# Protein structure prediction and structure-based protein function annotation

By
Ambrish Roy

Dissertation submitted to the graduate degree program in Bioinformatics & the Graduate Faculty of the University of Kansas

______________________________
Chairperson Dr. John Karanicolas

______________________________
Co-chair/Adviser Dr. Yang Zhang

______________________________
Dr. Ilya Vakser

______________________________
Dr. Eric. J. Deeds

______________________________
Dr. Mark. L. Richter

Date Defended: 29th November, 2011

The Dissertation Committee for Ambrish Roy
certifies that this is the approved version of the following thesis:

# Protein structure prediction and structure-based protein function annotation

__________________________________
Chairperson Dr. John Karanicolas

__________________________________
Co-chair/Adviser Dr. Yang Zhang

Date Approved:

# Abstract

Nature tends to modify rather than invent function of protein molecules, and the log of the modifications is encrypted in the gene sequence. Analysis of these modification events in evolutionarily related genes is important for assigning function to hypothetical genes and their products surging in databases, and to improve our understanding of the bioverse. However, random mutations occurring during evolution chisel the sequence to an extent that both decrypting these codes and identifying evolutionary relatives from sequence alone becomes difficult. Thankfully, even after many changes at the sequence level, the protein three-dimensional structures are often conserved and hence protein structural similarity usually provide more clues on evolution of functionally related proteins.

In this dissertation, I study the design of three bioinformatics modules that form a new hierarchical approach for structure prediction and function annotation of proteins based on sequence-to-structure-to-function paradigm. First, we design an online platform for structure prediction of protein molecules using multiple threading alignments and iterative structural assembly simulations (I-TASSER). I review the components of this module and have added features that provide function annotation to the protein sequences and help to combine experimental and biological data for improving the structure modeling accuracy. The online service of the system has been supporting more than 20,000 biologists from over 100 countries.

Next, we design a new comparative approach (COFACTOR) to identify the location of ligand binding sites on these modeled protein structures and spot the functional residue constellations using an innovative global-to-local structural alignment procedure and functional sites in known protein structures. Based on both large-scale benchmarking and blind tests (CASP), the method demonstrates significant advantages over the state-of-the-art methods of the field in recognizing ligand-binding residues for both metal and non-metal ligands. The major advantage of the method is the optimal combination of the local and global protein structural alignments, which helps to recognize functionally conserved structural motifs among proteins that have taken different evolutionary paths.

We further extend the COFACTOR global-to-local approach to annotate the gene-ontology and enzyme classifications of protein molecules. Here, we added two new components to COFACTOR. First, we developed a new global structural match algorithm that allows performing better structural search. Second, a sensitive technique was proposed for constructing local 3D-signature motifs of template proteins that lack known functional sites, which allows us to perform query-template local structural similarity comparisons with all template proteins. A scoring scheme that combines the confidence score of structure prediction with global-local similarity score is used for assigning a confidence score to each of the predicted function. Large scale benchmarking shows that the predicted functions have remarkably improved precision and recall rates and also higher prediction coverage than the state-of-art sequence based methods. To explore the applicability of the method for real-world cases, we applied the method to a subset of ORFs from *Chlamydia trachomatis* and the functional annotations provided new testable hypothesis for improving the understanding of this phylogenetically distinct bacterium.

Together, these developed softwares during my dissertation form an integrated pipeline to facilitate structural and functional annotation of genome sequences, and can thereby improve our understanding of the bioverse and well being of living organisms.

# Dedications

I dedicate the work presented in this thesis to all the biologists who contribute and improve our understanding of life.

I dedicate the thesis itself to my parents, my elder brother and my wife for their unflagging love and support, without which this work would have never been completed.

# Acknowledgements

First and foremost, I would like to thank my parents (Mala and Raman Roy) who are my first teachers. Their teachings, morals and aspirations underlying this thesis cannot be overstated.

I am grateful to all my teachers for their contribution in spawning my interest in research. Especially, I would like to thank Dr. Yang Zhang for being my mentor and giving me new ideas, for the constant motivation, for the financial aid and picking out flaws in my thinking. All these has helped me to strive for excellence and enabled me to grow as a scientist.

I thank all the members of Yang Zhanglab (especially Dr. Sitao Wu, Dr. Dong Xu, Dr. Andrea Bazzoli and Dr. Srayanta Mukherjee) for their valuable discussion, scientific suggestions, and co-authoring papers with me.

The love and confidence of my elder brother (Amritanshu Roy) and my wife (Ms. Neha Sarode) cannot be expressed in terms of words, but it was the strongest force to push me through most difficult phases of academic and personal life.

Final thanks goes out to the faculty and staff working at Center for Bioinformatics at the University of Kansas & Center for Computational medicine and Bioinformatics at the University of Michigan for providing the infrastructure and managing my nomadic life.

# Table of Contents

# List of Figures

# List of Tables

## List of Abbreviations used

| Abbreviation | Full form |
|---|---|
| BLAST | *B*asic *L*ocal *A*lignment *S*earch *T*ool |
| BP | *B*iological *P*rocess |
| BS | *B*inding *S*ite |
| CASP | *C*ritical *A*ssessment of techniques for protein *S*tructure *P*rediction |
| CC | *C*ellular *C*omponent |
| CM | *C*omparative *M*odeling |
| C-score | *C*onfidence score |
| DAG | *D*irected *A*cyclic *G*raph |
| EC | *E*nzyme Commission |
| FM | *F*ree *M*odeling |
| FG-MD | *F*ragment *G*uided *M*olecular *D*ynamics simulation |
| GO | *G*ene *O*ntology |
| HMM | *H*idden *M*arkov *M*odel |
| I-TASSER | *I*terative *T*hreading *ASSE*mbly *R*efinement |
| LOMETS | *L*ocal *M*eta *T*hreading *S*erver |
| MF | *M*olecular *F*unction |
| MUSTER | *MU*lti *S*ources *T*hread*ER* |
| NCBI | *N*ational *C*enter for *B*iotechnology *I*nformation |
| NR | *N*on *R*edundant |
| ORF | *O*pen *R*eading *F*rame |
| PDB | *P*rotein *D*ata *B*ank |
| PPA | *P*rofile *P*rofile *A*lignment |
| PSI-BLAST | *P*osition-*S*pecific *I*terative *B*asic *L*ocal *A*lignment *S*earch *T*ool |
| RefSeq | *R*eference *S*equence |
| RMSD | *R*oot *M*ean *S*quare *D*eviation |
| TrEMBL | Translated EMBL |
| SC | *S*ide-chain *C*enter |
| SG | *S*tructural *G*enomics |
| TASSER | *T*hreading *ASSE*mbly *R*efinement |
| TBM | *T*emplate *B*ased *M*odeling |
| TM | *T*emplate *M*odeling |
| UniProtKB | *U*niversal *P*rotein *R*esource *K*nowledge*B*ase |
| 3D | 3 *D*imension |

# Chapter 1

## Background and overview

With the advancement in sequencing technology, genome sequences have been rapidly pooling up in the sequence databases. However, to fully understand these “blueprints of life”, it is imperative to decode the genetic information into biologically meaningful knowledge. Proteins, despite being established as the “work horses” of the cell, are lagging far behind in terms of structural and functional information in this sequencing frenzy era. This is in part due to the time-consuming, expensive and technically difficult nature of experimental methods for characterizing these proteins. For instance, as of October 2011, more than 17 million protein sequences from over 18,000 species had already been deposited in UniProtKB/TrEMBL[1] database and the rate of deposition was over 200 times faster (Figure 1.1) than the rate at which protein structure were being deposited in the Protein Data Bank[2] (PDB), not to mention that the functional characterization of these proteins by experimental methods lags further behind. Thus, one of the most challenging tasks in modern molecular and cell biology is to characterize these protein sequences for better understanding of physiological processes and systems[3].

This dissertation work is a software engineering component geared towards the development of automated methods for annotating protein sequences, which can help biologists to construct hypothesis and design experiments for drawing reliable conclusions.

## 1.1. Proteins: from the view of a computational structural biologist

In this thesis, we will study and develop algorithms that will tackle two main problems: first, to improve the automated pipeline for protein structure prediction, and second, to develop a structure based approach for automated function prediction. Before we move ahead and approach these problems, we will briefly introduce some basics about protein bioinformatics, which is a preliminary requirement for understanding the later chapters.

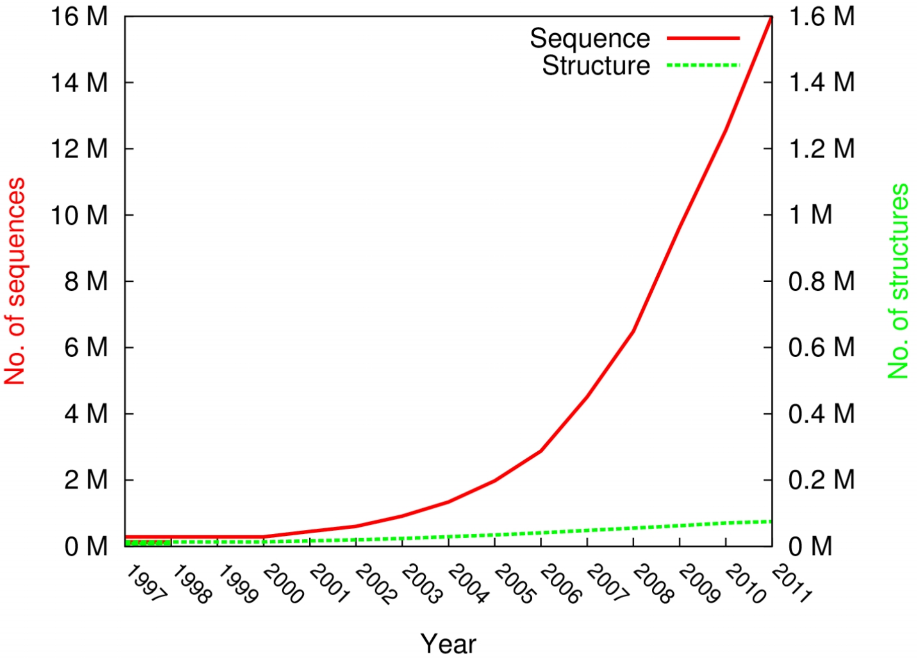


**Figure 1.1** Protein sequence to structure: Catch me if you can!

### 1.1.1. Levels of protein structure

Proteins are linear polymers of amino acids and can be described at four different levels of abstraction (Figure 1.2), namely:

(a) Primary structure: Is the sequence of amino acids in a protein, where each amino acid is represented by its one-letter code.

(b) Secondary structure: The polar groups of protein backbone form hydrogen bonds and generate repeating structural fragments called secondary structure. α-helix and β-sheet are two most commonly observed secondary structure elements in a protein.

(c) Tertiary structure: Also commonly referred as the 3D-structure of a protein, consists of multiple secondary structure elements, where every atom in the protein is represented by its 3D-coordinates.

(d) Quaternary structure: The quaternary structure of a protein represents the arrangement of multiple folded protein molecules in a multi-subunit complex. These spatial conformations are stabilized by non-covalent interactions.

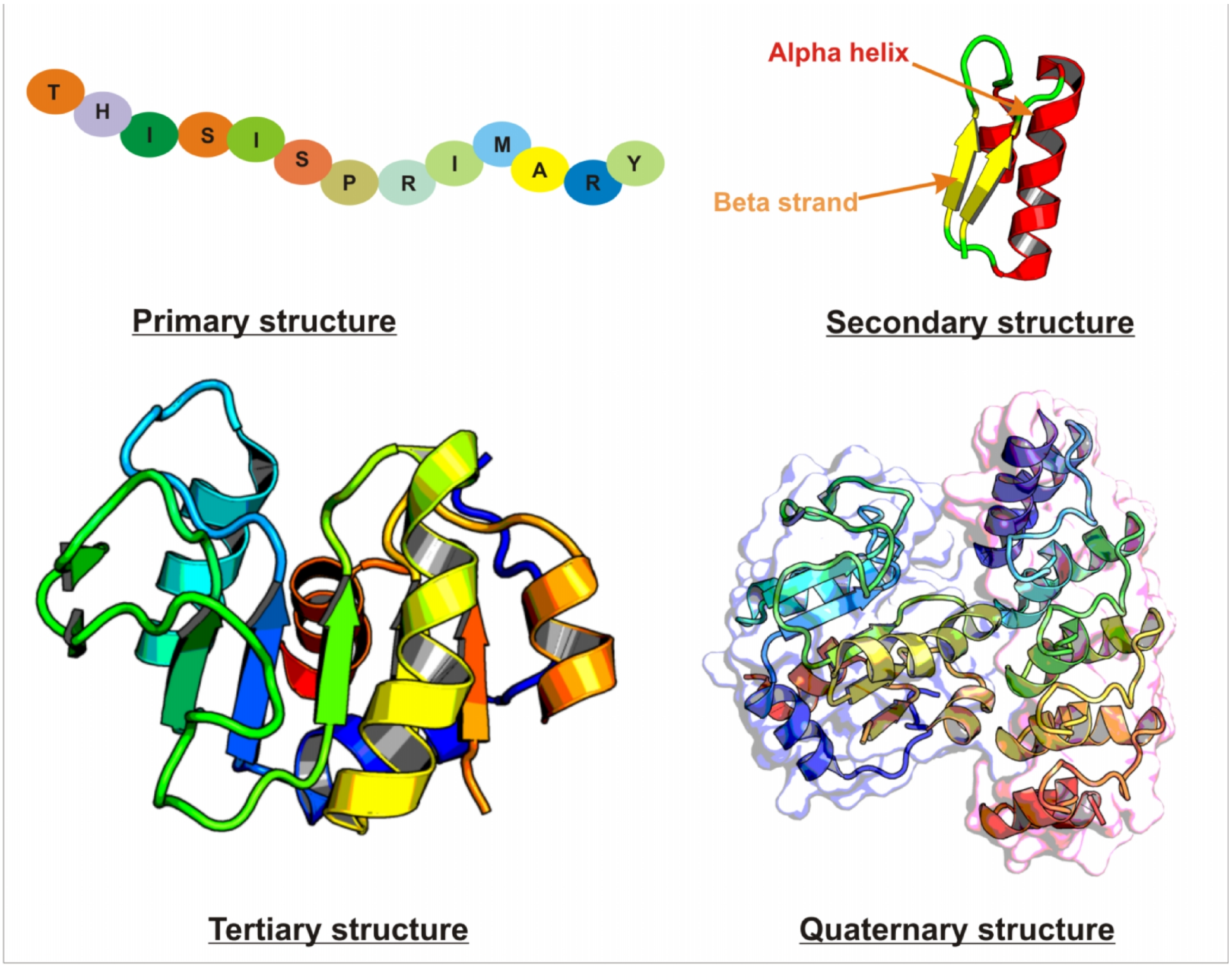


**Figure 1.2** Abstraction levels of protein structure.

### 1.1.2. Protein databases

The field of Biology over the years has become a data-rich science and computer has naturally become the storage medium. A large number of databases have therefore been developed to distribute the gathered information freely to the biologists, in a computer readable format. In Table 1.1, we enlist and then briefly summarize the available information in some of the protein databases that have been used in this work.

**Table 1.1 Publicly available protein databases used in this work.**

| Database | Information | Web link |
|---|---|---|
| PDB | Tertiary & Quaternary structure | http://www.pdb.org |
| NCBI nr | Non-redundant protein sequence database | ftp://ftp.ncbi.nih.gov/blast/db |
| Enzyme | Nomenclature of enzymes | http://enzyme.expasy.org |
| PDBSProtEC | Mapping of PDB chain to Enzyme nomenclature | http://www.bioinf.org.uk/pdbsprotec |
| Gene Ontology (GO) | Controlled vocabulary (GO terms) to describe the attributes of gene and gene product | http://www.geneontology.org |
| UniProtKB-GOA | Assignment of GO terms to UniProt records and GO assignment for PDB entries | http://www.ebi.ac.uk/GOA |
| NCBI RefSeq | Complete genomic DNA, gene transcripts and protein sequences | http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome |

**(a)** **PDB**[2]: The Protein Data Bank (PDB) stores experimentally solved tertiary and quaternary structure of biological macromolecules including nucleotides and proteins. As of October, 2011, PDB contains over 70,000 protein structures, and more than 3,000 protein-nucleic acid complexes.

**(b)** **NCBI nr**: NCBI provides a non-redundant set of protein sequences collected from other databases, namely: GenPept, SwissProt, PIR, PDF, PDB, and NCBI RefSeq.

**(c)** **Enzyme**[4]: Is a repository of enzyme nomenclature based on Enzyme Classification (EC) numbers. It also contains information about catalytic activity, required cofactor (if any) and mapping to SwissProt sequences entries.

**(d)** **PDBSProtEC**[5]: A database containing assignment of EC numbers to PDB chains based on PDB to SwissProt and SwissProt to Enzyme database mapping.

**(e)** **Gene Ontology**[6]: The Gene Ontology (GO) database provides annotations of genes, gene products and sequences using a structured, controlled vocabularies and classifications

**(f)** **UniProtKB-GOA**[7]: The UniProtKB-GOA consortium provides GO annotation mapping to proteins in UniProt Knowledgebase (UniProtKB). Annotation for PDB chains is also provided based on UniProtKB accession and match between InterPro and PDB mapping

**(g)** **NCBI RefSeq**[8]: Annotated and curated collection of nucleotide and protein sequences.

### 1.1.3. Analyzing similarity between two proteins

Before we proceed to describe the different levels of abstractions at which we can compare proteins, first we need to understand why we need to compare proteins. The reason is that nature is conservative and has evolved over time because of small incremental modifications, rather than large changes. Thus, by comparing two proteins, we can detect similarities, which allow us to infer the structure and function of isolated proteins using already characterized protein. That said, we will now outline sequence and structure based methods and evaluation metrics for comparing proteins.

#### 1.1.3.1. Sequence based alignment

Similarity between two sequences (nucleotides or proteins) can be obtained by either using dynamic programming or based on heuristic search. Even though dynamic programming methods find optimal alignment, they are relatively slow; heuristic approaches are much faster and therefore used for identifying similar sequences in large database, but they are less precise. Many sequence alignment algorithms have been developed, however, here we will only present an overview of the most basic and widely used algorithms that are relevant to this work.

### (a) Global sequence alignment (Needleman-Wunsch algorithm)

In 1970, Needleman and Wunsch[9] described the first method for aligning two sequences based on dynamic programming. The algorithm has three steps: (a) an initialization step; (b) a matrix filling step; and (c) traceback.

In the initialization step, for any two sequences A and B with lengths *x* and *y*, a matrix M of size (*x*+1, *y*+1) is initialized and M(0, 0) is set to 0. M(0, *j*) is initialized to score (numerically equal to *j* x *d*, where *d* is empirically determined gap penalty) resulting from aligning B[1] to a gap of length *j* and analogous to that M(*i*, 0) is initialized values (numerically equal to *i* x *d*) resulting from aligning A[1] to a gap of length *i*. For each position in the matrix M with *i, j* >0, the M(*i*, *j*) scores signify how favorably residue/nucleotide A[*i*] are replaced by B[*j*] or alternatively a deletion or insertion occurs.

In the next step, the matrix M is filled from top right to bottom left, where the M(*i, j*) score is filled based on the rule given in Equation 1.1.

$$M\,(i,j) = max \begin{cases} M\,(i-1,j-1) + score\,(a,b) & \text{substitution at } (i,\, j \\ M\,(i-1,j) - d & \text{deletion at position } j \\ M\,(i.j-1) - d & \text{deletion at position } i \end{cases} \tag{1.1}$$

where, *score (a, b)* is taken from substitution matrix like BLOSUM[10] and PAM[11], and *d* is the gap penalty.

During the matrix filling procedure, the chosen condition from Equation 1.1 is recorded for each cell. These choices are traced back in the last step, starting from the bottom right cell of matrix M to the top right cell and the alignment is printed. Thus, Needleman-Wunsch (NW) algorithm always generates global alignment of sequences and penalizes end gaps. In most part of this work, we will use this algorithm as implemented in NW-align program (http://zhanglab.ccmb.med.umich.edu/NW-align/) and sequence identity is defined as:

$$\text{Seq. ID} = \frac{\text{No. of identical residues in the alignment}}{\text{Length of query protein}} \tag{1.2}$$

### (b) Local sequence alignment (Smith-Waterman algorithm)

The Smith-Waterman (SW) algorithm[12] also uses dynamic programming algorithm but produces local alignment. Here, we will briefly summarize the changes in this algorithm compared to the Needleman and Wunsch algorithm described above.

In the initialization step, M($i$, 0) and M(0, $j$) are initialized to 0.

During the matrix filling procedure, a new option is introduced (Equation 1.3) which helps to set the negative scoring cells to 0 and aids in identifying positively scoring cells which render the local alignment.

$$M\left(i,j\right) = max\begin{cases} M\left(i-1,j-1\right) + score\left(a,b\right) & \text{substitution at } (i,\, j) \\ M\left(i-1,j\right) - d & \text{deletion at position } j \\ M\left(i,j-1\right) - d & \text{deletion at position } i \\ 0 & \text{stop the local alignment} \end{cases} \tag{1.3}$$

During the traceback, instead of starting at the bottom right matrix element, as in NW-algorithm, in SW-algorithm, the traceback starts from the cell with maximum value and then traced back until a cell with value 0 is encountered.

### (c) Heuristic database search (BLAST algorithm)

As mentioned earlier, heuristic approach generate sub-optimal alignment but is very fast and hence useful for searching database. FASTA[13] and BLAST[14] are two commonly used programs for heuristic search in large databases and generating pairwise alignment against top ranking hits;

while ClustalW[15] is a multiple sequence alignment tool. Here we will only discuss BLAST, because its derivative (PSI-BLAST[16] discussed in next section) has been used extensively in this work. The BLAST (*B*asic *L*ocal *A*lignment *S*earch *T*ool) algorithm consists of three main steps:

First, the query sequence is split into *k* letter words, where the default value of *k* for proteins is 3. All possible combinations of *k* letter words ($20^3$) are then scored against these query words using substitution matrices (BLOSUM[10] or PAM[11]). In the second step, words scoring higher than a threshold cutoff *T* are marked as hits, and the database sequence are searched using these selected hits, with the requirement that there should be an exact match. Matches in the databases are extended bi-directionally using un-gapped alignment, until the alignment score drops below a threshold cut-off *S*. These extended hits are called as HSP (*H*igh scoring *S*egment *P*airs). In the third step, high scoring HSPs with score ≥ *Sg*, are extended further using gapped alignment and highest scoring hits are reported with the final alignment.

### (d) Heuristic database search using sequence-profile alignment (PSI-BLAST)

The choice of substitution matrix (BLOSUM or PAM) used for scoring the alignment in heuristic algorithms (for e.g. in BLAST) determines how well the method can distinguish true hits from random ones during the database search. Therefore, the sensitivity of these heuristic algorithms can be improved by choosing a matrix, which can best describe the characteristics of the query protein family.

This concept forms the basis of PSI-BLAST (*P*osition *S*pecific *I*terative BLAST) algorithm[16], which scores the database sequences using the query profile. A profile[17] or PSSM (*P*osition *S*pecific *S*coring *M*atrix) is a matrix of dimensions 20 × *L*, where *L* is the length of query sequence and the 20 columns contain substitution scores for the 20 standard amino acids. PSI-BLAST uses an iterative search procedure, where the first round of PSI-BLAST search is essentially the same as running BLAST. Statistically significant hits (those with *e-value* lower than user-specified threshold) are selected and multiple sequence alignment is constructed from these sequences, which is then converted to PSSM. A PSSM contains information about both the query sequence and the substitution matrix, and is used for the next round of database search.

### 1.1.3.2. Structure based alignment

Compared to sequence comparisons, protein structure comparisons can provide useful insight into their functionality because (a) Protein residues located far apart in the primary sequence can come very close in 3D space; (b) structure is known to be more conserved than the sequence[18], as seemingly dissimilar sequences can fold into the same 3D conformation/structure.

Additionally in this work, we will use predicted models and structure alignment therefore is essential for evaluating the quality of predicted structures.

Given a set of two proteins *P1* and *P2* and a scoring scheme, the aim of any structure alignment method is to find a set of equal sized substructure in these proteins with highest score. In this section we will briefly review, two scoring schemes RMSD and TM-score, which have been used in the later chapters.

**(a)** **RMSD**: Root mean square deviation (RMSD) is one of the most widely used structure similarity measure. For calculating RMSD, one needs to first identify an optimal transformation (rotation and translation) that can superposes the complete structure or substructures of *P1* onto *P2* or vice-versa. The formulation for finding this optimal superposition is based on the Kabsch algorithm[19]. RMSD on the superposed structures is then defined as:

$$RMSD = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\delta_i} \tag{1.4}$$

where, $L$ is the length of common structure or the aligned length and $\delta$ is distance between the $i^{\text{th}}$ elements in the alignment.

**(b)** **<u>TM-score</u>**: Template modeling score (TM-score)[20], is another widely used metric for evaluating protein structure similarity. As in RMSD, for evaluating TM-score between two protein structures, they must be superposed onto each other. In TM-score, the transformation matrix for superposing the entire protein structure is acquired after Kabsch[19] superposition of fragments of various lengths gleaned from N-to-C terminus of proteins. After each superposition, TM-score is evaluated and recorded. TM-score is defined as:

$$\text{TM-sccre} = max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \tag{1.5}$$

where, $L$ is the length of query protein and $L_{ali}$ is the aligned length, $d_i$ is the distance between $i^{\text{th}}$ pair of aligned residues and $d_0$ is the normalization scale for defining aligned residue pairs given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$. Finally, during the iteration, the superposition with highest TM-score is reported as the final alignment. TM-score was originally designed for evaluating the quality of predicted protein structure (*model*) as it is more sensitive towards the global topology. Also, it down-weights large distances between aligned residue pairs compared to the smaller ones, while in RMSD all the residue are weighted equally.

In this work, the structural similarity between predicted model and its experimental form is evaluated using TM-score program (http://zhanglab.ccmb.med.umich.edu/TM-score); while the global structural similarity between two different proteins (*query* and *template*) is evaluated using TM-align (http://zhanglab.ccmb.med.umich.edu/TM-align). In TM-align, a wide variety of initial structure-based alignments are used as initial seed, which are then refined further based on heuristic iterations of the Needleman-Wunsch dynamic programming [9] with the distance scoring matrix defined by TM-score superposition. The best alignment is scored using TM-score (Equation 1.5), where $L$ is the length of query protein.

## 1.2. Protein structure: experimental determination & prediction

In Figure 1.1, we showed that the growth of experimental protein structures elucidation (dotted lines) has fallen behind in the race with the exponentially increasing number of sequenced proteins. In this section, we will first provide a brief overview of the two most commonly used methods for protein structure determination for understanding the limitation of the experimental approaches, and then review the developments of protein structure prediction methods for solving this problem.

### 1.2.1 Experimental methods for protein structure determination

X-ray crystallography and NMR spectroscopy are the two most commonly used methods for determining atomic resolution of protein structures. Cryo-EM and Small angle X-ray and Neutron scattering (SAXS), are two other rapidly developing methods; however the resolution of the structure obtained using these two methods is generally limited and can therefore only be used for studying large macro-molecular complexes.

The inevitable step of structure determination using X-ray requires the target protein to be crystallized. However, getting a protein crystal is not a trivial problem because: (a) the protein needs to be highly purified and also available in sufficiently large quantity; and (b) the crystallization procedure is like voodoo, the exact conditions that are required for getting a protein crystal from a protein solution is unknown. Once a good protein crystal is obtained, it is

X-rayed from various angles and the resulting scattered X-rays are recorded at a detector (photographic film or a electron counter). The collected data is then fed into a computer that computes a spatial electron density map. The position of the protein atoms is then manually fitted into these densities.

In NMR spectroscopy, the protein is in solution, so crystallization is not required. However, highly purified protein in large quantity is still a prerequisite. Also, the protein needs to be stable in solution for several days. Atoms with odd number of neutron or proton are associated with a nuclear spin, and since the nuclei are charged, they develop a magnetic field. The spin of the nuclei in these atoms is randomly oriented, but when they are placed in a strong magnetic field, the spin gets aligned about the direction of magnetic field. In NMR spectroscopy of proteins, the spin of the magnetic nuclei of protein atoms are influenced by varying the magnetic field in multiple directions and one measures the absorption spectra as the atom spin returns to being aligned with the magnetic field. Based on the collected absorption pattern it is possible to determine how many bonds exist between two atoms and the approximate spatial distance between the two atoms. Based on this information, protein structures can be solved by satisfying the spatial distance restraints[21]. Nevertheless, the procedure of generating spatial distance restraints from the absorption spectra is non-automated and tedious, and hence generally not suitable for large proteins.

The limitations of experimental methods for protein structure elucidation has actuated the structural genomics (SG) project to increase the throughput of experimental structure elucidation [22; 23; 24] and provide a framework for inferring molecular function [25; 26]. While the SG aims to structurally characterize the protein universe by an optimized combination of experimental structure determination and comparative modeling (CM) of protein structures, 3D structures of at least 16,000 optimally selected proteins would be required for CM to cover 90% of protein domain families [27], and at the current rate it appears that this goal can be achieved only in about next 10 years[28]. This underscores the need of computational methods for protein structure prediction, so that 3D structural models can be built and provide insight for functional analysis. Also, the development of better structure prediction methods would dramatically enlarge the scope of structural genomics project.

### 1.2.2 Computational approaches for protein tertiary structure prediction

The goal of protein tertiary structure prediction is to estimate the spatial position of every atom of a protein. Protein structure prediction methods can be classified into three categories: Comparative modeling (CM) [29; 30], threading [31; 32; 33; 34; 35; 36] and *ab initio* modeling [37; 38; 39; 40; 41]. In CM, the protein structure is constructed by matching the sequence of the protein of interest (query) to an evolutionarily related protein of known structure (template) in the PDB [2], where the residue equivalency between query and the template is obtained by aligning sequences or sequence profiles. Threading-based methods match the query protein sequence directly to 3D structures of solved proteins with the goal of recognizing similar protein folds, which may have no clear evidence of an evolutionary relationship with the query protein. The last resort for predicting the protein structure, when no good template is detected in the PDB library, is to predict the structure using *ab initio* modeling. Predictions based on this method assume that the native structure of a protein corresponds to its global free energy minimum [42] and the conformational space is sampled to attain this state as guided by well designed energy force fields. This is the most difficult category of protein-structure prediction and if successful will provide the eventual solution to protein folding problem. However, the success of *ab initio* modeling is currently limited to small proteins with less than 100 amino-acids [37; 38; 39; 40; 41].

As a general trend in the field of protein structure prediction, the borders between the conventional categories of methods have become blurred. For instance, both comparative modeling and threading based methods use sequence-profile and profile-profile alignments for identifying templates. Similarly, most of the contemporary *ab initio* based methods often use evolutionary information either for generating sparse spatial restraints or for identifying local structural building blocks. Recent community-wide blind tests have demonstrated significant advantages of the composite approaches in protein structure predictions [43; 44; 45], which combines the various techniques from threading, *ab initio* modeling and atomic-level structure refinements [46; 47]. In the later chapters we will focus on the methodology of I-TASSER [38; 47; 48], which serves an example of composite approach for generating 3D structural models and predicting the function of a given query sequence.

### 1.2.2.1. CASP experiments

The community-wide **C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction (CASP) experiments (http://predictioncenter.org) provides a standard platform to assess state-of-the-art methods for protein structure and function prediction. During this biennial event, the organizers release a large number of protein sequences for which structure and function is unknown. The participants are then asked by the organizers to predict the structure and function of these proteins and submit their predicted models before provided deadlines. Finally, after the experiment is over, the experts of the field evaluate the predicted models based on obtained experimental results.

## 1.3. Functional annotations based on sequence-to-structure-to-function paradigm

Understanding the relationship between protein sequence, structure and function is often considered as the ‘holy grail’ of computational biology. In this section, we will review the sequence-to-structure-to-function paradigm and the potential challenges for any developed method based on this paradigm. But before we proceed, it is necessary to introduce the taxonomies of protein function and the concept of homology, which is crucial for understanding the discussion in later chapters.

### 1.3.1. Protein function

The definition of protein function is subjective and contextual. In this work, we have used three standard vocabularies for defining protein’s function: (a) EC number and (b) Gene Ontology terms and (c) ligand binding residues. For the benchmarking experiments, we have used the protein sequence and functional annotations from the PDB database.

#### 1.3.1.1. Enzyme Commission number

Enzymes are proteins that catalyze (i.e. increase the rates of) chemical reactions in physiological processes. Based on the reactions they catalyze, enzymes are categorized into hierarchical families using a numerical classification scheme known as Enzyme Commission (EC) number.

