Protein Function Prediction Based on Sequence and Structure Information

Thesis by
Fatima Zohra Smaili

In Partial Fulfillment of the Requirements
For the Degree of
Master of Science

King Abdullah University of Science and Technology, Thuwal,
Kingdom of Saudi Arabia

(May, 2016)
ACKNOWLEDGMENTS

First and foremost, I thank Allah, the Almighty, for endowing me with courage and patience to carry on this project, and for blessing me with many great people who have been my greatest support.

I acknowledge with deep gratitude and appreciation, the inspiration, encouragement, valuable time and guidance given to me by my supervisor Professor Xin Gao. I am deeply grateful to him for his continuous support and personal involvement in all phases of this project.

Warm thanks are due to Professor Yang Zhang and his lab members at Zhang Lab in the Department of Computational Medicine and Bioinformatics at University of Michigan, for giving me the opportunity of an internship in this prestigious lab to work on this project and for providing me with priceless help and advice.

Special thanks to my committee members, Professor Vladimir Bajic and Professor Stefan Arolld for accepting to review this thesis.

Finally, I would like to express my deepest gratitude to my mother, father and sister for their love, patience and unconditional support throughout all the phases of my academic career.
TABLE OF CONTENTS

ACKNOWLEDGMENTS .......................................................................................................................... 1
LIST OF FIGURES ................................................................................................................................. 3
LIST OF TABLES ................................................................................................................................. 4
ABSTRACT ............................................................................................................................................. 5
1. INTRODUCTION ............................................................................................................................ 6
2. MATERIALS AND METHOD .......................................................................................................... 9
  2.1. I-TASSER protein structure prediction .................................................................................... 9
  2.2. Detection of functional homologous templates ....................................................................... 10
    2.2.1. Global similarity search .................................................................................................... 10
    2.2.2. Local similarity search .................................................................................................... 13
  2.3. Scoring function for functional annotation ............................................................................... 16
    2.3.1. Enzyme commission number ............................................................................................ 16
    2.3.2. Gene ontology terms ........................................................................................................ 17
  2.4. Template libraries ..................................................................................................................... 18
    2.4.1. Global enzyme template library ...................................................................................... 18
    2.4.2. Local enzyme template library ....................................................................................... 19
    2.4.3. GO template library ......................................................................................................... 20
  2.5. Extending the search with protein-protein interaction information ........................................ 20
  2.6. Assessment of protein function predictions ............................................................................ 21
3. RESULTS ............................................................................................................................................ 26
  3.1. Data set ..................................................................................................................................... 26
  3.2. Prediction of enzyme functions .............................................................................................. 27
    3.2.1. Example of dihydrolipicolinate reductase ......................................................................... 27
    3.2.2. Example of staphopain .................................................................................................. 29
    3.2.3. Overall analysis of enzyme function prediction ............................................................... 30
  3.3. Prediction of gene ontology terms ........................................................................................... 31
    3.3.1. Example of calmodulin .................................................................................................... 32
    3.3.2. Overall analysis of GO terms prediction: .......................................................................... 34
4. CONCLUSION ................................................................................................................................. 40
5. REFERENCES: ................................................................................................................................. 41
LIST OF FIGURES

Figure 1: A schematic diagram of the local similarity search procedure for functional site identification. The residues of the query protein (yellow) in the active site region is shown in magenta, while those of the template protein (grey) is shown in blue [32]. .............................................................................................................................. 16

Figure 2: DAG for GO term Intracellular Membrane-bound Organelle: 0043231 [51]. .......................... 23

Figure 3: Representative examples of structure-based function inference using high- and low-resolution structure predictions. Left: native query protein structure and its active sites; Middle: I-TASSER model; Right: The best structural analog with highest score identified from the function prediction libraries and sequence identity to query sequence. (a) DHPR; (b) Staphopain. ................................................................. 30

Figure 4: Precision-recall curve to compare the EC prediction performance of the EC-score vs LOMETS and HHSEARCH. ......................................................................................................................... 31

Figure 5: Structure-based GO term inference for calmodulin protein. Left: native structure; Middle: I-TASSER model; Right: Best structural analog identified from function libraries obtained using the I-TASSER model as query, and sequence identity to query protein sequence. ........................................................................ 32

Figure 6: An illustrative example of molecular function term predictions for calmodulin (CaM) protein. GO-score calculation for two molecular function terms: calcium binding (GO: 0005509) and protein dimerization activity (GO: 0046983). ................................................................................................................................. 33

Figure 7: Precision-recall curve to compare the GO prediction performance of GO-score vs LOMETS and HHSEARCH. ............................................................................................................................................. 34

Figure 8: Precision-recall curve to compare the performance of HHSEARCH combined with the GO-score to regular HHSEARCH................................................................................................................... 35

Figure 9: Precision-recall curve to compare the performance of LOMETS combined with the GO-score to regular LOMETS. ................................................................................................................... 36

Figure 10: Precision-recall curve to compare the performance of our GO-score method that used I-TASSER to HHSEARCH combined with the GO-score and LOMETS combined with the GO-score................. 37

Figure 11: Precision-recall curve comparing the GO-score method to GO-score combined with protein-protein interaction information from STRING. .................................................................................................................... 38
LIST OF TABLES

Table 1: S-values for GO terms in DAG for term Intracellular Membrane-bound Organelle: 0043231…24
The number of available protein sequences in public databases is increasing exponentially. However, a significant fraction of these sequences lack functional annotation which is essential to our understanding of how biological systems and processes operate. In this master thesis project, we worked on inferring protein functions based on the primary protein sequence. In the approach we follow, 3D models are first constructed using I-TASSER. Functions are then deduced by structurally matching these predicted models, using global and local similarities, through three independent enzyme commission (EC) and gene ontology (GO) function libraries. The method was tested on 250 “hard” proteins, which lack homologous templates in both structure and function libraries. The results show that this method outperforms the conventional prediction methods based on sequence similarity or threading. Additionally, our method could be improved even further by incorporating protein-protein interaction information. Overall, the method we use provides an efficient approach for automated functional annotation of non-homologous proteins, starting from their sequence.
1. INTRODUCTION

