DNA Research **16**, 13–30, (2009) doi:10.1093/dnares/dsn029

Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to *Escherichia coli* Genome

Uttam Roymondal[1], Shibsankar Das[2], and Satyabrata Sahoo[3,*]

Department of Mathematics, Raidighi College, South 24 Parganas, Raidighi, West Bengal, India[1]; Department of Mathematics, Uluberia College, Uluberia, Howrah, West Bengal, India[2] and Department of Physics, Raidighi College, South 24 Parganas, Raidighi, West Bengal, India[3]

(Received 21 March 2008; accepted 16 October 2008; published online 8 January 2009)

Abstract

We present an expression measure of a gene, devised to predict the level of gene expression from relative codon bias (RCB). There are a number of measures currently in use that quantify codon usage in genes. Based on the hypothesis that gene expressivity and codon composition is strongly correlated, RCB has been defined to provide an intuitively meaningful measure of an extent of the codon preference in a gene. We outline a simple approach to assess the strength of RCB (RCBS) in genes as a guide to their likely expression levels and illustrate this with an analysis of *Escherichia coli* (*E. coli*) genome. Our efforts to quantitatively predict gene expression levels in *E. coli* met with a high level of success. Surprisingly, we observe a strong correlation between RCBS and protein length indicating natural selection in favour of the shorter genes to be expressed at higher level. The agreement of our result with high protein abundances, microarray data and radioactive data demonstrates that the genomic expression profile available in our method can be applied in a meaningful way to the study of cell physiology and also for more detailed studies of particular genes of interest.

Keywords: codon usage; gene expression; predicted highly expressed genes; *Escherichia coli*

1. Introduction

Regulation of gene expression plays a central role in defining cell fate and controlling organ formation. Genomic function can be understood at the nucleotide level, but, the complexity and diversity of genomic function, leading to an emergent picture of the genome as an interacting system with many degrees of freedom, bring experimental and theoretical challenges to the quantitative measurement of the biological state, many of which are of statistical nature. Genes encode proteins, and proteins perform functions in the cell. Hence a gene takes part in biological function only if it is expressed, i.e. the protein produced from it is present in the cell. Gene regulation takes place during transcription, the process by which the cell reads the information contained in a gene and copies it to the messenger RNA which is subsequently used to make a functional protein. This is a most fundamental level of biological process which involves the interaction of DNA and proteins. Its regulation takes place through the binding of proteins to DNA at specific loci in the vicinity of the gene to be regulated. The transcription of one gene may be enhanced or reduced by the expression of the gene itself. The process is complex and not yet understood completely. Genes with high expression levels include those required for an organism's viability and the ability to identify these genes is crucial for drug development. Certainly the high cost and technical

Edited by Hiroyuki Toh

* To whom correspondence should be addressed. Tel. +91 3324-180575. Fax. +91 3174-270761. E-mail: dr_s_sahoo@yahoo.com

expertise required is an obstacle to many investigators who are interested in pursuing such studies. Although a variety of software tools and technologies have been developed for gene expression studies, a universal standard making these studies more suitable for comparative analysis and for inter-operability with other information sources is yet to emerge. Large-scale, high-throughput experimental methods require material and information processing systems to match. The analysis of high-throughput gene expression data is in an early stage of development. Development of advance technology for whole genome expression studies is thus becoming increasingly recognized. Predicting expression level of genes through computational methods is appealing because it circumvents expensive and difficult experiment.

In recent years there has been increasing reports[1–23,43,44] on predicted highly expressed genes in several micro-organisms which provide a wealth of information about gene expression. It is suggested that the essential genes primarily include the ensembles of highly expressed genes that encode proteins [transcription/translational factors (TF), ribosomal proteins (RP), proteases and chaperons (CH), degradation, cellular localization, biosynthesis, metabolism, photosynthesis, respiration and glycolysis, etc] vital for cell physiology. Perhaps, the essential functions of these gene products correspond to the biased amino acid composition that might minimize the substantial biosynthesis energy costs indicating the high biological significance of these genes. Besides other mechanisms, it is also suggested that codon bias can influence gene expression by optimization of the translational rate and thus, highly expressed genes can be characterized on the basis of biased codon usages compared with average genes. In several previous studies,[3,7–13,17] a number of different patterns of codon usage have been hypothesized and many indices have been proposed to measure the degree of codon bias. Among these, the codon adaptation index (CAI) has been widely applied to the prediction of highly expressed genes in various organisms.[3,15,16,24–27] CAI was proposed as a measure of codon usage in a gene relative to that in a reference set of genes.[3] The previous studies suggest that CAI index correlates better with expression level of a gene than other codon usage indices, such as the effective number of codons,[7] codon bias index,[8] the frequency of optimal codons,[9] intrinsic codon bias index,[10] maximum likelihood codon bias,[11] synonymous codon bias orderliness,[12] and measure independent of length and composition (MILC),[13] etc. The parameters underlying the CAI model rely on the codon composition of only a limited set of highly expressed genes and are based on a fairly simple assumption that the functional class of genes are highly expressed. To define the parameters in the CAI model, Sharp and Li[3] considered the codon frequency of only 24 highly expressed genes of which 50% were genes of RPs and the rest mostly metabolic enzymes. A related method, the codon usage model, is based on similar principles, but the parameters are based on a somewhat broader set of highly expressed genes. In application of this model, Karlin and coworkers[17–23] have shown that it is a reasonable assumption that for RP genes, CH and TF are highly expressed. Gene expressivity is strongly correlated with protein abundances. A number of studies have also revealed that codon compositions in highly expressed genes are influenced by tRNA abundances.[1–6] Generally, highly expressed genes, producing abundant proteins, use a subset of optimal codons which are recognized by the most abundant tRNA species. It is well established that highly expressed genes have strongly biased usage of alternative synonymous codons and that of preferred codons, which are thought to be translated most efficiently by the most abundant tRNAs, and the lowly expressed genes have less biased codon usage patterns.[1,2] The observations strongly suggest that natural selection has shaped the codon usage pattern accommodating optimal gene expression levels for most situations of its habitat, energy sources, and life cycle. Codon usages vary considerably within and between organisms. The effect of natural selection on codon usage quantifies the level of gene expression. However, the resulting bias in the codon usage has two main components. One is the correlation with tRNA availability and the other is non-random choices between pyrimidines for third base. A critical analysis of codon usage in a gene shows that mutational bias also plays a role in codon selection. Several studies have analysed the relationship between the GC-content of isochors and the expression patterns of the genes they contain.[28] The G + C composition resulting from mutational bias has been hypothesized to determine the major trends in codon usage of high or low G + C organisms. Within a genome, codon bias tends to be much stronger in highly expressed gene than in genes expressed at lower levels, suggesting that there might be some selective advantage to concentrate essential genes on GC rich domains of the genome. Surprisingly, to address this important issue, some studies have also given conflicting results.[29–33] Several papers reported very weak correlations, either negative or positive between the GC-content and gene expression. The discrepancy among the studies might be due to the methods used to measure the expression parameter of the data sets analysed or the differences in the way correlations were computed.

In fact, the characterization of regulatory elements underlying gene expression is largely an unsolved problem. The hypothesis that codon usage modulates gene expression has been accepted in general. Many researches in this field have formulated their own measures, which has led to a large number of available methods[3,7–12,17] for gene expressivity analysis. Unfortunately, these methods are not universally applicable as they exhibit strong artefacts 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 artefacts and we attempt here to verify the usefulness of such a measure by employing it to predict gene expressivity in *Escherichia coli* (*E. coli*).

2. Materials and methods

The genome sequence for *E. coli* K-12 MG1655 is obtained from Genebank accession no. NC_000913. All ORF (open reading frames) listed as coding for proteins (confirmed and hypothetical) are considered in this study. Our approach in estimating gene expression level is related to codon usage difference of a gene with respect to biased nucleotide composition at the three codon sites. Let $f(x,y,z)$ be the normalized codon frequency for the codon triplet (x,y,z) of a gene. Then the relative codon bias (RCB) of a codon triplet (x,y,z) in a gene is defined as

$$d_{xyz} = \frac{f(x,y,z) - f_1(x)f_2(y)f_3(z)}{f_1(x)f_2(y)f_3(z)}, \qquad (1)$$

where $f_1(x)$ is the normalized frequency of x at the first codon position, $f_2(y)$ is the normalized frequency of y at the second codon position, and $f_3(z)$ is the normalized frequency of z at the third codon position of the gene. The frequencies f_1, f_2, f_3 have been derived from the set of codon samples of a gene and the normalization of frequency is done over the gene length in codons, in an attempt to compensate for the expected increase of RCB with the total number of codons. We quantify the degree of codon bias of a gene in such a way that comparisons can be made both within and between genomes. As defined earlier, d_{xyz} contains somewhat more quantitative information than others, since it considers codon usage as well as the base compositional bias. Then the expression measure of a gene is

$$\mathrm{RCBS} = \left(\prod_{i=1}^{L}(1 + d^i_{xyz})\right)^{1/L} - 1, \qquad (2)$$

where d^i_{xyz} is the codon usage difference of i^{th} codon of a gene. L is the number of codons in the gene.

RCB is the difference of observed frequency of a codon from the expected frequency under the hypothesis of random codon usage where the base composition were biased at three sites as that in the sequence under study, divided by the expected frequency. RCBS is the overall score of a gene indicating the influence of RCB of each codon in a gene. Our analysis is based on the hypothesis that RCB reflects the level of gene expression. The expression measure of a gene in this approach is denoted by RCBS. RCBS value close to 0 indicates a lack of bias for the codons and is thus useful for comparing different sets of genes.

