Codon usage pattern and predicted gene expression in *Arabidopsis thaliana*

Satyabrata Sahoo\(^a\)*, Shib Sankar Das\(^b\), Ria Rakshit\(^c\)

\(^a\) Department of Physics, Dhruba Chandra Halder College, Dakshin Barasat, South 24 Parganas, W.B., India
\(^b\) Department of Mathematics, Uluberia College, Uluberia, Howrah, W.B., India
\(^c\) Department of Botany, Baruipur College, South 24 Parganas, W.B., India

**ABSTRACT**

The extensive research for predicting highly expressed genes in plant genome sequences has been going on for decades. The codon usage pattern of genes in *Arabidopsis thaliana* genome is a classical topic for plant biologists for its significance in the understanding of molecular plant biology. Here we have used a gene expression profiling methodology based on the score of modified relative codon bias (MRCBS) to elucidate expression pattern of genes in *Arabidopsis thaliana*. MRCBS relies exclusively on sequence features for identifying the highly expressed genes. In this study, a critical analysis of predicted highly expressed (PHE) genes in *Arabidopsis thaliana* has been performed using MRCBS as a numerical estimator of gene expression level. Consistent with previous other results, our study indicates that codon composition plays an important role in the regulation of gene expression. We found a systematic strong correlation between MRCBS and CAI (codon adaptation index) or other expression-measures. Additionally, MRCBS correlates well with experimental gene expression data. Our study highlights the relationship between gene expression and compositional signature in relation to codon usage bias and sets the ground for the further investigation of the evolution of the protein-coding genes in the plant genome.

1. Introduction

*Arabidopsis thaliana* has proven to be a model experimental organism for essentially developing plant biology at the molecular level. Undoubtedly, any useful insight in understanding the expression of functional proteins of *Arabidopsis thaliana* will contribute to the development of plant research as well as in the field of modern biotechnology. It is well known that the synthesis of every protein molecule is directed by the arrangement of genetic codes in a genomic DNA sequence. The genetic code uses sixty-one codons to encode 20 amino acids and three codons to terminate translation in the process of protein synthesis. The degeneracy of the genetic code suggests that there must be many alternative nucleotide sequences to encode the same protein. The codon usage pattern varies significantly between different organisms, and also between genes which are expressed at different levels in the same organism. A number of hypotheses prevail regarding the factors which influence the codon usage pattern. Attempts have been made to explain the codon distributions in the protein-coding genes as well as the changes in codon usages among different synonymous codons in each organism (Sharp et al., 1988; Brandis and Hughes, 2016; Sharp and Li, 1987; Ikemura, 1981; Hockenberry et al., 2014; Lee et al., 2010). It is well discussed in the literature that organisms might be subjected to codon biases of different origins. In fact, it is rather difficult to decide the most common dominant codon bias of a genome. Some researchers have speculated that codon bias that tends to reduce the diversity of isoacceptor tRNAs may reduce the metabolic load (Gustafsson and Govindarajan, 2004; Akashi, 1994; Ikemura, 1985). Many other analyses have also revealed that there are many other factors like nucleotide compositional constraint, codon anticodon interaction, amino acid conservation etc. which may also influence the codon usage pattern of a genome. Whatever may be the molecular basis for codon bias, it is evident that codon bias can have a significant impact on the expression of functional proteins. Translational selection pressure or protein secondary structure may have profound effect on codon bias. It is generally thought that a balance between mutation and natural selection on translational efficiency is expected to yield a

---

**Abbreviations:** MRCBS, Score of Modified relative codon bias; PHE, Predicted Highly Expressed; CAI, Codon adaptation index; RCB, Relative codon bias; SAGE, Serial Analysis of Gene Expression; RCA, Relative Codon Adaptation; RCRS, Relative Codon Bias Strength; MBP, Megabase pair; TAIR, The Arabidopsis Information Resources; MIT, Mitochondrion; CP, Chloroplast Ptd CP; RP, Ribosomal protein; MADS, Minichromosome maintenance 1, Agamous, Deficiens and Serum response factor; GEO, Gene Expression Omnibus; RMA, Relative Molecular Abundance

*Corresponding author.
E-mail address: dr_s_sahoo@yahoo.com (S. Sahoo).

https://doi.org/10.1016/j.gene.2019.100012
Received 30 October 2018; Received in revised form 30 January 2019; Accepted 21 February 2019
Available online 06 March 2019
2590-1583/ © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
correlation between codon bias and rate of gene expression, such that highly expressed genes often have stronger relative codon bias (RCB) than genes expressed at lower levels (Kurland, 1991; Hirooka et al., 2009). Our objective of this work is to identify and analyze PHE genes and codon usage pattern in Arabidopsis thaliana. Our analyses on E. coli, yeast, synechocystis and archaean genomes support the hypotheses that each genome has evolved a codon usage pattern promoting its gene expression level (Roymondal et al., 2009; Das et al., 2009; Das et al., 2012; Sahoo and Das, 2014a; Das et al., 2017).