Over 60 million protein sequences have been deposited in the UniProtKB/TrEMBL [1] nowadays. However, this increase in the amount of protein sequences available in the databases does not reflect a parallel increase in our biological knowledge as most of these sequences lack functional annotation. In fact, the functional annotation of these sequences is necessary to understand physiological processes and systems in living entities. At the same time, it is one of the most challenging tasks in modern molecular and cell biology [2]. Experiment-based annotation is the most reliable way to functionally annotate proteins, but as experiments are usually expensive and time consuming, there is an increasing need to provide reliable and fast automated protein function annotation. In this regard, significant effort has been made to identify evolutionary relatives and transfer functional annotations between homologous protein pairs. More specifically, powerful sequence-alignment methodologies have been used to make such sequence similarity based functional annotation transfers possible. In particular algorithms like BLAST/PSI-BLAST [3, 4] and hidden-Markov model (HMM) based techniques [5, 6, 7] have been frequently used to transfer functional annotations between homologous proteins [8, 9]. The underlying assumption of these sequence based methods is that proteins with similar sequences fold similarly [10, 11] and as structural similarity induces functional similarity, the detection of evolutionary relationships between proteins is used to predict functions of non-annotated protein sequences [12, 13]. It has been estimated that by using such approaches, functional inferences can be drawn for nearly 40-60% of the open reading frames (ORFs) in a given genome [14]. However, there are numerous cases in which proteins with relatively close sequences do not fold similarly, and therefore
annotations cannot be transferred reliably using sequence based method [8, 9]. Several methods have been developed to partially address the problem by identifying locally conserved motifs that are assumed to be responsible for the functional aspect of the protein, namely Prosite [15], PRINT [16] and Blocks [17]. Unfortunately, since the patterns in these databases are generated using multiple sequence alignments, these sequence motifs can be used only to search for conserved elements within the same protein family and carry little information about functionally conserved residues across different protein families.

The tertiary structure of a protein is believed to be responsible for its biological function [18, 19]. In fact, as residues located far apart in primary sequence may be close in 3D space, and only a few spatially conserved residues are generally responsible for a protein’s function [20, 21], the three dimensional (3D) structure of proteins can provide a better insight into which protein fragments contribute most to the functionality of a protein compared to the primary sequence. This fact is the motivation behind the increase in the throughput of experimental structure elucidation [22, 23, 24] and provide more reliable methods to transfer annotations and infer molecular functions [25, 26]. The resultant increase in the number of known structure in the PDB [27] has motivated the interest into developing many approaches which aim to identify the protein functions from the structure.

Most of these protein structure-based function prediction methods involve comparing the global fold of a query protein with other template proteins of known function as proteins with similar folds are believed to share the same functions. However, the relationship between protein structure and function is not always so straightforward, as many diverse folds are known to perform the same function while in many cases similar folds perform different functions [28, 29]. Therefore, many prediction methods have been relying on local structure similarity search methods
rather than global similarity search to identify functional homologous proteins. Most of these approaches scan the query protein against a library of known conserved spatial motifs or known active sites (e.g. binding sites) with known function [30]. Local similarity search methods have been proven to be capable of detecting functional similarity between proteins of different folds but they also have probability of producing false positive matches [31].

In my thesis, I have worked on improving an existing protein function prediction method, COFACTOR [32] that combines both global and local similarity search methods in order to get the best of the two approaches and address their respective issues. In this thesis work, we follow a sequence-to-structure-to-function paradigm. Starting from the protein amino acid sequence, we first generate structure predictions by the iterative threading assembly refinement method (I-TASSER) [33], which are then used for identifying the functional homologues using a combination of global and local structure similarity searching methods. We then further improve the prediction by including protein-protein interaction information extracted from the STRING database [34].

Since the terminology of a “protein function” could have multiple definitions which may lead to some ambiguity, I would like to clarify that the definition of function followed in this work is enzyme commission (EC) numbers [35] and gene ontology (GO) terms [36]. EC numbers are used to categorize enzymes into hierarchical families using a numerical classification. Specifically, the EC number (which is composed of four numbers separated by periods i.e. A.B.C.D) refers to the reaction catalyzed by a specific enzyme. That is to say, two enzymes share the same number if they catalyze the same reaction [35]. On the other hand, the gene ontology (GO) is a database that provides a set of terminologies to describe proteins and other types of gene products based on their functions within a cell [36]. Three ontologies, Biological Process, Cellular Component and Molecular Function are defined in this database. Each one of these three GO aspects 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 more general functions are present on the top while more specific functions are further down the graph.

Our prediction results have been mainly compared to LOMETS [37], a mega-threading algorithm, and HHSEARCH [6], an HMM based method used to detect sequence similarity. HHSEARCH and LOMETS are representative of the commonly used methods for protein function inference.

2. MATERIALS AND METHOD

2.1. I-TASSER protein structure prediction

The first step of protein function prediction is the generation of the predicted 3D model of the query protein using I-TASSER. A detailed description of the I-TASSER methodology can be found at [33], but for the sake of completion, I would like to include a brief outline of the methodology.

Starting from a query protein amino acid sequence, I-TASSER first threads the sequence through a representative PDB27 structure library using LOMETS [37], a locally installed meta-threading server combining 7 state-of-the-art threading programs (FUGUE [38], HHSEARCH [6], MUSTER [39], PROSPECT [40], PPA [33], SP3 [41] and SPARKS [42]). The top template hits from each threading program are then selected for further consideration, based on the statistical significance \((Z\text{-score})\) of the threading alignments. Continuous fragments from the selected templates are then excised and used in the assembly of full-length structural models while
threading unaligned regions (mainly loops and tail regions) are built by ab initio modeling [33, 43]. In the fragment assembly phase, numerous conformations are generated based on a modified replica-exchange Monte Carlo simulation [44] using knowledge-based statistical energy terms derived from known protein structures [43] and spatial restraint collected from LOMETS [37]. The conformations with low temperatures are clustered with the purpose of identifying the best low-free energy conformations. Cluster centroids are then obtained by averaging the 3D coordinates of all structural decoys in each cluster. Fragment assembly simulations are performed again starting from the cluster centroids. The purpose of this second iteration is to remove steric clashes as well as to refine the global topology of the cluster centroids. The decoys generated during this second round of simulations are clustered again. The lowest free energy structure from the largest cluster is selected as input for REMO [45] which generates the final full-atomic models from the Cα traces via the optimization of hydrogen-bonding networks.