3. Results

Our data set includes 4174 complete protein coding sequences from *E. coli*. Expression profiles of the genes are determined by calculating the score of RCB (RCBS value) for each gene and their distributions are shown in Fig. 1. The majority of genes (63%) have RCBS values lying between 0.2 and 0.4, and the mean and median values are 0.3870 and 0.3295, respectively. Only ~18% genes have RCBS values >0.5. The analysis of RCBS values among different gene class shows that the gene classes (RP, CH, TF), which serve the representatives of highly expressed genes have RCBS > 0.5 in most of the cases. This suggests that significantly stronger codon bias is a result for

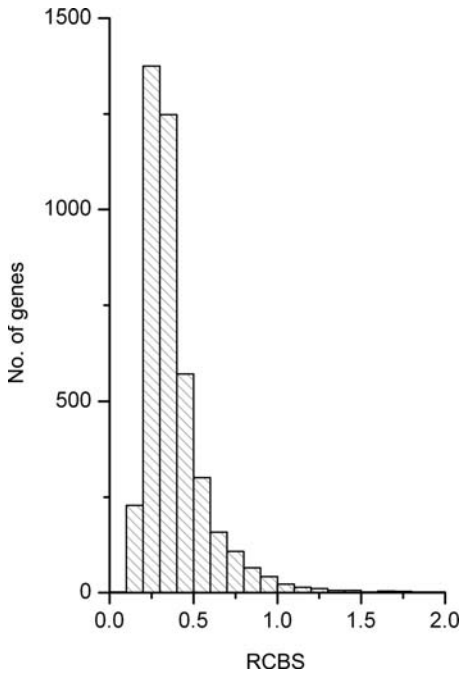

Figure 1. Distribution of RCBS for all coding genes in the genome of *E. coli*.

translational efficiency as well. This finding is consistent with others,[3,17,18] as most of the previous expression measures have considered those as representative standards for highly expressed genes in their calculation. There is also experimental evidence in support of RP, CH and TF as standard derivatives for the highly expressed genes as it is observed that many RPs augmented by abundant TF and CH proteins are needed to assure properly translated, modified and folded protein products which expedite and regulate cellular activities in most prokaryotic genomes. Our data support the proposition that each genome has evolved a codon usage pattern accommodating gene expression level, and RCBS value >0.5 exhibits favourable codon usage. So, we chose this index as an effective expression measure on the basis that it has been shown to correlate highly to expression levels and the predicted expression level based on RCBS (RCBS > 0.5) values suggests that almost 18% of genes in the *E. coli* genome qualify as highly expressed genes. In our study, the genes are segregated into different functional categories such as metabolism, information transfer, regulation, transport, cell process, cell structure, location of gene products, extra-chromosomal, DNA sites and cryptic genes in accordance with Munich Information Center for Protein Sequence (MIPS) classification. Functional analysis shows that highly expressed genes involved in the location of gene products are the largest functional class followed by genes involved in information transfer, metabolism, cell structure, cell process, extra-chromosomal, regulation and transport function, respectively. A total of 750 genes are identified as highly expressed genes in *E. coli* with 163 genes involved in energy metabolism, 75 genes involved in translation, 34 genes in transcription, and 29 in CH and folding (Supplementary Table SI). In addition, the functional class of amino acid biosynthesis, nucleotide biosynthesis, fatty acid biosynthesis and other cofactor and small molecule, etc includes 67 highly expressed genes. Besides, there are several ($\sim$185) genes encoding predicted proteins and 15 other genes of unknown function, which are thought to be highly expressed genes in our approach. We observe that 24 genes encoding predicted proteins and 12 genes encoding proteins of unknown function are highly expressed genes with RCBS > 1.0. The highly expressed genes of *E. coli* with RCBS > 1.0 are reported in Supplementary Table SII (hypothetical protein or predicted protein genes are not listed). Of these, 11 encode proteins that function in energy metabolism, 18 are RP genes, 11 encode TF and the remaining encode proteins that function in different cell process.

In order to compare our results, we have also calculated CAI values for the same genes. Fig. 2 shows the relationship between RCBS and CAI values. Here, the CAI scores have been calculated according to the original publication of Sharp and Li,[3] which stem from 24 highly expressed genes. It can be clearly seen that for genes with high CAI values (>0.5), there is strong correlation between them ($r = 0.4614$). But for proteins with CAI values significantly <0.3, correlation is worse ($r = -0.0572$). The novel method of quantitatively predicting gene expressivity is then compared with the other widely accepted measure of Karlin and Marzek.[17] In Fig. 3, we plot RCBS values against $E(g)$ of Karlin et al.[18] The correlation is surprisingly good with $r = 0.6706$, $P < 0.001$. We analyse further the relationship between the length of the coding regions and the expression level of genes. In Fig. 4 we plot RCBS as a function of the gene length. We observe that shorter genes assume the higher value of RCBS while longer genes tend to have lower RCBS.

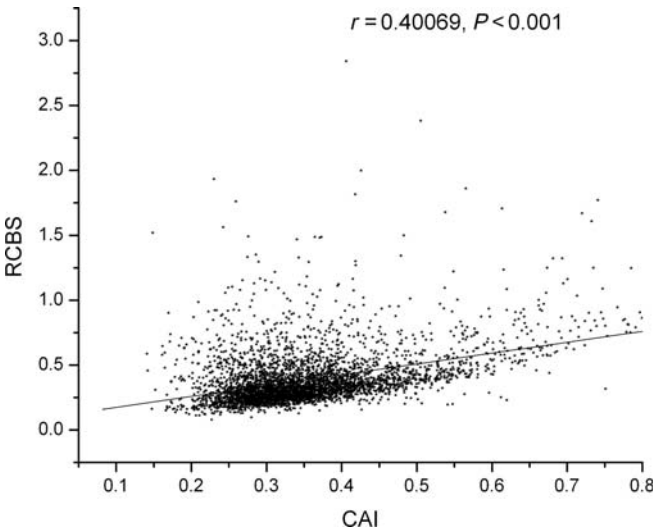

Figure 2. RCBS plotted against CAI for *E. coli* genes.

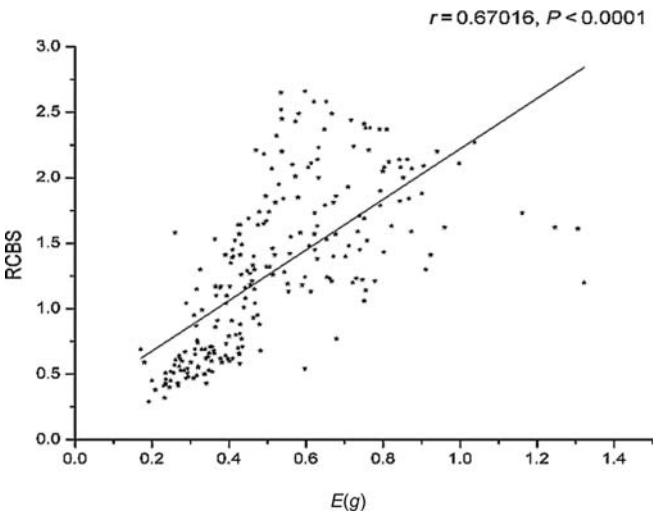

Figure 3. RCBS plotted against $E(g)$[18] for *E. coli* genes.

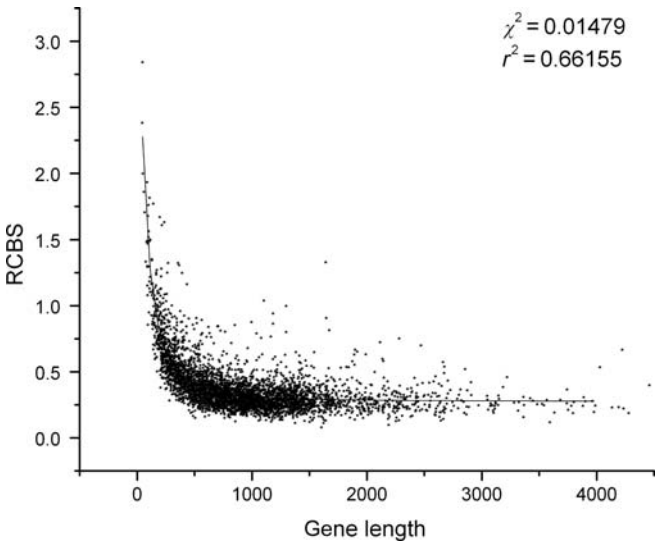

Figure 4. RCBS plotted against the length of 4174 genes from the *E. coli* genome.

There is a strong correlation between RCBS and gene length ($r^2 = 0.65878$ and $\chi^2 = 0.0149$). This effect is not due to systematic bias of gene size. To investigate the effect of protein length on gene expression as measured by RCBS, the data is split into three groups: short ($L < 150$), intermediate ($150 < L < 300$) and long ($L > 300$). Several observations can be made. Genes are sorted according to their expression level. It should be noted that genes of the same expression level may have wide variation in length and also that genes of the same length may have a wide range of RCBS. We observe that the estimate of expression level, as derived from RCBS, ranges from a low value to high value for each of the three length groups. It is evident from our data that RCBS ranges from 0.245 to 3.416 for $L < 150$, whereas it ranges from 0.123 to 0.907 for $150 < L < 300$ and from 0.079 to 1.328 for $L > 300$. It is noted that the selective pressure on codon usage appears to be lower in genes encoding long rather than short proteins. Our studies, although less extensive, suggest that selection on codon usage as well as sequence composition is primarily responsible for RCBS. For a simple explanation, we select a set of *E. coli* sequences of equal length and randomize the above sequences 500 times, keeping their (i) codon usage; and (ii) sequence composition conserved. RCBS calculated for those sequences are found to vary in a wide range. We repeat the experiment on different sets of genes with varying length. The results are summarized in Supplementary Tables SIIA and SIIB. Supplementary Table SIIA describes the results of 14 arbitrary nucleotide sequences of different length, each randomized 500 times. In Supplementary Table SIIB, we present the results of the same experiment on a few selected genes of different length. We observe that the smaller sequences have a greater probability of resulting in high value of RCBS (>0.5), but there is nothing to prevent longer sequences from having high RCBS. Although the values for shorter sequences are more variable due to sampling effect, the intrinsic effect of gene length on RCBS reduces with the increase in length. A thorough exploration of theoretical values of RCBS suggests that RCBS can be an effective measure of gene expression, as its value depends on codon usage pattern along with DNA compositional bias of a gene.

