Gene 497 (2012) 344–352

Contents lists available at SciVerse ScienceDirect

Gene

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

Short Communication

Gene expression profile of the cynobacterium synechocystis genome

Shibsankar Das [a], Uttam Roymondal [b], Brajadulal Chottopadhyay [c], Satyabrata Sahoo [d,*]

[a] Department of Mathematics, Uluberia College, Uluberia, Howrah, W.B., India
[b] Department of Mathematics, Raidighi College, Raidighi, South 24 parganas, W.B., India
[c] Department of Physics, Jadavpur University, 188 Raja S.C. Mullik Road, Kolkata-32, W.B., India
[d] Department of Physics, Raidighi College, Raidighi, South 24 parganas, W.B., India

ARTICLE INFO

Article history:
Accepted 19 January 2012
Available online 31 January 2012

Keywords:
Codon usage
Impact codons
Codon adaptation index
Gene expression
Predicted highly expressed genes
Synechocystis

ABSTRACT

The expression of functional proteins plays a crucial role in modern biotechnology. The free-living cynobacterium *Synechocystis* PCC 6803 is an interesting model organism to study oxygenic photosynthesis as well as other metabolic processes. Here we analyze a gene expression profiling methodology, RCBS (the scores of relative codon usage bias) to elucidate expression patterns of genes in the *Synechocystis* genome. To assess the predictive performance of the methodology, we propose a simple algorithm to calculate the threshold score to identify the highly expressed genes in a genome. Analysis of differential expression of the genes of this genome reveals that most of the genes in photosynthesis and respiration belong to the highly expressed category. The other genes with the higher predicted expression level include ribosomal proteins, translation processing factors and many hypothetical proteins. Only 9.5% genes are identified as highly expressed genes and we observe that highly expressed genes in *Synechocystis* genome often have strong compositional bias in terms of codon usage. An important application concerns the automatic detection of a set of impact codons and genes that are highly expressed tend to use this narrow set of preferred codons and display high codon bias .We further observe a strong correlation between RCBS and protein length indicating natural selection in favor of shorter genes to be expressed at higher level. The better correlations of RCBS with 2D electrophoresis and microarray data for heat shock proteins compared to the expression measure based on codon usage difference, $E(g)$ and codon adaptive index, CAI indicate that the genomic expression profile available in our method can be applied in a meaningful way to study the mRNA expression patterns, which are by themselves necessary for the quantitative description of the biological states.

1. Introduction

Cyanobacteria are prokaryotic organisms that serve as model microorganisms for the study of photosynthesis, carbon and nitrogen assimilation, evolution of plant plastids, and adaptability to environmental stresses. The regulation of gene expression plays an important role to study all such biological functions (Pakrasi, 1995). The description of the biological state by the quantitative measurement of the system constituents is an essential but largely unexplored area of biology. Although the recent technical advances including the development of cDNA microarray and DNA chip technology and serial analysis of gene expression (SAGE) are becoming increasingly recognized for the whole genome expression studies, but predicting expression level of genes through computational methods is appealing because it circumvents expensive and difficult experiment. With the development of a number of varieties of software tools like codon adaptation index (CAI) (Sharp and Li, 1986) and codon usage model (Karlin and Mrazek, 2000, 2004; Karlin et al., 2001, 2003, 2005a, 2005b, 2006; Mrazek et al., 2001), it is now feasible to obtain a predictive expression profile for genomes under study. But universal standards to make these studies more suitable for comparative analysis and for inter-operability with other information sources have yet to emerge. In most of these previous studies, expression level of a gene is determined by relating the codon usage difference to a prior definition of a reference set, consisting of highly expressed genes. The major problem of these methods is to find out the proper reference set and in some instances the use of alternative reference set results very poor. Although the composition of the reference set is based on the functional assignment of the genes, but there is no specific algorithm to construct a reference set of individual species. The outcome becomes highly dependent on the genome examined. With many different mRNA expression data sets available, it is worthwhile to obtain a single unified reference set, with the intention of reducing the errors and noise contained in the individual data set and to obtain a unified normal expression set of individual species. For a better

