MedCrave
Step into the World of Research

Journal of Investigative Genomics

Mini Review

Open Access

CrossMark

In *Silico* prediction of gene expression based on codon usage: a mini review

Abstract

The present review deals with the computational approach for estimating the average expression level of proteins in microorganisms. Based on the notion that codon assignment as well as codon usage patterns might play a key role to the solution of practical problems of gene expression, alternative models are proposed and developed to identify the highly expressed genes in diverse genomes. Facts and ideas presented here aims at deriving biological information from genome sequences by means various statistical analyses and appropriate design of algorithms.

Keywords: codon usage, codon bias, gene expression, predicted highly expressed genes

Volume 4 Issue 2 - 2017

Ria Rakshit,[1] Satyabrata Sahoo[2]
[1]Department of Botany, Dhruba Chand Halder College, India
[2]Department of Physics, Dhruba Chand Halder College, India

Correspondence: Satyabrata Sahoo, Department of Physics, Dhruba Chand Halder College, Dakshin Barasat, South 24 Parganas, WB, India, Email dr_s_sahoo@yahoo.com

Received: July 03, 2017 | **Published:** August 23, 2017

Introduction

The complexity and diversity of genomic function can only be understood at the nucleotide level. The genetic code uses sixty one codons to translate twenty amino acids and three to stop translation. The genes in DNA sequence encode proteins, which perform all the functions necessary in the cell. Regulation of gene expression, thus, plays a central role in defining cell viability and is dominated by a number of genes that are involved in protein biosynthesis, mutation, and folding. The essential functions of many of the gene products strongly depend on codon usage bias that corresponds to abundant tRNAs and thus, biased amino acid composition of proteins indicating the high biological significance of these genes. Recently there has been considerable interest in finding the informational content in gene sequences. As automated sequencing techniques have started to produce a rapidly growing amount of raw DNA sequences, the extraction of information from these sequences becomes a scientific challenge. These sequences carry relevant biophysical information in the form of a one dimensional chain of four nucleotide bases. The availability of the complete genomic sequences has made it possible for researchers to develop approaches that focus on the systematic properties of regularatory and metabolic networks, and to investigate gene expression and regulation in the context of a global cellular work. Despite much effort spent, it is still an open question what kind of correlation exists in gene sequences which rule the tentative function of the particular gene. Hence, our basic task in this review is to search for correlation which accommodates the gene expression level for most situations of its habitat, energy sources and life style.

The immense progress made recently in molecular biology has revealed that genomes are of extraordinary complexity. The sequencing of DNA has shed some light on one of the main characteristic features of the genomic sequence, namely its local and global compositional heterogeneity. The biochemical techniques alone are not sufficient to uncover the genetic information from these sequences. On the other hand, scientists in this field are trying combinations of different methods used in different scientific field to understand this complex structure and the computational tools based on concepts used in many scientific fields have recently found to be relevant. A relevant contribution is due to statistical methods, namely Markovian approximation, correlation function and Fourier transform etc. A considerable attention in this field has been devoted to diverse physical phenomena exhibiting cluster behaviour in space or time domains. Such cluster behaviour appears in processes in fluid mechanics, solid state physics, data transmission system and Brownian motion. The possible relevance of scale invariance and fractal concepts to the structural complexity of genomes has been the subject of considerable increase interest today. During the past few years, there has been intense discussion about the existence, the nature and the origin of the correlations in DNA sequences. The study of the mutual information function, the power spectrum, the scaling properties of the moments and the fluctuations Voss[1] have pointed towards hidden correlations in these sequences. The genomes have parts that code for proteins and parts that do not. In relatively higher organisms, it is the non-coding parts that make up the bulk of the sequence. The coding regions are few and far between. It has been proposed that the long range fractal sort of correlations appear to be present more in the non-coding parts than in the CDS. In this view the CDS, aside of its 3-periodicity, has random nucleotide distribution. The initial reports on long range correlations in NCDS have generated contradicting responses. Some support Peng et al.,[2] Nee,[3] Li et al.[4] the existence of the long range correlations in NCDS, while some disagrees. However the conclusions are inconsistent with one another and doubt the nature of correlation existing in the different parts of the sequences.

Recently there have been considerable studies to predict the highly expressed genes in prokaryotes and in eukaryotes. Genes can be expressed as a wide range of levels. Numerous studies have investigated gene expression by diverse technology. It is well established that the gene expression is regulated at several levels:

a. Initiation of transcription, promoter strength, promoter configuration and transcription factors

b. Transcription termination, mRNA stability, and turnover rates

c. Codon usage

d. Translation initiation and elongation

e. Protein folding, degradation, and cellular localization.

