VOL. 10, NO. 23, DECEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A COMPUTATIONAL METHOD FOR PROTEIN DOMAIN PREDICTION BY USING DOUBLE STAGE NEURAL NETWORK

U. H. Kalsum[1], Nazri Mohd Nawi[2] and Shahreen Kasim[2]
[1]Faculty of Computer Science and Information Technology, Kolej Poly-Tech Mara Batu Pahat, Johor, Malaysia
[2]Faculty of Computer Science and Information Technology, Soft Computing and Data Mining, Universiti Tun Hussein Onn, Batu Pahat, Johor, Malaysia
E-Mail: umi_kalsum@gapps.kptm.edu.my

ABSTRACT

Protein domains are basic unit of protein structure that can develop its self by using its own shape and their own function. Protein domain prediction is important for multiple reasons, which include predicting the protein function in order to manufacture new protein with new function. However, there are several issues that need to be addressed in protein domain prediction which the protein domain can exist in more than one categories of single or multiple domain. Therefore, this study proposed a computational method to predict protein domain using double stage neural network in order to handle these issues. The proposed computational method consists of three phases: dataset generating, profile descriptor for PDP and classification. The pre-processing phase involves datasets generation, splitting protein sequence into subsequence, perfume multiple sequence alignment (MSA) and extracting the MSA. All these process are introduce in order to increase the domain signal. The profile descriptor for PDP phase used several measures such as entropy, correlation, protein sequence termination, contact profile, physio-chemical properties and intron-exon boundaries to generate protein structure information in order to show clear domain signal. The classification phase involves classification by double stage neural network (DNN) and performance evaluation. The performance of the proposed method is evaluated in terms of sensitivity, specificity and accuracy on single domain and multiple-domain using dataset SCOP 1.75. The results showed that the proposed method achieved better results compared with single neural network (SNN) in single domain and multi domain predictions.

Keywords: protein domain, protein structure, neural network.

INTRODUCTION

Protein domain is structural and functional units of protein that exist in protein sequence. A protein domain also known as comprises of protein domain boundary that relates to a part in amino acid residue (protein sequence) where each residue in the protein chain is defined as domain position start and end. These domains structural exists independently. The independent structural of protein domain means that it can often be found in proteins with the same domain content, but in different orders or in different proteins sequences. The knowledge of protein domain is important to analyse the different functions of protein sequences.

Nowadays, it is not only important to predict a protein function accurately from large numbers of protein sequences with unknown structure [1], but it is also very important to predict protein domain boundaries of protein sequence. The problem in protein domain prediction is to determine the start and end of protein domain boundaries in protein sequences, since the protein sequences alone contain structural information that is only available in small portion of the protein space. Therefore, the prediction of protein domain based on protein sequence only is more likely to introduce incorrectly folded regions and making the decision a difficult task. The protein sequence may be contained of single-domain or several domains with different or matching copies of protein domain. Each category of domain may have specific function associated with it. Therefore, predicting the protein domain into their related category such as single-domain or multiple domains is becoming very important for researchers to understand the protein structure, function and biological processes.

Currently there are several computational protein domain boundaries determination's methods available such as method based on similarity and multiple sequence alignment, known protein structure, dimensional structure, used model based and protein sequence information. Mostly, previous protein domain predictor methods produce good results in cases of single-domain prediction. Methods that use dimensional structure to assume protein domain based on the same general principle that assumes domains to be structurally compact and separate substructures with higher density of contacts within the substructures than with their surroundings such as work done by Sadowski [2] and Barenboim *et al.*[3]. Methods that depend on known protein structure to identify the protein domain boundary since structural data are available for only a relatively small number of proteins. Several methods handle the problem of domain start and end signal by employing structure prediction methods by using other types of known information such as work done by Porter and Rose [4]. HHMP fam [5] and HMMSMART [6] have been applied in methods that used comparative model such as HMM to identify other member of protein domain family. Methods based on similarity and use multiple sequence alignments to represent protein domain. A sequence database search provides information on pairwise similarities such as work done by Jimenez-Roldan *et al.*[7] and Katagiri *et al.*[8]. Lastly, methods that

are based only on sequence information to provide an appealing alternative, especially for large-scale domain boundaries determination. Sequence information can be utilized in many ways, the most obvious of which is sequence similarity such as work done by Drew *et al*. [9] and Garcia *et al.* [10].