The EC numbers do not specify the enzymes, but the function characterized by the enzyme. The first number represents the type of enzymatic activity such as hydrolases (enzymes that cleave the substrate by hydrolysis), isomerases (enzymes which participate in intra-molecular rearrangement of the substrates) etc. All the enzymes can be categorized into six main classes based on their enzymatic activity (Table 1.2). The second number corresponds to the nature of

**Table 1.2 The six main classes of enzyme**

| Enzyme Class | Class Name | Reaction catalyzed |
|---|---|---|
| Class 1 | Oxidoreductases | Oxidoreduction reactions |
| Class 2 | Transferases | Transfers a group from one compound to the other |
| Class 3 | Hydrolases | Hydrolytic cleavage of C-O, C-N, C-C and phosphoric anhydrite bond |
| Class 4 | Lyases | Cleavage of C-C, C-O, C-N, and other bonds by elimination, leaving double bonds or rings |
| Class 5 | Isomerases | Intramolecular rearrangement of the substrates |
| Class 6 | Ligases | Joining of two molecules by utilizing ATP or a other triphosphate as energy source. |

chemical bonds or groups in the substrate on which the enzyme acts. The third number refers to nature of the cofactors required by the enzyme to catalyze the reaction. The fourth number corresponds to the nature of the substrate on which they act.

### 1.3.1.2. Gene Ontology

The gene ontology (GO) is currently the most effective approach for machine-legible and automatic functional annotation for both enzyme and non-enzymatic proteins. The Gene Ontology[6] (GO) is a widely used vocabulary for describing three different perspectives or “aspects” of gene functions: molecular function (MF), biological process (BP) and cellular component (CC). Each GO aspect is represented by a tree-like structured directed acyclic graph (DAG), where nodes in the graph represent a GO term and describe a component of gene product function (Figure 1.3), while the edges between the nodes are equivalent to the relationships (*is-a* or *part-of*) between the GO terms. The GO terms are held in a form of functional hierarchy, where functions that are more general are present on the top while functions that are more specific are further down the graph. Each GO term or node in the DAG can have multiple parents (for example transition metal binding and transcription regulator activity have two parent nodes), and is allowed to have multiple children. Moreover, since the edges are directed and the DAG is acyclic, we can never arrive at the same node if the edges are followed. Each protein can have multiple ascribed GO-terms, and all ancestors GO-terms are implied when a GO-term is assigned.

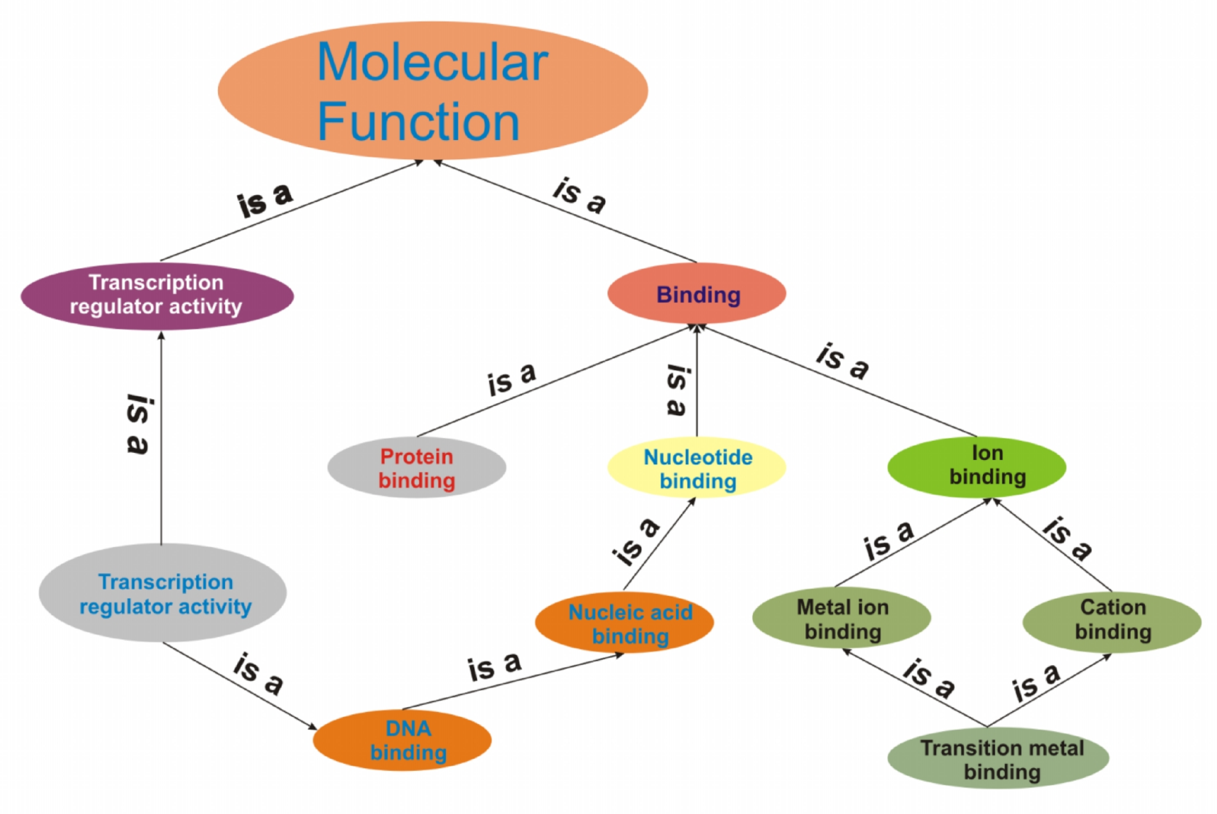


**Figure 1.3** An excerpt of a directed acyclic graph for molecular function.

## 1.3.2. The concept of homology

*Nature is an engineer*. The reason is quite obvious; it prefers to use existing resources within the cell for designing a new but related function, rather than inventing something from scratch, for e.g. the well known Rossmann fold is frequently used in many proteins for binding di-nucleotide co-enzymes[49]. This engineering process is quasi-static in nature. First, it requires gene duplication event to occur, so that a copy of the gene is available during the modification and the cell can alleviate the selection pressure. For a period of time, both the copies of the gene co-exist in the same species (S0 in Fig. 1.4). Eventually over a long period of time, either one or both the genes might be modified to perform a new but similar function. Since both the genes and their protein product are related by divergence from a common ancestor, they are ordained as ***homologs*** and their relationship as ***homology***. *Orthologs* and *paralogs* are two subcategories of homologs. While orthologs proceeds from speciation event (A1-A2 & B1-B2 in Fig 1.4), paralogs (A-B in Fig 1.4) arise after gene duplication; however, both have the potential to acquire new functional capabilities during the course of evolutionary divergence[50]. Nevertheless, orthologs most often perform same function and paralogs perform biologically distinct function. Homologous genes can also be transferred from one species to another via horizontal gene transfer i.e. without evolutionary decent, these are referred to as *xenologs*.

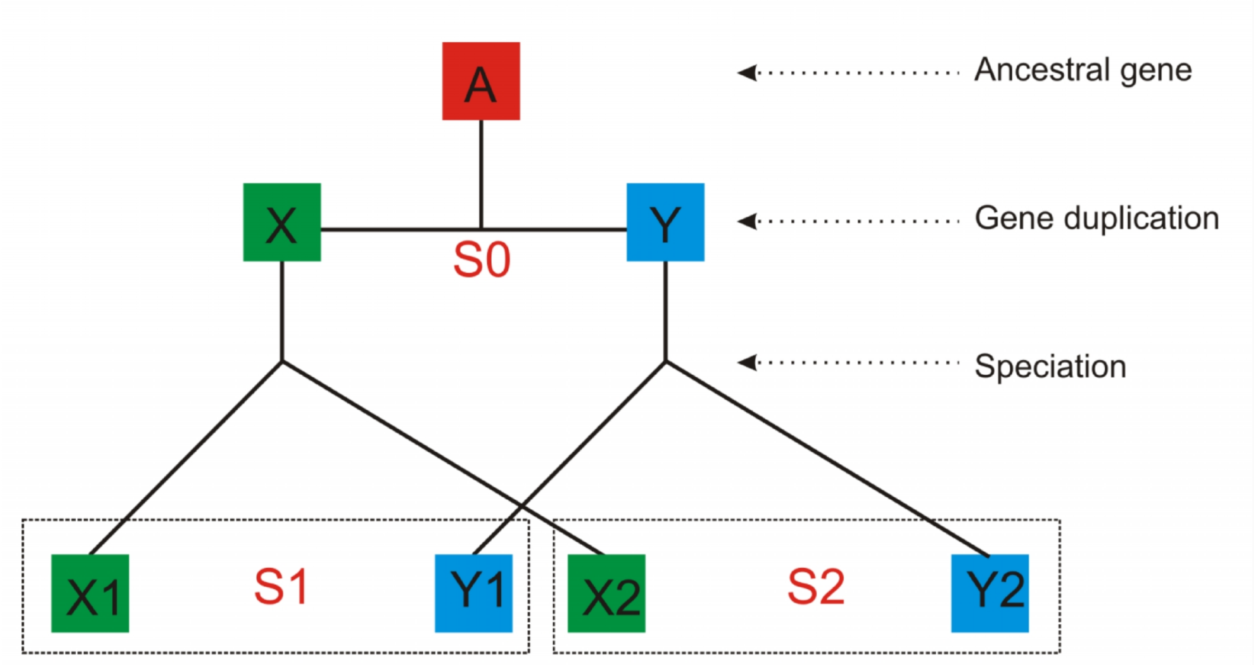


**Figure 1.4** Schematic diagram to explain the concept of homology. Evolutionary descent of an ancestral gene to paralogs and orthologs following gene duplication in species S0, and then speciation to yield species S1 and S2. Genes A, A1, A2, B, B1 and B2 have descended from a common ancestral gene. This picture has been adapted from 51.

Since all life forms on earth have a common ancestry, all the genes/proteins should ideally be expected to be homologous. But in parlance, they are not. The key concept missing in the descriptions provided above, is the time-scale required for the nature's engineering process, which is often difficult trace back. More appropriately, homology only pertains to genes/proteins when the evolutionary linkage is recent or detectable.

### 1.3.2.1. Functional annotations using homology

Once an evolutionary linkage is established, protein chemistry is often implicated. However, these linkages should be properly inspected before any functional inferences are drawn, because incorrect annotations can have far-reaching consequences like erroneous and costly experimental validation or incorrect functional assignments in the databases[52].

In most cases, homology between proteins is established based on sequence identity[14] or on sequence-profile comparision[16], where profile is a matrix containing family specific information of functionally and structurally important residues (refer to section 1.3.1.1 d). Thus, profile-based methods are generally more powerful in detecting homologous proteins than single sequence, because non-conservative substitutions during evolution chisel away the ancestor sequence to such an extent, that the sequence identity obtained based on a single sequence comparison between evolutionarily related and random protein is indistinguishable. Even when using these profiling tools, one needs to define a sequence similarity threshold for grouping proteins into families, which is often difficult to decide because some protein families are functionally more promiscuous than the others[53]. Moreover, the profile-based methods still have a lower limit of sequence identity threshold ("*twilight zone*")[54], where structure becomes an absolute necessity to warrant evolutionary relatedness.

Proteins domains are autonomously folding and functional units. As such, annotating protein at the domain level should be much more accurate[55]. However, identification of functional domains from sequence alone is still not satisfactory, and is usually successful only when a homologous template protein structure is available in the PDB library[56]. Identifying domain boundaries based on structural information is much easier and accurate[57; 58], which buttress the need of structures, either by experimental techniques or based on structure modeling. This is also one of the motivations for structural genomics projects: to make the 3D structures available for novel uncharacterized proteins.

Structure-based methods for homology detection are more powerful than sequence-alone-based methods, because in many cases evolution retains the folding pattern long after sequence similarity becomes undetectable[59]. The most general way of predicting the function from structure is to use global structure comparison and identify structural neighbors in structure databases with functional annotation. Although, putative relatives of a query can be identified based on these comparisons, the evidence is usually insufficient for transferring the functional annotations, as two proteins may have similar fold yet very different functions. For example, the classic fold alpha/beta barrel, is inhabited by five different classes of enzymes and also non-enzymatic proteins (Fig 1.5).

Since natural selection guides to optimize function, identifying conserved functional residues in structure can be very useful for assigning function. Many groups[60; 61; 62; 63] have focused on developing methods for identifying these active/binding site residues, as these are often more conserved than the overall fold and can be used for assigning the function. Nevertheless, these approaches often require a template library of known active/binding sites that can be pre-compiled either manually[64] or based on automation. But none of these libraries can be complete, because functional sites in many experimentally solved protein structures are still unknown.

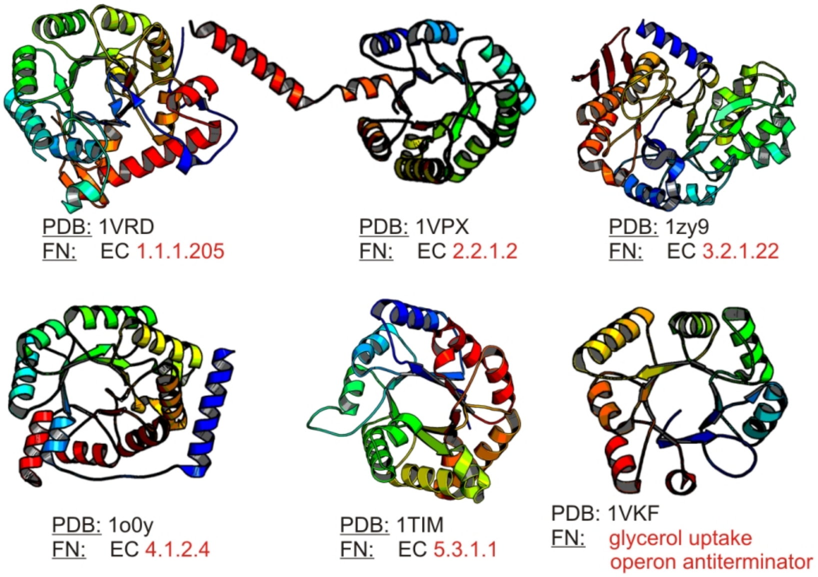


**Figure 1.5** Functional promiscuity in classic alpha/beta barrel.

## 1.4. Scope and outline of this work

Biological function of a protein is determined by its 3D shape, which dictates how the protein interacts with ligands or other protein molecules. However, experimental approaches for structural determination and functional characterization of proteins lag far behind the rapid increase in genome-wide sequence data. In this dissertation work, we have developed a unified automated platform for structural and functional annotation of protein sequences, based on the sequence-to-structure-to-function paradigm. The developed protocol is then tested on large-scale benchmarking experiments and real blind tests. The main aspects of the work are presented below:

- In the second chapter, we review the details and recent developments of tertiary structure prediction using the I-TASSER server, which forms the basis of functional annotations in later chapters. A large indescribable part of work related to this chapter is the maintenance of this prized server.
- In the third chapter, we present a new comparative approach (COFACTOR) to accurately recognize functional sites of protein-ligand binding interactions using low-resolution predicted protein structures by I-TASSER. We tested this algorithm in the recent community-wide CASP9 experiment, and it was ranked as the best method for binding-site prediction and outperformed all other participating algorithms.
- In the fourth chapter, we extend the developed COFACTOR algorithm for predicting other aspects of protein function, namely Enzyme commission (EC) numbers and Gene Ontology (GO) terms. A new approach that combines/utilizes both sequence and structure for functional homolog detection is introduced. As an illustration on the genome-wide functional annotations, the method was applied to the ORFs in the *Chlamydia trachomatis* genome, which revealed new testable insights distinct from the sequence-based annotations.
- In the last chapter, we conclude this thesis by providing a summary of the results and provide suggestions for future development.

## Bibliography

1. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142-8.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
3. Wiley, S. R. (1998). Genomics in the real world. *Curr Pharm Des* **4**, 417-22.
4. Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-5.
5. Martin, A. C. (2004). PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics* **20**, 986-8.

6. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

7. Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* **37**, D396-403.

8. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research* **37**, D32-6.

9. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53.

10. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-9.

11. Dayhoff, M. O., Schartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Natl. Biomed. Res. Found., Washington, DC.* (Dayhoff, M. O., ed.), pp. 353-358.

12. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7.

13. Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol* **266**, 227-58.

14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.

15. Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 2**, Unit 2 3.

16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

17. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* **84**, 4355-8.

18. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823-6.

19. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **32**, 922-923.

20. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.

21. Herrmann, T., Guntert, P. & Wuthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology* **319**, 209-27.

22. Gerstein, M., Edwards, A., Arrowsmith, C. H. & Montelione, G. T. (2003). Structural genomics: current progress. *Science* **299**, 1663.

23. Chandonia, J. M. & Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science* **311**, 347-51.

24. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science* **294**, 93-6.

25. Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nat Biotechnol* **18**, 283-7.

26. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* **311**, 395-408.

27. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nat Struct Biol* **8**, 559-66.

28. Zhang, Y. (2009). Protein structure prediction: when is it useful? *Curr Opin Struct Biol* **19**, 145-55.

29. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.

30. Fiser, A., Do, R. K. G. & Sali, A. (2000). Modeling of loops in protein structures. *Protein Science* **9**, 1753-1773.

31. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-70.

32. Wu, S. & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547-56.

33. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-9.

34. Xu, Y. & Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. *Proteins* **40**, 343-54.

35. Skolnick, J., Kihara, D. & Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* **56**, 502-18.

36. Wu, S. & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* **35**, 3375-82.

37. Bradley, P., Misura, K. M. S. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868-1871.

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

39. Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 5482-5485.

40. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for ab initio protein structural genomics. *Journal of Molecular Biology* **306**, 1191-1199.

41. Kihara, D., Lu, H., Kolinski, A. & Skolnick, J. (2001). TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A* **98**, 10125-30.

42. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-30.

43. Jauch, R., Yeo, H. C., Kolatkar, P. R. & Clarke, N. D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**, 57-67.

44. Kopp, J., Bordoli, L., Battey, J. N., Kiefer, F. & Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69**, 38-56.

45. Battey, J. N., Kopp, J., Bordoli, L., Read, R. J., Clarke, N. D. & Schwede, T. (2007). Automated server predictions in CASP7. *Proteins* **69**, 68-82.

46. Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmstrom, L., Wollacott, A. M., Wang, C., Andre, I. & Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**, 118-128.

47. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69 Suppl 8**, 108-17.

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

49. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature* **250**, 194-9.

50. Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics* **39**, 309-38.

51. Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol* **2**, INTERACTIONS1002.

52. Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in genetics : TIG* **17**, 429-31.

53. Roy, A., Srinivasan, N. & Gowri, V. S. (2009). Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biol* **9**, S41-55.

54. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94.

55. Reid, A. J., Yeats, C. & Orengo, C. A. (2007). Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics* **23**, 2353-60.

56. Ingolfsson, H. & Yona, G. (2008). Protein domain prediction. *Methods in molecular biology* **426**, 117-43.

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

58. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-108.

59. Whisstock, J. C. & Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics* **36**, 307-40.

60. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J Mol Biol* **351**, 614-26.

61. Kristensen, D. M., Ward, R. M., Lisewski, A. M., Erdin, S., Chen, B. Y., Fofanov, V. Y., Kimmel, M., Kavraki, L. E. & Lichtarge, O. (2008). Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **9**, 17.

62. George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., MacArthur, M. W., Porter, C. T., Al-Lazikani, B., Thornton, J. M. & Swindells, M. B. (2005). Effective function annotation through catalytic residue conservation. *Proc Natl Acad Sci U S A* **102**, 12299-304.

63. Xie, L. & Bourne, P. E. (2007). A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* **8 Suppl 4**, S9.

64. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci* **5**, 1001-13.

# Chapter 2

## I-TASSER server: an integrated platform for protein structure and function prediction

"I have a protein of interest but I don't know its structure/function" is one of the most common problems that most molecular and cell biologists face in their research. This impediment has been aggravated in recent years due to the fact that the percentage of protein sequences in UniProtKB/TrEMBL [1] with a solved protein structure in the PDB library [2] plunged to 0.4% by the end of 2011; this number was 0.6% at the end of 2009, 1.2% in 2007 and 2% in 2004 (Figure 1.1). Recent advances in computer algorithms for predicting protein structure and function have considerably alleviated this problem, and provided biologists with valuable information about their proteins of interest [3]. In this chapter, we will review the I-TASSER (*I*terative *T*hreading *ASSE*mbly *R*efinement) [4; 5; 6; 7] methodology for protein structure prediction, highlight the recent developments of popularly used I-TASSER server and provide guidelines for improving the structure and function modeling experiment. The server is freely available for the academic community at http://zhanglab.ccmb.med.umich.edu/I-TASSER.

## 2.1. An overview of the I-TASSER pipeline

I-TASSER [4; 5; 6; 7] is a hierarchical protein structure modeling approach based on the multiple threading alignments and an iterative implementation of the Threading ASSEmbly Refinement (TASSER) program [8]. Figure 2.1 shows the schematic representation of I-TASSER pipeline for protein structure and function prediction, which consist of four consecutive steps of threading, structure assembly, structure refinement and function prediction. In this chapter, we will only focus on the structure prediction module of the pipeline, as the function prediction module will be discussed in the following two chapters.

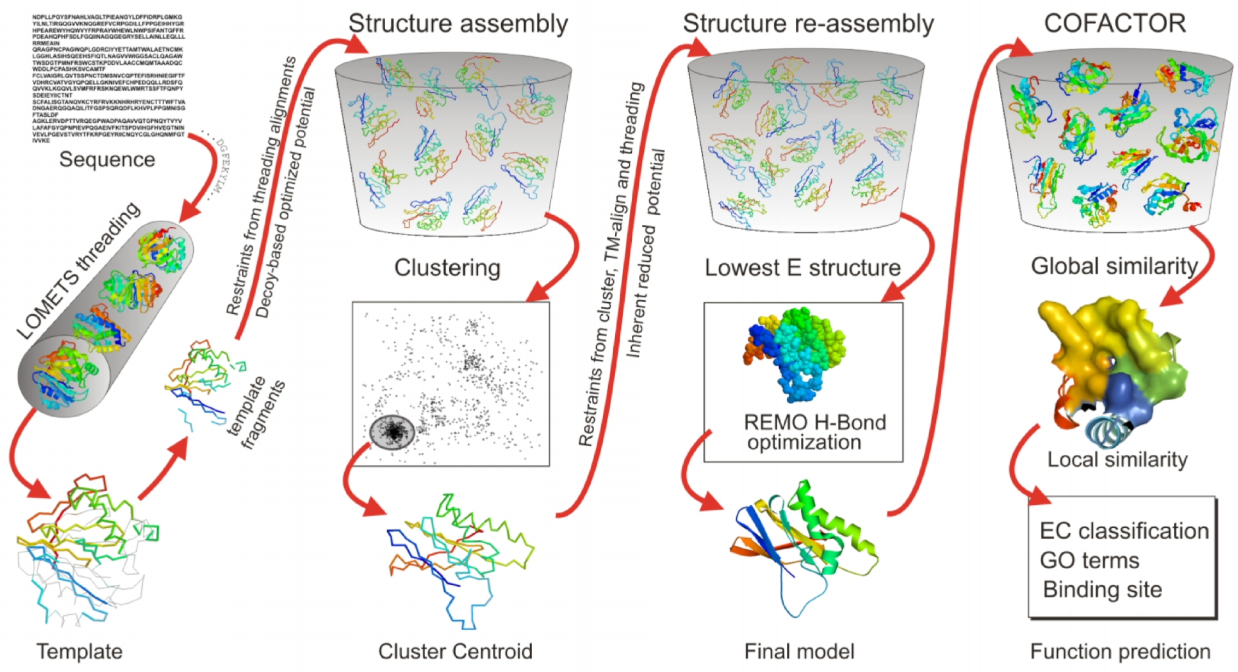


**Figure 2.1** Schematic diagram of the I-TASSER protein structure and function prediction protocol.

### 2.1.1. Threading of query sequence

Threading refers to a bioinformatics procedure for identifying template proteins from solved structure databases that have a similar structure or similar structural motif as the query protein sequence. In the first stage of I-TASSER, the query sequence is matched against a non-redundant (nr) sequence database by PSI-BLAST [9], to identify evolutionary relatives. A sequence profile is then created based on a multiple sequence alignment (MSA) of the homologs, which is also used to predict the secondary structure using PSSpred (http://zhanglab.ccmb.med.umich.edu/PSSpred). Assisted by the sequence profile and the predicted secondary structure, the query sequence is then threaded through a representative PDB structure (sequence identity cutoff of 70%) with the objective of identifying the global or local threading alignments using either MUSTER [10] (single threading server) or LOMETS [11] (meta-threading server). In this section, we will first describe the methodology of MUSTER threading algorithm and then give an overview and advantage of using LOMETS.

#### 2.1.1.1. MUSTER threading program

MUSTER is a sequence profile-profile alignment (PPA) method assisted by the predicted structural information like secondary structure, structure profiles, solvent accessibility, backbone dihedral torsion angles, and hydrophobic scoring matrix. The scoring function of MUSTER for aligning the *i*th residue of the query and the *j*th residue of the template is defined as [10]

$$Score(i,j) = \mathrm{E}_{\mathrm{seq-prof}} + \mathrm{E}_{\mathrm{sec}} + \mathrm{E}_{\mathrm{struc-prof}} + \mathrm{E}_{\mathrm{sa}} + \mathrm{E}_{\mathrm{phi}} + \mathrm{E}_{\mathrm{psi}} + \mathrm{E}_{\mathrm{hydro}} + \mathrm{E}_{\mathrm{shift}}. \qquad (2.1)$$

The first term $\mathrm{E}_{\mathrm{seq\_prof}}$, is the alignment score of the sequence profile-profile alignment. The second term $\mathrm{E}_{\mathrm{sec}}$ , computes the match between the predicted secondary structure of query and known secondary structure of templates. The third term $\mathrm{E}_{\mathrm{struc\_prof}}$, calculates the score of aligning the structured-derived profiles of templates to the sequence profile of query. The fourth term $\mathrm{E}_{\mathrm{sa}}$ , computes the difference between the predicted solvent accessibility of query and solvent accessibility of templates. The fifth and sixth terms ($\mathrm{E}_{\mathrm{phi}}$ and $\mathrm{E}_{\mathrm{psi}}$), calculate the difference between the predicted torsion angles (phi and psi) of query and those of templates. The experimental torsion angles for templates are calculated using STRIDE [12], while torsion angles of query are predicted by ANGLOR [13]. The seventh term $\mathrm{E}_{\mathrm{hydro}}$ , is an element of hydrophobic scoring matrix [14] that encourages the match of hydrophobic residue (V, I, L, F, Y, W, M) in the query and the templates. Finally, the last term $E_{\text{shift}}$ is a constant, which is introduced to avoid alignment of unrelated residues in local regions. While the first term is sequence-based information, the second to seventh terms are related to structural information. If only the first two terms plus $E_{\text{shift}}$ in equation 1 are involved, the corresponding threading program is called profile-profile alignment (PPA) [11], which is the precursor of MUSTER.

The sequence and structural information are then combined into a single-body energy term, which can be conveniently used in the Needleman-Wunsch [15] dynamic programming algorithm (Section 1.1.3.1) for identifying the best match between the query and the templates. A position-dependent gap penalty in the dynamic programming is employed, i.e. no gap is allowed inside the secondary structure regions (helices and strands); gap opening ($g_o$) and gap extension ($g_e$) penalties apply to other regions; ending gap-penalty is neglected.

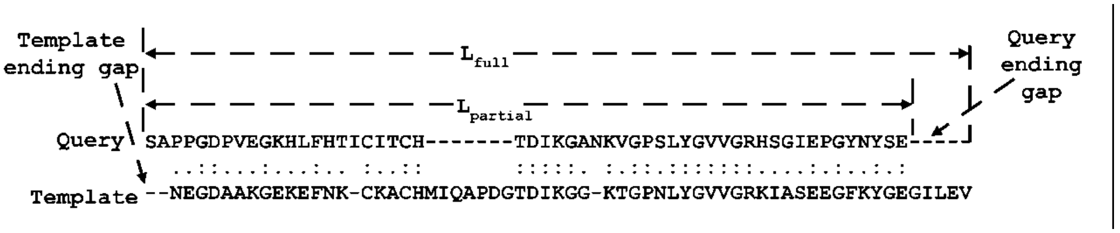

Template
ending gap
$L_{full}$
Query
ending
gap
$L_{partial}$
Query SAPPGDPVEGKHLFHTICITCH-------TDIKGANKVGPSLYGVVGRHSGIEPGYNYSE----
..::...:...:.. :..::       :::::. :.::.::::::::...:.:..:.:
Template --NEGDAAKGEKEFNK-CKACHMIQAPDGTDIKGG-KTGPNLYGVVGRKIASEEGFKYGEGILEV

**Figure 2.2** Illustration of MUSTER threading alignment score calculation. Full (Lfull) and partial (Lpartial) alignment lengths are used to normalize the threading alignment score (Rscore). Symbols '-', '.' and ':' indicate an unaligned gap, an aligned non-identical residue pair and an aligned identical residue pair, respectively. The query and template sequences are taken from 1hroA (first 53 residues) and 155c_ (first 61 residues), respectively, as an illustrative example. (Taken from Wu, S. and Zhang, Y. Proteins 72(2008): 550).

Following the dynamic programming alignments, the alignments of different structural templates are ranked based on their alignment score and the length of the alignment. In PPA [11], the templates are ranked based on a raw alignment score ($R_{\text{score}}$) divided by the full alignment length ($L_{\text{full}}$) (including query and template ending gaps) as shown in Figure 2.2. In MUSTER, however, $R_{\text{score}}/L_{\text{partial}}$ is used as an another possible ranking scheme, where

$L_{\text{partial}}$ is the partial alignment length excluding query ending gap as shown in Figure 2.2. A combined ranking is then taken as follows: If the sequence identity of the first template selected by $R_{\text{score}}/L_{\text{partial}}$ to the query is higher than that selected by $R_{\text{score}}/L_{\text{full}}$, then the template ranking is done based on $R_{\text{score}}/L_{\text{partial}}$. Otherwise, the templates are ranked based on $R_{\text{score}}/L_{\text{full}}$.

### 2.1.1.2. LOMETS: Meta-threading server

As observed in the CASP experiments [16], although the average TM-score of MUSTER outperforms many of the state-of-art algorithms, a single threading program can never be better than all other threading algorithms on every target. This inconsistency naturally leads to the prevalence of the meta-server [11; 17], which is designed to collect and combine prediction results from a set of individual threading programs.

On the I-TASSER web-server, this idea has been implemented using LOMETS [11], a locally installed meta-threading server. The threading programs in LOMETS represent a diverse set of the state-of-the-art algorithms using different approaches, namely: Sequence profile alignments (PPA-I [11], PPA-II [11], SPARKS2 [11], SP3 [18]), structural profile alignments (FUGUE [19]), pairwise potentials (PROSPECT2 [20]), and the hidden Markov models (HHsearch [21], SAM-T02 [22]). In the individual threading programs, the templates are ranked by a variety of sequence-based and structure-based scores. The top template hits from each threading program are then selected for further consideration. The quality of the template alignments (and therefore the difficulty of modeling the targets) is assessed based on normalized *Z-score,* which is defined as:

$$Norm.\,\text{Z-score} = \frac{Z\text{-}score}{Z_0} \tag{2.2}$$

where, *Z-score* is the score in standard deviation units relative to the statistical mean of all alignments generated by the program; and $Z_0$ is a program-specific *Z-score* cutoff determined based on large-scale threading benchmark tests [11] to differentiate 'good' from 'bad' templates.

For each target, LOMETS first threads the query sequence through the PDB library to identify template threading alignments by each threading program and then ranks them purely based on consensus. The idea behind the consensus approach is simple: there are more ways for a threading program to select a wrong template than that to select a right one. Therefore, the odds of multiple threading programs working collectively to make a common wrong selection is lower than the chance to make a common correct selection.

**Table 2.1** Performance comparison of component threading programs and LOMETS meta-server on 620 non-homologous testing proteins. (Taken from Wu, S. and Zhang, Y. Nuc. acid res. 35(2007): 3375).

| Threading servers or meta-servers | TM-score (MODELLER models) | | RMSD (Å) (MODELLER models) | |
|---|---|---|---|---|
| | First model | Best in top five models | First model | Best in top five models |
| PPA-I | 0.4117 | 0.4531 | 16.66 | 14.02 |
| SP3 | 0.4138 | 0.4551 | 13.86 | 12.83 |
| PPA-II | 0.4076 | 0.4512 | 14.89 | 13.02 |
| SPARKS2 | 0.3973 | 0.4441 | 13.60 | 12.23 |
| PROSPECT2 | 0.3914 | 0.4384 | 13.01 | 12.02 |
| FUGUE | 0.3721 | 0.4173 | 19.26 | 15.82 |
| HHSEARCH | 0.3827 | 0.4224 | 22.38 | 19.04 |
| SAM-T02 | 0.3575 | 0.3971 | 21.75 | 17.53 |
| LOMETS | 0.4434 | 0.4669 | 10.99 | 10.61 |

Table 2.1 shows the improvement of LOMETS over individual threading programs. For the purpose of eliminating the dependence on the alignment coverage, the full-length models have been built here by MODELLER [23], using the templates from each threading program. Based on 620 non-homologous testing proteins, the models generated by LOMETS threading alignments achieves an average TM-score of 0.4434, which is at least 8% higher than that by any individual threading program.

### 2.1.2. Structure assembly and refinement

Following the threading procedure, continuous fragments in threading alignments are excised from template structures, and are used to assemble structural conformations of the sections that aligned well, while the unaligned regions (mainly loops/tails) are built by *ab initio* modeling [6; 24].