In order to test the RCBS as an expression level predictor, we chose to compare our results with the experiments. We collected data sets (listed in Supplementary Tables SIII and SIV) which consist of mRNA or protein abundance data obtained by different methods—mostly cDNA microarrays[27,34,35] or 2D gel electrophoresis data[36–39] for abundances of many *E. coli* proteins are available for comparison with the predicted levels of expression. In Fig. 5, we compare the predicted levels of expression in *E. coli* with 2D gel patterns[34] and expression measure $E(g)$ of Karlin et al.[18] The relationship between RCBS values and mRNA levels seen in Fig. 5 agrees better than with the findings of Karlin et al.[18] The correlation between expression level (as relative molecular abundance) and RCBS value is found to be 0.4533 whereas that with $E(g)$ value is 0 .2618. Among the 20 most abundant proteins, 17 were identified as highly expressed genes with three exceptions for *metE, folA* and *ilvE*. The results are in good agreement with those predicted by $E(g)$. Among the 20 least abundant proteins, only three mismatch with our

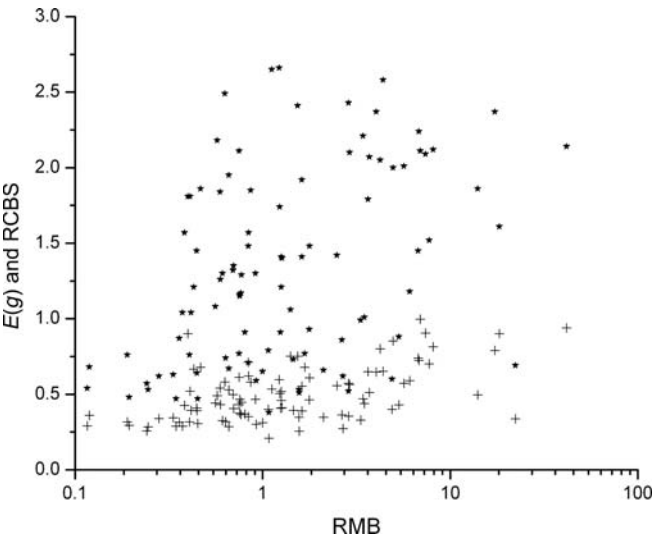

Figure 5. RCBS (+) and $E(g)$ (*) plotted against relative molecular abundance of 96 genes from *E. coli* genome.[18] RMB denotes relative molecular abundance. X-axis is taken in logarithmic scale.

predicted results whereas there are seven mismatches with the results of Karlin et al.[18] Although *pck*, *nusb*, *vals*, *args*, *rpll*, *thrs*, *leus* are less abundant, according to 2D gel patterns, the high $E(g)$ values of Karlin et al.[18] support naming the genes highly expressed. But our data support only *nusb, vals* and *rpll* to be highly expressive genes. Of the remaining 55 proteins 22 were identified as highly expressed genes. This agreement with molecular abundance data supports our predicted results better than others. In a step forward we compare RCBS and the concentrations of various proteins in *E. coli* along with their CAI values[24] (Supplementary Table SIV). Concentration is expressed as the number of protein molecules per cell. Concentration being used as a measure of gene expression, we find that our result is surprisingly good. The RCBS values along with the CAI values are plotted against the logarithm of concentration in Fig. 6. The predicted gene expression level using RCBS value is found to correlate well with the protein concentration data[24] ($r = 0.708211$). The correlation is better than the quantitative measure of CAI ($r = 0.615546$). It suggests that a quantitative estimate of the expression level by RCBS values performs better than other indices of expression measure. Thus, regardless of the state of cell growth, one can measure the relative expression level for each gene under various growth conditions, different genetic states or over a time course during environmental change.

In Fig. 7 we plotted radioactive data and microarray data against RCBS (Supplementary Table SV) for 117 genes as identified by heat shock treatment.[35] Among these, 26 genes show high (RCBS > 0.5), 84 genes moderate ($0.2 <$ RCBS < 0.5) and only seven genes show a low (RCBS < 0.2) level of expression. Despite the fact that the quality of experimental data seems to be a very important factor, we observe a good correlation between RCBS and microarray (radioactive) data ($r_{\text{micro}} = 0.2415$, $r_{\text{radio}} = 0.2098$).

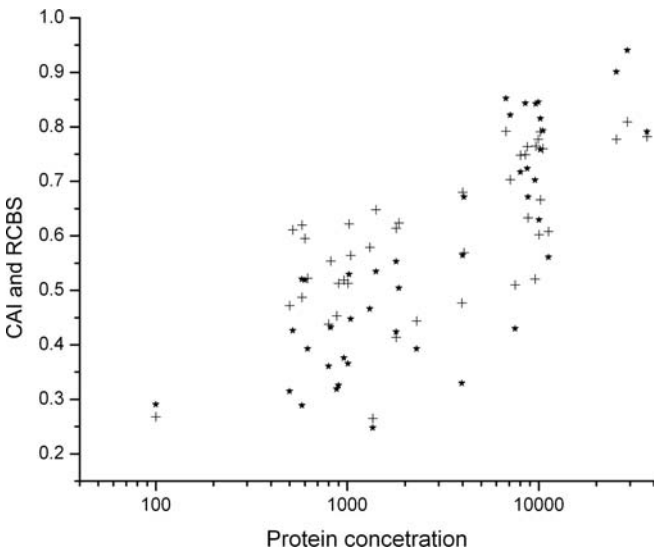

Figure 6. CAI (+) and RCBS (*) plotted against protein concentration of 45 genes from the *E. coli* genome.[24] X-axis is taken in logarithmic scale.

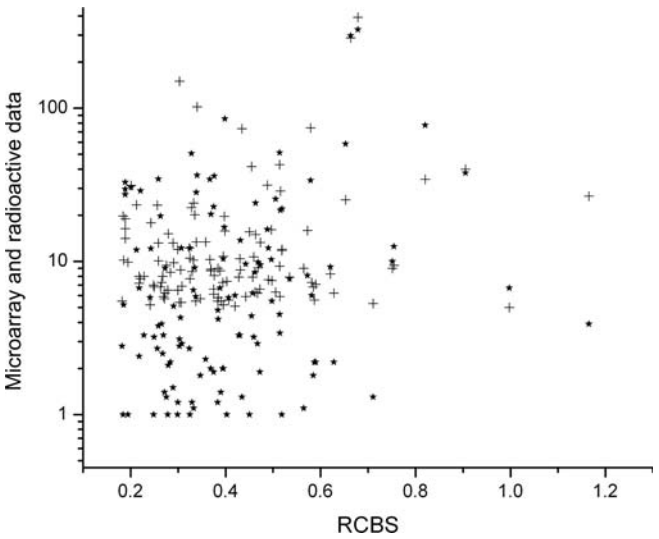

Figure 7. Radioactive data (+) and microarray data (*)[35] plotted against RCBS for *E. coli* genes. Y-axis is taken in logarithmic scale.

In another analysis we compared our expression measure (RCBS) with the genomic expression profiles of the *E. coli* genome growing on rich (Luria broth glucose) and on minimal culture (glucose) medium (Supplementary Tables SVA and SVB).[34] Of the 76 genes expressed at significantly higher levels on Luria broth plus glucose medium, 54 genes show a high expression level in our expression measure, whereas only 12 genes out of 107 genes expressed on minimal glucose medium have a high level of expression. We observe that the correlation coefficient of minimal culture data with RCBS ($r = 0.3011$) is good, but very much worse for Luria broth glucose data. The agreement of predicted and actual protein expression level varied greatly between all examined combinations of prediction method and data set. The discrepancy is thought to lie in the quality of experimental data. The preliminary analysis on the quality of experimental data shows that these kinds of experiments are inherently noisy and of low reproducibility. The reproducibility of microarray data can be evaluated through the computation of correlation coefficients within and among the data sets from different microarray experiments. Two data sets from different sources can be chosen for analysis in this study. In the first, the data set was obtained from ExpressDB and the comparison made between expression levels in *E. coli* grown to either mid-log phase (LP) or stationary phase (SP). In the

second, the data set was obtained from the ASAP database, where *E. coli* is cultured in lysogeny broth (LB). It can clearly be seen that the pair wise correlation coefficient among the gene expression levels from different experiments ($r_{\text{LP-SP}} = 0.52$, $r_{\text{LB-LP}} = 0.017$, $r_{\text{LB-SP}} = -0.039$)[34] vary broadly indicating the very noisy nature of microarray experiments and their lack of accuracy. The quality of experimental data seems to be a very important factor in this kind of analysis. Large variances may reduce the significance of statistical tests and might hide interesting trends in complex data. Microarray data tend to suffer from noise introduced at each step of different experimental protocols, while protein abundance data and mRNA expression level do not agree well in all cases. The other probable reason for incoherent results is that prediction of gene expression from genomic data, based solely on codon usage, is oversimplified. Other factors, such as promoter strength and gene copy number should also be taken into account.

We now discuss our results in more detail for different functional classes of genes. The highly expressed genes are then classified into different functional categories, e.g. RPs, CH and degradation proteins, transcription and TF, energy metabolism, electron transport, recombination and repair, outer membrane proteins, aminoacyl tRNA synthetases, etc. (The distribution of highly expressed genes of different functional class in the genomes of *E. coli* is displayed in Supplementary Table SI.) All, but one RP, the major CH/degradation proteins and translation/transcription processing factors attain high expression levels. Supplementary Table SII presents the 52 genes with the highest predicted expression levels in *E. coli*. The gene for *trp* operon ladder peptide *trpL* involved in amino acid (tryptophan) biosynthesis attains the highest RCBS value 3.42, among all *E. coli* genes.

3.1. RP genes

RPs are very important in cell biology as thus provide a range of activities required for all steps of protein biosynthesis. Following the analysis based on the definition RCBS and Equation (1) and (2), we observe that virtually all RP genes qualify as highly expressed genes. The genes encoding RPs, which are expected to be expressed at high levels during rapid cell growth, were identified with RCBS values >0.5 (Table 1). All but one RP in *E. coli* are expressed at significantly higher levels; the only exception is *rimK*, RP S6 modification protein, where it is thought to contribute to the ribosome maturation and modification. The RCBS values for highly expressed RP genes range from 0.50 to 1.77. In fact, all RP genes in *E. coli* do not reach the top expression level. Seventeen out of 56 are among the highest 86 highly expressed genes. The highest expression level occurs for L34, with an RCBS value of 1.77. The RPs are the major component, together with the ancillary proteins, involved in protein synthesis. The genes coding for RPs, protein synthesis factors and RNA polymerase subunits are all intermingled and organized into a small number of operons. We observe that the genes for some major translational or transcription processing factors, including *tufA*, *tufB*, *fusA*, *fkpA*, *slyD*, *rpoB* and *rpoC*, which are within or near the large RP operon, are predicted as highly expressed genes. Although RPs play an exclusive role in determining ribosome structure, several are multifunctional. *RplA, rplD* and *rplT*, the 50S ribosomal subunit proteins (L1, L4 and L20 respectively), and *rpsH*, the 30S ribosomal subunit protein S8 have a regulatory role. The S1 gene, a giant RP gene (labelled as *rpsA*) is essential to *E. coli* and putatively contributes to the initiation of protein synthesis. S9 (*rpsI*) participates in certain repair activities, and S16 (*rpsP*) acts as an endonucleases.