Abbreviations: CAI, Codon adaptation index; ORF, Open reading frame; PHE, Predicted highly expressed; PS, Photosystem; RCA, Relative codon adaptation; RCBS, The score of relative codon usage bias; RP, Ribosomal protein; SAGE, Serial analysis of gene expression; TF, Translation/transcription factor; WT, Wild type; 2D, Two dimensional.

* Corresponding author.
E-mail addresses: ssdas80@gmail.com (S. Das), urmandal@gmail.com (U. Roymondal), bdc_physics@yahoo.com (B. Chottopadhyay), dr_s_sahoo@yahoo.com (S. Sahoo).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.01.023

understanding of the potential expression level of the genes in the complete genome of cynobacteria we developed a statistical approach that relates relative codon usage among the genes to potential expression of the individual genes.

With the recent advent of expression measure from the score of relative codon bias (RCBS), we here report the performance of this measure to predict the highly expressed genes in *Synechocystis* genome. The objective of this work is to identify and analyze the major predicted highly expressed genes of *Synechocystis* with respect to relative codon usage. Our analysis (Das et al., 2009; Roymondal et al., 2009) of *Escherichia coli* and *Yeast* genome support the hypothesis that gene expression level relates to relative codon usage difference, indicating that codon usage contributes importantly to setting the level of expression of the gene. Although codon usage bias has long been known as a factor that affects average expression level of proteins in fast-growing microorganisms, but its role in dynamic regulation of expression is not fully understood. In one view high CAI induces strong expression of proteins, whereas other argue that strong expression is induced by the selection of optimal codons. A systematic study of synonymous nucleotide variation on gene expression by Kudla et al. (2009) has revealed that mRNA folding and associate rates of translation initiation might play a predominant role in shaping expression levels of individual genes, whereas codon bias influences global translation efficiency but not the expression of proteins. This finding seems to contradict the well known correspondence between the codon bias and expression level of individual gene. In a different study (Sharp et al., 2005), it is also shown that highly expressed genes in some species have no discernible difference in codon usage from other genes. So, in the absence of selected codon usage bias, the expression level of a gene is unlikely to be predictable from comparisons of codon usage. Finally, there are limitations to the use of codon usage bias in estimating gene expression level (Henry and Sharp, 2007). This apparent contradiction may be resolved with an alternative model of codon preference (Anderson and Kurland, 1990). The use of preferred codons, especially in genes expressing at high levels encoding mRNAs that must be translated more often, allows more efficient use of ribosomes which are quickly released to be available to translate other mRNA. But in a recent work by Sharp et al. (2010) it is shown that codon usage in highly expressed genes is strongly selected, whereas selection is very weak in genes expressed at lower levels. There has been much debate about exactly why translationally optimal codons are selected. However, the hypothesis that codon usage modulates the gene expression has been accepted in general. Many researches in this field have formulated their own measures (Anderson and Kurland, 1990; Freire-Picos et al., 1994; Gouy and Gautier, 1982; Henry and Sharp, 2007; Holm, 1986; Kudla et al., 2009; Morton, 1994; Sharp et al., 2005, 2010; Shields and Sharp, 1987; Supek and Vlahovicek, 2005; Urrutia and Hurst, 2001; Wan et al., 2004; Wright, 1990a) for gene expressivity analysis, but they exhibit strong artifacts of their formulation with varying sequence length, or overall codon bias, or codon bias discrepancy. Our aim is to develop a measure that will be free from any such possible artifacts and we attempt here to verify the usefulness of such measure by employing it to predict gene expressivity in *Synechocystis*.

