Analysis of Methods for Predicting Protein Fold and Remote Homologue Recognition

Prepared by

Sanjay Srivastava
ssrivast@stanford.edu

Biochemistry 218: Computational Molecular Biology
Professor Doug Brutlag
Stanford University

1. PROTEIN FOLD AND REMOTE HOMOLOGS

1.1 INTRODUCTION

Life is a complex system of information storage and processing. Information is transmitted vertically from cell to cell and at the same time, information is expressed horizontally within a cell of an individual organism. Bioinformatics and, in particular, computational biology aims at understanding biology at the molecular level. Among many of the on-going efforts to learn biology, the protein folding has been one of those challenging areas in computational biology. The challenge is to predict the native 3D structure or a fold of a protein from its amino acid sequence and its secondary structure.
The three-dimensional structures of number of globular and membrane proteins have been found experimentally and have been classified and annotated in various protein structural databases. It is believed that there is a limited number of protein folds in nature. The protein structure determination, therefore, falls into two categories: fold prediction and fold recognition. The functional and structural annotation of unknown protein structure is initially often performed by pairwise sequence similarity searches against protein sequence databases. The reason is that it three-dimensional (3D) structure is conserved in evolution and is based on protein primary sequence. If an unknown protein sequence can be aligned with a protein with a known structure it can be assumed that the new protein would have similar 3D structure. Accordingly if the probe has a sequence homolog, the protein structure prediction method is then called fold recognition otherwise the problem is referred as fold prediction.

1.1.1 Prediction of Protein 3D Structures

The laws of physics and the evolutionary changes have affected the Three-dimensional (3D) structure of proteins. According to the laws of physics, a protein molecule in solution is a system of atoms that interact through a variety of forces, such as chemical bonds, hydrogen bonds, Coulomb interactions, and Lennard-Jones forces. Under appropriate conditions, these forces fold almost any random starting conformation of a protein into a stable, well-defined 3D structure (i.e., the native state) in a matter of milliseconds or seconds. Evolution, on the other hand, resulted in families of proteins that share similar sequences, similar structures, and often have related functions. Different proteins evolved through duplication, speciation, and horizontal transfer, followed by accumulation of mostly neutral mutations.

Each of the two sets of principles that apply to the natural protein sequences and databases of known protein structures gave rise to two approaches to protein structure prediction methods (Baker, *et al*., 2001). The first approach, *ab initio* methods, predicts the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures (Bonneau, *et al*, 2001). The *ab initio* methods attempt to find the global free energy minimum by an exploration of many conceivable protein conformations. These methods assume that the native structure corresponds to the global free

energy minimum accessible. The second class of methods, which include threading and comparative modeling, rely on detectable similarity spanning most of the modeled sequence and at least one known structure (Blundell, *et al*, 1987). When the structure of one protein in the family has been determined by experiment, the other members of the family can be modeled on the basis of their alignment to the known structure. Comparative modelling is the most widely used method for predicting the 3-D structure of a target protein. The method is however limited to predicting the structure of proteins which are closely related to a template protein of known structure. The comparative modelling process can be divided into five basic steps: alignment of the target sequence with the sequence of a protein of known 3-D structure; building of a framework structure based on the alignment; loop building; addition and optimization of side chains; and finally model refinement. Apart from cases where the target has a very close homologue of known structure, the vast majority of comparative models still display quite serious errors in alignments. These methods are consequently loosing ground (Jones, 2001).

Threading methods were initially designed to predict protein folds when no suitable template structure could be found to create a model for comparative modeling prediction method. With threading methods, proteins that are remote homologues of known protein sequence can be examined for compatibility with the structural core of a known protein core. The query sequence is threaded into a database of proteins cores to look for matches and optimization of the potential energy function that includes parameters such as pair-wise potentials of side-chain interactions and scores for buried hydrophobic residues. In addition, these methods can include parameters derived from comparative modeling and *ab initio* predictions.