For a given threading alignment, I-TASSER first builds an initial full-length model by connecting the continuous secondary structure fragments (≥ 5 residues) through a random walk of $C_\alpha$-$C_\alpha$ bond vectors of variable lengths from 3.26 to 4.35Å. To guarantee that the last step of this random walk can quickly arrive at the first $C_\alpha$ of the next template fragment, the distance $l$ between the current $C_\alpha$ and the first $C_\alpha$ of the next template fragment is checked at each step of the random walk, and only walks with $l < 3.54n$ are allowed, where $n$ is the number of remaining $C_\alpha$-$C_\alpha$ bonds in the walk. If the template gap is too big to be spanned by a specified number of unaligned residues, a big $C_\alpha$-$C_\alpha$ bond is kept at the end of the random walk and a spring-like force that acts to draw sequential fragments close will be applied during Monte Carlo simulations, until a physically reasonable bond length is achieved.

To improve the efficiency of conformational search, I-TASSER adopts a reduced model to represent the protein chain, with each residue described by its Cα atom and its side-chain center of mass. Because the regions not aligned during the threading process usually have a lower modeling accuracy, the structure modeling in these regions is confined to a lattice system of grid size 0.87 Å [24], which helps to reduce the entropy of conformational search. Although this grid size may introduce considerable uncertainty of conformational representations in comparative modeling (which usually has an error range of 1-2 Å), it does not generate observable effect in the *ab initio* modeling, as it often has an error range of 4-6 Å. The threading aligned regions usually have a higher accuracy. Modeling in these regions is therefore off lattice and the template fragments are kept rigid during the simulations, which helps to maintain the fidelity of the high-resolution structures in these regions.

Next, the assembled structure is refined using a replica-exchange Monte Carlo simulation technique [25], which implements several replica simulations in parallel at different temperatures, with the temperatures periodically exchanged between the replicas; the energy barriers are flattened by a hyperbolic function to speed up the jumps of simulations between different energy basins. The overall simulation is guided by a composite knowledge-based force field, which is described in the next section.

## 2.1.2.1. I-TASSER force field

The I-TASSER simulations are guided by a composite knowledge-based force field, which includes: (1) general statistical terms derived from the PDB (C-alpha/side-chain correlations [24], H-bonds [26] and hydrophobicity [27]); (2) variety of statistical short-range and long-range correlation terms that are extracted from multiple threading alignments [11]; and (3) sequence-based contact predictions from SVMSEQ [28]. Partly because of the consideration of the hydrophobic interactions and the bias towards radius of gyration in the energy force field, the current I-TASSER procedure is designed to best fold single-domain globular proteins (the procedure for modeling multiple-domain proteins is discussed section 2.3). Readers are recommended to read Zhang and Skolnick [8; 29; 30] for further details about these energy terms.

### 2.1.2.2. Iterative strategy

The conformations generated in the low-temperature replicas during the refinement simulation are clustered by SPICKER [31], with the purpose of identifying low free-energy states. The cluster centroids are obtained by averaging all the clustered structures after superposition, and are ranked based on the structure density of the clusters. However, the cluster centroids generally have a number of non-physical steric clashes between $C_\alpha$ atoms and can be over-compressed. Starting from the selected SPICKER cluster centroids, the TASSER Monte Carlo simulation [25] is performed again (see Figure 2.1). While the inherent I-TASSER potential remains unchanged in the second run, external constraints are added, that are derived by pooling the initial high-confident restraints from threading alignments, the distance and contact restraints from the combination of the centroid structures and the PDB structures identified by the structure alignment program TM-align [32] using the cluster centroids as query structures. The conformation with the lowest energy in the second round is selected as the final model.

The main purpose of this iterative strategy is to remove the steric clashes of the cluster centroids. On a benchmark test set of 200 proteins with < 300 residues it was found that the average number of steric clashes (residue pairs with $C_\alpha$ distance < 3.6Å) for the cluster centroids of the first cluster dramatically reduces from 79 to 0.8. As strong distance map and contact restraints are implemented in this step, the topology of the models also improves. In these test cases, the average TM-score increased from 0.5734 to 0.5801 (1.2%) and the $C_\alpha$-RMSD to native decreased from 6.67Å to 6.52Å compared with the cluster centroid of the first round.

### 2.1.3. Construction of full atomic model

The models generated after I-TASSER Monte-Carlo simulations [25] and SPICKER clustering [31] are reduced models, where each residue is represented by the $C_\alpha$ atom and side-chain center of mass. To increase the biological usefulness of protein models, full-atomic models are constructed by REMO [33] from these cluster centroids, while optimizing the hydrogen-bonding network, where a H-bonding list is pre-constructed based on secondary structure predictions and the 3D backbone model. REMO can quickly build the initial full-atomic models from Cα traces but often the models have distortions in local structure and side-chain atoms. Finally, all the models generated by REMO are submitted to FG-MD (Fragment Guided-Molecular Dynamics) [34], with the purpose of improving the local geometry and hydrogen bonding, and reduce backbone and side-chain steric clashes in the model. FG-MD simulations are carried out in vacuum, as implemented in LAMMPS [35] package. The force field consists of energy terms from Amber99 [36], $C_\alpha$ repulsive potential, statistical hydrogen-bonding potential and distance restraints collected from both the template and structural fragments searched by TM-align[32] in the PDB library, using initial model as the probe. The distance restraints are generated by combination of distance maps from initial model, TM-align global template and TM-align fragments at each location. FG-MD refinement simulation is the last step of structure predictions pipeline.

In the last stage, the function of the query protein is inferred by structurally matching the predicted 3D models against the proteins of known structure and function in the PDB, using global and local structure alignment algorithms. We will discuss this module in detail in the following chapters.

## 2.2. Estimating the accuracy of predicted protein structure

Assessing the quality of a prediction is important because this assessment eventually determines how biologists will use the predicted model in their research. For estimating the accuracy of the structure predictions, a confidence score named C-score is defined, based on the quality of the threading alignments and the convergence of the I-TASSER's structural assembly refinement simulations, mathematically formulated as:

$$\mathrm{C-score} = \ln\left[\frac{M}{M_{tot}} \times \frac{1}{\langle RMSD \rangle} \times \sum_{i=1}^{N} Norm.\,Z-score(i)\right] \tag{2.3}$$

where $M$ is the multiplicity of structure decoys in the structural clusters identified by the SPICKER [31]; $M_{\mathrm{tot}}$ is the total number of decoys submitted to the clustering; ⟨RMSD⟩ is the average RMSD of the clustered decoys to the cluster centroids; *Norm.Z-Score*(*i*) is the normalized Z-score (Eq. 2.2) of the top threading alignment obtained from *i*th threading server in LOMETS [11]; *N* is the number of servers used in LOMETS.

The C-score scheme has been extensively tested in large-scale benchmarking tests [37; 38]. When tested on predicted structures, the Pearson correlation between C-score and the TM-score (the absolute difference between model to the native structure) was found to be 0.91, which is a significantly higher value, keeping in mind that the mathematic range of the Pearson correlation is between 0 (for random variables) and 1 (for identical variables). When a C-score cut-off of -1.5 is used to select models of correct topology, both the false positive and the false negative rate are below 0.1, which means that more than 90% of the quality predictions are correct. On combining C-score and protein length, the accuracy of the I-TASSER models can be predicted with an average error of 0.08 for the TM-score and 2 Å for the RMSD (root mean square deviation) [38]. Again, considering the big quality variations of protein structure predictions (i.e. TM-score in 0-1 and RMSD in 0~30Å), these estimation errors are very low and the assessments should provide quantitative guidance of model quality to the users.

## 2.3. Structure modeling of multi-domain proteins

Since the I-TASSER force field has been designed for modeling single-domain proteins, the procedure for modeling multiple-domain proteins is a slightly different, but fully automated process. First, the domain boundaries are defined based on the LOMETS threading programs, i.e. if a segment of query sequence of >80 residues has no alignment with template proteins in top two threading hits, the target is treated as a multiple-domain protein and the domain boundary is defined at the borders of the aligned/unaligned sections. Next, two types of assembly simulations are implemented: one simulation is conducted for modeling the whole-chain structure, which provides a guide for domain orientations; another simulation is carried out for modeling the single-domain structures individually. Finally, to obtain the full-length model, the models of individual domains are docked together using the whole-chain I-TASSER model as a template. The docking simulation is performed using a quick Metropolis Monte Carlo simulation where the energy is defined as the RMSD of the individual domain models to the whole-chain I-TASSER template plus the reciprocal of the number of inter-domain steric clashes. The purpose is to generate a global model, which has a similar domain orientation to the whole-chain I-TASSER model but with the minimum number of steric clashes. This procedure is applied only to proteins that have some domains that are not aligned in the top-scoring templates. If multi-domain templates are available and all domains of query protein are aligned, the whole chain will be modeled in I-TASSER using the full-chain template.

If the domain boundary information is available to the user, e.g. from some experimental data, it is recommended that the user should first split the sequence into individual domains and then submit each domain individually to the server. This will not only speed up the I-TASSER prediction process but also result in a more reliable structure and function prediction, since the current pipeline of the I-TASSER methodology has been optimized for modeling single-domain proteins[5]. Domain boundaries in protein sequences can also be predicted by using freely available external online programs such as PFAM (http://pfam.sanger.ac.uk/) or NCBI CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

## 2.4. Using biological knowledge to improve structure modeling

### 2.4.1. Provide external restraints

The structure assembly simulations in I-TASSER are mainly guided by spatial restraints collected from the LOMETS threading templates. For query proteins, that have good threading hit (*Norm. Z-score* >1) in the template library, derived spatial restraints are mostly of high accuracy and I-TASSER will generate high-resolution structural models for these proteins. In contrast, for query proteins that have weak or no threading hit, collected spatial restraints often contain errors because of the uncertainty of the template and the alignment.

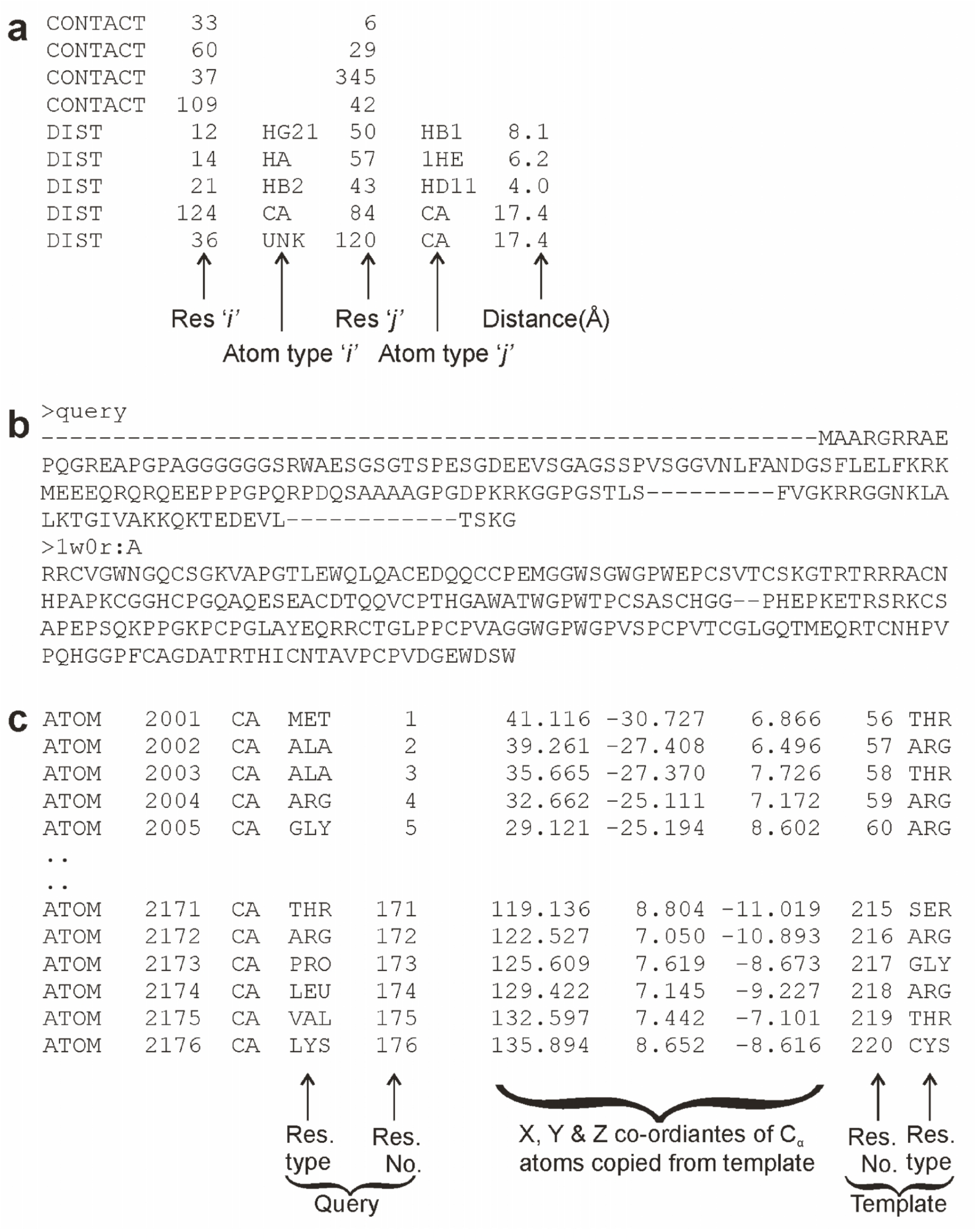
The I-TASSER server allows the user to specify additional restraints based on experimental evidence or biological insights. Because restraints from experiments normally have a higher accuracy than those derived from threading alignments, user-specified spatial information can be very useful for improving the quality of the structure assembly, especially for the non-homologous protein targets. Our benchmark test shows that by using as few as *N*/8 NOE restraints, obtained from the NMR experiments (where *N* is the length of the protein), the current simulation procedure is able to successfully fold 75% of the proteins of up to 200 residues, which could not be folded without using spatial restraints because of the lack of appropriate templates [39]. Users can provide external restraints to the I-TASSER server in two ways:

1. *Specify contact/distance restraints*

   Specify a restraint file with experimentally characterized inter-residue contacts/distances, for example from NMR or cross-linking experiments. An example file is shown in Figure 2.3a, where Column 1 specifies the type of restraint, i.e. "DIST" or "CONTACT". For distance restraint (DIST), columns 2 and 4 contain residue positions (*i*, *j*), columns 3 and 5 contain the atom-types in the residue and column 6 specifies the distance between the two specified atoms. For contact restraints (CONTACT), columns 2 and 3 contain the positions (*i*, *j*) of residues, which should be in contact. The distance between the side chains center of these contacting residue pairs is decided based on observed distances in known structures in PDB. I-TASSER will try to draw these atom pairs close to the specified distance during the structure refinement simulations.

2. *Specify a protein structure template*

   I-TASSER normally starts with a set of protein templates identified by the LOMETS threading programs, where the template library consists of a representative PDB subset at a pair-wise sequence identity cutoff of 70%. Users can specify a solved protein structure as the template, as the desired template may not be included in our library or the desired template may not be identified by LOMETS even though it is in the library.



```
CONTACT   33          6
CONTACT   60         29
CONTACT   37        345
CONTACT  109         42
DIST      12   HG21  50   HB1    8.1
DIST      14   HA    57   1HE    6.2
DIST      21   HB2   43   HD11   4.0
DIST     124   CA    84   CA    17.4
DIST      36   UNK  120   CA    17.4
```

```
>query
------------------------------------------------------MAARGRRAE
PQGREAPGPAGGGGGGSRWAESGSGTSPESGDEEVSGAGSSPVSGGVNLFANDGSFLELFKRK
MEEEQRQRQEEPPPGPQRPDQSAAAAGPGDPKRKGGPGSTLS---------FVGKRRGGNKLA
LKTGIVAKKQKTEDEVL------------TSKG
>1w0r:A
RRCVGWNGQCSGKVAPGTLEWQLQACEDQQCCPEMGGWSGWGPWEPCSVTCSKGTRTRRRACN
HPAPKCGGHCPGQAQESEACDTQQVCPTHGAWATWGPWTPCSASCHGG--PHEPKETRSRKCS
APEPSQKPPGKPCPGLAYEQRRCTGLPPCPVAGGWGPWGPVSPCPVTCGLGQTMEQRTCNHPV
PQHGGPFCAGDATRTHICNTAVPCPVDGEWDSW
```

```
ATOM   2001  CA  MET     1      41.116 -30.727   6.866   56 THR
ATOM   2002  CA  ALA     2      39.261 -27.408   6.496   57 ARG
ATOM   2003  CA  ALA     3      35.665 -27.370   7.726   58 THR
ATOM   2004  CA  ARG     4      32.662 -25.111   7.172   59 ARG
ATOM   2005  CA  GLY     5      29.121 -25.194   8.602   60 ARG
..
..
ATOM   2171  CA  THR   171     119.136   8.804 -11.019  215 SER
ATOM   2172  CA  ARG   172     122.527   7.050 -10.893  216 ARG
ATOM   2173  CA  PRO   173     125.609   7.619  -8.673  217 GLY
ATOM   2174  CA  LEU   174     129.422   7.145  -9.227  218 ARG
ATOM   2175  CA  VAL   175     132.597   7.442  -7.101  219 THR
ATOM   2176  CA  LYS   176     135.894   8.652  -8.616  220 CYS
```

Res. type
Res. No.
Query
X, Y & Z co-ordiantes of $C_\alpha$ atoms copied from template
Res. No.
Res. type
Template

**Figure 2.3** Example of external restraint files for specifying (a) residue-residue contact/distance restraints; (b) query-template alignment in FASTA format; and (c) query-template alignment in 3D format.

To specify a template, users can either upload a PDB formatted structure file or input a PDB ID and the I-TASSER server will obtain the structure from the PDB library. Once a template is specified, the I-TASSER simulation will start from the template with restraints mainly collected from it; but the simulation will also use the threading-based LOMETS restraints with the purpose to model the unaligned regions as well as adjust the reassembly of aligned regions.

The weight of the LOMETS restraints varies depending on the target type. Here, the query proteins are categorized into easy or hard targets based on the statistical significance of the threading alignments. The templates for easy targets are usually from homologous proteins and the alignments have a higher accuracy, while templates for hard targets are mostly from non-homologous proteins and the alignments have a lower accuracy. Because the accuracy of the LOMETS restraints is different for different targets, the weight of implementing the LOMETS restraints is stronger for easy targets than that in the case of hard targets, that have been systematically tuned based on large-scale benchmark training [11]. LOMETS threading programs use a representative PDB library to find plausible folds for the query protein. Although using a representative structure library helps to reduce the time required to compute the sequence-structure alignments, it is possible that a good template protein is missed in the library or the template may not have been identified by LOMETS threading programs, even though it is present in the library. In these cases, the user should specify the desired protein structure as the template.

To specify protein structure as an additional template, users can either upload a PDB formatted structure file or specify the PDB ID of a deposited protein structure in PDB library. The I-TASSER will generate the query-template alignment using MUSTER program and will collect spatial restraints from both the user specified template and

LOMETS templates to guide the structure assembly simulation. Because the accuracy of the LOMETS restraints is different for different targets, the weight of the LOMETS restraints is stronger in easy (homologous) targets than that in hard (non-homologous) targets that have been systematically tuned in our benchmark training.

Users can also specify their own query-template alignments. The server accepts alignment in two formats: the FASTA format (Figure 2.3b) and the 3D format (Figure 2.3c). The FASTA format is standard and is described at http://zhanglab.ccmb.med.umich.edu/FASTA. The 3D format is similar to the standard PDB format, but two additional columns derived from the templates are added to the ATOM records (see Figure 2.3c):

*Columns 1-30:* Atom (C-alpha only) and residue names for the query sequence.

*Columns 31-54:* Coordinates of C-alpha atoms of the query copied from the corresponding atoms in the template.

*Columns 55-59*: Corresponding residue number in the template based on alignment

*Columns 60-64*: Corresponding residue name in the template

### 2.4.2. Exclude template proteins

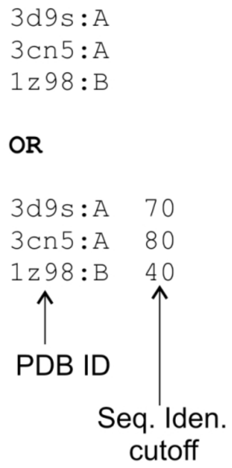
Proteins are flexible molecules and can adopt multiple conformational states to change their biological activity. For example, structures of many *protein* kinases and membrane proteins have been solved in both *active* and *inactive* conformation. Also, presence or absence of bound ligand can cause large structural movements. While all the conformational states of the template are alike for the threading programs, it is desirable to model the query using templates in only one particular state. A new option on the server allows the user to exclude template proteins during structure modeling. This feature would also allow the user to choose the homology level of templates to be used for the modeling. Users can exclude template proteins from the I-TASSER library by:

A. *Specifying a sequence identity cutoff*

Users can use this option to exclude homologous proteins from the I-TASSER template library. The homology level is set, based on the sequence identity cutoff i.e. the number of identical residues between the query and the template protein divided by the sequence length of the query sequence. For example, if the user types in "70%" in the provided form, all templates proteins that have a sequence identity >70% to the query protein will be excluded from the I-TASSER template library.

*B. Exclude specific template proteins*

Specific template proteins can be excluded from the I-TASSER template library by uploading a list containing PDB IDs of the structures to be excluded. An example file is shown in Figure 2.4. As the same protein can exist as multiple entries in the PDB library, I-TASSER server will by default exclude the specified templates (in Column1) as well as all other templates from the library that have an identity >90% to the specified templates. Users can also specify a different identity cutoff, e.g. 70%, where all templates with identity >70% to specified template proteins will be excluded.



```
3d9s:A
3cn5:A
1z98:B
```

**OR**

```
3d9s:A  70
3cn5:A  80
1z98:B  40
```

PDB ID
Seq. Iden. cutoff

**Figure 2.**4 An example file used for excluding template during I-TASSER structure modeling procedure. The first column contains the PDB ID of the template proteins to be excluded. The second column is used to specify the sequence identity cut-off, which will be used for other similar templates in the template library.

## 2.5. Conclusion

The biological usefulness of the predicted protein models relies on the accuracy of the structure prediction [40]. For example, high-resolution models with RMSD in range of 1-2 Å, typically generated by CM using close homologous templates, usually meet the highest structural requirements and are sometime suitable for computational ligand-binding studies and virtual compound screening [41; 42; 43]. Medium-resolution models, roughly in the RMSD range of 2-5 Å are typically generated by threading and CM from distantly homologous templates, and can be used for identifying the spatial locations of functionally important residues, such as active sites and the sites of disease-associated mutations [44; 45; 46; 47]. However, many of the functionally important sites are located on the loops that show large structural variability although the scaffold of the protein structure is conserved. Thus, accurate modeling of loop regions is still an important yet unsolved problem in template-based modeling [48; 49]. Nevertheless, even models with the lowest resolution, from an otherwise meaningful prediction, i.e. models with an approximately correct topology, predicted using either *ab initio* approaches or based on weak hits from threading, have a number of uses including protein domain boundary identification [50; 51], topology recognition, and family/superfamily assignment [52; 53].

The success of the I-TASSER methods in the blind CASP experiments [4; 5] and the large-scale benchmarking tests [6; 30; 37; 53] makes it a useful tool for automated protein structure and function annotation. In the past 36 months, the online I-TASSER server has generated >60,000 full-length structure and function predictions for over 20,000 registered scientists from 103 countries. Compared to a number of other useful on-line structure prediction tools [17; 21; 54; 55; 56; 57; 58; 59; 60], the uniqueness of the I-TASSER sever is in its significant accuracy and reliability of full-length structure prediction for protein targets of varying difficulty and the comprehensive structure-based function predictions. Especially, the inherent template fragment reassembly procedure has the power to consistently drive the initial template structures closer to the native structure [6; 16; 61]. For example, in CASP8, the final models generated by the I-TASSER server had a lower RMSD to the native structure than the best threading template for 139 out of 164 domains, with an overall RMSD reduction by 1.2 Å (on average from 5.45 Å in templates to 4.24 Å in the final models) [5].

It needs to be mentioned that despite extensive benchmark tests [4; 5; 38], there can be considerable uncertainty and error in the automated estimation of the quality of structure and function predictions. The final and essential validation of the predictions should therefore be made based on the experimental data collected by the users. Before the entire structure becomes available, other indirect structural information from the data like mutagenesis experiments, affinity labeling, NMR dipolar coupling, cryo-electron microscopy, circular dichroism and dual polarization interferometry experiments, can provide important information for validating the predicted models and help in deciding whether the predictions can be useful for further experimental design and study.

The purpose of this chapter was to provide a detailed overview of the I-TASSER pipeline, because the predicted models will be used for function prediction experiments in the later chapter. Meanwhile, since the I-TASSER server is based on the general sequence-to-structure-to-function paradigm, the described methodology can be valuable for the developers of other similar bioinformatics systems.

## Bibliography

1. The UniProt, C. (2008). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
3. Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* **18**, 342-8.
4. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**, 108-117.
5. Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* **77**, 100-113.
6. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17.
7. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.
8. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7594-7599.
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.
10. Wu, S. & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547-56.
11. Wu, S. & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* **35**, 3375-82.
12. Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566-79.
13. Wu, S. & Zhang, Y. (2008). ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS ONE* **3**, e3400.
14. Silva, P. J. (2008). Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis. *Proteins* **70**, 1588-94.
15. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53.
16. Battey, J. N., Kopp, J., Bordoli, L., Read, R. J., Clarke, N. D. & Schwede, T. (2007). Automated server predictions in CASP7. *Proteins* **69**, 68-82.
17. Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* **51**, 434-41.
18. Zhou, H. & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**, 321-8.
19. Shi, J., Blundell, T. L. & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* **310**, 243-57.

20. Xu, Y. & Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. *Proteins* **40**, 343-54.
21. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-60.
22. Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. & Hughey, R. (2003). Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl 6**, 491-6.
23. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.
24. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145-1164.
25. Zhang, Y., Kihara, D. & Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48**, 192-201.
26. Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E. & Skolnick, J. (2006). On the origin and completeness of highly likely single domain protein structures *Proc. Natl. Acad. Sci. USA* **103**, 2605-10.
27. Chen, H. & Zhou, H. X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* **33**, 3193-9.
28. Wu, S. & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* **24**, 924-31.
29. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* **85**, 1145-64.
30. Zhang, Y. & Skolnick, J. (2004). Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys J* **87**, 2647-55.
31. Zhang, Y. & Skolnick, J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25**, 865-71.
32. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
33. 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* **76**, 665-76.
34. Zhang, J., Liang, Y. & Zhang, Y. (2011). Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling. *Structure* **in press**.
35. Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J Comput Phys* **117**, 1-19.
36. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. J., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.
37. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A* **101**, 7594-9.
38. Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40.
39. Li, W., Zhang, Y. & Skolnick, J. (2004). Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys J* **87**, 1241-8.
40. Zhang, Y. (2009). Protein structure prediction: when is it useful? *Curr Opin Struct Biol* **19**, 145-55.
41. Ekins, S., Mestres, J. & Testa, B. (2007). In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol* **152**, 21-37.
42. Becker, O. M., Dhanoa, D. S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., Nudelman, R., Kauffman, M. & Noiman, S. (2006). An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J Med Chem* **49**, 3116-35.
43. Brylinski, M. & Skolnick, J. (2008). Q-Dock: Low-resolution flexible ligand docking with pocket-specific threading restraints. *J Comput Chem* **29**, 1574-88.
44. Arakaki, A. K., Zhang, Y. & Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* **20**, 1087-96.
45. Yue, P. & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *J Mol Biol* **356**, 1263-74.

46. Boyd, A., Ciufo, L. F., Barclay, J. W., Graham, M. E., Haynes, L. P., Doherty, M. K., Riesen, M., Burgoyne, R. D. & Morgan, A. (2008). A random mutagenesis approach to isolate dominant-negative yeast sec1 mutants reveals a functional role for domain 3a in yeast and mammalian Sec1/Munc18 proteins. *Genetics* **180**, 165-78.

47. Ye, Y., Li, Z. & Godzik, A. (2006). Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput*, 439-50.

48. Keedy, D. A., Williams, C. J., Headd, J. J., Arendall, W. B., 3rd, Chen, V. B., Kapral, G. J., Gillespie, R. A., Block, J. N., Zemla, A., Richardson, D. C. & Richardson, J. S. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins* **77 Suppl 9**, 29-49.

49. Tress, M., Ezkurdia, I., Grana, O., Lopez, G. & Valencia, A. (2005). Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* **61 Suppl 7**, 27-45.

50. Moult, J. (2008). Comparative modeling in structural genomics. *Structure* **16**, 14-6.

51. Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Seo, J. H., Lee, J., Baker, D., Chivian, D., Kim, D. & Ezkurdia, I. (2007). Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins* **69 Suppl 8**, 137-51.

52. Malmstrom, L., Riffle, M., Strauss, C. E., Chivian, D., Davis, T. N., Bonneau, R. & Baker, D. (2007). Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol* **5**, e76.

53. Zhang, Y., Devries, M. E. & Skolnick, J. (2006). Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput Biol* **2**, e13.

54. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846-856.

55. McGuffin, L. J. & Jones, D. T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**, 874-81.

56. Wallner, B. & Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* **21**, 4248-54.

57. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein server. *Nucleic Acids Res* **32**, W321-6.

58. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-8.

59. Kim, D. E., Chivian, D. & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526-31.

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

61. Kopp, J., Bordoli, L., Battey, J. N., Kiefer, F. & Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69**, 38-56.

# Chapter 3

## COFACTOR: predicting protein-ligand binding sites by global structural alignment and local geometry refinement

Proteins bind with other molecules to bolster or inhibit biological functions. The binding partner, commonly referred to as *ligand*, can be metal ions, small organic/inorganic molecules or macromolecules like proteins or nucleic acids. In all these protein-ligand interactions, only a few key residues are involved in the partner recognitions and for the affinity that tethers the ligand to its receptor molecule. Identification of these key residues is imperative for understanding protein's function, analyzing molecular interactions and guiding further experimental procedures [1]. Although experimental techniques/methods provide the most accurate assignment of the binding locations, the procedure can be time and labour-intensive.

Computational approaches to recognize these functional sites in proteins are generally classified into sequence- and structure-based methods. Most of the sequence-based approaches [2; 3; 4; 5] are based on the presumption that functionally important residues are preferentially conserved during the evolution, because natural selection acts on function. In many cases however, the sequence or evolutionary conservation of residues does not necessarily translate into their involvement in ligand binding, as these residues may play a structural role in maintaining the global scaffold. Nevertheless, the advantage of sequence-based methods is that 3D structure is not a prerequisite and they require negligible time to generate predictions.

Structure-based methods for ligand binding-site identification start with the 3D structure of protein molecules. Most of the early approaches followed the Emil Fisher's assumption that ligand binding in proteins is like "an insertion of key into a lock" [6]; hence shape and physiochemical complementarity are often used to detect concave pockets on proteins surface [7; 8; 9; 10; 11; 12; 13]. There are other methods that use calculated interaction energies [14; 15; 16] or protein structure dynamics [17; 18] to examine the click of "lock and key". With the recent increase in number of known protein-ligand complexes in Protein Data Bank [19], it is becoming evident that homologous proteins with similar global topology often bind similar ligands using a conserved set of residues [20]. Accordingly, many contemporary methods utilize both geometric match and evolutionary information to identify binding site pockets and residues. Some of them use known protein-ligand complexes as templates [21; 22; 23; 24; 25; 26], while others utilize purely sequence-based homology information [8; 11; 27].