Moreover, the relative influence of each of these factors varies from genome to genome, and from gene to gene. So far an accounting

of gene expression focuses on (1) favorable codon usage, (2) G+C content at the different codon sites and on (2) tRNA availabilities. Codon preferences vary considerably within and between organisms Grantham et al.[5] Sharp et al.[6] Karlin et al.[7] Within a genome, codon bias tends to be much stronger in highly expressed genes than in genes expressed at lower levels Sharp et al.,[6] Li[8] Lafay et al.,[9] Dos Reis et al.[10] Across genomes, the G+C composition resulting from mutational bias has been hypothesized to have major influence Wright[11] Friberg et al.[12] in codon usage bias of different organisms Knight et al.[13] To dissect the patterns and causality of codon usage, many indices have been proposed to measure the degree and direction of codon bias Sharp[6] Li[8] Wright.[11] These indices Sharp & Li[8] Ikemura[14] Roymandol et al.,[15] Das et al.[16,17] Sahoo et al.[18,19] Das et al.,[20] have been shown to correlate with mRNA expression levels Coghlan & Wolfe[21] Dos Reis et al.[10] Martin-Galiano et al.[22] Unfortunately, these methods are not universally applicable, as their behaviour tends to be context-dependent. 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 immune precipitation, DNA chip technology and Serial Analysis of Gene Expression (SAGE) have been developed for the purpose. All these experiments require a wide range conditions to match, massive investment of time and resources. To overcome these major obstacles for identifying highly expressed genes in vast majority of organisms, we must look beyond the direct experimental methods. Following this, we focused our study in developing computational methodology that can be used to study large scale gene expression profile of an organism.

Methods

Based on the hypothesis that highly expressed genes are often characterized by strong compositional bias in terms of codon usage, there are a number of measures currently in use that quantify codon usage bias in genes, and hence provide numerical indices to predict the expression levels of genes. In the present communication, we have limited our discussion to a variety of computational tools like CAI, E(g),RCBS, RCA, and MRCBS only.

The CAI mode[8]

The Codon Adaptation Index, CAI is given by

$$CAI = \left(\prod_{1}^{N} w_i\right)^{\frac{1}{N}} \quad (1)$$

Where, N is the number of codons in the gene and relative adaptiveness, w_i is defined as

$$w_i = \frac{f_i}{f_{aa,\max}} \quad (2)$$

f_i is the frequency of the i^{th} codon, and $f_{aa,max}$ 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.

The codon usage difference model[7]

The codon bias of a gene *g* relative to a set of gene *G* is defined as

$$B\left(g|G\right) = \sum_{aa} p_{aa}\left(g\right)\left(\sum_{(x,y,z)=aa}\left|f\left(x,y,z\right) - g\left(x,y,z\right)\right|\right) \quad (3)$$

Where $p_{aa}\left(g\right)$ is the fraction of amino acid aa in gene g; *f(x, y, z)* the frequency of a codon triplet *(x, y, z)* in gene g normalized such that *f(x, y, z)* = 1 if *(x, y, z)* is the most common synonymous codon; *g(x, y, z)* is the corresponding normalized codon frequency in gene set *G*.

Then expression level of gene is measured by

$$E\left(g\right) = \frac{B\left(g|C\right)}{\frac{1}{2}B\left(g|RP\right) + \frac{1}{4}B\left(g|Ch\right) + \frac{1}{4}B\left(g|Tf\right)} \quad (4)$$

Where the gene set *C* comprises all genes in the genome, *RP*, the ribosomal proteins, *Ch*, chaperones, and *Tf*, translation processing factors. *E(g)* is close to zero if gene g has a codon composition close to the average composition of the genome, while *E(g)* would take on very large values if the codon composition of gene *g* is close to the composition of ribosomal genes, chaperones and translation processing factors . The idea is that highly expressed genes tend to have higher values of *E* than lowly expressed genes.

