747 model encoding 121 residues, our algorithm used only a single template (3D33_A) in comparison to the best CASP10 model (Mufold-MD) which employed 2 templates (3SD2_A and 3D33_A). Assessing these models against the native target structure for the structural domain (Residue segment 24-34 and 43- 121), our predicted model shows the GDT-TS, GDT-HA, TM_Score and RMSD scores of 77.528, 59.551, 0.784 and 1.119 in comparison to the respective scores of 69.382, 48.596, 0.739 and 1.414 for the best predicted CASP model. A comparison of best CASP model, our model and the actual target structure reveals that our set of templates constructs topologically more accurate T0747 model, as best represented by the violet colored encircle region in Figure 5.6. This region shows the structural closeness of our predicted model with its experimental structure in comparison to the best predicted CASP10 model.

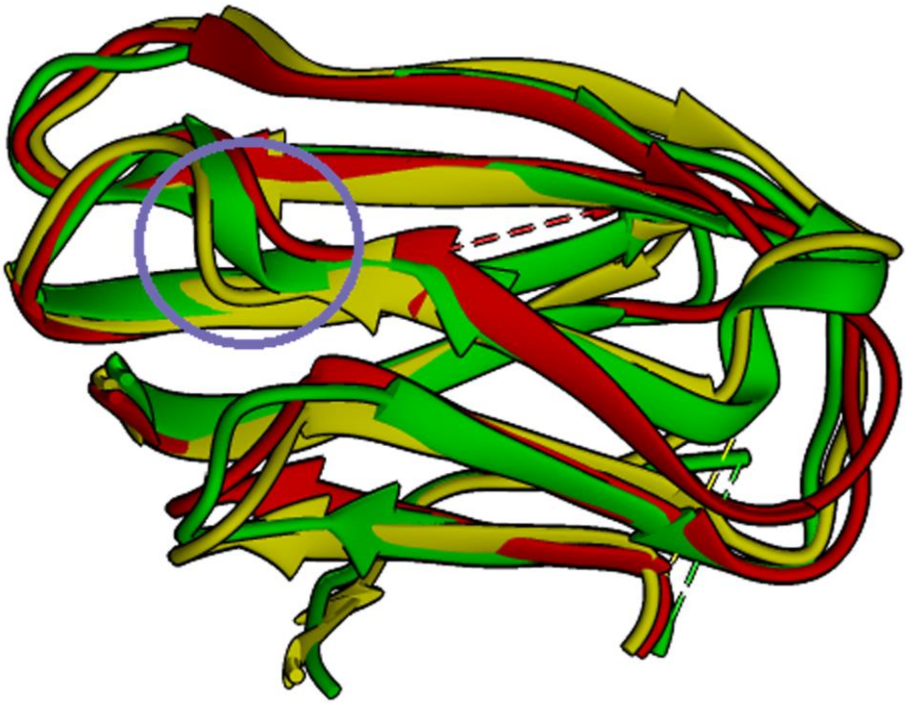


Fig. 5.6 Superimposed constructed model (Green), best CASP10 model (Yellow) and the experimental native conformation (Red) for the CASP10 target T0747. The main chain backbone conformation is shown by tube representation.

5.4 Domains without improved modelling accuracy

Our algorithm could not predict more accurate models for five CASP8, nine CASP9 and nine CASP10 domains. It has been noticed that a few of these domains are modelled through non-comparative modelling algorithm and are not discussed here. For example LEE and SAMUDRALA in CASP8, fams-ace3, United3D and McGuffin in CASP9 and LEE and Mufold in CASP10 did not employ the true comparative modelling algorithm to model the target sequence. Moreover, certain algorithms viz. EB_AMU_Physics of CASP8 and LTB of CASP9 are not available in the official CASP abstract booklet and are also not discussed in the following passage that briefly summarizes the major CASP algorithmic steps which are not employed by our modelling algorithm. These algorithms are streamlined to explain their conceptual methodology in a better way.

MULTICOM-RANK algorithm similarly employed the significant templates (E-value lesser than 10^{-20} and target coverage span more than 75%) along with the statistically insignificant ones to maximally cover the target sequence. Phyre2_A and Jones-UCL however compared the HMM profiles (Eddy 2011) of the target and template sequences to estimate their residue contact maps through PSICOV (Jones et al. 2012) and to construct their reliable alignment for selecting the best set of template fragments and constructing the threading based target model with the FRAGFOLD (Kosciolek & Jones 2014) as well as simulated annealing based algorithms. Seok-Server algorithm in CASP9 employed the HHPred to search the reliable hits for constructing their MSA through PROMALS3D (Pei et al. 2008). It further constructed the target models on the basis of the selected top-ranked templates through MODELLER-CSA protocol (Joo et al. 2008; Joo et al. 2009) to later refine their unreliable segments (loops or terminal ends) and rebuild their side chains for improving the structural topology. Contrarily, the MUFOLD-MD and MULTICOM-NOVEL algorithms extensively employed the template search methods to even select the evolutionarily distant as well as reliable templates for constructing the accurate target models. MULTICOM-CONSTRUCT also screened the reliable templates (Söding 2005; Karplus et al. 1997; Finn et al. 2011) to construct the alternative target-template alignments and build about 150,000 – 250,000 models for each target sequence. FEIG algorithm, although employed the template search algorithms (Söding et al. 2005; Jaroszewski et al. 1998; Zhang et al. 2005) along with the threading based models of TASSER and I-TASSER methodologies (Zhang et al. 2005; Wu et al. 2007) to construct the target structures. The MULTICOM algorithm in CASP8 moreover evaluated all of its predicted models through the ModelEvaluator (Wang et al. 2009) tool to consider the top 50% structures for extracting their best set of structural fragments and for efficiently annealing them the finest way to construct an accurate target structure. Baker algorithm in CASP10 however parsed the target sequences into several structural domains to separately model them with the reliable templates and ROSETTA *de-novo* fragment assembly approach (Bonneau et al. 2002; Leaver-Fay et al. 2011) for annealing them into an overall target structure by an extensive conformational sampling of about 100000-300000 models. Zhang-IRU however, directly assembled the 9-mer template fragments followed by the 3-mer template segments with Edafold (Simoncini et al. 2012) probabilistically to further repack the side–chains in the overall target structure and construct the minimal energy target model through ROSETTA (Gront et al. 2011). Pcons-net similarly employed the threading methods (Söding 2005; Wallner et al. 2007; Remmert et al. 2012; Wu

& Zhang 2007) along with the ROSETTA *de-novo* modelling methodology (Leaver-Fay et al. 2011) to construct a compact target structure. PconsM however realigned the top 20 Pcons-net predicted target models with HHsuite (Söding 2005) to construct their diverse set of alternative structural alignments and to merge their different structural segments in all the possible combinations for building the accurate target model (Larsson et al. 2009).

Altogether for some of the CASP target domains, the most accurate CASP algorithms threaded the target sequence through the top-scoring set of numerous templates or alternative target-template alignments, fixed the loop segments and side-chains additionally in the constructed target models and extensively simulated the target structure through computationally complex MD or simulated annealing methodologies. Due to consideration of these non-ideal comparative modelling steps or the additionally employed modelling tactics over the already constructed target structures, some of the CASP algorithms predicted more accurate models than our constructed structures for only some of the selected target domains.

5.5 Modelling steps where we improved the prediction accuracy

Template ranking helps a lot in improving the accuracy of TBM algorithm. The template ranking is likely to improve the protein model quality by the satisfaction of two conditions. One is that gapped regions in the alignment to the first best selected template can be covered by the complementary alignment with the other template. If multiple such complementary templates are available for the considered gapped region, the template ranking helps us to choose the best template for the considered sequence segment. Another reason for the modelling accuracy is even more consistently accurate. Ranking helps us to choose the best set of structurally similar templates. It has been observed that when any of these conditions is not satisfied by the selected set of templates for a target sequence, predicted models normally diverge from their native conformations.

The template ranking method employs the complete available information for a template from both pairwise and multiple sequence alignments. It aims to use the minimal number of correct templates for a target sequence. The algorithm solves the modelling problems that are caused due to the consideration of incorrect templates or the erroneous set of alternative target-template alignments and the illogical threading combination of structural fragments extracted from the templates unrelated with the considered target sequence (Cheng et al. 2005; Jones et al. 2005; Zhang et al. 2005; Wu & Zhang 2008; Moshe et al. 2009; Tianyun et al. 2009).

It is very important that the template ranking algorithm should correctly rank the templates through a set of multiple structural and alignment parameters. The approach has been implemented to make the ranking and modelling algorithm robust enough for selecting the correct templates from a set of promiscuously similar templates. We have not allowed iterative improvements or overtraining of the algorithm parameters for only a few CASP targets and rather we employed the same set of parameters by considering the complete available alignment information to make the algorithm robust enough at handling the diverse target proteins. Considering all the functionally similar HHPred templates has also allowed the feasible application of our algorithm even for the difficult modelling cases, where we have only a few templates sharing an utmost 35% identity with the target sequence.

It is interesting to observe that an erroneous topological pool of coordinates in a model normally generated from the consideration of multiple unrelated templates is successfully stabilized through our template ranking approach employed along with a structurally balanced sequence alignment. The template ranking algorithm also aids us to easily avoid the consideration of promiscuous template matches for the target sequences. The approach would assist the development of accurate protein modelling tools and servers to allow us quickly bridge the ever-increasing sequence-structure gap.

