Indian Journal of Medical Microbiology, (2008) 26(4): 313-21

Special Article

COMPARATIVE ANALYSIS OF CODON USAGE PATTERNS AND IDENTIFICATION OF PREDICTED HIGHLY EXPRESSED GENES IN FIVE *SALMONELLA* GENOMES

UK Mondal, S Sur, AK Bothra, *A Sen

Abstract

Purpose: To anlyse codon usage patterns of five complete genomes of *Salmonella*, predict highly expressed genes, examine horizontally transferred pathogenicity-related genes to detect their presence in the strains, and scrutinize the nature of highly expressed genes to infer upon their lifestyle. **Methods:** Protein coding genes, ribosomal protein genes, and pathogenicity-related genes were analysed with Codon W and CAI (codon adaptation index) Calculator. **Results:** Translational efficiency plays a role in codon usage variation in *Salmonella* genes. Low bias was noticed in most of the genes. GC3 (guanine cytosine at third position) composition does not influence codon usage variation in the genes of these *Salmonella* strains. Among the cluster of orthologous groups (COGs), translation, ribosomal structure biogenesis [J], and energy production and conversion [C] contained the highest number of potentially highly expressed (PHX) genes. Correspondence analysis reveals the conserved nature of the genes. Highly expressed genes were detected. **Conclusions:** Selection for translational efficiency is the major source of variation of codon usage in the genes of *Salmonella*. Evolution of pathogenicity-related genes as a unit suggests their ability to infect and exist as a pathogen. Presence of a lot of PHX genes in the information and storage-processing category of COGs indicated their lifestyle and revealed that they were not subjected to genome reduction.

Keywords: *Codon bias and expression, pathogenicity, Salmonella*

Food-borne disease has been defined by the World Health Organization (WHO) as an ailment of transmittable or toxic nature caused by, or thought to be caused by, the consumption of food or water.[1] A number of bacteria are known to be linked with food-borne diseases. Prominent amongst them are *Salmonella, Shigella*, *Listeria, Staphylococcus, Vibrio,* etc. *Salmonella* is a gram-negative, motile, rod-shaped bacterial pathogen extensively occurring in animals, primarily in poultry and swine. Environmental sources of the bacterium throughout the world include water, soil, insects, factory surfaces, kitchen surfaces, animal faeces, raw meats, raw poultry, and raw sea foods.[2] *Salmonella* causes substantial morbidity and mortality globally. The human-adapted serovars are responsible for typhoid, a systemic and life-threatening disease; whereas non-human-adapted serovars are normally accountable for gastroenteritis.[3]

The infection machinery of *Salmonella* involves a number of bacterial virulence genes, many of which are liable for invading, surviving, and replicating within host cells.[4] Recent work has exposed that a sizeable portion of *Salmonella typhimurium* genes are positioned in distinct chromosomal regions called pathogenicity islands.[4,5] The pathogenicity islands enclose genes associated with diseases and are often sources of toxins. Their G+C content differs from the rest of the chromosome, signifying that horizontal gene transfer acquired them.[6] Besides these, *vir* genes, *hrp* genes, invasions, *pip* genes, SPI, SOP, and toxin genes are also associated with pathogenicity.

Like other branches of biology, the study of pathogenic microorganisms has undergone a paradigm shift. The incredible deluge of information from genome-sequencing projects is revolutionizing the science of bacterial pathogenicity. The accessibility of the complete genome sequences of *Salmonella* provides a scope to undertake bioinformatics-based approaches focusing on synonymous codon usage and investigating the gene expression profile of the organism.

The non-random usages of synonymous codons are well accredited.[7] Synonymous codon usage is species specific and differs appreciably between the genes in the same organism.[8] Mutational pressure and natural selection operating at the level of translation are the primary reasons behind codon usage variation among the genes in different organisms.[9] Codon bias is quite high in the highly expressed genes compared to lowly expressed ones inside a genome.[10-13] The bias of highly expressed genes is influenced by translational selection; in contrast to lowly expressed genes, which is governed by mutational bias.[8]

*Corresponding author: (email: <senarnab_nbu@hotmail.com>)
Chemiinformatics Bioinformatics Laboratory (UKM, AKB), Department of Chemistry, Raiganj University College, Raiganj - 733 134, West Bengal, India; and NBU Bioinformatics Facility (SS, AS), University of North Bengal, Siliguri - 734 013, West Bengal, India
Received: 26-05-08
Accepted: 15-08-08

In order to inspect the patterns and cause of codon usage, many indices have been projected to assess the degree and direction of codon bias.[11] Amongst them, the codon adaptation index (CAI) was proposed as an estimate of codon usage within a gene relative to a reference set of genes (by and large, ribosomal protein genes).[11] This index has been revealed to relate better with mRNA expression levels.[14] Over and above codon adaptation index, the effective number of codons (Nc),[15] which is described as the amount of equal codons producing the same codon usage bias as observed; and the incidence of optimal codons (Fop),[9] defined as the proportion of synonymous codons that are optimal codons, are also used.

The objective of this study was to execute a comparative analysis of the synonymous codon usage patterns, predict expression levels for the protein coding genes in these pathogenic bacteria with special reference to the genes linked with pathogenicity, examine horizontally transferred pathogenicity-related genes to detect their presence in the strains, and scrutinize the nature of highly expressed genes to infer upon their lifestyle. We consider that the result of this study would be helpful for the microbiologists working on this bacterium.