2.2. Detection of functional homologous templates

The predicted model of the query protein obtained from I-TASSER is scanned against two template libraries of protein structures; the first one is a library of enzymes with known enzyme commission (EC) numbers and the second one is a library of proteins annotated with at least one Gene Ontology (GO) terms. The homologous proteins to the query protein in these two libraries are detected using two types of structure similarity search programs: global similarity search and local similarity search. More details about these two search approaches are described below:

2.2.1. Global similarity search

Templates with a similar global structure to the predicted structure of the query protein are searched across the template libraries using TM-align [46]. TM-align is a widely used protein
structure alignment program which starts from a range of initial alignment, which are then refined based on heuristic iterations of the Needleman-Wunsch dynamic programming approach [47]. For a given pair of structures the best alignment in TM-align is identified by the TM-score:

$$\text{TM - score} = \max \left[ \frac{1}{L} \sum_{i=1}^{L_{ali}} \frac{1}{1 + \left( \frac{d_{ij}}{d_0} \right)^2} \right],$$

where $L$ is the length of the query protein and $L_{ali}$ is the number of the aligned residue pairs between the template and the query structure. Here, $d_{ij}$ is the distance between C$_\alpha$ atoms of aligned residues $i$ and $j$ after superposition of the structures, and the distance scale $d_0 (L)$ is given by $d_0 = 1.24^{3\sqrt{L-15}} - 1.8$. The major advantage of TM-score over the often-used root mean squared deviation (RMSD) in assessing structural similarity is that TM-score accounts for both the structure similarity in the aligned regions and the coverage of the alignment in a single parameter. In addition, even when alignments with same coverage are evaluated, TM-score is more sensitive to the global topology of the structures than RMSD, because it down-weights large distances between aligned C$_\alpha$ pairs compared to the smaller ones. For RMSD comparisons, all distances are taken into account with an equal weight, 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 may be similar.

Another important parameter we take into consideration when searching for templates with similar global fold to the query protein is the quality of the structural models. Appraising the accuracy of the structure modeling in the scoring scheme helps to reduce the number of false positive predictions. In this particular case, the quality of the predicted I-TASSER model generated in the previous step is evaluated using the C-score, which is defined as:
\[ C\text{-score} = \ln \left[ \frac{M}{M_{\text{tot}}} \times \frac{1}{\text{RMSD}} \times \sum_{i=1}^{N} \frac{Z(i)}{Z_0(i)} \right] \]

where \( M \) is the multiplicity of structure decoys in the SPICKER cluster, \( M_{\text{tot}} \) is the total number of decoys submitted for the SPICKER clustering, \( \text{RMSD} \) 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 
\[ \text{C-score norm} = \frac{\text{C-score} + 5}{7}. \]

The final scoring function for assessing functional similarity between the query and the template protein based in the global similarity search is defined as:

\[ \text{Fh-score} = C\text{-score} \times \left( \text{TM-score} + \frac{1}{1 + \text{RMSD}_{\text{ali}}} \times \text{Cov} \right) + 3 \times \text{ID}_{\text{ali}} \times \text{Cov}, \]

where \( C\text{-score} \) estimates the quality of the I-TASSER protein structure prediction, \( \text{TM-score} \) measures the global structural similarity between the model and the template proteins, \( \text{RMSD}_{\text{ali}} \) is the RMSD between the model and the template structure in the structurally aligned region, \( \text{Cov} \) represents the coverage of the structural alignment, and \( \text{ID}_{\text{ali}} \) is the sequence identity between the query and the template based on the alignment generated by TM-align. Note that I did not design this scoring function myself but I am using the same scoring function designed for COFACTOR [32] previously.
2.2.2 Local similarity search

Although the global structural similarity search can be used for recognizing protein pairs with
similar global folds, in many cases proteins with the same fold can have diverse functions. In these
cases, functional similarity between protein pairs arises only due to the conservation of local active
site residues. This is the motivation behind using local structural search as a complementary
measure to the global search, for a more reliable detection of functional homologues.

The local structural search procedure, which is the same used procedure followed in
COFACTOR [32], includes three steps. The first step is the structural match of the specific
catalytic/active residue pairs. For a given pair of query and template proteins, we first scan the
known catalytic/active residues of the template through the query sequence. The residues whose
amino acid type is similar to that of the template’s catalytic/active residues are marked as potential
active site locations in the query. The structures of all combined sets of marked residues in the
query are excised from the predicted model and used as a candidate active site. The structure of
the candidate site is superimposed on the structure of the known catalytic/active residues in the
template. To increase the reliability of the structure superimposition, for each residue $i$, the
coordinates of C\(_\alpha\) atoms and the side-chain centers of mass of two neighboring residues, i.e. $i$-1
and $i+1$th residues are also included in the superimposition.

The second step is to identify the key local environment residues around the active sites in
the query and the template. For this purpose, we superimpose the complete structure of the query
and template proteins based on the rotation matrix acquired from the superimposition of the
candidate catalytic/active residue structures. A sphere of radius $r$ is then defined around the
geometric center of the template’s local 3D fragments, where $r$ is the maximum distance of the
template residues in the local 3D fragment from the geometric center. The sphere here represents
a local environment or probable active site region, under which the query and template’s chemical and structural similarity are compared. Because a sphere comprising of a very small number of catalytic/active residues can easily generate false positive hits, when the template’s active site region is small, $r$ is gradually incremented by 0.5 Å, until the number of residues inside the sphere is $>30$. Similarly, when the template’s active site region is very large (for example in multifunctional enzymes, which have more than one active site), it is difficult to find a match between the query and template using a very large sphere, so the maximum value of $r$ is set to 20Å.