3.2. Genes for transcription/translation processing factors

There are $\sim$100 genes encoding enzymes, factors and structural components that make up the translational apparatus. Out of these100 genes 75 are identified as highly expressed genes with RCBS values >0.5. Thus the majority of genes involved in translation are predicted to have a high expression level. Of these 75 translational genes, which are expressed at higher level, 55 encoded RPs. Highly expressed genes for transcription/translation processing factors are reported in Table 1 and can be compared with the data available.[18]

There are $\sim$260 known genes that encode factors involved in translation and ribosome modification including the initiation and elongation factors, 34 of which are indicated to be at a higher expression level. As with RPs, genes coding for elongation factors (*efp*, *yeip*, *fusA*, *tsf*, *tufA*, *tufB*), ribosome recycling factor (*frr*) and translation initiation factor (*infA*) register as highly expressed genes which play important roles in translation. The expression level of *infB*, fused protein chain initiation factor is moderately high (RCBS = 0.49017). The regulation of *infB* which is downstream and co-transcribed with moderately expressed TF gene *nusA* (RCBS = 0.46579), is complex and is thought to be the result of auto regulation of the extent of the read through at upstream terminators by moderately expressed *nusA*. The expression level of *infB* is higher than *nusA*. The elongation factor *efp* has been shown to be essential in *E. coli* for protein synthesis and viability. The expression levels of other elongation factors (*fusA*, *tsf*, *tufA*, *tufB*) are gradually higher. Interestingly, the

Table 1. RCBS of the highly expressed genes of different functional class in the *E. coli* genome

Functional class	Gene	RCBS	Gene	RCBS	Gene	RCBS	Gene	RCBS
Ribosomal	*rplN*	0.50496	*rpsJ*	0.74635	*rplS*	0.87367	*rpmA*	1.08922
	rpsD	0.56061	*rplX*	0.75111	*sra*	0.88011	*rpmC*	1.09439
	rpsS	0.60728	*rpsF*	0.75859	*rplI*	0.90076	*rplO*	1.16165
	rpsM	0.61255	*rplD*	0.76302	*rpmB*	0.90877	*rpsI*	1.24694
	rpsG	0.62318	*rplM*	0.79227	*rpsN*	0.91121	*rpmG*	1.2494
	rplF	0.62913	*rplC*	0.79299	*rplP*	0.92341	*rpsT*	1.24983
	rplE	0.67119	*rplQ*	0.80176	*rpsP*	0.92858	*rplL*	1.3063
	rpsH	0.67126	*rpsB*	0.80995	*rplY*	0.9446	*rplT*	1.3222
	rpsK	0.67627	*rpsA*	0.81499	*rpsL*	0.95959	*rpsO*	1.32324
	rpsE	0.7021	*rplJ*	0.82165	*rplW*	1.00068	*rpmJ*	1.49921
	rplB	0.71682	*rpsC*	0.84223	*rpmD*	1.00368	*rpsU*	1.60846
	rplV	0.7302	*rplK*	0.84341	*rpsQ*	1.03424	*rpmI*	1.66876
	rplR	0.7344	*rplA*	0.84538	*rpmF*	1.04844	*rpmH*	1.77046
	rplU	0.73917	*rpmE*	0.85618	*rpsR*	1.05606	–	–
Translational	*Efp*	0.70878	*raiA*	0.50131	*rrfE*	1.03184	*ssrS*	0.70761
	Efs	1.31636	*rrfA*	1.11799	*rrfF*	1.02752	*tsf*	0.85208
	Frr	0.77909	*rrfB*	1.03184	*rrfG*	1.11995	*tufA*	0.94012
	fusA	0.72335	*rrfC*	1.11995	*rrfH*	1.11995	*tufB*	0.86312
	infA	0.7532	*rrfD*	1.11995	*rrlA*	1.06128	*yeiP*	0.52763
Transcriptional	*alpA*	0.64494	*glnB*	0.81972	*pspA*	0.71495	*rpoZ*	0.874
	chaB	0.91144	*greA*	0.61192	*pspB*	0.77923	*sfsB*	0.66054
	Crl	0.68275	*greB*	0.52545	*relB*	0.68232	*slmA*	0.53879
	cspA	1.2802	*Hha*	0.88747	*relE*	0.54866	*soxR*	0.59593
	cspC	1.12974	*Hns*	0.73934	*rof*	0.65143	*soxS*	0.60395
	cspE	0.87402	*metJ*	0.5234	*rpoB*	0.53467	*suhB*	0.53095
	deaD	0.62977	*nusB*	0.66651	*rpoC*	0.66692	*tdcR*	0.60661
	flgM	0.58028	*nusG*	0.62894	*rpoD*	0.53475	*trpR*	0.6079
	flhC	0.504	*osmE*	0.55743	*rpoH*	0.51287	–	–
CH and folding	*ccmD*	0.81384	*groL*	0.90549	*hybG*	0.62208	*secB*	0.66081
	dksA	0.5747	*groS*	0.82021	*iscA*	0.66931	*skp*	0.85476
	dnaK	0.65259	*hscB*	0.62877	*iscX*	0.73575	*slyD*	0.60592
	dsbA	0.59085	*hslO*	0.51531	*lolA*	0.51362	*stpA*	0.74434
	fklB	0.63123	*hslU*	0.49623	*narJ*	0.50787	*tig*	0.79986
	fkpA	0.55943	*htpG*	0.5791	*ppiB*	0.65291	–	–
	fkpB	0.51531	*hyaE*	0.56129	*ppiC*	0.70111	–	–
	fliT	0.51569	*hybF*	0.51315	*rmf*	0.96923	–	–
Outer membrane	*csgA*	0.73214	*ompC*	1.03758	*slyB*	0.59077	*yqiG*	0.69853
	mipA	0.52949	*ompF*	0.63223	*tsx*	0.58718	–	–
	nmpC	0.51413	*ompX*	0.90683	*yddL*	0.57797	–	–
	ompA	0.79079	*pagP*	0.50225	*yqhH*	0.53974	–	–
Post-translational	*rimI*	0.50362	*Def*	0.50521	*napD*	0.65324	*npr*	0.66442
DNA repair/replication/recombination	*cspD*	0.49781	*Hole*	0.70777	*ihfB*	0.58392	*rusA*	0.53058
	dinI	0.66454	*hupA*	0.97108	*priC*	0.58088	*ssb*	0.71106
	dinJ	0.57421	*hupB*	0.74465	*rdgC*	0.51482	*xseB*	0.865
	fis	0.93575	*ihfA*	0.55962	*recA*	0.60858	*yebG*	0.59001
RNA modification	*rluB*	0.55764	*Pnp*	0.59733	*deaD*	0.62977	*rbfA*	0.72106
DNA degradation	*rusA*	0.53058	*xseB*	0.865	–	–	–	–
Degradation of Proteins/peptides/glycopeptides	*hflC*	0.4998	*degP*	0.51382	*yhbO*	0.53736	*yajG*	0.55166
Degradation of small molecules	*Pta*	0.58128	*frwB*	0.57401	*tnaC*	1.33277	–	–
Nucleoprotein and basic protein	*Hfq*	0.51407	*Hns*	0.73934	*skp*	0.85476	*tpr*	1.29474
	dps	0.55438	*stpA*	0.74434	*fis*	0.93575	–	–
	ihfB	0.58392	*hupB*	0.74465	*hupA*	0.97108	–	–
Aminoacyl tRNA synthase	*aspS*	0.52912	*lysS*	0.54138	*pheM*	2.38353	*valS*	0.52017
	ygjH	0.5786	–	–	–	–	–	–
Energy metabolism								
Glycolysis	*eno*	0.99727	*gapA*	0.87498	*pfkA*	0.67783	*pykF*	0.62056
	fbaA	0.7547	*gpmA*	0.65413	*pgk*	0.76595	*tpiA*	0.80293
TCA cycle	*mdh*	0.55763	*sucB*	0.51856	*sucC*	0.50409	*sucD*	0.62233

Continued

Table 1. Continued

Functional class	Gene	RCBS	Gene	RCBS	Gene	RCBS	Gene	RCBS
Pentose phosphate pathway	*talB*	0.58526	*tktA*	0.63261				
ATP synthase	*atpA*	0.64784	*atpC*	0.51365	*atpD*	0.64873	*atpE*	1.08527
	atpF	0.60762						
Pyruvate dehydronage	*aceE*	0.57263	*aceF*	0.55269	*lpd*	0.56421		
Aerobic respiration	*cyoC*	0.53164	*hyaE*	0.56129	*nuoA*	0.54378	*nuoK*	0.61103
	cyoD	0.61485	*nirD*	0.70885	*nuoI*	0.59343		
Anaerobic respiration	*frdC*	0.73468	*hybG*	0.62208	*menB*	0.60086	*pflB*	0.75126
	frdD	0.72395	*hydN*	0.69364	*narH*	0.52986	*ubiC*	0.52458
	glpE	0.54693	*hypA*	0.67865	*narJ*	0.50787		
	hybF	0.51315	*hypC*	0.56922	*yfiD*	0.87609		
Electron transport	*ackA*	0.61336	*Fdx*	0.61409	*fldA*	0.60624	*cybC*	0.56769
Flagellum biogenesis	*flgB*	0.54626	*fliJ*	0.67522	*fliS*	0.52105	*fliT*	0.51569
	fliE	0.66739	*fliQ*	0.5854				
Transport of small molecules	*nupC*	0.50273	*potC*	0.51092	*tsx*	0.58718		
Salvage of nucleocides and nucleotides	*Apt*	0.73291	*deoC*	0.63634	*upp*	0.51826	*hpt*	0.69492
	deoB	0.55136	*deoD*	0.57449	*gpt*	0.56649		
Central intermediary metabolism	*citD*	0.59133	*folX*	0.51347	*gloA*	0.76667	*ulaD*	0.52297
	citE	0.51485	*Mutt*	0.63455	*aspA*	0.52318	*gcvH*	0.72458
	fixX	0.60213						
Carbohydrate metabolism	*eda*	0.62187	*gntK*	0.50361	*ulaB*	0.51605	*uxaC*	0.57269
	gatB	0.53522	*Lpd*	0.56421	*ulaD*	0.52297	*uxuA*	0.59595
	paaB	0.60215						
Phosphorus metabolism	*pstA*	0.51705	*pstS*	0.5871	*ppa*	0.6365	*psiF*	0.66563
	phnG	0.5443						
Nitrogen metabolism	*cynS*	0.53274	*glnK*	0.65458				
Sulphur metabolism	*cysP*	0.51334						
Amines metabolism	*eutS*	0.57934						
Amino acid biosynthesis	*artM*	0.51962	*glnH*	0.54244	*ilvG*	1.32851	*metJ*	0.5234
	dapD	0.51627	*glnP*	0.596	*ilvL*	1.51982	*pheL*	2.8411
	fliY	0.51995	*glyA*	0.57258	*ilvM*	0.84298	*sdaC*	0.62785
	glnA	0.5114	*hisL*	1.99822	*ivbL*	1.76046	*thrL*	1.7054
	glnB	0.81972	*ilvC*	0.54397	*leuL*	1.93311	*trpL*	3.41556
	trpR	0.60479						
Fatty acid biosynthesis	*accA*	0.57451	*dgkA*	0.55757	*fabI*	0.54893	*ymcE*	0.60055
	acpS	0.55661	*fabA*	0.67664	*fabZ*	0.58465		
Nucleotide biosynthesis	*adk*	0.76156	*Ndk*	0.79214	*purC*	0.5899	*pyrL*	1.1651
	guaB	0.58481	*purA*	0.53711				
Cofactor and small molecule biosynthesis	*gapA*	0.87498	*mioC*	0.50538	*moaE*	0.58446	*ubiC*	0.52458
	glyA	0.57258	*moaC*	0.50171	*ribE*	0.59736		
	menB	0.60086	*moaD*	0.61154	*This*	0.78241		
Macromolecule biosynthesis	*accB*	0.55326	*dgkA*	0.55757	*grxC*	0.79395	*mipA*	0.52949
	acpP	0.82199	*fimA*	0.57714	*hipB*	0.62205	*nrdH*	0.66531
	ccmD	0.81384	*glgS*	0.89234	*iscR*	0.50455	*pagP*	0.50225
	cybC	0.56769	*grxA*	0.55662	*Lpp*	1.632	*trxA*	0.75124
	yfgJ	0.72071						
Inner membrane	*ccmD*	0.81384	*metI*	0.53708	*yccF*	0.58505	*yidH*	0.53297
	cyoC	0.53164	*mscL*	0.57954	*ydgC*	0.55456	*yiiR*	0.51556
	cyoD	0.61485	*narH*	0.52986	*yeaL*	0.50064	*yijD*	0.50746
	dgkA	0.55757	*nuoA*	0.54378	*yeaQ*	0.71217	*yjeO*	0.54162
	frdC	0.73468	*nuoK*	0.61103	*ygdD*	0.62392	*yjeT*	0.68009
	frdD	0.72395	*nupC*	0.50273	*yhdT*	0.74646	*yncH*	0.7111
	glnP	0.596	*Pal*	0.86696	*yhhL*	0.62656	*ynfA*	0.60738
	lpp	1.632	*yaaH*	0.7921	*yiaB*	0.65847		
	mdtJ	0.61263	*ybaN*	0.55105	*yiaW*	0.64364		