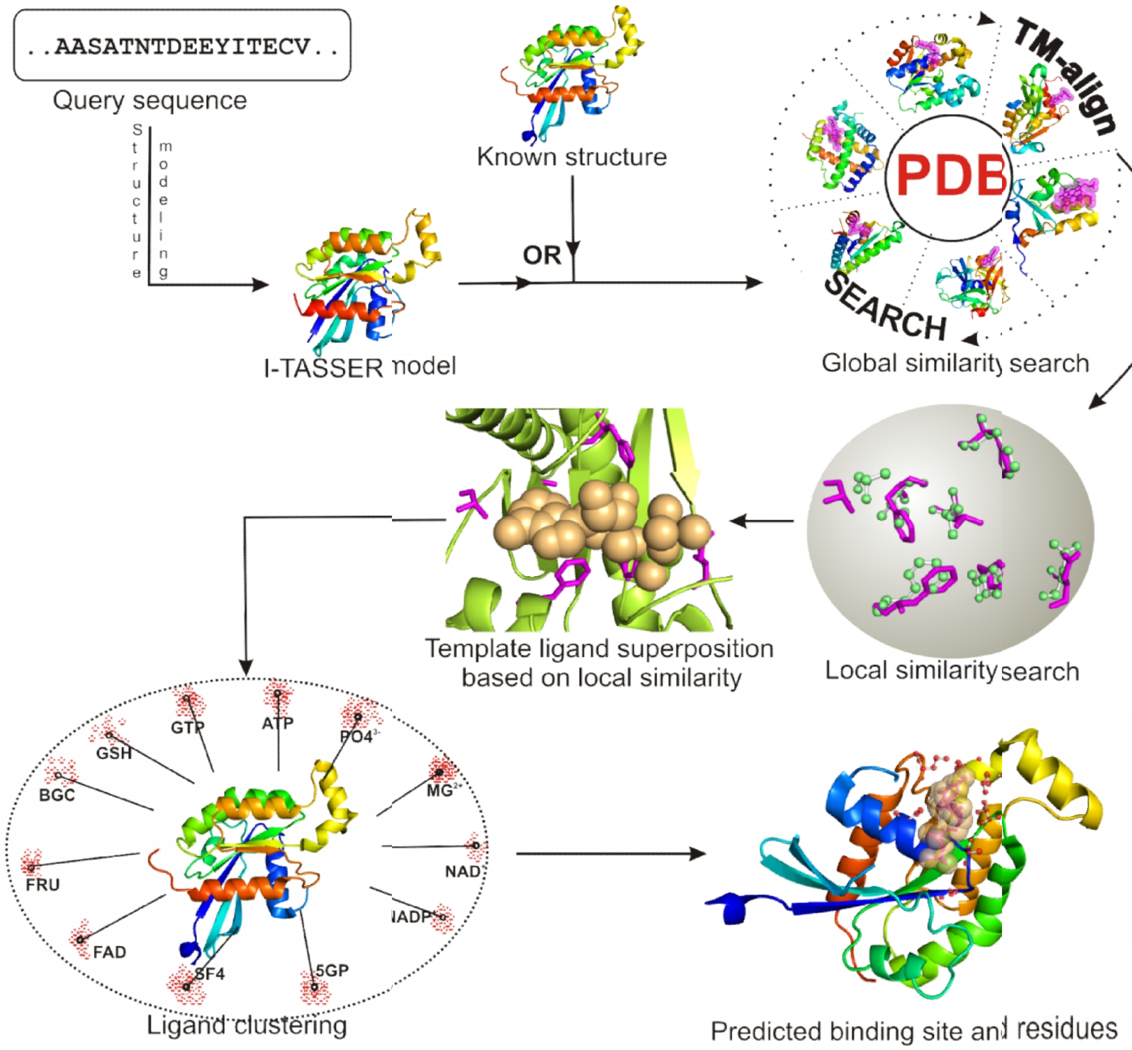


**Figure 3.1** Schematic overview of COFACTOR algorithm for ligand binding site prediction.

Following the sequence-to-structure-to-function paradigm, in this chapter we develop a new approach named COFACTOR, in which 3D and a combined global-and-local similarity search scheme is utilized to identify binding pockets and ligand-interacting residues in query protein. Figure 3.1 shows a schematic diagram describing the procedure of COFACTOR algorithm. Starting from the query sequence, the 3D structure model is first generated using the I-TASSER fragment assembly simulations [28; 29]. Experimental structure can also be used in the following steps. Template proteins with bound ligands in the PDB library are collected based on their global structural similarity to query protein, using the TM-align structure alignment program [30]. Meanwhile, to examine the ligand-binding details, the binding pockets of templates are scanned through the query model to identify the best local geometric and sequence matches. The binding pose of the ligand in the query structure is predicted based on the local alignment of predicted and template binding site residues. Finally, superposed ligands from multiple templates are clustered to procure the ligand-binding predictions.

The algorithm is evaluated using both the I-TASSER models and the experimental structures of target proteins. Large-scale benchmarking results show that COFACTOR can correctly identify ligand-binding locations and interacting residues in a large fraction of test cases for both natural and drug-like molecules. The algorithm was also tested in the recent community-wide CASP9 experiments, and the results highlight the potential applicability of the method for genome-scale functional annotations.

## 3.1. COFACTOR algorithm

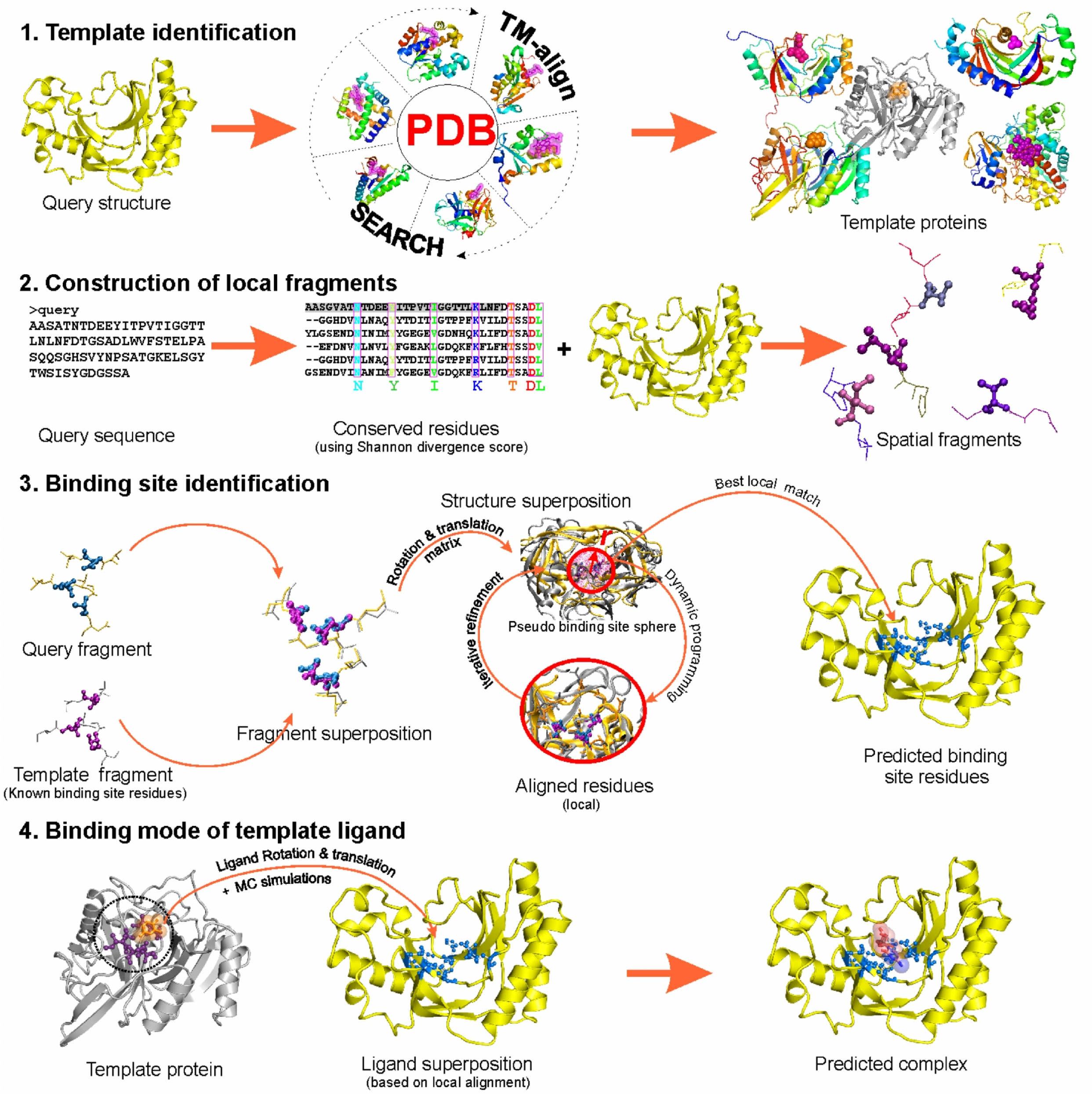

1. Template identification
Query structure
TM-align
PDB
SEARCH
Template proteins
2. Construction of local fragments
>query
AASATNTDEEYITPVTIGGTT
LNLNFDTGSADLWVFSTELPA
SQQSGHSVYNPSATGKELSGY
TWSISYGDGSSA
AASGVATNTDEEYITPVTIGGTTLKLNFDTSADL
--GGHDVNLNAQYYTDITIGTPPFKVILDTSSDL
YLGSENDNINIMYYGEGEVGDNHQKLIFDTSADL
--EFDNVNLNVLYFGEAKLGDQKFKFLFHTSSDV
--GGHDVNLNAQYYTDITLGTPPFRVILDTSSDL
GSENDVINANIMFYGEGEVGDQKFRLIFDTSADL
N Y I K T DL
+
Query sequence
Conserved residues
(using Shannon divergence score)
Spatial fragments
3. Binding site identification
Structure superposition
Best local match
Rotation & translation matrix
r
Dynamic programming
Iterative refinement
Pseudo binding site sphere
Query fragment
Fragment superposition
Template fragment
(Known binding site residues)
Aligned residues
(local)
Predicted binding
site residues
4. Binding mode of template ligand
Ligand Rotation & translation
+ MC simulations
Template protein
Ligand superposition
(based on local alignment)
Predicted complex

**Figure 3.2** Steps of COFACTOR algorithm for protein-ligand binding site predictions. (1) Template proteins from the ligand-binding library are identified using TM-align global similarity search. (2) Conserved residues in query sequence are identified based on Shannon diverge score which are used to glean local 3D-fragments from the query structure. (3) Each local 3D-motif of query is iteratively aligned with known binding site residues fragments from template where the binding pocket similarity between query and template is evaluated using BS-score. (4) The template ligand is transferred onto the query structure, which is refined by a short Monte Carlo (MC) simulation to improve the local geometry.

The COFACTOR algorithm consists of four major steps (see Figure 3.2). First, structural analogs of the query protein are identified by performing a global structure similarity search using TM-align [30], where the structural analogs are ranked based on TM-score [31]. The underlying hypothesis is that proteins with similar structure usually have similar function, and hence they may bind similar ligands. However, this is not always true since many observations have demonstrated that proteins with similar functions can have different global topology. This necessitates local structural comparisons, since similar ligands often have similar binding pockets, which is the goal of the next steps.

In the second step, multiple sequence alignment (MSA) for the query sequence is constructed by PSI-BLAST search through the non-redundant (NR) sequence database. Conserved residues in query sequence are then identified from the MSA, based on their Jensen–Shannon divergence score [2]. These residues mark potential binding site locations in query structure. The structures of all combined sets of these marked residues will be used as candidate binding site motifs.

In the third step, for any given template ($t$) with known binding site ($b$), residue triplets ($l_{tb}$, $m_{tb}$, $n_{tb}$) are selected from binding site residues (Figure 3.3). Similarly, conserved residues triplets ($a$, $b$, $c$) are selected from query as candidate binding site motif (Figure 3.3B). The structure of these candidate sites ($a$, $b$, $c$) is superposed on the known binding site residues ($l_{tb}$, $m_{tb}$, $n_{tb}$). As a pre-filter, we discard any candidate binding site motif for which a pair-wise residue distance ($d_{ab}$, $d_{bc}$ or $d_{ac}$) > $2r$, where $r$ is the maximum distance of any template binding site residues from the geometric center ($C_{tb}$) of template binding site. Furthermore, to increase the reliability of the structure superimposition, for each residue $i$, the coordinates of $C_\alpha$ atom and side-chain center of mass of two neighboring residues, i.e. the *i*-1th and *i*+1th residues, in both template and query are also included in the superposition.

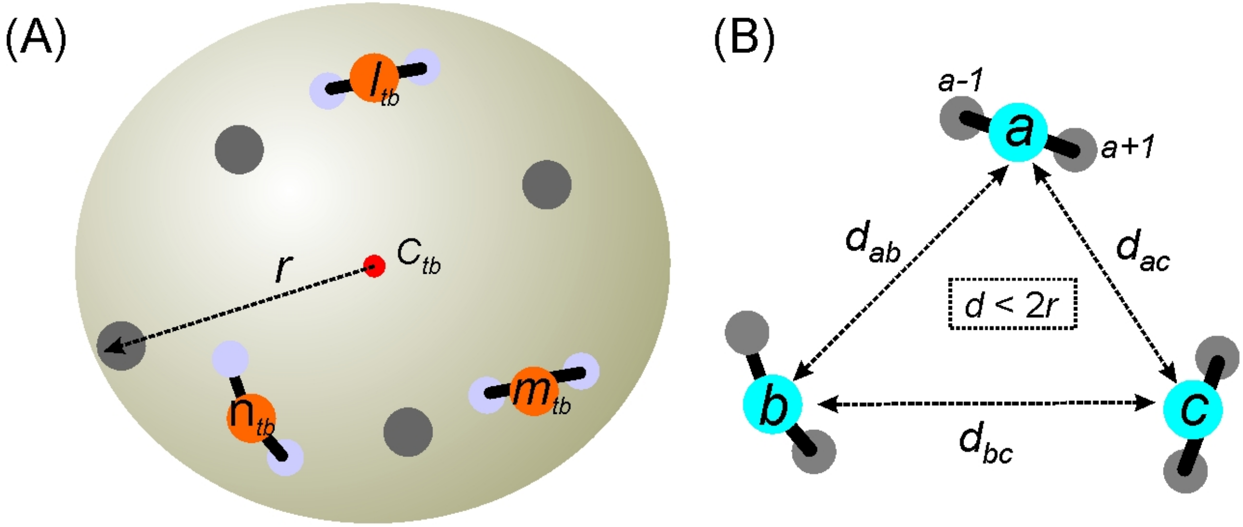


**Figure 3.3** Schematic diagram showing putative binding residues in template and query sequence. (A) template binding site (tb) defined using a sphere of radius r from geometric centroid of binding site. The selected binding site residue triplets (l, m, n) are highlighted in orange. (B) conserved residues triplets (a, b, c) of query protein with inter-residue distance (d) < 2r. In both query and template, for any residue i, two flanking residues i-1 and i+1 are also selected.

To account for similar local environment in query and template, a requirement for accommodating similar ligand molecules, the query structure is superposed onto the entire structure of the template based on the rotation matrix acquired from the superposition of the candidate binding site motifs and template residues. A sphere of radius *r* is then defined around the geometric center ($C_{tb}$). The sphere here represents a probable binding pocket, inside which the sequence and structural similarity of query and template are compared. Because a sphere comprising of very small number of residues can easily generate false positive hits, when the defined binding site region on the template is small (i.e. the number of residues within the sphere is less than 15), *r* is gradually incremented by 0.5 Å, until the number of residues inside the sphere is larger than 15.

A heuristic procedure, similar to that used in TM-align [30], is then used to refine the local match between the query and template structures, inside the sphere. Starting from the initial superposition of query and template protein structures based on the candidate motif, a Needleman-Wunsch dynamic programming [32] is performed to generate a new alignment within the selected sphere areas of query and template, where the alignment score $S_{ij}$ for aligning $i$th residue in query and $j$th residue in template is given by.

$$S_{ij} = \left[ \frac{1}{1 + \left( \frac{d_{ij}}{d_0} \right)^2} + M_{ij} \right] \tag{3.1}$$

Here, $d_{ij}$ is the $C_\alpha$ distance between $i$th residue in the query and $j$th residue in the template, $d_0$ is the distance scale chosen to be 3.0 Å, $M_{ij}$ is the substitution scores between the $i$th and $j$th residues taken from the BLOSUM62 mutation matrix. The element value in the BLOSUM62 matrix was normalized in between [0, 1], in order to keep both the distance and mutation scores in Eq. 3.1 in the same scale. Gap penalty is empirically set as -1. Based on the initial seed alignment, the areas within the spheres are re-superimposed and a new scoring matrix $S_{ij}$ is then constructed, which will result in a newer alignment from dynamic programming. This procedure is repeated until the final alignment is converged. For each alignment, a raw alignment score is defined for evaluating the binding site similarity (BS-score):

$$\text{BS} - \text{score} = \frac{1}{N_t} \sum_{i=1}^{N_{ali}} \frac{1}{1 + \left( \frac{d_{ii}}{d_0} \right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{ali}} M_{ii} \tag{3.2}$$

where $N_t$ represents the number of residues within the binding site sphere of the template, $N_{ali}$ is the number of aligned residue pairs. The procedure is repeated for all possible candidate binding site motifs (*a*, *b*, *c*) and known binding site residues triplets ($l_{tb}$, $m_{tb}$, $n_{tb}$) in this template binding site. Finally, BS-score, that determines the best local match between query and the known template binding site, is obtained:

$$\mathrm{BS-score}_{\max} = \max_{\forall abc \rightarrow lmn} \left[ \frac{1}{N_t} \sum_{i=1}^{N_{ali}} \frac{1}{1+\left(\dfrac{d_{ii}}{d_0}\right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{ali}} M_{ii} \right] \tag{3.3}$$

This marks a binding site prediction from a known binding pocket of one template that was scanned for all conserved motifs in the query.

The ligand pose will be copied from the template structure by superposing the structure of the known binding site residues in the template onto the predicted binding site residues in the query. To remove the overlap between the ligand and the query protein structures, a quick Metropolis Monte Carlo simulation is conducted to improve the local geometry of ligand-binding, in which the energy is defined as the sum of the number of contacts made by ligand with predicted binding site residues, the reciprocal of the number of ligand-protein clashes, and the contact distance error which is calculated as difference between inter-atomic ligand-protein contact distance in template and that in query model.

The last step of the procedure is to rank the predicted binding sites based on multiple templates. To do so, all the locally superposed ligands on the query structure are clustered based on their spatial proximity (distance cutoff = 8Å). The binding pockets with larger cluster size are supposed to have higher chance to be correct. As each binding pocket can bind multiple ligands (for example, an ATP binding pocket in enzymes can also bind MG, $PO4^{3-}$ and ADP), ligands within the same pocket are clustered again based on their chemical similarity (Tanimoto coefficient cutoff = 0.7) using the average linkage clustering procedure. From each ligand-specific cluster, a confidence score is defined as:

$$\mathrm{FC-score_{LB}} = \frac{2}{1+e^{-\left(\frac{\mathrm{N}}{\mathrm{N_{tot}}}\times(x-0.7)\right)}} - 1, \tag{3.4}$$

where N is the multiplicity of ligand decoys in the cluster and $\mathrm{N_{tot}}$ is the total number of predicted ligands using the templates. $x$ is an experience function to combine local and global structure similarities, and the evolutionary relation between target and template proteins:

$$x = (\mathrm{BS-score}/4) + \mathrm{TM-score} + 2.5\mathrm{ID_{Str}} + \frac{2}{1+\langle D\rangle} \tag{3.5}$$

The BS-score and TM-score measure local and global similarity of the query to the template. $\mathrm{ID_{str}}$ is sequence identity between the query and the template in the structurally aligned region. And <D> is the average distance of the predicted ligand to all other predicted ligands in the same cluster.

The FC-score definition of Eq. 3.5 thus represents a combination of the cluster size and structural and sequence similarities of target and template proteins. The parameters have been chosen to keep $\mathrm{FC-score_{LB}}$ in the range of [0, 1]. The ligand binding prediction with the highest $\mathrm{FC-score_{LB}}$ is finally selected.

## 3.2. COFACTOR binding site library

Constructing a binding site library containing biologically relevant ligands is not a trivial problem due to the large number of crystallization artifacts in the existing structures in PDB [33]. Utilization of homology information provides some respite to this problem, since in most cases homologous proteins bind ligands near similar locations [34].

To construct a comprehensive library with biologically relevant ligands, all protein chains with ligand interacting residues were first screened through the PDB library. Commonly used crystallization buffers, non-biological ions and heavy metal were pre-filtered. Protein sequence was extracted from the co-ordinates file of filtered complexes, while translating modified amino acids to their parent amino acid. Thereupon, sequences were clustered using CD-HIT [35] at 40% sequence identity cutoff, with the purpose of grouping them into homologous families. For orphan protein chains that formed single entry cluster, we tried to identify its homologous cluster by first performing a PSI-BLAST [36] search against already clustered proteins, else proteins with similar structure were identified using TM-align [30] search (TM-score >0.7).

The longest protein in each cluster was selected as the cluster representative, and all the cluster members were structurally superposed on the cluster representative using TM-align. Pair-wise distance between center of mass of ligands in superposed complexes was calculated. To judge whether a ligand is biologically relevant or not, we implemented the following filtering criteria: (a) the ligand should either have at least one ligand present in a superposed homologous structure (sequence identity <90%) within 5Å; (b) if the ligand is metal or inorganic ions, it should have at least 3 binding site residues; (c) if it is a non-metal ligand, 5 or more binding site residues is a prerequisite. Complexes that satisfied any of these three criteria were re-clustered based on ligand type and redundant binding site were removed by comparing binding site residues at 90% sequence identity cutoff. We also consulted Binding MOAD database [37], which contains both drug and natural ligands, to check for ligands that may have been missed during this automated procedure.

At present, the binding site library contains 45,381 entries, containing 13,763 metal ligands, 1,417 biopolymers and 30,201 monomeric ligands that include both drug-like and natural ligands. The library is freely available at

http://zhanglab.ccmb.med.umich.edu/COFACTOR/library.

## 3.3. Evaluation of COFACTOR algorithm

Benchmarking proteins were collected from ligAsite benchmark set (v9.1) [38], which contains 364 protein chains bound to small molecule ligands, including 63 “drug-like” and 382 “natural” ligands. To increase the sample size of drug-like compounds, we further added 137 proteins bound with drug-like molecules from references [39] and [40]. Metal ions were filtered out from this analysis, as the control methods (FINDSITE and ConCavity) could not predict binding sites for metal ions. We also excluded ligands bound at the interface of protein chains, since the current I-TASSER protein structure modeling could be performed only for single protein chains and both COFACTOR and FINDSITE do not incorporate oligomeric state of the protein.

The results are controlled by two recently developed structure-based methods, FINDSITE [22], and ConCavity [27]. FINDSITE predicts binding sites by matching the target structure with template proteins identified by threading [22], while ConCavity assigns binding residues as those closest to the spatial cavities surrounding the protein surface [27].

### 3.3.1. ConCavity and FINDSITE as experimental controls

Two recently developed structure-based methods, FINDSITE [22] and ConCavity [27], are used as controls in our benchmarking experiments.

ConCavity [27] was designed to identify solvent-accessible pockets formed by surface residues. The identified pockets are ranked based on sequence conservation of the residues associated with the pocket. Residues in the predicted pocket are smudged using a Gaussian filter to identify potential ligand interacting residues. ConCavity program was used to detect ligand binding sites in I-TASSER models and experimentally determined structures, using default parameters and by providing evolutionary sequence conservation information, estimated based on Jensen–Shannon divergence (JSD) score [2] of residues. JSD scores for each residue were computed using multiple sequence alignment of query protein with identified homologues in NR sequence database using PSI-BLAST [36]. A predicted pocket by ConCavity is represented by a set of 3D grid points; hence we used geometric center of predicted grid points as the location of predicted binding pocket.

FINDSITE [34] is a template-based method that first uses PROSPECTOR [41] to identify the threading template proteins in the PDB library. Then homologous template proteins of the identified threading templates are collected from the FINDSITE binding site library and superposed on the query structure (I-TASSER models or experimental solutions) using Fr-TM-align [42]. FINDSITE predicts binding pocket as a single point, calculated as the center of mass of all the threading template ligands superposed on query structure. Binding site residues are also predicted based on concurrence of residues that make contact with ligands in the cluster.

### 3.3.2. Evaluation metrics

The performance of protein-ligand binding predictions can be evaluated based on their ability to detect the spatial location of ligand binding pocket, competency to delineate protein residues that interact with the ligand or the shape of the predicted binding pocket. Here, we evaluate COFACTOR on all these criteria as well as the chemical similarity between predicted and native ligand.

The binding pocket predictions are evaluated by calculating the distance between the center of mass of the bound ligand in the experimental structure and the center of the predicted binding pocket in the query. We used 4.5Å as a cutoff to evaluate correct binding pocket predictions, which is close to the average radius of gyration of the 582 experimental ligands in the benchmark set.

The ligand-binding residue predictions is evaluated mainly by the Matthews Correlation Coefficient (MCC) between predicted and experimental binding residues:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{3.6}$$

where *TP*, *TN*, *FP* and *FP* are abbreviations for true positive, true negative, false positive, and false negative binding residue predictions. MCC ranges between 1 and −1, where a MCC of 1 indicates a prefect prediction, 0 a random prediction, and -1 an inverse prediction. We also define the accuracy of the binding site prediction as

$$\mathrm{Acc} = TP/(TP+FP) \tag{3.7}$$

which measures the ratio of the correctly predicted binding residues over the total number of predicted residues. The coverage is defined as

$$Cov = TP/(TP+FN) \tag{3.8}$$

which measures the portion of the correctly predicted binding residues over the total number of binding residues in the experimental structure. Here, true binding site residues are defined as those that have any heavy atom within a distance of 0.5Å plus the sum of the van der Waals radius of protein atom and ligand atoms in the experimental structure.

The shape similarity between the predicted ligand and the bound ligand in experimental structure is evaluated based on the volume overlap, measured as Jaccard Coefficient (JC):

$$\mathrm{JC} = \left|\frac{\text{predicted vol.} \cap \text{native vol.}}{\text{predicted vol.} \cup \text{native vol.}}\right| \tag{3.9}$$

For FINDSITE, we selected the ligand from a template amongst the clustered templates, that had highest sequence identity to query protein and best-predicted pocket; while for ConCavity, predicted pocket grid points were presumed as H-atom of ligand. Volume calculation is done on 3-D grid of 1Å spacing.

Chemical similarity between two compounds (*A* and *B*) is evaluated using the Tanimoto coefficient:

$$TC = \frac{N_A + N_B - N_{AB}}{N_{AB}} \qquad (3.10)$$

where $N_A$ and $N_B$ are the number of chemical and structural features that are present in each ligand and $N_{AB}$ is the number of common features between *A* and *B*. Features of all the ligands were defined using Open Babel package [43].

## 3.4. Benchmarking results of binding site predictions

The performance of protein-ligand binding predictions can be evaluated based on their ability to detect the spatial location of ligand binding pocket and the competency to delineate protein residues that interact with the ligand. In the first evaluation, the prediction errors are evaluated by measuring the spatial distance between the center of the predicted binding pocket and the ligand in experimental structure. In the second evaluation, the assignment accuracy of ligand-interacting residues in the protein sequence is gauged. Here, we evaluate COFACTOR on both criteria. Two recently developed structure-based methods, FINDSITE [22], and ConCavity [27] were used as controls for the result. FINDSITE predicts binding sites by matching the target structure with template proteins identified by threading [22], while ConCavity assigns binding residues as those closest to the spatial cavities surrounding the protein surface [27].

### 3.4.1. Ligand-binding pocket predictions

The ability of the algorithms to identify ligand-binding pocket is tested on 501 benchmarking proteins, collected from three previous experiments [38; 39; 40] that harbor 582 ligands. The experimental structures of the protein-ligand complexes were collected from the PDB library [33].

Figure 3.4 shows the cumulative fraction of predicted binding pockets as a function of distance between the center of mass of the native ligand and the center of the predicted binding pocket. If we make a cutoff at the pocket distance of < 4.5Å, which is close to the average radius of gyration of all ligands in the benchmark set (4.41 Å), the binding pocket predictions by COFACTOR are correct in 67% cases, when the low-resolution I-TASSER structure models were used. The control methods FINDSITE and ConCavity correctly predicted binding pocket for 60% and 39% cases, respectively. These differences are statistically significant, where the p-value of paired student t-test for the COFACTOR prediction is $9.69e^{-7}$ to FINDSITE and $1.62e^{-12}$ to ConCavity results.

Compared to ConCavity, both COFACTOR and FINDSITE are not very sensitive to the accuracy of the protein structure predictions, as long as the global topology of the target model is correct. When the apo-form experimental structures of the target proteins were used, the accuracy of the binding pocket predictions by COFACTOR and FINDSITE increased only marginally to 74% and 61%, respectively, whereas that of ConCavity changed significantly changed 39% to 53%. This difference in structural sensitivity is probably due to the fact that the cavity-based methods such at ConCavity are sensitive to the local geometry of the target structures while the template-based methods rely more on the global similarity of the target-template topologies. Although homologous templates have been excluded from the I-TASSER template library, majority of I-TASSER models (91%) have a correct topology with TM-score >0.5, which explains the independence of the average performance of COFACTOR and FINDSITE on the models chosen of the target structures.

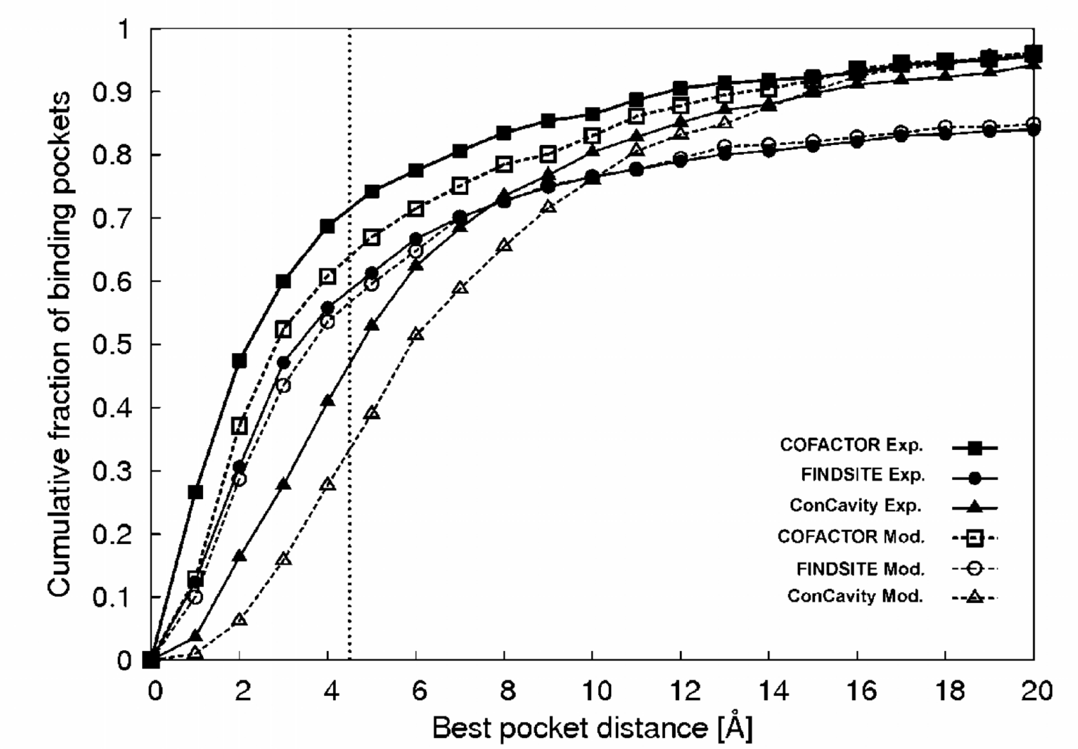


**Figure 3.4** Comparison of different methods in identifying ligand binding pocket using either I-TASSER models or experimental structures. Results are presented as the cumulative fraction of predicted binding site pockets versus distance between the center of the native ligand position and the center of the best in top five predicted ligand-binding poses.

We observed that in 9% of the cases FINDSITE didn't generate any pocket predictions due to lack of good threading templates in its binding-site library. As a result, ConCavity shows an improved performance over FINDSITE in difficult cases, i.e. ConCavity outperforms FINDSITE in cumulative fraction of binding pocket when the pocket distance increases. If we consider only the 446 proteins (with 514 binding sites) where all the three methods successfully generated a prediction, the average binding-pocket distance of the best in top-five predictions by COFACTOR, FINDSITE and ConCavity using I-TASSER models are 3.9Å, 5.0Å and 6.5Å, respectively. When the experimental structures are used, the average distance errors are reduced to 4.1Å, 5.0Å and 7.0Å, respectively. This data shows that for both easy and hard targets the binding pockets identified by COFACTOR are on average closer to the actual binding pocket.

### 3.4.2. Ligand binding-site residue assignments

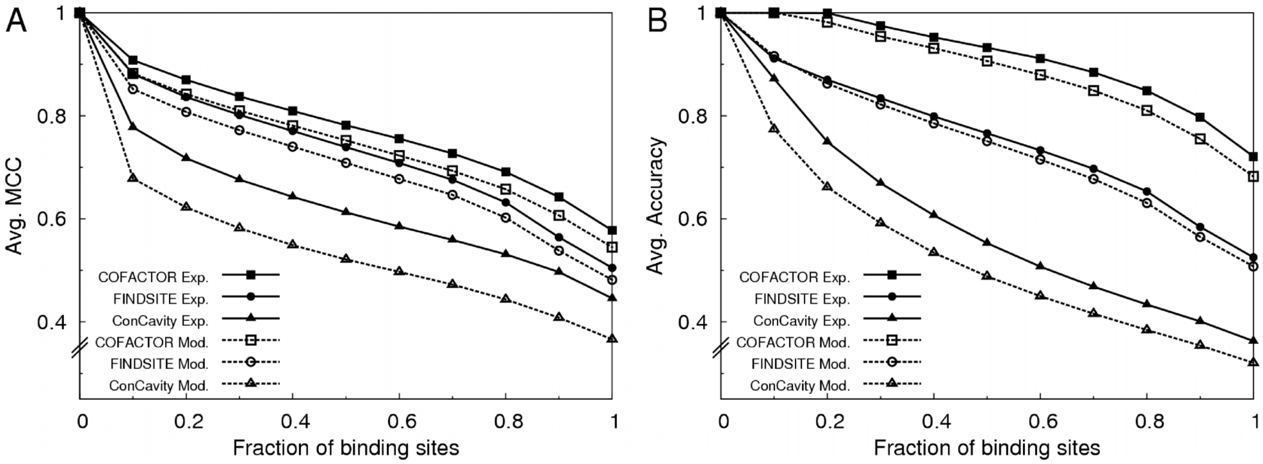


**Figure 3.5** Performance of different methods in detecting ligand interacting residues. (A) Average Matthews's correlation coefficient (MCC); (B) average accuracy of predicted binding site residues.

