Int. J. Agricult. Stat. Sci., Vol. 9, No. 1, pp. 303-310, 2013 *ISSN : 0973-1903*

STATISTICAL AND COMPUTATIONAL METHODS FOR DETECTION OF SYNONYMOUS CODON USAGE PATTERNS AND GENE EXPRESSION

Md. Samir Farooqi*, R. K. Sanjukta, D. C. Mishra, D. P. Singh[1], Anil Rai, K. K. Chaturvedi, Anil Kumar, Sanjeev Panwar and Naveen Sharma

Centre for Agricultural Bioinformatics, I. A. S. R. I., Library Avenue, Pusa, New Delhi - 110 012, India.
[1]National Bureau of Agriculturally Important Microorganisms, Mau, India.
E-mail: samir@iasri.res.in

Abstract

This paper discusses the various statistical and computational methods applied for detection of synonymous codon usage patterns in prokaryotes and eukaryotes. Synonymous codons (those codons which are encoding the same amino acid) are not equally used by an organism. This phenomenon is called codon usage bias, which exists in a wide range of biological systems. Quantification of codon usage bias helps to understand the evolution patterns of living organisms. The heterogeneous pattern of codon usage within the prokaryotes and eukaryotes genome can be explained by a balance between genome-wide mutational biases and selection against these mutations. Advanced statistical and computational techniques are being applied for better interpretation of codon usage pattern in any individual gene sequence or within and across the genome sequences of organisms, codon usage indices and various multivariate statistical analysis techniques are used for prediction of functional and structural characteristics of coding sequences, phylogenetic analysis, codons and amino acids usage patterns among genes.

Key words : Codon usage pattern, GC content, Gene expression, Synonymous codon, Codon bias.

1. Introduction

A codon is the three-unit sequence (UUA, AGC, etc.) of mRNA nucleotides *i.e.* Adenosine (A), Guanine (G), Thymine (T), Cytosine (C) that codes for a specific amino acid. Since, there are only twenty commonly used amino acids and sixty-four ($4 \times 4 \times 4$) possible codon sequences, the genetic code is described as both degenerate and unambiguous. Each codon codes for only one amino acid, but each amino acid may have more than one matching codon. It is well known that these 64 codons from four nucleotides codes performing to 20 amino acids are responsible for synthesis of proteins during translation process of proteins [Hassan *et al.* (2009)]. Apart from this, one family consists of start and stop codons. The availability of alternate codons for translation into same amino acids is known as

**Corresponding author.* *Received Sept. 12, 2012* *Revised Feb. 28, 2013* *Accepted May 14, 2013*

synonymous codons. There are five families of synonymous codons and it has been observed that an organism does not use synonymous codons randomly for synthesis of proteins. This unequal usage of synonymous codons by an organism is called codon usage bias, which exists in a wide range of biological systems. Different factors are responsible for codon usage biasness in an organism, these factors form a pattern called synonymous codon usage pattern, which can explains the causes of variation present in the genes. Some of the factors that contribute to the codon usage variations are :

(i) Natural selection (ii) Mutational bias (iii) Translational efficiency, accuracy and (iv) gene expression.

In case, codon usage bias varies across species as well as among genes within species indicates the role of natural selection [Sharp *et al.* (1988)]. Almost all organisms are subject to directional mutational presence and in the absence of selection; it is this presence that shapes gene codon usage pattern. There is often mutational selection balance in operation that shapes the overall frequency with each codon is used [Sharp and Matassi (1994)]. Codons that are recognised by abundance of t-RNA in the cell are translated 3-6 fold faster than their synonymous counterparts. This is called translational efficiency of those codons. It is also well known that non optimal codons (codons whose frequency is low in highly expressed genes) are mistranslated 8-10 times more often than its optimal synonymous [Nomura *et al.* (1987) and Sharp (1990)]. Finally highly expressed genes in any organisms or crops have on average a more extreme codon bias [Sharp and Li (1986)].