Materials and Methods

The complete genome sequences for five *Salmonella* strains [(*Salmonella enterica* Paratyphi, *Salmonella enterica* Typhi CT18, *Salmonella enterica* Typhi Ty2, *Salmonella enterica cholerasuis* SC-b67, and *Salmonella typhimurium* LT2 (hence forth, these strains will be referred to as SEP, SECT18, SETY2, SECSCb67, and STLT2 respectively)] were obtained from the IMG website (www.img.jgi.doe.gov).[16] All of the protein coding genes, genes associated with pathogenicity, and ribosomal protein genes were examined using Codon W software (http://bioweb2.pasteur.fr)[9] and CAI Calculator 2 (http://www.evolvingcode.net/codon/CalculateCAIs.php).[17]

The software Codon W[9] was employed to inspect G or C in the third position of codons (GC3s), as well as to determine the effective number of codons (Nc)[15] and the frequency of optimal codons (Fop).[9] Nc is a straightforward measure of codon bias.[17] It ranges from 20 (when merely one codon is used per amino acid) to 61 (when each and every codon is used in equal likelihood). Fop[9] determines the section of synonymous codons that are optimal codons. Its value varies from 0 (meaning a gene has no optimal codons) to 1.0 (when a gene is exclusively comprised of optimal codons).

The 'codon adaptation index' (CAI)[9] values were computed using the web-based application 'the CAI Calculator 2' (http://www.evolvingcode.net/codon/cai/cais.php)[17] taking the ribosomal genes as a reference. It quantifies the relative adaptiveness of a gene's codon usage, which is its codon usage as compared to the codon usage of highly expressed genes. The relative adaptiveness of each codon is the quantity of the usage of each codon compared to that of the most plentiful codon inside the same synonymous family.[9] The CAI value varies from 0 to 1.0, with higher CAI values signifying that the gene of concern has a codon usage pattern resembling that in the reference genes.

Z test was performed to check whether the values of the above-mentioned indices in the pathogenicity-related genes and ribosomal protein genes varied from those in the protein coding genes.

An analysis of the horizontally transferred pathogenicity-related genes among the studied strains was carried out to detect whether they are present in all the strains or native to a particular strain. The information about horizontally transferred genes was obtained from the website (http://cbcsrv.watson.ibm.com/HGT/).[18] Tsirigos and Rigoutsos[18] devised a new computational method for identifying horizontally transferred genes in 123 microbial genomes. It relied upon a gene's compositional features and necessitated having knowledge on codon boundaries. In addition to the single genes, the method was applicable to the clusters of genes transferred horizontally. The technique conveys a typicality score to each gene reflecting the gene's similarity with the containing genome, using specific features.[18]

First of all, the pathogenicity-related genes acquired by horizontal gene transfer mechanisms in the studied strains were sorted out. Using the Integrated Microbial Genomes database (www.img.jgi.doe.gov),[16] the sorted pathogenicity-related genes for each strain were subjected to IMG Genome BLAST against the studied strains to find out the sequence homologs. The minimum percent identity was set at 90%; and the maximum E (expect) value 1e-2.

Correspondence analysis (COA) was performed using Codon W (http://bioweb2.pasteur.fr).[9] This method explores the major trends in codon and amino acid variations among the genes.

Results

Codon usage patterns

Our first endeavour in the study of the codon usage patterns among various *Salmonella* genomes was to settle on the degree of variation in codon use. Most bacteria with a balanced AT/GC genome content have a sizeable amount of codon variation. Codon heterogeneity is usually associated with gene expression level. Thus, highly expressed genes contain a higher frequency of codons that are considered translationally optimal.[7,12,14] The GC3s and Nc values for all of the genes in these genomes were calculated to determine if codon heterogeneity exists among genes of various *Salmonella* species. Two different indices, namely,

effective number of codons (Nc) and GC3, have been used to detect codon usage variation among the genes. The Nc vs. GC3 plots have been recommended to be an effective means to explore the codon usage variations among genes in the same genome.[15] The Nc values of the *Salmonella* genes ranged from 25±1 to 61±0, and the GC3 values ranged from 0.14±0.3 to 0.91±0.6.

From Fig. 1, it is seen that the pathogenicity-related genes are lying below the expected curve. Genes which are anticipated to be highly expressed are clustered at one end of the Nc/GC3 plots. This phenomenon has been previously reported in *E. coli* and *Streptomyces*.[17]

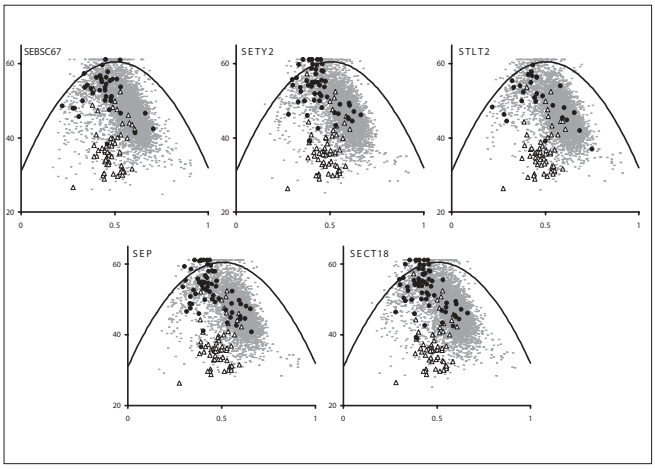

Figure 1: The effective number of codons used (Nc) in each gene (Y axis) plotted against the G+C content at synonymous third position of codons (GC3) (X axis) for all *Salmonella* genomes. The continuous curve in each plot symbolizes the null hypothesis that the GC bias at the synonymous site is solely due to mutation but not selection. Protein coding genes = "-"; PRG (pathogenicity-related genes) = æ '; ribosomal protein genes = "and #8710"