Previously, single Neural Network (SNN) is used in protein domain detection such as in the work of Armadillo [11], DOMpro[12] and DOBO[13]. The development of SNN followed a heuristic path, with application and extensive experimentation preceding theory. A neural network is an adaptable system that can learn relationships through repeated presentation of data and is capable of generalizing to new, previously unseen data [14]. It is so powerful because it can learn any desired input-output mapping if they have sufficient numbers of processing elements in the hidden layers [15].

This propose method used multi-stage filtering and the output of the previous training to be as input to the next stage which we called it as double-stage Neural Network (DNN) approach. Multi-stage filtering is believe in order to improve the clustering [16]. The purpose of this paper is to discover the capability of the proposed DNN approach in predicting the protein domain boundaries. Next section will introduce about the approach method beginning from data generation, protein structure information generating and DNN works.

EXPERIMENTAL SET UP

The proposed computational method contains three phases namely the data generating phase, profile descriptor for PDP phase and classification phase. The preview of the proposed computational method is shown in Figure-1. This method begins with phase 1 that contains three parts: generate the dataset, splitting the protein sequence and performing multiple sequence alignment. The splitting protein sequence task is applied to split the protein sequence with more than 600 residues into protein subsequences based on ordered and disordered regions [18]. Multiple sequence alignment generated from the protein subsequences is expected to produce higher similarity. While, extracting multiple sequence alignment (MSA) task is applied in order to make sure the protein domain boundaries are shown clearly.

Profile descriptor for PDP in phase 2 will use several measures to predict the information of protein sequences based on information from extraction of MSA. This information will be used for DNN as input and reflected in protein domain prediction. Lastly, the results from DNN will be evaluated in term of sensitivity, specificity and accuracy.

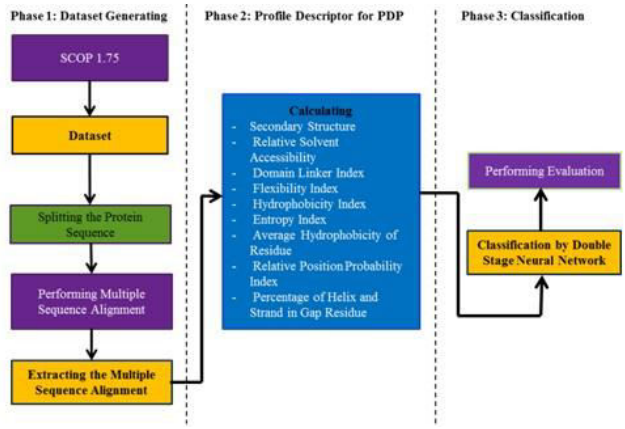

Figure-1. Proposed approach protein domain prediction using double stage neural network.

DATASET GENERATING

The performance of the proposed method will be tested using SCOP database version 1.75 with 9536 protein sequences where protein sequences that are shorter than 40 amino acids are discarded. The protein sequences are then matched with PDB where 95% of identical match in PDB is retained, which means the valid protein sequences are left at 1070.

Then, the 1070 protein sequences are split randomly in the 80:20 ratios into training and testing datasets. The training dataset is used for optimizing the DNN parameter and for classifying the protein domain into single-domain and multiple-domain. The testing dataset is used to evaluate the performance of the DNN. Lastly, multiple sequence alignment is performed using Clustal Omega algorithm [19] where alignments are represented as a protein sequence of alignment column that associated with one position in protein sequence.

Then extracting the pairwise alignments generated by Clustal Omega. In the extraction, a domain boundaries signal is defined as a gap which begins at the N or C terminal. The gap with 45 residues or more will remove and the continuous sequence over 45 residues will remain for generating the protein domain signal. The extractions of pairwise alignment are expected to increase PSI-BLAST e-value [20]. The information of domain boundaries are very importance in protein domain prediction.

PROFILE DESCRIPTOR FOR PDP

Several measures are used to generate the structural information of protein sequences. The information is used for DNN as an input and reflects in protein domain prediction. The measures are secondary structure, relative solvent, domain linker index, flexibility index, hydrophobicity index, entropy index and percentage of helix and strand in gap residue. The protein secondary structure is classified into three: alpha-helices, beta-sheet and coil and contain domain signal. We obtain entropy index and relative solvent according to the profile of amino acids. Entropy index measuring the conservation of an alignment can be computed by information entropy.