For all the selected CASP8, CASP9 and CASP10 targets, we have improved the modelling accuracy with a huge respective sum GDT-TS, GDT-HA and TM_Score margin of 327.241, 480.346 and 1.681 in comparison to the best predicted CASP models, as orderly enlisted in the Tables 5.1, 5.2 and 5.3. Lastly, the significantly high modelling accuracy of the constructed target models prove that our predicted structures are definitely the result of correct structural reorientation of the secondary structure segments extracted from the correct templates and it may also be due to the higher coverage of the target sequence through the reliable set of structurally similar templates.

Chapter VI

Conclusion and Future Perspectives

6.1 Conclusion

With the increasing number of templates in PDB, a multi-template based TBM algorithm has become increasingly useful to predict protein structures. It has already been shown that a target sequence can be more accurately modelled through the consistent distance restraints of multiple templates in comparison to the single best available template (Zemla, 2003; Zhang et al. 2005). Despite the development of several TBM or threading algorithms, the template selection and combination step still has some limitations such as the inability to select and employ the best set of templates for constructing the accurate target models.

In comparison to the conventional TBM algorithms that evaluate pairwise alignment or MSA of a target sequence against the templates to select the reliable templates, we have evaluated both the pairwise alignment as well as MSA information of templates against a target sequence through a diverse set of scoring parameters. Employing these pairwise and MSA scores of templates, we have developed a template ranking-cum-selection methodology for selecting the best set of templates. Our algorithm evaluates the pairwise sequence similarity of a target sequence against the template hits through the key scoring parameters viz. coverage span and aligned charges. Moreover, it also evaluates the structural similarity of a hit against the seed template through several important scoring parameters viz. TM_Score, GDT-TS and the proportion of topologically correct residues fitting within 8Ấ distance deviation. The consideration of all these scoring parameters empowers our developed algorithm to correctly rank the templates and select their best set for a target sequence.

To sample the conformational space of a target sequence, we have developed an iterative sampling algorithm. Unlike the conventional single long model sampling technique that does not consistently improves the topology of a target structure, our algorithm employs the best set of model assessment measures (TM_Score and Z_Score) to guide the optimal iterative sampling of the target conformation and construct a model that is more accurate than the first constructed target model.

Among the 33, 52 and 31 domains present in the selected CASP8, CASP9 and CASP10 TBM-HA targets, our developed algorithm has predicted models with improved GDT-TS score for 31, 43 and 25 domains in comparison to their best structures predicted during the CASP. Overall for all the selected CASP target domains, our predicted models have shown an average respective GDT-TS, GDT-HA and TM_Score improvement of 3.531, 4.814 and 0.022 over the best structures predicted during the CASP.

We have programmed our complete template-ranking-cum-selection and sampling scripts in the “C”, Python and PERL languages. The developed algorithm significantly improves the accuracy of predicted protein models and it would certainly pave way for the development of automated protein modelling tools that can be integrated with the genome sequencing experiments.

6.2 Future Perspectives

To understand the cellular system the best possible way, the detailed knowledge of protein structures is essential to understand all of their functions. Despite the phenomenal research currently being done in this field, protein structure prediction is one of the most challenging problems of bioinformatics to structurally understand the complete proteome of a cell system. We have improved the modelling accuracy of several CASP target domains and have attempted to solve the modelling problems that are usually caused due to the inaccurate template selection and model sampling steps. However, further advancements are still needed to overcome the shortcomings of each of the steps of a TBM algorithm for predicting the accurate protein structures.

Although several HMM based template search algorithms have been developed (Söding 2005; Biegert & Söding, 2009; Hildebrand et al. 2009; Remmert et al. 2012), the entire set of correct and biologically related templates is not resulted by any of these methodologies for a target sequence. Though being probabilistic, these HMM profile based template search algorithms are unable to search the correct templates that share a distant evolutionary relationship with the target sequence. Moreover, these algorithms also result in some false positive and spurious templates that share a significant homoplastic sequence similarity with the target sequence due to the following three reasons. Firstly, the template selection error starts due to greedy nature of PSI-BLAST (employed by HHPred to construct the target-template scoring profiles) as it also considers non-homologous template fragments at the sequence-ends of the high-scoring centrally aligned segments of a target sequence (Gonzalez & Pearson, 2010). To solve this problem, HHSearch aligns the target-template sequences and considers only the high-scoring segments by pruning the terminal unaligned segments, although yet another problem arises. HHSearch aligns the high-scoring local folds or conserved domains of templates with the target sequence segments to result in an overall sequence profile of the target sequence and these local folds of different templates are usually never the complimentary sub-structures which can be reliably juxtaposed to result in an accurate target model. Current TBM algorithms merely search and combine the local template folds to maximally cover the target without worrying for their overall correct mutual agreement with other target residues, possibly covered with the other templates (Cheng 2007; Gront et al. 2011; Simoncini et al. 2012). Secondly, the prefixed values of all the parameters viz. E-value cut-off, E-value inclusion threshold and the number of considered profile iterations, that are normally used to construct a HMM profile of a target or a template sequence, may wrongly result in some unreliable hits as high-scoring false-positive hits. The reliability of such distant templates again becomes too questionable especially while predicting the structure of a protein sequence whose function is not well understood. Lastly, the HMM-based algorithms predict secondary structure of target sequence through PSIPRED (Pirovano et al. 2007) and consider its structural context to optimally localize the gaps in the target-template sequence profile for searching the reliable templates. Consideration of only the PSIPRED makes the entire template search results of these HMM based algorithms biased towards it. Moreover, PSIPRED does not properly estimates the distant evolutionary relationship between the target and template sequences, and it consequently fails to select the correct set of distantly related templates. A better HMM sequence profile based template search approach is thus needed to search all the correct templates that are biologically related with a target sequence. and to improve the spatial satisfaction of unaligned target segments in accordance with the best set of reliable structural folds of the selected templates. Hence, if we are able to implement the robust set of protein secondary structure prediction algorithms the best way in the optimally scored HMM based computational target-template sequence profile construction and evaluation algorithm, we can attempt to screen the plausible as well as reliable templates for a target sequence.

Constructing a biologically correct alignment is yet another challenge. It has been observed that the protein models constructed through the sequence profile based MSA of templates are more accurate than the ones constructed through the structural alignment of templates (Hildebrand et al. 2009). Through the target-template sequence profile, the TBM algorithms consider the top-ranked templates to construct the target model, although the conformational topology of secondary strctures extracted for all the corresponding target residues as per the considered target-template alignment is not normally retained in the target model. The algorithm that evaluates all the optimal as well as sub-optimal target-template alignments, constructed through sequence as well as structural topology parameters, is required to select or further construct the biologically significant alignment through the efficient implementation of protein secondary structure prediction algorithms. It would thus allow us to properly use the available templates to build accurate target model.

Model building algorithms also suffer from several problems. Several modelling algorithms employ the threading methodologies like TASSER (Zhou & Skolnick 2009) and ROSETTA (Leaver-Fay et al. 2011) to link the structurally conserved folds of different templates for constructing the target model which is not always an energetically stable conformation devoid of atomic clashes. Moreover, some modelling algorithms employ the Cα backbone topology of selected templates to construct the complete target structures by lastly adding the backbone and side-chain atoms. However, this method is very complex as it considers the templates to evaluate different combinations of template fragments, side-chain atoms and backbone atoms. As this method separately add the backbone and side-chain atoms to the Cα backbone model of a target sequence, the probability of erroneous alteration of the conserved Cα backbone, extracted from the selected templates, greatly increases and it may further ruin the overall topology and accuracy of the constructed target model. Additionally, this method fails to construct the target models that are free of atomic clashes (Zhou & Skolnick 2009). Hence, the modelling methodology should efficiently assess the topological similarity and conformational diversity of the top-ranked templates to select thir best set for maximally covering the target sequence and constructing its accurate structure.

As multiple templates are normally employed for modelling a target sequence, model refinement algorithms still need to be improved to refine the target model the best way so that the structural topology extracted from the considered templates is not disturbed and the side-chains, loops and secondary structures are packed the best possible way in the target model. These refinement algorithms are required to optimally sample the conformational space of the target sequence and construct its minimal energy conformation that is structurally closer to its native conformation. Hence, while energetically refining the energetically unfavorable and incorrect segments of the target model, the algorithms should quickly evaluate the conformational space of a target sequence and should also retain the conformational topology intact for all of its secondary structures that are correctly constructed by the modelling algorithm.

Model assessment has also been given huge emphasis in the recent CASP rounds and many assessment measures like TM_Score, GDT, DISOPRED, and SphereGrinder have been developed. Currently, the TM_Score and GDT-TS scores are considered as the most reliable model assessment parameters. However as these scoring schemes consider the Cα distance deviation between the equivalent residues of the two protein structures and the count of such topologically correct residues fitting within a specified distance deviation to calculate the overall structural similarity score, the residue segments that are structurally much more similar or are topologically correct are not clearly highlighted. Selecting an accurate structure also becomes a really difficult task when the two predicted models have an equivalent TM_Score or GDT score against the employed template(s). Hence the algorithm that effectively evaluates the topological cartesian distance between every successive residue Cα atom and also correctly assesses the angular deviation of every successive residue backbone plane is urgently required to solve the logical shotcomings of the TM_Score and GDT based model scoring measures. Furthermore, we also need to revise our assessment algorithm to correctly rank the target models that contain several structural domains.

Altogether, the algorithmic improvement is still needed at every single step of a TBM algorithm to construct an accurate target model.

References

Al-Lazikani, B., Shienerman, FB. & Honig, B. (2001). Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domain of Janus kinase. *Proceedings of the National Academy of Science USA,* 98(26), 14796-14801.

Altschul, SF. & Erickson, BW. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology,* 48(5-6), 603-616.