Table 1 shows that the mean Nc values of the total protein coding genes in the studied strains are in the range of 46-47, with the mean standard deviation value hovering around 6. With the exception of the ribosomal protein genes, the mean Nc values of the other categories of genes in the studied strains are quite high. From Table 1 it is observed that there is a good deal of variation of GC3 values among different categories of genes in the studied strains. Variation in the mean Nc values and GC3 values for the different gene groups was observed within the same species as well as other species. Ribosomal protein genes and the protein coding genes had higher Fop values compared to the pathogenicity-related genes.

Z test did not reveal any significant difference between the different types of genes undertaken in the study at significance level of 0.05%. Z test gives a standard normal cumulative distribution function. For a given hypothesized population mean, Z test returns the probability that the sample mean would be greater than the average of observations in the data set (array) - that is, the observed sample mean. From table 2 it is clearly seen that two-tailed probability values of Z test for CAI values in pathogenicity-related genes, ribosomal protein genes, and protein coding genes reveal trivial differences in SEP and STLT2 and are more or less same in SECSB67, SETY2, and SECT18. There is no significant correlation between the *P* values of the different sets of genes. The correlations have been depicted in Table 2.

Table 1: Mean values of effective number of codons (Nc), guanine cytosine percentage (GC), guanine cytosine ratio at third position (GC3), codon adaptation index (CAI), and frequency of optimal codons (Fop) of the genes in five *Salmonella* strains

Strain	Genes	Nc	GC%	GC3%	CAI	Fop
SEP	PCG	46.2±6.07	52.5±0.055	52.53±0.095	0.450±0.088	0.528±0.086
	RPG	36.90±6.05	51.20±0.02	48.6±0.062	0.705±0.102	0.738±0.090
	PRG	52.2±5.26	46.2±0.053	44.4±0.092	0.378±0.054	0.435±0.066
SECT18	PCG	47.26±6.28	52.1±0.056	55.4±0.098	0.445±0.086	0.522±0.085
	RPG	36.91±6.24	51.11±0.028	48.51±0.060	0.704±0.103	0.738±0.091
	PRG	52.61±5.73	46.32±0.050	44.30±0.087	0.377±0.055	0.434±0.064
SETY2	PCG	46.82±6.08	52.42±0.055	55.87±0.094	0.448±0.088	0.526±0.085
	RPG	36.91±6.24	51.17±0.028	48.57±0.061	0.704±0.103	0.738±0.103
	PRG	52.61±5.88	46.25±0.051	44.16±0.088	0.375±0.056	0.429±0.63
SECSCB67	PCG	47.12±6.22	52.31±0.05	55.64±0.096	0.460±0.083	0.522±0.084
	RPG	37.23±6.25	51.01±0.031	48.29±0.063	0.699±0.117	0.727±0.109
	PRG	52.08±5.17	45.67±0.056	44.31±0.098	0.412±0.037	0.449±0.055
STLT2	PCG	46.69±5.95	52.65±0.053	56.23±0.093	0.444±0.087	0.527±0.083
	RPG	36.52±5.92	51.19±0.029	48.55±0.062	0.710±0.100	0.743±0.089
	PRG	50.43±5.77	46.71±0.071	46.33±0.127	0.391±0.049	0.458±0.069

SEP= *Salmonella enterica* Paratyphi; SECT18= *Salmonella enterica* Typhi CT18; SETY2= *Salmonella enterica* Typhi Ty2; SECSCB67= *Salmonella enterica cholerasuis* SC-b67; STLT2= *Salmonella typhimurium* LT2; PCG= protein coding genes; RPG= ribosomal protein genes; PRG= pathogenicity-related genes

Table 2: Two-tailed probability values of Z test for codon adaptation index values in pathogenicity-related genes, ribosomal protein genes, and protein coding genes

Strain	Genes	Probability values of z test	Correlation with *P* values
SEP	PCG	0.854	0.192
	RPG	0.455	0.079
	PRG	0.498	-0.105
SECT18	PCG	0.494	-0.005
	RPG	0.480	-0.128
	PRG	0.496	0.142
SETY2	PCG	0.500	0.193
	RPG	0.479	-0.094
	PRG	0.495	-0.240
SECSCB67	PCG	0.500	-0.08
	RPG	0.491	0.263
	PRG	0.473	-0.170
STLT2	PCG	0.654	0.200
	RPG	0.494	0.165
	PRG	0.945	-0.l32

SEP= *Salmonella enterica* Paratyphi; SECT18= *Salmonella enterica* Typhi CT18; SETY2= *Salmonella enterica* Typhi Ty2; SECSCB67= *Salmonella enterica cholerasuis* SC-b67; STLT2= *Salmonella typhimurium* LT2; PCG= protein coding genes; RPG= ribosomal protein genes; PRG= pathogenicity-related genes

Analysis of horizontally transferred pathogenicity-related genes

From the web location (http://cbcsrv.watson.ibm.com/HGT/),[18] it was observed that the studied *Salmonella* strains contained 616, 555, 562, 604, and 558 horizontally transferred genes for SECSCb67, SEP, SECT18, STLT2, and SETY2 respectively. Among these the numbers of pathogenicity-related genes were 15, 33, 29, 11, and 18 for SECSCb67, SEP, SECT18, STLT2, and SETY2.