Comparative modeling and threading methods are knowledge based protein structure prediction methods. These methods rely on classification and structures of known protein structures. Since structures of many proteins have been determined experimentally and have been classified based on their hierarchies. Few databases exist with detailed structural and functional annotations. In 1971, the Protein Data Bank (PDB) was established. Since then, few more structural databases have been created. These include Structural Classification of Proteins (SCOP) database (Murzin *et al*, 1995), CATH (Orengo *et al.*, 1997) FSSP, and pCLASS. The majority of proteins in these databases falls into globular or membrane proteins categories. Based on content and arrangement of structure elements, globular proteins are classified into the following categories: all α proteins, all β proteins, α/β proteins, and $\alpha+\beta$ proteins. These categories are often called a *class* of protein. This kind of protein classification was further refined to represent hierarchy of protein structures. The SCOP database, for example, follows the class classification at the top level and subdivides it consisting of folds, superfamilies, and families. Where fold is a specific arrangement of secondary structure elements such as α helix bundle. Family is a group of proteins that have significant level of structural similarity but not necessarily significant sequence similarity and proteins in a superfamily have sequence similarity, which suggests common evolutionary origin.

1.2 COMMON PROTEIN PREDICTION METHODS

Few methods have been deployed to find homologies in protein sequences and ultimately protein structure. The most widely used methods involve either using pairwise searches or using structural information. The pairwise search method uses a single sequence of the unknown protein is scanned against each sequence in a database using programs such as BLAST, FASTA or any dynamic programming. With PSI-BLAST, even remote homologies can be found (Altschul, *et al*, 1997). In this iterative search method, the unknown sequence is used to identify close homologues that are then aligned to generate a weighted profile formalized as a position-specific scoring matrix (PSSM) (Henikoff & Henikoff, 1994). The Psi-BLAST searches are more sensitive than pairwise sequence homologies search techniques and are widely used. The reason being that PSI-BLAST searches are bale to find structural similarities even among proteins sequences that are not homologs.

Structural techniques have been used at a variety of levels (Jones, 1997). These can be classified in two types: Profile method and optimization of a pesudo-potential based on the relative attraction of the residue pairs. In the profile method the environment of each amino acid residue, secondary structure, and area of side chains in the structural core is determined. Based on this information, the protein is characterized and aligned. The proteins in the database are similarly aligned and the query protein sequence is then scored against the proteins in the database (Bowie *et al*, 1991). For pesudo-potential based method, the contact potential of adjacent amino acid residues in the structural core are determined and the conformations that would give rise to a most stable structure is predicted (Jones, *et al*., 1992).

In this project, I will describe and analyze one sequence-based protein structure prediction method and another one that incorporate both the sequence and structural based profiling. In particular, I will use Hidden Markov Model (HMM) and 3-Dimensional Position Specific Scoring Method (3D-PSSM) in this project. I chose these methods in this study because these two methods have fared well at CASP (Critical Assessment of methods for Structure prediction of Proteins) meetings. CASP is an international experiment, which is organized every 2 years since 1994 (Shortle, *et al*. 1995). The purpose of CASP is that the prediction community is challenged to predict the structure of a number of proteins whose structures are not yet publicly known, and these 'blind' predictions are then evaluated by a number of independent assessors who compare the predictions with the experimentally determined structures.

1.2.1 Three-Dimensional Position-Specific Scoring Method, 3D-PSSM

The 3D-PSSM method takes a protein sequence and attempt to predict its 3-Dimensional (3D) structure and probable function (Lawrence, *et al*., 2000). It uses a library of known protein structures onto which the sequence is "threaded" and scored for compatibility. The 3D-PSSM combines multiple-sequence profiles such as PSI-Blast (1D-PSSMs), more general profiles containing more remote homologues (3D-PSSMs), matching of secondary structure elements,

and propensities of the residues in the query sequence to occupy varying levels of solvent accessibility; specifically, solvation potentials.

The general procedure in predicting a protein fold is as follows:

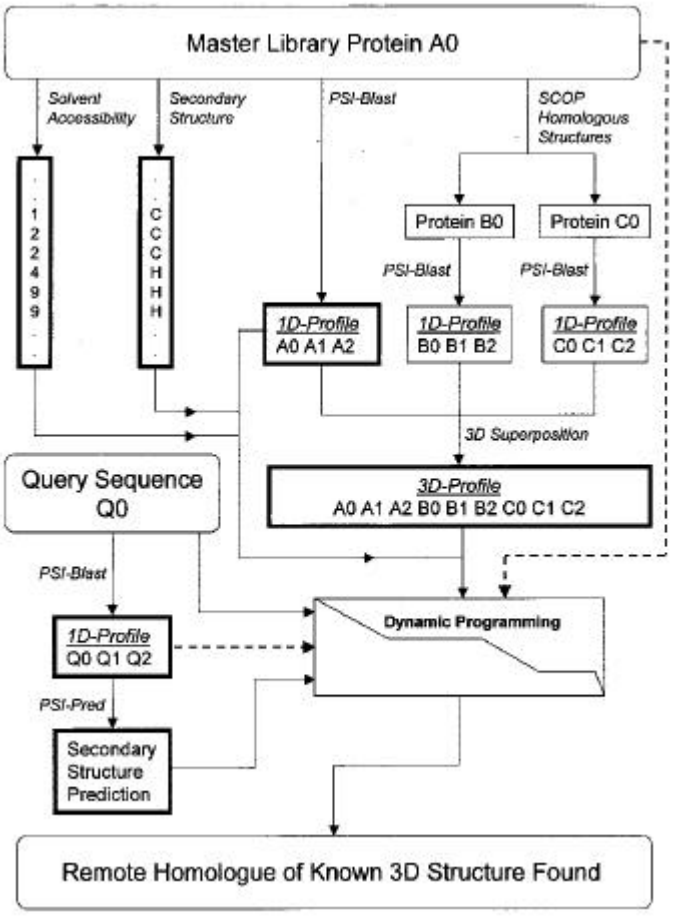

Figure 1 A Schematic of 3D-PSSM Procedure

First, a representative profile library of known protein structure is created. The library of known structures is taken from the SCOP database that classifies proteins into homologous Superfamilies. From this library, a representative parent protein sequence is chosen and is called (Master A0). Upon selection, the method generates 1D-PSSM or 1D-profile using PSI-Blast (A0,A1,A2…). Next, it uses STRIDE (Fisherman & Argos, 1995) to make a three state (helix, coil, sheet) secondary structure assignment of each residue in *Master A* protein sequence. In addition, using DSSP technique, propensities of residues to varying level of solvent exposure are determined. Finally, it generates a 3D-PSSM of the representative sequence. To generate 3D-PSSM for the representative Master A0 protein, it uses protein sequences in the homologous superfamilies family. Proteins in same fold families may not have similar sequences or functions but share similar tertiary structures. Using this information, accurate residue equivalencies for a fold can be determined. Representatives

from the profile family is chosen to create ID-profiles such as B0, B1, B2… and C0, C1, C2… and used to create a 3D-PSSM for the representative protein, Master A0.

The query sequence undergoes similar processes. The PSI-Blast is first used to derive its sequence homologues to create 1D-profile and PSI-Pred is used to predict its secondary structure. Lastly, The query sequence is aligned using a dynamic programming algorithm, FOLDFIT (Russell, *et al*., 1998) against each library entry. The query is first aligned to the template using the 1D-PSSM of the library entry, then the 3D-PSSM and finally the process is reversed and the library sequence is aligned to the 1D-PSSM of the query. The highest scoring pass is taken as the final result. The server is publicly available at http://www.sbg.bio.ic.ac.uk/~3dpssm and was used for this project.

1.2.2 Hidden Markov Model, HMM

This method develops HMM models to recognize remote homologies. A hidden Markov Model (Rabiner, 1989) has been used to describe a series of observations by a stochastic process. Before solving protein alignment and structure problem, HMMs were used in speech recognition where observations were sounds forming a word. In that case the model described the random process that produces these sounds with high probability. In speech recognition, the "alphabet" forms the phonemes in a particular language and in protein folding, the "alphabets" are the 20 amino acids that form a protein sequence. A model in a set of proteins is the one that sets high probability for a sequence in a particular set. Once a model has been constructed using unaligned sequences, dynamic programming, while searching a database, can be used to generate multiple alignments of protein sequences. The model can therefore, distinguish different families of proteins.

One method for predicting the structure of a target sequence involves: constructing an HMM from the target and identified homologs and scoring the sequences in the Protein Data Bank (PDB). Later, the process scores each target sequence against a library of HMMs constructed on a representative subset of PDB (Karplus, K. 1997). Typically use of profile HMMs in homology detection is more complicated because of the need to first construct the profile HMM. The procedure consists of three steps. (i) A multiple sequence alignment is made of known members of a given protein family. The quality of the alignment and the number and diversity of the sequences it contains are crucial for the eventual success of the whole procedure. (ii) A profile HMM of the family is built from the multiple sequence alignment. The model-building program uses information derived from the alignment together with its prior knowledge of the general nature and structure of proteins. (iii) Finally, a model-scoring program is used to assign a score with respect to the model to any fold sequence of interest; the better the score, the higher the chance that the query fold sequence is a member (homolog) of the protein family represented by the model. In this way each sequence in a database can be scored to find the members of the family present in the database.

