Review Article

Open Access

An overview of the protein thermostability prediction: databases and tools

Abstract

Thermophilic proteins are characterized as high thermal stability proteins while mesophilic proteins are stable at lower temperatures. These types of proteins have numerous applications regarding protein engineering, drug design and industrial processes. Studies showed that thermal stability is strongly related to structural and sequential properties in thermophilic proteins. Some computational studies were being taken to identify the mentioned properties in heat resistant proteins. This paper reviews the studies of protein thermostability prediction and gives an introduction to the thermal stability related tools and databases.

Keywords: Rotein thermostability, Thermophilic proteins, Mesophilic proteins, Databases, computational methods, Bioinformatics

Volume 3 Issue 5 - 2016

Maryam Mahmoudi,[1] Seyyed Shahriar Arab,[1] Javad Zahiri,[2] Yasaman Parandian[2]

[1]1Department of Biophysics, Tarbiat Modares University (TMU), Iran

[2]Bioinformatics and Computational Omics Lab (BioCOOL), Trabiat Modares University (TMU), Iran

Correspondence: Seyyed Shahriar Arab, Department of Biophysics, School of Biological Sciences, Tarbiat Modares University (TMU), Tehran, Iran Email sh.arab@modares.ac.ir

Received: March 12, 2016 | **Published:** July 5, 2016

Introduction

Environmental temperature plays an important role in the cell life.[1] There are four classes of organism in relation to their optimal growth temperature namely hyperthermophile (>80◦C), thermophile (45-80◦C), mesophile (20-45◦C) and psychrophile (<20 ◦C).[2] Thermal stability is defined as the ability of material to resist changes in physical structure or chemical irreversibility, or spatial structure stability of polypeptide chains at high temperatures.[3] Studies showed that thermal stability of thermophilic proteins is related to a series of protein sequential and structural properties.[4] A small number of these mentioned properties are going to be introduced in this paper. Also, the amino acid compositions difference had been studied in mesophilic and thermophilic proteins.[3,5-7] For instance, Zhang and Gromiha research shows that Lys, Arg, Glu and Pro were higher and Ser, Met, Asp and Thr were lower in number of thermophilic than the of mesophilic proteins number .[6,8] (Figure 1). Protein secondary structure stability like alpha-helix is considered as a necessary factor for thermal stability.[6] Studies suggested that thermal-stability is increased by certain characteristics in proteins. These characteristics are: increased number of hydrogen bonds.[7] salt bridges, ion pairs .[9] aromatic clusters.[8] sidechain-sidechain interactions, electrostatic interactions of charged residues .[9] and hydrophobic interactions.[5]

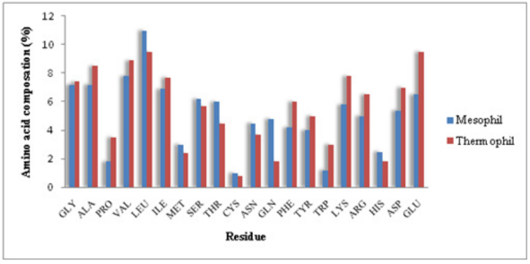

Figure 1 Comparison of Amino acid composition in thermophilic and mesophilic proteins.[6,8]

Protein's Thermal Stability Prediction Methods

Protein's thermal stability can be predicted based on sequence or structure. Both mentioned methods and their corresponding advantages and limitations have been discussed here in further detail. Table 1 demonstrates an overview of the thermal stability prediction methods.

Table 1 An overview of protein thermostability prediction studies

Sequence/Structure Feature	Algorithm	Reference
Amino acid sequence	Support vector machine	10
Primary structure	LogitBoost	12
Amino acid sequence and residues and dipeptide composition	Neural network	8
Primary, secondary and tertiary structure information	Decision tree	11
Amino acid distribution and dipeptide composation	Support vector machine	13
Amino acid composition-based similarity distance	KNN-ID	1
Dipeptide composation	Statistical Methods	14
Amino acid sequence	Genetic Algorithm	15
Thermodynamic parameters	Statistical Potentials	16

Sequence based prediction

This method utilizes sequence information of proteins; for instance, distribution of amino acid and di-peptide composition for discrimination of thermophilic and mesophilic proteins. Studies revealed the differences between amino acid and di-peptide composition in thermophilic and mesophilic proteins. For example, the frequency of Lys, Arg, Glu and Pro was higher in thermophilic than mesophilic protein.[8,10]. These studies also show that the occurrences of EE, KK, RR, PP, KI, VV, VE, KE, and VK were higher in thermophilic proteins while QQ, AA, EQ, LL, NN, QT had lower occurrences.[6] In addition, the frequency of charged, hydrophobic and aromatic amino acids in thermophilic protein is higher than mesophilic ones.[3] Moreover, the correlation between protein amino acid composition and its biological function has been proven.[1] So, the protein sequence analysis provides valuable information to predict protein thermostability; particularly whenever the structural information of proteins is not available.