IMG genome BLAST results revealed homologs having sequence identity with a number of similar proteins in other strains. In SECSCb67 the pathogenicity-related genes like putative shiga-like toxin A subunit, vir K, pathogenicity island-encoded protein SPI3, virulence gene, cytoplasmic cell invasion proteins, secreted proteins in SOP, and outer membrane-associated proteins found 18 horizontally transferred homologs (percent identity ranging from 95 to 100) in STLT2, SEP, SECT18, and SETY2.

In SEP pathogenicity-related genes like type III secreted protein effector, putative pathogenicity island proteins, putative pathogenicity island lipoproteins, putative pathogenicity island effector protein, outer membrane invasion protein, outer membrane virulence proteins, toxin-like proteins, putative vir K proteins, virulence proteins, cell adherence invasions, virulence-associated secretary proteins, pathogenicity island 1 effector proteins, and oxygen-regulated invasins had 52 horizontally transferred homologs (percent identity, 96-100) in SECT18; SELT2, SETY2, and SECSCB67.

The SECT18 pathogenicity-related genes like putative auto transporter virulence proteins, putative pathogenicity island protein, putative pathogenicity island lipoproteins, putative pathogenicity island effector protein, outer membrane invasion protein, outer membrane virulence proteins, virulence proteins, cell invasion proteins, pathogenicity island 1 and 2 effector protein, cell adherence protein, hypothetical proteins associated with virulence, and invasion-associated proteins found 51 horizontally transferred homologs (percent identity, 95-100) in SELT2, SEP, SETY2, and SECSCB67.

Among the pathogenicity-related genes of SELT2, putative shiga-like toxin A protein, pathogenicity island-encoded protein A, virulence protein PAGD precursor, virulence proteins, and invasion protein transcriptional activators found 16 horizontally transferred homologs (percent identity, 95-100) in SETY2, SEP, SECT18, and SECSCB67.

In SETY2, the pathogenicity-related genes like putative pertussis-like toxin subunit A, outer membrane invasion protein, putative pathogenicity island effector protein, putative pathogenicity island protein, putative auto transporter/virulence factor, virulence protein, hypothetical protein associated with virulence, and invasion-associated secreted protein had 35 horizontally transferred homologs (percent identity, 95-100) in SELT2, SEP, SECT18, and SECSCB67.

Correlating codon usage bias with tRNA content in Salmonella genomes

Eduardo Rocha[19] discussed the correlation between codon usage bias and tRNA content in bacterial genomes. The optimal generation times of the five studied *Salmonella* genomes were obtained from personal communications with Prof. J. Parkhill, Sanger Institute, Welcome Trust Genome Campus. The studied *Salmonella* genomes had an optimal generation time of 0.5 to 1 hour and could be regarded as fast growers on the basis of Rocha's[19] observations. He reported that fast growers have a median of 61 tRNA genes compared to 44 for slow growers, and the former tend to have stronger codon usage bias contrary to the latter. SECSCb67, STLT2, SETY2, SECT18, and SEP had 85, 86, 78, 80, and 82 tRNA genes respectively. The studied *Salmonella* strains had on an average 37 distinctive anticodon tRNA genes, i.e., they had more similar tRNAs.

Multivariate statistical analysis

Multivariate statistical analysis was performed to study the codon usage variation among the genes. Correspondence analyses of codon count of the protein coding genes, ribosomal protein genes, and pathogenicity-related genes

for the *Salmonella* strains were performed. Fig. 2 reveals the positions of the genes on the planes defined by the first and second principal axes generated by COA of codon count for the protein coding genes, pathogenicity-related genes, and ribosomal protein genes. It is seen from fig. 2 that the scatter plot of SEP, STLT2, and SECT18 revealed a small core region and two ascending horns, as reported for other eubacteria like *E. coli,*[20] whereas that of SEBSC67 and SETY2 revealed a core region with two descending horns. Barring SEBSC67, the left horns in all the other genomes were less dispersed than the right horn. In case of SEBSC67, it is seen that the genes related to pathogenicity are located in the positive side of the Axis 1 compared to the same for other *Salmonella* genomes, where they reside on the negative side of Axis 1. Other genes are more or less clustered on the right side of the axis. The highly expressed genes are clustered together in the right horn of SEP, STLT2, SETY2, and SECT18 and left horn of SEBSC67 on the first axis of the COA of simple codon count.

No significant observation was noticed on correlating the CAI values of the protein coding genes of *Salmonella* strains with Axis 1. No correlation was observed between the positions of the genes on the Axis 1 produced by COA of codon count and the GC3 levels. However, we have found negative correlations between the positions of genes in Axis 1 produced by COA of codon count and Nc values of the protein coding genes in SECScb67 and SECT18 and SETY2 (results not shown). Very little positive correlations were obtained between positions of genes in Axis 1 and Nc values in SEP and STLT2. The genes with negative coordinates on the principal axis have more biased usage of codons compared to the genes with positive Axis 1 coordinates.