Relative codon bias strength (RCBS):[15,16]

The expression measure of a gene, RCBS is given by

$$RCBS = \left(\prod_{i=1}^{L}\left(1 + d_{xyz}^{i}\right)\right)^{1/L} - 1 \quad (5)$$

Where

$$d_{xyz}^{i} = \left\{f_{xyz} - f_1\left(x\right) f_2\left(y\right) f_3\left(z\right)\right\} / f_1\left(x\right) f_2\left(y\right) f_3\left(z\right)$$

is the relative codon usage bias of i^{th} codon of a gene, f_{xyz} the normalized codon frequency for the codon *xyz* , $f_n(m)$ the normalized frequency of base *m* at codon position m in a gene. L is the number of codons in the gene.

Relative codon adaptation (RCA)[23]

The relative codon adaptation (RCA) for an entire genome is computed as

$$RCA = \left(\Pi_{(i=1)}^{L} RCA_{xyz}\left(i\right)\right)^{(1/L)} \quad (6)$$

$$MRCBS = \prod_{i=1}^{N}\left(MRCBS_{xyz}\right)^{1/N} \quad (7)$$

Where f_{xyz} is the observed frequency of codon *xyz* in any particular reference gene set, $f_n(m)$ the observed frequency of base *m* at codon position *n* in the same reference set, and L the length in codons of the query sequence. Like CAI and RCBS, RCA is computed as the geometric mean of the RCA_{xyz} term for each codon *xyz* in the sequence of interest. It depends on the size of the reference set as relative frequency used to calculate RCA of the codons.

Citation: Rakshit R, Sahoo S. In *Silico* prediction of gene expression based on codon usage: a mini review. *J Investig Genomics*. 2017;4(2):42–45.
DOI: 10.15406/jig.2017.04.00063

Modified relative codon bias strength (MRCBS)[17–20]

The modified relative codon bias strength, MRCBS measures the expression level of a gene and is defined as,

$$MRCBS = \prod_{i=1}^{N} \left(MRCBS_{xyz} \right)^{1/N} \quad (8)$$

$$MRCBS_{xyz} = \frac{RCBS(xyz)}{RCBS_{aa,\max}}, \; RCBS(xyz) = \frac{f_{xyz}}{f(x)_1 f_2(y) f_3(z)} \quad (9)$$

Where f_{xyz} is the relative observed frequency of codon *xyz* in any particular reference gene set, $f_n(m)$ the relative observed frequency of base *m* at codon position *n* in the same reference set, $RCBS_{aa,max}$ is the maximum RCBS of codon encoding same amino acid *aa* in the same reference set, and N the length in codons of the query sequence. $MRCB_{xyz}$ is independent of the size of the reference set as it is the ratio of the RCBS of the codon xyz to the maximum of RCBS of codon encoding same amino acid.

Discussion

The Codon Adaptation Index (CAI) is a simple, effective measure of synonymous codon usage bias, specifically in the direction of the bias seen in highly expressed genes. Calculation of CAI score depends on the knowledge of codon bias of a set of highly expressed genes and the relative merits of each codon of a gene with respect to the reference set of highly expressed genes have been taken into consideration in calculating the score of a gene under study. Thus, the index estimates the degree of selection of a codon in moulding the pattern of codon usage and in that respect it is useful for predicting the level of expression of a gene. The expression measure of a gene, E(g), is devised to predict the gene expression level from the codon usage difference model. It calculates the codon usage difference B(g/G) of a gene relative to three classes of gene, namely, RP(ribosomal proteins), TF (Transcription processing factors) and CH (Chaperon degradation). Predition of highly expressed genes is based on the score of E(g), provided they have high value of B(g/G) with respect to all protein coding genes, but low value with respect to RP,TF and CH. But, there are limitations in this model in predicting the gene expression profile. It has been observed that genes with strong selected codon bias are not likely to have the high value of E(g) and thus, are unlikely to be predictable as highly expressed.

Relative codon bias (RCB) is the difference of observed frequency of a codon from the expected frequency under the hypothesis of random codon usage that the base composition was biased at three sites as that in the genome 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 as a guide to their likely expression level and has been shown to be an improvement in accuracy of quantitative measurement of gene expression over CAI and other codon bias indices in some micro organisms. It is a novel method to estimate codon usage bias and, thus, to predict the gene expression level without any reference set. It is further observed that a strong correlation exists between RCBS and protein length indicating natural selection in favour of shorter genes to be expressed at higher level. This may not be always true, but the result may be due to the artifact of this model which constantly overestimates the intrinsic bias of short sequences. A statistical analysis to assess the strength of relative codon bias in genes as a potential numerical measure to predict the gene expression level suggests a decrease of the informational entropy in the highly expressed genes. The relative codon adaptation (RCA) is a reference-set-based codon bias index like CAI. Like RCBS, it directly takes into account the genomic base composition in contrast to uniform background as hypothesized in case of CAI. However, there is an important difference between CAI and RCA. In CAI, the relative adaptiveness of a codon (w_i) is computed as the ratio between the frequency of that codon in the reference set and the largest frequency among its synonymous codons and thus, the background nucleotide distribution in this model is assumed to be uniform. The inability to take into account background nucleotide composition is a fundamental problem of many CBIs. The RCA index first computes the expected frequency of a codon based on its positional base frequencies. It then measures codon adaptation as the deviation of the observed codon frequency from the expected codon frequency. Like RCBS, RCA takes explicitly into account sequence composition to provide more robust and accurate estimates of gene expression. This improvement is reflected by the fact that unlike RCBS, RCA do not present significant length dependency.