Structure based prediction

The studies of protein thermostability prediction are based on protein structures utilized protein secondary and tertiary information for discrimination of thermophilic and mesophilic proteins. Important features considered in this studies include amount of secondary structure, ion pairs, hydrogen bonds, disulfide bonds and accessible surface area.[11] Although the thermal stability is directly related to the protein structure stability .[11] Regarding the fact that structural and sequential features affect the thermal stability, applying the both mentioned features at the same time leads to a more accurate, precise prediction. The protein structural information may not be always available; This restrains structure based protein thermostability prediction.

Protein's thermal stability prediction procedure: Several machine learning methods have been applied to predict protein thermostability. Here, we briefly review these methods. Figure 2 provides an illustration of these methods. As illustrated in the figure, in order to predict the thermal stability of proteins, at first, a dataset of thermophil and mesophil proteins is collected from the related databases. Then, proteins are analyzed based on their sequential and structural characteristics. The goal in this stage is to select those features which are significantly important regarding protein thermostability prediction. It should be noted here that considering the structural and sequential features at the same time can produce more precise results. In the next stage, the dataset is going to be divided into the train and test datasets. The train dataset is then used for learning the machine learning algorithm while the test dataset is used to evaluate the model.

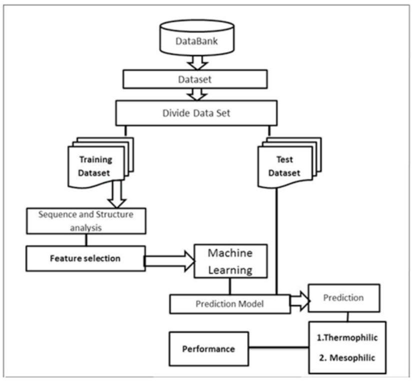

Figure 2 Thermal stability prediction procedure.

Prediction algorithms based on machine learning methods

The following section introduces a few machine learning algorithms. The selected algorithm is going to distinguish the thermophile from mesophile proteins.

a. **Support vector machines (SVMs):** Support vector machines is an machine learning method for classification two classes of data and many kind of kernel functions can be used for classification in this algorithm.[17]

b. **Artificial Neural Networks (ANN):** The ANN concept is inspired by the neural structure of the brain. In this model of prediction, the system is supposed to learn from data - a large number of inputs and solve a wide variety of tasks. ANN software packages can be downloaded from Open NN (Available online: http://www.cimne.com/flood/download.asp).[15]

c. **Decision Tree:** A decision tree is popular machine learning algorithm in bioinformatics and computational biology. It uses a tree-like graph or model of decisions and their possible consequences to classify input instances.

Performance measures

Assessing a prediction tool is a critical task. Table 2 describes commonly used measures for performance prediction assessment: accuracy, sensitivity, specificity, strength, MCC, precision, F-measure and area under the ROC curve (AUC). These measures based on the following four basic parameters:

Table 2 Commonly used measures for performance assessment in protein thermostability prediction.

Expression	A brief description
$Accuracy = \dfrac{\mathrm{TP + TN}}{\mathrm{TP + TN + FP + FN}}$	percent of correct prediction
$Sensitivity = \dfrac{TP}{TP + FN}$	percent of correctly predicted positive
$Specificity = \dfrac{\mathrm{TN}}{\mathrm{TN + FP}}$	percent of correctly predicted negative
$Precision = \dfrac{TP}{TP + FP}$	Positive Predictive Value
$F - measure = \dfrac{2 \times Presion \times Sensitivity}{Presion + Sensitivity}$	The harmonic mean of sensitivity and specificity

a. **True positive (TP):** The number of thermophile proteins, which have been correctly predicted as thermophile.

b. **True negative (TN):** The number of mesophile proteins, which have been correctly predicted by the prediction method as mesophile.

c. **False positive (FP)**: The number of mesophile proteins, which have been incorrectly predicted as thermophile.

d. **False negative (FN):** The number of thermophile proteins, which have been incorrectly predicted by the prediction method as mesophile.

Databases

To build a model capable of predicting the proteins thermal stability; at first, a dataset is created using the related databases. This dataset contains information about the structure and sequence of thermophilic and mesophilic proteins. Table 3 describes a few databases that have been used in studies of protein's thermal stability prediction. According to Table 3, PGT and ProTherm DBs are specifically used to predict the thermal stability. PDB database is used to extract structural information while Uniport gives the sequential information of thermophilic and mesophilic proteins.