In the third step, the best alignment of the local active site residues in the spheres between the query and the template is identified using a heuristic procedure similar to that used in TM-align [46]. Starting from the initial superposition of query and template protein structures, a Needleman-Wunsch dynamic programming [47] is performed to generate the best alignment for the selected sphere areas of query and template, where the alignment score matrix $S_{ij}$ for aligning the $i$th residue in the query and the $j$th residue in the template is given by

$$S_{ij} = \left[ \frac{1}{1 + \frac{d_{ij}}{d_0}} + M_{ij} \right],$$

where $d_{ij}$ is the $C_α$ distance between residues $i$ and $j$, $d_0$ is the distance cutoff, chosen to be 3.0 Å, $M_{ij}$ is the substitution score between the $i$th and the $j$th residues taken from the BLOSUM mutation matrix with the value normalized by the diagonal element in the mutation matrix. The gap penalty is set as -1. For a given scoring matrix $S_{ij}$, a new alignment will be generated by dynamic programming. Then a new superposition and scoring matrix will be constructed based on the new alignment, 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 active site match (AcM):

$$\text{AcM} = \frac{1}{N_t} \sum_{i=1}^{N_{ali}} \frac{1}{1 + \left( \frac{d_{ii}}{d_0} \right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{ali}} M_{ii},$$

where $N_t$ represents the number of residues present in the active site sphere of the template and $N_{ali}$ is the number of aligned residue pairs. The maximum AcM score encountered during the heuristic iterations is recorded for each candidate active site. Finally, the candidate active site with the highest AcM score is selected for evaluating the similarity between the query and the template’s active site. The three steps conducted in the local search method are summarized in Figure 1[32] below:
2.3. Scoring function for functional annotation

2.3.1. Enzyme commission number

The prediction of enzyme functions, i.e. EC numbers, involves combining local and global search methods. The final score, the EC-score, used to sort the hits from the enzyme library is a combination of the Fh-score (obtained from the global similarity search) and the AcM score.
(obtained from the local similarity search). However, the local similarity search result is taken into consideration only when the predicted active site is a reliable, i.e. AcM score > 1.1. Thus the EC-score is defined as:

\[
EC\text{-score} = \begin{cases} 
C\text{-score} \times \left[ \text{TM-score} + \left( \frac{1}{1 + \text{RMSD}_{\text{ali}}} \times \text{Cov} \right) \right] + 2 \times \text{ID}_{\text{ali}} \times \text{Cov} + \frac{\text{AcM}}{2} & \text{if } \text{AcM} > 1.1 \\
C\text{-score} \times \left[ \text{TM-score} + \left( \frac{1}{1 + \text{RMSD}_{\text{ali}}} \times \text{Cov} \right) \right] + 3 \times \text{ID}_{\text{ali}} \times \text{Cov}, & \text{if } \text{AcM} < 1.1
\end{cases}
\]

2.3.2. Gene ontology terms

For the time being, the functionally important residues for most of the proteins in the GO template library are unknown. Therefore, only the global similarity search is taken into consideration when sorting the hits from the GO library. However, since each single protein is usually associated with multiple GO terms and the global search may result in many close template structures, a query protein can have multiple GO term predictions with high Fh-scores. Therefore, the confidence score of each predicted GO term is calculated as follows:

\[
\text{GO-score}(\lambda) = \frac{1}{N} \sum_{i=1}^{N_{\lambda}} \text{Fh-score}_i(\lambda),
\]

where \(\lambda\) represents a given gene ontology term, \(N_{\lambda}\) is the number of templates which are associated with the GO term \(\lambda\), and \(N\) is the total number of templates selected for generating the consensus. When multiple close templates are available, we only consider the templates with an Fh-score >1. For those query proteins with less than 10 templates of Fh-score >1, the top 10 templates are selected for generating the consensus prediction regardless of the Fh-score. Also, given the
The hierarchical nature of the GO DAG, we consider that when a template protein is annotated with a GO term, all its ancestor GO terms are automatically implied. Therefore, once a GO term $\lambda$ is scored, we score all its ancestor terms as well. The score of any ancestor GO term $\mu$ of term $\lambda$ is calculated as:

$$GO\text{-}\text{score}(\mu) = GO\text{-}\text{score}(\lambda) \times \left(1 + \frac{N_\mu}{N_0}\right)$$

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

Subsequently, all GO terms are sorted based on their depth in the GO DAG and GO terms with GO-score $>0.5$ are predicted. Once a GO term is predicted, all its ancestors are automatically eliminated from this prediction list because they are automatically implied. If this procedure does not yield any confident GO term prediction, the GO terms are ranked based on their GO-scores and the top 10 GO term predictions are reported.

2.4. Template libraries
The libraries used for the functional template search are the same ones used by COFACTOR [32]. To represent the typical functional space, three independent libraries of protein structures with known biological function have been created. The two global template libraries comprise of proteins that have an association with EC number(s) or GO specific terms, while the local template library consists of enzymes with known active site residues.

2.4.1. Global enzyme template library
With the purpose of annotating enzyme function based on global structure match, a library of global enzyme structures has been created based on the structures in the PDB [27] that are annotated with an EC number. Since many proteins in the PDB library represent only single
domain fragments of multiple domain proteins, automatic annotations using the SwissProt mapping are erroneous in many cases. Accordingly, proteins that have multiple EC numbers after automatic annotation have been manually verified and corrected in the library. Finally the protein chains having the same enzyme commission (EC) number and >90% sequence identity to other enzymes in the library are excluded to avoid redundancy and to reduce the search space.

For the time being, the compiled enzyme library contained 7674 protein chains with 203 unique first 3-digit EC numbers and 1,900 unique 4-digit EC numbers.

### 2.4.2. Local enzyme template library

Functional similarity in evolutionarily diverged enzymes or enzymes belonging to different structural families arises mainly due to the spatial conservation of only a few catalytically important residues. For example, in case of aspartic proteases, proteins having the same function can be recognized based on a local structural motif that consists of eight conserved residues in the active site of the enzyme [48]. However, the number of proteins or protein families for which such discerning functional insight is available is limited and is presently catalogued in the Catalytic Site Atlas (CSA) [49].

The local 3D template library was initially generated by extracting all the literature derived non-redundant protein entries from CSA, which presently contains 909 proteins covering 171 unique 3-digit EC numbers and 753 unique 4-digit EC numbers. To increase the library size, we used the developed local structure alignment program to identify the conserved active site residues in the global enzyme template library, where the proteins in CSA were used as the initial templates. Only reliable hits with an AcM score >1.1 and aligned catalytic residues with larger than >50% identity are added to the local template library, producing a total of 2,385 non-
redundant proteins. It is obvious that the set of proteins in this local library is a subset of the set of protein in the enzyme global library discussed above.