Advanced statistical and computational methods are being applied for identification of factors responsible for codon usage statistics and for better interpretation of codon usage pattern in any individual organism or within and across different genome sequences of organisms. Gene sequence features such as codon bias, codon context, codon expansion (*e.g.* trinucleotide repeats) can be better understood at the genomic scale by combining statistical methodologies with advanced computer algorithms and data visualization techniques. This paper discusses various statistical and computational methods applied for analysis of synonymous codon usage patterns in an organism. In this study, data related to genes regulating for cation uptake and iron carrying compounds of *Salinibacter ruber* (*S. ruber*) DSM 13855 were analysed and used as an example to describe the tools and techniques employed in codon usage analysis. *S. ruber* is an extremely halophilic bacteria, phylum *Bacteroidetes*. It was first isolated from the salt-saturated environment of saltern crystallizers in Spain [Anton *et al.* (2002)]. *S. ruber* is a brightly red-coloured, aerobic, motile, straight or slightly curved, gram-negative rod. It is one of the most important halophilic species of bacteria which requires minimally 150 g/l salt for growth, grows optimally between 200–300 g/l salt and tolerates salt concentrations upto saturation. The optimum temperature for growth is 35–45°C [Anton *et al.* (2002)].

2. Measures of Codon usage Bias

Codon usage indices are useful measures for the tabulation and investigation of codon usage in an organism. Indices reduce the codon usage data into a useful summary. There

are two basic categories of codon usage indices. First category measures the overall deviation of codon usage from some expected usage and second measures a bias towards a particular subset of optimal codons (codons, whose frequency is high in highly expressed genes). In many species, highly expressed genes preferentially use a subset of codons (*i.e.* optimum codons) [Liu *et al.* (2010)]. Several indices estimate extent to which the codon usage of a gene has been altered towards preferential usage of these optimal codons. Some of the measures for codon usage bias are as follows :

(a) Effective number of codons (N_c) : It is used to examine the relationship between codon bias and mutational bias. It measures the bias with respect to equal usage of codons within synonymous groups [Wright (1990)]. Further, it also provides information about gene expression. Gene sequences in which N_c values <30 are highly expressed while those with N_C value >55 are poorly expressed genes [Hassan *et al.* (2009)]. It can take values from 20 to 61, when only one codon or all synonyms codons with equal frequencies are used for translation to a particular amino acid, respectively.

$$N_c = 2 + s + [29/\{s^2 + (1-s)^2\}]; \text{ where, } s = GC_{3s}$$

(b) Relative Synonymous Codon Usage (RSCU) : RSCU is defined as the ratio of observed frequency of codon usage to expected frequency of codon usage; if all the synonymous codons for those amino acids are used equally [Sharp *et al.* (1986)], RSCU is used to observe the synonymous codon usage variation among the genes. RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency, whereas, the reverse is true for RSCU value less than 1.0.

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_1} X_{ij}}$$

Where, X_{ij} is the number of occurrences of the jth codon, which can be translated into ith amino acid and n_i is the number (from 1 to 6) of alternative codons for translation into ith amino acid.

(c) Frequency of optimal codons (F_{op}) : It is a simple ratio between frequency of optimal codons and total number of synonymous codons. Its values range from 0 (when a gene contains no optimal codons) to 1 (when a gene is entirely composed of optimal codons) [Ikemura (1981), Ikemura and Ozeki (1982), Ikemura (1985)].

$$F_{op} = \frac{N_{\text{optimal codons}}}{N_{\text{synonymous codons}}}$$

(d) Codon Bias Index (CBI) : Codon bias index (CBI) is a measure of directional codon bias towards a subset of optimal codons [Bennetzen and Hall (1982)]. CBI is similar to F_{op} except that N_{ran} is used as a scaling factor.