Altschul, SF. & Koonin, EV. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends in Biochemical Science,* 23(11), 444-447.

Altschul, SF., Gish, W., Miller, W., Myers, EW. & Lipman, DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* 215(3), 403–410.

Altschul, SF., Madden, TL., Schaffer, AA., Zhang, J., Zhang, Z., Miller, W. & Lipman, DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25(17), 3389-3402.

Andreeva, A., Howorth, D., Brenner, SE., Hubbard, TJP., Chothia, C. & Murzin, AG. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32(1), D226-D229.

Anfinsen, CB. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096), 223-230.

Arendall, WB. III, Tempel, W., Richardson, JS., Zhou, W., Wang, S., Davis, IW., Liu, ZJ., Rose, JP., Carson, WM., Luo, M., Richardson, DC. & Wang, BC. (2005). A test of enhancing model accuracy in high-throughput crystallography. *Journal of Structural and Functional Genomics*, 6(1), 1–11.

Baker, TS. & Johnson, JE. (1996). Low resolution meets high: towards a resolution continuum from cells to atoms. *Current Opinion in Structural Biology*, 6, 585-594

Barbato, A., Benkert, P., Schwede, T., Tramontano, A. & Kosinski, A. (2012). Improving your target-template alignment with MODalign. *BIOINFORMATICS*, 28(7), 1038-1039.

Bateman, A., Coin, L., Durbin, R., Finn, RD., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, EL., Studholme, DJ., Yeats, C. & Eddy, SR. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32(1), D138-D141.

Battey, JN., Kopp, J., Bordoli, L., Read, RJ., Clarke, ND. & Schwede, T. (2007). Automated server predictions in CASP7. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 68-82.

Bayley, MJ., Jones, G., Willett, P. & Williamson, MP. (1998). Genfold: A genetic algorithm for folding protein structures using NMR restraints. *Protein Science*, 7(2), 491–499.

Baysal, C. & Meirovitch, H. (2000). *Ab Initio* Prediction of the Solution Structures and Populations of a Cyclic Pentapeptide in DMSO Based on an Implicit Solvation Model. *Biopolymers*, 53(5), 423-433.

Berman, HM., Westbrook, J., Feng, Z., Gilliland, G., Bhat, TN., Weissig, H., Shindyalov, IN. & Bourne, PE. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.

Berrondo, M. & Gray, JJ. (2011). Computed structures of point deletion mutants and their enzymatic activities. *PROTEINS: Structure, Function and Bioinformatics*, 79(10), 2844-2860.

Biegert, A. & Söding, J. (2009). Sequence context-specific profiles for homology searching. *Proceedings of the National Academy of Science USA*, 106(10), 3770-3775.

Blundell, TL., Sibanda, BL., Sternberg, MJ. & Thornton, JM. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature,* 326(6111), 347-352.

Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marín-López, MA., Fernandez-Fuentes, N. & Oliva, B. (2014). ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Research*, 42(D1), D315-319.

Bonneau, R. & Baker, D. (2001). Ab-initio protein structure prediction: Progress and prospects. *Annual Review of Biophysics Biomolecular Structure*, 30, 173–189.

Bonneau, R., Strauss, CEM., Rohl, CA., Chivian, D., Bradley, P., Malmström, L., Robertson, T. & Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*, 322(1), 65-78.

Boratyn GM. I., Schäffer AA., Agarwala R., Altschul SF., Lipman DJ. & Madden TL. (2012). Domain enhanced lookup time accelerated BLAST. *Biology Direct,* 7, 12.

Borjesson, U. & Hunenberger, PH. (2001). Explicit-solvent molecular dynamics simulation at constant pH: methodology and application to small amines. *The Journal of Chemical Physics*, 114(22), 9706-9719.

Borjesson, U. & Hunenberger, PH. (2004). pH-dependent stability of a decalysine a-helix studied by explicit-solvent molecular dynamics simulations at constant pH. *Journal of Physical Chemistry B*, 108(35), 13551-13559.

Bower, MJ., Cohen, FE. & Dunbrack, RL. Jr. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modelling tool. *Journal of Molecular Biology*, 267(5), 1268–1282.

Braun, W. & Go N. (1985). Calculations of protein conformations by proton-proton distance constraints. A new efficient algorithm. *Journal of Molecular Biology*, 186(3), 611-626.

Bray, JE., Todd, AE., Pearl, FM., Thornton, JM. & Orengo, CA. (2000). The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Engineering*, 13(3), 153-165.

Brenner, SE., Chothia, C. & Hubbard, TJ. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Science USA*, 95(11), 6073–6078.

Brooks, BR., Bruccoleri, RE., Olafson, BD., States, DJ., Swaminathan, S. & Karplus, M. (1983). CHARMM –A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), 187-217.

Bruccoleri, RE. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 26(1), 137–168.

Bruccoleri, RE. & Karplus, M. (1990). Conformational sampling using high temperature molecular dynamics. *Biopolymers*, 29(14), 1847–1862.

Burgi, R., Kollman, PA. & van Gunsteren, WF. (2002). Simulating proteins at constant pH: an approach combining molecular dynamics and Monte Carlo simulation. *PROTEINS: Structure, Function and Bioinformatics*, 47(4), 469-480.

Canutescu, AA., Shelenkov, AA. & Dunbrack RL. Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12(9), 2001–2014.

Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. *BIOINFORMATICS*, 31(12), i116-i123.

Carrio, M., Gonzalez-Montalban, N., Vera, A., Villaverde, A. & Ventura, S. (2005). Amylod-like properties of bacterial inclusion bodies. *Journal of Molecular Biology*, 347(5), 1025-1037.

Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, 10(7), 1470–1473.

Chakrabarti, P. & Pal, D. (2001). The interrelationships of side-chain and main-chain conformations in proteins. *Progress in Biophysics & Molecular Biology*, 76, 1-102.

Chapple, CE., Herrmann, C. & Brun, C. (2015). PrOnto database : GO term functional dissimilarity inferred from biological data. *Frontiers in Genetics*, 6, 200.

Chen, J. & Brooks, CL. III. (2007). Can molecular dynamics simulations provide high-resolution refinement of protein structure?. *PROTEINS: Structure, Function and Bioinformatics*, 67(4), s922–930.

Cheng, J. (2007). DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Research*, 35, W354–W356.

Cheng, J., Randall, AZ., Sweredoski, MJ. & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(2), W72–W76.

Chiti, F. & Dobson, CM. (2006). Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry,* 75(1), 333–366.

Chothia, C. & Lesk, AM. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4), 823-826.

Clarke, ND., Ezkurdia, L., Kopp, J., Read, RJ., Schwede, T. & Tress, M. (2007). Domain definition and target classification for CASP7. *PROTEINS: Structure, Function, and Bioinformatics*, 69(8), 10–18.

Clore, GM., Brunger, AT., Karplus, M. & Gronenborn, AM. (1986). Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination, A model study of crambin. *Journal of Molecular Biology*, 191(3), 523-551.

Colovos, C. & Yeates, TO. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science*, 2(9), 1511–1519.

Cornell, WD., Cieplak, P., Bayly, CI., Gould, IR., Merz, KM. Jr., Ferguson, DM., Spellmeyer, DC., Fox, T., Caldwell, JW. & Kollman, PA. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of American Chemical Soceity*, 117, 5179-5197.

Cozzetto, D., Giorgetti, A., Raimondo, D. & Tramontano, A. (2008). The evaluation of protein structure prediction results. *Molecular Biotechnology*, 39(1), 1–8.

Cozzetto, D. & Tramontano, A. (2005). Relationship between multiple sequence alignments and quality of protein comparative models. *PROTEINS: Structure, Function and Bioinformatics*, 58(1), 151–157.

Cramer, CJ. & Truhlar, DG. (1999). Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chemical Reviews,* 99(8), 2161–2200.

Cunha, KC., Rusu, VH., Viana, IF., Marques, ET., Dhalia, R. & Lins, RD. (2015). Assessing protein conformational sampling and structural stability via de novo design and molecular dynamics simulations. *Biopolymers*, 103(6), 351-361.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, MD., Bhat, D., Chivian, D., Kim, DE., Sheffler, WH., Malmström, L., Wollacott, AM., Wang, C., Andre, I. & Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 118-128.

Davis, IW., Leaver-Fay, A., Chen, VB., Block, JN., Kapral, GJ., Wang, X., Murray, LW., Arendall, WB. III., Snoeyink, J., Richardson, JS. & Richardson, DC. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research,* 35, W375–W383.

Dayhoff, MO., Schwartz, R. & Orcutt, BC. (1978). A model of Evolutionary Change in Proteins. *In* Dayhoff, MO. (ed) Atlas of Protein Sequence and Structure, 5th ed., vol. 3. National Biomedical Research Foundation, Maryland.

Dunbrack, RL. Jr., Gerloff, DL., Bower, M., Chen, X., Lichtarge, O. & Cohen, FE. (1997). Meeting review: the Second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, California, December 13-16, 1996. *Folding & Design*, 2(2), R27-42.

Dunbrack, RL. Jr. (2006). Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology*, 16(3), 374–384.

Eddy, SR. (1998). Profile hidden Markov models. *BIOINFORMATICS*, 14(9), 755-763.

Eddy, SR. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10), e1002195.

Edgar, RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high through-put. *Nucleic Acids Research*, 32(5), 1792-1797.

Edgar, RC. & Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. *BIOINFORMATICS*, 20(8), 1309–1318.

Edman, P. (1949). A method for the determination of amino acid sequence in peptides. *Archieves of Biochemsitry*, 22(3), 475.

