Recent Research in Science and Technology 2013, 5(2): 24-28
ISSN: 2076-5061
Available Online: http://recent-science.com/

Comparative codon usage pattern analysis of *Eubacterium eligens, E. limosum* and *E. rectal*

Kishor Shende*[1], Anita Mishra[1], Ragini Gothalwal[1] and Mudasir Iqbal[2]

[1]Bioinformatics Center, Dept. of Biotechnology, Barkatullah University, Bhopal- 462026 (M.P.) India
[2]Rajeev Gandhi College, Bhopal 462026 (M.P.) India

Abstract

Analysis of codon usages data has both practical and theoretical importance in understanding the basics of molecular biology. Current study was aimed to study the codon usage preferences pattern and biases in three Eubacterium species viz. *E. eligens, E. limosum* and *E. rectal*. *Eubacterium* is the second most common genus after the genus *Bacteroides* in human intestinal tract, which are anaerobic, non-spore forming, and gram-positive rods. ORF sequence file were obtained from NCBI ftp site and analyzed by CodonW. *E. eligens* and *E. rectal* are AT-rich consisting of 62.04% and 57.82% AT-content. *E. limosum* has equal content of GC (48.62%) and AT (51.38%). In *E. rectale* and *E.* eligens showed maximum preferences for 'A' or 'T' ending codons whereas in *E. limosum*, 'A' or 'T' and 'G' or 'C' ending codons are preferred equally. In case of *E. limosum* the 'GC' drifts has influenced the codon usage pattern in this bacteria. All three *Eubacterium sp.* has shown low level codon usage heterogenesity among the genes. Nc-GC3 plot indicated major influence of compositional constraints in *Eubacteria eligens* and *E. rectal*. Beside E. rectal has additional factor influencing the codon usage. In *E. limosum* translational selection and additional factors are dominant over compositional constraints. The study indicates the close relationship between *E. eligens* and *E. rectal* and *E. limosum* diverging from them due to GC drift.

Keywords: Codon usage bias, *Eubacterium*, compositional bias, translational selection, Nc-GC3

INTRODUCTION

In general, the term "genetic code" refers to the sequences of nucleotides in DNA and RNA that determine the various amino acid sequences of proteins. There are 64 possible codons, three are stop codons (viz: TGA, TAA, TAG as per Universal Codon Table), which do not code for amino acids but instead indicate the end of a translation process. The remaining 61 codons specify the 20 amino acids that make up proteins. Analysis of codon usages data has both practical and theoretical importance in understanding the basics of molecular biology [Moriyama and Hartl, 1993]. Different organisms often show particular preferences for one of the synonymous codon that encode the same amino acid which is known as codon usage bias. It is well known that synonymous codon usage bias is non-random and species specific [Grantham *et al.* 1981]. Each species show a specific pattern of codon usages [Grantham *et al.* 1980a, Grantham *et al.* 1980b]. Moreover codon usage patterns differ significantly among different genes within the same taxa [Wada *et al.*, 1992]. The codon usage patterns differ significantly depending upon several factors such as compositional mutational bias, natural selection for translation optimization. Analysis of codon usage patterns in an organism helps in understanding the basis of molecular biology of gene regulation and gene expression. This can indirectly help in understanding the pathogenicity of bacterium.

The human colonic microbiota consists of at least 500 bacterial species [Eckburg, *et al.* 2005] and plays an important role in maintaining human health by preventing colonization by pathogens, degrading dietary and in situ-produced compounds, producing nutrients, and shaping and maintaining the normal mucosal immunity [Hooper, 2004]. The highest proportion of human fecal organisms detected fell within the *Clostridium coccoides-Eubacterium rectale* group (7.2 3 1010 cells/g dry weight of feces), which forms part of clostridial cluster [Franks, *et al.* 1998]. The genus *Eubacterium* contains anaerobic, non-spore forming, gram-positive rods which are distinguished from other genera mainly on the basis of negative metabolic characteristics [Moore, *et al.*, 1986]. *Eubacterium limosum* is industrially important in conversion of CO to multi-carbon compound. Therefore *E. limosum* has been considered for a model strain for bio energy production from sugars (obtained from biomass) [Chang *et al.*, 2007]. E. eligens and E.rectale are capable of fermentation of substrate like polygalacturonate, amylopectin and L-fucose [Salyers, et al., 1977].