With the advent of modern technologies, several high-throughput experiments are widely used to identify the highly expressed genes. The most commonly used technique to study large scale gene expression is cDNA microarray. Besides, other novel techniques like 2D gel electrophoresis, Mass spectrometry, Chromatin immunoprecipitation, DNA chip technology and Serial Analysis of Gene Expression (SAGE) have been developed for the purpose. All these experiments require wide range of conditions to match, massive investment of time and resources. To overcome these major obstacles for identifying highly expressed genes in the vast majority of organisms, we must look beyond the direct experimental methods. Following this, we focused our study on developing a computational methodological that can be used to study the large-scale gene expression profile of an organism. Based on the hypothesis that highly expressed genes are often characterized by strong compositional bias in terms of codon usage (Ikemura, 1981; Ikemura, 1985; Kurland, 1991; Sahoo and Das, 2014b; Karlin and Mrazek, 2006; Karlin et al., 2005; Carbone et al., 2003; Supek Fand Vlahovicek, 2005; Supek Fand Vlahovicek, 2010), a number of varieties of software tools like Codon Adaptation Index (CAI) (Sharp and Li, 1987), Relative Codon Adaptation (RCA) (Fox and Erill, 2010), Relative Codon Bias Strength (RCBS) (Roymondal et al., 2009; Das et al., 2009) etc. have been developed to provide numerical indices to predict the expression level of genes. There are no universal standards to make these results more suitable for comparative analysis. However, most of these commonly used computational approaches depend on the knowledge of codon bias of a reference set of highly expressed genes. But, MRCBS has been devised as an alternative model to predict gene expression level from their codon compositions in such a way that the score of the expression indicator may be calculated without any knowledge of previously set selective highly expressed genes as a reference set. In fact, MRCBS performs better to capture the highly expressed genes compared to the performances of several other commonly used measures (Das et al., 2012; Sahoo and Das, 2014a; Das et al., 2017; Sahoo and Das, 2014b).

Here, we investigated the gene expression profile and the variation in synonymous codon usage pattern of Arabidopsis thaliana genome. It is a small flowering plant with a relatively short life cycle and is the first plant to have its genome completely sequenced (The Arabidopsis Genome Initiative, 2000). Since 1943, Arabidopsis thaliana started to be widely used as experimental biological material in plant research laboratories around the world. The small size of its genome with approximately 135 MBPs and 5 chromosomes makes it a useful model for plant sciences. An extensive study has been done by plant biologist to assign functions of its 2500 genes and 3500 proteins they encode. The latest information on Arabidopsis research is available from Arabidopsis Information Resources (TAIR). The small genome size and the availability of the complete DNA sequence of Arabidopsis thaliana have attracted the attention of a wide range of scientists, including evolutionary biologists and biotechnology companies. The rapid life cycle, unusual properties of inheritance and the vast information about their genealogy suggest that this organism may be used as a useful tool for the plant biologist. Finally, its important role in the study of plant-pathogen interaction makes them very attractive to biotechnology companies for industrial and research uses. Thus, the gene expression profile of Arabidopsis thaliana is expected to make important contributions in plant sciences.

2. Materials and methods

The whole genome sequence of Arabidopsis thaliana along with the gene annotations was taken from NCBI GenBank have been considered in our study. All gene sequences under study along with those annotated as hypothetical have been extracted from the Gene Bank Accession Nos: NC_003070.9(Chromosome 1), NC_003071.7(Chromosome 2), NC_003074.8(Chromosome 3), NC_003075.7(Chromosome 4), NC_003076.8(Chromosome 5), NC_001284.2(Mitochondrion MT), NC_000932.1(Chloroplast Pltd).

In the present communication, we have reported the codon usage pattern and gene expression in Arabidopsis thaliana genome. For this purpose, a variety of computational tools like CAI, Relative codon adaptation (RCA), GC3 and MRCBS have been used in this study.

1. The codon adaptation index, CAI is given by (Sharp and Li, 1987)

\[
CAI = \left( \prod_{i=1}^{w} \frac{f_i}{f_{i,\text{max}}} \right)^{1/N}
\]

where, \( N \) is the number of codons in the gene and relative adaptiveness, \( w_i \) is defined as

\[
w_i = \frac{f_{i}}{f_{i,\text{max}}}
\]

\( f_i \) is the frequency of the \( i^{th} \) codon, and \( f_{i,\text{max}} \) is the maximum frequency of the codon most often used for encoding amino acid \( aa \) in a set of highly expressed genes of the particular genome. The score measured by CAI ranges from 0 to 1 indicating that the higher are the CAI values, the genes are more likely to be highly expressed.