To evaluate the ability of COFACTOR to detect the binding site residues, in Figure 3.5 we plotted the average Matthews correlation coefficient (MCC) and accuracy of the predicted binding residues as a function of the coverage of the predicted binding residues under consideration, where MCC and binding accuracy were defined in Eqs. 3.6 and 3.7, respectively.

When using the I-TASSER predicted models, COFACTOR could identify binding-site residues for 90% of the targets with an average MCC of 0.61. The average MCC for all targets was 0.54. The average accuracy of the binding residue prediction was 76% (68%) for 90% (all) targets. Compared to the control methods (FINDSITE and ConCavity), COFACTOR showed an overall improvement of 13-46% on MCC (Figure 3.5A), and 33-112% improvement on the prediction accuracy (Figure 3.5 B). The reason for the obviously low accuracy and MCC for ConCavity is that the algorithms defines all the conserved residues lining with the predicted pockets as potential ligand interacting residue, which although improves the prediction coverage (Table 3.1), considerably increases the rate of false positive prediction and results in the low MCC and accuracy. When using experimental structure, the MCC and accuracy of the binding site residues by COFACTOR slightly improved to 64% (58%) and 80% (72%), for 90% (all) targets (Figure 3.5).

**Table 3.1 Average Matthews's correlation coefficient (MCC), accuracy (Acc) and coverage (Cov) of ligand-binding residue predictions by ConCavity, FINDSITE and COFACTOR, using I-TASSER models and experimental apo structures as receptor structure.**

| Protein structure | Ligands | Methods | First prediction | | | Best in top 5 | | |
|---|---|---|---|---|---|---|---|---|
| | | | MCC | Acc | Cov | MCC | Acc | Cov |
| I-TASSER models | Natural (382 ligands) | ConCavity | 0.34 | 0.30 | 0.58 | 0.37 | 0.34 | 0.62 |
| | | FINDSITE | 0.43 | 0.47 | 0.45 | 0.50 | 0.55 | 0.53 |
| | | **COFACTOR** | **0.46** | **0.60** | **0.42** | **0.55** | **0.70** | **0.48** |
| | Drug-like (200 ligands) | ConCavity | 0.33 | 0.27 | 0.56 | 0.36 | 0.30 | 0.58 |
| | | FINDSITE | 0.39 | 0.38 | 0.44 | 0.44 | 0.43 | 0.49 |
| | | **COFACTOR** | **0.45** | **0.53** | **0.41** | **0.54** | **0.63** | **049** |
| | Overall (582 ligands) | ConCavity | 0.34 | 0.29 | 0.57 | 0.37 | 0.32 | 0.60 |
| | | FINDSITE | 0.41 | 0.44 | 0.45 | 0.48 | 0.51 | 0.51 |
| | | **COFACTOR** | **0.46** | **0.57** | **0.42** | **0.54** | **0.68** | **0.48** |
| Experimental structures | Natural (382 ligands) | ConCavity | 0.42 | 0.35 | 0.69 | 0.45 | 0.38 | 0.73 |
| | | FINDSITE | 0.44 | 0.48 | 0.47 | 0.52 | 0.56 | 0.55 |
| | | **COFACTOR** | **0.49** | **0.64** | **0.45** | **0.58** | **0.73** | **0.48** |
| | Drug-like (200 ligands) | ConCavity | 0.41 | 0.31 | 0.69 | 0.44 | 0.34 | 0.73 |
| | | FINDSITE | 0.42 | 0.40 | 0.47 | 0.47 | 0.45 | 0.52 |
| | | **COFACTOR** | **0.49** | **0.59** | **0.44** | **0.58** | **0.66** | **0.52** |
| | Overall (582 ligands) | ConCavity | 0.41 | 0.34 | 0.69 | 0.45 | 0.36 | 0.73 |
| | | FINDSITE | 0.43 | 0.45 | 0.47 | 0.51 | 0.53 | 0.54 |
| | | **COFACTOR** | **0.49** | **0.62** | **0.44** | **0.58** | **0.72** | **0.51** |

## 3.4.3. Drug-like versus natural ligands

If we define bio-molecules binding to enzyme active and allosteric sites as "natural" ligands and artificially designed molecules as "drug-like" ones, 382 out of 582 ligands are classified as natural ligands, while the remaining 200 are drug-like in our benchmark set. Based on the results shown in Figure 3.6A, we find that there is little difference in the average MCC of predicted binding site residues for the different ligand types. The difference becomes notable for prediction accuracy (Table 3.1), where ligand interacting residues for natural ligands were predicted with accuracy 6-7% higher than for drug-like compounds.

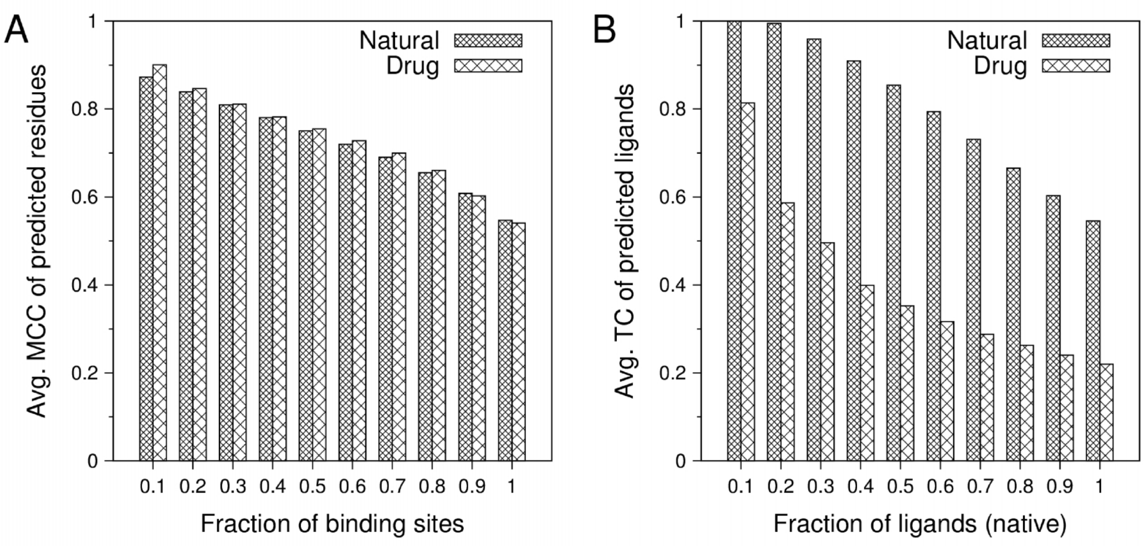


**Figure 3.6** COFACTOR ligand-binding predictions for natural ligands and drug-like compounds. (A) Matthews's correlation coefficient MCC in identifying ligand interacting residues as a function of the fraction of binding sites. (B) Chemical similarity between the native bound ligands and the predicted ligands as assessed by Tanimoto coefficient (TC). For both analyses, I-TASSER models are used as the apo receptor structure.

In Figure 3.6B, we further analyzed the chemical similarity between the predicted ligands by COFACTOR and the native ligands in experimental structure, measured by the Tanimoto coefficient (TC). It is worth noting, that for nearly 73% of the proteins with bound "natural" ligands, the predicted ligands by COFACTOR shared a high chemical similarity (TC > 0.75), and therefore can be used for a more detailed level elucidation of protein function. For targets with bound drug-like molecules, even though the predicted residues had an overall high average MCC (54%), close to that of the natural counterpart, the predicted and solved ligands were chemically similar in only 22% cases. This observation recapitulates the fact that the majority of these drugs are targeted near the active/allosteric sites, where even though they are chemically dissimilar to the substrate molecules, they are tethered by similar set of binding residues. These high accuracy predicted binding site residues by COFACTOR therefore can also be used for creating binding-site based 3D-pharmacophore models for ligand-screening and structure-based drug design even for proteins with unknown structure.

### 3.4.4. Ligand shape comparison

In Table 3.2, we compare the shape of the predicted binding pocket/ligand with that of the native ligands (average volume 743 Å$^{3}$) bound in the experimental structure, as an assessment of predicted ligand conformation. Predicted ligands by COFACTOR, FINDSITE and ConCavity using the I-TASSER model (experimental structure) have an average Jaccard Coefficient (JC) of 0.34 (0.42), 0.27 (0.29) and 0.19 (0.24) respectively, while the average volume of ligand/pocket predicted by the three methods are 945 (893), 964 (962) and 2208 (2307) respectively. The result demonstrates that although the volume of predicted ligands by COFACTOR are on average smaller, the best match for the shape of the predicted ligands match the best with the native ligands, which is important for shape similarity based studies such as docking and ligand screening [44]. Moreover, the average number of protein-ligand clashes is generally fewer in complexes generated by COFACTOR (Table 3.2).

**Table 3.2 Comparison between predicted and bound ligands.**

| Protein structure | Methods | Exp. ligand volume (Å$^{3}$) | Prediction volume (Å$^{3}$) | Jaccard Coefficient | Average # of clashes |
|---|---|---|---|---|---|
| I-TASSER models | ConCavity | 743 | 2208 | 0.19 | 282 |
| | FINDSITE | | 964 | 0.27 | 63 |
| | **COFACTOR** | | **945** | **0.34** | **32** |
| Experimental structures | ConCavity | 743 | 2307 | 0.24 | 287 |
| | FINDSITE | | 962 | 0.29 | 49 |
| | **COFACTOR** | | **893** | **0.42** | **20** |

## 3.5. Blind test of COFACTOR in CASP9

The ninth community-wide critical assessment of techniques for protein structure prediction (CASP9) released 129 query protein sequences for blind test of protein structure and function prediction methods. The function prediction section was focused on evaluating the ligand binding-site predictions, where the predictors were asked to identify ligand-interacting residues in the provided protein sequence.

**Table 3.3 Binding site predictions by COFACTOR for 31 CASP9 targets.**

| Target | TM-score** | Native ligand(s) | Predicted ligand(s) | C-score$_{LB}$ | MCC | Acc | Cov |
|---|---|---|---|---|---|---|---|
| T0515* | 0.89 | PLP, LYS | ORX, PLP | 0.61, 0.45 | 0.68 | 0.64 | 0.75 |
| T0516 | 0.89 | PF1 | PF1, HMH | 0.79, 0.88 | 0.84 | 0.85 | 0.85 |
| T0518 | 0.80 | NA | CA, MN | 0.41, 0.39 | 0.38 | 0.38 | 0.43 |
| T0521 | 0.52 | 2 CA | 4 CA | 0.67, 0.76, 0.66, 0.60 | 0.08 | 0.10 | 0.22 |
| T0524* | 0.87 | GAL | GAL | 0.75 | 0.66 | 0.73 | 0.62 |
| T0526* | 0.88 | GLA | GAL | 0.55 | 0.46 | 0.42 | 0.56 |
| T0529 | 0.23 | MN | ZN, AMP | 0.72, 0.23 | 0.55 | 0.31 | 1.00 |
| T0533 | 0.79 | PHE | 2 PHE | 0.88, 0.09 | 0.88 | 1.00 | 0.79 |
| T0539 | 0.64 | ZN, ZN | ZN, ZN | 0.85, 0.77 | 1.00 | 1.00 | 1.00 |
| T0547* | 0.71 | PLP, LYS | PLP, LYS, AZ1, ORX, P3T | 0.61, 0.61, 0.61, 0.54,0.54 | 0.77 | 0.74 | 0.82 |
| T0548 | 0.56 | ZN | SAL, ZN | 0.21, 0.67 | 0.69 | 0.50 | 1.00 |
| T0565* | 0.74 | DGL, ALA | DLG, ALA, UNL | 0.88, 0.50, 0.52 | 0.86 | 1.00 | 0.75 |
| T0570 | 0.88 | MG, GOL | CA, GOL, PO4 | 0.83, 0.21, 0.34 | 0.87 | 0.88 | 0.88 |
| T0582 | 0.85 | ZN | ZN | 0.64 | 1.00 | 1.00 | 1.00 |
| T0584* | 0.83 | IPR, DST | IPR, RIS, MG, MG, PO4 | 0.51, 0.25, 0.68, 0.75, 0.58 | 0.75 | 0.63 | 0.92 |
| T0585 | 0.78 | ZN | ZN | 0.85 | 0.77 | 1.00 | 0.60 |
| T0591 | 0.89 | LLP | PLP, PLP | 0.83, 0.81 | 0.76 | 0.65 | 0.91 |
| T0597 | 0.86 | ANP | MG, ATP, AMP | 0.93, 0.83,0.80 | 0.70 | 0.80 | 0.63 |
| T0599* | 0.95 | ISC | MG, ISC | 0.88, 0.83 | 0.83 | 0.75 | 0.92 |
| T0604 | 0.41 | FAD | FAD | 0.72 | 0.45 | 0.54 | 0.42 |
| T0607* | 0.86 | ZN, ZN, BES | MN, MN, BIB | 0.93, 0.83, 0.68 | 0.50 | 0.71 | 0.36 |
| T0609* | 0.78 | GAL | GAL | 0.74 | 0.82 | 0.75 | 0.90 |
| T0613* | 0.96 | GAR, NHS | UNL, THH | 0.48, 0.58 | 0.70 | 0.77 | 0.67 |
| T0615* | 0.71 | MN, GPX | MN, PO4 | 0.83, 0.77 | 0.50 | 0.83 | 0.33 |
| T0622* | 0.69 | NAD | NAD,ATP | 0.66, 0.71 | 0.76 | 0.67 | 0.93 |
| T0625 | 0.74 | ZN | ZN | 0.79 | 1.00 | 1.00 | 1.00 |
| T0629 | 0.34 | 6 FE | ZN | 0.45 | 0.37 | 1.00 | 0.14 |
| T0632 | 0.74 | COA | COA, PHB | 0.68,0.60 | 0.46 | 0.67 | 0.38 |
| T0635 | 0.91 | CA | MG, PO4 | 0.96, 0.90 | 0.60 | 0.38 | 1.00 |
| T0636* | 0.93 | HAS, PLP | HAS, PMP, PMP | 0.51, 0.32, 0.51 | 0.79 | 0.78 | 0.82 |
| T0641 | 0.91 | STE | PLM | 0.82 | 0.83 | 0.80 | 0.89 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Average** | | | | | **0.69** | **0.72** | **0.72** |

*Holo structure of these proteins was solved with non-native ligand, however CASP9 assessors inferred the native ligand binding information from homologous PDB structures.
**TM-score of I-TASSER models for the target protein.

During CASP9, we first generated the 3D structural models using I-TASSER and the structure-based ligand binding site predictions were generated using the COFACTOR algorithm. Although we generated predictions for all the 129 targets, only 31 proteins were solved in their holo form and were used in the official assessment [45]. The definition of the binding site residues in our analysis follows the CASP9 assessor's rendition. The COFACTOR prediction results on the 31 proteins are listed in Table 3.3.

Overall, the models by COFACTOR (named "I-TASSER_FN" in the server section and "Zhang" in the human section) were ranked at the top 2 positions based on the mean MCC Z-scores with and without bootstrapping experiment (Figure 3.7). As CASP9 assessors concluded, among all 33 participant groups "Two groups (FN096, Zhang; FN339, I-TASSER_FUNCTION) performed better than the rest, while the following ten prediction groups performed comparably well."[45]

Overall, for the 31 evaluated proteins, the binding-site residues were predicted with an average MCC of 69%, which is slightly higher than the above benchmark test since CASP9 has more easy targets [45]. For the best 24 proteins, more than 50% ligand interacting residues were correctly identified. We observed that most of the high accuracy predictions are for binding-sites harboring non-metal ligands (average accuracy of 75.5%), while the binding-site residues for metal ions have a slightly lower average accuracy 69.8%. The metal ion binding residues also show large variations in their prediction coverage. One of the major reasons for the moderate predictions in cases involving metal ions is the relatively lower quality of receptor models. The average TM-score is 0.66 ± 0.21 for the metal-bound proteins while that for non-metal proteins

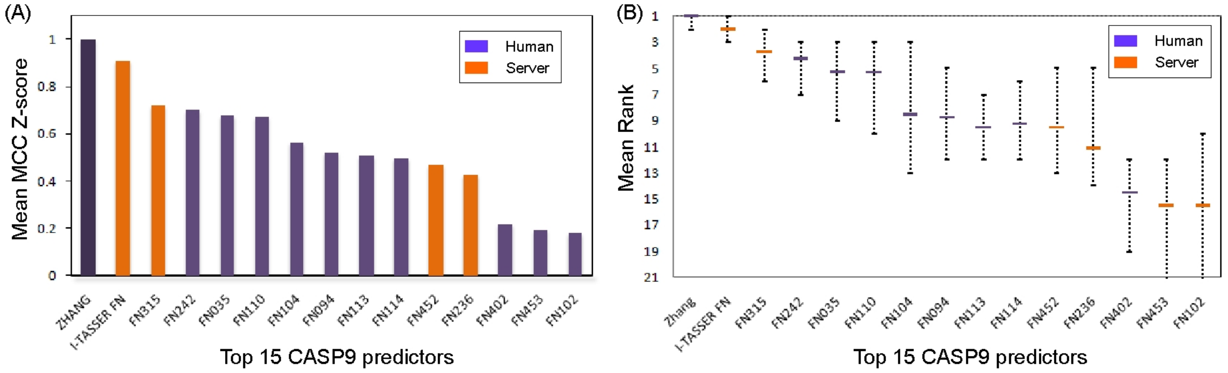


**Figure 3.7** Performance of the top 15 groups in the binding site prediction category in CASP9. Data taken from the CASP9 assessors [46]. (A) Mean MCC Z-score. (B) Mean Rank of CASP9 predictors based on bootstrapping experiment. The top two groups ('Zhang' as human group and 'I-TASSER FN' as automated server group) used COFACTOR to predict the binding site residues in protein structures obtained from I-TASSER predictions.

is 0.82 ± 0.12. Also, in some of these metal-binding proteins COFACTOR additionally predicted non-metal ligand binding sites (for example $PO4^{3-}$ in T0635) and was the source of over-prediction. Nevertheless, similar to observations in the benchmarking analysis, in most of the cases, the predicted and native ligands are highly similar, implying the applicability of COFACTOR for a more detailed elucidation of protein function.

Figure 3.8 shows two representative examples of easy and hard test cases, T0609 and T0518, for which COFACTOR's predictions significantly outperformed other groups. Target T0609 (PDB ID: 3os7) is a putative galactose mutarotase crystallized with tartaric acid. Although the crystal structure was solved without the native ligand, the CASP9 assessors inferred that the protein binds Beta-D-Galactose (GAL) in the same binding cleft as the crystallized tartaric acid. Figure 3.8A shows the successful prediction (MCC=0.82, accuracy=0.75) by COFACTOR for this target, where four of the five binding site residues were correctly identified (shown in green). This prediction was deduced from a distant homologue protein Gal10 bifunctional protein (PDB ID: 1z45) from *S.cerevisiae*, which also binds GAL. Most groups in CASP9 missed the prediction because the template by threading has a poor alignment quality; while COFACTOR used the I-TASSER full-length models (TM-score = 0.78), which correctly detected the template with correct alignment by TM-align. This is an excellent example showing the advantage of COFACTOR by using a better quality of receptor models generated by I-TASSER.

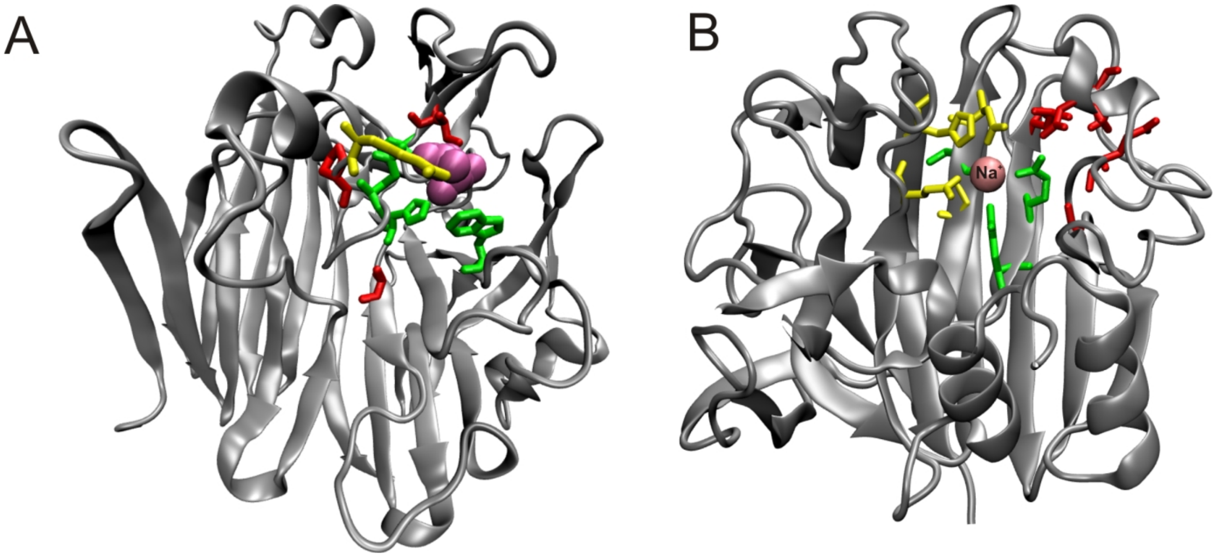


**Figure 3.8** Examples of successful predictions by COFACTOR in CASP9. (A) T0609; (B) T0518. Correctly predicted residues are shown in green (true positive), false positive predictions highlighted in red, and false negatives residues shown in yellow.

T0518 (PDB ID 3nmb) is a putative sugar hydrolase crystallized with sodium ion. Although the receptor was an easy target for structure modeling (TM-score of I-TASSER model is 0.80) and a close homolog (PDB ID: 3imm) had a very similar $Na^+$ binding site, most predictors in CASP9 failed to predict the binding site because $Na^+$ was considered a crystallization artifact. The COFACTOR template library also missed this template protein. However, a local similarity was detected between the I-TASSER model and peanut-lectin (PDB IDs 2dv9 and 2tep). Two binding sites for $Mn^{2+}$ and $Ca^{2+}$ were then predicted by COFACTOR although with a low confidence score in the same binding cleft. Out of the seven native ligand-binding residues (Fig. 3.8B), three residues were correctly identified (shown in green). Five were incorrectly annotated as binding residues (shown in red), while four correct residues (shown in yellow) were missed during the prediction. Nonetheless, T0518 represents a typical successful example, where although a close template was not present in the template library, COFACTOR correctly identified a remote homolog of the protein using local comparisons and provided a reasonable prediction that could be useful for understanding the function.

## 3.6. Why does COFACTOR work?

An important question is: why COFACTOR outperforms most of the existing state-of-the-art methods in overall binding site prediction accuracy, although both COFACTOR and these other methods have exploited the sequence and structural information in their predictions?

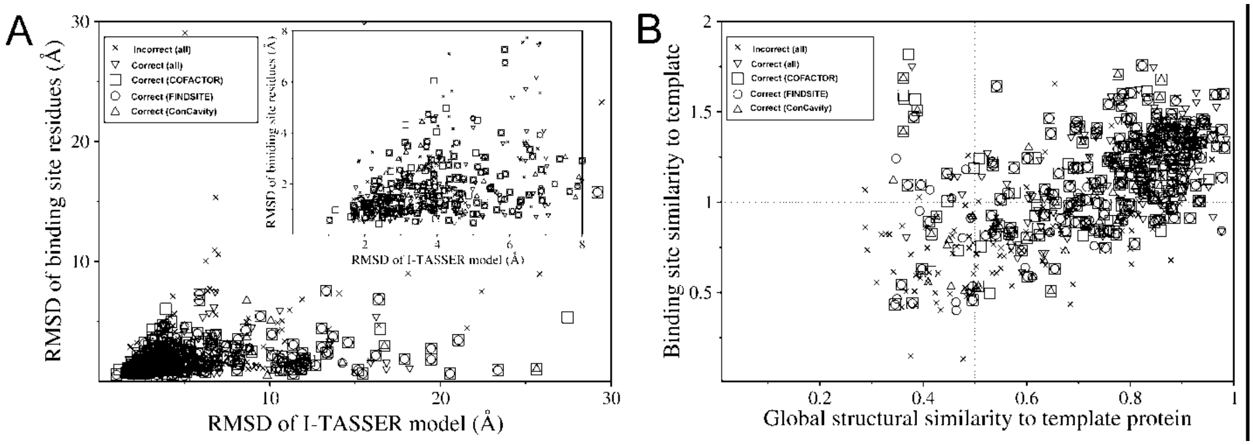


**Figure 3.9** Dependence of COFACTOR on model quality and structural similarity of template proteins. (A) Structural accuracy of ligand binding residues versus the accuracy of full-length receptor models. Ligand binding pocket predictions using higher resolution receptor models are shown in the inset. (B) Local versus global similarity of template to target structures. The local similarity is evaluated by BS-score (Eq. 3.3), while global structural similarity is measured by TM-score of template and the I-TASSER model. In both the plots, the correct predictions with a distance error < 4.5 Å by different methods are represented by different symbols.

In Figure 3.9A, we analyzed the dependence of binding pocket predictions by COFACTOR and the two control methods (FINDSITE and ConCavity) on the accuracy of predicted receptor structure. The local structure quality of predicted receptors is evaluated by the RMSD of known ligand binding residues, while that of global structure is measured by the RMSD of full-length receptor models. For targets with approximately correct global topology (RMSD < 8 Å), all three methods have a reasonable ability to predict the ligand-binding pocket. Nevertheless, COFACTOR generates 12% and 98% more correct (distance error < 4.5Å) binding pocket predictions than FINDSITE and ConCavity (Fig. 3.9A Inset) respectively. Moreover, in these correct predictions, the average distance error of pocket prediction by COFACTOR is lower (1.9 Å), compared to that by FINDSITE (2.1 Å) and ConCavity (3.0 Å), which highlights the fact that a combination of local and global structural alignment improves the accuracy of binding site predictions for easy modeling proteins.

Even for the harder cases, when the global topology of the receptor model is incorrect (global RMSD > 8Å) but the ligand binding pocket is correctly formed (local RMSD < 8Å), COFACTOR had 9% and 89% more correct predictions, compared to the control methods FINDSITE and ConCavity (lower-right area of Figure 3.9A) respectively. Since the topology of the receptor models is incorrect, methods that rely only on global comparisons will have difficulty identifying the correct template, which was improved in COFACTOR by using local structural comparisons.

In Figure 3.9B, we analyzed the performance of COFACTOR in relation to global and local similarity between target and template structures. When query and template proteins have a similar fold (TM-score > 0.5) and the local match near the binding pockets is significant (BS-score > 1.0), in 80% cases the predictions generated by COFACTOR were correct and the average distance error was 1.78Å. Conversely, for protein that uses template proteins of the same fold but with relatively poorer local match (BS-score < 1.0), the prediction accuracy rapidly decreased to 58% and ligand distance error increased to 2.3Å. This highlights the importance of local comparisons while selecting templates in template-based binding site prediction methods. For example, FINDSITE, which does not uses local comparisons, has 9% and 20% lower correctly predicted cases in these two regions. This inference is buttressed, when we examine cases in the upper-left region of Figure 3.9B in which COFACTOR used templates that had different fold (TM-score < 0.5) compared to the query model, as no good template of same fold was available. When a good match near the binding pocket (BS-score >1) was found, in 61% cases the binding pocket prediction was correct, which is 31% and 70% higher than the control methods FINDSITE and ConCavity respectively.

In Figure 3.10, we show a successful example from carnitine CoA-transferase (PDB ID: 1xvt) that demonstrates the strength of local structural matches. In this example, the correct template protein is from the glucose-6-phosphate dehydrogenase (PDB ID: 2bh9) that however has a completely different overall fold with a TM-score to the target of 0.36. Nevertheless, the structure of both template and target contains a pocket with 3-layer (aba) sandwich architecture in their N-terminal region, which forms a NADP+ (bound NAP in 2bh9) binding site in Glucose-6-phosphate dehydrogenase and a CoA binding site in carnitine CoA-transferase. Although there is no global structural similarity, COFACTOR identifies this local structural similarity of the two proteins with a high BS-score, which results in the first model of ligand-binding residues with an MCC of 56% and accuracy of 75%. The predicted ligand (NAP) for the query contains the same adenine and ribo-phosphate moiety as “native” ligand (bound CoA in 1xvt).

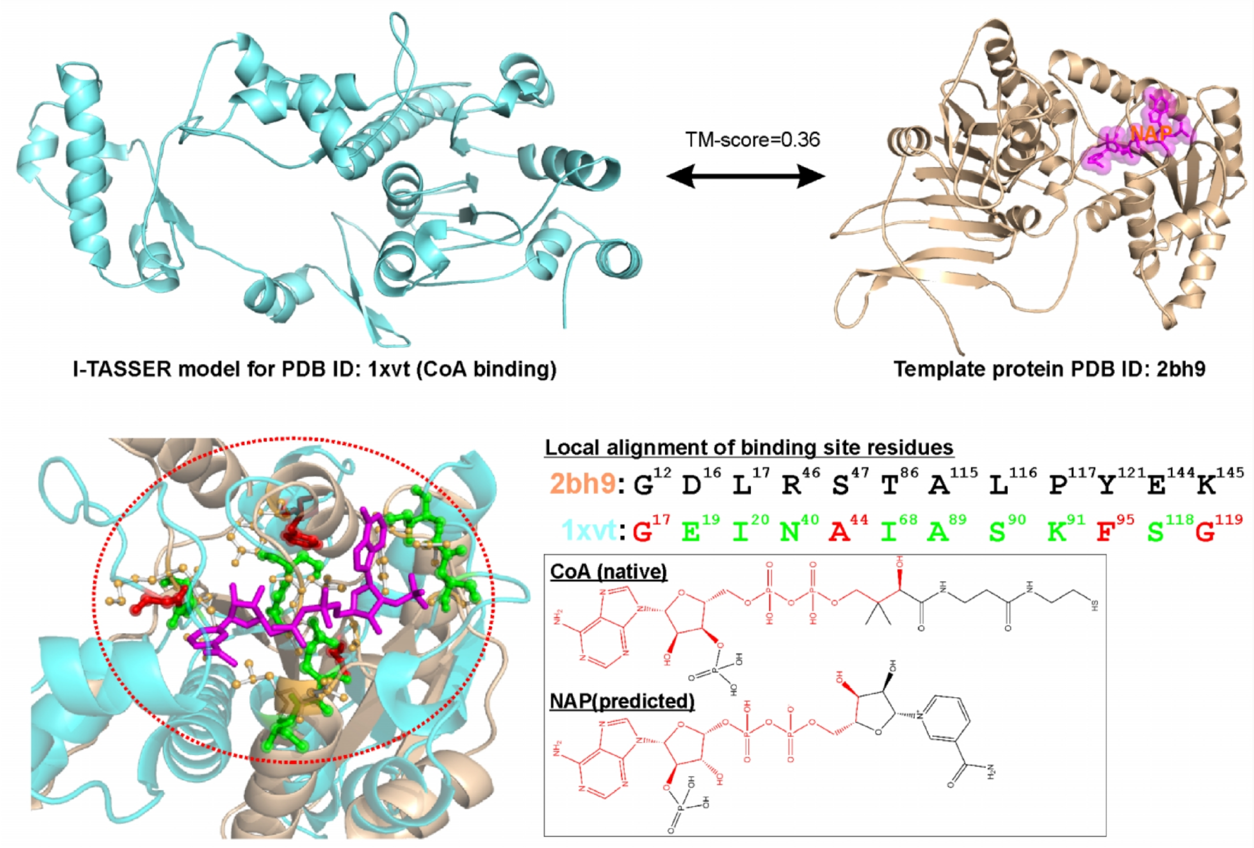

TM-score=0.36
NAP
I-TASSER model for PDB ID: 1xvt (CoA binding)
Template protein PDB ID: 2bh9
Local alignment of binding site residues
2bh9: G12 D16 L17 R46 S47 T86 A115 L116 P117 Y121 E144 K145
1xvt: G17 E19 I20 N40 A44 I68 A89 S90 K91 F95 S118 G119
CoA (native)
NAP(predicted)

**Figure 3.10** A representative example of COFACTOR binding-site prediction based on local structural comparisons. Binding site residues of the carnitine CoA-transferase (PDB ID: 1xvt) was detected using glucose-6-phosphate dehydrogenase (PDB ID: 2bh9) as template with MCC 56% and accuracy 75%. The true positive residues are shown in green and false positive ones are in red. Inset shows that CoA (native ligand) and NAP (predicted ligand) have similar chemical structure (adenine and ribo-phosphate moiety shown in red).

## 3.7. Conclusion

A new approach, COFACTOR, for high accuracy prediction of protein-ligand interaction has been developed. The anatomy of results obtained on a large-scale dataset containing functionally diverse proteins, shows that the algorithm could accurately identify binding pockets in 67% of cases with an average error of 2 Å, when predicted protein structures were used and homologous templates were completely excluded from both structure and protein-ligand template libraries. In

90% of the cases, without knowing the ligand *a priori*, the ligand interacting residues were assigned with an average Matthews correlation coefficient of 61% and 76% accuracy.