The domain linker index and flexibility are used to get flexibility information of protein sequences and the flexibility is used to predict the protein domain. The left and right of sequence termination are calculated to identify the strong signal of protein domain boundaries. The strand in gap residue contains intron-exon structure at DNA level and intron-exon position is used to detect the protein domain boundaries. All the information is used to predict protein domain boundaries.

DOUBLE STAGE NEURAL NETWORK

Double stage neural network (DNN) is multilayer Neural Network (NN). DNN involves training and validation process. First stage in DNN involves the task to create structural of protein domain for each protein sequence. This task used windows of protein structure as an input obtain from sequence profile description. While, second stage in DNN used the output from first stage in order to predict protein domain. The prediction of protein domain either single or multiple domains identified based on its homology where, the homologues are obtained from extraction of MSA and structural domain.

The DNN learn the task of windows identification by training the windows as input and windows ID as output. The input used 15 windows and each residue in windows contain 15 nodes. The first stage has 225 inputs (15*15) and the output from training are tested using testing data. The windows ID from first stage are used to select DNN model for next stage. The input for the first layer of DNN is 15 nodes. 3 node from secondary structure, 1 node from relative solvent accessibility, 2 node from domain linker index, 2 node from flexibility index, 2 node from hydrophobicity index, 2 node from entropy index, 2 node from average hydrophobicity of residue near N or C terminal and 1 node from relative position probability index.

The second stage in this algorithm is in order to develop a set of DNN model where each model corresponding to the windows ID from first stage. Input for second stage is the windows from first stage. The node used in second stage is 19 nodes including 4 nodes from percentage of helix, strand residue from n and C terminal. Then, train the DNN in order to learn the data from first stage. The output for each model is prediction of protein domain. After that, the results are tested using testing data.

EVALUATION

The method is tested and evaluated for the result of single-domain and multiple-domain prediction based on sensitivity (SN) and specificity (SP). Protein sequences which contain only one domain are predicted as single-domain prediction. Meanwhile, if a protein sequence contains more than one domain, it is defined as multiple-domain prediction. In order to analyze the performance, the proposed method is compared with Single Neural Network (SNN).

The sensitivity and specificity are calculated based on the length distribution of true domain boundary regions and false domain boundary regions. The sensitivity measures the proportion of actual positives which are correctly identified as single-domain and multiple-domain. The specificity measures the proportion of negatives which are correctly identified as not a single-domain or multiple-domain. Higher sensitivity and specificity represents better results in protein domain prediction. The SN and SP are defined as follows:

$$SN = \left(\frac{TP}{TP + FN} \right) * 100 \quad (1)$$

where TP is the number of true positive and FN is the number of false negative of protein domain.

$$SP = \left(\frac{TP}{TP + FP} \right) * 100 \quad (2)$$

where FP is the number of false positive of protein domain.

Overall performance of protein domain prediction for single or multiple domain will evaluate based on the accuracy (AC) of prediction. The accuracy will calculate the percentage of the correctly predicted protein domain. The AC is defined as follows:

$$AC = \left(\frac{(TP - TN)}{(TP + FN + TN + FP)} \right) * 100 \quad (3)$$

RESULT AND DISCUSSION

The proposed computational method is tested and compares its performance with SNN. The properties of protein sequence are detected from several measures as mention previously generate a strong signal [18] of protein domain boundaries since the result of sequence profile is used as input to DNN model. The data generating task is effect to generate the protein sequence profile since we only take valuable protein sequence as dataset. The process of splitting long protein sequence into a subsequence also gives more domain signal since before this the long sequence only detect one domain signal. The extracting MSA give more clearly about domain boundaries. These task help in generating sequence profile in order to make sure the profiles are clearer about domain signal. Then, the properties of protein sequences are used as input in the DNN model. Finally, the results generated by DNN are evaluated in term of sensitivity and specificity. The results showed that DNN is ranked highly or better than SNN in some aspects.

The datasets obtained from SCOP 1.75 that have been defined in the previous section are used to test and evaluate the DNN and SNN. The results of the accuracy prediction including sensitivity and specificity for single-domain and multiple-domain are presented in Table-1 and Figure-2. It is easy to see that predicting multiple-domain is more difficult than predicting single-domain. The DNN achieved a higher sensitivity of 78% for single-domain and 87% for the multiple-domain compared to SNN. The DNN achieved a higher specificity of 79% for the single-domain and 82% for the multiple-domain compared to

SNN. The DNN produces better result of 85% accuracy for single-domain and 81% accuracy for multiple-domain compared with SNN.