Eisenmenger, F., Argos, P. & Abagyan, R. (1993). A method to configure protein side-chains from the main-chain trace in homology modelling. *Journal of Molecular Biology*, 231(3), 849–860.

Engh, RA. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica A*, 47, 392–400.

English, CA. & García, AE. (2014). Folding and unfolding thermodynamics of the TC10b Trp-cage miniprotein. *Physical Chemistry Chemical Physics*, 16(7), 2748-2757.

Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J. & Brooks CL. III. (2000). Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *PROTEINS: Structure Function and Genetics,* 41(1), 86–97.

Ferrara, P., Apostolakis, J. & Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *PROTEINS: Structure, Function and Bioinformatics*, 46(1), 24-33.

Fidelis, K., Stern, PS., Bacon, D. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Engineering*, 7(8), 953–960.

Filippis, D., Sander, VC. & Vriend, G. (1994). Predicting local structural changes that result from point mutations. *Protein Engineering*, 7(10), 1203–1208.

Fine, RM., Wang, H., Shenkin, PS., Yarmush, DL. & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *PROTEINS: Structure, Function and Bioinformatics*, 1(4), 342–362.

Finn, RD., Clements, J. & Eddy, SR. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(2), W29-W37.

Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert Review of Proteomics*, 1(1), 97-110.

Fiser, A. & Sali, A. (2003). Modeller: generation and refinement of homology based protein structure models. *Methods in Enzymology*, 374, 461–491.

Fiser, A. & Sali, A. (2003). ModLoop: automated modelling of loops in protein structures. *BIOINFORMATICS*, 19(18), 2500-2501.

Fiser, A., FIEG, M., Brooks, CL. III. & Sali, A. (2002). Evolution and physics in comparative protein structure modelling. *Accounts of Chemical Research*, 35(6), 413-421.

Gadad, BS., Britton, GB. & Rao, KS. (2011). Targeting oligomers in neurodegenerative disorders: lessons from α-synuclein, tau, and amyloid-β peptide. *Journal of Alzheimer's disease*, 24(2), 223–232.

García-Moreno, EB. & Fitch, CA. (2004). Structural interpretation of pH and salt-dependent processes in proteins with computational methods. *Methods in Enzymology*, 380, 20–51.

George, DCP., Chakraborty, C., Haneef, SAS., NagaSundaram, N., Chen, L. & Zhu, H. (2014). Evolution- and Structure-Based Computational Strategy Reveals the Impact of Deleterious Missense Mutations on MODY 2 (Maturity-Onset Diabetes of the Young, Type 2). *Theranostics*, 4(4), 366-385.

Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16(2), 172–177.

Goh, GB., Hulbert, BS., Zhou, H. & Brooks, CL. III. (2014). Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. *PROTEINS: Structure, Function and Bioinformatics*, 82(7), 1319-1331.

Gonnet, GH., Cohen, MA. & Benner, SA. (1994). Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. *Biochemical and Biophysical Research Communications*, 199(2), 489-496.

Gonzalez, MW. & Pearson, WR. (2010). Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Research*, 38(7), 2177–2189.

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4), 903–919.

Goyal, S., Jamal, S., Shanker, A. & Grover, A. (2015). Structural investigations of T854A mutation in EGFR and identification of novel inhibitors using structure activity relationships. *BMC Genomics*, 16(5), S8.

Greer, J. (1980). Model for haptoglobin heavy chain based upon structural homology. *Proceedings of the National Academy of Science USA*, 77(6), 3393-3397.

Gront, D., Kmiecik, S. & Kolinski, A. (2007). Backbone Building from Quadrilaterals: A Fast and Accurate Algorithm for Protein Backbone Reconstruction from Alpha Carbon Coordinates. *Journal of Computational Chemistry*, 28(9), 1593-1597.

Gront, D., Kulp, DW., Vernon, RM., Strauss, CEM. & Baker, D. (2011). Generalized fragment picking in rosetta: design, protocols and applications. *PloS ONE*, 6(8), e23294.

Guex, N. & Peitsch, MC. (1997). SWISS-MODEL and the Swiss-Pdb-Viewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15), 2714-2723.

Guo, JT., Ellrott, K. & Xu, Y. (2008). A historical perspective of template-based protein structure prediction. *Methods in Molecular Biology,* 413, 3-42.

Hanley, JM., Haugen, TH. & Heath, EC. (1983). Biosynthesis and processing of rat haptoglobin. *Journal of Biological Chemistry*, 258(12), 7858-7869.

Hao, F., Xavier, P. & Alan, EM. (2012). Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: Application in the refinement of de novo models. *PROTEINS: Structure, Function and Bioinformatics*, 80(7), 1744-1754.

Hardin, C., Taras, VP. & Zaida, LS. (2002). Ab-initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2), 176–181.

Havel, TF. & Snow, ME. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. *Journal of Molecular Biology*, 217(1), 1-7.

Hellberg, S., Sjöström, M., Skagerberg, B. & Wold, S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *Journal of Medicinal Chemistry*, 30(7), 1126-1135.

Henikoff, S. & Henikoff, JG. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science USA*, 89(22), 10915–10919.

Henikoff, S. & Henikoff, JG. (1993). Performance evaluation of amino acid substitution matrices. *PROTEINS: Structure, Function and Bioinformatics*, 17(1), 49–61.

Henikoff, S. & Henikoff, JG. (1997). Embedding strategies for effective use of information from multiple sequence alignments. *Protein Science*, 6(3), 698-705.

Heringa, J. (1999). Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computers & Chemistry*, 23(3-4), 341–364.

Herzyk, P. & Hubbard, RE. (1993). A reduced representation of proteins for use in restraint satisfaction calculations. *PROTEINS: Structure, Function and Bioinformatics*, 17(3), 310-324.

Higo, J., Collura, V. & Garnier, J. (1992). Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. *Biopolymers*, 32(1), 33–43.

Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *PROTEINS: Structure, Function and Bioinformatics*, 77(9), 128–132.

Holak, TA., Nilges, M., Prestegard, JH., Gronenborn, AM. & Clore, GM. (1988). Three-dimensional structure of acyl carrier protein in solution determined by nuclear magnetic resonance and the combined use of dynamical simulated annealing and distance geometry. *European Journal of Biochemistry / FEBS*, 175(1), 9-15.

Holley, LH. & Karplus, M. (1991). Neural networks for protein structure prediction. *Methods in Enzymology*, 202, 204-224.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *Journal of Molecular Biology*, 218(1), 183-194.

Hsieh, M. & Luo, R. (2004). Physical Scoring Function Based on AMBER Force Field and Poisson–Boltzmann Implicit Solvent for Protein Structure Prediction. *PROTEINS: Structure, Function, and Bioinformatics*, 56(3), 475–486.

Huang, JT., Wang, T., Huang, SR. & Li, X. (2015). Reduced alphabet for protein folding prediction. *PROTEINS: Structure, Function, and Bioinformatics*, 83(4), 631-639.

Hwang, JK. & Liao, WF. (1995). Side-chain prediction by neural networks and simulated annealing optimization. *Protein Engineering*, 8(4), 363–370.

Iwata, Y., Kasuya, A. & Miyamoto, S. (2002). An efficient method for reconstructing protein backbones from alpha-carbon coordinates. *Journal of Molecular Graphics & Modelling*, 21(2), 119-128.

Iverson, GM., Reddel, S., Victoria, EJ., Cockerill, KA., Wang, YX., Marti-Renom, MA., Sali, A., Marquis, DM., Krilis, SA. & Linnik, MD. (2002). Use of single point mutations in domain I of beta 2-glycoprotein I to determine fine antigenic specificity of antiphospholipid autoantibodies. *Journal of Immunology*, 169(12), 7097-7103.

Jacobsen, JL. (2008). Unbiased sampling of globular lattice proteins in three dimensions. *Physical Review Letters*, 100(11), 118102.

Jacobson, MP., Kaminski, GA., Friesner, RA. & Rapp, CS. (2002). Force field validation using protein side-chain prediction. *The Journal of Physical Chemistry B,* 106(44), 11673–11680.

Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Science,* 9(8), 1487-1496.

Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Science,* 7(6), 1431-1440.

Jauch, R., Yeo, HC., Kolatkar, PR. & Neil, DC. (2007). Assessment of CASP7 structure predictions for template free targets. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 57–67.

Jeanmougin, F., Thompson, JD., Gouy, M., Higgins, DG. & Gibson, TJ. (1998). Multiple sequence alignment with Clustal X. *Trends in Biochemical Science,* 23(10), 403-405.

Jefferys, BR., Kelley, LA. & Sternberg, MJE. (2010). Protein Folding Requires Crowd Control in a Simulated Cell. *Journal of Molecular Biology*, 397(5), 1329-1338.

John, B. & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research*, 31(14), 3982-3992.

Jonathan, JW., Liam, JM., Kevin, B., Bernard, FB. & David, TJ. (2004). The DISOPRED server for the prediction of protein disorder. *BIOINFORMATICS*, 20(13), 2138-2139.

Jones, DT. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287(4), 797–815.

Jones, DT. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292, 195-202.

Jones, DT., Bryson, K., Coleman, A., McGuffin, LJ., Sadowski, MI., Sodhi, JS. & Ward, JJ. (2005). Prediction of Novel and Analogous Folds Using Fragment Assembly and Fold Recognition. *PROTEINS: Structure, Function and Bioinformatics*, 61(7), 143–151.

Jones, DT., Buchan, DW., Cozzetto, D. & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *BIOINFORMATICS*, 28(2), 184-190.

Jones, TA. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *The EMBO Journal*, 5(4), 819-822.