Continued

Table 1. Continued

Functional class	Gene	RCBS	Gene	RCBS	Gene	RCBS	Gene	RCBS
Transport	*yjdM*	0.76533	*glnH*	0.54244	*ptsH*	0.93025	*csgF*	0.54377
	yjgA	0.5484	*glnP*	0.596	*potC*	0.51092	*secG*	0.75473
	fliY	0.51995	*mscL*	0.57954	*pmrD*	0.5388	*mokC*	0.62148
	cyoC	0.53164	*sugE*	0.51943	*yrbC*	0.54592	*yajC*	0.69682
	metI	0.53708	*mdtI*	0.74374	*frwB*	0.57401	*tatA*	0.72924
	metQ	0.56475	*mdtJ*	0.61263	*fryB*	0.70188	*tatE*	0.71983
	feoA	0.76102	*chbA*	0.55214	*yedE*	0.50339	*cysP*	0.51334
	gatB	0.53522	*chbB*	0.65397	*ygaH*	0.5262	*npr*	0.66442
	gspI	0.54627	*nuoI*	0.59343	*yqaE*	1.13838	*sdaC*	0.62785
	crr	0.6849	*nupC*	0.50273	*marB*	0.61754		
Regulator	*chpS*	0.57732	*csrC*	0.51672	*hipB*	0.62205	*yfeC*	0.5528
	cpxP	0.50596	*dsrA*	1.78721	*Spf*	1.34529	*yiaG*	0.51628
	csgA	0.73214	*dsrB*	0.75282	*sufE*	0.58559	*yifE*	0.54534
	csrA	0.83793	*feoC*	0.86637	*yddM*	0.5642	*yrbA*	0.62229

Table 2. Predicted expression levels of highly expressed prophage genes

Gene	Description	RCBS
yeeT	CP4-44 prophage; predicted protein	0.76113
alpA	CP4-57 prophage; DNA-binding transcriptional activator	0.64494
ypjK	CP4-57 prophage; predicted inner membrane protein	0.7551
yfjU	CP4-57 prophage; predicted inner membrane protein	1.07646
yfjM	CP4-57 prophage; predicted protein	0.56069
yafW	CP4-6 prophage; antitoxin of the YkfI–YafW toxin–antitoxin system	0.54248
tfaS	CPS-53 (KpLE1) prophage; conserved protein	0.60714
yfdT	CPS-53 (KpLE1) prophage; predicted protein	0.54524
yfdS	CPS-53 (KpLE1) prophage; predicted protein	0.59437
yffM	CPZ-55 prophage; predicted protein	0.72955
ninE	DLP12 prophage; conserved protein	0.61069
rusA	DLP12 prophage; endonuclease RUS	0.53058
emrE	DLP12 prophage; multidrug resistance protein	0.65874
borD	DLP12 prophage; predicted lipoprotein	0.50128
rzoD	DLP12 prophage; predicted lipoprotein	0.98537
essD	DLP12 prophage; predicted phage lysis protein	0.77232
ybcO	DLP12 prophage; predicted protein	0.56517
ybcW	DLP12 prophage; predicted protein	0.67154
ylcG	DLP12 prophage; predicted protein	1.05554
yciH	e14 prophage; 5-methylcytosine-specific restriction endonuclease B	0.67815
yciX	e14 prophage; predicted DNA-binding transcriptional regulator	0.79718
yciO	e14 prophage; predicted inner membrane protein	0.50282
rluB	e14 prophage; predicted integrase	0.55764
ymiA	e14 prophage; predicted protein	1.3517
ylcH	hypothetical protein, DLP12 prophage	1.56134
insM	KpLE2 phage-like element; iron-dicitrate transporter subunit	0.6455
insA	KpLE2 phage-like element; IS1 repressor protein InsA	0.52239
yqiG	KpLE2 phage-like element; IS2 insertion element repressor InsA	0.69853
yjhD	KpLE2 phage-like element; IS30 transposase	0.6955
relB	Qin prophage; bifunctional antitoxin of the RelE–RelB toxin–antitoxin system/transcriptional repressor	0.68232
dicB	Qin prophage; cell division inhibition protein	0.66801

Continued

Table 2. Continued

Gene	Description	RCBS
cspB	Qin prophage; cold shock protein	0.52261
cspF	Qin prophage; cold shock protein	0.5891
cspI	Qin prophage; cold shock protein	0.80085
dicC	Qin prophage; DNA-binding transcriptional regulator for DicB	0.69275
ydfK	Qin prophage; predicted DNA-binding transcriptional regulator	0.50987
ynfN	Qin prophage; predicted protein	0.69704
gnsB	Qin prophage; predicted protein	0.82038
ydfD	Qin prophage; predicted protein	0.83742
ydfA	Qin prophage; predicted protein	0.95351
ydfB	Qin prophage; predicted protein	1.34218
essQ	Qin prophage; predicted S lysis protein	0.62869
hokD	Qin prophage; small toxic polypeptide	0.75743
relE	Qin prophage; toxin of the RelE–RelB toxin–antitoxin system	0.54866

regulation *tufB* is partially dependent upon the *fis* gene, global DNA binding transcriptional and the *fis* gene has significantly higher expression level (RCBS = 0.93575). Small RNA molecules are very important in cell biology and can regulate translation. It is found that genes coding 5S RNAs (*rrfA*, *rrfB*, *rrfC*, *rrfD*, *rrfE*, *rrfF*, *rrfG*, *rrfH*) and 23S RNA (*rrlA*) have distinctive RCBS values >1.0. Gene expression is controlled by a regulator that interacts with a specific sequence of a target RNA. *Ffs* coding for the 4.5S sRNA component of signal recognition particle works with the *ffh* protein (RCBS = 0.3524) and is involved in co-translational protein translocation into and possibly through membranes. *SsrS* coding for 6S sRNA inhibits RNA polymerase promoter binding. It acts as a template for RNA-directed pRNA synthesis by RNAP and mimics an open promoter. *RaiA* codes for cold shock protein associated with 30S ribosomal subunit. *Ffs,ssrS* and *raiA* involved in translational process are predicted to be highly expressed genes in our approach.

Moreover we identify four other genes which are involved in the post-translational process and are expressed at higher level. These are *riml* coding acetylase for 30S ribosomal subunit S18, *def* coding peptide deformylase, *hypC* coding protein required for maturation hydrogenases 1 and 3, *napD* coding for assembly protein for periplasmic nitrate reductage, and *npr* coding for phosphohistidinoprotein-hexose phosphotransferage component of N-regulated peroximal targeting signal (PTS) system.

Transcription is the first stage in gene expression and the principal step at which it is controlled. The gene for major cold shock protein (*cspA*) attains a significantly high expression level (RCBS = 1.28). The gene *cspA* is a regulator needed for adaptation to atypical conditions and gives a response to temperature stimulus. *CspC* coding for other stress proteins and a member of the *cspA* family is also a highly expressed gene. Among other genes involved in the transcription process RNA polymerase plays a vital role. RNA synthesis is catalysed by the enzyme RNA polymerase. Transcription starts when RNA polymerase binds to the promoter. Among the DNA-directed RNA polymerase *rpoB, rpoC, rpoD, rpoH* and *rpoZ* subunits in *E. coli* qualify the high expression level. RNA polymerase must be able to handle situations when transcription is blocked, e.g. when DNA is damaged. In the case of *E. coli* RNA polymerase, the proteins *greA* and *greB*, which have been predicted to have a high expression level, release polymerase from elongation arrest. *Rho*, transcription termination factor, attains a moderate expression level (RCBS = 0.4749). Termination and anti-termination are closely connected and involve proteins that interact with RNA polymerase. Anti-termination is used as a control mechanism and controls the ability of the enzyme to read past a terminator into genes lying beyond. The *nus* loci code for proteins that form part of the transcription apparatus. The *nusA*, *nusb*, *nusG* functions are concerned solely with the transmission of transcription. Transcription anti-termination protein (*nusB*) and transcription termination factor (*nusG*) have high expression levels. *NusB* is required for *rho*-dependent terminators whereas *nusG* may be considered with the general assembly of all the *nus* factors into a complex with RNA polymerase. *NusA* required for intrinsic terminators has a moderate expression level (RCBS = 0.4658).