The method is developed from the detailed investigation on protein domain prediction, strength of domain signal, misleading of domain signal and classification algorithm. Based on the investigation, the proposed method generates information of protein structure from MSA extraction in order to give a strong signal of protein boundaries. This shows that the protein structure information does have an effect on the protein domain boundaries prediction. The DNN focuses on the use of scores of measures to detect the protein domain region in order to classify a domain into single domain or multiple domains. The first stage of DNN used in order to generate the domain structural. After that, the second stage of DNN used to predict protein domain into single domain or multiple domain based on domain structural from first stage of DNN. These tasks showed an improvement result compare with SNN.

Table-1. Sensitivity and specificity for single and multiple domain prediction.

	Single Domain			Multiple Domain		
	SN	SP	AC	SN	SP	AC
SNN	0.68	0.73	0.71	0.76	0.71	0.73
DNN	0.78	0.87	0.85	0.79	0.82	0.81

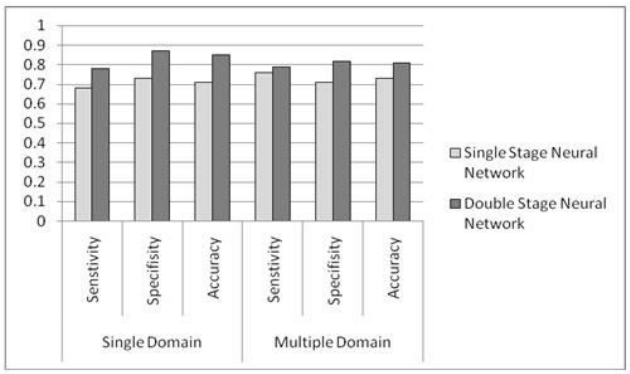

Figure-2. Sensitivity and specificity chart for single and multiple domain prediction.

CONCLUSIONS

Currently, several computational methods including tools have been developed to predict protein domain. However, those computational methods produce good result only in single-domain prediction since the current methods does not emphasis the strength and misleading of domain signal. Therefore, this method introduce for increasing the domain signal in order to predict protein domain accurately.

The method begins with searching the seed protein sequences as dataset from SCOP 1.75. The dataset is split into training and testing with 80:20 ratios. Then, applied split technique to protein sequence and performed MSA. After that, extract the MSA to remove gap data. Then, generate protein sequence information using several measures such as secondary structure, relative solvent, domain linker index, flexibility index, hydrophobicity index, entropy index and percentage of helix and strand in gap residue to increase signal of protein domain. Then, DNN is applied to predict and classified the protein domain into single and multiple domain. Lastly, the result from proposed approach evaluated in term of sensitivity and specificity and compare with SNN evaluation. The proposed approach has shown the improvement of prediction with 85% accuracy in single domain prediction and 81% accuracy in multiple domain prediction compare with SNN. Therefore, the increasing information of protein sequence information is believed increase the domain signal. When domain signal is increasing, the prediction of protein domain into single or multiple domains is more accurate.

ACKNOWLEDGEMENTS

This project is funded by the Universiti Tun Hussein Onn Malaysia under Research Acculturation Collaborative vote no. 1447. Many thanks to GATES IT Solution Sdn Bhd for ideas and collaboration.

REFERENCES

[1] Loh Swee Kuan, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, Shahreen Kasim, Choon Yee Wen, Zuwairie Ibrahim, Bambang Susilo, Yusuf Hendrawan, and Agustin Krisna Wardani. 2015. A Review of Software for Predicting Gene Function. International Journal of Bio-Science & Bio-Technology. Vol. 7, No. 2, pp. 57-70.

[2] Sadowski and Michael I. 2013. Prediction of Protein Domain Boundaries from Inverse Covariances. Proteins. Vol. 81, No. 2, pp. 253-260.

[3] Barenboim Maxim, Majid Masso, Iosif I. Vaisman, and D. Curtis Jamison. 2008. Statistical Geometry Based Prediction of Nonsynonymous SNP Functional Effects Using Random Forest and Neuro-fuzzy Classifiers. Proteins. Vol. 71, No. 4, pp. 1930-1939.

