Bangladesh J. Environ. Sci., Vol. 46, 73-76, 2024 @ BAED ISSN 1561-9206

COMPUTATIONAL TOOLS FOR SEQUENCE BASED IDENTIFICATION AND CHARACTERIZATION OF ANTIFREEZE PROTEINS

A. Mimi[1], M. R. Amin[1], M. A. Haque[2], S. M. Ahmed[3] and M. Z. Tareq[1]
[1]Genome Research Centre, BJRI, Dhaka-1207
[2]Dept. of Biotechnology and Genetic Engineering, Islamic University, Kushtia-7003
[3]ACI Molecular Genetics, ASRBC, ACI Agri-business, ACI Limited, Gulshan, Dhaka

ABSTRACT

Living organisms including fishes, microbes, and animals can live in extremely cold weather. To stay alive in cold environments, these species generate antifreeze proteins (AFPs), also referred to as ice binding proteins. Moreover, AFPs are extensively utilized in many important fields including medical, agricultural, industrial, and biotechnological. Several predictors were constructed to identify AFPs. However, due to the sequence and structural heterogeneity of AFPs, correct identification is still a challenging task. In this study, deals with several bio-informatical or computational tools which have been proposed for prediction of AFPs more precisely can predict AFPs more accurately and can participate in a significant role in medical, agricultural, industrial, and biotechnological fields.

Key words: Antifreeze protein, bioinformatics, physicochemical properties, homology modeling

Introduction

Computational tools provide researchers a cost effective way to understand physicochemical and the structural aspects of a protein for the successful design of many biological experiments with in a limited time and these methods are not amenable to high throughput techniques. Physicochemical characterization studies give more information about the properties such as molecular weight (Mwt.), theoretical isoelectric point (pI), aliphatic index (AI), grand average hydrophathy (GRAVY) and Instability Index (II). These properties are essential and vital for the characterization of proteins and their properties (Pradeep *et al*., 2012). Many lines of evidences have indicated that computational approaches can provide useful information for both drug discovery and basic research in a timely manner (Shao *et al*., 2009) such as protein subcellular location prediction [Chou and Shen 2007(a), Chou and Shen 2008(a)] structural bioinformatics (Chou 2004), identification of proteases and their types [Chou and Shen 2008(b)], identification of membrane proteins and their types [Chou and Shen 2007(b)], molecular docking (Chou *et al*., 2003; Li *et al*., 2007 and Wang *et al*., 2008), identification of enzymes and their functional classes [Shen and Chou 2007(a)], and signal peptide prediction [Chou and Shen 2007(c); Shen and Chou 2007(b)]. Numerous structure and function studies of AFPs have been reported experimentally from time to time while computational study of AFPs are much more limited. Thus it is necessary to know about the various bioinformatical or computational tools for the prediction of antifreeze proteins. Although, each prediction system made efforts to predict antifreeze proteins. However, due to the variant behavior of AFP structure and sequences, it is still highly desirable to predict AFP more accurately. In this study, we will focus on the In silico characterization and homology modeling of AFPs from different sources.

Use of Bioinformatics and Computational tools for AFPs

Several computational approaches and online servers provide great opportunities for the characterization and analysis of protein to accelerate experimental approaches as well as widening scientific thoughts. Computational tools provide researchers a cost effective way to understand physicochemical and the structural properties of a protein for the successful design of many biological experiments with in a short range of time. A series of methods have been established for identification of AFP. For example,

Kandaswamy *et al*. developed a method, called AFP-Pred to discriminate AFP from non-AFP. They used short peptides, secondary structure properties, physicochemical features, and RF (Random Forest) as training model (Kandaswamy *et al*., 2011). In another approach (AFP-PSSM), these authors utilized evolutionary information in conjunction with SVM (Zhao and Yin, 2012). Yu *et al*. adopted multi-respective several composition features such as TPC, DPC, and AAC. They selected the best patterns via genetic algorithm and prediction was carried out by SVM. They also established a web server, called iAFP (Yu and Lu, 2011). Onward, Mondel *et al*. proposed AFP-PseAAC predictor employing PseAAC (Pseudo Amino Acid Composition) with SVM (Mondal and Pai, 2014). In RAFP-Pred predictor, authors split each protein sequence into two sub-sequences. Features from each part were abstracted by AAC and DPC. Info-Gain algorithm was implemented for selection of optimal features and the model was trained using RF classifier. Usman *et al*. proposed AFP-LSE predictor. They used autoencoder with Composition of K-spaced amino acid pairs and achieved a balanced accuracy of 0.903. In another work, Usman *et al*. constructed AFP-SRC improved method. Similarly, PoGB-pred approach is developed by Alim *et al*. (2021). They employed PseAAC, AAC, and DPC as feature descriptors and PCA for reducing the feature dimension. Recently, Miyata *et al*. designed a novel predictor using new datasets. They applied CD, DC, AAC for feature encoding and Light eXtreme Gradient Boosting machine for model learning. A novel computational method, named AFP-LXGB has been proposed for prediction of AFPs more precisely (Khan *et al*., 2022, Usman *et al*., 2020). verified that AFP-LXGB can predict AFPs more accurately and can participate in a significant role in medical, agricultural, industrial, and biotechnological fields.

