Journal of Engineering and Applied Sciences 12 (24): 7534-7538, 2017
ISSN: 1816-949X

Function Based Predictions of Protein Fold Recognition using Go-Term

[1]E. Loganathan, [2]K. Dinakaran, [3]S. Gnanendra and [4]P.Valarmathie
[1]Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India
[2]Department of Computer Science and Engineering, P.M.R Engineering College, Chennai, Tamil Nadu, India
[3]Department of Biotechnology, Yeungnam University, Gyeongsan, South korea
[4]Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, Tamil Nadu, India

Abstract: Machine learning-based methods are the most prominently employed in methods in the development of novel protein fold recognition tools. The most recent fold recognition method was developed by combining the four descriptors (e-Values) of Position Specific Iteration BLAST (PSI BLAST), reverse PSI-BLAST (RPS-BLAST), alignment of Secondary Structure Elements (SSE) and PROSITE motifs. In this present study, we emphasized to improve the fold recognition methods by including gene-ontology terms as additional descriptors which can aid in the determination of function based predictions. This method of descriptor combinations have resulted high sensitivity in determining the protein folds when compared to the methods developed with single descriptors. Also, the inclusion of GO-term descriptor have highly increased the sensitivity of the methods in fold recognition which significantly envisages the usage of GO-terms as prominent descriptors that can be employed in the protein fold predictions.

Key words: Secondary structure, protein FOLD, gene ontology, MeSH terms, alignment, development

INTRODUCTION

Globally, the huge deposition of completed genome brings the drastic gap in the protein sequence, structure and their functions. Thus, based on the available empirical protein sequences the protein structure prediction is dramatically increasing and has established as a routine application in determining the new protein structures by many life sciences (Petrey and Honig, 2005). The protein structure prediction based on experimental protein sequence (template) classically needs three steps. template identification with remote homology for the target sequence. Query and template sequence similarity. Generation of template based query protein 3D structure. In this pipeline, the identification of template (experimentally available protein 3D structure) protein sequence that are similar to the query sequence are considered as the most significant step. In line with this, the most commonly used alignment tools such as Basic local alignment search tool (Altschul *et al*., 1990), Fast algorithm (Pearson, 1990), Smith-waterman algorithm (Smith and Waterman, 1981) or Needleman-wunsch algorithm tools (Needleman and Wunsch, 1970) are employed to understand the percentage of similarity shared by query-template sequences. In general, these protein modeling that are based on similarity are known as theoretical modeling or comparative modeling. While employing these comparative modeling it is believed that the template and query alignment should share more that 40% similarity to generate the reliable protein structure. On the other hand, the query-template alignments with lesser similarity (<40%) are known as remote homologous. In certain cases, it is entrusted that the query template with lesser similarity (often referred as remote homology) can share the similar folds in the protein 3D structure. In this scenario, the most frequently used sequence alignment tools fails to identify the remote homologous proteins. However, the development of profiles based alignment methods such as PSI BLAST (Altschul *et al*., 1997), RPS-BLAST (reverse PSI-BLAST) (Altschul *et al*., 1997), IMPALA (Integrating matrix profiles and local alignments) (Schaffer *et al*., 1999) and HMM (Hidden Markov Models) profiles (Sonnhammer *et al*., 1997) has significantly, identified the remote homologous proteins. Yet, these methods also exhibited poor performance when the template and query sequence alignment shares 20-35% or lesser similarity (Twilight zone) (Rost, 1999). Thus, the development of sensitive tools that can detect the remote homologous proteins has drastically increased in number

Corresponding Author: E. Loganathan, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India

during last decade. Now these remote homologous tools including FFAS-3D (Fold and Function Assignment System) (Jaroszewski *et al*., 2015), 3D-PSSM (Position-Specific Scoring Matrix) (Kelley *et al*., 2000), Fugue (Shi *et al*., 2001), mGenThreader (McGuffin and Jones, 2003), ORFeus (Ginalski *et al*., 2003), MUSTER (MUlti-Sources ThreadER) (Wu and Zhang, 2008) and SP5 (Sparks 5) (Zhang *et al*., 2008) are considered as powerful fold recognition tools in the identification of template for comparative modeling.

In general, these fold recognition tools takes the strategy of screening all the structures of fold library with the given query sequence and can identify the best template based on the sequence structure compatibility. At present the fold recognition methods are based on structure seeded profiles (3D-PSSM and Fugue), profile-profile alignment (PSI and RPS) and machine learning (mGenThreader). In general, the structure seeded profile methods (3D PSSM) uses sequence and structure profiles along with the predicted secondary structure. Whereas the profile-profile alignment methods (PSI BLAST) involves dynamic programming for alignment and structural information. While, the machine learning methods combine various parameters related to sequence and structure. In recent years, Support Vector Machines (SVMs) are widely used as machine learning method which can build binary classifiers to predict the sequence that belongs to structural fold.