2. Materials and methods

The complete genome sequence for *Synechocystis* is obtained from Genebank. All ORF listed as coding for proteins are considered in this study. We have used the score of the relative codon usage bias (RCBS) to calculate gene expressivity (Das et al., 2009; Roymondal et al., 2009).

2.1. Expression measure

The expression measure of a gene, RCBS is given by

$$\mathrm{RCBS} = \left(\prod_{i=1}^{L}\left(1+d_{xyz}^{i}\right)\right)^{1/L} - 1$$

where $d_{xyz}^{i} = \{f(x,y,z) - f_1(x)\,f_2(y)\,f_3(z)\}/f_1(x)\,f_2(y)\,f_3(z)$ is the relative codon usage bias of ith codon of a gene, $f(x, y, z)$ the normalized codon frequency for the codon xyz , $f_1(x)$ the normalized frequency of x at the 1st codon position, $f_2(y)$ the normalized frequency of y at the 2nd codon position, and $f_3(z)$ the normalized frequency of z at the 3rd codon position of the gene. L is the number of codons in the gene.

With a view of evolving codon assignments as well as codon usage patterns as the adaptive response of genomes, we had taken the criteria RCBS > T, a threshold score, as a benchmark for identifying the highly expressed genes. As our aim is to develop a measure which will be free from all artifacts, discussed in previous section, here we propose an algorithm for finding the threshold score of a genome. The algorithm is iterative. First, we calculate the CAI of the all genes in a genome by taking all Ribosomal protein (>80aa) genes as a reference set. We take this as a standard of our method and call this as an evaluation set. Then we set an arbitrary level of threshold T in the RCBS data set to differentiate highly expressed genes from lowly expressed genes. The genes whose RCBS values are greater than T are considered as predicted highly expressed (PHE) genes. Taking these PHE genes as a reference set, we calculate CAI for all genes of the genome and call this as a test set. Then we evaluate the correlation co-efficient between the test set and the evaluation set as a merit of the method. We iterate the procedure by changing the threshold T and repeat the subsequent steps until we reach at an optimal merit of the method.

2.2. Impact codons

The distribution of the nucleotides over the three codon position is different both within and between species. We quantify the degree of codon bias by assigning an impact score $(1 + d_{xyz}^{i})$ to each codon of a gene sequence, since it considers codon usage as well as the base compositional bias. If $\bar{X}$ and μ denote the sample mean and population mean of the impact score for a particular codon respectively; and σ the population standard deviation, then z score of a test statistics is given by $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$, where N is the total no of codons. The impact codons are then identified, based on the level of significance from the z score of test statistic. The scores of the impact codons differ markedly from the results expected in the absence of codon bias and it seems reasonable to assume that relative codon usage bias in the highly expressed genes is strongly influenced by the presence of impact codons.

3. Results

Having defined our formalism, we apply it to diverse genes from *Synechocystis* sp. PCC 6803. Our dataset includes 3172 complete protein coding sequence . Expression profiles of the genome are determined by calculating the score of relative codon usage difference (RCBS) for each gene. The distribution of the score is unimodal, with the majority of genes have RCBS lying between 0.1 and 0.4, and the mean and median RCBS are 0.36262 and 0.32052 respectively. The optimal threshold for the *Synechocystis* genome comes out to be 0.56 .If we take this as the benchmark for identifying highly expressed genes, only about 9.5% genes in *Synechocystis* can be classified as predicted highly expressive(PHE) genes. In Table 1 we have compiled a list of PHE genes commonly used for essential metabolic functions. We have adopted functional classes of genes from Kazusa Research Institute Web site (http://www.kazusa.or.jp/cyano). Functional analysis shows that most ribosomal proteins, transcription-translation factors, biosynthesis of cofactors, prosthetic groups and carriers and photosynthesis and respiration proteins attain high expression levels.

3.1. Ribosomal protein(RP) genes

RP genes are very important in cell biology as these provide a range of activities required for all steps of protein biosynthesis.