2. The relative codon adaptation (RCA) for an entire genome is computed as (Fox and Erill, 2010)

\[
RCA = \left( \prod_{i=1}^{L} \frac{RCA_{\text{AA}}(i)}{RCA_{\text{AA},i}} \right)^{1/L}
\]

where \( L \) is the length of a gene and \( RCA_{\text{AA},i}(i) \) is defined by.

\[
RCA_{\text{AA},i}(i) = \frac{f_{i}(x,y,z)}{f(x,y,z)}
\]

\( f_{xy} \) is the observed relative frequency of a codon \( xyz \) in any reference gene set, \( f_{i}(m) \) is the observed relative frequency of base \( m \) at position \( i \) in the same reference set.

3. GC3 measures the frequency of G or C at the third position of synonymous codons and can be used as an index of codon bias. It is measured by

\[
GC_3 = \frac{\sum_{i} f_{i}}{\sum_{i} f_{i}}
\]

where \( N = \text{any base, } S = G \text{ or } C, \text{and } f_{xy} \text{ is the observed frequency of codon } xyz \).

4. The score of modified relative codon bias, MRCBS measures the expression level of a gene and is defined as (Das et al., 2012; Sahoo and Das, 2014a; Das et al., 2017; Sahoo and Das, 2014b),

\[
MRCBS = \left( \prod_{i=1}^{N} \frac{\text{MRCBS}_{\text{xyz},i}}{\text{MRCBS}_{\text{xyz},i}} \right)^{1/N}
\]

where

\[
\text{MRCBS}_{\text{xyz},i} = \frac{RCBS_{\text{xyz}}}{RCBS_{\text{xyz},i}} \cdot RCBS_{\text{xyz}} = \frac{f_{i}(x,y,z)}{f(x,y,z)}
\]
where \( f_{\text{xyz}} \) is the normalized codon frequency of a codon \( \text{xyz} \) and \( f_{\text{xyz}}(m) \) is the normalized frequency of base \( m \) at codon position \( n \) in a gene. \( \text{RCBS}_{\text{max}} \) is the maximum value of \( \text{RCBS} \) of codon encoding the same amino acid \( \text{aa} \) in the same reference set, and \( N \) is the codon length of the query sequence. The score of the modified relative codon bias ranges from 0 and 1. The numerical value computed by this method may be used to rank the set of genes with respect to codon bias towards gene expression. It is suggested that the threshold score of the modified relative codon bias identifies the highly expressed genes. But due to evolving codon assignments as well as codon usage patterns as the adaptive response of genomes, threshold score for identifying highly expressed genes varies from genome to genome and the methodology used to calculate threshold score was described in (Sahoo and Das, 2014a).

In this work, the different expression level predictors have been computed by comparing its codon usage bias with the profile of universally functional genes. The predicted highly expressed genes (PHE) are then characterized on the basis of the strength of the codon usage bias derived from the algorithms as described in the literature and a gene is identified as PHE gene provided its MRCBS exceeds the threshold value. Pearson \( r \) correlation coefficients between different codon usage bias indices have been computed for a systematic analysis of the gene expression profile of the genome under study.

The impact score of a codon \( (\text{xyz}) \) in a gene sequence is then defined by MRCBS(\( \text{xyz} \)) and is used to describe the codon usage profile of the genome under study. If and \( \mu \) denote the sample mean and population mean of the impact score for a particular codon respectively; and \( \sigma \) the population standard deviation, then \( z \) score of a test statistics is given by

\[
    z = \frac{\bar{\text{xyz}} - \mu}{\sigma/\sqrt{N}}
\]

where \( N \) is the total no of codons. The impact codons are then identified by the impact score of a codon based on the level of significance from the \( z \) score of the test statistic.

3. Results and discussion

In the present study, we have analyzed gene expression profile of Arabidopsis genome and predicted highly expressed (PHE) genes with respect to MRCS. We have measured the expression pattern and codon usage bias of all protein-coding genes in the genome under study. Our study includes 12,645 protein-coding sequences of chromosome 1, 7596 protein-coding sequences of chromosome 2, 9474 protein-coding sequences of chromosome 3, 7426 protein-coding sequences of chromosome 4, 10,993 protein-coding sequences of chromosome 5, 117 protein-coding sequences of mitochondrion MT and 85 protein-coding sequences of chloroplast Ptd CP. Some basic information of Arabidopsis genome is given in Table 1. The expression level of all protein-coding genes was calculated by MRCS and compared with other codons usage models like CAI and RCA. Threshold score for identifying highly expressed genes in Arabidopsis thaliana has been calculated to be 0.77. GC content of the genome under study is 44.26%. The overall GC3 score is 0.4215. Many researchers have argued that GC content or GC3 may be viewed as the primary influence on the codon usage pattern and thus on the expression profile. Table 2 displays the statistics of PHE genes and the top 20 PHE genes of Arabidopsis thaliana genome along with their functions and scores calculated in our approach (MRCBS).