Profile hidden Markov models (HMMs) are amongst the most successful procedures for detecting remote homology between proteins and predicting protein srtutures. There are two

popular profile HMM programs, HMMER (Eddy, et al, 1998) and Sequence Alignment Modeling (SAM).(Karplus, K. et al, 1998). SAM uses a single target sequence and creates a HMM using an iterative method to refine the model. One of its main uses is to produce multiple alignments of homologs of the target sequence. Developed by the bioinformatics group at the University of California, Santa Cruz (http://www.cse.ucsc.edu/research/compbio/sam.html). The SAM model-scoring program calculates E-values directly using a theoretical function that takes as its argument the difference between raw scores of the query sequence and its reverse.

1.3 3-DIMENSIONAL STRUCTURE PREDICTIONS FOR TRYPSIN HOMOLOGS

The prediction of the three-dimensional (3D) structure of a protein from its one dimensional (1D) involves the kind of fold that the given amino acid sequence may adopt. If similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed, because the 3D structures of proteins from the same family are more conserved than their primary sequences. With this in mind, I chose a trypsin molecule (1TRY) and its sequence homolgs as a starting query sequences for this project. Later I used 3D-PSSM and HMM based SAM-T99 server to predict their 3D structures. It was expected that all the homologs of the starting sequence would have similar fold as the starting trypsin molecule. In addition, the idea of this study was to highlight any sensitivity and accuracy differences between two prediction methods. In particular, I was interested to examine which technique is able to identify remote homologies more accurately as both these techniques would be able to predict the fold. The reason for accurate prediction is that their profile library does contain 1TRY protein.

The 1TRY molecule has been isolated from Fusarium oxysporum and its sequence and X-ray structure has been determined (Rypniewski, *et. al*, 1993). Trypsin or serine proteases are ubiquitous. They are found in viruses, bacteria and eukaryotes and include proteins with exopeptidase, endopeptidase, oligopeptidase and omega-peptidase activity. Serine proteases of the chymotrypsin family share a common fold and are involved in diverse and important functions, including digestion and degradative processes, blood coagulation, fibrinolysis, cellular and humoral immunity, embryonic development, and fertilization (Rawlings *et al.*, 1994).

According to SCOP structural classification, 1TRY is classified as follows:

1. Class: All beta proteins
2. Fold: Trypsin-like serine proteases
   barrel, closed; n=6, S=8; greek-key
   duplication: consists of two domains of the same fold
3. Superfamily: Trypsin-like serine proteases
4. Family: Eukaryotic proteases

5. Protein: Trypsin(ogen)
6. Species: Mold (*Fusarium oxysporum*)

The structural view of 1TRY is shown in Figure 2.

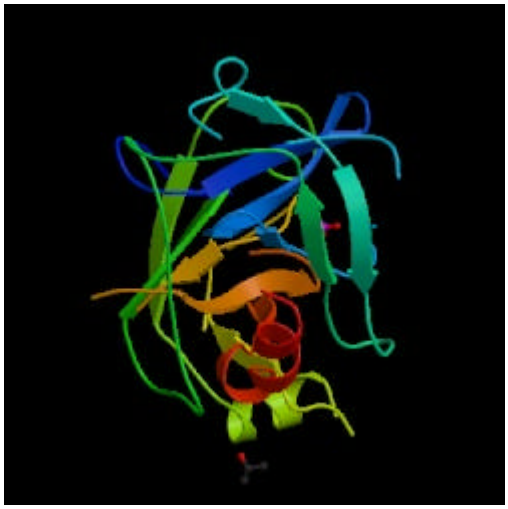

Figure 2. Structure of Inhibited Trypsin From Fusarium-Oxysporum At 1.55-Angstrom.

Strutural homologs of trypsin were identified first, using Smith-Waterman Algorithm on DeCypher machine (http://decypher.stanford.edu/index_by_algo.htm). Its sequence homologs that were used in this project are listed in **Table 1**.

Table 1 1TRY sequence homologs used in this study.