We have analyzed the predicted binding sites for both “natural” and “drug-like” molecules, but no significant differences were observed between the predictions for the two classes of molecules. In particular, for 70% of the proteins with bound “natural” ligand, the predicted ligand shared a high chemical similarity to the bound ligand in native state, which suggests a potential application of the method for a more elaborate functional elucidation of uncharacterized proteins. Successful predictions were also observed for “drug-like” compounds, which open up the possibility for structure-based drug design even for proteins with no available structural information.

We have compared our benchmarking results with two recently developed structure-based methods (FINDSITE and ConCavity). Starting from the same set of structural models, the MCC of ligand-binding residue predicted by COFACTOR is 13% and 46% higher than that by FINDSITE and ConCavity respectively; while the distance error in locating ligand-binding pocket by COFACTOR is 1.1 Å and 2.6 Å lower than that by the aforementioned two control methods. In the recent community-wide CASP9 experiment [45], COFACTOR achieved an average MCC of 0.69 and accuracy of 0.72, which significantly outperforms all other methods from 33 participating groups (Figure 3.7).

The major advantage of COFACTOR over the existing methods is the optimal combination of global and local structural comparisons for identifying ligand-binding sites. Firstly, it outperforms the popular cavity-based methods [27; 47; 48] in cases where only low-resolution protein models are available. This is because global topology comparisons can reliably identify the correct functional templates, as their accuracy is not sensitive to the local structural errors.

Secondly, for proteins that have functional templates with different global topology but similar conserved binding pockets, local structural comparisons help COFACTOR to correctly recognize the ligand-binding residues, which cannot be achieved by the purely global structural comparison methods [22; 24; 25].

The latter advantage of local structural comparison is particularly important for functional annotations of proteins in the so-called "twilight-zone" regions, where the protein structure prediction methods often have difficulties in generating correct global fold due to the lack of appropriate templates. However, many methods, including I-TASSER [28; 49], can almost always predict/generate models with correct super-secondary structures [50; 51], especially in the functionally conserved regions, which provide important insight for local-structure based functional inferences. Thus, combining the presented method with the state-of-the-art protein structure predictions represents an automated and optimal method for genome-wide structural and functional annotations for majority of the proteins that lack experimental structures.

## Bibliography

1. Rausell, A., Juan, D., Pazos, F. & Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A* **107**, 1995-2000.
2. Capra, J. A. & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-82.
3. Wang, K., Horst, J. A., Cheng, G., Nickle, D. C. & Samudrala, R. (2008). Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. *PLoS Comput Biol* **4**, e1000181.
4. Valdar, W. S. (2002). Scoring residue conservation. *Proteins* **48**, 227-41.
5. Pei, J. & Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700-12.
6. Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27**, 2985-2993.
7. Levitt, D. G. & Banaszak, L. J. (1992). POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph* **10**, 229-34.
8. Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**, 323-30, 307-8.
9. Hendlich, M., Rippmann, F. & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* **15**, 359-63, 389.

10. Brady, G. P., Jr. & Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* **14**, 383-401.
11. Huang, B. & Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* **6**, 19.
12. Weisel, M., Proschak, E. & Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J* **1**, 7.
13. Le Guilloux, V., Schmidtke, P. & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168.
14. Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28**, 849-57.
15. Laurie, A. T. & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**, 1908-16.
16. Wade, R. C., Clark, K. J. & Goodford, P. J. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J Med Chem* **36**, 140-7.
17. Lin, J. H., Perryman, A. L., Schames, J. R. & McCammon, J. A. (2002). Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc* **124**, 5632-3.
18. Landon, M. R., Amaro, R. E., Baron, R., Ngan, C. H., Ozonoff, D., McCammon, J. A. & Vajda, S. (2008). Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem Biol Drug Des* **71**, 106-16.
19. Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M. & Bourne, P. E. (2010). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*.
20. Russell, R. B., Sasieni, P. D. & Sternberg, M. J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* **282**, 903-18.
21. Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**, 163-4.
22. Brylinski, M. & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **105**, 129-34.
23. Xie, L. & Bourne, P. E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A* **105**, 5441-6.
24. Oh, M., Joo, K. & Lee, J. (2009). Protein-binding site prediction based on three-dimensional protein modeling. *Proteins* **77** **Suppl 9**, 152-6.
25. Wass, M. N., Kelley, L. A. & Sternberg, M. J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* **38 Suppl**, W469-73.
26. Tseng, Y. Y. & Li, W. H. (2011). Evolutionary approach to predicting the binding site residues of a protein from its primary sequence. *Proc Natl Acad Sci U S A* **108**, 5313-8.
27. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* **5**, e1000585.
28. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.
29. Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40.
30. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
31. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.
32. Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol* **162**, 705-8.
33. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
34. Brylinski, M. & Skolnick, J. (2009). FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput Biol* **5**, e1000405.
35. Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-9.

36. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

37. Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. A. (2005). Binding MOAD (Mother Of All Databases). *Proteins* **60**, 333-40.

38. Dessailly, B. H., Lensink, M. F., Orengo, C. A. & Wodak, S. J. (2008). LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res* **36**, D667-73.

39. Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N. & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **50**, 726-41.

40. Perola, E., Walters, W. P. & Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**, 235-49.

41. Skolnick, J. & Kihara, D. (2001). Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. *Proteins* **42**, 319-31.

42. Pandit, S. B. & Skolnick, J. (2008). Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **9**, 531.

43. Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. & Willighagen, E. L. (2006). The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model* **46**, 991-8.

44. Giganti, D., Guillemain, H., Spadoni, J. L., Nilges, M., Zagury, J. F. & Montes, M. (2010). Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J Chem Inf Model* **50**, 992-1004.

45. Schmidt, T., Haas, J., Cassarino, T. G. & Schwede, T. (2011). Assessment of ligand binding residue predictions in CASP9. *proteins*, In press.

46. Schmidt, T., Haas, J., Gallo Cassarino, T. & Schwede, T. (2011). Assessment of ligand binding residue predictions in CASP9. *Proteins*, in press.

47. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**, W89-93.

48. Sael, L. & Kihara, D. (2010). Binding ligand prediction for proteins using partial matching of local surface patches. *Int J Mol Sci* **11**, 5009-26.

49. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**, 108-117.

50. Jauch, R., Yeo, H. C., Kolatkar, P. R. & Clarke, N. D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**, 57-67.

51. Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J. L. & Levy, Y. (2009). Assessment of CASP8 structure predictions for template free targets. *Proteins* **77 Suppl 9**, 50-65.

# Chapter 4

## Remote homolog detection and function prediction using COFACTOR

## Introduction

Various computational approaches have been developed till date to provide clues about the biological role of functionally uncharacterized gene products accumulating in public databases. The *de facto* approach is to identify evolutionarily related proteins of known function using sequence-based search[1; 2; 3] and confer the functional annotation from a close relative (homolog). However, such annotation transfers are usually error-prone, mainly for three reasons: (a) a sequence similarity threshold that can warrant the correctness of annotation transfer cannot be defined since homologous proteins in different species evolve at different rates owing due to both different selection pressures and mutation rates; (b) even protein pairs sharing very high sequence identity can exhibit functional divergence[4; 5]; and (c) many early sequenced genomes were annotated using homology transfer approach, introducing the distinct possibility of misannotation of template proteins in function databases[6]. Nevertheless, provided that the annotation of the template proteins are correct, it is estimated that reliable functional inferences can be drawn for nearly 35-50% of open reading frames (ORFs) in a given genome[7; 8], using sequence-based homology transfer, The more challenging problem, however, is to identify the function of ORFans, for which we either don't find any hit using sequence-based search or they don't share significant sequence similarity to proteins with known function. This problem can be partially addressed by identifying locally conserved motifs within the sequences using databases like Prosite[9], Print[10] and Blocks[11]. However, the patterns in these databases are mostly generated using multiple sequence alignments and can only be used to search conserved elements within the same protein family and carry little information about functionally conserved residues across different protein families.

During the process of evolution, nature remembers the protein folding pattern. As a result, protein structures are found to be more conserved[12] compared to sequence and provide much more information for inferring evolutionary and functional relationships[13; 14], especially when the sequences diverge beyond the *twilight zone*[15] of sequence identity. In most cases, when two proteins share a common scaffold, they perform same or similar function. However, this cannot be generalized for all protein folds because proteins inhabiting the same fold but performing different function has been reported[16]; thus making the predictions based on normal structure-function relationship skewed. The reason for functional promiscuity of some folds is that random mutations occurring during evolution are more likely to change the function first, rather than the structure.

Natural selection, the primary means for protein evolution, acts to optimize the function. As a result, functionally important residues remain more conserved than the overall global fold. Many contemporary structure-based approaches have therefore been devised to identify local structural similarity for drawing functional inferences[17; 18]. Depending on whether or not these methods use *a priori* knowledge of functional residues/region in the local similarity search, they can be broadly classified into *template-free* and *template-based* methods[19]**. I**n *template-free* methods, the protein-structure pairs are compared to identify similar functional patches (e.g. shape, electrostatic surface, binding site/clefts profiles etc.). The geometrical and physiochemical features of these identified patches are then correlated with known biochemical function. In *template-based* methods, query protein structure is scanned against a precompiled library of known active site or ligand binding residues to identify similar functional motifs[20; 21; 22]. These libraries of conserved *local 3D-motifs* can be compiled either manually[23] after doing a literature search or using automated scripts[24]. As the manual creation of these local templates is extremely laborious, time consuming and hence not very comprehensive, methods like PROFUNC[20] uses both available template libraries as well as automatically created tri-peptide fragments created using query itself (reverse templates) to match against a library of non-redundant structures; Evolutionary Trace (ET)[25] approach uses predicted functional residues to automatically create the template, while GASP[26] method uses genetic algorithm to build the templates based on their ability to discriminate between different protein families against a background of representatives from the SCOP[27] database. These local similarity search methods have great potential to detect functional similarity between evolutionarily diverged and sometimes evolutionarily unrelated proteins. However, the performance of these aforementioned methods is largely dependent on the resolution of the query structure and the binding state (apo/holo) of the compared protein pairs. Moreover, there is no gold standard for assessing the significance of local match, and most methods use similarities between random pairs of structures to fit a distribution and use it as a background for evaluating the statistical significance[20].

As the function of proteins is accompanied by binding, in the current work we extend our COFACTOR methodology for ligand binding-site detection, to predict two other unambiguously defined concepts of functions: Enzyme Commission (EC) numbers[28] and Gene Ontology (GO)[29] terms. Since most of the proteins lack known 3D-structure and anticipating that the comparative modeling methods will continue to improve as the PDB library saturates[30], we will start from the amino acid sequence and generate low-to-medium resolution 3D-model using the iterative threading assembly refinement method (I-TASSER)[31; 32; 33]. The predicted model is matched against two independent template function libraries for identifying functional homologues using a combination of global and local structure-based similarity search algorithms and the significance of the match is judged based on novel scoring function that combines sequence-structure similarity scores in a single term. The algorithm is benchmarked on a large benchmark set of 450 non-homologous proteins collected from PDB library and commonly-used homology based approaches namely sequence-profile alignment[2], profile-profile alignment[34] and HMM-HMM alignment[35] as experimental controls. To demonstrate the suitability of this approach for genome-scale functional annotations, the method is applied to a subset of proteins encoded by the phylogenetically distinct bacteria, *Chlamydia trachomatis* and the functional annotations suggest new insights into this phylogentically distinct bacterium.

## 4.1 Materials and methods

### 4.1.1 Definition of function

Although the definition of protein function is subjective and contextual, we have used two standard vocabularies for defining a protein's function: (a) EC number and (b) Gene Ontology terms. For the benchmarking proteins, these annotations were taken from the PDB database.

### 4.1.2 Data sets

We evaluate our function prediction approach on two different datasets. The first set contains 450 non-homologous proteins collected from the PDB library with diverse functions, while ensuring that the pair-wise sequence identity is below 30% and there is no non-self BLAST[1] hit within the dataset. In this set, 318 proteins are enzymes, with unique EC numbers covering all the 6 enzyme classes. The GO term predictions are evaluated on 337 proteins annotated with at least one GO term. These include 205 enzymes and 132 non-enzymatic proteins. Of these 337 proteins, 308 were annotated with at least one molecular function term; 295 were annotated to be involved in a biological process and cellular location was annotated for 213 proteins in the PDB GOA annotation[36]. The Gene Ontology predictions are evaluated on each of the three subsets individually, and also as a combined set.

The second set represents a blind test case on 17 medium-sized proteins taken from the ORFs of the *Chlamydia trachomatis* genome. As an experimental control, we include in this set 7 proteins (CT243, CT296, CT381, CT390, CT610, CT780 and CT828) from *Chlamydia trachomatis* for which the structure (or that of a very close homologous protein) has already been experimentally determined. The remaining 10 proteins don't have known structure and are ascribed as hypothetical proteins in RefSeq[37] database. During the structure modeling and function prediction experiment of these proteins, no template proteins (except for solved protein structures belonging to the *Chlamydia* genus) were excluded from the template library.

### 4.1.3 Detection of functional homologs

For identifying functional homologs of a query protein, its predicted 3D-structure is scanned against two representative template libraries of experimentally determined protein structures associated with (1) Enzyme Commission (EC) number; and (2) Gene Ontology (GO) terms. The template proteins in these libraries are first scanned based on global structure similarity search algorithm. In the following step, a local structure similarity refinement search is performed on selected hits with the purpose of filtering out template proteins that do not share functional site similarity with the query protein. During both global and local structure similarity search (described below), template proteins in the database are scored against the query protein structure using an innovative structure-sequence similarity measure, which is devised to capture the functional homology between the query and the template protein. As we are using modeled structures in this study, confidence score of functional annotation using predicted protein structure is scaled based on confidence score of structure modeling; in order to provide reliable functional annotations using identified hits.

#### 4.1.3.1 Global similarity search

In most cases, the function of a protein is dictated by both its structure as well as by its sequence. Based on this tenet, we extend the popularly used TM-align[38] algorithm to quickly search the function libraries and identify template proteins that have both high structure similarity and similar sequence profile as the query protein.

TM-align uses TM-score[39] rotation matrix and dynamic programming to find the best structural alignment between two protein structures. The main advantage of TM-align is that it uses TM-score rotation matrix, over often-used RMSD matrix. TM-align is more sensitive to the global topology of structure, because TM-score inherently down-weights large distances between aligned $C_\alpha$ pairs compared to the smaller ones. On the other hand, RMSD weights all the residues equally, and therefore local errors in the model (e.g. an incorrectly-oriented tail) will result in a large RMSD value, even when the global topology of the two structures might be similar. Moreover, TM-score accounts for both the structural similarity in the aligned region and the alignment coverage in a single parameter, which is important for differentiating alignment between single-to-single and single-to-multiple domain proteins.

In the modified algorithm, the initial seed alignments (gapless threading, secondary structure match and the combination of the two) are the same as in the original TM-align program. In addition to this, a new seed alignment based on profile-profile alignment of query and template protein sequence has been implemented. These seed alignments are then refined further based on heuristic iterations of the Needleman-Wunsch dynamic programming[40], where the score for aligning *i*th residue of the query to the *j*th residue in the template is given by:

$$S_{ij} = \frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} + w1\sum_{k=1}^{k=20} F(i,k)\times P(j,k) + w2\delta_{ij}\,, \tag{4.1}$$

where $d_{ij}$ is the distance between $C_\alpha$ atoms of query and template, $d_0$ is given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ and $L$ is the length of the query protein. $F(i,k)$ is the frequency of *k*th amino acid at *i*th position of the multiple sequence alignment (MSA) of PSI-BLAST hits obtained for the query sequence against the non redundant sequence database using an E-value cutoff of 0.001. $P(j,k)$ is the log-odds profile (Position Specific Substitution Matrix) for the template protein sequence for the *k*th amino acid at the *j*th position with values normalized between 0 and 1. $\delta$ is a step function for evaluating sequence identity between the amino-acid pairs and equals to 1 when $k(i) = k(j)$ and 0 otherwise. *w1* (= 0.1) and *w2* (= 0.9) are parameters optimized on a benchmark set of 100 protein structures (both model and experimental) with distinct functions (EC and GO terms) using an objective function to increase the number of proteins with similar function in top 1/5/10 ranked hits. For a given pair of structures, the global similarity ($G_{sim}$) to measure both topological similarity and sequence profile conservation is given as:

$$G_{sim} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}} S_{ii}\right], \tag{4.2}$$

where $L$ is the length of the query protein, $L_{ali}$ is the number of the aligned residue pairs and $S_{ii}$ is the score obtained from Eq.4.1 for *i*th aligned residue pair.

### 4.1.3.2 Local similarity search

The global similarity search described above can be used for recognizing protein pairs with highly similar functions. However, some folds are highly promiscuous, especially when function is analyzed using EC number vocabulary, because even closely related homologues can belong to entirely different classes of enzymes[16]. In these cases, function can be predicted precisely only by evaluating the similarity between constellations of active site residues required for the catalysis. Moreover, in many cases local-3D functional motif remain conserved during the evolution to maintain the function, even when the global similarity dwindles or become undetectable. This begets a need for identifying the local structural similarity, which is complimentary to the global search and also act as an additional filter for hits obtained using global search, thus providing a more reliable way of annotating the function of the query proteins. It needs to be mentioned that proteins can also share same function because of convergent evolution of protein families (analogs). However, in this work we only focus on predicting the function using distant evolutionary relatives (homologs).

The local similarity search for identifying functional sites in query structure is done using the COFACTOR algorithm (discussed in last chapter), where the local similarity ($L_{sim}$) score for evaluating similarity between query and template functional sites is defined as:

$$L_{sim} = \frac{1}{N_t} \sum_{i=1}^{i=N_{ali}} \frac{1}{1+\left(\frac{d_{ii}}{d_0}\right)} + \frac{1}{N_t} \sum_{i=1}^{i=N_{ali}} M_{ii} \, , \qquad (4.3)$$

where $N_t$ represents the number of residues present in the active/binding site sphere of the template, $N_{ali}$ is the number of aligned residue pairs, $d_{ii}$ is the $C_\alpha$ distance between residue $i$th aligned residues, $d_0$ is the distance cutoff chosen to be 3.0 Å, $M_{ii}$ is the substitution scores between $i$th aligned residues taken from BLOSUM62 matrix with values normalized between [0, 1], in order to keep both the distance and mutation scores in Eq. 4.3 on the same scale. The highest $L_{sim}$ encountered during the heuristic iterations is recorded for each candidate *local 3D-motif* of the query. Finally, the motif with highest $L_{sim}$ is selected for evaluating the local similarity between the query and the template's functional site and sequence similarity between aligned residues of query and known functional site residues in template is recorded as $SS_{BS}$.

### 4.1.4 Confidence score of functional annotations

In this study, since we have employed predicted 3D structures the quality of functional inference relies on the quality of the structural models. Appraising the accuracy of the structure model in the scoring scheme is necessary to reduce the number of false positive predictions. When the native structure is unknown, the quality of the I-TASSER model can be estimated using C-score[41], which is defined as:

$$\mathrm{C-score} = \ln\left[\frac{M}{M_{tot}} \times \frac{1}{\langle RMSD \rangle} \times \sum_{i=1}^{N} \frac{Z(i)}{Z_0(i)}\right], \tag{4.4}$$

where $M$ is the multiplicity of structure decoys in the SPICKER cluster, $M_{\mathrm{tot}}$ is the total number of decoys submitted for the SPICKER clustering, ⟨RMSD⟩ is the average RMSD of the clustered decoys to the cluster centroids, $Z(i)$ is the Z-score of the top threading alignment obtained from $i$th server in LOMETS, $Z_0(i)$ is the Z-score cutoff to distinguish good and bad threading alignments for the server, and $N$ is the number of servers used in LOMETS. The C-score value is typically in the range of [-5, 2], where a higher score reflects a model of better quality. To normalize the range to [0, 1], we transform the C-score according to C-score=(C-score+5)/7. In a large scale test of 500 non-homologous proteins, the correlation coefficient between C-score and the TM-score of the first I-TASSER model is 0.91[41].

The confidence score of functional annotations (FC-score$_{EC}$, FC-score$_{GO}$) using any template proteins with values ranging between [0, 1] is defined using an exponential function, mathematically formulated as:

$$\mathrm{FC-score} = \frac{2}{1+e^{-(x)}} - 1, \tag{4.5}$$

where $x$ is an experienced scoring function to combine local and global similarities between query and template proteins, defined as:

$$x = \mathrm{C-score} \times \left[ \left( w \times L_{sim} \times SS_{BS} \right) + G_{sim} - \delta \right]. \tag{4.6}$$

Here C-score is the confidence score of predicted I-TASSER structure, $L_{sim}$ and $G_{sim}$ are measures of local and global similarity between the query and the template, $SS_{BS}$ is the sequence similarity of binding site residues, $w$ is a scaling factor for local similarity, and δ is the functional similarity threshold. The values of $w$ (=0.50) is the weight for local similarity and δ (=0.70) is the inflexion point of the exponential curve; both the values have been decided based on benchmark training of 100 protein structures with annotated function (GO terms and EC numbers).

### 4.1.5 Consensus Gene Ontology predictions

Template proteins in the GO library are mostly associated with multiple GO terms, describing different aspects of biological and cellular functions. Simply transferring GO annotation based on identified hits can be error prone, especially when the template proteins have additional functional domains. Based on the assumption that each domain contributes independently to the protein function, we reconcile the GO terms ascribed to the top ranking hits, such that the consensus predictions identifies the intersection of function and provides specific annotation to the query protein. For reconciling the GO annotations we make use of the hierarchical nature of

Gene Ontology functions, and consider that when a template protein is annotated with a GO term, all its ancestor GO terms are automatically ascribed. The following steps provide details of the consensus approach and is adapted from the PIPA algorithm[42].

1. For any query protein we first collect a set $F$ of GO terms from the top $N$ templates, where each GO term $\lambda \in F$. Each GO term is assigned a confidence score (GOscore), given by the following equation

   $$\text{GOscore}(\lambda) = 1 - \prod_{i=1}^{i=N_\lambda} [1 - \text{FCscore}(\lambda)], \tag{4.7}$$

   where $N_\lambda$ is the number of templates which are associated with the GO term $\lambda$, and $N$ (=5) is the total number of templates selected for find the concurrence of function.

2. Next, we score all ancestor GO terms of $\lambda$. The score for any ancestor GO term $\mu$ is scored as:

   $$\text{GOscore}(\mu) = GOscore(\lambda)\left(1 + \frac{N_\mu}{N_0}\right), \tag{4.8}$$

   where $N_\mu$ & $N_0$ are the number of leaf nodes under node $\mu$ and the root node.

3. Subsequently, all GO-terms are sorted based on their depth in the Directed Acyclic Graph (DAG) and GO-terms with GOscore >0.30 are predicted. Once a GO term is predicted its ancestor GO-terms are automatically eliminated from the sorted list. This is because ancestor GO-terms are automatically implied based on true-path rule of Gene Ontology, which states that any gene associated with a GO term is also associated with the ancestors GO-terms leading back to the ontology root. If this procedure does not yield any confident GO-term prediction (GOscore <0.30), then the GO-terms are anyways ranked based on their GOscore and top 10 GO term predictions are reported.

### 4.1.6 Template libraries

To represent the functional space, two independent template libraries of protein structures with known biological function (EC number or GO terms) has been created. These libraries are freely available from our website http://zhanglab.ccmb.med.umich.edu/COFACTOR/library.

***EC template library:*** With the purpose of annotating enzyme function, a library of template enzyme structures with annotated EC number(s) has been collected from PDB[43] and their corresponding EC number mapping using PDBSprotEC[44]. Template proteins which are also enlisted in Catalytic Site Atlas (CSA)[45] are marked as preferred templates because active site residues are already known for these templates and they can be used for both local and global structure similarity comparisons. Finally, redundant protein chains having the same enzyme commission[28] (EC) number and >90% sequence identity to other enzymes in the library are excluded, while preferentially keeping the templates with known active sites, to reduce the search space and eliminate biased predictions;. At the time of this work, the compiled enzyme library contained 8392 protein chains with 223 unique first 3 digits of EC number and 1,947 unique 4-digit EC numbers.

***GO template library:*** The gene ontology (GO) is currently the most effective approach for machine-legible and automatic functional annotation. To this end, a second library of protein chains that have an associated GO term has been created by downloading the structures from the PDB along with the GO mapping taken from the Gene Ontology Annotation database (http://www.ebi.ac.uk/GOA/) and SIFTS project (http://www.ebi.ac.uk/pdbe/docs/sifts/). Redundant protein chains, annotated with same GO terms and >90% sequence identity to any other protein in the library, are excluded. Currently, this library contains 24,035 unique protein chains associated with 13,757 unique GO terms.

#### 4.1.6.1 Construction of *ab-initio* signature motifs for template proteins

Even though we preferentially keep templates with experimentally characterized functional sites while constructing the library, only 72% templates in EC library and 42% templates in the GO library are imputed with known active/binding sites. This can greatly limit the applicability of the local approach similarity search to only those proteins for which templates with known functional sites are identified.

We tackle this problem of template proteins with unknown functional sites by generating *ab-initio* signature motifs for the template proteins that lack known functional site information. Here, we summarize the procedure of identifying potential functional sites in template, which can thereupon be used for generating *local 3D-templates*. Identification of potential functional site in the template, although *ab-initio*, is still based on the prior knowledge that functional sites in proteins are constellated by conserved residues, are solvent exposed for ligand binding and form a concave pocket which in most cases is located towards the protein core i.e. away from the surface. To identify residues that match the three aforementioned criterions, we first perform a PSI-BLAST search against the non-redundant (NR) sequence database using the amino-acid sequence of template protein. The identified hits and alignments are used for creating a multiple sequence alignment (MSA), and $Z\text{-}score_{\mathrm{JSD}}$ of Jensen–Shannon divergence (JSD) score is calculated for every template residue $\boldsymbol{i}$.

Template residues that are located towards the protein core are identified based on spatial $Z\text{-}score_{\mathrm{C}}$ is calculated based on residue centrality(C) [46] defined as:

$$C_i = \frac{L_t - 1}{\sum_{i \neq k} d(i,k)}, \tag{4.9}$$

where $L_t$ is the length of template protein and *d(i,k)* is the shortest distance between residue ***i*** and ***k***.

The accessibility of residue ***i*** for ligand binding is evaluated based on its solvent accessibility *SA(i),* which is calculated using the DSSP[47] program.

Finally, we select potential functional (*PF*) site residues based on criteria described in Equation 4.10.

$$PF(i) = \begin{cases} 1 & \text{if } Zscore_{JSD}(i) \geq -0.2 \text{ \& } Zscore_{C}(i) > -0.2 \text{ \& } SA(i) > 0.2 \\ 0 & \text{else} \end{cases} \tag{4.10}$$

All residues with *PF(i)*=1, are clustered together using an average spatial distance cutoff of 15Å. The clustered residues are ranked based on their cluster size and the top 5 clusters are selected as potential functional sites. The local search procedure described above, constructs the *local 3D-fragments* using clustered residues in these potential sites and the best local match is used for evaluating the local similarity between query and template protein.

### 4.1.7 Assessment of protein function prediction

For the EC number predictions, we first evaluate the functional similarity between the template proteins and query protein at different levels of Enzyme Commission nomenclature. A precision-recall analysis is then performed for the predicted EC number using identified templates and by varying their threshold of prediction confidence score. We consider a match between the first three digits of EC number as true positive, because in most cases the last digit of Enzyme Commission nomenclature represents only substrate specificity or serial number of enzyme.

Gene Ontology predictions are evaluated by measuring semantic similarity (SS) between GO terms and functional similarity between gene products[48]. Semantic similarity measures degree of relatedness between the GO terms based on the DAG structure of Gene Ontology and information content of the term. In this analysis, we used Wang et al.'s hybrid approach[49] for measuring SS, as it determines the SS between two GO terms based on their location in the GO graph and the relation with their ancestor terms. Given this way of measuring SS between two GO terms, we evaluate functional similarity (*Fsim*) of predicted GO terms $\mathrm{GO}_1 = \{go_{11}, go_{12}, go_{13}...go_{1m}\}$ with the annotated GO terms of query protein $\mathrm{GO}_2 = \{go_{21}, go_{22}, go_{23}...go_{2n}\}$ using the best match average score strategy[49], which is defined as :

$$Fsim(\mathrm{GO}_1,\mathrm{GO}_2) = \frac{\sum\limits_{1\le i\le m} SS_{\max}(go_{1i}, GO_2) + \sum\limits_{1\le j\le n} SS_{\max}(go_{2j}, GO_1)}{m+n}, \tag{4.11}$$

where $SS_{\max}(go, GO)$ represents the maximum SS between *go* and any of the terms in the set *GO*. Both *Fsim* and SS range between 0 and 1.

## 4.2 Results

### 4.2.1 Overview of function predictions using COFACTOR

We provide an overview of the function prediction (EC numbers and GO terms) methodology using two example proteins CT867 and CT043 taken from *Chlamydia trachomatis* serovar D genome. We specifically select these two proteins, because although they are ascribed as 'hypothetical protein' in RefSeq database[50], available experimental data aids in objective assessment of the predicted functions using the previously described COFACTOR algorithm.

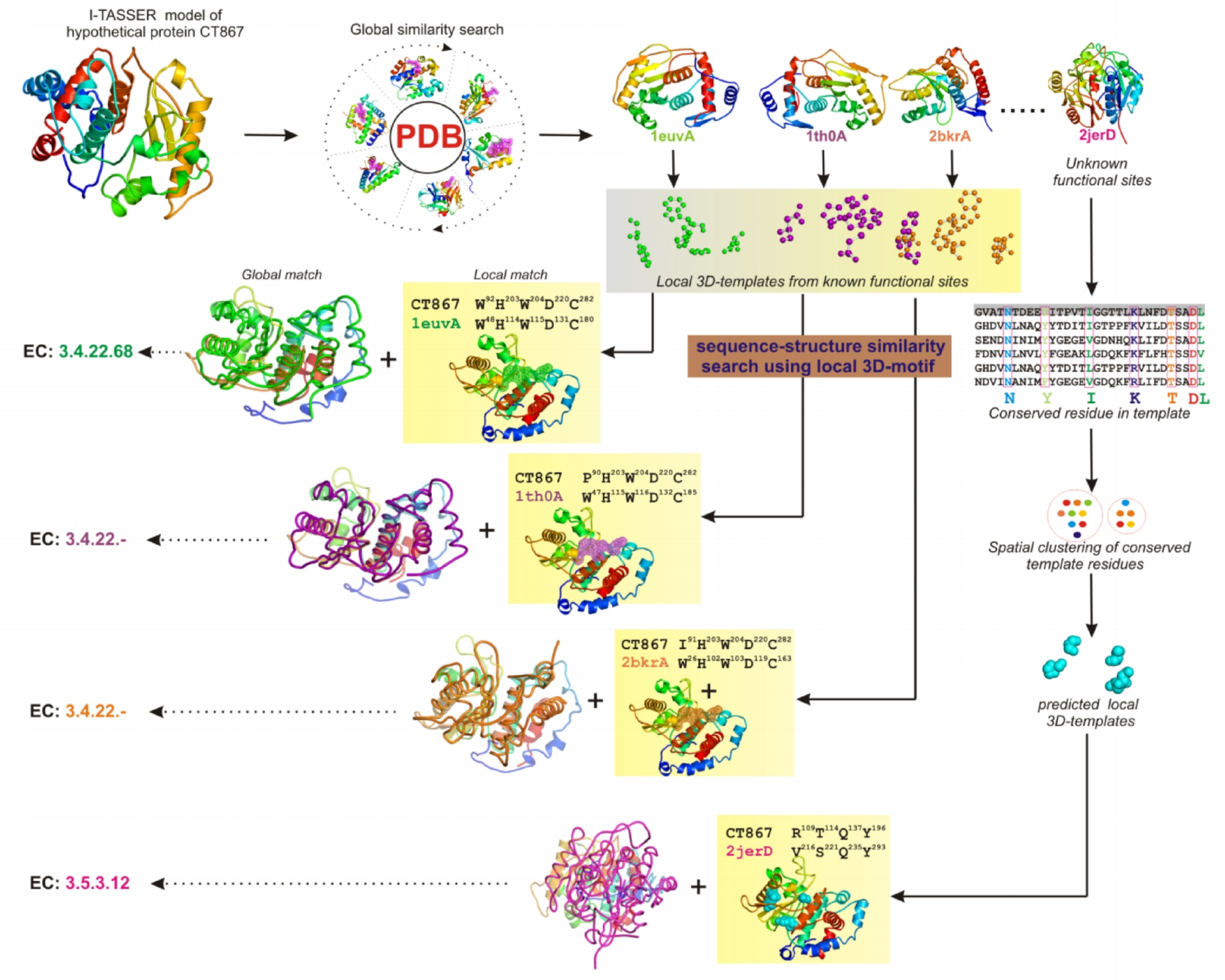

I-TASSER model of hypothetical protein CT867
Global similarity search
PDB
1euvA
1th0A
2bkrA
2jerD
Unknown functional sites
Local 3D-templates from known functional sites
Global match
Local match
EC: 3.4.22.68
CT867 $W^{92}H^{203}W^{204}D^{220}C^{282}$
1euvA $W^{48}H^{114}W^{115}D^{131}C^{180}$
sequence-structure similarity search using local 3D-motif
GVATNTDEEYITPVTIGGTTLKLNFDTSADL
GHDVNLNAQYYTDITIGTPPFKVILDTSSDL
SENDNINIMYYGEGEVGDNHQKLIFDTSADL
FDNVNLNVLYFGEAKLGDQKFKFLFHTSSDV
GHDVNLNAQYYTDITLGTPPFRVILDTSSDL
NDVINANIMFYGEGEVGDQKFRLIFDTSADL
N Y I K T DL
Conserved residue in template
EC: 3.4.22.-
CT867 $P^{90}H^{203}W^{204}D^{220}C^{282}$
1th0A $W^{47}H^{115}W^{116}D^{132}C^{185}$
Spatial clustering of conserved template residues
EC: 3.4.22.-
CT867 $I^{91}H^{203}W^{204}D^{220}C^{282}$
2bkrA $W^{26}H^{102}W^{103}D^{119}C^{163}$
predicted local 3D-templates
EC: 3.5.3.12
CT867 $R^{109}T^{114}Q^{137}Y^{196}$
2jerD $V^{216}S^{221}Q^{235}Y^{293}$

**Figure 4.1** EC number predictions for hypothetical chlamydial protein CT867 using I-TASSER model and COFACTOR algorithm.