Joo, H., Qu, X., Swanson, R., McCallum, CM. & Tsai, J. (2010). Fine grained sampling of residue characteristics using molecular dynamics simulation. *Computational Biology and Chemistry*, 34(3), 172-183.

Joo, K., Lee, J., Kim, I., Lee, SJ. & Lee, J. (2008). Multiple sequence alignment by conformational space annealing. *Biophysical journal*, 95(10), 4813-4819.

Joo, K., Lee, J., Seo, JH., Lee, K., Kim, BG. & Lee, J. (2009). All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. *PROTEINS: Structure, Function and Bioinformatics*, 75(4), 1010-1023.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-2637.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A34, 827–828.

Karplus, K., Sjölander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L. & Sander, C. (1997). Predicting protein structure using hidden Markov models. *PROTEINS: Structure, Function and Bioinformatics*, 29(S1), 134--139.

Kevin, D., Patrick, W., Glenn, LB., Viktors, B., Keith, U., Jonathan, A., Michael, R., Erik, S., Bill, B., David, RG., Trisha, ND., Dennis, S., Lars, M. & Richard, B. (2011). The Proteome Folding Project: Proteome-scale prediction of structure and function. *Genome Research*, 21(11), 1981–1994.

Kim, SH. (1998). Shining a light on structural genomics. *Nature Structural Biology*, 5, 643-645.

Kinch, LN., Shi, S., Cheng, H., Cong, Q., Pei, J., Mariani, V., Schwede, T. & Grishin, NV. (2011). CASP9 target classification. *PROTEINS: Structure, Function and Bioinformatics*, 79(10), 21-36.

Knudsen, B. & Miyamoto, MM. (2003). Sequence alignments and pair hidden Markov models using evolutionary history. *Journal of Molecular Biology*, 333(2), 453-460.

Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *Journal of Molecular Biology*, 239(2), 249–275.

Koehl, P. & Delarue, M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nature Structural Biology*, 2(2), 163–170.

Koh, IY., Eyrich, VA., Marti-Renom, MA., Przybylski, D., Madhusudhan, MS., Eswar, N., Graña, O., Pazos, F., Valencia, A., Sali, A. & Rost, B. (2003). EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Research*, 31(13), 3311–3315.

Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51(2), 349-371.

Kopp, J. & Schwede, T. (2004). The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Research*, 32(1), D230-234.

Kopp, J. & Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, 5(4), 405-416.

Kosciolek, T. & Jones, DT. (2014). De Novo Structure Prediction of Globular Proteins Aided by Sequence Variation-Derived Contacts. *PLoS ONE*, 9(3), e92197.

Kosinski, J., Cymerman, IA., Feder, M., Kurowski, MA., Sasin, JM. & Bujnicki, JM. (2003). A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *PROTEINS: Structure, Function and Bioinformatics*, 53(6), 369-379.

Krupa, P., Mozolewska, MA., Joo, K., Lee, J., Czaplewski, C. & Liwo, A. (2015). Prediction of Protein Structure by Template-Based Modeling Combined with the UNRES Force Field. *Journal of Chemical Information and Modeling*, 55(6), 1271-1281.

Kryshtafovych, A. & Fidelis, K. (2008). Protein structure prediction and model quality assessment. *Drug Discovery Today*, 14(7-8), 386-393.

Kryshtafovych, A., Fidelis, K. & Moult, J. (2007). Progress from CASP6 to CASP7. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 194-207.

Kryshtafovych, A., Fidelis, K. & Moult, J. (2009). CASP8 results in context of previous experiments. *PROTEINS: Structure, Function and Bioinformatics*, 77(9), 217-228.

Kryshtafovych, A., Venclovas, C., Fidelis, K. & Moult, J. (2005). Progress over the first decade of CASP experiments. *PROTEINS: Structure, Function and Bioinformatics*, 61(7), 225– 236.

Kuziemko, A., Honig, B. & Petrey, D. (2011). Using structure to explore the sequence alignment space of remote homologs. *PLoS Computational Biology*, 7(10), e1002175.

Larsson, P., Skwark, MJ., Wallner, B. & Elofssson, A. (2009). Assessment of global and local model quality in CASP8 using Pcons and ProQ2. *Proteins: Structure, Function and Bioinformatics*, 77(S9), 167-172.

Laskowski, RA., MacArthur, MW., Moss, DB. & Thornton, JM. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography,* 26, 283-291.

Lathrop, RH., Rogers, RG. Jr., Smith, TF. & White, JV. (1998). A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology*, 60(6), 1039-1071.

Lawrence, AK. & Michael, JES. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols*, 4(3), 363 - 371.

Lazaridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, 10(2), 139–145.

Leaver-Fay A., Tyka, M., Lewis, SM., Lange, OF., Thompson, J., Jacak, R., Kaufman, K., Renfrew, PD., Smith, CA., Sheffler, W., Davis, IW., Cooper, S., Treuille, A., Mandell, DJ., Richter, F., Ban, YE., Fleishman, SJ., Corn, JE., Kim, DE., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, JJ., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, JJ., Kuhlman, B., Baker, D. & Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology,* 487, 545-574.

Lee, J., Lee, D., Park, H., Coutsias, EA. & Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins: Structure, Function, and Bioinformatics*, 78(16), 3428-3436.

Lesk, AM. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, 136(3), 225-270.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology*, 226(2), 507-533.

Lewis, DF., Bailey, PT. & Low, LK. (2002). Molecular modelling of the mouse cytochrome P450 CYP2F2 based on the CYP102 crystal structure template and selective CYP2F2 substrate interactions. *Drug Metabolism and Drug Interactions*, 19(2), 97-113.

Li, J., Adhikari, B. & Cheng, J. (2015). An Improved Integration of Template-Based and Template-Free Protein Structure Modeling Methods and its Assessment in CASP11. *Proteins and Peptide Letters*, 22(7), 586-593.

Li, Y. & Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *PROTEINS: Structure, Function and Bioinformatics*, 76(3), 665-676.

Liang, S. & Grishin, NV. (2002). Side-chain modeling with an optimized scoring function. *Protein Science*, 11(2), 322–331.

Lisa, NK., Qi, Y., Hubbard, TJ. & Grishin, NV. (2003). CASP5 Target Classification. *PROTEINS: Structure, Function, and Genetics*, 53(6), 340 –351.

Liu, Y. & Beveridge, DL. (2002). Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *PROTEINS: Structure, Function and Bioinformatics*, 46(1), 128-146.

Looger, LL. & Hellinga, HW. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *Journal of Molecular Biology*, 307(1), 429–445.

Lopes, A., Alexandrov, A., Bathelt, C., Archontis, G. & Simonson, T. (2007). Computational Sidechain Placement and Protein Mutagenesis With Implicit Solvent Models. *PROTEINS: Structure, Function, and Bioinformatics*, 67(4), 853–867.

Lu, H. & Skolnick, J. (2003). Application of statistical potentials to protein structure refinement from low resolution ab-initio models. *Biopolymers,* 70(4), 575-584.

Lucas-Lenard, J. (1971). Protein biosynthesis. *Annual review of biochemistry*, 40, 409-448.

Luheshi, M., Crowther, DC. & Dobson, CM. (2008). Protein misfolding and disease: from the test tube to the organism. *Current Opinion in Chemical Biology*, 12(1), 25–31.

Lushington, GH. (2015). Comparative modeling of proteins. *Methods in Molecular Biology*, 1215, 309-330.

Luthy, R., Bowie, JU. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364), 83–85.

Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y. & Yang, Y. (2014). Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, 35(28), 2040-2046.

MacKerell, AD., Bashford, D., Bellott, M., Dunbrack, RL., Evanseck, JD., Field, MJ., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, FT., Mattos, C., Michnick, S., Ngo, T., Nguyen, DT., Prodhom, B., Reiher, WE., Roux, B., Schlenkrich, M., Smith, JC., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18), 3586–3616.

Madhusudhan, MS., Marti-Renom, MA., Sanchez, R. & Sali A. (2006). Variable gap penalty for protein sequence-structure alignment. *Protein Engineering, Design & Selection*, 19(3), 129-133.

Madhusudhan, MS., Webb, BM., Marti-Renom, MA., Eswar, N. & Sali, A. (2009). Alignment of multiple protein structures based on sequence and structure features. *Protein Engineering, Design & Selection*, 22(9), 569–574.

Manavalan, B., Lee, J. & Lee, J. (2014). Random Forest-Based Protein Model Quality Assessment (RFMQA) Using Structural Features and Potential Energy Terms. *PLOS ONE*, 9(9), e106542.

Mao, B., Guan, R. & Montelione, GT. (2011). Improved technologies now routinely provide protein NMR structures useful for molecular replacement. *Structure*, 19(6), 757-766.

Mao, B., Tejero, R., Baker, D. & Montelione, GT. (2014). Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *Journal of the American Chemical Society*, 136(5), 1893-1906.

Marenich, AV., Cramer, CJ. & Truhlar, DG. (2009). Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B,* 113(18), 6378–6396.

Margelevicius, M. & Venclovas, C. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics*, 11, 89.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *PROTEINS: Structure, Function and Bioinformatics*, 79(10), 37-58.

Marko, AC., Stafford, K. & Wymore, T. (2007). Stochastic pairwise alignments and scoring methods for comparative protein structure modeling. *Journal of Chemical Information and Modeling*, 47(3), 1263-1270.

Martin, AC., MacArthur, MW. & Thornton, JM. (1997). Assessment of comparative modeling in CASP2. *PROTEINS: Structure, Function and Bioinformatics*, 1, 14-28.

Martinez, JC. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Structural Biology*, 6(11), 1010–1016.