Eubacterium sp are pathogenic but on other side provide benefits to human by colonizing in intestinal tract as part of normal microbial flora. Genome sequences are available for these three species at NCBI (National Center for Biotechnology Information). The aim of the present study was to perform the codon usage preference analysis among three species as, Eubacterium eligen, E. rectal and E. limosum.

METHODS AND MATERIAL

The ORFs sequences file of the genomes of Eubacterium eligens ATCC, E. rectale ATCC 33656 and *E. limosum* KIST612

*Corresponding Author

Kishor Shende
Bioinformatics Center, Dept. of Biotechnology, Barkatullah University, Bhopal-462026 (M.P.) India

Email: kishor556@hotmail.com

[Mahowald, *et al.*, 2009] were retrieved from the ftp site of NCBI. Codon usage analysis were performed using software CodonW **(**http://codonw.sourceforge.net**)** [Peden, 1994], compositional analysis using in-house developed C-program. Statistical analysis was performed using PAST [Hammer, 2001] and Systat-13 (SigmaPlot Corp.) software.

RESULTS AND DISCUSSIONS

Table-1 shows general information of *Eubacterium* species. *E. limosum* is the largest genome having size 4.3 mbp. *E. rectale* and *E. eligens* are having 3.45 mbp and 2.83 mbp genome size. *E. limosum* is marginally equal in GC (48.62%) and AT (51.38%) content. *E. rectale* and *E. eligens* are AT-rich with 57.82% and 62.04% AT-content respectively. All three species are pathogenic (host specific habitat), non-motile, facultative anaerobe and mesophillic

Table 1. General features of *Eubacterium* species

	***E. eligens* ATCC 27750**	***E. limosum* KIST612**	***E. rectale* ATTC 33656**
RefSeq ID	NC_012778	NC_014624	NC_012781
Size	2.83419 mbp	4.3 mbp	3.44969 mbp
GC content of genome	37.6%	48.62%	42.15%

Nucleotide composition of ORFs

Table-2 shows base composition of complete ORFs of all three *Eubacterium* species. *E. eligens* and *E. rectale* are consisting of 62.04% and 57.82% AT-content respectively, whereas *E. limosum* is consisting of marginally equal composition of AT (51.38%) and GC (48.62%) content. *E. limosum* is comparatively richer in GC-content than *E. rectale* and *E. eligens.*

Table 2. Base compositions of ORFs of Eubacterium species

Organisms	**A**	**T**	**G**	**C**	**Total**	**AT content (%)**	**GC content (%)**
E. eligens	652920	559663	450710	291294	1954587	**62.04**	**37.96**
E. limosum	1055584	931780	999762	880708	3867834	**51.38**	**48.62**
E. rectale	990896	809893	763051	550580	3114420	**57.82**	**42.18**

Table 3. Relative Synonymous Codon Usage (RSCU) Values of three *Eubacterium* spp.