Table 3 List of databases in protein thermostability prediction

	Data bases	Note	Ref. Num
General Databases	UniProt	The Universal Protein Resource (UniProt) provides a stable, comprehensive, freely accessible, central resource on protein sequences and functional annotation. This DB is used to extract the sequential information of thermophilic and mesophilic proteins. Availability: http://www.uniprot.org.	18
	PDB	The Protein Data Bank contains information of the 3D structures of large biological molecules, including proteins and nucleic acids. This DB is used to extract structural information of thermophilic and mesophilic proteins. Availability: http://www.rcsb.org.	19

Citation: Mahmoudi M, Arab SS, Zahiri J, et al. An overview of the protein thermostability prediction: databases and tools. *J Nanomed Res.* 2016;3(5):1–6. DOI: 10.15406/jnmr.2016.03.00072

Table continued...

	Data bases	Note	Ref. Num
Specific Databases	Pro Therm	ProTherm is a thermodynamic database that contains experimentally determined thermodynamic parameters of protein stability. This DB is specifically used to predict the thermal stability. Availability: http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.htm	20
	PGT	PGT contains Prokaryotic Growth Temperature database (PGTdb). This DB is specifically used to predict the thermal stability. Availability: http://pgtdb.csie.ncu.edu.tw	2

Conclusion

Due to the recent pervasive use of thermostable proteins and enzymes in industry, protein engineering and other theoretical/experimental studies play a significant role in identification of protein thermal stability. Regarding the high expense rate of laboratory procedures, the employment of theoretical methods for predicting the thermal stability with high accuracy could be so helpful. So far, most computational thermophilic and mesophilic protein identification studies have been solely based on the protein sequence. Regarding the fact that both structural and sequential features affect the thermal stability, applying the both mentioned features at the same time leads to a more accurate, precise prediction.

Acknowledgments

None.

Conflicts of interest

None.

References

1. Zuo YC, Chen W, Fan GL et al. A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins. Amino acids. 2013;44(2):573–580.
2. Huang SL, Wu LC, Liang HK et al. PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics. 2004;20(2):276–278.
3. Zhou XX, Wang YB, Pan YJ et al. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino acids. 2008;34(1):25–33.
4. Vieille C, Zeikus GJ Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol and Mol Biol Rev. 2001;65(1):1–43.
5. Zhang G, Fang B Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Process Biochemistry. 2006;41(3):552–556.
6. Zhang G, Fang B Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process biochemistry. 2006;41(8):1792–1798.
7. Jahandideh S, Abdolmaleki P, Jahandideh M et al. Sequence and structural parameters enhancing adaptation of proteins to low temperatures. J Theor Biol. 2007;246(1):159–166.
8. Gromiha MM, Suresh MX Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins. 2008;70(4):1274–1279.
9. Kumar S, Tsai CJ, Nussinov R Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochem. 2001;40(47):14152–14165.
10. Zhang G, Fang B Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Prot pept lett. 2006;13(10):965–970.
11. Wu LC, Lee JX, Haung HD et al. An expert system to predict protein thermostability using decision tree. ESWA. 2009;36(5):9007–9014.
12. Zhang G, Fang B LogitBoost classifier for discriminating thermophilic and mesophilic proteins. J biotechnol. 2007;127(3):417–424.
13. Lin H, Chen W Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods. 2011;84(1):67–70.
14. Ku T, Lu P, Chan C et al. Predicting melting temperature directly from protein sequences. Comput biol chem. 2009;33(6):445–450.
15. Wang L, Li C Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification. Biotechnol lett. 2014;36(10):1963–1969.
16. Pucci F, Dhanani M, Dehouck Y et al. Protein thermostability prediction within homologous families using temperature–dependent statistical potentials. PloS One. 2014;9(3):91659.
17. Si J, Zhao R, Wu R An Overview of the Prediction of Protein DNA–Binding Sites. Int J Mol Sci. 2015;16(3):5194–5215.
18. Consortium U The universal protein resource (UniProt). Nucleic Acids Res. 2007;35:D193–D197.
19. Berman HM, Westbrook J, Feng Z et al. The protein data bank. Nucleic Acids Res 28(1): 235–242.
20. Kumar MD, Bava KA, Gromiha MM et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res. 2006;34(1):204–206.

Citation: Mahmoudi M, Arab SS, Zahiri J, et al. An overview of the protein thermostability prediction: databases and tools. *J Nanomed Res.* 2016;3(5):1–6. DOI: 10.15406/jnmr.2016.03.00072