Gene 551 (2014) 92–102

Contents lists available at ScienceDirect

Gene

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

Sequence determinants of prokaryotic gene expression level under heat stress

Heng Xiong, Yi Yang, Xiao-Pan Hu, Yi-Ming He, Bin-Guang Ma *

Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

ARTICLE INFO

Article history:
Received 19 April 2014
Accepted 25 August 2014
Available online 26 August 2014

Keywords:
Gene regulation
Temperature adaptation
Sequence analysis
Transcriptome
Proteome

ABSTRACT

Prokaryotic gene expression is environment-dependent and temperature plays an important role in shaping the gene expression profile. Revealing the regulation mechanisms of gene expression pertaining to temperature has attracted tremendous efforts in recent years particularly owning to the yielding of transcriptome and proteome data by high-throughput techniques. However, most of the previous works concentrated on the characterization of the gene expression profile of individual organism and little effort has been made to disclose the commonality among organisms, especially for the gene sequence features. In this report, we collected the transcriptome and proteome data measured under heat stress condition from recently published literature and studied the sequence determinants for the expression level of heat-responsive genes on multiple layers. Our results showed that there indeed exist commonness and consistent patterns of the sequence features among organisms for the differentially expressed genes under heat stress condition. Some features are attributed to the requirement of thermostability while some are dominated by gene function. The revealed sequence determinants of bacterial gene expression level under heat stress complement the knowledge about the regulation factors of prokaryotic gene expression responding to the change of environmental conditions. Furthermore, comparisons to thermophilic adaption have been performed to reveal the similarity and dissimilarity of the sequence determinants for the response to heat stress and for the adaption to high habitat temperature, which elucidates the complex landscape of gene expression related to the same physical factor of temperature.

1. Introduction

Prokaryotes are widely distributed on the earth due to their prominent adaptation capability to environments. Microbes can be found in cold (temperature < 10 °C) and very hot (temperature > 100 °C) environments that are not suitable or even lethal for human and other eukaryotes to live. The molecular mechanisms for this superior capability have attracted intensive research efforts in recent years (Berezovsky and Shakhnovich, 2005; Hickey and Singer, 2004; Kumar and Nussinov, 2001; Ma et al., 2010; Mallik and Kundu, 2013; Mamonova et al., 2013; Sabath et al., 2013; Sterner and Liebl, 2001; Taylor and Vaisman, 2010; van Wolferen et al., 2013; Zeldovich et al., 2007). One important aspect of the prokaryotic molecular adaptation is about gene expression. Because gene expression is environment-dependent, how the gene expression level is adapted to temperature is intriguing. The prokaryotes living under moderate temperature condition are constantly used in the studies of temperature adaptation by applying an environmental stress and checking the gene expression profiles.

Gene expression is a fairly complex process that is regulated at multiple levels by different mechanisms. According to the central dogma of molecular biology (Crick, 1970; Koonin, 2012), gene expression is a flow of information from genetically encoded DNA to the messenger RNA to polypeptide and finally to the folded protein. Both the transcription and translation processes are involved in the regulation of gene expression. Affected by the external environments, the expression patterns of mRNA and proteins are determined by the intrinsic genetic information encoded in the DNA sequences of a bacterium genome. The sequencing technology has helped in accumulation of thousands prokaryotic genomes and the gradual maturity and declined cost of high-throughput approaches for the detection of gene transcription and translation levels have resulted in numerous transcriptomes and proteomes of bacteria under heat stress conditions (Barreiro et al., 2009; Chhabra et al., 2006; Fleury et al., 2009; Gunasekera et al., 2008; Han et al., 2007; Jain et al., 2011; Koide et al., 2006; Lenco et al., 2009; Luders et al., 2009; Ribeiro et al., 2011; Ternan et al., 2012; van der Veen et al., 2007). Although the host organisms of these transcriptomes and proteomes are diverse in life styles, the common building blocks for the macromolecules (namely nucleotides and amino acids) and the same physical factor of temperature may confer commonality of the gene regulation mechanisms for adaptation to heat stress. The previous studies of the adaptation and regulation mechanisms of gene expression were concentrated on the

Abbreviations: DEGs, differentially expressed genes; CAI, Codon Adaptation Index; OGT, Optimal Growth Temperature; COG, Clusters Of Orthologous Groups of proteins.

* Corresponding author.
E-mail address: mbg@mail.hzau.edu.cn (B.-G. Ma).

http://dx.doi.org/10.1016/j.gene.2014.08.049
0378-1119/

characterization of the expression profiles of individual organisms and little effort has been devoted to the investigation of the commonness of the differentially expressed genes and their regulation mechanisms from the viewpoint of intrinsic sequence features with respect to the consequences of high temperature.

In this work, we try to answer the following questions pertaining to heat stress condition. Do the differentially expressed genes (DEGs) have peculiarity in their nucleotide and protein sequence features? Do up-regulated and down-regulated genes under heat stress have different sequence features? Are these features common or conserved among different organisms? Meanwhile, we want to know the similarity and difference of the sequence features of these DEGs in response to heat stress and those of the genes in thermophilic organisms who live under high temperature environment perennially. We gathered the transcription and translation profiles under heat stress conditions reported in recent years and examined the sequence features along the gene expression information flow to reveal how the genotypic sequence determinants are interpreted into phenotypic responses to the elevated living temperatures.