Codons	***E. rectale***	***E. elegens***	***E. limosum***		***Codons***	***E. rectale***	***E. elegens***	***E. limosum***
GCG (Ala)	0.56	0.37	0.88		CCG (Pro)	**1.50**	0.88	**1.38**
GCA (Ala)	**1.92**	**2.08**	0.81		CCA (Pro)	1.11	**1.51**	0.82
GCT (Ala)	0.88	1.19	0.72		CCT (Pro)	1.17	**1.53**	0.67
GCC (Ala)	0.64	0.36	**1.59**		CCC (Pro)	0.22	0.08	1.13
TGT (Cys)	**1.07**	**1.31**	0.82		CAG (Gln)	**1.62**	**1.79**	**1.66**
TGC(Cys)	0.93	0.69	**1.18**		CAA (Gln)	0.38	0.21	0.34
GAT (Asp)	**1.37**	**1.52**	**1.01**		AGG (Arg)	1.12	0.90	0.66
GAC (Asp)	0.63	0.48	0.99		AGA (Arg)	**2.36**	**3.73**	0.93
GAG-Glu	**1.15**	0.82	0.88		CGG (Arg)	0.35	0.16	**1.44**
GAA-Glu	0.85	**1.18**	**1.12**		CGA (Arg)	0.37	0.23	0.31
TTT (Phe)	**1.39**	**1.34**	**1.41**		CGT (Arg)	1.07	0.73	1.03
TTC (Phe)	0.61	0.66	0.59		CGC (Arg)	0.72	0.26	**1.64**
GGG (Gly)	0.31	0.25	0.62		AGT (Ser)	0.96	1.16	0.68
GGA (Gly)	**1.63**	**1.77**	0.78		AGC (Ser)	1.23	0.92	**1.93**
GGT (Gly)	0.99	1.32	0.71		TCG (Ser)	0.59	0.34	0.52
GGC (Gly)	1.07	0.65	**1.89**		TCA (Ser)	**1.74**	**2.05**	0.75
CAT (His)	**1.32**	**1.55**	**1.07**		TCT (Ser)	0.90	1.28	0.73
CAC (His)	0.68	0.45	0.93		TCC (Ser)	0.58	0.26	1.39
ATA (Ile)	**1.08**	**1.13**	0.26		ACG (Thr)	0.49	0.30	0.59
ATT (Ile)	1.08	**1.43**	**1.34**		ACA (Thr)	**2.12**	**2.36**	0.98
ATC (Ile)	0.83	0.45	1.40		ACT (Thr)	0.68	1.09	0.47
AAG (Lys)	**1.10**	**1.19**	0.87		ACC (Thr)	0.71	0.26	**1.97**
AAA (Lys)	0.90	0.81	**1.13**		GTG (Val)	1.14	0.58	**1.34**

TTG (Leu)	0.69	0.48	0.54		GTA (Val)	**1.25**	1.30	0.62
TTA (Leu)	0.95	1.28	0.70		GTT (Val)	1.07	1.83	1.00
CTG (Leu)	1.31	0.77	**2.79**		GTC (Val)	0.54	0.29	1.03
CTA (Leu)	0.25	0.14	0.12		TGG (Trp)	1.00	1.00	1.00
CTT (Leu)	**2.09**	**3.07**	1.00		TAT (Tyr)	**2.96**	**3.41**	**2.05**
CTC (Leu)	0.71	0.26	0.85		TAC (Tyr)	1.41	0.97	1.69
ATG (Met)	1.00	1.00	1.00					
AAT (Asn)	**1.32**	**1.50**	0.97					
AAC (Asn)	0.68	0.50	**1.03**					

Table 4. Maximally preferred codons

Amino Acid	***E. rectale***	***E. eligens***	***E. limosum***
Ala	GCA	GCA	GCC
Cys	TGT	TGT	TGC
Asp	GAT	GAT	GAT,GAC
Glu	GAG	GAA	GAA
Phe	TTT	TTT	TTT
Gly	GGA	GGA	GGC
His	CAT	CAT	CAT,CAC
Ile	ATA, ATT	ATT, ATA	ATC,ATT
Lys	AAG	AAG	AAA
Leu	CTT	CTT	CTG
Asn	AAT	AAT	AAC,AAT
Pro	CCG	CCA	CCG
Gln	CAG	CAG	CAG
Arg	AGA	AGA	CGC
Ser	TCA	TCA	AGC
Thr	ACA	ACA	ACC
Val	GTA	GTA	GTG
Tyr	TAT	TAT	TAT

