Journal of Theoretical Biology 390 (2016) 117–126

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Insights into the molecular basis of piezophilic adaptation: Extraction of piezophilic signatures

Abhigyan Nath*, Karthikeyan Subbiah*

Department of Computer Science, Banaras Hindu University, India

HIGHLIGHTS

- Insilco filter to extract molecular signatures of piezophilicity.
- Enhanced classification of piezophilic proteins by SVM.
- Rule extraction and biological interpretation of rules.
- Ranking of amino acids according to their discriminative ability.

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form
6 November 2015
Accepted 21 November 2015
Available online 2 December 2015

Keywords:
Piezophiles
Amino acid composition
Molecular signature
SVM
PART

ABSTRACT

Piezophiles are the organisms which can successfully survive at extreme pressure conditions. However, the molecular basis of piezophilic adaptation is still poorly understood. Analysis of the protein sequence adjustments that had taken place during evolution can help to reveal the sequence adaptation parameters responsible for protein functional and structural adaptation at such high pressure conditions. In this current work we have used SVM classifier for filtering strong instances and generated human interpretable rules from these strong instances by using the PART algorithm. These generated rules were analyzed for getting insights into the molecular signature patterns present in the piezophilic proteins. The experiments were performed on three different temperature ranges piezophilic groups, namely psychrophilic–piezophilic, mesophilic–piezophilic, and thermophilic–piezophilic for the detailed comparative study. The best classification results were obtained as we move up the temperature range from psychrophilic–piezophilic to thermophilic–piezophilic. Based on the physicochemical classification of amino acids and using feature ranking algorithms, hydrophilic and polar amino acid groups have higher discriminative ability for psychrophilic–piezophilic and mesophilic–piezophilic groups along with hydrophobic and nonpolar amino acids for the thermophilic–piezophilic groups. We also observed an overrepresentation of polar, hydrophilic and small amino acid groups in the discriminatory rules of all the three temperature range piezophiles along with aliphatic, nonpolar and hydrophobic groups in the mesophilic–piezophilic and thermophilic–piezophilic groups.

1. Introduction

Water has played a vital role in the evolution of life on our planet Earth. More than 70% of Earth is covered by oceans (Abe and Horikoshi, 2001). The marine environment provides a habitat for a wide variety of living forms, including many extremophiles like acidophiles, alkaliphiles, psychrophiles, thermophiles, piezophiles, etc. The high pressure habitats are the most prominent among others and they represent the largest ecosystem on the planet Earth. The molecular mechanism governing the life at deep sea under high pressure environments is yet to be fully understood.

Initially organisms thriving in high pressure environments were called as barophiles (ZoBell and Johnson, 1949). Later the term Piezophile was coined by Yayanos (1995). Piezophiles are pressure loving microorganisms that thrive at a pressure greater than the atmospheric pressure and they belong to both bacteria and archea. These piezophilic organisms are classified into three main categories on the basis of pressure, namely: piezotolerant, piezophilic and hyperpiezophilic. These three groups are in turn sub-classified into psychro, meso, thermo and hyperthermo-piezophilic groups on the basis of temperature (Fang et al., 2010).

* Corresponding authors. Tel.: +91 9956015187; fax: +91 9473967721.
E-mail addresses: abhigyannath01@gmail.com (A. Nath), karthinikita@gmail.com (K. Subbiah).

http://dx.doi.org/10.1016/j.jtbi.2015.11.021
0022-5193/© 2015 Elsevier Ltd. All rights reserved.

One of the marked adaptations to maintain the life processes at high pressure and cold surroundings is an increase in the unsaturated fatty acid contents in cell membranes (Simonato et al., 2006). High pressure also affects protein structure and leads to denaturation (Hayakawa et al., 1996) and this may be attributed to the weakening of the hydrophobic effect (Grigera and McCarthy, 2010; Mozhaev et al., 1996). The primary structure of proteins is generally not affected by pressure up to 20 kbar (Balny, 2000), but high pressure mainly affects the interactions maintaining the native folded state that leads to weakening of hydrophobic interactions, causes reduction in volume and favours unfolding of protein structures (Siddiqui and Thomas, 2008).

Previous research has reported the genes involved in inorganic ion transport to be positively selected in hyperthermophilic piezophilic *Pyrococcus abyssi* (Gunbin et al., 2009). Further Campanaro et al. (2008) have reported that the genes involved in solute transport, protein translocation, motility and DNA synthesis are also positively selected. Di Giulio (2013) calculated the Pressure Asymmetry Index (PAI) of amino acids, where higher PAI values indicate more barophilic amino acids (Arginine, Glycine, Serine, Aspartic acid, Valine, Glutamic acid, Lysine, Alanine, Asparagine, Leucine, Histidine, Metheonine, Isoleucine, Threonine, Phenylalanine, Proline, Cysteine, Glutamine, Tryptophan and Tyrosine, in the decreasing order of PAI values). Pradel et al. (2013) did a comparative analysis by taking mesophilic–piezophilic *Desulfovibrio piezophilus* and mesophilic–nonpiezophilic counterpart and reported the specific preferential amino acid replacements (Glutamic acid (E), Lysine (K), Asparagine (N), Alanine (A), Histidine (H), Arginine (R), Threonine (T)). Some of the recent studies have also speculated the origin of the genetic code in high hydrostatic pressure environments (Di Giulio, 2005a, b; Di Giulio, 2013); consequently this makes the study of high pressure adaptation both interesting and important.