Although CAI has been universally accepted as a standard measure of codon bias for prediction of gene expression levels, the determination of the reference set of highly expressed gene is a major problem to calculate CAI. Besides, CAI is also found to be relatively noisy in the short region to capture local codon bias pattern.[24,25] However, the determination of highly expressed gene as a reference set is not required for calculating score of RCBS, but RCBS has partial dependence on gene length (for genes having length<300 *aa*).[23] Like CAI, RCA also depends on the knowledge of codon bias of highly expressed genes and it also depends on the size of the reference set as relative frequency is used to calculate relative adaptation of the codons.[19,20] 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 what the most common dominant codon bias of a genome is, if it exists at all. Instead, it is more appropriate to set numerical criteria to detect the tendency of a gene toward a bias and to measure the strength of this bias. The numerical coefficients can be used to rank different genomes with respect to given bias, and to detect whether a genome has tendency for a bias or not. Here, we suggest a threshold that is an indicator for strong bias. It will allow us to automatically identify the highly expressed genes. It is important to quantify how much information each genomic sequence carries with respect to expressivity. Thus, MRCBS has been devised as an alternative model to predict gene expression level from their codon compositions in such a way that score of the expression indicator may be calculated without any knowledge of previously set selective highly expressed genes as a reference set. With a view of evolving codon assignments as well as codon usage patterns as the adaptive response of genomes, a threshold score has been formulated in this model as a benchmark for identifying the highly expressed genes. The predicted highly expressed genes (PHE) are then characterized on the basis of the strength of the codon usage bias as derived from this model and a gene is identified as PHE gene provided its MRCBS exceeds threshold value. The significant performance of the methodology for estimating expression levels in archaeal genome makes this index a superior choice for predicting gene expression profile in different organisms. It has been further observed that MRCBS correlates well with CAI than other codon bias measures in archaeal genomes.

The experimental methods are very expensive and laborious. Results of gene expression profiling by computational methods might be used as reference data for validating and better understanding experimental data. The observation that highly expressed genes will

Citation: Rakshit R, Sahoo S. In *Silico* prediction of gene expression based on codon usage: a mini review. *J Investig Genomics*. 2017;4(2):42–45.
DOI: 10.15406/jig.2017.04.00063

preferentially choose a small subset of codons recognized by the most abundant tRNA species suggests that highly expressed genes are often characterized by strong compositional bias in terms of codon usage. Based on this hypothesis, a number of varieties of software tools like Codon Adaptation Index (CAI),[8] Relative Codon Adaptation (RCA),[23] Relative Codon Bias Strength (RCBS),[16,17] MRCBS[18–20] etc. are currently in use. These provide numerical indices to predict the expression levels of genes. There are no universal standards to make these results more suitable for comparative analysis.

Conclusion

Despite the continuing debate on different methods on rather struggling questions, it is now well admitted that certain kind of correlation do exist in genomic sequences and it is not just an artifact of the non uniformity in the composition of genes. But, their biological interpretation still remains a continuing debate and furthermore, it is still an open question whether the correlation properties are different for protein-coding and non-coding regions of nucleotide sequence and how they can be related to the expression and regulation of genes. It is, thus, essential to develop a novel method to quantify the level of expression of a gene and to understand the nucleotide structure of the gene in a genomic DNA sequences which is believed to contain vast information of our life cycle. In conclusion, we would like to emphasize the notion that codon assignment as well as codon usage patterns as the adaptive response of genomes to the solution of practical problems of gene expression. Of course, there are many limitations of the expression data that might confuse the relationship between expression levels and codon composition, because the available experimental data is subject to many biophysical and biochemical constraints. But, the value of the codon-based indicator MRCBS can perhaps be appreciated by comparing it to the other commonly used measures of gene expression. The approach is validated on a number of slow-growing and fast-growing bacteria and archaeal genomes, *Saccharomyces cerevisiae, Caenorhabditis elegans* and *Drosophila melanogaster*. The guiding line to this review aims at designing an appropriate algorithm to derive biological information from genome sequences by means of a purely mathematical analysis.