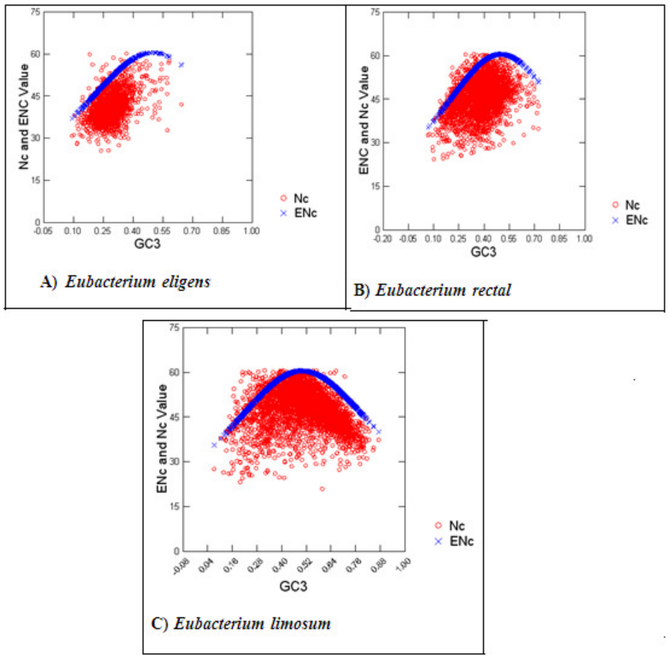
Measure of Synonymous Codon Usage Bias

Codon usage values within the datasets of differing amino acid compositions were normalized by Relative Synonymous Codon Usage (RSCU) value which is a ratio of observed codon usage to the average codon usage for that amino acid. RSCU removes the influence of amino acid composition that is present in raw codon usage data [Sharp and Li, 1986]. Codon usage frequency and RSCU values were calculated by CodonW software and tabulated (Table-3). Table-4 lists the maximally preferred codons to their synonymous codons coding for specific amino acids. Table-3 indicates that, 'A' or 'T' ending codons are maximally preferred in case of *E. eligens* and *E. rectal* for the amino acids as, Ala, cys, Asp, Phe, Gly, His, Ile, Leu, Asn, Arg, Ser, Thr, Val, Tyr. 'G' or 'C' ending codons are preferred maximally for the amino acids as, Glu, Lys, Gln. This dominance is high in case of *E. eligens* as it is comparatively more 'AT' rich than *E. rectale*. In *E. limosum,* 'G' or 'C' ending codons are maximally preferred for the amino acids as, Ala, Asp, Gly, Leu, Pro, Gln, Arg, Ser, Thr, Val and 'A' or 'T' ending codons for the amino acids as, Glu, His, Ile, Lys, Asn, Tyr. The dominance of 'G' or 'C' ending codons in *E. limosum* is due to 'GC' drift that has influenced the codon usage pattern. This indicates the compositional constraints as major factor in shaping the codon usage bias. Frequency of tRNA copies greatly affects the translation process in living system. Maximum number of isoacceptor tRNA copies of synonymous codons will increase the rate of translation process. *E. eligens* has 47 copies of tRNA where as *E. limosum* and *E. rectal* has 57 numbers of copies of tRNA in each. Low number of copies in *E. eligens* may be one factor supporting the fact that translational selection is one factor in shaping the codon usage bias in *E. eligens*.