Previously a few machine learning based classification schemes have been developed for thermophilic proteins using sequence and structural data (Li et al., 2010; Lin and Chen, 2011; Zhang and Fang, 2007). Amino acid composition plays an important role in environmental adaptation (De Vendittis et al., 2008) and some researchers had developed successful classifiers for thermophilic and psychrophilic proteins using only the amino acid composition (Gromiha and Suresh, 2008; Nath and Subbiah, 2014). The availability of structural data is still scarce for piezophilic proteins so there is a need for sequence feature based predictor. Previous studies have shown the effectiveness of machine learning methods in learning about the molecular basis of adaptations in psychrophilic proteins (Nath and Subbiah, 2014).

The amino acid features responsible for piezophilic adaptation in carrying out significant comparative analysis can be obtained from the protein sequences of two organisms having differences only in their optimal growth pressure; optimal growth temperature being in the same range. For example genomic comparisons have shown that proteins from thermophilic organisms tend to have a higher percentage of charged amino acids (Glutamate (E), Lysine (K) and Arginine (R)) than those from mesophiles (Liang et al., 2005).

In the current study, firstly machine learning algorithms are used to discriminate between piezophilic and nonpiezophilic protein sequences to select the best classifier and then the classifier is used as a filter to sieve out the correctly predicted protein sequences (strong instances that includes only true positives-correctly predicted piezophilic proteins and true negatives-correctly predicted nonpiezophilic proteins). The PART algorithm is further applied on these sets of strong instances (correctly predicted) to generate rules which are human interpretable in order to draw any conclusion on the biological basis of piezophilic adaptation.

This study focuses on protein sequence adaptation at the molecular level to high pressure environments. As the molecular basis of the piezophilic proteins is still largely unknown, this study will facilitate to fill the gap in understanding the molecular adaptation at high pressure environments. This may in turn help in protein engineering for suitable design of enzymes that can work at high pressure for application in various industrial processes

2. Materials and methods

2.1. Dataset

We created a two dimensional habitat space based data set, first on the basis of pressure (piezophiles and nonpiezophiles), then on the basis of temperature preference of the species (psychrophilic, mesophilic and thermophilic).

We took the entire proteome sequences of organisms, which have been previously used for comparative analysis: *Psychrophilic–Piezophilic(PP): Shewanella violacea* DSS12 (Masanari et al., 2014); *Psychrophilic–Nonpiezophilic (PNP): Shewanella frigidmarina* NCIMB400 (Campanaro et al., 2008*); Mesophilic–Piezophilic (MP): Desulfovibrio piezophilus* C1TLV30*; Mesophilic–Nonpiezophilic (MNP): Desulfovibrio salexigens* DSM 2638 (Pradel et al., 2013*); Thermophilic–Piezophilic (TP-I): Pyrococcus yayanosii* CH1*; Thermophilic–Nonpiezophilic (TNP-I): Pyrococcus furiosus* COM1*: Thermophilic–Piezophilic (TP-II): Thermococcus barophilus and Thermophilic–Nonpiezophilic(TNP-II):Thermococcus kodakarensis* KOD1 (Di Giulio, 2013*)*. The detailed description of the piezophilic strains is given in Supplementary material. All these sequences are pre-processed by removing those sequences having the keyword putative, predicted, hypothetical, fragment and having nonstandard residues of ‘B’,’J’,’O’,’U’,’X’ or ‘Z’. Further the CD-HIT (Li and Godzik, 2006) program is applied on the pre-processed sequences to cluster them with less than 40% sequence identity. We have finally obtained the following number of sequences in each group: 2464 (PP)/2684 (PNP), 2125 (MP)/2566 (MNP), 1058 (TP-I)/1025 (TNP-I), and 1099(TP-II)/1249(TNP-II) after preprocessing and redundancy reduction.

The imbalance data set problem occurs when there is a large difference between the total number of instances belonging to positive class and the negative class, which results in *classifier bias*. The imbalance data sets often tends to produce a majority classifier by over-predicting the presence of majority class and this issue is rarely discussed in bioinformatics literature (Nath and Subbiah, 2015; Wei and Dunbrack, 2013). This lowers the sensitivity and increases the specificity of the classifier performance even though the overall accuracy is high. To deal with imbalance data set problems, we have created balanced training set by randomly selecting equal number of positive (piezophilic proteins) instances and negative (nonpiezophilic proteins) instances and keeping the rest of the sequences from both the classes in the testing set. As a result the balanced training set consists of equal number of piezophilic and nonpiezophilic sequences. The proportions of sequences in each group are given in Table 1. The complete sequence ids used in the training and testing sets for different piezophilic and nonpiezophilic groups are provided in Supplementary material.

2.2. Input sequence features

The main aim of this study is to characterise piezophilic protein sequences on the basis of amino acid composition for understanding how the subtle amino acid compositional changes facilitate the adaptation to high pressure, so we have selected amino