Marti-Renom, MA., Madhusudhan, MS. & Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Science*, 13(4), 1071-1087.

Marti-Renom, MA., Stuart, AC., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29, 291-325.

Masuda, T. (2015). Molecular dynamics simulation for the reversed power stroke motion of a myosin subfragment-1. *Biosystems*, 132-133, 1-5.

Matsui, E., Abe, J., Yokoyama, H. & Matsui, I. (2004). Aromatic residues located close to the active center are essential for the catalytic reaction of flap endonuclease-1 from hyperthermophilic archaeon Pyrococcus horikoshii. *The Journal of Biological Chemistry*, 279(16), 16687-16696.

McCallus, DE., Ugen, KE., Sato, AI., Williams, WV. & Weiner, DB. (1992). Construction of a recombinant bacterial human CD4 expression system producing a bioactive CD4 molecule. *Viral immunology*, 5(2), 163-172.

Meier, A. & Söding, J. (2014). Context similarity scoring improves protein sequence alignments in the midnight zone. *BIOINFORMATICS*, 30(22), 1-8.

Meier, A. & Söding, J. (2015). Context similarity scoring improves protein sequence alignments in the midnight zone. *BIOINFORMATICS*, 31(5), 674-681.

Melo, F. & Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *Journal of Molecular Biology*, 277(5), 1141–1152.

Melo, F., Sanchez, R. & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Science*, 11(2), 430–448.

Metzker, ML. (2010). Sequencing technologies - the next generation. *Nature reviews Genetics*, 11(1), 31–46.

Mihaesco, E., Guglielmi, P., Brouet, JC. & Mihaesco, C. (1983). Biochemical and biosynthetic studies of a crystallizable human gamma 1 heavy-chain disease. *Scandinavian Journal of Immunology*, 18(2), 145-152.

Minde, DP., Anvarian, Z., Rüdiger, SG. & Maurice, MM. (2011). Messing up disorder: how do missense mutations in the tumor suppressor protein APC lead to cancer?. *Molecular Cancer,* 10, 101.

Miranker, A. & Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *PROTEINS: Structure, Function, and Bioinformatics*, 11(1), 29-34.

Montelione, GT. & Anderson, S. (1999). Structural genomics: keystone for a Human Proteome Project. *Nature Structural Biology*, 6(1), 11-12.

Montelione, GT., Nilges, M., Bax, A., Guntert, P., Herrmann, T., Richardson, JS., Schwieters, CD., Vranken, WF., Vuister, GW., Wishart, DS., Berman, HM., Kleywegt, GJ. & Markley, JL. (2013). Recommendations of the wwPDB NMR Validation Task Force. *Structure*, 21(9), 1563-1570.

Moo-Penn, WF., Schmidt, RM., Jue, DL., Bechtel, KC., Wright, JM., Horne, MK. III., Haycraft, GL., Roth, EF. Jr. & Nagel, RL. (1977). Hemoglobin S Travis: a sickling hemoglobin with two amino acid substitutions [beta6(A3)glutamic acid leads to valine and beta142 (h20) alanine leads to valine). *European Journal of Biochemistry*, 77(3), 561-566.

Moshe, B., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, JL. & Levy, Y. (2009). Assessment of CASP8 structure predictions for template free targets. *PROTEINS: Structure, Function and Bioinformatics*, 77(9), 50-65.

Moult, J. & James, MN. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *PROTEINS: Structure, Function and Bioinformatics*, 1(2), 146–163.

Moult, J. & Melamud, E. (2000). From fold to function. *Current Opinion in Structural Biology*, 10(3), 384-389.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3), 285-289.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. & Tramontano, A. (2007). Critical assessment of methods of protein structure prediction—Round VII. *PROTEINS: Structure, Function, and Bioinformatics,* 69(8), 3–9.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) – round X. *PROTEINS: Structure, Function, and Bioinformatics,* 80(2), 1-6.

Moult, J., Fidelis, K., Rost, B., Hubbard, T. & Tramontano, A. (2005). Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round 6. *PROTEINS: Structure, Function, and Bioinformatics*, 61(7), 3–7.

Murphy, ME., Moult, J., Bleackley, RC., Gershenfeld, H., Weissman, IL. & James, MN. (1988). Comparative molecular model building of two serine proteinases from cytotoxic T lymphocytes. *PROTEINS: Structure, Function, and Bioinformatics*, 4(3), 190-204.

Murzin, A. & Hubbard, TJP. (2001). Prediction Targets of CASP4. *PROTEINS: Structure, Function, and Genetics*, 45(8), 8–12.

Murzin, AG., Brenner, SE., Hubbard, T. & Chothia, C. (1995). SCOP: A Structural Classification of Proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), 536-540.

Navaratnam, N., Fujino, T., Bayliss, J., Jarmuz, A., How, A., Richardson, N., Somasekaram, A., Bhattacharya, S., Carter, C. & Scott J. (1998). Escherichia coli cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. *Journal of Molecular Biology*, 275(4), 695–714.

Nguyen, KD., Pan, Y. & Nong, G. (2011). Parallel progressive multiple sequence alignment on recon-figurable meshes. *BMC Genomics*, 12(5), S4.

Notredame, C., Higgins, DG. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1), 205–217.

Onuchic, JN., Nymeyer, H., Garcia, AE., Chahine, J. & Socci, ND. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advances in Protein Chemistry*, 53, 87-130.

Orengo, CA., Michie, AD., Jones, S., Jones, DT., Swindells, MB. & Thornton, JM. (1997). CATH--a hierarchic classification of protein domain structures. *Structure*, 5(8), 1093-1108.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, DG. & Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2), 385-395.

Park, H. & Seok, C. (2012). Refinement of unreliable local regions in template-based protein models. *Proteins: Structure, Function, and Bioinformatics*, 80(8), 1974-1986.

Park, H., Ko, J., Joo, K., Lee, J., Seok, C. & Lee, J. (2011). Refinement of protein termini in template-based modeling using conformational space annealing. *Proteins: Structure, Function, and Bioinformatics*, 79(9), 2725-2734.

Pearl, FMG., Bennett, CF., Bray, JE., Harrison, AP., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. & Orengo, CA. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, 31(1), 452-455.

Pearlman, DA., Case, DA., Caldwell, JW., Ross, WS., Cheatham, TE., Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995). AMBER, a package of computer-programs for applying molecular mechanics,<normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3), 1-41.

Pearson, WR. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183, 63–98.

Pearson, WR. (2013). Selecting the Right Similarity-Scoring Matrix. *Current Protocols in Bioinformatics*, 43, 3.5.1–3.5.9.

Pei, J., Kim, B. & Grishin, NV. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Research*, 36(7), 2295-2300.

Peng, J. & Xu, J. (2010). Low-homology protein threading. *BIOINFORMATICS*, 26(12), i294-i300.

Pentony, MM., Winters, P., Penfold-Brown, D., Drew, K., Narechania, A., DeSalle, R., Bonneau, R. & Purugganan, MD. (2012). The Plant Proteome Folding Project:

Structure and Positive Selection in Plant Protein Families. *Genome Biology and Evolution*, 4(3), 360-371.

Petrey, D., Chen, TS., Deng, L., Garzon, JI., Hwang, H., Lasso, G., Lee, H., Silkov, A. & Honig, B. (2015). Template-based prediction of protein function. *Current Opinion in Structural Biology*, 32C, 33-38.

Pirovano, W., Feenstra, K.A. & Heringa, J. (2007). PRALINE™: a strategy for improved multiple alignment of transmembrane proteins. *BIOINFORMATICS*, 24(4), 492-497.

Pokarowski, P., Kolinski, A. & Skolnick, J. (2003). A Minimal Physically Realistic Protein-Like Lattice Model: Designing an Energy Landscape that Ensures All-Or-None Folding to a Unique Native State. *Biophysical journal*, 84(3), 1518-1526.

Ponder, JW. & Richards, FM. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193(4), 775–791.

Qian, B., Ortiz, AR. & Baker, D. (2004). Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proceedings of the National Academy of Science USA*, 101(43), 15346–15351.

Qian, B., Raman, S. & Das, R. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, 450(7167), 259-264.

Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27(5), 1664-1670.

Razzokov, J., Naderi, S. & Schoot, P. (2014). Prediction of the structure of a silk-like protein in oligomeric states using explicit and implicit solvent models. *Soft Matter*, 10(29), 5362–5374.

Reddy, BV., Li, WW., Shindyalov, IN. & Bourne, PE. (2001). Conserved key amino acid positions (CKAAP) derived from the analysis of common substructures in proteins. *PROTEINS: Structure, Function and Bioinformatics*, 42(2), 148-163.

Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173-175.

Reva, B., Finkelstein, A. & Topiol, S. (2002). Threading with chemostructural restrictions method for predicting fold and functionally significant residues: application to dipeptidylpeptidase IV (DPP-IV). *PROTEINS: Structure, Function and Bioinformatics*, 47(2), 180–193.

Ring, CS., Sun, E., McKerrow, JH., Lee, GK., Rosenthel, PJ., Kuntz, ID. & Cohen, FE. (1993). Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proceedings of the National Academy of Science USA*, 90(8), 3583–3587.

Robson, B. & Platt, E. (1987). Modelling of alpha-lactalbumin from the known structure of hen egg white lysozyme using molecular dynamics. *Journal of Computer-Aided Molecular Design*, 1(1), 17-22.

Rodriguez, A., Mokoema, P., Corcho, F., Bisetty, K. & Perez, JJ. (2011). Computational Study of the Free Energy Landscape of the Miniprotein CLN025 in Explicit and Implicit Solvent. *The Journal of Physical Chemistry B*, 115(6), 1440–1449.