Detection of PHX genes in Salmonella

Codon adaptation index (CAI) is a gauge of directional synonymous codon usage bias. The index uses a reference set of highly expressed genes from a species to evaluate the relative merits of each codon, and a score for a gene is determined from the frequency of use of all codons in that gene. The index assesses the degree to which selection has been successful in moulding the pattern of codon usage. The CAI value was calculated using the ribosomal protein genes, which are known to be highly expressed as a reference. The CAI values for all genes in different *Salmonella* strains were calculated, and their distributions are shown in fig. 3.

The average CAI values for different gene groups associated with diverse functions varied. Ribosomal protein genes showed high CAI values, indicating high levels of gene expression. These CAI values ranged from 0.203 to 0.877, 0.14 to 0.872, 0.191 to 0.874, 0.196 to 0.872, and 0.188 to 0.872 for SEBSC67, SECT18, SEP, SETY2, and STLT2 respectively. The majority of the genes for the *Salmonella* genomes had CAI values between 0.3 and 0.5.

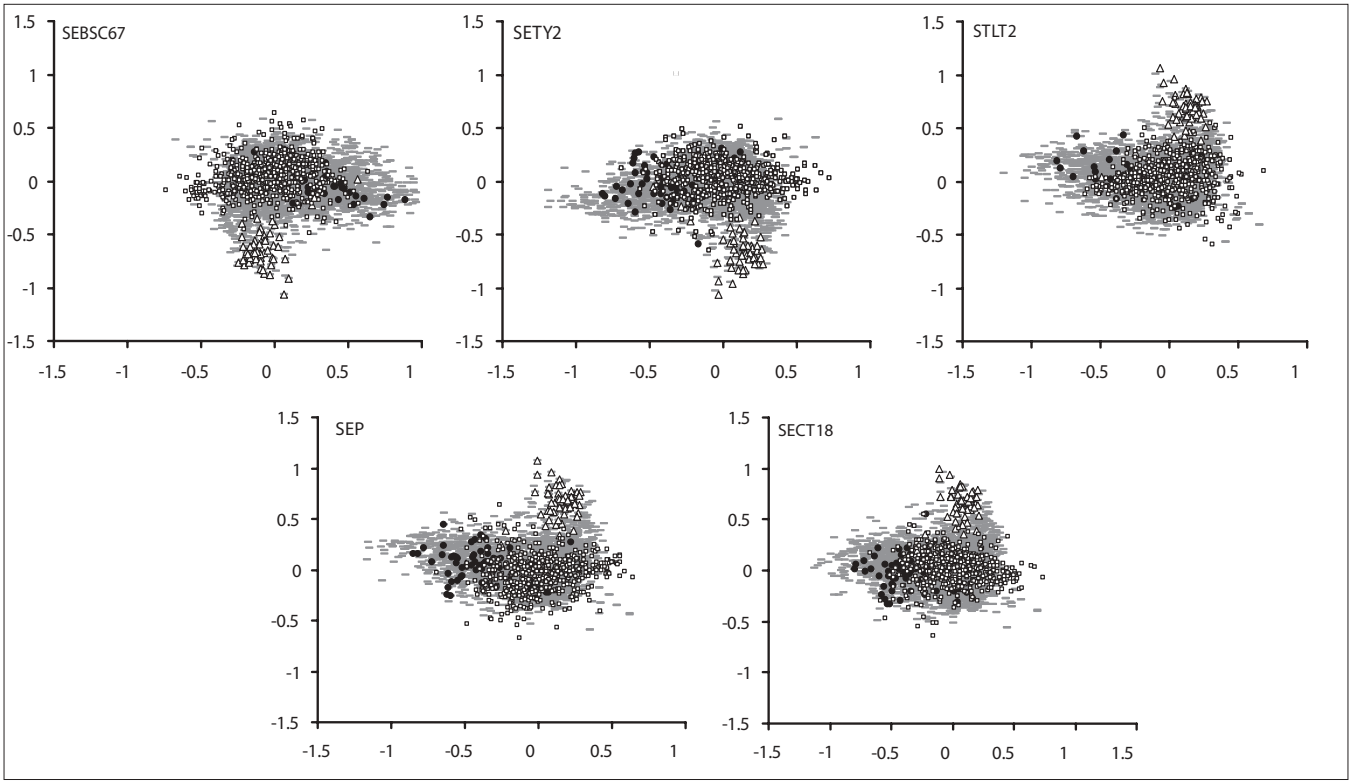

Figure 2: Correspondence analysis of codon usage patterns on codon count for various *Salmonella* genomes. In all the plots, X and Y axes correspond to axes 1 and 2 of the analysis. Horizontally transferred genes are represented by white boxes; rest of the buttons are as per fig. 1