### 2.4.3. GO template library

The gene ontology (GO) is currently the most effective approach for machine-legible and automatic functional annotation. To this end, a third 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 the GO2PDB mapping at the Jena library (http://www.imb-jena.de/IMAGE.html). Protein chains that are associated with the same GO terms and have >90% sequence identity with other proteins in the library are excluded. Currently, this library contains 13,027 unique protein chains. Additionally, multiple domain proteins have been split into individual domains using an automated domain parser program [50]. The split domains are associated with the same GO terms as their parent protein. Both single-domain and multiple-domain structures are kept in the library for the convenience of structural search, since the query can be a single- or multiple-domain protein. Overall, there are 26,004 entries in this library associated with 11,686 unique GO terms.

### 2.5. Extending the search with protein-protein interaction information

The combination of global and local similarity search approach described above to make protein function prediction gives satisfying results. As will be shown in the results section, the method performs much better than both HHSEARCH and LOMETS which are the leading methods in sequence similarity and threading approaches respectively. However, there is always room for improvement if we use more information sources. In particular, protein-protein interaction information can be useful in predicting GO terms. In fact, proteins in a living cell interact
continuously with each other. Each group of interacting proteins participate in the same biological process and are located in the same cellular component. This is why protein-protein interaction information can be very useful in improving the prediction of GO terms especially Cellular Component and Biological Process terms. In this regard, we exploited the information provided by STRING [34] database which is a protein interaction network to extend our prediction set. Given a query protein, we collect the set of directly interacting proteins with it from STRING. Then, we calculate the prediction score of STRING for a GO term \( \lambda \) \( (P_{\text{STRING}}(\lambda)) \) as the frequency of the GO term \( \lambda \) among the annotated interaction partners of the query proteins:

\[
P_{\text{STRING}}(\lambda) = \frac{\text{Number of partners annotated with } \lambda}{\text{Total number of GO terms among the partners}}
\]

We then combine this score obtained from STRING with the GO-score introduced previously to calculate the final confidence score of each GO term \( \lambda \) as the consensus of the two scores:

\[
\text{Final Score}(\lambda) = 1 - [1 - \text{GO-score}(\lambda)] \times [1 - P_{\text{STRING}}(\lambda)]
\]

2.6. Assessment of protein function predictions
For the EC number predictions, we first evaluate the functional similarity between the template proteins and the query protein at different levels of enzyme commission nomenclature. A precision-recall analysis of the predicted functions using identified templates and by varying their Fh-score thresholds is then performed. 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 the serial number of the enzyme.
For assessing the GO term prediction we make use of the GO term semantic similarity terminology introduced by Wang et al. in [51]. For the sake of completion, we would like to include a summary on how the method works. As mentioned earlier, the GO database defines three types of ontologies: Biological Process, Cellular Component and Molecular Function. Each one of these ontologies is represented as a DAG in which the terms (nodes) are linked through two kinds of semantic relations: ‘is-a’ is a simple subclass-class relation where A is-a B means that A is a subclass of B, and ‘part-of’ which is a partial ownership relation where C part-of D means that whenever C is present, it is always a part of D. The GO DAG features other kind of relations namely the ‘regulates/positively regulates/negatively regulates’ relations, but only ‘is-a’ and ‘part-of’ are considered as semantic relations that contributes to the definition of the semantics of a GO term, and therefore they are the only relations taken into consideration when we try to evaluate the semantic similarity between two terms.

To measure the semantic similarity of two GO terms, this method first encodes the semantics of one GO term in numeric format based on the location of the term in the GO graph and the semantic relations to its ancestors. To define the semantics of a GO term, they use the DAG starting from this specific term and ending at a root term (could be Biological Process, Cellular Component or Molecular Function depending on where the term is located) to represent this term. As an example, Figure 2 [51] below shows the DAG for GO term Intracellular Membrane-bound Organelle: 0043231.
Figure 2: DAG for GO term Intracellular Membrane-bound Organelle: 0043231 [51].

In this DAG, a dotted arrow represents the ‘part-of’ relation, while a solid arrow represents the ‘is-a’ relation. To put it in a formal way, a GO term $\lambda$ can be represented as $DAG_\lambda = (\lambda, T_\lambda, E_\lambda)$ where $T_\lambda$ is the set of GO terms in $DAG_\lambda$, including term $\lambda$ and all its ancestors in the GO graph up to the root, and $E_\lambda$ is the set of semantic relations (‘is-a’ and ‘part-of’ relations) connecting the GO terms in $DAG_\lambda$. To define the semantics of a GO term quantitatively, this method first defines the semantic contribution of a GO term $t$ to a GO term $\lambda$ for every term $t$ in $DAG_\lambda$ as the S-value of $t$ with respect to $\lambda$, $S_\lambda(t)$. This S-value is defined as:

$$
S_\lambda(\lambda) = 1
$$

$$
S_\lambda(t) = \max\{w_c \times S_\lambda(t') \mid t' \in \text{childrenof}(t)\} \text{ if } t \neq \lambda
$$
where $w_e$ is the semantic contribution of relation $e$. The semantic contribution for ‘is-a’ and ‘part-of’ relations are set to 0.8 and 0.6 respectively after parameter training. In $DAG_\lambda$, the contribution of term $\lambda$ to its own semantics is 1. As we go up in the $DAG_\lambda$, the terms become more and more general and thus contribute less and less to the semantic of the term $\lambda$. As an illustration, Table 1 [51] below summarizes the S-value of each term in the DAG corresponding to GO term Intracellular Membrane-bound Organelle: 0043231 that has been shown above:

<table>
<thead>
<tr>
<th>GO terms</th>
<th>0043231</th>
<th>0043229</th>
<th>0043227</th>
<th>0005622</th>
<th>0005623</th>
<th>0043226</th>
<th>0005575</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-value</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.48</td>
<td>0.288</td>
<td>0.64</td>
<td>0.512</td>
</tr>
</tbody>
</table>

Table 1: S-values for GO terms in DAG for term Intracellular Membrane-bound Organelle: 0043231.