[4] Porter L. L. and Rose G. D. 2012. A thermodynamic definition of protein domains. In: Proceedings of the National Academy of Sciences. S. Walter Englander (Eds.). pp.109:930.

[5] Bateman A. 2004. The Pfam Protein Families Database. Nucleic Acids Research. Vol. 32, No. 1, D138-D141.

[6] Ponting C. P., Schultz J., Milpetz F. and Bork P. 1999. SMART: Identification and Annotation of Domains from Signaling and Extracellular Protein

Sequences. Nucleic Acids Research. Vol. 27, No. 1, pp. 229-232.

[7] Jimenez-Roldan J. E., Freedman R. B., Romer R. A. and Wells S. A. 2012. Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses. Physical Biology. Vol. 9, No. 1, p. 016008.

[8] Katagiri D., Hideyoshi F., Saburo N. and Tyuji H. 2008. Ab Initio Protein Structure Prediction with Force Field Parameters Derived from Water-Phase Quantum Chemical Calculation. Journal of Computational Chemistry. Vol. 29, No. 12, pp. 1930-1944.

[9] Drew K., Winters P., Butterfoss G. L., Berstis V., Uplinger K., Armstrong J., Riffle M., Schweighofer E., Bovermann B., Goodlett D. R., Davis T. N., Shasha D., Malmström L. and Bonneau R. 2011. The Proteome Folding Project: proteome-scale prediction of structure and function. Genome Research. Vol. 21, No. 11, pp. 1981-1994.

[10] Garcia B., Ricardo A., Agapito L. and Araceli S. 2008. Protein-Protein Functional Association Prediction using Genetic Programming. In: Proceedings of the 10th annual conference on genetic and evolutionary computation of the Atlanta. Maarten Keijzer (Eds.). pp. 347-348.

[11] Dumontier M., Rong Y., Howard J. F., and Christopher W. H. 2005. Armadillo: Domain Boundary Prediction by Amino Acid Composition. Journal of Molecular Biology. Vol. 350, No. 5, pp. 1061-1073.

[12] Cheng J., Michael J. S. and Pierre B. 2006. DOMpro: Protein Domain Prediction using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. Data Mining and Knowledge Discovery. Vol. 13, No. 1, pp. 1-10.

[13] Eickholt J., Xin D. and Jianlin C. 2011. DoBo: Protein Domain Boundary Prediction by Integrating Evolutionary Signals and Machine Learning. BMC Bioinformatics. Vol. 12, No. 43, pp. 1471-1504.

[14] Mohorianu S., Lozovan M. and Baciu C. 2007. A new simulation method based on artificial neural networks for a special class of nanomagnetic materials design. Journal of Optoelectronics and Advantage Materials. Vol. 9, No. 5, pp. 1499-1504.

[15] Kalsum U. Hassan, Razib M. Othman, Rohayanti Hassan, Hishammuddin Asmuni, Jumail Taliba and Shahreen Kasim. 2012. Recurrent Neural Networks and Soft Computing. First Ed. Mahmoud Elhefnawi and Mohamed Mysara. InTech, Europe. pp. 134-150.

[16] Shahreen Kasim, Safaai Deris, Razib M Othman. 2013. Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data. Computers in biology and medicine. Vol. 43, No. 9, pp. 1120-1133.

[17] Hamp T., Birzele F., Buchwald F. and Kramer S. 2010. Improving structure alignment-based prediction of SCOP families using Vorolign kernels. Bioinformatics. Vol. 27, No. 2, pp. 204-210.

[18] Kalsum Hassan U., Zuraini A. Shah, Razib M. Othman, Rohayanti Hassan, Shafry M. Rahim, Hishammudin Asmuni, Jumail Taliba, and Zalmiyah Zakaria. 2009. SPlitSSI-SVM: an algorithm to reduce the misleading and increase the strength of domain signal. Computers in Biology and Medicine. Vol. 39, No. 11, pp. 1013-1019.

[19] Sievers F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. Mcwilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. 2011. Armadillo: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. Vol. 7, p. 539.

[20] Henikoff J. G., Greene E. A., Pietrokovski S. and Henikoff S. 2000. Increased Coverage of Protein Families with the BLOCKS Database Servers. Nucleic Acids Research. Vol. 28, No. 1, pp. 228-230.