Codon usage profile of Arabidopsis genome has been described in terms of average impact score of 27,046 complete protein-coding sequences of the genome [Fig. 1]. Although most of the amino acids can be specified by more than one codon, only a subset of potential codons is used [Table 3] in highly expressed genes. There are no impact codons coding His, Thr and Val in the presently studied Arabidopsis genomes. The impact codons in Arabidopsis are found to be mostly used in coding Phe (ttt,ttc), Leu (ttg,ctt,ctc), Ile (atc), Met (atg), Tyr (tac), Gln (caag,cag), Asn (aac), Lys (aaq,aag), Asp (gat), Glu (gaa,gag), Ser (tct,tca,agc), Pro (ctt,cca), Ala (gct), Cys (tgc), Trp (ttg), Arg (aga), Gly (ggt,gga).

Importantly, these codons do not reflect any simple compositional bias. Not all of the preferred (impact) codons are GC rich and GC/GC3 may not be the accurate representation of the trend in codon usage. It may be thought that the selection of the preferred codons causing the optimization of the translational rate possibly depends on the codon–anticodon interaction.

The large data set analyzed here revealed a strong bias towards usage of a different set of preferred codons in genes with high cytoplasmic mrna levels. In contrast, genes with low mrna levels showed very little synonymous codon usage bias. Usage bias was proposed as a result from translational selection, since using a codon that is translated via an abundant tRNA species were hypothesized to boost translational efficiency. Codon frequencies are found to vary between genes in the same genome. The standard version of the genetic code includes 61 sense codons and three stop codons. Although almost all organisms have made the same codon assignments for each amino acid, the preferred use of individual codons varies greatly among genes. The overall nucleotide composition of the genome which influences the codon usage pattern introduces selective forces acting on highly expressed genes to improve the efficiency of translation. It is now widely accepted that synonymous codon preferences in a unicellular organism are affected by the cellular amount of isoacceptor tRNA species. But we observe that not all tRNA genes corresponding to impact codons have been detected by tRNAscanSE. However many tRNAs can translate more than one codon, but with variable ability and it is suggested that impact codons have favored translational efficiency. Since the highly expressed genes use a preferred set of optimal codons in accordance with their respective tRNA levels, this observation might find another important application in tRNA finding algorithm.

Expression profiles of the genes are determined by calculating MRCS for each gene and their distributions are shown in Fig. 2. The majority of genes (90%) have MRCS values lying between 0.65 and 0.75, and the mean and median values are 0.3870 and 0.3295, respectively. Only 3.3% genes have MRCS values > 0.77. It was observed that percentage of PHE genes vary between 3% to 4% in Arabidopsis thaliana chromosomes, whereas no highly expressed genes are predicted in CP/MT genomes. The overall variation of GC or GC3 content of the genes is depicted in Suppl. Figs. 1 and 2 respectively. It indicates that majority of genes have GC3 score lying between 0.3 and 0.6 and (88.5%) of genes have GC content lying between 0.4 and 0.5. We observed that the percentage of PHE genes varies from chromosome to chromosome and is independent of GC content or GC3 score of these genes. In fact, we have failed to find any correlation between gene expression and GC content or GC3 score. It is well studied that highly expressed genes display more biased codon usage than the lowly expressed genes [Table 3]. We observed that PHE genes of Arabidopsis thaliana mostly include ribosomal protein (RP) genes, translation initiation factors, translation elongation factors, MADS box transcription factor, membrane traffic protein, trans-membrane protein,
chaperon, heat shock protein, histone, ubiquitin, nucleic acid binding protein and many stress and energy metabolism genes. However, all RP genes of Arabidopsis thaliana do not comprise the PHE gene class. Table 2 reports the statistics of PHE gens. The percentage of PHE genes in Arabidopsis thaliana is 3.3%, whereas only 17.7% genes fall in the class of RP genes. It is remarkable that 99.21% RP genes in Yeast genome and almost all RP genes in E. coli genome fall in PHE class of genes. An average of 65.56% RP genes in the archaeal genome is PHE. Out of 561 RP genes 255 RP genes are PHE. Thus a very poor fraction of RP genes of Arabidopsis thaliana has highly predicted expression level in contrast to E.coli, Yeast and Archaea. The top 20 genes with the highest predicted expression levels for Arabidopsis thaliana genomes are displayed in Table 2. Our analysis predicted 1063 highly expressed genes in Arabidopsis thaliana. A list of well-characterized PHE genes has been displayed in Suppl. Table 1. It is worth noticing that these genes are separated into different functional categories. Table 4 displays a set of well-characterized PHE genes segregated into different functional categories.