NC-GC3 Plot

Effective number of codons provides a way to quantify how different the codon usage of a particular gene is from the equal use of synonymous codons. Nc is an estimate of the strength of general codon usage bias and might be influenced by mutation biases and/or selection for particular codons [Wright, 1990]. The GC3 value is fraction of codons which are synonymous at the third codon position and have either a G or a C at that position. Enc value is expected Effective Number of codons (Nc). Figure-1 (A), (B) and (C) are the Nc and ENc plot against the GC3 values in *E. eligens, E. rectale* and *E. limosum* respectively. *Eubacteria eligens* (figure-1 A) shows all the points lying beneath the expected curve, with low GC3 value

(0.10 to 0.5). *E. eligens* is AT-rich bacteria hence low GC3 value. The plot indicate the major influence of compositional constraints on codon usage bias and other factors also has influence but to less extent. *Eubacterium rectale* in figure-1(B) show that maximum points are lying below and beneath the curve. Very few points are on the expected curve. GC3 value ranges from 1 to 0.7, but maximum points are in the range of 0.2 to 0.5 indicating compositional bias. *E. rectale* is AT-rich bacterium. Very few points are well below the expected curve indicating additional factors influencing the codon usage bias. But still the compositional constraint is important factor in shaping the codon usage. *Eubacterium limosum* is having nearly equal composition of AT and GC. Maximum points are lying well below the expected curve (figure-1 C). Less number of points are lying beneath the expected curve and very few on the curve. This indicates that translational selection is dominant factor over compositional constraints in shaping codon usage. Influence of translational selection was also observed in *Burkholderia mallei* relative to compositional bias [Zhao *et al.* 2007]. Correlation between codon usage bias and gene expression was studied and observed the strong influence of natural selection at translational level in *C. elegans* [Stenico *et al*, 1994], *Sinorhizobium meliloti* [Peixoto *et al.*, 2003], *Cyprinidae* [Romero *et al.*, 2003], *thermophilic prokaryotes* [Singer and Hickey, 2003], *D. melanogaster,* and *A. thaliana* [Duret and Mouchirroud, 1999].



A) *Eubacterium eligens*

B) *Eubacterium rectal*

C) *Eubacterium limosum*

Fig 1. Nc and GC3 Plot

CONCLUSION

Directional mutation pressure on DNA sequences and natural selection affecting gene translation are the two major factors that have been widely accepted to account for both interspecific codon usage variation and intragenomic codon usage variability [Zhao *et al.*, 2007]. In unicellular organisms, such as *Escherichiacoli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, and *Dictyostelium discoideum*, the codon usage is attributable to the equilibrium between natural selection and compositional mutation bias [Sharp *et al.,* 1993]. The ORFs *E. eligens* and *E. rectal* are richer in AT-content than *E. limosum*. Therefore preferences for 'A' or 'T' ending codons is dominant in *E. eligens* and *E. rectal* and equal preferences of 'A' or 'T' and 'G' or 'C' ending codons in *E. limosum*. The 'GC' drift in *E. limosum* has influenced the codon usage bias indicating compositional constraints as major factor in shaping the codon usage. *Eubacterium* eligens has low level of variation in codon usage and less percent of heterogenesity. *Eubacterium rectal* has less heterogeinity and codon usage variation among the genes. But it shows high level of variation among the highly expressed genes. Even *Eubacterium limosum* also has less amount of codon usage variation. In *Eubacteria eligens* and *Eubacterium rectal,* compositional constraint is a major influencing factor to codon usage bias and other factors have less influence. *Eubacterium limosum* it is translational selection is major factor as compared to compositional constraints in shaping codon usage. This closely related species are showing great variation in codon usage pattern indicating high level of genetic compositional drift.

REFERENCES

[1] Mahowald M.A., Rey F.E., Seedorf H., Turnbaugh P.J., Fulton R.S., Wollam A., Shah N., Wang C., Magrini V., Wilson R.K., Cantarel B.L., Coutinho P.M., Henrissat B., Crock L.W., Russell A., Verberkmoes N.C., Hettich R.L. and Gordon J.I. 2009. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc. Natl. Acad. Sci. U.S.A. 106 (14):5859-5864.

[2] Moriyoma E.N., Hartle D.L. 1993. Codon usage bias and base composition of nuclear genes in drosophila, *Genetics;* 134:847-858.

[3] Grantham R.C., Gautier, Gouy M., Jacobzone M., Mercier R. 1981. Codon catalogue usage is a genome strategy for genome expressivity. *Nucleic Acids Research.* 9: r43-r75.

