Gene 421 (2008) 32–36

Contents lists available at ScienceDirect

Gene

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

How does gene expression level contribute to thermophilic adaptation of prokaryotes? An exploration based on predictors

Ji Wang [a], Bin-Guang Ma [b,c,*], Hong-Yu Zhang [b], Ling-Ling Chen [b], Shi-Cui Zhang [a,*]

[a] Department of Marine Biology, Ocean University of China, Qingdao 266003, P. R. China
[b] Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, P. R. China
[c] Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Bergen 5008, Norway

ARTICLE INFO

Article history:
Received 2 February 2008
Received in revised form 16 June 2008
Accepted 17 June 2008
Available online 22 June 2008

Received by M. Di Giulio

Keywords:
Gene expression level
Thermophilic adaptation
Prokaryotes
PHX genes
CAI

ABSTRACT

By analyzing the predicted gene expression levels of 33 prokaryotes with living temperature span from <10 °C to >100 °C, a universal positive correlation was found between the percentage of predicted highly expressed genes and the organisms' optimal growth temperature. A physical interpretation of the correlation revealed that highly expressed genes are statistically more thermostable than lowly expressed genes. These findings show the possibility of the significant contribution of gene expression level to the prokaryotic thermal adaptation and provide evidence for the translational selection pressure on the thermostability of natural proteins during evolution.

1. Introduction

Although most species grow at normal temperature ranging from 20 °C to 50 °C, some microorganisms have been found able to thrive far beyond this range. The currently known range of optimal growth temperature (OGT) for microorganisms varies from less 10 °C to more than 100 °C. Attempts to trace the imprint of ambient temperature in the genotypic traits of microorganisms have led to substantial understanding of the molecular basis of thermal adaptation. It has been well recognized that all the cell components such as lipids, nucleic acids and proteins should adapt to environmental temperature. For example, an alteration of lipid composition may be needed to maintain an acceptable permeability and fluidity of the cytoplasmic membranes at high temperatures (Stetter, 1999; Albers et al., 2000). The GC or AG content may be increased to stabilize DNA or RNA chains in response to heat stress (Galtier and Lobry, 1997; Wang and Hickey, 2002). Heat-resistance of natural proteins can reflect in both the structure and sequence levels (Kumar et al., 2000; Sterner and Liebl, 2001; Hickey and Singer, 2004). Factors reported to contribute to the enhanced thermostability of proteins involve roughly the following aspects: (i) enhancement of electrostatic or hydrophobic interactions due to a bias of amino acid composition (Vetriani et al., 1998; Kumar and Nussinov, 2001; Haney et al., 1999; Jaenicke and Böhm, 1998; Di Giulio, 2000; Cambillau and Claverie, 2000; Zeldovich et al., 2007); (ii) increment of hydrogen bonds (Borders et al., 1994; Vogt et al., 1997); (iii) variations of secondary structures (Menéndez-Arias and Argos, 1989; Querol et al., 1996); (iv) more compactness of native conformation (Hurley et al., 1992; Thompson and Eisenberg, 1999). All these findings form solid knowledge of thermophilic adaptation.

However, because thermophilic adaptation is a complex issue with contribution from many factors and some factors may attract relatively less attention and thus be underestimated. One of such factors may be the gene expression level. Although it has been reported that synonymous codon usage is subject to the natural selection of environmental temperature (Lynn et al., 2002) and noticed that the difference of synonymous codon usage increases when only ribosomal proteins (usually highly expressed) were taken into account (Lobry and Necsulea, 2006), which implicate the gene expressivity effect, a systematical investigation of gene expression level respective to thermal adaptation lacks. Some fundamental questions concerning this issue are still elusive. For example, what is the general relationship between genes' expression level

Abbreviations: PHX, predicted highly expressed; OGT, optimal growth temperature; CIT, Comprehensive Indicator of Thermostability for proteins; CAI, Codon Adaptation Index; COG, Cluster of Orthologous Groups of proteins.

* Corresponding authors. Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, P. R. China. Tel./fax: +86 533 2780271.
E-mail addresses: bgMa@sdut.edu.cn, mbgmbg@163.com (B.-G. Ma), sczhang@ouc.edu.cn (S.-C. Zhang).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.020

and their thermostability? Are highly expressed proteins more thermostable than lowly expressed ones? Or in one word, how does the gene expression level contribute to the prokaryotic thermal adaptation?

Our former work has revealed that there exists a strong positive correlation between the percentage of predicted highly expressed (PHX) genes and the OGTs of 6 deep-sea microorganisms (Xu and Ma, 2007). This enlightens us about the contribution from gene expression level to thermophilic adaptation. Because our previous work focused on a relatively small amount of deep-sea species, the omnipresence of the observed correlation is unknown. Furthermore, why this correlation exists needs further exploration. In this study, 33 species are investigated, including both marine and terrestrial organisms, and by calculating the percentage of PHX genes, it was found that there indeed exists a significantly positive correlation between the percentage of PHX genes and the organismal OGT. Moreover, by defining a comprehensive indicator of thermostability for proteins, the reasons for why this correlation exists is further examined, with the finding that PHX genes are more thermostable than non-PHX genes. These new findings are helpful to answer some fundamental questions pertaining to prokaryotic thermal adaptation.