Rossmann, MG. & Argos, P. (1981). Protein folding. *Annual review of biochemistry*, 50, 497-532.

Rossman, MG. (2000). Fitting atomic models into electron-microscopy maps. *Acta Crystallographica D*, 56, 1341-1349.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design & Selection*, 12(2), 85-94.

Rotkiewicz, P. & Skolnick, J. (2008). Fast method for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry*, 29(9), 1460-1465.

Roux, B. & Simonson, T. (1999). Implicit solvent models. *Biophysical Chemistry*, 78(1–2), 1–20.

Ruan, KH., Milfeld, K., Kulmacz, RJ. & Wu, KK. (1994). Comparison of the construction of a 3-D model for human thromboxane synthase using P450cam and BM-3 as templates: implications for the substrate binding pocket. *Protein Engineering*, 7(11), 1345-51.

Russell, RB. & Barton, GJ. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *PROTEINS: Structure, Function and Bioinformatics*, 14(2), 309-323.

Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, 9(2), 232–241.

Sadowski, MI. & Jones, DT. (2007). Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *PROTEINS: Structure, Function and Bioinformatics*, 69(3), 476-485.

Sadreyev, RI., Shi, SY., Baker, D. & Grishin, NV. (2009). Structure similarity measure with penalty for close non-equivalent residues. *BIOINFORMATICS*, 25(10), 1259-1263.

Sali, A. & Blundell, TL. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3), 779-815.

Samudrala, R. & Moult, J. (1998). A graph-theoretic algorithm for comparative modeling of protein structure. *Journal of Molecular Biology*, 279(1), 287–302.

Sanchez, R. & Sali, A. (1997). Evaluation of comparative protein structure modelling by MODELLER-3. *PROTEINS: Structure, Function and Bioinformatics*, 1, 50–58.

Sanger, F., Nicklen, S. & Tuppy, H. (1951). The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 1. THE IDENTIFICATION OF LOWER PEPTIDES FROM PARTIAL HYDROLYSATES. *Biochemical Journal*, 49(4), 463-481.

Sarig, O., Nahum, S., Rapaport, D., Ishida-Yamamoto, A., Fuchs-Telem, D., Qiaoli, L., Cohen-Katsenelson, K., Spiegel, R., Nousbeck, J., Israeli, S., Borochowitz, Z., Padalon-Brauch, G., Uitto, J., Horowitz, M., Shalev, S. & Sprecher, E. (2012). Short Stature, Onychodysplasia, Facial Dysmorphism, and Hypotrichosis Syndrome Is Caused by a POC1A Mutation. *The American Journal of Human Genetics*, 91(2), 337-342.

Schäffer, AA., Aravind, L., Madden, TL., Shavirin, S., Spouge, JL., Wolf, YI., Koonin, EV. & Altschul, SF. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14), 2994-3005.

Schiffer, CA., Caldwell, JW., Kollman, PA. & Stroud, RM. (1990). Prediction of homologous protein structures based on conformational searches and energetics. *PROTEINS: Structure, Function and Bioinformatics*, 8(1), 30-43.

Schoonman, MJ., Knegtel, RM. & Grootenhuis, PD. (1998). Practical evaluation of comparative modelling and threading methods. *Computers and Chemistry*, 22(5), 369-375.

Schwede, T., Kopp, J., Guex, N. & Peitsch, MC. (2003). SWISS-MODEL: An automated protein homology-modelling server. *Nucleic Acids Research*, 31(13), 3381-3385.

Scott, WRP., Hunenberger, PH., Mark, AE., Billeter, SR., Fennen, J., Torda, AE., Huber, T., Kruger, P. & van Gunsteren, WF. (1999). The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A*, 103(19), 3596-3607.

Serra, E., Ars, E., Ravella, A., Sánchez, A., Puig, S., Rosenbaum, T., Estivill, X. & Lázaro, C. (2001). Somatic NF1 mutational spectrum in benign neurofibromas: MRNA splice defects are common among point mutations. *Human genetics*, 108(5), 416–29.

Shatnawi, M. & Zaki, N. (2015). Inter-domain linker prediction using amino acid compositional index. *Computational Biology and Chemistry*, 55, 23-30.

Shindyalov, IN. & Bourne, PE. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9), 739-747.

Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *BIOINFORMATICS*, 16(9), 776-785.

Simoncini, D., Berenger, F., Shrestha, R. & Zhang, KYJ. (2012). A probabilistic fragment-based protein structure prediction algorithm. *PloS One*, 7(7), e38799.

Sinha, N. & Nossinov, R. (2001). Point mutations and sequence variability in proteins: Redistributions of preexisting populations. *Proceedings of the National Academy of Science USA*, 98(6), 3139-3144.

Sippl, MJ. (1993). Recognition of errors in three-dimensional structures of proteins. *PROTEINS: Structure, Function and Bioinformatics*, 17(4), 355-362.

Smith, RE., Lovell, SC., Burke, DF., Montalvao, RW. & Blundell, TL. (2007). Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. *BIOINFORMATICS*, 23(9), 1099–1105.

Smith, TF., Lo Conte, L., Bienkowska, J., Gaitatzes, C., Rogers, RG. Jr. & Lathrop, R. (1997). Current limitations to protein threading approaches. *Journal of Computational Biology*, 4(3), 217–225.

Snyder, DA., Bhattacharya, A., Huang, YJ. & Montelione, GT. (2005). Assessing Precision and Accuracy of Protein Structures Derived From NMR Data. *PROTEINS: Structure, Function, and Bioinformatics*, 59(4), 655– 661.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *BIOINFORMATICS*, 21(7), 951-960.

Söding, J., Biegert, A. & Lupas, AN. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33, W244-W248.

Srinivasan, R. & Rose, GD. (2002). Ab initio prediction of protein structure using LINUS. *PROTEINS: Structure, Function and Bioinformatics*, 47(4), 489-95.

Srinivasan, S., March, CJ. & Sudarsanam, S. (1993). An automated method for modelling proteins on known templates using distance geometry. *Protein Science*, 2(2), 277–289.

Standley, DM., Gunn, JR., Friesner, RA. & McDermott, AE. (1998). Tertiary structure prediction of mixed alpha/beta proteins via energy minimization. *PROTEINS: Structure, Function and Bioinformatics*, 33(2), 240–252.

Steinbach, PJ. (2004). Exploring Peptide Energy Landscapes: A Test of Force Fields and Implicit Solvent Models. *PROTEINS: Structure, Function, and Bioinformatics*, 57(4), 665–677.

Still, WC., Tempczyk, A., Hawley, RC. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of American Chemical Society*, 112(16), 6127–6129.

Styczynski, MP., Jensen, KL., Rigoutsos, I. & Stephanopoulos, G. (2008). BLOSUM62 miscalculations improve search performance. *Nature Biotechnology*, 26(3), 274–275.

Subbiah, S. & Harrison, SC. (1989). A simulated annealing approach to the search problem of protein crystallography. *Acta crystallographica A*, 45(5), 337-42.

Subramani, A. & Floudas, CA. (2012). Structure Prediction of Loops with Fixed and Flexible Stems. *The Journal of Physical Chemistry B*, 116(23), 6670-6682.

Sugita, Y. & Okamoto, Y. (2009). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2), 141–151.

Sun S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science*, 2(5), 762-785.

Sutcliffe, MJ., Haneef, I., Carney, D. & Blundell, TL. (1987). Knowledge based modeling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering*, 1(5), 377-384.

Teichmann, SA., Chothia, C., Church, GM. & Park, J. (2000). Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *BIOINFORMATICS*, 16(2), 117–124.

Tianyun, L., Jeremy, AH. & Samudrala, R. (2009). A novel method for predicting and using distance constraints of high accuracy for refining protein structure prediction. *PROTEINS: Structure, Function and Bioinformatics*, 77(1), 230-234.

Taylor, TJ., Tai, C., Huang, YJ., Block, J., Bai, H., Kryshtafovych, A., Montelione, GT. & Lee, B. (2014). Definition and classification of evaluation units for CASP10. *PROTEINS: Structure, Function and Bioinformatics*, 82(2), 14-25.

Tomasi, J., Mennucci, B., Cammi, R. & Cammi, R. (2005). Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8), 2999–3093.

Tomkins, GM. & Martin, DW. Jr. (1970). Hormones and gene expression. *Annual review of genetics*, 4, 91-106.

Tong, J., Pei, J., Otwinowski, Z. & Grishin, NV. (2015). Refinement by shifting secondary structure elements improves sequence alignments. *PROTEINS: Structure, Function and Bioinformatics*, 83(3), 411-427.

Topf, M., Baker, ML., Marti-Renom, MA., Chiu, W. & Sali, A. (2006). Refinement of Protein Structures by Iterative Comparative Modeling and CryoEM Density Fitting. *Journal of Molecular Modeling*, 357(5), 1655–1668.

Topham, CM., Mcleod, A., Eisenmenger, F., Overington, JP., Johnson, MS. & Blundell, TL. (1993). Fragment ranking in Modeling of protein structure, Conformationally constrained environmental amino acid substitution tables. *Journal of Molecular Biology*, 229(1), 194-220.

Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *PROTEINS: Structure Function and Genetics*, 53(6), 352-368.

Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Joo, HS., Lee, J., Baker, D., Chivian, D., Kim, D. & Ezkurdia, I. (2007). Assessment of predictions submitted for the CASP7 domain prediction category. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 137–151.

Tress, M., Ezkurdia, I., Grana, O., Lopez, G. & Valencia, A. (2005). An assessment of predictions submitted for the CASP6 comparative modeling category. *PROTEINS: Structure, Function and Bioinformatics*, 61(7), 27-45.