Physicochemical properties

The physicochemical properties of primary sequences of proteins helps in determining both the structure and biological functions. The sequence analysis of the proteins and nucleic acids is most fundamental element of bioinformatics. Research has shown that the natural biological properties of peptides are a complex combination of hydrophobicity, charge, molecular mass, reduction of the hydrophobic moment (Fjell *et al*., 2012) and other physicochemical characteristics which can provide useful information in classifying, predicting, and synthesizing new peptides (Manavalan *et al*., 2017). The physicochemical properties were calculated from the primary structure of antifreeze protein where the physicochemical parameters, theoretical isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient (Gill and Von Hippel, 1989), half-life (Tobias *et al*., 1991), instability index (Guruprasad *et al*., 1990), aliphatic index (Ikai *et al*., 1980) and grand average hydrophathy (GRAVY) (Kyte and Doolo-ttle, 1982) were computed using the Expasy's Prot-Param (Gasteiger *et al*., 2005) (http://us.expasy.org /tools/protparam.html) prediction server.

Functional characterization and secondary structure analysis

The secondary structure of a protein is primarily determined by the interactions between the amino acid residues, such as hydrogen bonding, van der Waals forces, and hydrophobic interactions. It is important for protein stability, function, and overall three-dimensional structure. The identification of transmembrane regions of a protein was identified by SOSUI server. The predicted transmembrane helices were visualized and analyzed using Helical Wheel Plots. SOPMA (Geourjon and Deleage, 1995) was employed for calculating the secondary structural features of the antifreeze proteins. Computational methods were also applied for determining disulphide bonds. Disulphide bonds are very essential in determining the functional linkage and the stability of a particular protein. The presence of SS bond and their bonding patterns were predicted by CYS_REC (Hossain, 2012).

Homology modeling and validation

Homology models of proteins are of great interest for planning and analyzing biological experiments when no experimental three dimensional structures are available. Many proteins are simply too large for NMR analysis and cannot be crystallized for X-ray diffraction. Protein modeling is the only way to obtain structural information if experimental techniques fail. Therefore, it is an obvious demand to bridge this

'structure knowledge gap' and computational methods for protein structure prediction have gained much interest in recent years (Schwede *et al*., 2003). The modeling of 3D structure of 2 antifreeze proteins was performed by three homology modeling programs Geno3D (Combet *et al*., 2002), Swiss-model (Arnold *et al*., 2006), CPH-models (Nielsen *et al*., 2010). The model-ed 3D structures were evaluated using the online server Rampage, ProQ (Protein quality server) and ProSA. The structure validation of antifreeze proteins was performed by online PROCHECK (Laskowski *et al*., 1996) and What IF (Vriend, 1990).

Conclusion

Accurate identification of AFPs may provide important clues to decipher the underlying mechanisms of AFPs in ice-binding and to facilitate the selection of the most appropriate AFPs for several applications. The tools may also use to uncover the relationship between proteins and ice-crystals as well as, more generally, the adaptation of organisms to their environments depend on the ability to understand the evolution of AFPs.

References

Alim, A., Rafay, A. and Naseem, I. 2021. PoGB-pred: Prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach. *Curr. Bioinform*., 16: 446-456.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. 2006.The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling, *Bioinformatics*., 22:195-201.

Chou, K. C. and Shen, H. B. 2007(a). Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem*., 370:1-16.

Chou, K. C. and Shen, H. B. 2007(b). MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem.Biophys. *Res. Commun*., 360:339–345.

Chou, K. C. and Shen, H. B. 2007(c). Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem.Biophys. Res. Commun*., 357:633-640.

Chou, K. C. and Shen, H. B. 2008(a). Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc*., 3:153-162.

Chou, K. C. and Shen, H. B. 2008(b). ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun*., 376:321-325.

Chou, K.C. 2004. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem*., 11:2105-2134.

Combet, C., Jambon, M., Deleage, G. and Geourjon, C. 2002. Geno3D: Automatic comparative molecular modelling of protein, *Bioinformatics*.,18: 213-214.