3.3. CH/degradation protein genes

CH/degradation proteins are vital in cell physiology. CHs are proteins that assist the non-covalent folding/unfolding and assembly/disassembly of other

macromolecular structures. One major function of CH is to prevent both newly synthesized polypeptide chains and assembled subunits from aggregating into non-functional structures. Many CHs are heat shock proteins, that is, proteins expressed in response to elevated temperatures or other cellular stresses. The reason for this behaviour is that protein folding is severely affected by heat and, therefore, some CHs act to repair the potential damage caused by misfolding. Other CHs are involved in folding newly made proteins as they are extruded from the ribosome. Although most newly synthesized proteins can fold in the absence of CHs, a minority strictly requires them. *DnaK* (HSP70), perhaps the best characterized CH in *E. coli*, is identified as a highly expressed gene. The Hsp70 proteins are aided by Hsp40 proteins (*DnaJ* in *E. coli*), which increase the ATP (adenosine triphosphate) consumption rate and activity of the Hsp70s. But, *dnaJ* has a low expression level (RCBS = 0.3988). It has been noted that increased expression of Hsp70 proteins in the cell results in a decreased tendency towards apoptosis. Although a precise mechanistic understanding has yet to be determined, it is known that Hsp70s have a high-affinity bound state to unfolded proteins when bound to adenosine diphosphate ribosyl, and a low-affinity state when bound to ATP. It is thought that many Hsp70s crowd around an unfolded substrate, stabilizing it and preventing aggregation until the unfolded molecule folds properly, at which time the Hsp70s lose affinity for the molecule and diffuse away. Other highly expressed heat shock proteins are *groS, groL, hslO* (Hsp33) *htpG* (Hsp90). *GroS* and *groL* are the small subunits of GroESL. These are the best characterized heat shock protein complexes in *E. coli*, identified as highly expressed genes. *HtpG* in *E. coli* is the least well-understood CH. Hsp90, a molecular CH, might be essential for activating many signalling proteins in the eukaryotic cell and is necessary for viability in eukaryotes. Since it is predicted to be a highly expressed gene, it is possibly necessary for prokaryotes as well.

Protein degradation plays an important role in cell cycle, in signal transduction and in maintaining the integrity of the proper folded state of a protein. Out of 100 genes involved in macromolecular degradation only six genes qualify as highly expressed genes. In Table 1, the predicted expression levels of highly expressed degradation genes are reported. Among these the genes encoding *xseB* (exonuclease VII small subunit) and *rusA* (DLP12 prophage, endonuclease RUS) are enzymes which regulate the degradation of DNA. These are also involved in DNA repair activity. *Pnp* and *csrA* are the only two proteins qualifying as highly expressed genes involved in RNA degradation. *Pnp*, polynucleotide phosphorylase/polyadenylase, is fundamental in RNA processing. Polyadenylation plays an important role in initiating degradation of some RNAs. Triple mutations that remove *Pnp* have a strong effect on stability. Poly(A)polymerase may create a poly (A) tail that acts as a binding site for the nucleases. *DegP*, serine endoprotease (Protease D0) encodes an enzyme which is involved in protein and peptide degradation and is predicted to be required for global protein degradation. It responds to temperature stimulus. *YhbO*, *YajG*, a predicted lipoprotein and *YhbO*, a predicted intercellular protease are thought to be involved in degradation of proteins and polysaccharides.

3.4. *Aminoacyl tRNA synthetases and modification genes*

There are 37 genes encoding the tRNA synthetases and other enzymes involved in tRNA modification. Results have been reported in Table 1. Compared with 19 *PHX* genes as predicted by Karlin et al.,[18] only three genes register as highly expressed genes in our expression measure. These include aspartyl tRNA synthetase (*aspS*), lysine tRNA synthetase (*lysS*) and valyl tRNA synthetase (*valS*). The gene encoding glysine tRNA synthetase (*glyS*) is also predicted to be a highly expressed gene marginally with RCBS = 0.4974. Among other tRNA synthetase genes *phes*, *glyQ*, *glnS*, *leus*, *serS*, *pros*, *tyrS*, *gltX* and *metG* have moderate expression levels. *PheM*, phenylalanyl tRNA synthetase operon leader peptide registers a high RCB score with RCBS = 2.1835.

3.5. *Outer membrane protein*

There are ~13 highly expressed genes encoding outer membrane proteins, as predicted by our expression measure. The expression levels of these genes have been displayed in Table 1. These include outer membrane protein (*ompA*, *ompC*, *ompF*, *ompX*), outer membrane lipoprotein (*slyB*), truncated outer membrane porin (*nmpC*), palmitoyl transferase for Lipid A (*pagP*), scaffolding protein for murein synthesizing machinery (*mipA*) and *tsx*. Moreover, *yqiG*, a predicted outer membrane user protein, *yqhH*, a predicted outer membrane lipoprotein, and *yddL*, a predicted putative outer membrane protein have been predicted as highly expressed genes in our analysis.

3.6. *Inner membrane protein*

Among the genes encoding inner membrane protein, murein lipoprotein (*lpp*) has the highest expression level (RCBS = 0.6320). Other than conserved inner membrane protein, 34 inner membrane protein genes have been listed in Table 1 as highly expressed genes. There are ~83 conserved inner membrane proteins in the *E. coli* genome. Out of

those, 17 have been predicted to be highly expressed genes (Supplementary Table SVII).

3.7. Amino acid biosynthesis

Overall, 20 of the 255 amino acid biosynthesis genes are expressed at a higher level. The *artM*, an arginine transporter subunit, *flyM*, a cystine transporter subunit, *glnH* and *glnP*, the glutamine transporter subunits are predicted to be expressed at higher levels. The *glnA* gene, which encodes glutamine synthetase, and *glnB*, which encodes regulatory protein for glumine synthetase, are expressed at higher levels. Interestingly, *hisL*, his operon ladder peptide; *ilvL*, ilvG operon ladder peptide; *ivbL*, ilvB operon ladder peptide; *leuL*, leu operon ladder peptide; *pheL*, *pheA* gene ladder peptide; *thrL*, thr operon ladder peptide; and *trpl*, *trp* operon ladder peptide are expressed at higher levels. The monocystronic gene *ilvC*, which is depressed exclusively by valine has a high value of expression score. The *dapD* product, *2,3,4,5*-tetrahydropyridine-2-carboxylate *N*-succinyl transferage, which encodes the enzyme for lysine biosynthesis process via diaminopimelate has a high expression level.

3.8. Nucleotide biosynthesis

According to MIPS classification, ~31 genes encode enzymes for nucleotide biosynthesis. In our study, we observe that five genes namely *purA*, *purC*, *adk*, *ndk* and *guaB* encoding enzymes which are involved in Purine ribonucleotide biosynthesis and *pyrL*, pyrBI operon leader peptide for Pyrimidine ribonucleotide biosynthesis, are highly expressed genes. *PyrL* has a significantly high expression level with RCBS = 1.16.

3.9. Genes for energy metabolism and metabolism of carbon compounds

Of the 392 genes involved in metabolism of carbon compound, 39 genes have a significantly high expression level. Of those, 27 are involved in carbohydrate metabolism, 10 are involved in amino acid metabolism, and two are involved in amines metabolism. *Lpd* is involved both in carbohydrate and amino acid metabolism. Rest one is involved in other carbon compound metabolism. No genes involved in fatty acid metabolism attain a high expression level, but seven of the 27 genes involved in fatty acid biosynthesis have a significantly high expression level. The data presented here indicate that *accA* (acetyl-CoA carboxylase), which encodes one component of acetyl coenzyme A carboxilase is a highly expressed gene. In addition, *ymcE*, which is cold shock protein and *aspS* also attain a high expression level. Although less is known about *fab* genes except the FadR activation on *fabA*, we predict that some of *fab* genes (*fabA*, *fabI*, *fabZ)* have a significant expression level. This is consistent with genomic expression profiling obtained from DNA microarray analysis of Tao et al.[34]

3.10. Energy metabolism genes

The genes involved in energy metabolism are primarily divided into four groups: glycolysis, pyruvate dehydronage, the pentose phosphate pathway and the TCA cycle. Of the 1530 genes that are involved in energy metabolism, 163 have been predicted to be highly expressed genes in our approach. Two basic metabolic pathways glycolysis and TCA cycle involve eight and four highly expressed genes respectively, whereas the genes in glycolysis and pyruvate metabolism are predominantly highly expressed genes. These include the genes for *eno*, *fbaA*, *gapA*, *gpmA*, *pfkA*, *pykF*, *tpiA*, *pgk*.

Unlike Karlin et al. the proteins involved in the initial steps of glycolysis (*pgi* coding glucophosphate isomerage and the proteins involved in the initial steps of TCA cycle (*gltA*, citrate synthase) are not highly expressed genes in our observation. Besides having the most TCA cycle, pyruvate dehydronage and glycolysis, *E. coli* genome has several highly expressed genes of anaerobic and aerobic respiration. Among NADH dehydrogenase nuo complex *nuoA*, *nuoI* and *nuoK* are highly expressed genes. Genes encoding α, β and ε subunits of F1 sector of membrane bound ATP synthase and b and c subunits of F0 sector of membrane bound ATP synthase genes have been predicted to be highly expressed genes. With respect to electron transport flavodoxin 1 (*fldA*) and cytochrome o ubiquinol oxidase subunit III (*cyoC*) are highly expressed gene with RCBS values 0.6062 and 0.5316, respectively. In addition, cytochrome c biogenesis protein (*ccmD)*, and cytochrome o ubiquinol oxidase subunit IV (*cyoD*) also register high expression level in our approach.

In marked contrast to Kerlin et al., *E. coli* has six highly expressed flagellar genes *flgB*, *fliE*, *fliJ*, *fliQ*, *fliS*, *fliT*. The flagellum secretion apparatus may be viewed as part of the CH family essential for bacterial viability. Assembly of a flagellum is required to export protein subunits to the outer surface of the cell. Recent evidence indicates that flagellum regulon can also influence bacterium–host interactions independent of motility.

3.11. Fatty acid biosynthesis

Fatty acid metabolism is crucial because not only does it provide various fatty acids and phospholipids necessary for cell growth, but it also serves as a source of precursors for biosynthesis of secondary metabolites. The highly expressed genes involved in

fatty acid biosynthesis included genes encoding beta-hydroxydecanoyl thioester dehydrase (*fabA*), NADH-dependent enoyl-[acyl-carrier-protein] reductase (*fabI*), (3R)-hydroxymyristol acyl carrier protein dehydratase (*fabZ*), holo-[acyl-carrier-protein] synthase 1(*acpS*), *accA*, cold shock gene (*ymcE*). Besides 3-oxoacyl-[acyl-carrier-protein] synthase I (*fabB*) has moderately high value of RCBS (RCBS = 0.4954).

3.12. Central intermediary metabolism

Several highly expressed genes in this functional class are also involved in carbohydrate metabolism. Besides other genes in this class which are also involved in nitrogen metabolism, phosphorus metabolism, amino acid metabolism, etc., our analysis identified the key genes involved in central intermediary metabolism, encoding aspartate ammonia-lyase (*aspA*), citrate lyase (*citD*, *citE*), glycine cleavage complex lipoylprotein (*gcvH*), Ni-dependent glyoxalase I (*gloA*), 3-keto-L-gulonate 6-phosphate decarboxylase (*ulaD*), D-erythro-7,8-dihydroneopterin triphosphate 2′;-epimerase and dihydroneopterin aldolase (*folX*) and D-erythro-7,8-dihydroneopterin triphosphate 2′;-epimerase and dihydroneopterin aldolase (*mutT*) as highly expressed genes. *FixX*, *4Fe-4S ferredoxin*-type protein is also registered as a highly expressed gene predicted to be involved in central intermediary metabolism.