Thus, in this study, we emphasized the development of old server methods through a machine learning method. For this, the features (descriptors) of two proteins are analyzed. For instance, the amino acid composition of a sequence considered a descriptor, the BLAST search value of query protein against template proteins considered as a descriptor. Likewise, thirteen descriptors including gene-ontology terms were also evaluated in this study for its capabilities in identification of folds.

MATERIALS AND METHODS

Datasets: The most reliable protein 3D structures are screened from the protein data bank and their information's regarding secondary structure are predicted a database namely PSSRDB was constructed. These secondary structure datas from PSSRDB are used in the prediction of fold recognition.

Descriptors: The results of PSI-BLAST, RPS-BLAST, SSEA-based descriptor and motif-based descriptor are used to evaluate the performance of the fold recognition proposed in this study. These descriptors are evaluated as composite prior to the individual evaluation of each descriptor.

Go-term descriptors: Every protein sequence can be linked with one or more gene-ontology terms categorized under Molecular Functions (MF), Biological Processes (BP) and Cellular Component (CC). General, these GO-terms can also be used to determine the homology of protein sequences. The functional homology score (often Fh score) is used to establish the similarity between query and template sequences based on their functions. Also, these Fh scores can be used to guess and annotated the functions of the proteins. In this study, the GO terms of most hit templates are used to predict the GO-term of the query proteins (Fh cutoff: 0.8).

RESULTS AND DISCUSSION

The server for fold recognition proposed in this research can generate the potential 3 dimensional structure of query protein. This method is based on homologous shared between query and template proteins sequences. The steps involved in this methods are the remote homology determination by PSI-BLAST searched) secondary structure predictions and consensus alignment Hidden Markov Model (HMM) of query sequence and screening against the PSSRDB (Protein Secondary Structure Representative Database) holding experimentally solved protein structures with secondary structure consensus data. Generation of query sequence 3D structure based on HMM alignment between query and templates structure refinement based on loop library amino acid side chain modeling using a rotamer library. The GO termand MeSH terms based homologous is used to optimize the choice of rotamer. The detailed work flow of the proposed system is outlined in Fig. 1.

The present proposed method is efficient in producing about 70% accurate models based on the template domains. This rate of accuracy is achieved due to the identification of remote homologous proteins by implementing the profile-profile and HMM alignment strategies. However, this proposed method has its own drawbacks while the query and template sequence homology falls in twilight zone (<15%).

Secondary structure prediction: The secondary structure of protein is predicted as 3-state: α-helix, β-strand or coil. The prediction of SSE were given as Hh, Ee, Cc while the unknown (disordered) regions in query protein are indicated by question marks (?). However, these regions might be functionally very important. Yet, there is not much attention paid to predict the structure at these regions. It has been observed that on average 78-80%