The semantic value of a GO term $\lambda$, $SV(\lambda)$, is defined as the aggregation of the S-values of all terms in $DAG_\lambda$ with respect to $\lambda$. It is calculated as:

$$SV(\lambda) = \sum_{t \in T_\lambda} S_\lambda(t)$$

Now, given two terms A and B, such that $DAG_A = (A, T_A, E_A)$ and $DAG_B = (B, T_B, E_B)$ are the DAGs corresponding to terms A and B respectively, the semantic similarity between these two terms, $SS(A, B)$ is defined as:

$$SS(A, B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)}$$
where $S_A(t)$ is the S-value of GO term $t$ with respect to term A, and $S_B(t)$ is the S-value of GO term $t$ with respect to B.

However, in our protein GO function prediction method, one single protein could be annotated with several GO terms and we can predict a set of GO terms for each protein. Therefore, we do not only need to calculate the semantic similarity of two GO terms, but rather we need the semantic similarity of two sets of GO terms (the predicted set and the actual annotation set). Given the semantic similarity definition between two terms that is introduced in [51] and that has been discussed above, we evaluate functional similarity ($FSim$) of predicted GO terms $GO_1 = \{go_{11}, go_{12}, go_{13}, \ldots go_{1m}\}$ with the annotated GO terms of query protein $GO_2 = \{go_{21}, go_{22}, go_{23}, \ldots go_{2n}\}$ using the best match average score strategy[51], which is defined as:

$$FSim(GO_1, GO_2) = \frac{\sum_{1 \leq i \leq m} SS_{\text{max}}(go_{1i}, GO_2) + \sum_{1 \leq j \leq n} SS_{\text{max}}(go_{2j}, GO_1)}{m + n},$$

where $SS_{\text{max}}(go, GO)$ represents the maximum SS between term $go$ and any of the terms in the set $GO$. We consider a match between our predicted set $GO_1$ and the actual annotated set $GO_2$ when $FSim(GO_1, GO_2) > 0.5$. 
3. RESULTS

3.1. Data set
To evaluate our method we use a data set of 250 non-homologous proteins, selected from the PDB library while ensuring that the pairwise sequence identity between these test proteins is less than 30%. Additionally, we omit all the proteins which are classified as “easy” by LOMETS [37] threading program. The objective behind that is to test our method on low-resolution structure predictions to make sure it is effective for those proteins with unknown structure and with no close homologs of known structure. Additionally, when applying our prediction method on each one of these test proteins, we filter our threading libraries (discussed in the section above) by omitting all templates with > 30% sequence identity to the query protein. Among these 250 proteins in our testing data set, 113 are enzymes which have the first three digits of their EC numbers covering all the 6 different enzyme families. This subset is used for evaluating the EC number predictions. The GO term predictions are evaluated on 225 proteins which are associated with at least one GO term, including 88 enzymes and 137 non-enzymatic proteins; the remaining 25 proteins do not have annotated GO terms and are not included in the GO test set. Of these 225 proteins, 205 were annotated with at least one Molecular Function term; 198 were annotated for Biological Process and 139 had at least one Cellular Component descriptions in the PDB GOA annotation [52]. As mentioned in the previous section, the structure model of each test protein has been generated using I-TASSER [33], and proteins having a sequence identity > 30% with the query protein are excluded from the templates libraries both in the I-TASSER threading library and in our function prediction template libraries.
3.2. Prediction of enzyme functions

Enzymes are proteins that catalyze (i.e., increase the rates of) chemical reactions in physiological processes. Based on the reactions they catalyze, the EC number is used to categorize these enzymes into hierarchical families. As shown previously, to predict the EC number from a query protein sequence, we predict its 3D structure and then combine global and local search methods to detect templates that share the same fold and share the same local active site with the query protein. The templates with the highest confidence score are used later for transferring the annotation.

For the purpose of illustration, we first present examples of EC prediction on two sample enzymes in our dataset: dihydrodipicolinate reductase and a cysteine protease staphopain, predicted with different structure modeling accuracies. There then follows an analysis of the overall prediction results from the EC testing data set.

3.2.1. Example of dihydrodipicolinate reductase

Dihydrodipicolinate reductase (DHPR) [EC number: 1.3.1.26] is a NAD(P)H dependent oxidoreductase involved in the anabolism of lysine and diaminopimelic acid in bacteria and plants. The structure of DHPR from E.coli is predicted using the automated I-TASSER pipeline with a confidence score (C-score) of 0.15; this indicates that the model may have a correct topology but the accuracy of intricate structural details may be unreliable. When compared to the experimental structure (PDB ID: 1ARZ), the modeled structure has a Cα RMSD of 5.9 Å and TM-score of 0.71 (Figure 3a). The predicted structure of this enzyme is examined using both local and global structure alignment programs and the template proteins in the enzyme libraries are ranked based on their similarity scores with the I-TASSER model.

The local similarity search finds only one hit, DHPR from M. tuberculosis (PDB ID: 1YL7), with an active site match (AcM) score of 1.65. It is usual to have just one or even no significant match
when predicted structures are used for local similarity search, because templates with relevant similarity at the level of the active site, are not always present in the local template library. Inaccuracies in the predicted structural models near the active site may also impede a good local match between the query and template proteins.

The global similarity search is performed independently on a different library. The templates are then ranked based on the Fh-score as explained previously. Since the local similarity search identifies only one template 1YL7 as the only template protein having a reliable AcM score, the final functional similarity score used for ranking 1YL7, the EC-score, will incorporate contributions from both the AcM score which represents the local similarity score and the Fh-score which is the global similarity score. The remaining templates which do not show a significant similarity at the level of the active sites to the query model will be ranked only based on their Fh-scores. The best scoring templates are then selected for inferring the protein function. For this query protein, all of the top five identified template proteins have the same global topology. However, the best template (1YL7) which shares the same EC number as the query protein has a significantly high Fh-score of 2.0. The other templates on the other hand produce a score lower than 0.92. Interestingly, even if we exclude the results obtained from the local-similarity search and rely completely on global similarity scores for ranking the templates, 1YL7 still achieves a significant score of 1.44, which is still much higher than the score of the rest of the identified hits. This proves the ability of the scoring function to detect the correct functional homologues even within functionally similar folds. Nevertheless, the combination of both global and local similarity helps to achieve a higher confidence level for the predicted function.
3.2.2. Example of staphopain

The second illustrative example is staphopain (EC: 3.4.22.48) from *Staphylococcus aureus* (PDB ID: 1CV8). The I-TASSER structure of this protein has a C-score of -2.21, which indicates a low resolution predicted structure. This is reflected in the $C_{\alpha}$ RMSD of 9.76 Å and TM-score of 0.45 for the model compared to the native structure (Figure 3b).