***Function predictions for CT867:*** Both structure and function of CT867 is unknown. A simple PSI-BLAST[2] search through the NCBI non-redundant (nr) sequence database annotates it as glycogen branching enzyme (EC No: 2.4.1.18) based on identified closest homolog with *e-value* 1e-107. All other identified homologs too have unknown functions. We modeled the tertiary structure (Figure 4.1) of this protein using the automated I-TASSER pipeline and the model was predicted with a confidence score (C-score) of -2.84, which indicates a low-resolution predicted structure[41] for this protein. Next, we scanned this model through two independent structure libraries annotated with EC number and GO terms using global structure alignment (see Materials and methods), and the proteins in each library (templates) were ranked based on their global similarity score ($G_{sim}$, Eq. 4.2) to the query model. Here, the global similarity score measures both topological similarity and conservation of sequence profile. As the confidence score of the predicted model is low (C-score < -1.5), top 40 scoring templates were selected from both enzyme and GO template libraries. Using more number of templates helps to improve the prediction coverage because structural inaccuracies in the model lower the sensitivity of global structural alignment; as a result, proteins belonging to same protein family/superfamily might get missed. All these templates were used in the next step of local refinement search

For EC number predictions, we first tried to collect functional motifs of these selected templates from a pre-compiled library of known active site residues. For most (72%) proteins in the current enzyme template library, active site residues are already known; for the rest, we predicted 5 potential functional sites in the template protein, based on spatial clustering of evolutionarily conserved residues in template protein, and used them for screening against the query structure. Figure 4.1 shows the procedure of constructing signature motifs for template protein 2jerD, as its active site was unknown. The predicted model of CT867 was scanned through motifs gleaned from these functional sites using a local structural similarity search procedure, which seeks to identify the optimal local sequence-structure match between the two protein structures. Only 3 template proteins (PDB ids: 1euvA, 1th0A, 2bkrA) exhibited a significant local match ($L_{sim} > 1.0$; Eq. 4.3) with query protein and all of them belong to Cysteine endopeptidase family (EC No: 3.4.22.-). In the end, the template proteins were ranked based on combined global and local similarity scores (FC-score; Eq. 4.5 & 4.6) and used to evaluate the confidence score of function prediction. For CT867, all the three templates with good local match were ranked higher, while all the lower ranked template proteins belonged to different enzyme classes. Also, the locally aligned residues in CT867 model with known active site residues of the peptidases are identical (Figure 4.1), which further strengthens the prediction as correct.

Similar to EC number, for GO terms prediction, functional motifs of the selected template proteins were collected from known ligand binding sites or were predicted where required. Local refinement search was then performed and the templates were ranked based on their combined global-local similarity score. CT867 is a single domain protein and all the templates identified in the GO library are peptidases with single domain. Identifying common function among these hits is trivial, so we keep the discussion related to consensus GO term prediction for the next presented example (CT043) described below. Nevertheless, the GO term predictions for CT867 are: Cysteine-type peptidase activity (GO:0008234), binding (GO:0005488) and proteolysis (GO:0006508).

Experimental characterization of CT828 has shown that the protein possesses deubiquitinating and deneddylating activity[51]. This is closely related to the best scoring template (PDB id: 1euvA), which is a yeast ULP1 protease catalyzing: (a) cleavage of SUMO to its mature form; and (b) deconjugation of SUMO from target protein.

***Function predictions for CT043*:** The second illustrative example is of CT043, which is also ascribed as a hypothetical protein. However, a recent experimental study[52] showed that CT043 encodes a T3SS chaperone that binds to the N-terminal region of the effector TARP[53].

I-TASSER predicted the tertiary structure of this protein with a C-score of -1.03, which suggests that the model may have a correct topology, but the intricate structural details might be incorrect. Next, potential functional homologs of this protein were identified by performing a global structural similarity search through both EC and GO libraries and the template proteins were ranked based on their global similarity score ($G_{sim}$). For CT043, since the model quality was reliable, only top 20 scoring templates were selected and scanned using the local similarity search procedure.

In the enzyme library, no template protein had significant global ($G_{sim} > 0.7$) or local match ($L_{sim} > 0.9$) to the query structure. Therefore the confidence score of predicted EC numbers based on both global and local similarity are close to random (FC-score < 0.07), and suggest that CT043 is probably a non-enzymatic protein.

In the GO library, the CT043 model recognizes multiple template proteins with high global similarity; most of these proteins function as Type III secretion chaperone in different gram-negative pathogenic bacteria. However, none of these template protein structures are solved with a bound ligand and therefore for local similarity comparisons functional sites were predicted. The model finds significant local matches ($L_{sim} >1.2$) using these local 3D-motifs of template proteins, indicating a similar constellation of conserved residues in both CT043 and these template proteins, which may or may not be involved in ligand binding.

Each template protein in the GO library is annotated with multiple functions (GO terms) and we could have adopted the simplest approach of transferring the function (GO terms) from the top ranked hits, as we did for EC number predictions. However, unlike EC numbers, which characterize only the catalytic domain of an enzyme, GO terms can have contributions from other domains as well. Therefore a simple annotation transfer can be erroneous. Reconciling consensus GO terms amongst the top hit is more appropriate. Also, when query protein is multi-domain and no template protein with similar domain combination is present in the library, identifying concurrence of function amongst the weak hits can still provide information about the common function contributed by similar domain (for example the frequently occurring PDZ domain which is involved in protein binding).

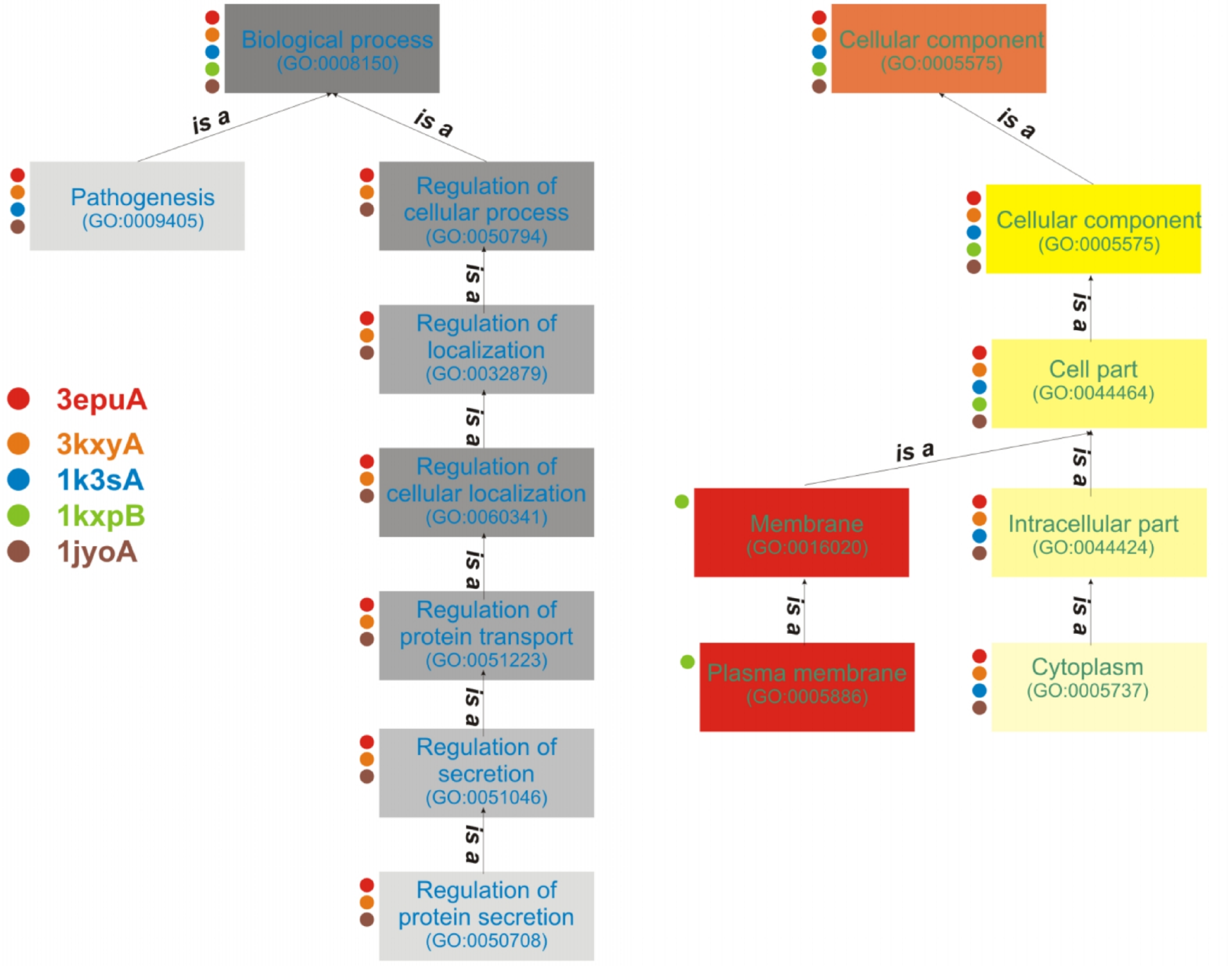


**Figure 4.2** Illustration of consensus GO term prediction for chlamydial protein CT043. Each box represents a function (GO term). Red boxes represent rejected functions during consensus prediction and colored dots beside each box represent a matched function, where the color corresponds to the template protein from which the GO term was contributed.

In our benchmarking training experiment, we observed that the top5 scoring hits identified by COFACTOR in the function library share the highest functional similarity (*Fsim,* Eq. 4.11) with the query protein. Therefore top5 template proteins (PDB ids: 3epuA, 3kxyA, 1k3sA, 1xkpB, 1jyoA) with highest global-local match were selected to identify the common function amongst these hits. Figure 4.2 shows an illustration of the procedure to identify most frequently represented and most likely functions for CT043 using these hits. Four of the five template proteins are involved in pathogenesis and are present in the cytoplasm; while three of them are involved in regulation of protein secretion. Outlier amongst these high scoring hits is 1xkpB, which is a synN chaperone in *Yersinia pestis* and is present in both membrane and cytoplasm. Consensus prediction therefore eliminated the membrane GO term (red box in Figure 4.2); while all the other three GO terms, namely: pathogenesis (GO:0009405), regulation of protein secretion (GO:0050708) and cytoplasm (GO:0005886) are retained by a voting procedure (see section 4.1.5) and predicted with high confidence.

### 4.2.2 Function predictions in benchmarking experiment

#### 4.2.2.1 Analysis of predicted EC numbers

We first analyze the EC number predictions using direct annotation (EC number) transfer from the identified remote homologs. The prediction results are obtained on a benchmark set of 368 non-homologous enzymes. EC number predictions by COFACTOR use I-TASSER models generated after removing template proteins with > 30% sequence identity to the target query protein. As our experimental controls, we use state-of-the-art methods for homolog detection, namely: profile-sequence alignment (PSI-BLAST[2]), profile-profile alignment (MUSTER[34]) and HMM-HMM alignment (HHsearch[35]). All programs searched through the same template library (EC library described in SI text) for predicting the function. To test the ability of different approaches, we also filter out template proteins that have >30% sequence identity to the target protein, because in this work the focus is only on function prediction using remote homologs.

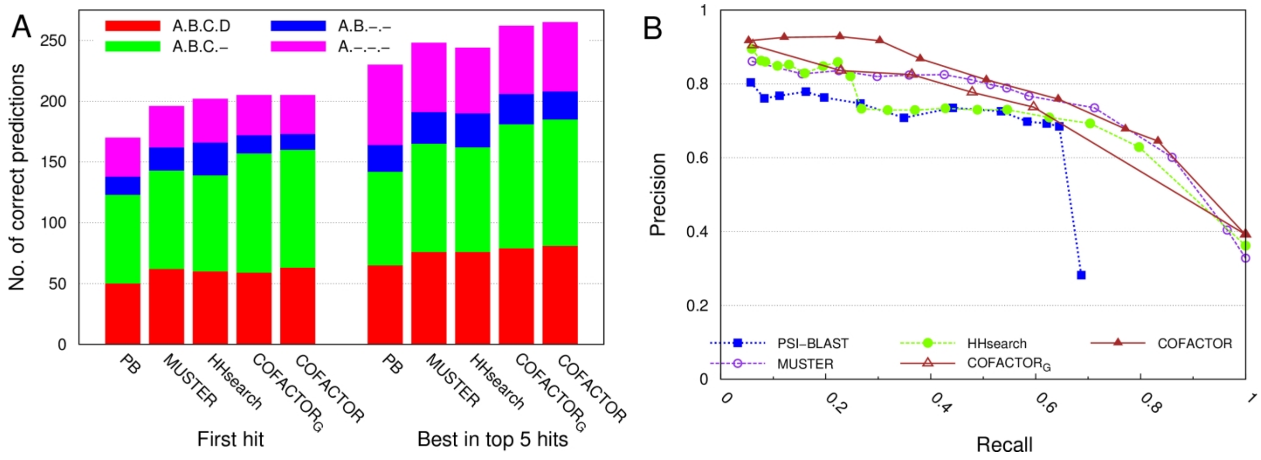


**Figure 4.3** Performance comparisons for EC number prediction. (A) Histogram analysis of functional inferences drawn for benchmarking enzymatic proteins at different level of Enzyme Commission number. (B) Precision-Recall analysis for predicting first three digits of EC number.

Figure 4.3A shows the histogram comparison of EC number predictions generated by each method using their first and best in top 5 hits. If we consider identity of first 3 digits of EC number as a criteria to evaluate the correctness of prediction, functional annotation was transferred correctly from the top hit of COFACTOR in 160 test cases, which is approximately 40%, 12% and 15% higher than the results obtained using the top hit of PSI-BLAST (123), MUSTER (143) and HHsearch (139), respectively. If we leave alone the local similarity search, and use results only from global search (COFACTOR$_G$), the prediction coverage is only slightly (5%) worse (153 hits). This data suggests that by combining global structural similarity with sequence and evolutionary profile (COFACTOR$_G$), we had already the upper bound of available template proteins in the library that have similar global topology and function (EC number). For the best in top 5 hits, COFACTOR still generates higher number (185) of correct predictions and also shows the largest improvement (25 new correct predictions). For the control methods PSI-BLAST, MUSTER and HHsearch, the best in top 5 hit provides correct functional inference for 142, 165 and 162 proteins, respectively; that are 30%, 12% and 14% lower than those obtained using COFACTOR. COFACTOR$_G$ on the other hand has 8% lower prediction coverage. Despite this improvement, overall coverage of the best prediction still appears to be low as the method failed to identify correct enzyme commission numbers in nearly half of the test cases. We analyze this by taking the best possible prediction from all the control methods and by running COFACTOR using experimentally determined structure of the 368 test proteins. After removing all template proteins with > 30% sequence identity to the query protein, as we did in this benchmarking experiment, the maximum prediction coverage achieved is 55%. Thus, the coverage of 50% obtained by COFACTOR using predicted protein structure represents a near-optimum annotation using the available template library and truly reflects a real-world scenario.

Next, we plot the precision-recall graph (Figure 4.3B), to analyze the ability of each methods score to identify correct function. In a precision-recall graph, an improved prediction method would produce a curve closer to the top-right corner. Figure3B shows that predicted EC numbers with higher score by all methods are more likely to be correct. Nevertheless, COFACTOR shows a striking performance and maintains a high precision across full recall range. When homologs are easily detectable by most methods (recall rate is < 0.3), COFACTOR consistently generates predictions with precision > 0.90, which is much higher than all the control methods. This improvement in precision is due to functional promiscuity in homologous proteins that can be distinguished by COFACTOR, by evaluating both binding pocket similarity and similarity of residue constellation involved in catalysis. This can be emphasized by analyzing the precision (< 0.9) of COFACTOR$_G$, which only uses global similarity, and fails to capture the functional promiscuity of the fold. At a recall rate of 0.5, the precision of COFACTOR, HHsearch, MUSTER and PSI-BLAST are 0.81, 0.73, 0.80 and 0.72 respectively. More importantly, for the same precision of 0.73, the recall rates of COFACTOR, HHsearch, MUSTER and PSI-BLAST are 0.72, 0.54, 0.71 and 0.44 respectively. This tells us that structure-based functional annotations are of greatest importance in the twilight zone where sequence based functional annotation becomes difficult.

#### 4.2.2.2 Analysis of predicted GO terms

Gene Ontology[29] (GO) is a widely used vocabulary for describing three different taxonomies or "aspects" of gene functions: molecular function (MF), biological process (BP) and cellular component (CC). Each GO aspect is represented as a structured directed acyclic graph (DAG), where nodes in the graph represent a GO term and describe a component of gene product function, while the edges between the nodes are equivalent to the relationships (*is-a* or *part-of*) between the GO terms. The GO terms are held in a form of functional hierarchy, where functions that are more general are present on the top while functions that are more specific are further down the graph.

Similar to EC number predictions, here we first assess the performance of different methods to identify proteins with similar function in the same template library. Functional similarity (*Fsim*; defined using Eq. 4.11) is evaluated based on annotated GO terms of query protein and annotated GO terms for the identified template proteins. Table 4.1 shows the average *Fsim* values for the predicted GO terms using the best and the best in top5 hit for the 337 tested proteins by COFACTOR and other control methods. After removing close homologs from the template library using a sequence identity cutoff of 30%, the average *Fsim* values of the top hit identified by COFACTOR, COFACTOR$_G$, HHsearch, MUSTER and PSI-BLAST are 0.51, 0.48, 0.45, 0.49 and 0.32, respectively. While using the best template in top5, the average *Fsim* values of all the methods significantly improve to 0.60, 0.58, 0.56, 0.59 and 0.37 respectively, suggesting that all the methods had a difficulty in selecting the best template amongst the top hits. We further investigate this for different aspects (MF, BP and CC) of GO term. As shown in the table, the average *Fsim* values for all the methods are highest for molecular function aspect, followed by biological process and least with cellular component. Nevertheless, in all these aspects COFACTOR has highest *Fsim* values and also in all the cases, the local similarity comparisons helped in selecting functionally more similar template proteins.

**Table 4.1. GO-term annotation coverage (*Fsim*) using identified hits by different approaches.**

| | **Overall** | | **Molecular function** | | **Biological Process** | | **Cellular component** | |
|---|---|---|---|---|---|---|---|---|
| **Method** | Top | Best | Top | Best | Top | Best | Top | Best |
| PSI-BLAST | 0.32 | 0.37 | 0.42 | 0.47 | 0.34 | 0.40 | 0.28 | 0.32 |
| MUSTER | 0.49 | 0.59 | 0.55 | 0.64 | 0.47 | 0.55 | 0.41 | 0.48 |
| HHsearch | 0.45 | 0.56 | 0.52 | 0.62 | 0.45 | 0.52 | 0.40 | 0.48 |
| $\text{COFACTOR}_G$ | 0.48 | 0.58 | 0.55 | 0.65 | 0.48 | 0.55 | 0.41 | 0.50 |
| **COFACTOR** | 0.51 | 0.60 | 0.59 | 0.67 | 0.49 | 0.56 | 0.42 | 0.51 |

### *Consensus GO term predictions*

The general practice for GO term prediction is to identify evolutionary relatives and transfer the GO terms. However, as we observed for CT043 (in Figure 4.2) even close homologs can have discrete functions, therefore simply copying GO terms would propagate a lot of false positive annotations. Finding common ancestral GO terms on the GO DAG are more likely to provide correct annotation and would result in improved precision. However, this approach sacrifices a lot of functional information, as more general GO terms would be predicted.

In COFACTOR, we first pool the annotated GO terms from top 5 templates and then assign each GO term a confidence score equal to the average FC-score of the template proteins with which these terms are associated (Eq. 4.7). We also propagate these confidence scores to the ancestor GO terms (Eq. 4.8) and finally pick those terms with GOscore exceeding a certain threshold (GOscore > 0.3). Once a deep level GO term is selected, all its ancestor GO terms in the DAG are progressively eliminated irrespective of their high GOscores. This procedure helps to balance both precision and recall. Moreover in PDB, GO term annotations for many proteins are incomplete, for example in Fig 4.2 template protein 1k3sA is only annotated for pathogenesis. Therefore, taking a consensus from the identified hits would also improve the coverage of GO term predictions.

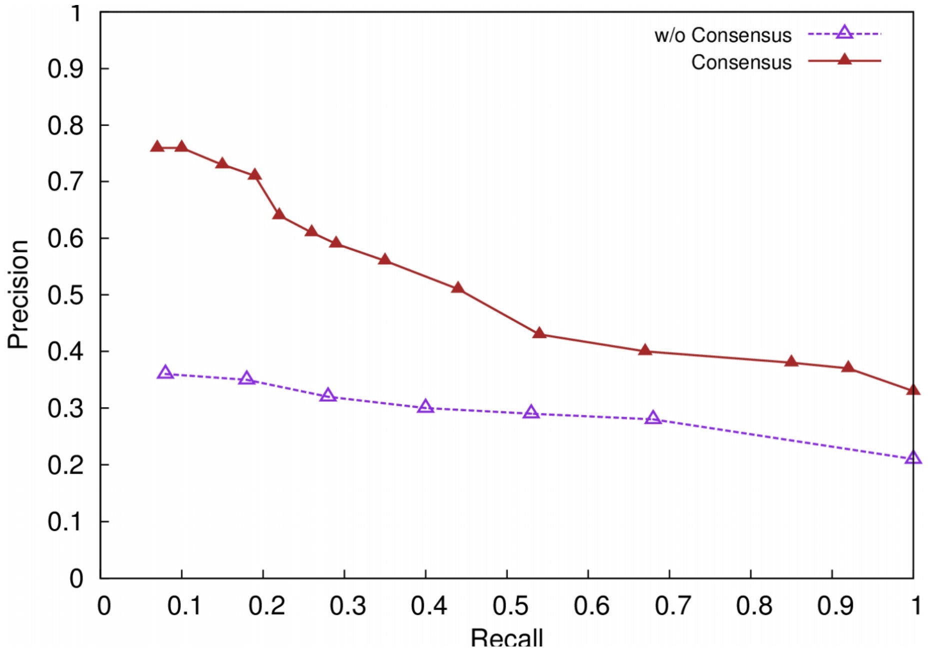


**Figure 4.4** GO annotations with and without the consensus approach are compared using precision recall graph. When the consensus is not used, each data point is obtained by selecting a different cutoff of FC-score, while for consensus predictions different cutoff of GOscore was used.

Figure 4.4 shows the precision-recall plot for functional annotation transfer using top 5 hits identified by COFACTOR and using the consensus approach. To make sure that we don't count predictions that are higher up in the DAG and therefore less specific, we defined true positive hit only when the predicted and annotated GO term of the query had a semantic similarity >0.5. The graph clearly shows that consensus GOscore outperforms the simple annotation transfer with a significantly high precision and recall. For example, at the recall rate 0.53, GOscore has a precision of 0.43, which is 48% higher than that by simple annotation transfer (29%). The results clearly suggest that consensus approach can effectively identify the concurrence of function among the selected hits to improve the precision of GO annotation.

### 4.2.3 Application to *Chlamydia trachomatis* proteome

Bacteria belonging to the *Chlamydia* genus are implicated in a large number of human diseases, including those with ocular and genital tract manifestations. Important to this study, these obligate intracellular organisms are phylogenetically distinct[54] and diverged relatively early, over a billion years ago[55]. Due to their characteristic developmental cycle that contains numerous selective stages for successful propagation, the evolution of proteins and their divergence has been severely constrained[56]. Furthermore, largely due to the sequestered nature of *Chlamydia* (metabolically active forms are contained within a *Chlamydia* specific vacuole inside a eukaryotic host cell), genetic exchange with other organisms has been limited. As a result of these factors, accurate functional assignment of chlamydial proteins using primary sequence homology has been limited.

To provide a practical assessment and expected application of the described method for structure and function predictions, 17 proteins from *C. trachomatis* were analyzed. While there is a relative paucity of structures available for chlamydial proteins, structure for seven proteins (CT243, CT296, CT381, CT390, CT610, CT780 and CT828) had already been determined, while the function has been experimentally characterized for five of these proteins. The structure and function for all the 17 proteins was modeled after excluding known protein structures from the same genus (*Chlamydia*) as the initial template. Figure 4.5 shows the predicted structures of these seven proteins with their predicted functional sites and evaluated structural similarity to experimentally determined structure.

CT243 is a multi-domain protein (2-domains), with a N-terminal UDP binding domain (residue 1-97) and C-terminal domain (residues 98-252) that acts as a UDP-3-O-glucosamine N-acyltransferase[57; 58] (EC: 2.3.1.-) for biosynthesizing lipid poly-saccharides (LPS). The domain boundaries can be easily recognized based on threading alignment and therefore we split this protein into its individual domains. The C-terminal catalytic domain is modeled with a C-score of −0.20 using I-TASSER. A search through the Enzyme can easily recognize H247 as the active site residue and the correct function of this protein. It is noteworthy that UDP-N-acetylglucosamine acyltransferase (PDB id: 2jf2) has a higher structural similarity score (TM-score=0.88) than UDP-3-O-glucosamine N-acyltransferase (PDB Id: 3eho), even though both templates possess the same catalytic residue. COFACTOR can differentiate between the function of two proteins based on local similarity and preferentially rank 3eho as the closest functional homolog in PDB library.

The crystal structure of CT296 was solved recently and it was shown that the *ab initio* I-TASSER model was predicted with a $C_\alpha$ RMSD of 2.72 Å (for 101/137 residues) to the experimentally determined form (PDB id: 3qh6A). Although both CT296 model and native structure have a structural scaffold similar to non-heme Fe(II) 2-oxoglutarate enzymes, the authors observed that key enzymatic residues were not conserved in CT296, suggesting a unique biochemical process is likely associated with CT296 function. We used the same model and ran COFACTOR predictions. The only reliable prediction was for the GO-term prediction for

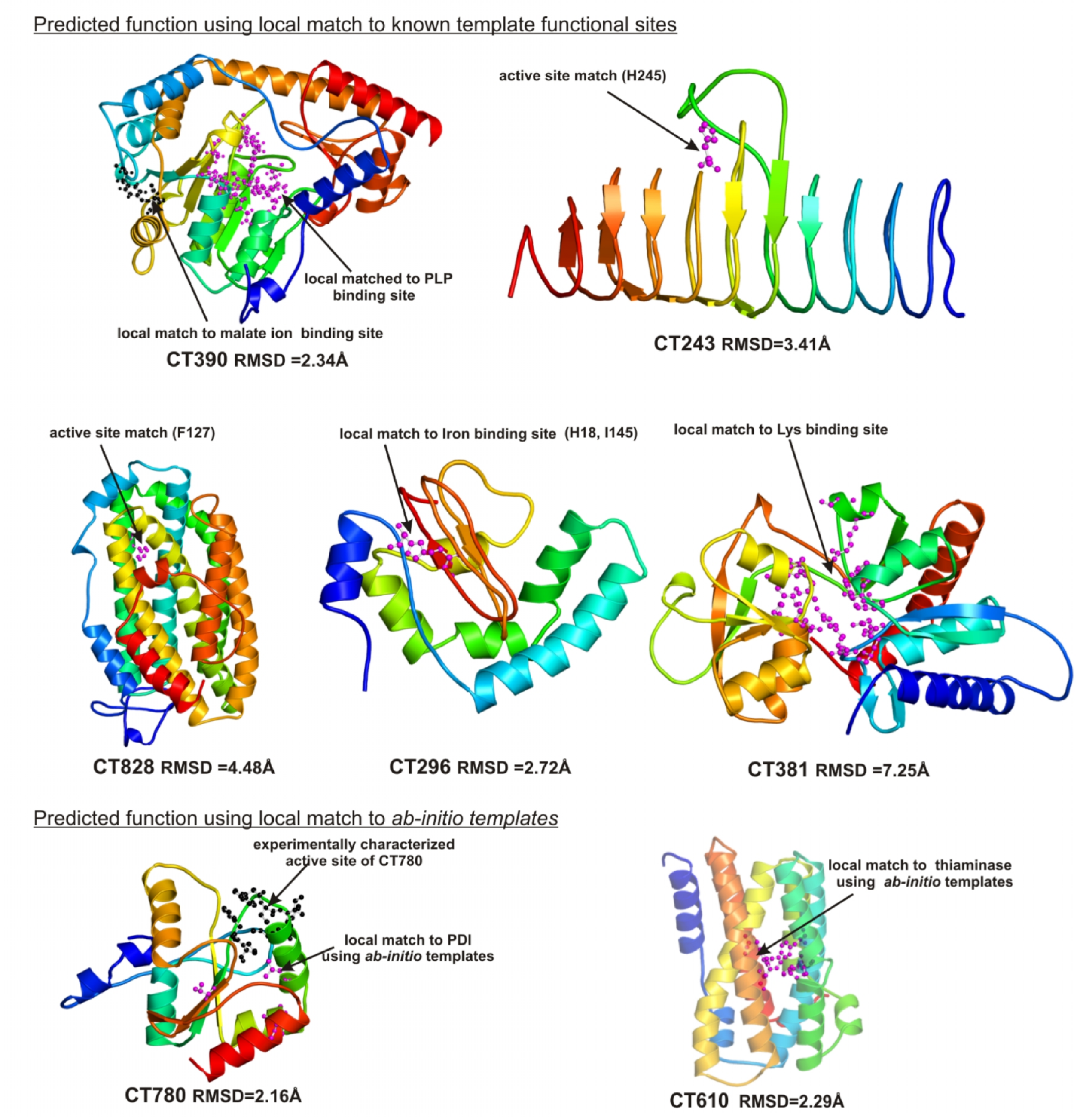

Predicted function using local match to known template functional sites
local matched to PLP binding site
local match to malate ion binding site
CT390 RMSD =2.34Å
active site match (H245)
CT243 RMSD=3.41Å
active site match (F127)
local match to Iron binding site (H18, I145)
local match to Lys binding site
CT828 RMSD =4.48Å
CT296 RMSD =2.72Å
CT381 RMSD =7.25Å
Predicted function using local match to *ab-initio templates*
experimentally characterized active site of CT780
local match to PDI using *ab-initio* templates
CT780 RMSD=2.16Å
local match to thiaminase using *ab-initio* templates
CT610 RMSD=2.29Å

**Figure 4.5** Benchmarking of structure and function prediction for chlamydial protein using I-TASSER model and COFACTOR algorithm. All these seven proteins have been already solved in PDB. COFACTOR identifies local matches between predicted model and template proteins in the library using both known and predicted functional sites of template. These matched functional sites were used for inferring the function and are shown in ball and stick.

oxidation-reduction process. Although none of the EC-number predictions are reliable they all have 1.14.-.-, which also suggests that CT296 acts on paired donors, with incorporation or reduction of molecular oxygen. We further analyzed the functional site matches and found that CT296 possesed a weak local similarity to di-iron binding site of these template proteins near His18, I145 and can be used for further analysis on this protein.

CT381 is an arginine binding protein with LAO (Lysine-, Arginine-, Ornithine-) domain. In our EC template library Cyclohexadienyl dehydratase (PDB id: 3kbr) has a very similar (TM-score = 0.70) to the modeled structure of this protein. However, COFACTOR can distinguish that the protein is non-enzymatic because both sequence profile and the local comparisons don't find a good match, and has low confidence score (FC-score <0.07) of EC number prediction. Search within the GO library, retrieves multiple template proteins with this domain, suggesting that this domain is more commonly utilized amongst non-enzymatic proteins. These template proteins bind to a wide variety of charged amino acids: namely Arg, Gln, His and Ornithine and are present in the periplasmic region. A consensus GO prediction can easily identify that all these proteins have a common function of amino-acid binding, are transporters and are present in the periplasmic space; and generate these predictions for CT381.