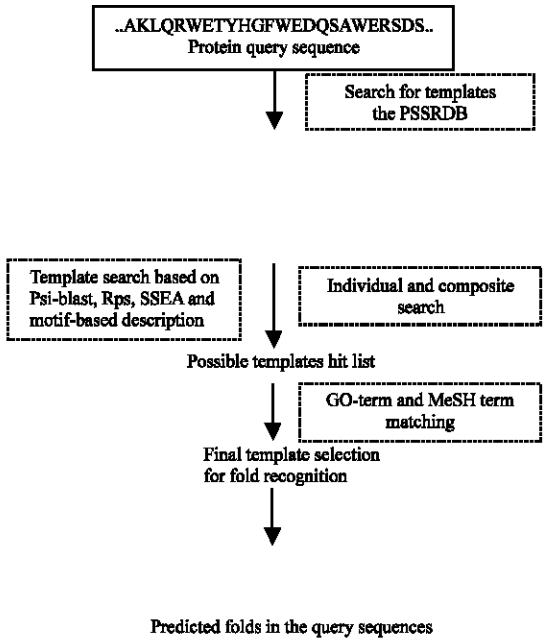

Fig. 1: The fold recognition flow chart proposed in this study

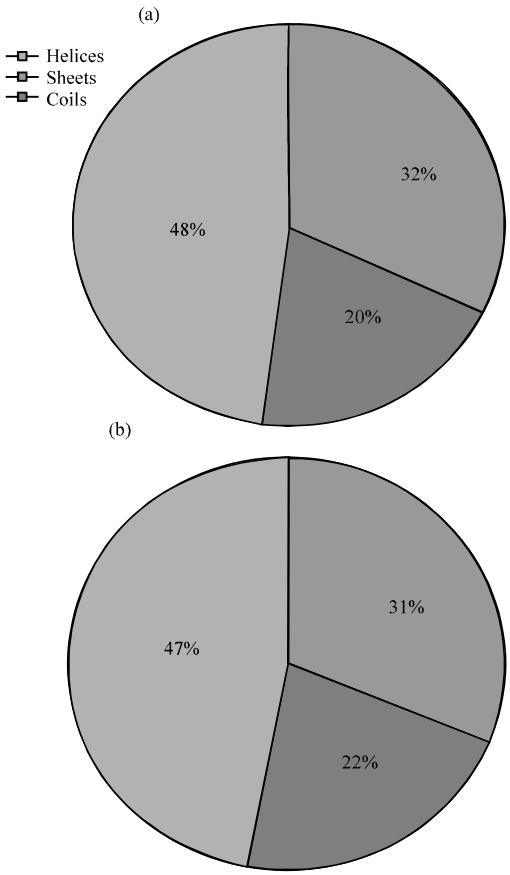

Fig. 2: The correctness in secondary structure prediction for the dataset from PSSRDB; a) trainung set and b) test set

accuracy is achieved in secondary structure prediction. The percentage of secondary structure elements predicted by using test and training dataset are shown in Fig. 2.

On the other hand, this level of exactness might be due to the large amount of varied sequence homologues in the PSSRDB. If query sequence shares less similarity, then precision rate falls to roughly 65-70%. In the secondary structure predictions the elements such as π-helices, or 3_{10}-helices, turns, bends are merged together and treated as α-helices and coils. In addition, the composition of amino acids and their properties also plays a vital role in secondary structure prediction. For instance, polar amino acids (A, G, P, S, T), hydrophobic (I, L, M, V), charged (D , E, H, K, N, Q, R) and aromatic (C, F, W, Y) (Fig. 2).

Template information: The homology shared between the template and query sequences is given in terms of alignment score. This is purely based on the number of matched and mismatched and gapped residues. Also, the secondary structure prediction and its similarity are considered for the choice of template. The template selected for the construction of models will be displayed. the generate models co-ordinate files in PDB format will be provided to download and can ve viewed in any of the support visualizations tools such as rasmol, jmol etc. The percentage of similarity between query and template alignment is also displayed as Structural homology (Sh) score (this is not the modeled structure accuracy). Based on which the users can also use the selected template of their choice to model the protein. In general, the proposed system will use the template with highest Sh score which can significantly produce the high accuracy models. For instance, the Sh core of (>90%) can produce the model with 2-3 A^0 of RMSD (Root Mean Square Deviation). This might be due to the possibility of exact fold recognitions. However, the fold recognition plays a crucial role in protein modeling while the template with low Sh core exists. In line with this, the proposed systems can generate the high accurate models with Sh score between 30-40% or above. Also, some times, the models with moderate accuracy can also be generated for query sequences that shares lower Sh score (<15%) with the templates. However, the proposed system is not enough significant in generating even the low accurate models when the Sh score of query-template falls below 14% or

lower. Also, the template secondary structure and its consensus taken form PSSRDB will also be presented for the users references. Further, predicted GO-terms of the query sequence and template sequences are also given as Fh (Function homology) score. These terms are used as major descriptors in predicting the functional domains of the query sequences and also used to identify the functional domains by providing the matched regions on template. Based on the hypothesis that the GO-terms are highly relevant to classify those proteins molecular functions, biological process and functions, the annotation of query protein function can be determined without any error. However, at certain cases the chances of error are high when the Fh score is very low. In this study, it is observed that 50% of the native GO terms are properly recognized from template with 0.8 Fh-score. Also, the MeSH (Medical Subject Heading) terms used as descriptors for identify subset of protein specificity.

Alignment: The models generated in this proposed systems is based on the query template alignment (Sh score), HMM alignment, secondary structure, GO term (Fh score) and MeSH term similarities. The alignment between query and template along with the predicted secondary structure and its consensus are used in combination to reveal the Sh scores. The alignments with most disordered secondary structure are also displayed separately. Thus, the user can make use of the both the template for modeling. In the predicted secondary structure of the templates, the DSSP (Dictionary of Secondary Structure of Proteins) notations secondary structure notations such as S (bend), T (hydrogen bonded turn), G (3_{10} helix), I(π helix) and B (β-bridge) characters are also included. The secondary structures predictions were given as eight state assignments (Hh, Gg, Ii, Bb, Ee, Tt, Ss, Cc) and then reduced to three states H (Helices), E (sheets), C (Coils), so as to improve the accuracy and the generation of secondary structure information representation in the form of summary (e.g., predicted secondary structure for a sequence is CCCHHcChHHHcCchHhHHhCCC, the generated summary will be ChcChHcCchHhHhC). Thus, the template and query conserved with these SSE (3 states) are considered for the modeling.