Fjell, C. D., Hiss, J. A., Hancock, R. E. and Schneider, G. 2012. Designing antimicrobial peptides: form follows function. *Nature Rev. Drug Discov*., 11:37

Geourjon, C. and Deléage, G. 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments, *Comput Appl Biosci.*, 11:681-684.

Gill, S. C. and Von Hippel, P. H. 1989. *Extinction coefficient, Anal Biochem*.,182: 319-328.

Guruprasad, K., Reddy, B. V. P. and Pandit, M. W. 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence, *Prot. Eng*.,4: 155-164.

He, X., Han, K., Hu, J., Yan, H., Yang, J. Y., Shen, H. B. and Yu, D. J. 2015. Target Freeze: Identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. *J. Membr. Biol*., 248: 1005-1014.

Hossain, M. 2012.Fish antifreeze proteins: Computational analysis and physicochemical characterization. *Int. Current Pharmaceu. J*., 1: 2, 18-26.

Ikai, A. J. 1980. Thermo stability and aliphatic index of globular proteins. *J. Biochem*., 88: 1895-1898.

Kandaswamy, K. K., Chou, K.-C., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S. and Pugalenthi, G. 2011. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol*., 270: 56-62.

Khan, A., Uddin, J., Ali, F., Ahmad, A., Alghushairy, O., Banjar, A. and Daud, A. 2022. Prediction of antifreeze proteins using machine learning. Scientific Reports, 12:20672

Khan, S., Naseem, I., Togneri, R. and Bennamoun 2016. M. Rafp-pred: Robust prediction of antifreeze proteins using localized analysisof n-peptide compositions. IEEE/ACM Trans. *Comput. Biol. Bioinf.*, 15: 244-250.

Kyte, J. and Doolottle, R. F. 1982. A simple method for displaying the hydropathic character of a protein, *J. Mol Biol*., 157: 105-132.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. and Thornton, J. M. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, *J Biomol. NMR*.,8: 477-486.

Li, Y., Wei, D. Q., Gao, W. N., Gao, H., Liu, B. N., Huang, C. J., Xu, W. R., Liu, D. K. Chen, H. F. and Chou, K. C. 2007. Computational approach to drug design for oxazolidinones as antibacterial agents. *Med. Chem*., 3: 576-582.

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O. and Lee, G.2017.DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget*. 8:77121

Miyata, R., Moriwaki, Y., Terada, T. and Shimizu, K. 2021. Prediction and analysis of antifreeze proteins. *Heliyon*., 7: e07953

Mondal, S. and Pai, P. P. 2014. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol*., 356: 30-35.

Nielsen, M., Lundegaard, C., Lund, O. and Petersen, T. N. 2010. CPHmodels-3.0-remote homology modeling using structure-guided sequence profiles, *Nucleic Acids Res*., 38: W576-W581.

Pradeep N. V., Anupama, K. G. and Vidyashree, P. Lakshmi 2012. In silico Characterization of Industrial Important Cellulases using Computational Tools. *Advances in Life Sci. Technol.*, 4: 8-14.

Pratiwi, R., Aijaz, M., Schaduangrat, N. and Prachayasittikul, V. 2017. CryoProtect: A web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem*., 1-15.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M. C. 2003. SWISS-MODEL: an automated protein homology-modeling server, *Nucleic Acids Research*., 13: 3381-3385.

Shen, H. B. and Chou, K. C. 2007(a). EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun*., 364: 53-59.

Shen, H. B. and Chou, K. C. 2007(b). Signal-3L: A 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun*., 363: 297-303.

Tobias, J. W., Shrader, T. E., Rocap, G. and Varshavsky, A. 1991. The N-end rule in bacteria, *Science*., 254: 5036,1374-1377.

Usman, M., Khan, S. and Lee, J. A. 2020. Afp-lse: Antifreeze proteins prediction using latent space encoding of composition of k-spacedamino acid pairs. *Sci. Rep*., 10: 1-13.

Usman, M., Khan, S. and Lee, J. A. 2020. AFP-LSE: Antifreeze proteins prediction Using Latent Space encoding of composition of k-Spaced Amino Acid pairs. *Scientific Reports*, 10:7197

Usman, M., Khan, S., Park, S. and Wahab, A. 2021. AFP-SRC: Identification of antifreeze proteins using sparse representation classifier. *Neural Comput. Appl*., 10:7197

Wang, J. F., Wei, D. Q., Chen, C., Li, Y. and Chou, K. C. 2008. Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein Pept. Lett*.,15: 27-32.

Yu, C. S. and Lu, C. H. 2011. Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PLoS ONE*., 6: e20445.

Zhao, X., Ma, Z. and Yin, M. 2012. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. *Int. J.Mol. Sci.*, 13: 2196-2207.