in codon usage through multidimensional hyperspace that account for the largest fractions of the variation among genes. Correspondence analysis can be carried out on any standardized indices, but in this article it came on RSCU values and simple codon counts in order to investigate the codon usage variation among the genes. Only the distributions of the genes along the first two major axes, which accounts for maximum variation, are considered. A scatter graph is plotted in a two dimensional space to see the position or classification of the genes on the basis of these major explanatory axis. Further, CAI index can be used to classify genes based on their expression level. Table 2, below depicts the percentage codon usage variation explained by each axis whereas Figs. 2 and 3 show the scatter plots of COA scores for axis 1 and axis 2 for 66 genes responsible for cation uptake and iron carrying compounds of *S. ruber* based on codon count and RSCU values, respectively. It can be observed that, it is possible to differentiate the genes through plotting *i.e.* axis 1 and axis 2 of correspondence analysis but pattern of distribution varies among organism and expression level of different gene under a particular environment. In case of RSCU values, genes are distributed around origin and there is a hardly any correlation between axis 1 and axis 2.

(d) Other Statistical techniques : Besides MVA, various other statistical tools and techniques such as cluster analysis, exploratory data analysis, correlation analysis, chi-square analysis etc. have been applied to identify codon usage patterns within and across genomes of different species. Cluster analysis based on RSCU values of codons was used to arrange

Table 2 : Percentage of codon usage variation explained by the axes.

Explanatory Axis determined by COA	% variation based on RSCU	% variation based on codon usage
Axis 1	20.50	15.32
Axis 2	9.04	10.67
Axis 3	7.65	9.46
Axis 4	6.09	5.35

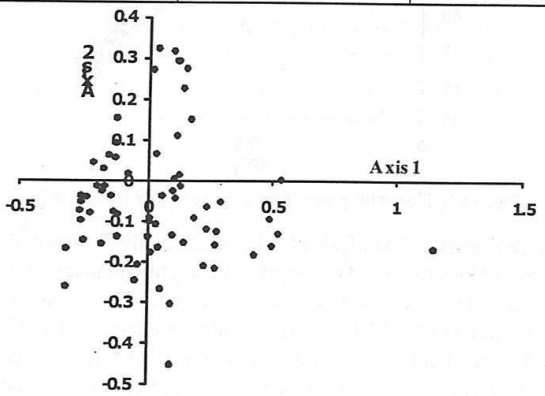

Fig. 2 : Scatter plots of Axis 1 and Axis 2 of correspondence analysis on codon count value.

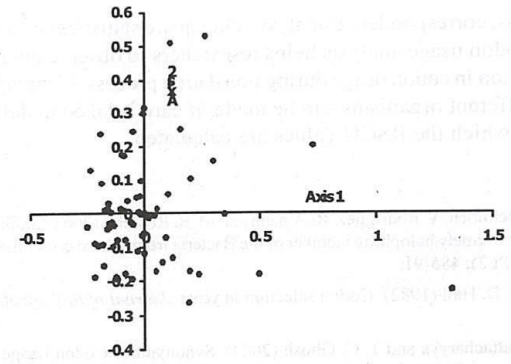

Fig. 3 : Scatter plots of Axis 1 and Axis 2 of correspondence analysis on RSCU value.

genes according to similarity pattern of gene expression. Correlation analysis has been widely applied in identifying the gene expression pattern and the direction of codon bias. Chi-square based analysis compares observed codon frequencies to those expected, if codon usage reflects only local base composition at synonymous sites. This method has been useful in distinguishing real trends in codon usage from random fluctuations. Chi-square based analysis is accurate and sensitive. Exploratory data analysis also plays an important role in the analysis of codon usage patterns in organisms. Various plots such as scatter plot, histogram etc. are being plotted between different indices to establish relationship and the direction of codon bias. Table 3 below shows high correlation values of GC_{3s}, GC and N_c with Axis 1 in genes of cation uptake and iron carrying compounds of *S. ruber*. The genes positions on axis 1 were strongly negatively correlated with GC_{3s} and GC content of genes, and significantly positively correlated with N_c. While genes positions on axis 1 were not significantly correlated with CAI values. These values indicate their role in the variation of codon usage GC and N_c are playing important role in determining the codon usage pattern in the functional group of this organism.