3.13. Genomic repair proteins

An event that introduces a deviation from the usual double-helical structure of DNA is a threat to the genetic constitution of the cell. The repair system is thus very important for the survival of the cell. The repair system can recognize a range of distortions in DNA as signal for action, and is likely to have several systems able to deal with DNA damage. Table 1 reports the highly expressed repair proteins in *E. coli* genome. Other repair proteins have low to moderate expression levels. Of the 51 genes involved in DNA repair, only six genes reach a high expression level. The principal pathway for recombination repair in *E. coli* is identified by the *rec* genes. *recA*, predicted to be highly expressed genes in our approach is not only involved in recombination–repair activities, but also has another quite distinct function. It can be activated by many treatments that damage DNA or inhibit replication in *E. coli*. This causes it to trigger a complex series of phenotype changes called the SOS response, which involves the expression of many genes whose products include repair function. The other highly expressed repair genes in *E. coli* are *xseB*, *dinI*, *yebG*, *dinJ*, *rusA*. *DinI*, DNA damage-inducible protein I, and *dinJ*, predicted antitoxin of *YafQ–DinJ* toxin antitoxin system act on damaged DNA and involved in repairing damaged DNA. *YebG*, a conserved protein regulated by *LexA* functions as DNA repair.

3.14. Regulatory protein

About 440 genes in *E. coli* encode regulatory proteins. Among these regulatory proteins 62 genes are predicted to be highly expressed genes. Several of the genes in this class also function in translation, transcription, DNA repair, replication/recombination, cell process, etc. The predicted expression levels of several other highly expressed genes of specific regulatory proteins are listed in Table 1.

3.15. Biosynthesis of vitamins, cofactors and small molecules

Vitamin biosynthesis proteins have largely low expression levels. Only *ribE*, riboflavin synthetase, is highly expressed. This is in contrast to the result of Karlin et al.[18] Pathways for the synthesis of vitamins of which only small amounts are generally needed to achieve adequate function, record low RCBS values ranging from 0.1801 to 0.5974. Some of the enzymes that utilize the vitamins as cofactors are highly expressed, e.g. *accB*, acetyl-CoA carboxylase, BCCP subunit of *E. coli* is registered as highly expressed gene in our approach with RCBS = 0.5533. Expression of the 10 highly expressed genes involved in the biosynthesis of cofactors and small molecules are listed in Table 1.

3.16. Biosynthesis of other macromolecules

Among the genes encoding proteins for macromolecular biosynthesis, *lpp* attains significantly high RCBS value (RCBS = 1.6320). In addition to it, other highly expressed genes involved in macromolecular biosynthesis genes are major type 1 subunit *fimbrin* (*fimA*), DNA-binding transcriptional repressor (*iscR*) and truncated cytochrome b562 cytochrome (*cybC*). *GlsG*, a predicted glycogen synthesis protein and *yfgJ*, another predicted protein thought to be involved in macromolecular biosynthesis also attain the score of high expression level.

Of the 39 cryptic genes in *E. coli* analysed in our model, only three register as highly expressed genes. Those are *csgA*, a criptic curlin major subunit which is involved in glycoprotein biosynthesis, *mokC*, a regulatory protein of *hokC*, and *gspI*, a putative transport protein. The expression levels of these genes are 0.7, 0.62 and 0.55, respectively.

Among the genes induced under starvation conditions only *dps*, Fe-binding and storage protein (RCBS=0.5544) which provides DNA protection during starvation proteins, *rpoH*, RNA polymerase, sigma 32 (sigma H) factor (RCBS = 0.5129) are predicted as

highly expressed genes in agreement with Karlin et al.[18] Other starvation protein genes [*otsA* (RCBS = 0.2349), *otsB* (RCBS = 0.2700), *rpoE* (RCBS = 0.2781), *rpoN* (RCBS = 0.2486), *rpoS* (RCBS = 0.4093), *katE* (RCBS = 0.2359), *surA* (RCBS = 0.3936), *bolA* (RCBS = 0.4342)] have low to moderate expression levels. The survival protein *surA* which is registered as PHX with $E(g) = 1.10$ does not qualify as a highly expressed gene in our approach. Besides, we also observe that a number of genes encoding prophases are recorded as highly expressed genes in our analysis. A phase DNA molecule is often integrated into the DNA molecule of bacterium forming a prophase. A list of highly expressed genes encoding different prophases in *E. coli* is displayed in Table 2.

Apart from these classified genes, a fraction of poorly characterized genes which are generally annotated based on strong sequence similarity is also found among predicted highly expressed genes. Many of these genes encode predicted proteins and some are poorly characterized hypothetical genes. (A list of highly expressed genes which are thought to encode predicted proteins is given in supplementary Supplementary Table SVII). Our analysis thus provides strong support for significant roles of these genes which may be highly relevant for *E. coli*.

The large data set analysed here shows a clear connection between relative codon usage difference and gene expression level. Codon frequencies are found to vary between genes in the same genome and between genomes. Thus overall nucleotide composition of the genome which influences codon usage pattern introduces selective forces acting on highly expressed genes to improve efficiency of translation. This is also evident from the observation that shorter coding sequence has greater RCBS value, i.e. shorter genes have high expression level[4,5,40,41] and this is consistent with the fact that the cost of producing a protein is proportional to its length.

Interestingly, we observe that besides highly expressed protein coding genes all tRNA genes (listed in Table 3) are also registered with very high RCBS values. This observation suggests that usage of preferred codons in these and highly expressed genes is positively correlated and the highly expressed genes use a preferred set of optimal codons in accordance with their respective tRNA levels. Moreover, this result might find another important application in tRNA genes. Besides measuring expression levels of a gene, RCBS score can be remarkably used to remove the false positives in tRNA finding algorithm. Moreover, several genes of unknown functions with predicted high expression

Table 3. Predicted expression levels of tRNA genes

Gene	RCBS	Gene	RCBS	Gene	RCBS	Gene	RCBS
alaX	1.35584	*glnW*	1.96033	*leuP*	1.06805	*serT*	1.15723
alaW	1.35584	*glnU*	1.96033	*leuX*	1.18771	*serU*	1.32755
alaV	1.5556	*gltW*	1.85009	*leuU*	1.23093	*serW*	1.45877
alaU	1.5556	*gltU*	1.85009	*leuZ*	1.3515	*serX*	1.45877
alaT	1.5556	*gltT*	1.85009	*lysT*	1.91913	*thrW*	1.175
argU	1.40468	*gltV*	1.85009	*lysW*	1.91913	*thrV*	1.27061
argX	1.67244	*glyW*	1.32551	*lysY*	1.91913	*thrT*	1.27325
argQ	1.76167	*glyV*	1.32551	*lysZ*	1.91913	*Thru*	1.7256
argZ	1.76167	*glyX*	1.32551	*lysQ*	1.91913	*trpT*	1.62046
argY	1.76167	*glyY*	1.32551	*lysV*	1.91913	*tyrU*	1.00445
argV	1.76167	*glyT*	1.33638	*metY*	1.22225	*tyrV*	1.0433
argW	1.99759	*glyU*	1.47125	*metZ*	1.32682	*tyrT*	1.0433
asnT	1.87865	*hisR*	1.21868	*metW*	1.32682	*valW*	1.37166
asnW	1.87865	*ileX*	1.41462	*metV*	1.32682	*valT*	1.37566
asnU	1.87865	*ileV*	1.42883	*metU*	1.36722	*valZ*	1.37566
asnV	1.87865	*ileU*	1.42883	*metT*	1.36722	*valU*	1.37566
aspU	1.38539	*ileT*	1.42883	*pheV*	1.38483	*valX*	1.37566
aspV	1.38539	*ileY*	1.45397	*pheU*	1.38483	*valY*	1.37566
aspT	1.38539	*leuW*	1.02415	*proL*	1.26942	*valV*	1.6125
cysT	1.35851	*leuT*	1.03107	*prom*	1.38923	*selC*	1.28639
glnX	1.65127	*leuV*	1.03107	*proK*	1.44416	–	–
glnV	1.65127	*leuQ*	1.03107	*serV*	1.14888	–	–

Table 4. Predicted expression levels of highly expressed hypothetical protein genes

Gene	RCBS	Gene	RCBS	Gene	RCBS
ytcA	0.51055	*ylcI*	0.77343	*ybhU*	1.09738
ybfK	0.51884	*yojO*	0.84734	*ynhF*	1.15141
ymjA	0.58644	*ygdT*	0.85155	*ydgU*	1.48121
yrhD	0.63276	*ypaB*	0.92206	*ypfM*	1.86114
ydbJ	0.63348	*yccB*	1.07903	*ylcH*	1.56134

levels may be attractive candidates for experimental characterization because we assume that they have important functions in those organisms. Table 4 lists such gene families of unknown functions. This kind of analysis is valuable in helping to identify the promising candidate genes to be focused for further experimental characterization.

4. Discussion

Our analysis supports that each genome has evolved codon usage patterns indicating gene expression levels. The three protein families – RPs, major translation/transcription processing factors, and CH/degradation proteins which are fundamental at many stages of the life style in promoting growth and stability, have been identified as highly expressed genes. Although the concept of predicting gene expression from codon usage was proposed a decade ago, only recently these methods have been successfully applied to the identification of highly expressed genes in various bacteria and eukaryotic organisms. But, any such codon usage-based prediction of gene expression relies on a prior definition of a reference set, consisting of highly expressed genes. For instance, CAI listed a set of 27 highly expressed genes for *E. coli*, which includes gene encoding 17 RPs, four elongation factors, four outer membrane protein, *recA*, and *dnaK*. For *yeast* a set of 24 highly expressed genes has been taken as a reference set. These include 16 genes encoding RPs, one for an elongation factor, two *enolase* genes, two *GA-3-PDH* genes, *ADH 1*, *PCK*, *pyruvate kinase*.[3] Karlin and coworkers[17–23] included transcription/translation-related factors and CHs in the reference set, in addition to the RP genes. MILC-based expression level predictor MELP[13] is based on a reference set consisting of all genes coding for RPs, longer than 100 codons. 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 for individual species. The outcome is highly dependent on the genome examined. In some instances, in the use of alternative reference sets results are very poor. In principle it is not possible to regulate protein expression level by the judicious use of certain codons. It is worth emphasizing that individual genes tend to favour characteristic codon distributions and there is a strong connection between protein expressivity and the degree of codon bias. So, we emphasize that codon assignment as well as codon preferences should be taken into account in a single measure which will have functional feedback between the constraints of gene expression and microstructure of genomes. To better understand potential expression levels of genes, we developed a methodology that relates codon usage as well as large-scale DNA compositional biases among gene classes to the expression potential of individual genes. The CAI[3] and codon usage models[13,17] are originally based on somewhat qualitative assumptions about the expression levels of relatively few genes. This is our motivation for using a quantitative measure (RCBS) to recalculate genome-wide expression data. The new approach begins with the assumption, based on the argument just presented; that the general codon usage features observed in highly expressed genes greatly differ from that of randomly generated sequences with their sequence composition conserved. Our proposition is based on the fact that the difference between the geometric average of normalized frequency of codons (f_{xyz}) in a sequence of nucleotides and that of $f_1(x) \times f_2(y) \times f_3(z)$ is >0.5 of the geometric average of $f_1(x) \times f_2(y) \times f_3(z)$ for highly expressed genes. The proposed threshold value (0.5) of RCBS is investigated for *E. coli* genome, Yeast genome and archeal genomes. The data (available on request) provide the evidence in favour of potential strength of our expression measure over the others. The most of the housekeeping genes fall in the category of highly expressed genes. The study also identifies a number of functionally unknown genes as highly expressed genes based on their codon profile. Thus, it often seems sufficient that our approach is a better alternative to the existing expression models. Surprisingly, we have found that there is a strong negative correlation between relative codon usage bias and protein length in contradiction with others.[24,42] Although our primary motivation in developing this novel method was to compensate the possible artefacts due to sequence length variability, we have observed that highly expressed genes (identified by RCBS) show negative correlation with gene length leading to a biological relevance. This is suggested to be due to more effective translational selection acting to reduce size of the abundant proteins, to minimize transcriptional and translational energy costs. Although the longer sequences appear to be better optimized in terms of having codons for more abundant tRNAs which increase their