Although the global structure of the predicted model has low resolution, the core region of the model is well predicted (74/173 residues have an RMSD of 2.46 Å to native structure). All the templates detected by the global similarity search have a rather dissimilar overall topology compared to the predicted structure, and they all have a low confidence score. Thus the confidence for predicting the function using the template proteins identified by the global similarity search alone is low. On the other hand, the local similarity search results in 9 hits with reliable active sites matches (AcM>1.1), and they all share the same EC number. When the global and local scores are combined, we can achieve a better confidence score for these templates. The best template among these has an EC-score of 1.31 and shares the correct EC number with the query protein (EC: 3.4.22.48) which results in a correct prediction. This example shows how this method can achieve correct predictions even when there is no homologous proteins which shares the same global fold as the query protein.
3.2.3. Overall analysis of enzyme function prediction

A direct comparison of the EC prediction results on our testing data set compared to HHSEARCH [6] and LOMETS [37] is performed. We compare the performance of these methods using statistical parameters, specifically the commonly used metrics: precision (positive predictive value) and recall (sensitivity) rates. Figure 4 shows the precision-recall graph corresponding to each of these three methods. Our method is referred to as the “EC-score” which is the confidence score to score our predictions. A match is considered to be “true” if the first three digits of the EC number from the hit are identical to that of the query protein, otherwise the hit is considered to be “false”. A perfect prediction method on the precision-recall graph would be represented by a single point with coordinates (1, 1). Thus, an improved prediction method would produce a curve closer to the top-right corner. As shown in Figure 4, the rate of true positive predictions using the EC-
score is much higher than that of HHSEARCH and also that of LOMETS at most recall rates. For example, at a recall rate of 0.58, the EC-score achieves a precision of 0.76, 20% higher than that of LOMETS (0.63) and 52% higher than that of HHSEARCH (0.50).

**Figure 4:** Precision-recall curve to compare the EC prediction performance of the EC-score vs LOMETS and HHSEARCH.

### 3.3. Prediction of gene ontology terms

As shown in the previous section, to identify the GO terms corresponding to a query protein, we rely only on a global structural similarity approach for detecting functionally homologous template proteins because at the time being the functionally important residues for GO annotated proteins are unknown and thus, we do not have a library of local active sites corresponding to GO annotations that we could possibly use for similarity search methods. Based on the templates with similar global folds found during our search, we derive the consensus GO terms from the templates with the top Fh-score. We then detect the frequent GO terms among the interaction partners of the query protein extracted from STRING protein-protein interaction network [34].
3.3.1. Example of calmodulin

To illustrate how the global similarity search method works, we would like to discuss the example for calmodulin protein, which is a protein that binds to many other proteins and regulates multiple biological functions [53]. This protein is annotated with the following GO molecular functions: Calcium binding (GO:0005509), protein binding (GO:0005515), titin binding (GO:0031432), Thioesterase binding (GO:0031996) and N-terminal myristolation domain binding (GO:0031997). The structural model of this protein from I-TASSER has a relatively low confidence score (C-score) -1.46, which suggests that the structural model should have a low resolution. When compared to the native structure (PDB ID: 1CTR), the predicted structure has a TM-score of 0.45 and RMSD of 14.3Å (Figure 5).

The candidate template proteins which may have similar functions to the query protein are detected by performing a global structural search of the I-TASSER models through the GO template library, which is then followed by ranking of all the template proteins based on their Fh-scores. Although the confidence score of the structure modeling is low, there are still four
templates that have an Fh-score above the threshold cutoff of 1 (Figure 6). To calculate the confidence level of annotating the query protein with any GO-term, we combine all the GO terms associated with these top templates as well as their ancestors GO terms, present in their DAG. The GO-score which shows the confidence level of each term is then calculated by complementing the count of each term with a weighting factor, which is equal to the Fh-score of the associated template. For the example in Figure 6, all the four selected templates, Apoptosis linked gene protein (PDB ID: 1hqv), C-terminal domain of toponinC (PDB ID:2jnf), Calpain 13 (PDB ID: 2i7a) and Flagellar calcium binding protein (PDB ID: 3cs1) have a common function of Calcium binding (GO:0005509), for which the confidence score (GO-score=1.13) is simply the average of Fh-score of all the selected templates. For protein dimerization activity (GO:0046983), however, only two templates (Apoptosis linked gene protein and Calpain 13) have this function and the GO-score equals to the sum of the Fh-score of the two templates normalized by the total number of templates, i.e. 0.61 \[=(1.44+1.02)/4\].

<table>
<thead>
<tr>
<th>Selected templates</th>
<th>Fh-score</th>
<th>Gene Ontology terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1hqvA</td>
<td>1.44</td>
<td>GO:0005509 GO:0048306 GO:0042803 GO:0060090</td>
</tr>
<tr>
<td>2jnfA2</td>
<td>1.06</td>
<td>GO:0005509</td>
</tr>
<tr>
<td>2i7aA</td>
<td>1.02</td>
<td>GO:0005509 GO:0004198 GO:0046983</td>
</tr>
<tr>
<td>3cs1A</td>
<td>1.00</td>
<td>GO:0005509</td>
</tr>
</tbody>
</table>

\[
\text{GO-score (GO:0005509)} = \frac{1}{4} \times \sum_{i=1}^{4} \text{Fh-score}_i = \frac{1.44 + 1.06 + 1.02 + 1}{4} = 1.13
\]

\[
\text{GO-score (GO:0046983)} = \frac{1}{4} \times \sum_{i=1}^{2} \text{Fh-score}_i = \frac{1.44 + 1.02}{4} = 0.61
\]

**Predicted functions using consensus**

GO:0005509 (1.13), GO:0046983 (0.61), GO:0048306 (0.36),
GO:0042803 (0.36), GO:0060090 (0.36), GO:0004198 (0.25)
Figure 6: An illustrative example of molecular function term predictions for calmodulin (CaM) protein. GO-score calculation for two molecular function terms: calcium binding (GO: 0005509) and protein dimerization activity (GO: 0046983).