It has been observed that PHE genes belonged to various functional classes and variably represented in the genome. These include carbohydrate kinase, dehydratase, dehydrogenase, ATP synthase, acyltransferase, methyltransferase, Amino acid transporter, actin/actin-related protein, calcium-binding protein, calmodulin, cysteine protease, chromatin/chromatin-binding protein, DNA directed DNA/RNA polymerase, enzyme modulator, extracellular matrix structural protein, ligase, non motor actin/microtubule-binding protein, non receptor

### Table 2
Characteristics of PHE genes and top 20 genes with the highest predicted expression levels for Arabidopsis thaliana genome.

| Average length | Average GC content | Average GC3 content | % of PHE RP genes | % of PHE hypothetical genes | Top 20 genes | Locus tag/gene name | Function | MRCBS |
|---------------|-------------------|---------------------|------------------|-----------------------------|--------------|--------------------|-----------|-------|
| 658           | 0.461             | 0.475               | 17.70%           | 8.63%                       | AT5G03710    | Replication factor C large subunit | 0.942377  |       |
|               |                   |                     |                  |                             | AT3G56020    | Ribosomal protein L41 family | 0.902928  |       |
|               |                   |                     |                  |                             | AT5G03850    | Nucleic acid-binding, OB-fold-like protein | 0.885142  |       |
|               |                   |                     |                  |                             | RPS28        | Ribosomal protein S28 | 0.884064  |       |
|               |                   |                     |                  |                             | AT3G46430    | ATP synthase | 0.877127  |       |
|               |                   |                     |                  |                             | AT3G08520    | Ribosomal protein L41 family | 0.872734  |       |
|               |                   |                     |                  |                             | AT2G04621    | Trans membrane protein | 0.869109  |       |
|               |                   |                     |                  |                             | AT5G56670    | Ribosomal protein S30 family protein | 0.868022  |       |
|               |                   |                     |                  |                             | AT3G10090    | Nucleic acid-binding, OB-fold-like protein | 0.866286  |       |
|               |                   |                     |                  |                             | RPL23AA      | Ribosomal protein L23AA | 0.86058  |       |
|               |                   |                     |                  |                             | AT2G19730    | Ribosomal L28e protein family | 0.860542  |       |
|               |                   |                     |                  |                             | RS27A        | Ribosomal protein S27 | 0.860165  |       |
|               |                   |                     |                  |                             | AT4G27090    | Ribosomal protein L14 | 0.856987  |       |
|               |                   |                     |                  |                             | AT2G14285    | Small nuclear ribonucleoprotein family protein | 0.856773  |       |
|               |                   |                     |                  |                             | AT3G11120    | Ribosomal protein L41 family | 0.855905  |       |
|               |                   |                     |                  |                             | AT5G16130    | Ribosomal protein S7e family protein | 0.854895  |       |
|               |                   |                     |                  |                             | AT2G31490    | Neuronal acetylcholine receptor subunit alpha-5 | 0.854269  |       |
|               |                   |                     |                  |                             | CAM3         | Calmodulin 3 | 0.852098  |       |
|               |                   |                     |                  |                             | RPS15        | Cytosolic ribosomal protein S15 | 0.848976  |       |
|               |                   |                     |                  |                             | CAM2         | Calmodulin 2 | 0.847033  |       |

**Fig. 1.** Average impact score of codons in Arabidopsis thaliana genome.
serine/ threonine protein kinase, oxidase, oxidoreductase, nucleotidyltransferase, reductase, peroxidase, phosphatase, peroxidase/phosphatase inhibitor, transfer/ carrier protein.

Besides, we have identified a number of PHE genes which play important roles in signal transduction mechanism, amino acid transport and metabolism, secondary metabolites biosynthesis and catabolism, cell membrane biogenesis, inorganic ion transport and metabolism, coenzyme transport and metabolism, carbohydrate transport and metabolism, intercellular trafficking, and energy production and conversion. These include vascularude, vascular ATP synthase, vascular calcium-binding protein, vascular ATPase, vesicle coat protein, seed storage albumin, arabinogalactan protein, cytochrome complex, cytochrome c oxidase/electron carrier and members of the cytochrome family, DELF family, dehydrin family. In addition, a number of PHE genes encoding plasma membrane intrinsic protein, plant defense, photosystem II, phytochrome associated protein, phytosulfokine, plant viral response protein have significant roles in plant. Among other PHE genes, copper chaperone, copper iron-binding protein, a copper transport protein, zinc-binding ribosomal family protein and ferredoxin like superfamily protein have important functions in this organism.