Performance of proposed fold recognition serve: The performance of the proposed fold recognition serveris assessed by secondary structure elements alignment, PSI-BLAST, reverse PSI BLAST (RPS-BLAST), motif, GO-terms and MeSH terms based methods individually (Table 1). In which the performance of the each methods on the PSSRDB dataset is very discouraging. However, the combination of these methods on this PSSRDB data set has shown remarkable increase in gthe performance in term of sensitivity. Thus, the comparison of all the methods was determined to explore the descriptors that can be used to explore the remote homology proteins. The combination of six methods has shown an raise in the sensitivity of the homologies predicted based on the query sequence (Table 2).

Table 1: The performance of individual descriptors used in the fold recognition servers

Descriptors	Sensitivity (%)
SSE	28.55
PSI-BLAST	36.83
RPS-BLAST	37.48
MOTIF	19.61
GO-term	30.04
MeSH term	23.34

Table 2: Sensitivity based on the different descriptors combination

Descriptors	Sensitivity (%)
SSE and PSI-BLAST (Set A)	48.02
Set A and RPS BLAST (Set B)	51.06
Set B and motif (Set C)	55.86
Set C and GO term (Set D)	68.01
Set D and MeSH term	72.04

For future development, the templates in the PSSRDB are concerned to regularly update and are provided withwide-ranging of proteins with experimentally determined structures will be included as fold library. Further, the inclusions of new descriptors such as physio-chemical properties of amino acids are in consideration to enhance the prediction accuracy of this proposed fold recognition server. However, the inclusion of new descriptors will certainly complicate the prediction results of server and interferes with the performance of individual descriptors. Thus, it is always important to carefully assess the new descriptors performance before including in the fold recognition server. While this sort of more descriptors inclusions into a fold recognition system are favored by machine learning techniques such as SVM, BPN and KNN methods which may ultimately lead to the better performance of the system. Thus, the development of fold recognition systems based on machine learning methods will result in high accuracy models and can establish the relationship between sequence, function and structure.

CONCLUSION

In the present research, we focused to suggest a protein fold recognition server that predicts based on the proteins affiliated GO term and MeSH term and by a composite model with these factors as descriptors. The propsed method explored a better performance based as it captures the functional and evolutionary information of

proteins. The combination of GO term with the other descriptors has resulted in the performance accuracy of 68.01% and the inclusion of MeSH terms (72.04%) to this combination has shown in 4.06% raise in the performance. This significantly, envisages that the inclusion of new descriptors might result in the better performance of fold recognition servers. Also that the development of machine learning based fold recognition methods can result in high accuracy models which can significantly establish the relationship among sequence, function and structure.

REFERENCES

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

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

Ginalski, K., J. Pas, L.S. Wyrwicz, M.V. Grotthuss and J.M. Bujnicki *et al*., 2003. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res., 31: 3804-3807.

Jaroszewski, L., L. Rychlewski, Z. Li, W. Li and A. Godzik, 2015. FFAS03: A server for profile-profile sequence alignments. Nucleic Acids Res., 33: W284-W288.

Kelley, L.A., R.M. MacCallum and M.J. Sternberg, 2000. Enhanced genome annotation using structural profiles in the program 3D-pssm. J. Mol. Biol., 299: 499-520.

McGuffin, L.J. and D.T. Jones, 2003. Improvement of the GenTHREADER method for genomic fold recognition. Bioinf., 19: 874-881.

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

Pearson, W. R., 1990. Rapid and sensitive sequence comparisons with fastap and fasta. Method. Enzymol., 183: 63-98.

Petrey, D. and B. Honig, 2005. Protein structure prediction: Inroads to biology. Mol. Cell, 20: 811-819.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng., 12: 85-94.

Schaffer, A.A.I., Y.P. Wolf, C. Ponting, V.E. Koonin and L. Aravind *et al*., 1999. IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinf., 15: 1000-1011.

Shi, J., T.L. Blundell and K. Mizuguchi, 2001. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure- dependent gap penalties. J. Mol. Biol., 310: 243-257.

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

Sonnhammer, E.L., S.R. Eddy and R. Durbin, 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins Struct. Function Genet., 28: 405-420.

Wu, S. and Y. Zhang, 2008. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Struct. Funct. Bioinf., 72: 547-556.

Yona, G . and M. Levitt, 2002. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. J. Mol. Biol., 315: 1257-1275.

Zhang, W., S. Liu and Y. Zhou, 2008. SP5: Improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PloS One, 3: 1-6.

Zhang, Z., S. Kochhar and M.G. Grigorov, 2005. Descriptor-based protein remote homology identification. Protein Sci., 14: 431-444.