2. Materials and methods

2.1. Data collection

Up to April, 2007, more than 400 prokaryotes have been sequenced and nearly 160 species have been determined the unique OGT. Among the 160 species, many of them have the same OGT, with 33 different OGTs in all. For each different OGT, a representative species is chosen, therefore, totally 33 prokaryotes are randomly sampled including both marine and terrestrial ones. Ten of them are archaea and the others are bacteria. Although more species could be analyzed, this sample size is big enough to explore universal trends if they exist. The genomes of all the 33 species were downloaded from NCBI (RefSeq Project (Pruitt et al., 2007)) on April 7th, 2007 and the OGT data are obtained from literature and some online resources (e.g., NCBI website, ATCC and DSMZ databases). The OGTs for the 33 species vary from 8 °C to 100 °C. Some basic information for the used organisms is listed in Supplementary Table 1.

2.2. Measures of gene expression level

Two theoretical indicators are employed to measure the level of gene expression based on the finding that highly expressed genes exhibit a stronger codon usage bias (Ikemura, 1981; Ikemura, 1985). One is $E(g)$ (Karlin et al., 1998; Karlin and Mrazek, 2000; Karlin et al., 2005) and the other is CAI (Sharp and Li, 1987). Both the two indicators have been widely used in the past few years and the predicted results are consistent with the experimental facts (Karunakaran et al., 2003; Jansen et al., 2003). According to the $E(g)$ measure (Karlin and Mrazek, 2000), genes that deviate strongly in codon usage from the average gene but are sufficiently similar in codon usage to ribosomal protein (RP) genes, to translation and transcription processing factor (TF) genes, and/or to chaperone-degradation protein (CH) genes are predicted highly expressed (PHX) (see Karlin's publication (Karlin and Mrazek, 2000; Karlin et al., 2005) for details). The percentage of PHX genes is calculated for each genome used in this study.

The "Codon Adaptation Index" (CAI) has been extensively used as a classical measure of gene expression level. A CAI value is between 0 and 1, and a higher value means a stronger codon usage bias (Sharp and Li, 1987), indicating a higher expression level. In this study, the "cai" program in the EMBOSS software package is adopted to calculate the CAI values. Compared with the $E(g)$ method, there is no definite criterion for CAI measure to distinguish PHX genes and non-PHX genes. As a result, genes which have the top 20% largest CAI values are considered as PHX genes based on the positive correlation between CAI value and gene expression level.

2.3. Comprehensive indicator of themostability for proteins

The best indicator for the thermostability of individual protein may be its melting temperature (T_m) (Kumar et al., 2000). However, few proteins have been measured of T_m compared with the need of large scale proteomic analysis. As a simpler alternative, it is well known that the themostability of proteins can be indicated by the bias of their amino acid composition. Some indicators for this purpose have been proposed at proteomic level such as *CvP*-bias (Suhre and Claverie, 2003) or the contents of seven residues IVYWREL (Zeldovich et al., 2007). Although these indicators have been demonstrated showing a significant correlation with the OGT of prokaryotic microorganisms, they are all incomprehensive (e.g. in the definition of *CvP*-bias, non-polar (hydrophobic) residues are not included, while for the indicator of the seven amino acids IVYWREL, the polar ones are neglected) and thus cannot quantitatively weigh the relative contribution of different categories of residues to the thermostability of protein.

Considering the fact that 20 kinds of amino acids can be classified into three categories: charged, polar and non-polar (hydrophobic), here we define a comprehensive indicator of thermostability for proteins (CIT) which includes all the three categories of residues and its form is as follows:

$$\mathrm{CIT} = aX_c + bX_p + cX_h + d, \quad (1)$$

where a, b, c, d are coefficients and X_c is the content sum of the 4 charged residues D, E, K, R, i.e., $X_c = X_D + X_E + X_K + X_R$, in which X_D, X_E, X_K and X_R are the frequencies of the 4 charged residues; similarly, X_p are the content sum of the 6 polar residues C, N, Q, S, T and Y and X_h is the content sum of the 9 non-polar (hydrophobic) residues A, F, G, I, L, M, P, V and W. Note that in the definition of CIT the frequency of Histidine (H) is not included, because the content sum of all the 20 kinds of residues is 1 and there are only 19 frequencies are independent. In principle anyone of the 20 kinds of amino acids can be removed from the CIT definition indistinguishably to ensure the independence of the frequency variables. Here the content of Histidine (H) is removed because it's a charged residue but with a p*K*a near 7, which makes it difficult to categorize it as a charged residue or a polar one. The coefficients in CIT can be determined by doing regression of OGT to the three content sums X_c, X_p, X_h on the dataset of the 33 proteomes used. The CIT value is calculated for each proteome in this work to correlate with OGT. Moreover, genes in each genome are parted into two groups: PHX genes and non-PHX genes, and the CIT values for both the two groups of genes are also calculated and compared with each other.

3. Results

3.1. Significant correlation between the percentage of PHX genes and OGT

The percentages of PHX genes that are predicted by $E(g)$ in each genome are presented in Supplementary Table 2. Of all the 33 species, *Thermofilum pendens* Hrk 5 has the highest percentage (20.77%) of PHX genes, and *Polaromonas naphthalenivorans* CJ2 has the lowest (2.30%). The average percentage of PHX genes is 9.98%. An evident correlation between the percentage of PHX genes and the OGT of a microorganism can be found in Fig. 1. The correlation is positive and statistically significant ($R=0.684$, $p<0.0001$), which means that more PHX genes are present in the organisms with higher OGT. This universal trend may result from at least two possible reasons: one is that some genes with specific functions are highly expressed in the organisms living at elevated temperature (functional interpretation); another