However, a fraction of poorly characterized hypothetical genes was also found among the PHE genes. Table 2 displays the general statistics of hypothetical or poorly characterized PHE genes in Arabidopsis genome. Genes of unknown function with high predicted expression levels may be attractive candidates for experimental characterizations. The characteristic codon distribution of these genes indicates that they may have important functions in these genes. A variety of PHE genes encoding proteins of unknown function may provide targets for identification of additional key features of Arabidopsis thaliana. The temporal and spatial organization of these genes for chromosome replication, genome segregation and cell division processes are less characterized in Arabidopsis genome. A detailed analysis of these putative/hypothetical PHE genes would generate a more comprehensive picture of the replication and division machineries, and of the regulatory features of the cell cycle.

3.1. Correlations among different codon bias indices

In this study, we compared the performances of several commonly used computation tools for predicting gene expression level. The expression profiles of the Arabidopsis genome were analyzed in terms of CAI, RCA and MRCBS. The CAI scores have been calculated by taking all RP (> 80aa) genes as PHE genes which are commonly referred as reference set. RCA frequencies are computed using the identical reference set as used in the calculation of CAI. The results indicate that there is a good correlation between RCA and CAI(r = 0.673761) while the correlation of RCA with MRCBS is significantly higher (r = 0.787772) [Fig. 3]. The novel method of quantitatively predicting gene expressivity MRCBS is then compared with CAI and correlation between them is found to be surprisingly good (r = 0.900204) [Fig. 4].
Table 4  
A list of potential PHE genes segregated into different functional categories.

| Functional Category                        | Gene Symbol(s)                                                                 |
|-------------------------------------------|-------------------------------------------------------------------------------|
| Transcription factor                      | AT4G10480, AT3G12390, AT5G09920, AT4G35900, AT2G17770, AT1G54830, AT5G53980, AT1G56170 |
| Elongation                                | AT1G56070, AT4G20360, AT3G12915, AT1G07930, AT1G30230, AT2G18110, AT5G19510, AT1G51210 |
| Translation initiation/elongation factor  | AT1G04410, AT5G43330                                                         |
| MADS box transcription factor             | AT1G69120, AT1G31140, AT1G50780, AT3G17692, AT1G07930, AT5G57860, AT1G56170, AT1G51210 |
| Chromatin/chromatin binding protein       | AT3G05590, AT1G01160, AT1G75060, AT5G12910, AT4G20980, AT4G26630, AT5G05470, AT5G69410 |
| Histone                                   | AT4G04040, AT5G09870, AT5G12910, AT5G10390, AT2G14285, AT3G11500, AT1G79550, AT3G12390 |
| mRNA processing/splicing                  | AT3G62840, AT3G44500, AT5G47210, AT5G08420, AT5G57820, AT1G50390, AT3G12390, AT1G79550 |
| Tubulin                                   | TU2, TU3, TU4, TU5, TUB2, TUB3, TUB4, TUB5, TUB7, TUB9, XS, TUB6, CRT1a, CRT1b |
| Calcium binding protein                   | AT3G13400, AT5G13930, AT1G66200, AT5G35630, AT2G04190, AT3G17820, CAM1, CAM2 |
| G protein coupled receptor/modulator      | AT5G42090, AT5G18520, AT2G30060, AT3G07880, AT2G04621, AT2G01870, AT1G76640, CAM1 |
| Transmembrane Protein                     | AT5G24090, AT2G22080, AT3G47820, AT2G02180, AT3G18000, AT2G25297, AT4G11840, CAM1 |
| Acyltransferase                           | AT5G11670, AT5G19875, AT3G03120, AT2G22140, AT3G39800, AT5G53980, AT2G18110, CAM1 |
| Basic helix-loop-helix transcription factor| AT2G447710, AT5G09225, AT3G09440, AT3G09440, AT2G17770, AT5G53980, AT2G18110, CAM1 |
| Basic leucine zipper transcription factor  | AT5G45960, AT5G59480, AT1G50390, AT5G35980, AT5G59480, AT3G13930, AT3G13930, CAM1 |
| Extracellular matrix structural protein    | AT3G41840, AT5G58390, AT1G55570, AT5G47210, AT5G53980, AT5G35630, AT5G35630, CAM1 |
| Homeodomain transcription factor          | AT5G53980, AT3G18800, AT5G09225, AT2G22080, AT2G25297, AT3G47820, AT3G47820, CAM1 |
| Cysteine protease                         | AT3G94840, AT3G46790, AT3G46440, AT3G94840, AT3G47820, AT3G47820, AT3G47820, CAM1 |
| Dehydratase                               | AT3G51160, AT3G46440, AT3G06630, AT3G47820, AT3G47820, AT3G47820, AT3G47820, CAM1 |
| Aminoacyl-tRNA synthetase                 | AT1G55800, AT1G55800, AT5G09225, AT3G47820, AT3G47820, AT3G47820, AT3G47820, CAM1 |
| Antibacterial response protein            | AT5G50840, AT5G50840, AT5G50840, AT3G47820, AT3G47820, AT3G47820, AT3G47820, CAM1 |
| ABC transporter                           | AT5G60790, AT5G60790, AT5G60790, AT3G47820, AT3G47820, AT3G47820, AT3G47820, CAM1 |
| Ubiquitin/ubiquitin like                  | UBIQ11, UBIQ11, UBIQ11, UBIQ11, UBIQ11, UBIQ11, UBIQ11, UBIQ11, CAM1 |
| Chaperone/heat shock protein              | AT1G72130, AT1G72130, AT1G72130, AT1G72130, AT1G72130, AT1G72130, AT1G72130, CAM1 |
| ABC transporter                           | AT5G19860, AT5G19860, AT5G19860, AT5G19860, AT5G19860, AT5G19860, AT5G19860, CAM1 |
These correlation coefficients can be used to express the strength of the existing prediction methods. It can be seen that MRCBS consistently yields better correlation than other. We also observe that there is no clear correlation between CAI or MRCBS with GC3 (rCAI = -0.05726, rMRCBS = 0.101083) or GC (rCAI = -0.15775, rMRCBS = 0.041383). So, GC content and GC3 may not be the accurate representation of the trend in codon usage bias. Similarly, no correlation between the length of the gene and MRCBS or CAI has observed in our study.