3.3.2. Overall analysis of GO terms prediction:
Our GO prediction method has also been compared to HHSEARCH and LOMETS to evaluate its performance. Figure 7 shows the precision-recall graph computed using matches (semantic similarity > 0.5) to the annotated GO terms of the query protein. The graphs clearly show that the GO-score outperforms both HHSEARCH and LOMETS. For instance, at the recall rate of 0.30, GO-score achieves a precision of 0.56 while LOMETS achieves a precision of 0.44 and HHSEARCH a precision of 0.44.

![Performance Comparison between GO-Score LOMETS and HHSEARCH](image)

Figure 7: Precision-recall curve to compare the GO prediction performance of GO-score v.s. LOMETS and HHSEARCH.
In addition to comparing the performance of the GO-score method with respect to LOMETS and HHSEARCH, we were also curious to know which part of the algorithm contributes most to the prediction accuracy. Since I-TASSER is currently the best structure predictor, we were wondering if the scoring function does also contribute to the performance of our method. To do that, we decided to combine HHSEARCH with our scoring function (GO-score) and compare its performance to the original HHSEARCH. Similarly, we compared LOMETS combined with our GO-score to the original LOMETS. Figures 8 and 9 show the performance of HHSEARCH combined with the GO-score to the original HHSEARCH, and LOMETS combined with the GO-score to the original LOMETS respectively.

**Figure 8:** Precision-recall curve to compare the performance of HHSEARCH combined with the GO-score to the original HHSEARCH.
As Figure 8 and Figure 9 show, the performance of HHSEARCH and LOMETS when combined with our scoring function is much better than the performance of HHSEARCH and LOMETS alone which shows that the design of the scoring function does contribute to improve the prediction accuracy.

Additionally, we also wanted to see how our method using I-TASSER and the GO-score would perform compared to these combined LOMETS /HHSEARCH methods with the GO-score. Here again, we draw the precision-recall curve to compare our GO-score method which uses I-TASSER to HHSEARCH combined with the GO-score and LOMETS combined with the GO-score as shown in Figure 10 below:
The precision-recall curve shows that even when we combine HHSEARCH or LOMETS with our scoring function, our method which combines I-TASSER with the GO-score still performs slightly better. This is to show that our protein structure predictor, I-TASSER does also contribute to the prediction accuracy. Therefore I-TASSER and the GO-score contribute together to achieve an accurate GO function prediction.

As discussed in the methodology section previously, we also wanted to make use of more information sources to improve our GO prediction. In particular, while structure based prediction, which is what we used so far, achieves a good performance to predict GO terms of type Molecular Function, we were assuming that including protein-protein interaction information may contribute
to improving Cellular Component and Biological Process GO terms, as two proteins that interact should theoretically be part of the same cellular component and be part of the same biological process. To do so, we incorporated protein-protein interaction information from STRING database to our GO-score prediction to see if we can improve our performance as explained in the methodology section. Figure 11 below shows the precision-recall graph of the combination of the GO-score with STRING information compared to the performance of the GO-score method alone:

"GO-Score" vs "GO-Score+STRING" Performance

![Graph showing precision-recall curve](image)

**Figure 11:** Precision-recall curve comparing the GO-score method to GO-score combined with protein-protein interaction information from STRING.

The precision-recall curve above shows that the contribution of protein-protein information from STRING is very significant, especially for large recall rates. For instance at a recall rate of 0.71, GO-score combined with STRING achieves a precision of 0.52 which is 49% higher than the precision achieved by GO-score alone (0.34).

Also, we were especially curious to see how protein-protein interaction information improves each of the three aspects of GO terms: Molecular Function, Biological Process and Cellular
Component. Figure 12 below shows the precision-recall curve for GO-score combined with STRING vs. GO-score alone for each of these three aspects.

![Precision-recall curves](image)

Figure 12: Precision-recall curve comparing the GO-score method to GO-score combined with STRING for each of three GO aspects. (a) Molecular Function; (b) Biological Process; (c) Cellular Component.

The precision-recall curve in Figure 12 confirms our guess on how protein-protein interaction information is very useful in predicting Biological Process and Cellular Component GO terms. As shown in the figure, while there is some improvement in predicting Molecular Function terms, this improvement is not substantial. However, for Cellular Component GO terms and especially Biological Process GO terms the improvement form STRING protein-protein interaction network is very significant.
4. CONCLUSION

In this thesis work, we have worked on deducing biological function of protein molecules based on the sequence-to-structure-to-function paradigm. From the amino acid sequences, 3D structures are first predicted by the iterative threading assembly refinement (I-TASSER) methods. Functional analogs are then identified by performing global and local structural similarity searches through the functional libraries, with the scoring function involving the confidence score of structural predictions, sequence and structural similarity of I-TASSER model with the functional templates, and the local active site matches. We have also tried to improve the performance of GO prediction by incorporating protein-protein interaction information, especially in order to improve those GO terms under Biological Process and Cellular Component aspect.

Our method has been tested on a set of hard proteins with known EC number and gene ontology terms. The results show the significant advantage of our structure-based function inference is found over the conventional sequence-based predictions using HHSEARCH or threading-based prediction using LOMETS, especially for the proteins lacking homologous templates. Also, incorporating protein-protein interaction information from STRING shows a significant improvement over our initial method especially for Biological Process and Cellular Component GO terms. Overall, these results demonstrate great promise towards utilization of the current method for automated, genome-wide structural and functional annotations.

As a future work, we are working on enhancing the method by including amino acid motif features to improve the overall precision of the method.
5. REFERENCES:

   http://www.ebi.ac.uk/uniprot/TrEMBLstats .
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local
   Bioinformatics 21, 951-960.
8. Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of
9. Rost, B. (2002). Enzyme function less conserved than anticipated. J Mol Biol 318, 595-
   608.
    structure in proteins. EMBO J 5, 823-826.
    Mol Biol 291, 977-995.
    genomics: quantifying the relations between protein sequence, structure and function
    Structure 13, 121-130.
    227-258.
    descriptors. Brief Bioinform 3, 265-274.