Accession Number	Definition
P35049	fusarium oxysporum. trypsin precursor (ec 3.4
P04814	drosophila melanogaster (fruit fly). trypsin
P54625	drosophila erecta (fruit fly). trypsin beta p
P54624	drosophila erecta (fruit fly). trypsin alpha
P42277	drosophila melanogaster (fruit fly). trypsin
P54626	drosophila erecta (fruit fly). trypsin delta/gamma precursor
P42276	drosophila melanogaster (fruit fly). trypsin
P35004	drosophila melanogaster (fruit fly). trypsin
O97370	euroglyphus maynei (house-dust mite). mite al
P29786	aedes aegypti (yellowfever mosquito). trypsin

Regardless of the prediction method used, it was expected that each of these homologs would belong to similar SCOP classification as 1TRY because they have sequence similarity and have similar functions. The results are tabulated in
Table 2

Table 2.3D Structure Prediction. At least one true positive was noted for each sequence homologs

Accession Number	Class	Fold	Superfamily	Family	3D-PSSM	SAM T99
P35049	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P04814	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P54625	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P54624	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P42277	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P54626	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P42276	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P35004	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
O97370	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes
P29786	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases	Yes	Yes

As shown in Table 2., both these prediction methods were successful in identifying the correct folds. It is not suprising because both these methods attempt to combine sequence profile alignment methods with fold recognition. This should, in principle, produce an

alignment method that produces accurate alignments both where the target and template proteins are in the same superfamily. It is expected that even an increasingly sensitive sequence comparison methods such as PSI-BLAST will equal or even surpass the abilities of fold recognition methods such as SAM and 3D-PSSM. This is true especially in this study because predicted homologs are in the training data set. The reason is that am using the publicly available servers and they use "nr" protein database that probably contain my query sequences. In view of that, I will highlight their differences in their ability to recognize remote homologies.

It was noted that despite recognizing the correct folds, the two methods predicted different structural neighbors within the predicted folds. One reason for different set of predicted structural neighbors could be in the manner the profile model is created and/or the database they use to search the model profile sequences. Both these methods rely on creation of profile models and searching the database and scoring the results. Because resulting structural homologs are different, it is evident that the two methods differ. To learn more about the model creation, I performed a ClustalW multiple alignment method on the resulting predicted protein sequences obtained by 3D-PSSM and SAM-T99. As an example of such an analysis, **Figures 3** and **4** displays the phylogenic trees of 1TRY's predicted structural neighbors.