3.2. Correlation of protein and mRNA expression levels with MRCBS

In this study we choose to compare our results with the experimental datasets. The value of codon-based expression indicator can perhaps be appreciated by comparing them with the experimental gene expression data in general. Of course, the codon-based expression indicator yields static value, whereas gene expression is a dynamic process with very different expression levels under different conditions. The expression data that we have used in this study stems from Gene Expression Omnibus (GEO) datasets. In GEO dataset (GEO accession: GSM2473182) protein expression levels were quantified by RMA (Relative Molecular Abundance) signal intensity. For the entire group of selected genes (20,900 genes) for which the complete data set can be generated along with the codon based expression indicator, the Pearson correlation coefficient between CAI and MRCBS comes out to be 0.901964. The pair-wise correlation coefficient between protein expression level and MRCBS, CAI, RCA and GC turns out to be 0.268321, 0.253094, 0.283545 and 0.206581 respectively. Correlation is worse with GC3 (0.049775). It has been observed that for genes with high RMA signal intensity (> 7.59), the pair-wise correlation coefficients are better (0.386227, 0.337139, 0.303723, 0.251336 and 0.290886) [Suppl. Figs. 3–7].

In another analysis we have compared our results with the radioactive data (González-Pérez et al., 2011). We have collected 1797 Arabidopsis genes for which there are orthologous in yeast and humans and that have mRNA half-life data (Calderwood et al., 2016). For these dataset, the predicted gene expression level using MRCBS value is found to correlate well with RMA signal intensity (r = 0.50923) [Fig. 5]. The correlation is better than the quantitative measure of CAI (r = 0.470608), RCA (r = 0.442278), GC3 (r = 0.405765) and GC (r = 0.362806) [Suppl. Figs. 8–11]. It suggests that a quantitative estimate of the expression level by MRCBS values performs better than other indices of expression-measure. The novel method of quantitatively predicting gene expressivity is then compared with mRNA half-life data. We observe that the correlation coefficient of mRNA half-life data with MRCBS (r = 0.3504) is good [Fig. 6], but worse compared to RMA signal intensity. Although the pair-wise correlation coefficient among the gene expression levels from two experimental datasets (r = 0.525273) is good, it can be clearly seen that the agreement of predicted and actual protein expression level quantified by mRNA half-life data varied greatly between all examined combinations of prediction method and data set (rCAI = 0.31067, rGC3 = 0.310397, rGC = 0.281694 and rRCA = 0.279249) [Suppl. Figs. 12–15].

To assess the value of MRCBS for predicting protein expression levels in Arabidopsis thaliana, we plotted the two experimental sets of data versus MRCBS along with RCA and CAI. The distribution patterns for both the protein expression data with respect to these expression indicators are highly similar. Comparing the performance of the MRCBS, the CAI and RCA as numerical indices of the gene expression level in terms of the Pearson correlation coefficient with the expression data, we observed that MRCBS generally performs better than CAI and RCA.