[4] Grantham R.C., Gautier, Gouy M. 1980a. Codon frequencies in 119 genes confirm consistent choices of degenerate base according to genome type. *Nucleic Acids Research*. 8:1892-1912.

[5] Grantham R. 1974. Amino acid difference formula helps to explain protein evolution. *Science*. 185:862-864

[6] Grantham R.C., Gautier, Gouy M., Mercier R., Pave A. 1980b. Codon catalogue usage and the genome hypothesis. *Nucleic Acids Research*. 8: r49-r62

[7] Wada K.N., Wada Y., Ishibashi F., Gojobori T., Ikemura T. 1992. Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Research.* 20:2111-2118.

[8] Hanseong R., Hyeokh J.K., Dachee K., Geon C.D., Shinyoung P., Sujin K., Geon C.I. 2011. Complete Genome Sequence of a Carbon Monoxide-Utilizing Acetogen, Eubacterium limosum KIST612.

[9] Franks A.H., Harmsen H.J.M., Raangs G.C., Jansen G.J., Welling G.W. 1998. Variations of bacterial populations in human feces quantified by fluorescent in situ hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. *Appl. Environ. Microbiol.* 64:3336–3345.

[10] Moore W.E.C, Holdeman L.V. Genus Eubacterium Prévot 1938. In: Sneath P H A, Mair N S, Sharpe M E, Holt J G, editors. Bergey's manual of systematic bacteriology. Vol. 2. Baltimore, Md: The Williams & Wilkins Co.; 1986. pp. 1353–1373.

[11] Chang I.S., Kim D., Kim B.H., Lovitt R.W. 2007. Use of an industrial grade medium and medium enhancing effects on high cell density CO fermentation by Eubacterium limosum KIST612. *Biotechnol Lett* (2007) 29:1183–1187. DOI 10.1007/s10529-007-9382-x

[12] Eckburg P.B., Bik E.M., Bernstein C.N., Purdom E., Dethlefsen L., Sargent M., Gill S.R., Nelson K.E., Relman D.A. 2005. Diversity of the human intestinal microbial flora. *Science* 308:1635–1638.

[13] Hooper L.V. 2004. Bacterial contributions to mammalian gut development. *Trends Microbiol.* 12:129–134.

[14] Salyers A.A., West S.E., Vercellotti J.R., Wilkins T.D. 1977. Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Appl. Environ. Microbiol*. 34(5):529.

[15] Sharp P.M. and Li W.H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol.*24, 28-38. 199

[16] Peden J. 1999. Analysis of codon usage, PhD Thesis. University of Nottingham.

[17] Hammer Ø., Harper D.A.T., Ryan P.D. 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica.* 4(1): 9pp.

[18] Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* 87, 23-29. 201

[19] Zhao S., Zhang Q., Chen Z., Zhao Y., Zhong J. 2007. The Factors Shaping Synonymous Codon Usage in the Genome of Burkholderia mallei *Journal of Genetics and Genomics.* 34(4): 362-372

[20] Stenico M., Lloyd A.T., Sharp P.M.1994. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 22(13): 2437-2446.

[21] Peixoto L., Zavala A., Romero H., Musto H. 2003. The strength of translational selection for codon usage varies in the three replicons of Sinorhizobium meliloti. *Gene.* 3(20): 109-116.

[22] Romero H., Zavala A., Musto H., Bernardi G. 2003. The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene.* 317(1-2): 141-147.

[23] Singer G.A., Hickey D.A. 2003. Thermophilic prokaryotes have characteristic characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene.* 317(1-2): 39-47.

[24] Duret L., Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA. 96(8): 4482-4487

[25] Sharp P.M., Stenico M., Peden J.F., Lloyd A.T. 1993. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans.* 21(4): 835-841.

[26] http://codonw.sourceforge.net

[27] http://www.ncbi.nlm.nih.gov