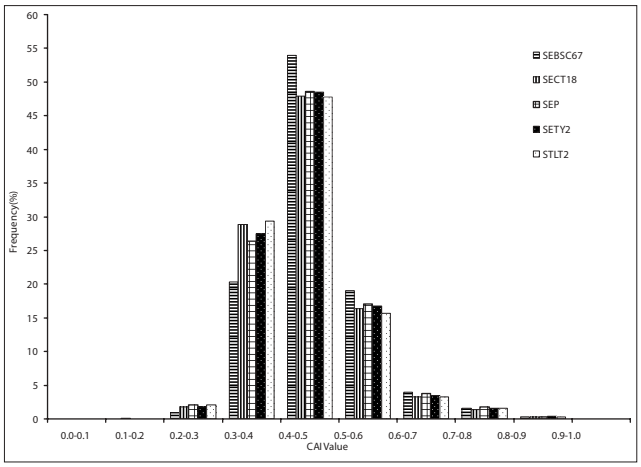

Figure 3: The frequency of distribution of the CAI values for all coding genes in the *Salmonella* genomes

As visualized by Wu *et al.,*[17] the top 10% of the genes, in terms of CAI values, were classified as the predicted highly expressed genes (PHX), and corresponded to CAI cutoffs of 0.562, 0.55, 0.558, 0.552, and 0.55 for SEBSC67, SECT18, SEP, SETY2, and STLT2 respectively. SEBSC67 had 477 PHX genes, including 51 ribosomal protein genes; SECT18 had 492 PHX genes, with 54 ribosomal protein genes; SEP had 423 PHX genes, with 54 ribosomal protein genes; SETY2 had 448 PHX genes, with 54 ribosomal protein genes; and SLT2 had 470 PHX genes, with 53 ribosomal protein genes.

Functional analysis of the PHX genes

To figure out the functional distribution of the PHX genes amongst the five *Salmonella* genomes, the clusters of orthologous groups of proteins were considered. For these *Salmonella* genomes, 20 COG categories were analysed. Fig. 4 illustrates the allocation of the PHX into each COG category on the basis of total PHX genes (a) and the total genes within that COG group (b), expressed as a percentage. To support the analysis, each of the COG categories were clustered in the following four COG groups: *information and storage processing* comprising of COGs connected to J-translation; K-transcription; L-DNA replication, recombination, repair (COG 1); *cellular processes* encompassing COGs linked to V-defence mechanism; T-signal transduction; M-cell envelope biogenesis; N-cell motility and secretion; U-intracellular trafficking; D-cell division; O-post-translational modification, protein turnover and chaperones (COG 2); *metabolism* consisting of COGs related to C-energy production and conversion; G-carbohydrate transport and metabolism; E-amino acid transport and metabolism; F-nucleotide transport and metabolism; H-coenzyme metabolism; P-inorganic ion transport and metabolism; I-lipid metabolism; Q-secondary metabolites, biosynthesis, and transport (COG 3); *general function prediction and unknown function* — R-general function prediction and S-unknown function (COG 4). The CAI values of each and every gene present in various COG groups were calculated, and the PHX genes were documented on the basis of the cut off values for various *Salmonella* genomes. Fig. 4 exemplifies the percentage of PHX genes in different COG categories clustered in the four COG functional groups. The *Salmonella* genomes had the following distribution in the COG functional groups: SEBSC67 — 17.20, 9.10, 15.19, and 6.57%; SEP — 18.11, 9.72, 12.84, and 5.6%; SECT18 — 16.5, 10.20, 14.85, and 6.09%; SETY2 — 17.7, 9.48, 14.46, and 5.36; and STLT2 — 17.46, 9.15, 13.87, and 5.45 for the COG functional groups 1 to 4 correspondingly.

Discussion

The Nc and GC3 values for all genomes suggested that they exhibited differences in codon usage as anticipated. If synonymous codon bias were to be absolutely dictated by GC3s, Nc values should fall on the expected curve of the GC3 and Nc plot. However, we found that except for a few, the values obtained for majority of the genes were well below the expected curve (Fig. 1). This result clearly indicates that codon usage bias for the greater part of *Salmonella* genes is affected independently of overall base composition. On an average, the high Nc values of the protein coding genes and pathogenicity-related genes suggest that they are lowly biased. The clustering of highly expressed genes at one end of the Nc/GC3 plots in all the *Salmonella* genomes points out that codon usage in the studied *Salmonella* strains has a strong probability of being determined by translational selection.

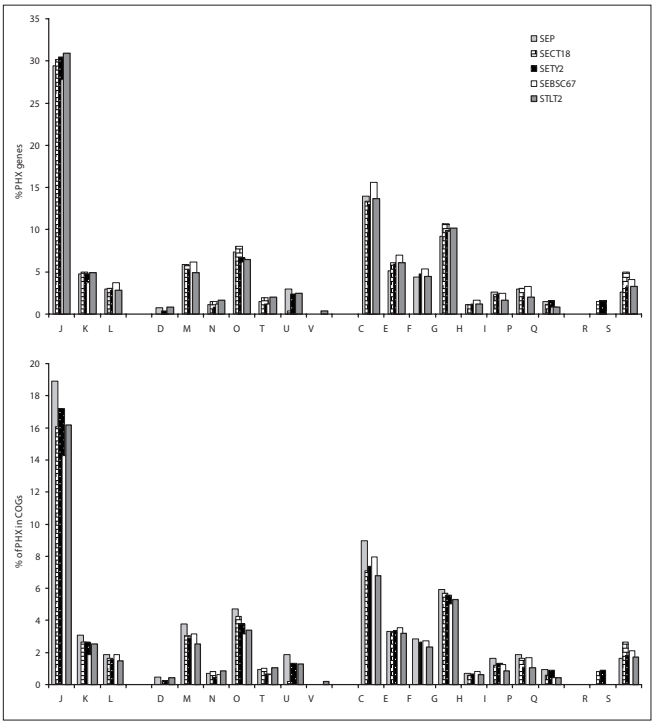

Figure 4: Distribution of *Salmonella* predicted highly expressed genes within functional COG groups (as in text)