Acknowledgements

The authors would like to acknowledge the editor and reviewer for their valuable suggestions to improve the manuscript.

Conflict of interest

The authors confirm that there is no conflict of interest for publication of the resent article.

References

1. Voss RF. Evolution of long range correlation and 1/f noise in DNA base sequences. *Phys Rev Lett.* 1992;68(25):3805–3808.
2. Peng CK, Buldyrev SV, Goldberger AL, et al. Long range correlations in nucleotide sequences. *Nature*. 1992;356(6365):168–170.
3. Nee S. Uncorrelated DNA walks. *Nature.* 1992;357(6378):450.
4. Li W, Kaenko K. Long range correlation and partial 1/f spectrum in a non-coding DNA sequences. *Europhys Lett.* 1992;17(7):655–660.
5. Grantham R, Gautier C, Gouy M, et al. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*. 1981;9(1):r43–r74.
6. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol*. 1986;24(1–2):28–38.
7. Karlin S, Campbell AM, Mrázek J. Comparative DNA analysis across diverse genomes. *Annu Rev Genet*. 1998;32:185–225.
8. Sharp PM, Li WH. The codon adaption index- a measure of directional synonymous codon usage bias, and its potential application. *Nucleic Acids Res.* 1987;15(3):1281–1295.
9. Lafay B, Atherton JC, Sharp PM. Absence of translationally selected aynonymous codon usage bias in Helicobacter pylori. *Microbiology.* 2000;146(4):851–860.
10. Dos Reis M, Wernisch L, Savva R. Unexpected correlation between gene expression and codon usage bias from microarry data for the whole E.Coli K-12 genome. *Nucleic Acids Res.* 2003;31(23):6976–6985.
11. Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990;87(1):23–29.
12. Friberg M, von Rohr P, Gonnet G. Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae. *Yeast.* 2004;21(13):1083–1093.
13. Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*. 2001;2(4):RESEARCH0010.
14. Ikumura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Evol Biol.* 1985;2(1):13–34.
15. Roymondal U, Das S, Sahoo S. Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia coli Genome. *DNA Res.* 2009;16(1):13–30.
16. Das S, Roymondal U, Sahoo S. Analyzing gene expression from relative codon usage bias in Yeast genome: A statistical significance and biological relevance. *Gene.* 2009;443(1–2):121–131.
17. Das S, Roymondal U, Chottopadhyay B, et al. Gene expression profile of the cynobacterium synechocystis genome. *Gene.* 2012;497(2):344–352.
18. Sahoo S, Das S. Analyzing gene expression and codon usage bias in diverse genomes using a variety of models. *Curr Bioinform.* 2014;9(2):102–112.
19. Sahoo S, Das S. Analyzing gene expression and codon usage bias in Metallosphaera sedula. *J Bioinform Intell Control.* 2014;3(1):72–80.
20. Das S, Chottopadhyay B, Sahoo S. Comparative analysis of Predicted Gene Expression among Crenarchaeal Genomes. *Genomics Inform*. 2017;15(1):38–47.
21. Coghlan A, Wolfe KH. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerrevisiae. *Yeast*. 2000;16(12):1131–1145.
22. Martin-Galiano AJ, Wells JM, Campa AG. Relationship between codon biased genes, microarry expression values and psychological characteristics of Streptococcus pneumonia. *Micobiology*. 2004;150:2313–2325.
23. Fox JM, Erill I. Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression. *DNA Res.* 2010;17(3):185–196.
24. Lee S, Weon S, Lee S, et al. Relative codon adaptation index,a sensitive measure of codon usage bias. *Evol Bioinform Online*. 2010;6:47–55.
25. Hockenberry AJ, Sirer MI, Amaral LA, et al. Quantifying position-dependent codon usage bias. *Mol Biol Evol.* 2014;31(7):1880–1893.

Citation: Rakshit R, Sahoo S. In *Silico* prediction of gene expression based on codon usage: a mini review. *J Investig Genomics*. 2017;4(2):42–45.
DOI: 10.15406/jig.2017.04.00063