CT390 is II-diaminopimelate aminotransferase, which is used by Chlamydia to bypass three enzymatic steps in usual lysine biosynthesis pathway[59]. COFACTOR can easily identify the closest homolog in the PDB library (PDB id: 2z20) and can predict the correct EC number. Local comparisons with the templates identified in the GO library reveal an identical match to PLP and malate ion binding site and consensus GO term predictions suggest that CT390 can also bind cooper ion. However, we did not find any close homolog crystallized with copper ion, suggesting that this annotation for the template protein and CT390 might be incorrect.

The exact function of CT610 is still unknown, however the solution structure of CT610 (PDB id: 1rcw) and experimental investigation has revealed that the protein modulates host cell apoptosis and has active site similar to methane mono-oxygenase hydrolase[60]. We computationally generated a reliable structural model of this protein with C-score of 0.49 and used it for both EC and GO term predictions. Although, the fold (heme-oxygenase) of this protein is functionally promiscuous, COFACTOR identifies a near-perfect local match to TenA homologs (PDB ids: 2rd3 and 2qcx) and generates high confidence prediction that CT610 act as thiamine hydrolase (EC: 3.5.99.2) and is a transcriptional regulator. It is noteworthy that active site residues for both these template proteins are unknown and COFACTOR used *ab-initio* templates to find this local match and generate these predictions. The predicted GO term with highest confidence also suggests that CT610 may possess thiaminase activity (GO:0050334). While recent patent (Patent ID: 7736898) claims that thiamine and thiaminase genes can be used for inducing apoptosis in vivo by reducing the level of thiamin in the cell, our predictions using CT610 provides testable hypothesis for understanding the underlying mechanism of apoptosis caused by this protein.

Experimental structure of CT780 from *C. pneumoniae* has been determined (PDB id: 2ju5) and biochemical characterization has shown that this protein is a protein disulphide isomerase[61]. We modeled the structure of CT780 from *C.trachomatis* without using this homologous structure as template and the resultant model has a relatively lower confidence score (C-score = −1.04). The reason for this low confidence structure prediction is because this protein has a large threading unaligned N-terminal loop of 42 residues. During the structure modeling simulations, this region possesses large conformational freedom and therefore the sampled conformations remained diverged; thus lowering the structure density of the clusters and C-score of the predicted model. Corresponding to the C-score of the model, the functional prediction score was also low, although the best identified template protein (PDBid: 2bjx) has protein disulphide isomerase activity (EC: 5.3.4.1). Similar to the case of CT610, there was no known functional site information available for the template protein and COFACTOR used *ab-initio* generated signature fragments of template and found a good local match. Analysis of this match, reveals that even though these matches were near the active site (Figure 4.5), none of the known active site residues in CT780 locally aligned with the template fragments, suggesting that even though these residues are spatially and evolutionarily conserved they might have a structural or other functional role in this protein. Gene Ontology molecular function predictions (electron carrier activity and protein disulphide oxidoreducatase activity) for this protein is also in accord with the known function.

CT828 encodes the R2 subunit of ribonucleotide reductase (RNR), that is essential for synthesizing deoxyribonucleotides for DNA replication and is virtually ubiquitous in cellular organisms [62]. All class I RNR enzymes have the same EC number (1.17.4.1), however CT828 forms a distinct subclass as it lacks a tyrosyl radical site which is required for catalysis. The predicted EC number for CT828 using the I-TASSER model (C-score =0.61) finds top 5 enzyme homologs with RNR activity and can be used for predicting the correct function. However, the confidence score of prediction using top two templates (PDB ids: 1biq and 3ee4) is much higher. A detailed analysis of the prediction scores reveals that these top two ranked templates have an identical active site residue (F127 in CT828), while other templates (PDB ids: 2rcc, 1jku and 1h0n) have a tyrosine at the same position and therefore get lower scores. This example illustrates that COFACTOR predictions can be highly specific if accurate active site residue constellations are known for the template proteins. In combination, these observations demonstrate the accuracy and ability of this method to predict protein structure and subsequent function.

To begin addressing the challenge of functional assignments for Chlamydial proteins and establish a benchmark for future applications of this method for *genome-wide* functional annotation, we modeled all the *hypothetical* proteins from the *Chlamydia trachomatis* genome. Here, we only present results for 17 medium size proteins (7 with known structure and 10 are hypothetical) for which COFACTOR provides high confidence functional annotations and serves as a representative example from our genome-scale analysis. Table 4.2 provides a summary for the predicted structure and function for all the 17 proteins undertaken in this study. Experimental analyses for some of these proteins are already available and match with our predictions; for the rest, the predicted active site and binding-site residues can be used as testable hypothesis for further mutational analysis and better understanding of this organism.

**Table 4.2** Predicted function for Chlamydia trachomatis ORFs using I-TASSER model and COFACTOR algorithm.

| Name | Mod. C-score | Detected closest homolog in PDB | Seq. Id (%) | FN. conf. score | Predicted function using COFACTOR | Predicted functional residues | Annotated function OR (Ref) |
|---|---|---|---|---|---|---|---|
| Chlamydial proteins with *known structure* | | | | | | | |
| CT243 | -0.20 | 3eh0:A | 19.5 | 0.35 | UDP-3-O-glucosamine acyltransferase | H247 | UDP-3-O-glucosamine N-acyltransferase |
| CT296 | -2.43 | 2a1x:A | 27.0 | 0.06 | Phatanoyl-CoA 2-hydroxylase | H18,I145 | *Hypothetical protein* |
| CT381 | -0.66 | 2y7i:A | 25.7 | 0.93,<br>0.65,<br>0.93 | Transporter activity, Amino acid transport, Outer membrane bounded periplasmic space | N41,T43,Y44,E48,F82,T99, G100,M101,S102,R107,Q147, T150,Y151,Q152,P189 | Amino acid transporter |
| CT390 | 0.91 | 2z20:A | 42.0 | 0.44 | LL-diaminopimelate aminotransferase | G109,A110,K111,Y134, C176,N180, D208,Y211, S239,K242; Y77,N281 | LL-diaminopimelate aminotransferase |
| CT610 | 0.49 | 2rd3:D | 21.6 | 0.21 | Pyrroloquinoline-quinone synthase | Y47,H50,I51,F54, Y141,F214 | Pyrroloquinoline-quinone biosynthesis |
| CT780 | -1.04 | 2bjx:A | 20.7 | 0.99 | Disulfide isomerase | - | Disulfide isomerase |
| CT828 | 0.61 | 1biq:B | 24.0 | 0.38 | Ribonucleotide reductase | F127 | Ribonucleotide reductase |
| Chlamydial proteins with *unknown structure and function* | | | | | | | |
| CT043 | -1.05 | 3epu:A | 18.6 | 0.77<br>0.67 | Pathogenesis<br>Reg. of protein secretion | F13,L17,L19,P20,L32, D35,Y54,I123,E124 | *Hypothetical protein*<br>Ref: [52] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CT047 | 0.46 | 1jr3:D | 21.7 | 0.81; 0.71; 0.70 | Protein binding; DNA-directed polymerase activity; DNA clamp loader activity | R218,F250 | *Hypothetical protein* |
| CT077 | -0.25 | 3pnd:C | 28.2 | 0.71; 0.71; 0.30 | Thiamine biosynthetic process; plasma-membrane; Calcium binding; FAD binding | M33,I35,R112,I117,K121, L126,D170,K176, R241,H245,D254,T286; T174,D282,T286 | *Hypothetical protein* |
| CT263 | 0.36 | 1odi:A | 22.4 | 0.29 | Purine nucleoside phosphorylase | S160 | *Hypothetical protein* |
| CT309 | -0.38 | 1r5z:A | 19.9 | 0.15 | V-type ATPase | D118,F119 | *Hypothetical protein* |
| CT349 | 1.06 | 1ex2:A | 31.6 | 0.69, 0.69 | Nucleotide binding, metal ion binding | S10,R14,D31,E34, K54,K84 | Maf-like protein |
| CT355 | -1.38 | 3gpk:A | 9.3 | 0.44 | peptidyl-prolyl cis-trans isomerase | V199,K203,S238, S282,K285 | *Hypothetical protein* |
| CT373 | -0.33 | 1n2m:C | 20.0 | 0.22 | Pyruvoyl-dependent arginine decarboxylase | E120 | *Hypothetical protein* |
| CT663 | 0.19 | 3epu:A | 24.1 | 0.93 0.81 | Pathogenesis, regulation of protein secretion | Y116,Y200 | Hypothetical protein |
| CT867 | -2.84 | 1euv:A | 16.2 | 0.10 | ULP1 protease | W92,H203,W204, D220,C282 | *Hypothetical protein*[51] |

Although functional annotation for a large fraction (68%) of the Chlamydia trachomatis ORFs has been inferred using sequence based comparisons[63], Table 2 for the first time provides many new functional assignments using the predicted 3D structure. For example, CT043 has been experimentally characterized to act as a molecular chaperone. Our analysis provided a molecular three-dimensional structure for this protein, which is also a promising candidate for vaccine design. COFACTOR predicts CT043 to be a virulence chaperone with high confidence because ab-initio templates find a good local match between CT043 model and the template proteins. However, delineating exact binding site residue based on these matches can be erroneous (as we observed for CT780), therefore we filtered out these local matches by analyzing the protein-protein interaction between homologous protein ExsC-ExsE (PDBid: 3kxyA) and found that (F13, L17, L19, P20, L32, D35, Y54, I123 and E124) are likely to be important for peptide binding and can be targeted in future functional studies on this protein.

## 4.3 Conclusion

In this work, we describe a new approach for deducing the biological function (EC number and GO terms) of protein molecules using predicted protein structures and global-local structural match to solved structures with known function. A robust approach that can handle both functional promiscuity of fold and also provide correct functional annotation for multi-domain protein was presented.

Benchmarking experiment on a comprehensive benchmark set of 450 proteins annotated with (EC) numbers and gene ontology (GO) terms shows significant advantages of the structure-based function inference over conventional sequence-based predictions. The method outperforms commonly used homology based approaches for functional inferences by generating predictions with higher precision and recall, and with vast improvement in prediction coverage. For EC number predictions, the method correctly predicts the function for 50% of the test proteins, where the best possible annotation using the existing template library is nearly 55%. For GO term prediction, the average functional similarity (*Fsim*) of molecular function was 0.67, which was 43% better than most routinely used sequence based method (PSI-BLAST) for functional inference.

As an illustrative application of the methodology for genome scale prediction, we first benchmarked the predictions for the 7 chlamydial proteins with already known structures. For 5 of these proteins that have already been experimentally characterized, COFACTOR predictions show a close agreement with known experimental annotations and for the other two proteins local structural match by COFACTOR provides new functional insights beyond the sequence-based annotations. We also present high confidence prediction results for 10 hypothetical proteins from *Chlamydia trachomatics* genome, which can be considered as testable hypothesis for improving our understanding about this bacterium.

Although the accuracy of the predictions show modest dependency on the global quality of structure predictions, the combination of the global and local structural searches help generate a reasonable level of predictions across the entire range of proteins studied, including those with low resolution structures. Overall, the data demonstrates great promise towards utilization of the current method for automated, genome-wide structural and functional annotations.

## Bibliography

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.
2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.
3. Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435-41.
4. Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**, 863-82.
5. Rost, B. (2002). Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595-608.
6. Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in genetics : TIG* **17**, 429-31.
7. Arakaki, A. K., Tian, W. & Skolnick, J. (2006). High precision multi-genome scale reannotation of enzyme function by EFICAz. *BMC Genomics* **7**, 315.
8. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. & Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and molecular life sciences : CMLS* **60**, 2637-50.
9. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**, 265-74.
10. Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* **31**, 400-2.
11. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* **28**, 228-30.
12. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823-6.
13. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current opinion in structural biology* **6**, 377-85.
14. Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignments and functional genomics. *Proteins* **42**, 378-82.
15. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94.
16. Roy, A., Srinivasan, N. & Gowri, V. S. (2009). Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biol* **9**, S41-55.

17. Watson, J. D., Laskowski, R. A. & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current opinion in structural biology* **15**, 275-84.

18. Gherardini, P. F. & Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. *Brief Funct Genomic Proteomic* **7**, 291-302.

19. Gherardini, P. F. & Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. *Briefings in functional genomics & proteomics* **7**, 291-302.

20. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J Mol Biol* **351**, 614-26.

21. Arakaki, A. K., Zhang, Y. & Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* **20**, 1087-96.

22. Stark, A., Sunyaev, S. & Russell, R. B. (2003). A model for statistical significance of local similarities in structure. *J Mol Biol* **326**, 1307-16.

23. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci* **5**, 1001-13.

24. Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A. & Kolchanov, N. A. (2005). PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Research* **33**, D183-7.

25. Kristensen, D. M., Ward, R. M., Lisewski, A. M., Erdin, S., Chen, B. Y., Fofanov, V. Y., Kimmel, M., Kavraki, L. E. & Lichtarge, O. (2008). Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **9**, 17.

26. Polacco, B. J. & Babbitt, P. C. (2006). Automated discovery of 3D motifs for protein function annotation. *Bioinformatics* **22**, 723-30.

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

28. (1999). Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur J Biochem* **264**, 610-50.

29. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

30. Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* **18**, 342-8.

31. Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* **77**, 100-113.

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

33. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.

34. Wu, S. & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547-56.

35. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-60.

36. Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* **37**, D396-403.

37. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends in genetics : TIG* **16**, 44-7.

38. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.

39. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.

40. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53.

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

42. Yu, C., Zavaljevski, N., Desai, V., Johnson, S., Stevens, F. J. & Reifman, J. (2008). The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics* **9**, 52.

43. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.

44. Martin, A. C. (2004). PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics* **20**, 986-8.

45. Porter, C. T., Bartlett, G. J. & Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **32**, D129-33.

46. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein science : a publication of the Protein Society* **15**, 2120-8.

47. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.

48. Pesquita, C., Faria, D., Falcao, A. O., Lord, P. & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol* **5**, e1000443.

49. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274-81.

50. Pruitt, K. D. & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* **29**, 137-40.

51. Misaghi, S., Balsara, Z. R., Catic, A., Spooner, E., Ploegh, H. L. & Starnbach, M. N. (2006). Chlamydia trachomatis-derived deubiquitinating enzymes in mammalian cells during infection. *Mol Microbiol* **61**, 142-50.

52. Meoni, E., Faenzi, E., Frigimelica, E., Zedda, L., Skibinski, D., Giovinazzi, S., Bonci, A., Petracca, R., Bartolini, E., Galli, G., Agnusdei, M., Nardelli, F., Buricchi, F., Norais, N., Ferlenghi, I., Donati, M., Cevenini, R., Finco, O., Grandi, G. & Grifantini, R. (2009). CT043, a protective antigen that induces a CD4+ Th1 response during Chlamydia trachomatis infection in mice and humans. *Infection and immunity* **77**, 4168-76.

53. Brinkworth, A. J., Malcolm, D. S., Pedrosa, A. T., Roguska, K., Shahbazian, S., Graham, J. E., Hayward, R. D. & Carabeo, R. A. (2011). Chlamydia trachomatis Slc1 is a type III secretion chaperone that enhances the translocation of its invasion effector substrate TARP. *Mol Microbiol* **82**, 131-44.

54. Weisburg, W. G., Hatch, T. P. & Woese, C. R. (1986). Eubacterial origin of chlamydiae. *J Bacteriol* **167**, 570-4.

55. Stephens, R. S. (2002). *In Chlamydial Infections, Proceedings of the Tenth International Symposium on Human Chlamydial Infections*, *Antalya, Turkey*.

56. Stephens, R. S., Myers, G., Eppinger, M. & Bavoil, P. M. (2009). Divergence without difference: phylogenetics and taxonomy of Chlamydia resolved. *FEMS Immunol Med Microbiol* **55**, 115-9.

57. Bartling, C. M. & Raetz, C. R. (2008). Steady-state kinetics and mechanism of LpxD, the N-acyltransferase of lipid A biosynthesis. *Biochemistry* **47**, 5290-302.

58. Buetow, L., Smith, T. K., Dawson, A., Fyffe, S. & Hunter, W. N. (2007). Structure and reactivity of LpxD, the N-acyltransferase of lipid A biosynthesis. *Proc Natl Acad Sci U S A* **104**, 4321-6.

59. Watanabe, N., Clay, M. D., van Belkum, M. J., Fan, C., Vederas, J. C. & James, M. N. (2011). The structure of LL-diaminopimelate aminotransferase from Chlamydia trachomatis: implications for its broad substrate specificity. *Journal of Molecular Biology* **411**, 649-60.

60. Schwarzenbacher, R., Stenner-Liewen, F., Liewen, H., Robinson, H., Yuan, H., Bossy-Wetzel, E., Reed, J. C. & Liddington, R. C. (2004). Structure of the Chlamydia protein CADD reveals a redox enzyme that modulates host cell apoptosis. *J Biol Chem* **279**, 29320-4.

61. Mac, T. T., von Hacht, A., Hung, K. C., Dutton, R. J., Boyd, D., Bardwell, J. C. & Ulmer, T. S. (2008). Insight into disulfide bond catalysis in Chlamydia from the structure and function of DsbH, a novel oxidoreductase. *J Biol Chem* **283**, 824-32.

62. Roshick, C., Iliffe-Lee, E. R. & McClarty, G. (2000). Cloning and characterization of ribonucleotide reductase from Chlamydia trachomatis. *J Biol Chem* **275**, 38111-9.

63. Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., Koonin, E. V. & Davis, R. W. (1998). Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science* **282**, 754-9.

# Chapter 5
# Conclusion and Outlook

In this dissertation, we have presented three bioinformatics softwares that have significantly improved the state-of-the-art computational techniques for structural and functional characterization of protein molecules. The main parts of this work are briefly summarized below.

## 5.1 Conclusions

### 5.1.1 Protein structure predictions using the I-TASSER server

Genome sequencing projects have ciphered millions of protein sequences, which require knowledge of their structure and function to improve the understanding of their biological role. Although experimental methods provide more detailed and reliable information for a small fraction of these proteins, computational modeling is needed for the majority of protein molecules that are experimentally uncharacterized. In chapter **2**, we developed a hierarchical approach for high-resolution modeling of protein structure using multiple threading alignment[1] and replica-exchange monte-carlo simulations[2]. Given a protein sequence, a typical output from the I-TASSER[3; 4; 5; 6] includes secondary structure prediction, predicted solvent accessibility of each residue, homologous template proteins detected by threading and structure alignments, up to five full-length tertiary structural models, and structure-based functional annotations: including Enzyme Commission (EC) numbers[7] and Gene Ontology (GO)[8] terms and functional sites (active and ligand binding sites) in the protein. All the predictions are tagged with a confidence score that signifies how accurate the predictions in the absence experimental data. To facilitate the special requests of end users, we developed channels to accept user-specified inter-residue distance and contact maps to interactively modify the structure modelling procedure; it also allows users to specify any proteins as template, or to exclude any template proteins from the structure assembly simulations. The structural information could be collected by the users based on experimental evidences or biological insights with the purpose of improving the quality of I-TASSER predictions. The server was evaluated as the best method for protein structure and function predictions in the recent community-wide CASP experiments

### 5.1.2 Detection of functional sites in protein using the COFACTOR algorithm

Proteins perform their functions by interacting with other molecules. However, structural details for most of the protein-ligand interactions and the location of functional sites are unknown. In chapter **3**, we presented a new comparative approach (COFACTOR) to infer protein-ligand binding site locations and interactions from known protein structures, based on an optimal global-to-local structural alignment procedure. The method was tested in both benchmark and blind tests, and has demonstrated significant improvements over the current state-of-the-art methods. In a large-scale benchmark test on 501 proteins harbouring 200 drug-like and 382 natural ligands, the method successfully identified ligand-binding pocket locations for 67% of apo receptors with an average distance error of 2 Å. The average accuracy of binding-residue assignments using this algorithm is 33-112% higher than the two best-performing methods in the field (FINDSITE[9] and ConCavity[10]). A detailed analysis of the results obtained in the benchmarking experiment highlighted that for 70% of the proteins with bound "natural" ligand, the predicted ligand by COFACTOR shared a high chemical similarity to the bound ligand in the experimentally determined structure, which suggests a potential application of the method for a more elaborate functional elucidation of uncharacterized proteins. Successful predictions were also observed for "drug-like" compounds, which open up the possibility for structure-based drug design even for proteins with no available structural information. In the recent community-wide CASP9 experiments, COFACTOR achieved a binding-site prediction accuracy of 72% and

Matthews correlation coefficient of 69% in recognizing ligand-binding residues, for both metal and non-metal ligands, and significantly outperformed all other state-of-the-art methods. The above data demonstrates the power of combining global-local structure search procedure for inferring function using predicted protein structures.

### 5.1.3 Prediction of EC number and Gene Ontology terms using COFACTOR

Although the definition of protein is context dependent, in chapter **4** we sought to improve the COFACTOR approach and predict the function of protein molecules. We use two standard vocabularies for describing the function, namely Gene Ontology and EC number. We added two new features to COFACTOR: first, a new global search algorithm was developed which evaluates both topological similarity and conservation of sequence profile; and second, we developed a method for constructing local 3D-signature motifs of template proteins that lack known functional sites. This allowed us to perform query-template local structural similarity comparisons for all template proteins. Benchmarking experiment on 450 non-homologous and functionally diverse proteins showed that the first three digits of EC numbers can be correctly assigned for 50% of the test proteins, where the best possible annotation using the existing template library is nearly 55%. For GO term prediction, the average functional similarity (*Fsim*) of molecular function was 0.67, which was 43% better than routinely used sequence based method for functional inference. We also showed that by identifying the concurrence of function among the top5 hits, COFACTOR can generates predictions with 48% higher precision than by simply transferring the annotation from the identified hits. To explore the applicability of the method, we applied the method to a subset of ORFs from *Chlamydia trachomatis* and the function annotations suggest new insights into this phylogentically distinct bacterium.

## 5.2 Outlook and future directions

### 5.2.1 Structure modeling using I-TASSER

The automated I-TASSER pipeline for structure prediction described in this work performs very well for most proteins. Although the reassembly of structural fragments excised from the threading alignments often results in significantly improved models, the quality of final models is essentially still dependent on identified threading templates. In general, I-TASSER works very well when appropriate template proteins are detected, while *ab initio* folding (e.g. QUARK[11], ROSETTA[12]) generates better models when there is a lack of good templates. However, correct determination of target type (TBM or FM) is critical for choosing the appropriate methodology for structure modeling, especially in the weak-homology modeling region, where reasonable templates are available but their threading alignment scores are low. During CASP9 experiments, we observed that combining threading alignment score (Z-score) and structural similarity score (i.e. average pair-wise TM-score) between the templates identified by different threading programs provides a more accurate way for classifying the targets.[13] A more detailed study is needed to parameterize the weights of these measures and add new rules that can aid in improving the coverage of accurate protein structure modeling.

Domain splitting is another long-standing issue. I-TASSER currently uses threading-alignment to automatically determine the domain boundaries and splits them into individual domains. Although analysis of threading alignment to identify domain boundaries is a powerful method, for the extremely difficult targets, iterative threading might become necessary. For example in CASP9, T0529 was one such example, where the I-TASSER server (as “Zhang server”) infers incorrect domain boundary because most of the LOMETS programs generate weak alignments for the entire query sequence.

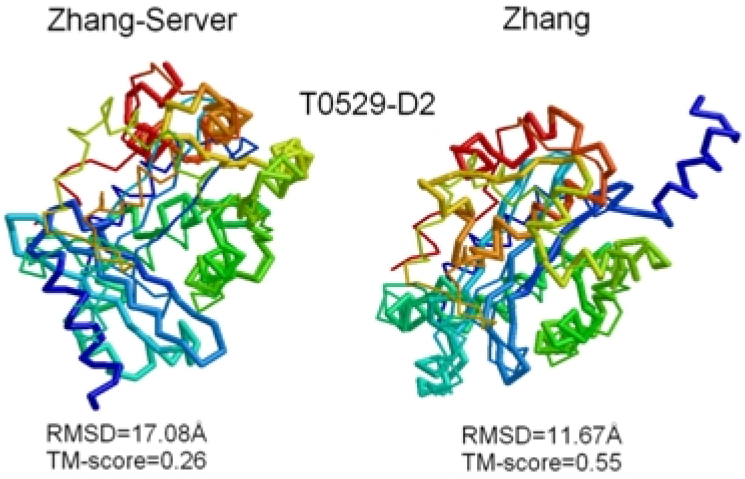


**Figure 5.1** Example of structure modeling for T0529-D2 by automated I-TASSER server as "Zhang-Server" (Left) and human group as "Zhang" (Right). Models (thick backbone) are superimposed on the native structure (thin backbone) with blue to red running from N- to C-terminal.

In the human prediction, region 378L-569L (T0529-D2) emerged as an independent domain with the alignments of a higher confident score in the second round of LOMETS when threading was run on the roughly split sequence based on the first round; this eventually results in the correct detection of the C-terminal domain, a template-based domain target.

Model selection is another classic issue. I-TASSER has the advantage in refining the models that are on average significantly better than the initial threading templates; mainly attributed to the use of consensus spatial restraints collected from multiple templates. However, I-TASSER sometimes fails to select the best model as the first model, when only minority of threading programs detect the best template and the majority of the threading alignments *consistently* hit an incorrect template. Here, the second condition is essential for I-TASSER's failure in selecting the best template, while it was observed that I-TASSER could often pick up the best template, if the templates by the majority of the threading programs are diverged. When the majority of the threading programs hit a common (incorrect or second best) template, the consensus restraints can be too strong and distract the template selection of the I-TASSER modeling.

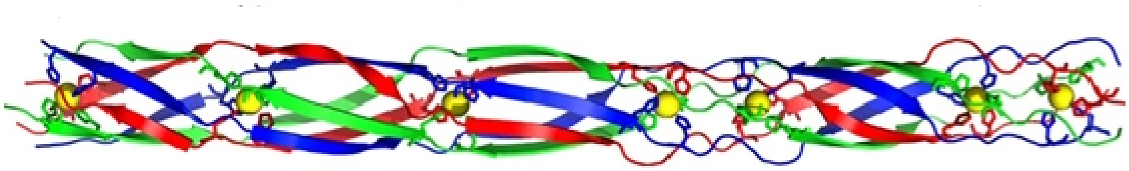


**Figure 5.2** Crystal structure of T0629-D2 which forms the needle domain of bacteriophage T4 long tail fiber protein . Iron ions are represented as yellow balls. The histidine doublets coordinating the iron ions are shown in sticks. This figure has been taken from Bartual *et al.*[14]

Several proteins are solved in their quaternary structure form. For example in CASP9, T0629-D2 was a long tail fiber protein from bacteriophage T4, with three identical protein chains intertwined together to form an elongated six-stranded antiparallel beta-strand structure. The core of this structure is stabilized by the alternate hydrophilic and hydrophobic regions, where the hydrophilic residues form coordination site for seven iron ions. All the structure modeling methods failed to generate a reasonable structure for this domain, highlighting the need to extend the current tertiary structure modeling method for quaternary structure modeling, to model these complex protein structures.

## 5.2.2 Function prediction using COFACTOR

During the benchmarking experiments of COFACTOR (Chapter **3),** we observed that the algorithm can identify the constellation of functional residues with very high accuracy. However like I-TASSER, the success of COFACTOR is also template library dependent. In chapter **4**, we tried to overcome this problem by predicting functional sites of template and screening them against the query structure, to identify the best local match. Although this approach helps to identify remote homologs with similar function, because homologous proteins still share similar spatially located evolutionary conserved residues, the Matthew's correlation coefficient (MCC) for predicting known ligand binding site residues is ~30%. We constructed these *ab-initio* local signature motifs for the template protein using very simple rules, as observed in known functional sites. Adding new features using machine-learning methods like SVM would likely improved the binding site prediction even in the absence of homologous co-crystallized protein-ligand complexes.

Another important aspect is the curation of already known protein-ligand complexes in PDB library. Development of automated machine learning methods or a discriminatory function that can distinguish "real ligand" and "crystallization artifacts" would remove many false positive predictions. Finally, due to the fact that definition of protein "function" is context dependent, different research groups have primarily focused on different aspects of protein function and compiled the structure-function libraries independently. This leads to technical difficulties for COFACTOR, which can in principle predict all three aspects of functions using the same common procedure using a universal approach. Unfortunately, since the libraries are not integrated, COFACTOR needs to separately search through three different libraries to accurately predict all three aspects. Hypothetically however, since the PDB already represents a common universal set, creation of an integrated library encompassing all three aspects is technically possible.

Many proteins perform their physiological function by interaction with DNA and/or RNA, and constitute a very important aspect of biological function. During this study, we mainly focused on methodological development of COFACTOR and testing it to predict small molecule binding sites and enzyme active sites. The same approach was never applied to identify nucleic acid binding site. Extending the COFACTOR algorithm towards accurately predicting nucleic acid binding sites represents a critical advancement, which can have great significance in the field.

One of the technical drawbacks of the COFACTOR algorithm is that the search engine is based on Needleman Wunsch dynamic programming algorithm[15]. The efficacy of the search is therefore dependent on the conservation of sequence order in the query and the template-binding site. While in a majority of cases this may be true, proteins can share similar binding site because of convergent evolution. Thus, it is imperative to add sequence order independent search engines like geometric hashing to further improve the COFACTOR algorithm.

Nevertheless, even in the current form all these developed softwares can be used for genome scale applications and improve our understanding of the bioverse.

## Bibliography

1. Wu, S. & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* **35**, 3375-82.
2. Zhang, Y., Kihara, D. & Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48**, 192-201.
3. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.
4. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17.
5. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69 Suppl 8**, 108-17.
6. Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40.
7. (1999). Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur J Biochem* **264**, 610-50.
8. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.
9. Brylinski, M. & Skolnick, J. (2009). FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput Biol* **5**, e1000405.
10. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* **5**, e1000585.
11. Xu, D. & Y, Z. (2011). QUARK Ab Intio Protein Structure Prediction. *Submitted*.
12. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for ab initio protein structural genomics. *Journal of Molecular Biology* **306**, 1191-1199.
13. Zhang, Y. (2009). Protein structure prediction: when is it useful? *Curr Opin Struct Biol* **19**, 145-55.

14. Bartual, S. G., Otero, J. M., Garcia-Doval, C., Llamas-Saiz, A. L., Kahn, R., Fox, G. C. & van Raaij, M. J. (2010). Structure of the bacteriophage T4 long tail fiber receptor-binding tip. *Proc Natl Acad Sci U S A* **107**, 20287-92.

15. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53.

# List of Publications

## List of publications related to this dissertation work

1. **Roy, A**., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-738

2. **Roy, A**. & Zhang, Y. (2011). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. (Submitted)

3. **Roy, A**. Mukherjee. S, Hefty, P.S. & Zhang, Y. (2011). A global-local structure similarity approach for remote homolog detection and function prediction. (In preparation).

4. **Roy, A**., Yang. J.Y. & Zhang, Y. Structure based functional annotation for protein molecules using the COFACTOR server. (In preparation).

5. **Roy, A**. & Jiao. C & Zhang, Y. (2011). Structure and function modeling of Marek's disease virus proteome (In preparation).

6. **Roy, A.**, Wu, S. & Zhang, Y. Composite approaches to protein tertiary structure prediction: A case-study by I-TASSER. In *Protein Structure Methods and Algorithms* 14 edit. (Rangwala, H. & Karypis, G., eds.). Hoboken, N.J. : Wiley, 2010

7. **Roy, A**., Xu, D., Poisson, J & Zhang, Y. (2011). A protocol for computer-based protein structure and function prediction. JoVE. doi: 10.3791/3259

8. Xu, D., Zhang, J., **Roy, A.**, Zhang, Y. (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics*. doi: 10.1002/prot.23111

## List of publications not related to this work

1. **Roy, A.**, Srinivasan, N. & Gowri, V. S. (2009). Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biol* **9**, S41-55

2. Menon, R., **Roy, A.**, Mukerjee, S., Belkin, S., Zhang, Y., Omenn, G. (2011). Functional Implications of Structural Predictions for Alternative Splice Proteins Expressed in Her2/neu-Induced Breast Cancers. *Journal of Proteome Research* (PMID 22003824)

3. Li, Y., **Roy, A.** & Zhang, Y. (2009). HAAD: A Quick Algorithm for Accurate Prediction of Hydrogen Atoms in Protein Structures. *PLoS ONE* **4**, e6701

4. Rankin, C. A., **Roy, A.**, Zhang, Y. & Richter, M. L. (2011). Parkin, a top level manager in the cell's Sanitation Department. *Open Biochem J.* 2011;5:9-26.

5. Trivedi, M., Davis, R. A., Shabaik, Y., **Roy, A**., Verkhivker, G., Laurence, J. S., Middaugh, C. R. & Siahaan, T. J. (2009). The role of covalent dimerization on the physical and chemical stability of the EC1 domain of human E-cadherin. *J Pharm Sci* **98**, 3562-3574

6. Somarelli, J. A., Mesa, A., **Roy, A.**, Zhang, Y. & Herrera, R. J. (2010). A three-dimensional model of the U1 small nuclear ribonucleoprotein particle. *Entomological Research* **40**, 104-112

7. Mukherjee, S., Szilagyi, A., **Roy, A.** & Zhang, Y. Genome-wide protein structure prediction. In *Multiscale approaches to protein modeling* 11 edit. (Andrzej Kolinski editor). Springer-London, 2010