Table 3 : Correlation GC_{3s}, GC, N_c and CAI with Axis 1.

	GC_{3s}	GC	N_c	CAI
Axis 1	-.96*	-.63*	.92*	.093

4. Conclusion

Various factors such as natural selection, t-RNA abundance, translational accuracy, mutational bias, selection of optimal codons etc. contribute to the codon usage bias in prokaryotes and eukaryotes. These factors vary for different genomes and depend on genome composition and biology. GC composition, GC (AT) skew and translational selection play an important role in shaping codons bias in the organisms. Mutational bias is mostly shaped by very high and very low GC content, whereas, the translation selection may depend on genome biology. Various computational and statistical methods like correlation analysis, principal

component analysis, correspondence analysis, chi-square statistics etc. are applied for codon usage analysis. Codon usage analysis helps researchers to observe inter specific as well as intraspecific variation in codon usage during translation process. Meaningful comparisons of codon usage in different organisms can be made, if care is taken in defining the reference set of genes from which the RSCU values are calculated.

References

Anton, J., A. Oren, S. Benlloch, V. Rodriguez, R. Amann and M. R. Rossello (2002). *Salinibacter ruber* gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds. *Int J Syst Evol Microbiol*, **3**, 52 (Pt 2): 485-91.

Bennetzen, J. L. and B. D. Hall (1982). Codon selection in yeast. *Journal of Biological Chemistry*, **257**, 3026-3031.

Gupta, S. K., T. K. Bhattacharyya and T. C. Ghosh (2004). Synonymous Codon Usage in *Lactococcuslactis* : Mutational Bias versus Translational Selection. *Journal of Biomolecular Structure & Dynamics*, **4**, 1-9.

Hassan, S., V. Mahalingam and V. Kumar (2009). Synonymous Codon Usage Analysis of Thirty Two Mycobacteriophage Genomes. *Advances in Bioinformatics*, 316-936, 11 pp.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons of its protein genes. *Journal of Molecular Biology*, **146**, 1-21.

Ikemura, T. and H. Ozeki (1982). Codon usage and transfer RNA contents: organism specific codon choice patterns in reference to the iso-acceptor contents. *Cold Spring Harbor Symposium Quantitative Biology*, **47**, 1087-1097.

Ikemura, T. (1985). Codon usage and t-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13-34.

Liu, H., R. He, H. Zhang, Y. Huang, M. Tian and J. Zhang (2010). Analysis of synonymous codon usage in *Zea mays*. *Mol Biol Rep*, **37**, 677–684.

Nomura, M., F. Sor, M. Yamagashi and M. Lawson (1987). Heterogeneity of GC content within a single bacterium and its implications for evolution. *Cold Spring Harbor Symposium Quantitative Biology*, **52**, 658-663.

Peden, J. F. (1999). Analysis of Codon Usage *Ph.D. Thesis*, University of Nottingham.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe and F. Wright (1988). Codon usage in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizo saccharomycespombe*, *Drosophila melanogaster* and *Homo sapiens* : a review of the considerable within-species diversity. *Nucleic Acids Research*, **16**, 8207–8211.

Sharp, P. M. and G. Matassi (1994). Codon usage and genome evolution. *Current opinion in genetics development*, **4**, 851-860.

Sharp, P. M. (1990). Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. *Molecular Microbiology*, **4**, 119-122.

Sharp, P. M. and W. H. Li (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, **24**, 28–38.

Sharp, P. M., T. M. F. Tuohy and K. R. Mosurski (1986). Codon usage in yeast : Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, **14**, 5125-5143.

Wright, F. (1990). The effective number of codons used in a gene. *Gene*, **87**, 23-29.