probability in proper and timely translation, it is easier for a ribosome to translate a short RNA sequences, as opposed to decrease in fidelity for longer translation. Therefore it is likely that there is a natural selection for the shorter genes to be expressed at higher level.[41]

To summarize, we have introduced a novel method, based on codon usage difference with regard to random base composition at three codon sites, to estimate the level of expression of a gene. In this article, predicted highly expressed genes are characterized for *E. coli* genome only, but the method equally applies to other microbes to be reported in separate communication. By comparing its performance with other commonly used measures of gene expression, we have established that RCBS is a generally applicable method, being resistant to species specific and introduces little noise into measurements. It is remarkable that the present model usually performs as well as other codon usage model of Kerlin et al.[18] sometime lead to a better correlation with expression data according to several other measures based on CAI.[3] The prediction of expression level in our approach can be appreciated by comparing them with the protein abundance data and microarray data. Thus, our method is effectively complementary to the experimental procedures of 2D gel electrophoresis and DNA microarray analysis in assessing gene expression levels. In contrast to other existing measures, our model describes the global enrichment of a codon in highly expressed genes with no restrictions on composition of the other codons. Of course, the codon-based expression indicators yield static value, whereas gene expression is a dynamic process with very different expression levels under different conditions. In our view codon usage pattern of genomes evolves as a result of interplay between mutational and selective forces and the proper account of the adaptive response to the codon assignment can lead to a practical solution of gene expression.

Acknowledgements: The authors would like to acknowledge the reviewers for their valuable suggestions and comments to improve the manuscript.

Supplementary data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

Financial support by the University Grants Commission, India, sanction No. F.PSW-060/05-06 (ERO), is gratefully acknowledged.

References

1. Gouy, M. and Gautier, C. 1982, Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.*, **10**, 7055–7073.
2. Holm, L. 1986, Codon usage and gene expression, *Nucleic Acids Res.*, **14**, 3075–3087.
3. Sharp, P. M. and Li, W. H. 1986, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.*, **15**, 1281–1295.
4. Ikemura, T. 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.*, **146**, 1–21.
5. 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.
6. Karlin, S., Mrazek, J. and Campbell, A. M. 1998, Codon usage in different gene classes of Escherichia coli genome, *Mol. Microbiol.*, **29** (6), 1341–1355.
7. Wright, F. 1990, The effective number of codons used in a gene, *Gene*, **87** (1), 23–29.
8. Morton, B. R. 1994, Codon use and rate of divergence of land plant chloroplast genes, *Mol. Biol. Evol.*, **11** (2), 231–238.
9. Shields, D. C. and Sharp, P. M. 1987, Synonymous codon usage in bacillus subtilis reflects both translational and mutational biases, *Nucleic Acid Res.*, **15** (19), 8023–8040.
10. Freire-Picos, M. A., Gonzalez-Siso, M. I., Rodriguez-Belmonte, E., Rodriguez-Torres, A. M., Ramil, E. and Cerdan, M. E. 1994, Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes, **139** (1), 43–49.
11. Urrutia, A. O. and Hurst, L. D. 2001, Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection, *Genomes*, **159** (3), 1191–1199.
12. Wan, X. F., Xu, D., Kleinhofs, A. and Zhou, J. 2004, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes, *BMC Evol. Biol.*, **4** (1), 19.
13. Supek, F. and Vlahovicek, K. 2005, Comparison of codon usage measure and their applicability in prediction of microbial gene expressivity, *BMC Bioinformatics.*, **6**, 182.
14. Karlin, S. and Mrazek, J. 1996, What drives codon choices in human genes?, *J. Mol. Biol.*, **262** (4), 459–472.
15. Jansen, R., Bussemaker, H. J. and Gerstein, M. 2003, Revisiting the codon adaptation index from a whole-genome perspective:analyzing the relationship between gene expression and codon occurance in yeast using a variety of models, *Nucleic Acids Res.*, **31** (8), 2242–2251.
16. Wu, G., Culley, D. E. and Zhang, W. 2005, Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism, *Microbiology*, **151**, 2175–2187.
17. Karlin, S. and Mrazek, J. 2000, Predicted highly expressed genes of diverse prokaryotic genomes, *J. Bacteriol.*, **182**, 5238–5250.
18. Karlin, S., Mrazek, J., Campbell, A. M. and Kaiser, D. 2001, Characterizations of highly expressed genes of four fast-growing bacteria, *J. Bacteriol.*, **183**, 5025–5040.

19. Karlin, S., Mrazek, J., Ma, J. and Brocchieri, L. 2005, Predicted highly expressed genes in archeal genomes, *PNAS*, **102**, 7303–7308.
20. Mrazek, J., Bhaya, D., Grossman, A. R. and Karlin, S. 2001, Highly expressed and alien genes of the Synechocystis genome, *Nucleic Acids Res.*, **29** (7), 1590–1601.
21. Karlin, S., Barnett, M., Campbell, A. M., Fisher, R. F. and Mrazek, J. 2003, Predicting gene expression levels from codon biases in α-probacterial genomes, *PNAS*, **100**, 7313–7318.
22. Karlin, S., Brocchieri, L., Mrazek, J. and Kaiser, D. 2006, Distinguishing features of δ-probacterial genomes, *PNAS*, **103**, 11352–11357.
23. Karlin, S. and and Mrazek, J. 2004, Comparative analysis of gene expression among low G+C gram-positive genomes, *PNAS*, **101**, 6182–6187.
24. Eyre-Walker, A. 1996, Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy, *Mol. Biol. Evol.*, **13**, 864–872.
25. Coughlan, A. and Wolfe, K. H. 2000, Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*, *Yeast*, **16**, 1131–1145.
26. Martin-Galiano, A. J., Wells, J. M. and de la Campa, A. G. 2004, Relationship between codon bised genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*, *Microbiology*, **150**, 2313–2325.
27. dos Reis, M., Wernisch, L. and Savva, R. 2003, Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome, *Nucleic Acids Res.*, **31** (23), 6976–6985.
28. Semon, M., Mouchirdoud, D. and Duret, L. 2005, Relationship between gene expression and GC-content in mamals:statistical significance and biological relevance, *Hum. Mol. Genet.*, **14**, 421–427.
29. Goncalves, I., Duret, L. and Mouchiroud, D. 2000, Nature and structure of human genes that generate retropseudogenes, *Genome Res.*, **10**, 672–678.
30. Duret, L. 2002, Evolution of synonymous codon usage in metazoans, *Curr. Opin. Genet. Dev.*, **12**, 640–649.
31. Ponger, L., Duret, L. and Mouchiroud, D. 2001, Determination of CpG islands: expression in early embryo and isochore structure, *Genome Res.*, **11**, 1854–1860.
32. Vinogradv, A. E. 2003, Isochores and tissue specificity, *Nucleic Acids Res.*, **31**, 5212–5220.
33. Urruita, A. O. and Hurst, L. D. 2003, The signature of selection mediated by expression on human genes, *Genome Res.*, **13**, 2260–2264.
34. Tao, H., Bausch, C., Richmond, C., Blattner, F. R. and Conway, T. 1999, Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media, *J. Bacteriol.*, **181**, 6425–6440.
35. Richmond, .C. S., Glasner, J. D., Mau, R., Jin, H. and Blattner, F. R. 1999, Genome-wide expression profiling in *Escherichia coli* K-12, *Nucleic Acids Res.*, **27** (8), 3821–3835.
36. VanBogelen, R. A., Abshire, K. Z., Pertsemlidis, A., Clark, R. L. and Neidhardt, F. C. 1996, Gene-protein database of *Escherichia coli* K-12, In: Neidhardt, F. C., Curtiss, R. III, Ingraham, J. L., Lin, E. C. C. and Umbarger, H. E. (eds.), Escherichia coli and Salmonella*:* Cellular and Molecular Biology*, 6th edn*, Washington, D.C: ASM Press, pp. 2067–2117.
37. Pederson, S., Bloch, P. L., Reeh, S. and Neidhardt, F. C. 1978, Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates, *Cell*, **14**, 179–190.
38. Bloch, P. L., Philips, T. A. and Neidhardt, F. C. 1980, Protein identification on O'Farrell two dimensional gel:-location of 81 *Escherichia coli* proteins, *J. Bacteriol.*, **141**, 1409–1420.
39. Philips, T. A., Bloch, P. L. and Neidhardt, F. C. 1980, Protein identification on O'Farrell two dimensional gel: location of 55 *Escherichia coli* proteins, *J. Bacteriol.*, **144**, 1024–1033.
40. Comeron, J. M., Kreitman, M. and Aguade, M. 1999, Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila, *Genetics*, **151**, 239–249.
41. Duret, L. L. and Mouchiroud, D. 1999, Expression pattern, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, *PNAS*, **96**, 4482–4487.
42. Moriyama, E. N. and Powell, J. R. 1998, *Nucleic Acids Res.*, **26**, 3188–3193.
43. Merk, I. R. 2003, A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency, *J. Mol. Evol.*, **57** (4), 453–466.
44. Wagner, A. 2000, Infering lifestyle from gene expression patterns, *Mol. Biol. Evol.*, **17** (12), 1985–1987.