On the whole, the GC3 content for these *Salmonella* genomes was moderate. Ribosomal protein genes and pathogenicity-related genes had lower GC3 values compared to the protein coding genes. Consequently, there are factors other than compositional constraints influencing codon usage variation among the genes. Higher Fop values of the ribosomal protein genes and protein coding genes compared to pathogenicity-related genes imply the presence of higher proportion of optimal codons in these genes. If mutational bias had wholly controlled codon bias, these genes would have had a low Fop value. Since that was not the condition for these *Salmonella* genomes, there may be additional factors like gene expression levels and GC3 compositional bias acting on codon usage bias.

It is seen from the results of the Z scores in table 2 that there is no significant correlation between the *P* values of the different categories of the genes in the studied genomes of *Salmonella*. So, the values for CAI in *Salmonella* genomes do not significantly differ in the categories of genes studied. These observations imply that there are inconsequential divergences in the characteristics of the studied genes.

The analysis of the pathogenicity-related genes revealed that not all of them were acquired by horizontal gene transfer mechanisms. Most of the pathogenicity-related genes acquired by horizontal gene transfer mechanisms were pathogenicity island encoded proteins, virulence proteins, secreted proteins, cell invasion proteins, toxin proteins, etc. Although the rest of the homologs for pathogenicity-related genes in all the strains showed percent identities ranging from 91 to 100, they were not found to be acquired by horizontal gene transfer mechanisms. These results indicated that they were native to those bacteria and they warded off the selective pressure of evolution. The horizontally transferred homologs, on the other hand, were gained from other organisms; and the high level of percent identity within the strains indicated that these genes are mobile within the genus. Most of them are associated with toxicity, virulence, pathogenicity islands, and invasion and are responsible for causing diseases resulting in epidemics. The high level of identity amongst them indicates that they evolved as a unit. Being a pathogenic bacterium, *Salmonella* has to fight against the host's defence systems, antibiotics, etc. The evolution of these genes as a unit suggests their ability to survive, infect, and exist as a pathogen.

Analysis of the correlation of codon usage bias with tRNA content in *Salmonella* genomes implies that these strains are well equipped to use small set of anticodons while maintaining high number of tRNAs. This is in line with Rocha's[19] observations. The ribosomal protein genes of these *Salmonella* strains, which are known to be highly expressed, showed high codon bias. This is expected since the codons associated with most abundant tRNAs have a propensity to be copious in highly expressed genes. The translation apparatus of *Salmonella* in all probability evolved with elevated codon bias in highly expressed genes compared to the rest of the genome. The mean CAI values of the studied *Salmonella* genomes varied widely from those of the ribosomal protein genes. This explains why selection for translational efficiency is the major source of variation of codon usage in *Salmonella* genomes. This has been previously exemplified by Rocha[19] in 102 bacterial genomes.

Multivariate statistical analysis data plotted in fig. 2 specify that the relative positions of the pathogenicity-related genes and ribosomal protein genes are same in all the studied strains. It is fascinating to see that the highly expressed genes are clustered together in all the strains, signifying that they share a similar codon bias that is somewhat diverse from the rest of the genes. These results indicate that the translational selection is quite strong enough to ward off the selection pressure due to mutation in the studied strains of *Salmonella*. Majority of the genes in the core region (±0.5 to +0.5) are associated with housekeeping functions and metabolic pathways and are highly conserved. Genes located away from this core region included a number of hypothetical protein genes, ribosomal protein genes, and translation factors. In all the strains, the horizontally transferred genes were clustered together in the core region.

Absence of any significant correlation of the CAI values with Axis 1 of correspondence analysis of the protein coding genes of *Salmonella* strains clearly shows that expression levels do not discriminate genes according to their codon usage along the major explanatory axis. This was expected since the average CAI values of the protein coding genes are much lower than those of the ribosomal protein genes. In fact, a comparison of the results of different indices (Table 1) for ribosomal protein genes and all the protein coding genes reveals wide differences. These results validate our point that *Salmonella* genomes with lower mean CAI values are controlled by translational selection. No correlation of the positions of genes on the Axis 1 produced by COA of codon count with GC3 indicates that GC3 levels have practically no effect in differentiating the genes according to the codon usage variation along the first major explanatory axis. Negative correlation of the positions of genes in Axis 1 produced by COA of codon count with Nc values of the protein coding genes in SECScb67 and SECT18 and SETY2 is attributed to the decrease in codon bias among the genes lying towards the left of Axis 1.