4. Conclusion

Our study demonstrates that MRCBS may be a useful tool for...
predicting highly expressed genes. The idea of supporting our method is based on the hypothesis that codon usage pattern is largely responsible for regulation of gene expression which can occur during transcription or at the level of protein translation. Although the concept of predicting gene expression level from the codon usage pattern was proposed a decade ago, only recently these methods have been successfully applied to identification of highly expressed genes in various bacteria and eukaryotic genomes. The improved reliability of MRCBS for estimating expression levels in Arabidopsis genome thus makes this index a superior choice for undertaking and benchmarking predictions of gene expression. In this study, various approaches to estimating gene expression level based on codon usage have been applied to Arabidopsis genome with the objectives of testing the present alternative method of studying whole-genome gene expression. Our results demonstrate significant heterogeneity in codon usage among genes in Arabidopsis genome. Furthermore, the predicted gene expression level using the quantitative measure CAI was found to correlate well with MRCBS. In addition, since the expression levels measured by current DNA microarray and proteomics technologies represent the accumulated results of expression and degradation, the results from this computational approach could be used as reference data for calibrating and better interpreting experimental data. For example, observation of low level of expression from proteomic or microarray data for a gene with a high PHE index might suggest the possible involvement of degradation in regulating expression levels of that gene. Although most of the PHE genes are essential genes responsible for the habitat, energy sources and life style of an organism, the study also identified a number of functionally unknown genes as PHE genes based on their codon usage profile. Further investigation of these genes by an integrated computational and experimental approach will enhance our knowledge of metabolism. Given that a large volume of experimental data is available on this plant, such novel method may be helpful on extracting meaningful information for understanding the details of functional genomics.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2019.100012.

**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Declaration of interests**

We, the authors declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**

Akashi, H., 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136, 927–935.

Brandis, G., Hughes, D., 2016. The selective advantage of synonymous codon usage bias in Salmonella. PLoS Genet. 12 (3), 1–16.

Calderwood, A., Kopriwa, S., Morris, R.J., 2016. Transcript abundance explains mRNA mobility data in Arabidopsis thaliana. Plant Cell 28, 610–615.

Carbajo, A., Zinoviev, A., Fekos, F., 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19, 2005–2015.

Das, S., Roymondal, U., Sahoo, S., 2009. Analyzing gene expression from relative codon usage bias in Yeast genome: a statistical significance and biological relevance. Gene 443, 121–131.

Das, S., Roymondal, U., Chottopadhyay, B., Sahoo, S., 2012. Gene expression profile of the cyanobacterium Synechocystis. Gene 497, 344–352.

Das, S., Chottopadhyay, B., Sahoo, S., 2017. Comparative analysis of predicted gene expression among crenarchaeal genome. Genome Inform. 15 (1), 38–47.

Fox, J.M., Erill, I., 2010. Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res. 17, 185–196.

González-Pérez, S., Gutiérrez, J., García-García, F., Osuna, D., Dopeno, J., Lorenzo, Ó., Revuelta, J.L., Aréllano, J.B., 2011. Early transcriptional defense responses in Arabidopsis cell suspension culture under high-light conditions. Plant Physiol. 156 (3), 1439–1456.

Gustafsson, C., Govindarajan, Minshull J., 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22 (7), 346–353.

Hiraoaka, Y., Kawamata, K., Haraguchi, T., Chikashige, Y., 2009. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes Cells 14, 499–509.

Hockenberg, A.J., Sirer, M.L., Amaral, L.A., Jewett, M.C., 2014. Quantifying position-dependent codon usage bias. Mol. Biol. Evol. 31, 1880–1893.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389–409.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.

Karlin, S., Mrázek, J., 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J. Bacteriol. 182, 5238–5250.

Karlin, S., Mrázek, J., Brocchieri, M.L., 2005. Predicted highly expressed genes in archaean genomes. Proc. Natl. Acad. Sci. U. S. A. 102, 7303–7308.

Kurland, C.G., 1991. Codon bias and gene expression. FEBS Lett. 285, 165–169.

Lee, S., Weon, S., Lee, S., Kang, C., 2010. Relative codon adaptation index, a sensitive measure of gene usage bias. Evol. Bioinformatics Online 6, 47–55.

Roymondal, U., Das, S., Sahoo, S., 2009. Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome. DNA Res. 16, 13–30.

Sahoo, S., Das, S., 2014a. Analyzing gene expression and codon usage bias in Metallosphaera Sedula. J. Bioinf. Comput. Biol. Control 3, 72–80.

Sahoo, S., Das, S., 2014b. Analyzing gene expression and codon usage bias in diverse genomes using a variety of models. Curr. Bioinform. 9, 102–112.

Sharp, P.M., Li, W.H., 1987. The codon adaptation index - a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sharp, P.M., Cove, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F., 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. 16 (17), 8207–8211.

Supch Fand Vlahovicek, K., 2010. Correction: comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinf. 11, 463.

The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.