2. Materials and Methods

2.1. Transcriptome and Proteome Data Under Heat Stress

The transcriptome data for 7 bacteria were collected from the literature and the basic information for this data set was listed in Table 1. The proteome data for 6 bacteria were collected from the literature and the basic information for this data set was presented in Table 2. The DNA and protein sequences were taken from NCBI (RefSeq).

2.2. Distribution of Differentially Expressed Genes in Operons

The locations of the differentially expressed genes in the genome were determined according to the RefSeq annotation (Pruitt et al., 2012). The operon information was taken from the DOOR database (Mao et al., 2009). The distance between the starting position of a gene and the middle point of its corresponding operon was used to indicate the location: $d = S_i - M$, where S_i is the coordinate of the start of the considered gene and M is the coordinate of the middle point of the operon where the gene i is located. If d is negative, then the gene i is located at the first-half in the operon; otherwise, the second-half.

2.3. Sequence Feature Calculation and Prediction

The mRNA secondary structures were predicted by the RNAfold program in the Vienna RNA secondary structure package (version 1.8.5) (Arsene et al., 2000) with the folding temperature as the temperature of heat stress reported in the original literature and other parameters as the default.

Codon frequencies in the mRNA sequence were counted and compared with the genomic frequencies taken from genomic tRNA database (Lowe and Eddy, 1997) to identify the deviation from the global genomic frequencies and the difference between up-regulated and down-regulated genes. Codon usage and Codon Adaptation Index (CAI) values were calculated by using the 'cusp' and 'cai' program in the EMBOSS software package with ribosomal proteins as the reference of highly expressed proteins following the standard CAI procedure (Sharp and Li, 1987).

The protein secondary structures were predicted by locally running the PSIpred program (version 3.3) (McGuffin et al., 2000) on a Linux platform with default parameters. In addition to the differentially expressed proteins, the same amount of proteins whose expression levels do not vary in the heat stress condition were randomly sampled for each proteome and used as contrast. The protein secondary structure contents were calculated as the fractions of C (coil), H (helix) and E (strand).

The protein folding rates were predicted by the seqRate server (Lin et al., 2010) with "Fold Type" set as 'Unknown'. The protein aggregation propensity was predicted by the Zyggregator online server (Tartaglia and Vendruscolo, 2010). In addition to the differentially expressed proteins, the same amount of proteins whose expression levels do not vary under the heat stress condition were randomly sampled from each proteome and used as contrast.

2.4. Comparison and Statistics

For the examined features in the 7 transcriptomes and 6 proteomes, comparisons were performed between "up-regulated" and "down-regulated" (up-down), "up-regulated" and "overall" (up-all), "down-regulated" and "overall" (down-all) by using Wilcoxon rank sum tests, respectively, and the corresponding p-values were listed in Tables S1 and S2. Similar statistical comparisons were also made for the content sum of hydrophobic amino acids in the 6 thermophilic bacteria (Fig. S4) and the corresponding p-values were listed in Table S4. The significance of the difference was considered at three levels of $p < 0.1$, $p < 0.05$ and $p < 0.001$ that were marked on the figures with "*", "**" and "***", respectively.

3. Results and Discussion

By mining the literature, we have collected the expression data of 7 transcriptomes (Table 1) and 6 proteomes (Table 2) for bacteria under heat stress condition. The sequence features of the DEGs were investigated at multiple levels (Fig. 1) including gene length and distribution in operons, the nucleotide composition in mRNA sequence, the features of mRNA secondary structure, the codon usage in the coding sequences, the amino acid composition in the protein sequences, the contents of protein secondary structures and the predicted folding rates and aggregation propensity of the protein sequences. We examined the differences between the up-regulated and down-regulated genes as well as the differences between the DEGs and the overall genes of the whole genome by using Wilcoxon rank sum tests (p-values listed in Tables S1, S2). Comparison to thermophilic adaption was performed to reveal the similarity and dissimilarity of the sequence determinants for the response to heat stress and for the adaption to high habitat temperature.

Table 1
Basic information for the 7 transcriptomes collected from literature.

Species	Abbr.	Stress	Control	DEG no.[a]	Reference
Clostridium difficile 630	CD	Heat (41 °C)	37 °C	337	Ternan et al. (2012)
Escherichia coli K-12	EC	Heat (43 °C)	30 °C	221	Gunasekera et al. (2008)
Corynebacterium glutamicum ATCC 13032	CG	Heat (40 °C)	30 °C	778	Barreiro et al. (2009)
Staphylococcus aureus ISP794 (NCTC8325)	SA	Heat (43 °C)	37 °C	89	Fleury et al. (2009)
Listeria monocytogenes EGD-e	LM	Heat (48 °C)	37 °C	147	van der Veen et al. (2007)
Yersinia pestis strain 201	YP	Heat (45 °C)	37 °C	573	Han et al. (2007)
Xylella fastidiosa 9a5c	XF	Heat (40 °C)	29 °C	438	Koide et al. (2006)

[a] The number of differentially expressed genes reported in the original literature.