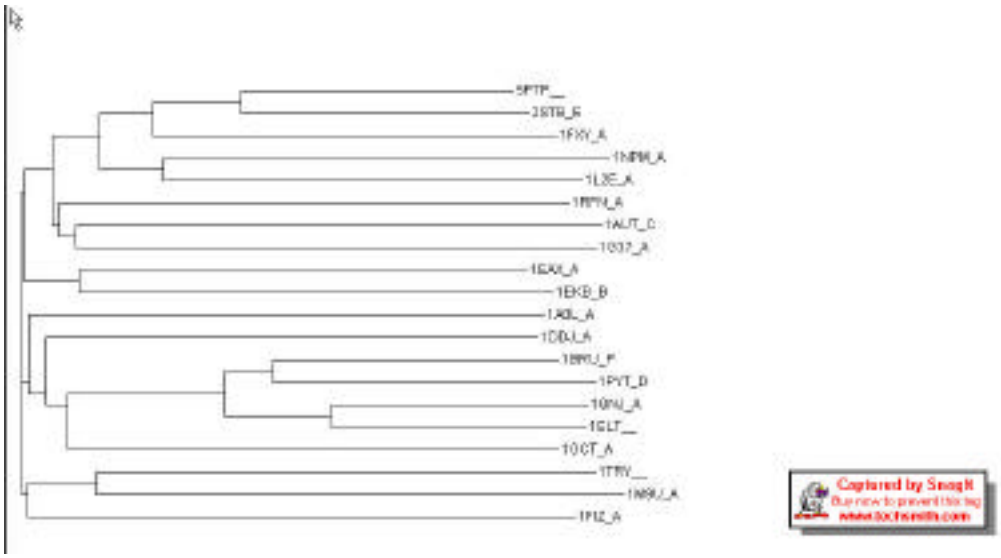

Figure 3 Structural neighbors of 1TRY as predicted by 3D-PSSM

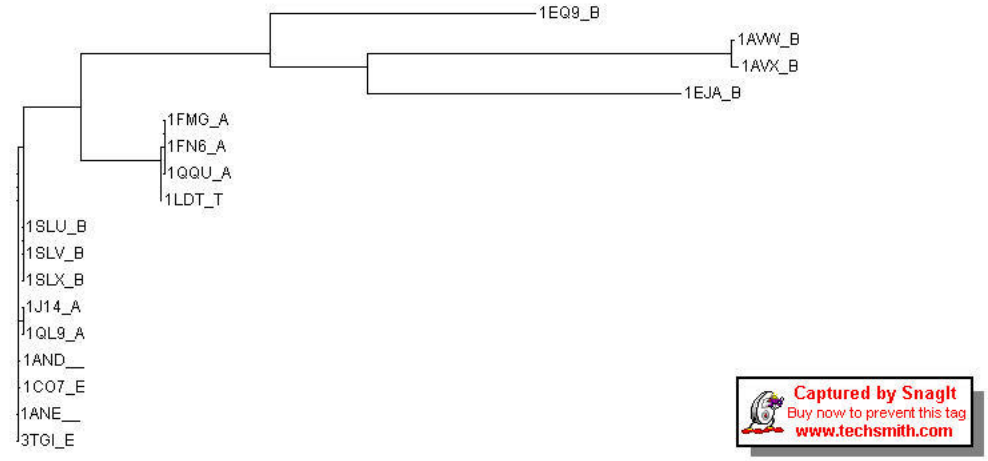

Figure 4 Structural neighbors of 1TRY as predicted by SAM-T99.

1.4 DISCUSSION AND CONCLUSIONS

Overall the 3D prediction for 1TRY protein and its homologs are given in **Table 2**. It appears that both methods were able to successfully predict the fold to be Trypsin-like serine proteases as expected. Nonetheless, the two methods identified different structural homologs. Multiple sequence alignment results shown in **Figure 2** and **3** clearly demonstrate that 3D-PSSM is better in recognizing remote homologies than SAM-T99. The phylogenic trees are orthologous because they are rooted trees where its leaves (proteins) belong to different species such as mouse, humans, etc and branch length of the two trees indicate that 3D-PSSM was able to identify more disparate remote homologies than SAM-T99 technique. The difference in the predicated structural homologs within the structural folds can be ascribed to the prediction algorithm and process that SAM and 3D-PSSM utilizes. The SAM-T99 protocol, uses a multiple alignment (or even a single seed sequence) to build an HMM, which is then used for searching for new members of the family. When new members are found, the HMM is retrained to include them, new multiple alignments are made, and the process iterated. SAM-T99 starts with a query sequence and searches the non-redundant protein database using WU-BLASTP (Altschul, *et al.*, 1990) to produce of potential homologs. This method has proven more effective in finding remote homologs than competing sequence-based methods such as PSI-BLAST and ISS (Karplus, *et al.*, 1998).

On the other hand, 3D-PSSM method explicitly uses information from structural alignments in searching remote homologs of known 3D structures. In this method, a library of known structures is taken and a 1D-PSSM for a representative protein is created by using PSI-BLAST. Later, its profile is generated by predicting its secondary structure and using solvation accessibility information at each residue. This information is used to create a 3D-PSSM fold library using remote sequence homologs in the superfamily. Latter allows inclusion of structural information even when proteins have dissimilar sequences. The query

sequence's 1D-PSSM and secondary structure is predicted and aligned against 1D-PSSM and 3D-PSSM of the template library created earlier.

In summary, SAM-T99 builds profile HMM model based on multiple sequence alignment and general nature of the proteins and 3D-PSSM uses sequence alignment and structural information to predict 3D Structural homologs. Based on the results obtained in this analysis, both methods are able to identify the correct fold for the query protein sequences, however 3D-PSSM fares better in recognizing remote homologs. According to LiveBench-2 evaluation results, 3D-PSSM indeed performs better than most threading servers in predicting protein structures (Janusz, M., *et. al.* 2001).