The plot of the frequency distribution of CAI values for the five *Salmonella* genomes showed more or less similar distribution patterns. All the genomes had a peak in the 0.4-0.5 CAI range. CAI values for all the genomes rose and fell steadily. SEBSC67 had the highest peak value, viz., 53.90%. It has been noted that the percentages of PHX genes in COG category 1 and COG category 3 for the *Salmonella* genomes are well above the expected value of 10%. This reveals

that the genes in these categories have reasonably superior expression levels than rest of the genes in the genomes. Functional analysis showed that the COG functional group 1 (information and storage processing) incorporated the maximum number of PHX genes in all the genomes. The COG groups translation, ribosomal structure biogenesis [J], and energy production and conversion [C] contained the highest number of predicted highly expressed genes. The distribution of high number of PHX genes in the translation, ribosomal structure biogenesis (J) functional groups of COG is attributed to the presence of high percentage of ribosomal protein genes which are highly expressed. Ribosomal protein genes which are PHX contributed to 67.94%, 66.66%, 67.08%, 70.42%, and 67.08% of PHX genes for SEP, STLT2, SETY2, SEBSC67, and SECT18 in the (J) functional group. Therefore, the weights of the ribosomal proteins played an important role in this case. Elevated number of PHX genes associated with translation, ribosomal structure biogenesis is beneficial for *Salmonella* to cause infections, overcome host immunity, and spread disease. The distribution patterns of the PHX genes in the various COG groups were approximately alike in all the five strains. Approximately 75% to 80% of the protein coding genes of the *Salmonella* strains belong to the COG category. This is significant because the huge number of genes in the COG groups of the *Salmonella* strains, in fact, helps them preserve their lifestyle, and it also divulges that *Salmonella* genomes are not subjected to genome reduction leading to gene loss. Being a pathogenic bacterium, it has to overcome host defence mechanisms to establish infection; and the presence of the genes responsible for pathogenicity and toxicity in the COG groups merely proves the fact.

The results from this study indicate variations existing among the genes of these genomes. Selection for translational efficiency is the major source of variation of codon usage in the genes of *Salmonella*. GC3 composition does not influence codon usage variation in the genes of these *Salmonella* strains. The horizontally transferred homologs, on the other hand, are gained from other organisms, and the high level of percent identity within the strains indicated that these genes are mobile within the genus. The evolution of these genes as a unit suggests their ability to survive, infect, and exist as a pathogen.

Correspondence analysis revealed clustering of the highly expressed genes together. Genes belonging to the COG categories are more or less conserved in the studied strains. Codon usage-based strategy has been applied to identify highly expressed genes in the studied strains of *Salmonella*. Genes related to information and storage processing include the highest number of PHX genes. Huge numbers of genes (approximately 75%-80%) in the COG categories of *Salmonella* genomes reflect their way of existence.

Acknowledgements

The authors are grateful to the Department of Biotechnology (DBT), Government of India, for providing financial help in setting up of Bioinformatics Informatics facility at the Department of Botany, University of North Bengal.

References

1. Adams MR, Moss MO. Food microbiology. New Age International (P) Limited, Publishers; 2000. p. 102-4.
2. Wray C, Sojka WJ. Experimental *Salmonella typhimurium* infections in calves. Res Vet Sci 1978;25:139-43.
3. Brown NF, Vallance BA, Coombes BK, Valdez Y, Coburn BA, Finley BB. *Salmonella* pathogenicity island 2 Is expressed prior to penetrating the intestine. PLoS Pathog 2005;1:e32.
4. Groisman EA, Ochman H. How *Salmonella* became a pathogen. Trends Microbiol 1997;5:343-49.
5. Hacker J, Blum-Oheler G, Mulhdorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function, impact on microbial evolution. Mol Microbiol 1997;23:1089-97.
6. Blum G, Ott M, Lishewski A, Ritter A, Imrich H, Tschape H, *et al*. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of *E. coli* wild type pathogen. Infect Immun 1994;62:606-14.
7. Ikemura T. Codon usage and transfer-RNA content in unicellular and multicellular organisms. Mol Biol Evol 1985;2:13-34.
8. Banerjee T, Basak S, Gupta SK, Ghosh TC. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. J Biomol Struct Dyn 2004;22:13-23.
9. Sen A, Sur S, Bothra AK, Benson DR, Normand P, Tisa LS. The implication of life style on codon usage patterns and predicted highly expressed genes for three *Frankia* Genomes. Anton van Leeuw 2008;93:335-46.
10. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986;24:28-38.
11. Sharp PM, Li WH. The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res 1987;15:1281-95.
12. Lafay B, Atherton JC, Sharp PM. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology 2000;146:851-60.
13. Dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucl Acids Res 2003;31:6976-85.
14. Ikemura T. Correlation between abundance of *E. coli* tRNAs and their occurrence of the respective codons in protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* system. J Mol Biol 1981;146:1-21.
15. Wright F. The “effective number of codons” used in a gene. Gene 1990;87:23-9.
16. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, *et al*. An experimental metagenome data management and analysis system. Bioinformatics 2006;22:e359-67.
17. Wu G, Culley DE, Zhang W. Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces*

avermitilis and the implications for their metabolism. Microbiology 2005;151:2175-87.
18. Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. Nucl Acids Res 2005;33:922-33.
19. Rocha EP. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Res 2004;14:2279-86.
20. Medigue C, Viari A, Henaut A, Danchin A. *Escherichia coli* molecular genetic map (1500 kbp) Update II. Mol Microbiol 1991;5:2629-40.

Source of Support: Nil, **Conflict of Interest:** None declared.