Tress, M., Ezkurdia, L. & Richardson, JS. (2009). Target domain definition and classification in CASP8. *PROTEINS: Structure, Function and Bioinformatics*, 77(S9), 10–17.

Trojanowski, S., Rutkowska, A. & Kolinski, A. (2010). TRACER. A new approach to comparative modeling that combines threading with free-space conformational sampling. *Acta Biochimica Polonica*, 57(1), 125-133.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure and Dynamics,* 8(6), 1267–1289.

Tyka, M., Jung, K. & Baker, D. (2012). Efficient Sampling of Protein Conformational Space using Fast Loop Building and Batch Minimization Highly Parallel Computers. *Journal of Computational Chemistry*, 33(31), 2483-2491.

Unger, R., Harel, D., Wherland, S. & Sussman, JL. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *PROTEINS: Structure, Function and Bioinformatics*, 5(4), 355-373.

van Vlijmen, HW. & Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. *Journal of Molecular Biology*, 267(4), 975–1001.

Wallner, B. (2014). ProQM-resample: improved model quality assessment for membrane proteins by limited conformational sampling. *BIOINFORMATICS*, 30(15), 2221–2223.

Wallner, B., Larsson, P. & Eloffson, A. (2007). Pcons.net: protein structure prediction meta server. *Nucleic Acids Research*, 35(2), W369–W374.

Wang, C., Ren-Xiang, Y., Xiao-Feng, W., Jing-Na, S. & Ziding, Z. (2011). Comparison of linear gap penalties and profile-based variable gap penalties in profile–profile alignments. *Computational Biology and Chemistry*, 35(5), 308-318.

Wang, Z., Tegge, AN., & Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Structure, Function and Bioinformatics*, 75(3), 638-647.

Weiser, J., Shenkin, PS. & Still, WC. (1999). Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *Journal of Computational Chemistry*, 20, 217–230.

Wilson, DB. (1971). Initiation of hemoglobin biosynthesis. *Series heamatologica*, 4(3), 70-83.

Wishart, DS., Arndt, D., Berjanskii, M., Tang, P., Zhou, J. & Lin, G. (2008). CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research*, 36(2), W496-W502.

Wohlers, I., Domingues, FS. & Klau, GW. (2010). Towards optimal alignment of protein structure distance matrices. *BIOINFORMATICS*, 26(18), 2273-2280.

Wolf, E., Vassilev, A., Makino, Y., Sali, A., Nakatani, Y. & Burley, SK. (1998). Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase. *Cell*, 94(4), 439-449.

Wu, S. & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10), 3375-3382.

Wu, S. & Zhang, Y. (2008). MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *PROTEINS: Structure, Function and Bioinformatics*, 72(2), 547–556.

Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Bioinformatics*, 5, 17.

Xiang, Z. & Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology*, 311(2), 421–430.

Xiang, Z., Soto, CS. & Honig, B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Science USA*, 99(11), 7432–7437.

Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score=0.5?. *BIOINFORMATICS*, 26(7), 889-895.

Xu, Z., Lazim, R., Sun, T., Mei, Y. & Zhang, D. (2012). Solvent effect on the folding dynamics and structure of E6-associated protein characterized from ab initio protein folding simulations. *The Journal of Chemical Physics*, 136(13), 135102.

Yang, AS. & Honig, B. (1999). Sequence to Structure alignment in comparative modeling using PrISM. *PROTEINS: Structure, Function and Bioinformatics*, 3, 66-72.

Yang, Y. & Zhou, Y. (2008). Specific interactions for *ab initio* folding of protein terminal regions with secondary structures. *PROTEINS: Structure, Function and Bioinformatics*, 72(2), 793-803.

Yelena, AA., Vorobjev, YN., Vila, JA. & Scheraga, HA. (2009). Identifying native-like protein structures with scoring functions based on all-atom ECEPP force fields, implicit solvent models and structure relaxation. *PROTEINS: Structure, Function, and Bioinformatics*, 77(1), 38–51.

Yoon, BJ. (2014). Sequence alignment by passing messages. *BMC Genomics*, 15(1), S14.

Zemla, A. (2003). LGA - a Method for Finding 3D Similarities in Protein Structures. *Nucleic Acids Research*, 31(13), 3370-3374.

Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y. & Xu, D. (2010). MUFOLD: A new solution for protein 3D structure prediction. *PROTEINS: Structure, Function and Bioinformatics*, 78(5), 1137-1152.

Zhang, Y. & Skolnick, J. (2004). Scoring Function for Automated Assessment of Protein Structure Template Quality. *PROTEINS: Structure, Function and Bioinformatics*, 57(4), 702–710.

Zhang, Y. & Skolnick, J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, 25(6), 865–871.

Zhang, Y. & Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Science USA*, 102(4), 1029–1034.

Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on TM-score. *Nucleic Acids Research*, 33(7), 2302-2309.

Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 108-117.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3), 342–348.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9, 40.

Zhang, Y., Arakaki, A. & Skolnick, J. (2005). TASSER: An Automated Method for the Prediction of Protein Tertiary Structures in CASP6. *PROTEINS: Structure, Function and Bioinformatics*, 61(S7), 91–98.

Zhang, Y., Kihara, D. & Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *PROTEINS: Structure, Function and Bioinformatics*, 48(2), 192-201.

Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal*, 85(2), 1145-1164.

Zheng, W., Eickholt, J. & Cheng, J. (2010). MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *BIOINFORMATICS*, 26(7), 882-888.

Zhou, H., Shashi BP., Lee, SY., Borreguero, J., Chen, H., Wroblewska, L. & Skolnick, J. (2007). Analysis of TASSER-based CASP7 protein structure prediction results. *PROTEINS: Structure, Function and Bioinformatics*, 69(8), 90-97.

Zhou, H. & Skolnick, J. (2009). Protein structure prediction by pro-sp3-TASSER. *Biophysical Journal*, 96(6), 2119-2127.

Zhou, R. (2003). Free energy landscape of protein folding in water: explicit vs. implicit solvent. *PROTEINS: Structure, Function and Bioinformatics*, 53(2), 148–161.

List of Publications

List of Publications

Publications:

1. **Runthala, A,** Chowdhury S, (2015) Refined template selection and combination algorithm significantly improves Template Based Modelling accuracy, Proteins and Peptide Letters (*In Review*).
2. **Runthala, A**, Chowdhury S, (2014) Unsolved problems of ambient computationally intelligent TBM algorithms, *Springer Hybrid Soft Computing Approaches: Research and Applications* (*Accepted for publication*).
3. **Runthala, A,** Chowdhury S, (2013) Protein Structure Prediction: Are we there yet?, *Knowledge-Based Systems in Biomedicine and Computational Life Science. Studies in Computational Intelligence. 450*: 79-115
4. **Runthala, A,** (2012) Protein structure prediction: challenging targets for CASP10, *Journal of Biomolecular Structure and Dynamics*. 30(5):607-615

Book Chapters:

- ✓ **Runthala, A,** (2012) Protein Modeling & CASP: Progress and Future Prospects?. *In* Berhardt LV (eds) *Advances in Medicine & Biology*, NOVA Publishers, USA, Vol. 46, pp: 259-270.
- ✓ **Runthala, A,** (2010) Hunting Drugs for Potent Antigens in the Silicon Valley. *In* Shukla A, Tiwari R (eds) *Intelligent Medical Technologies and Biomedical Engineering: Tools and Applications*, IGI GLOBAL Publishers, USA, pp: 203-225.

Papers presented:

- Presented a paper on “Iterative Optimal TM_Score and Z_Score Guided Sampling Significantly Improves Model Topology”, in International MultiConference of Engineers and Computer Scientists, 2014, Vol. I, pp: 123-128.

  Venue: Royal Garden Hotel, Kowloon, Hong Kong.

- Presented a paper on " Modified template scoring scheme improves template based modeling accuracy”, in International Conference on Biomolecular Forms and Functions, 2012.

  Venue: Indian Institute of Science, Bangalore.

- Presented a paper on "Prioritized ranked minimal set of closest structural folds significantly improve model predictions”, in National conference on Contemporary Trends in Biological and Pharmaceutical Research, 2011.

  Venue: Department of Biological science, BITS, Pilani.

Biography of Prof. Shibasish Chowdhury

Prof. Shibasish Chowdhury has done his Master’s in Physical Chemistry from Calcutta University. He then shifted to Molecular Biophysics Unit at Indian Institute of Science, Bangalore on “Computer modelling studies on G-rich unusual DNA structure”. Subsequently, he entered the protein folding field and worked as postdoctoral research fellow in the Department of Chemistry and Biochemistry, University of Delaware, USA for three years. He worked as Lecturer in Birla Institute of Technology and Science, Pilani for 2 years from 2004 to 2005, before being promoted to Assistant Professor till 2012 after which he got promoted to Associate Professor in 2013. His research interests include Protein folding, Biomolecular Modeling, Evolution and Bioinformatics.

Biography of Ashish Runthala

Mr. Ashish Runthala has done his Master’s in Biological Sciences from Birla Institute of Technology and Science, Pilani. He has done his Master’s in Biotechnology also from the same Institute. He has worked as Teaching Assistant and as Assistant lecturer in Biological Science department of this Organization before being promoted here as a Lecturer for pursuing his PhD. His doctoral thesis is entitled “Refinement and Improvement of Template Based Protein Modelling Algorithms” which he is completing under the guidance of Prof. Shibasish Chowdhury. His research interest includes Protein structure prediction algorithms, Biocomputing, Structural Bioinformatics and Functional Proteomics.