Protein folding in living cell is a complex and dynamic process involving a number of other molecules called chaperones. The cellular environment assists specific interactions with various molecules, therefore protein is only a part of the whole system and that amino acid sequence may not necessarily contain all the information to fold it up in its native active form. Homologous proteins with their relative sequences usually evolve from a common ancestor, and nearly always have similar three-dimensional structures and often have similar functions. However, a difference in activity is always a possibility (Russell *et. al.* 1998; Hegyi & Gerstein, 1999). In view of that, I believe that the protein-folding and especially predicting its function is a very challenging problem and cannot be solved for a many of proteins in nature without considering the network of specific molecular interactions. Automated threading procedures based such as HMM and in particular 3D-PSSM are quite successful in predicting protein folds as shown here and in CASP4 and CASP5 meetings. However, it is not clear that either these fold recognition methods described here or elsewhere will be able to distinguish individual protein structures within the same fold. As seen here, both these methods were able to predict the fold but predicted different set of structural homologs that had different functions. Besides the common folds and common motif overlap, these techniques were not able to distinguish them.

Although, 3D-PSSM also attempts to suggest functional similarity of the query with the template, both these techniques can be improved using human intervention. Most prediction methods, and fold recognition methods in particular, are designed to predict the structure of a single domain, and therefore fare very badly when confronted with a large multidomain target protein. For example, prediction abilities of automated SAM T-99 (Karplus K. *et al.* 2001) were improved when multi-domain proteins were split manually and folds for individual domains were determined accurately. In addition, most solvation and contact potential methods used in structural profiling as in 3D-PSSM use interactions of amino acids that are immediate neighbors in the protein core. I believe that one should consider long-range interactions in defining structural profiles as a folded protein could have such interactions.

1.5 REFERENCES

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. Basic local alignment search tool. 1990 *J. Mol. Biol.*, 215, 403–410.

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

Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. Science 2001, 294, 93-96.

Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J.; Thornton, J. M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. Nature 1987, 326, 347-352.

Bonneau, R.; Baker, D. Ab Initio Protein Structure Prediction: Progress and Prospects. Annu. Rev. Biophys. Biomol. Struct.2001, 30, 173-189.

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

Eddy,S.R. Profile hidden Markov models. *Bioinformatics*, 1998, 14, 755–763.

Karplus,K., Barrett,C. and Hughey,R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 1999, 14, 846–856.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J. Mol. Biol. 288, 147-164.
Jones, D.T., Critically assessing the state-of-the-art in protein structure prediction. Pharmacogenomics J. 2001;1(2):126-34. Review.

Janusz M. Bujnicki, Arne Elofsson, Daniel Fischer, and Leszek Rychlewski, LiveBench-2: Large-Scale Automated Evaluation of Protein Structure Prediction Servers, Proteins, 2001, Suppl 5:184–191.

Jones D.T., Taylor W.R., and Thronton J.M. A new approach to protein fold recognition. Nature, 1992, 358: 86-89.

Kevin Karplus, Rachel Karchin, Christian Barrett, Spencer Tu, Melissa Cline, Mark Diekhans, Leslie Grate, Jonathan Casper, and Richard Hughey, What Is the Value Added by Human Intervention in Protein Structure Prediction? Proteins, 2001, Suppl 5:86-91

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Hausler, D., Hugey, R., Holm, L., and Sander, C. Predicting Protein Structure Using Hidden Markov Models, *Proteins*, 1997, Suppl. 1, 134-139.

Kevin Karplus and Birong Hu, Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set**,** Bioinformatics, 2001,17(8), 713–720

Lawrence A. Kelley, Robert M. MacCallum and Michael J. E. Sternberg Enhanced Genome Annotation Using Structural Profiles in the Program 3D-PSSM J. Mol. Biol. 2000, 299, 499-520.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classiÆcation of protein databases for the investigation of sequences and structures. J. Mol. Biol. 1995, 247, 536-540.

Rabiner, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE, 1989, 77, (2), 257-286.

Rawlings,N.D., Barrett,A.J., Families of Serine Peptidases. *Meth. Enzymol.* 1994, 244: 19-61

Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. Supersites within superfolds. Binding site similarity in the absence of homology. J. Mol. Biol, 1998a, 282, 903-918.

Russell, R. B., Saqi, M. A. S., Bates, P. A., Sayle, R. A. & Sternberg, M. J. E.. Recognition of analgous and homologous folds. Assessment of prediction success and associated alignment accuracy using empirical matrices. Protein Eng., 1998b, 11, 1-9.

Rypniewski,W.R., Hastrup,S., Betzel,C., Dauter,M., Dauter,Z., Papendorf,G., Branner,S. and Wilson,K.S. The sequence and X-ray structure of the trypsin from Fusarium oxysporum Protein Eng., 1993, 6 (4), 341-348

Shortle D. Protein fold recognition. *Nature Struct Biol* 1995; 2: 91–93.