Comparative study of Codon usage pattern and compositional distribution between whole genome and virulence gene set of *Vibrio cholera N16961*.

Sushanta Deb, Surajit Basak✉

Department of Molecular Biology and Bioinformatics, Tripura University, Suryamaninagar, India

✉ Corresponding author email: basaksurajit@gmail.com

Preferred citation for this article:

Deb S. and Basak S., 2015, Comparative study of Codon usage pattern and compositional distribution between whole genome and virulence gene set of Vibrio cholera N16961., Computational Molecular Biology, Vol.5, No.6 1-4 (doi: 10.5376/cmb.2015.05.0006)

Abstract *Vibrio cholerae* is the pathogenic organism causes cholera, a severe diarrheal disease. Occurs frequently in southern Asia. *Vibrio cholerae* has both pathogenic and nonpathogenic strains that vary in their virulence gene content. Great variety of strains and biotypes of *Vibrio cholerae* are found. These varieties are involve in shuffling of different pathogenic factors   among them such as receiving and transferring genes for toxins, colonization factors, antibiotic resistance, capsular polysaccharides which giving resistance to chlorine 7 and new surface antigens, such as the 0139 lip polysaccharide and O antigen capsule. Different mode of transfer of these virulence gene i.e. lateral and horizontal transfer by phase, collection of pathogenic genes and other accessory genetic element, pave the way to understand how bacterial pathogen develop its Pathogenicity and become a new strain. To provide a insights into the genetic features and the relationship between the overall codon usage pattern of virulence gene set (VGS).We measure the GC content of VGS which shows that there is no any difference between GC content of whole genome and VGS .It also has been found that GC content shows the similar distribution among the CDS of both whole genome and Virulence gene set. A correlation analysis between the A3s, T3s, G3s, C3s, and GC3s, the ENC values, and the nucleotide contents (A%, T%, G%, C%, and GC %) indicated that mutational bias plays role in shaping the VGS codon usage bias.

Keywords Codon usage pattern; whole genome; virulence gene; Vibrio cholera

Introduction

Strains of the El Tor biotype causes sporadic infections and cholera epidemics as early as 1910, this biotype emerged in 1961 to cause the 7th pandemic which in turn causes the global elimination of classical biotype strains as a cause of disease. The Gram negative, *Vibrio cholerae* El Tor N16961 possesses a complete genomic sequence of is 4,033,460 base pairs (bp). The whole genome of *Vibrio cholerae* El Tor N16961 is divided in to two circular chromosomes of 2,961,146 bp and 1,072,314 bp. The total 3,885 open reading frames were encoded by the whole genome of *vibrio cholera*. Major part of recognizable genes which plays a chief role in cell functions (such as DNA duplication, transcription, protein synthesis and cell-wall biosynthesis) and pathogenicity (for example, toxins, surface antigens and adhesions) are resides on the primary chromosome. The *V. cholerae* genomic sequence open the scope for understanding how a free-living,

environmental organism evolved to become a significant human bacterial pathogen.

Pathogenic bacteria uses a lot of mechanisms to cause disease in human hosts. Bacterial pathogens have a wide range of molecules that bind host cell targets to produce different type of host responses. The molecular mechanism of pathogenic bacteria to interact with the host unique to every pathogens or conserved throughout the several different species. The availability of complete genome sequences for several bacterial pathogens facilitates to reveal the mystery behind the molecular strategy used by bacteria to infect host. The "horizontal gene transfer" is one of the major factor which changes the genomic feature of bacterial genome in a fast and dramatic way. Recent studies have shown that horizontal gene transfer plays an important role in the molecular evolution of novel bacterial pathogens. There is a section which may contain large blocks of virulence determinants (adhesions, invasions, toxins, antibiotic

resistance protein, etc.), and thus are referred to as pathogenicity islands. It has also been reported that several biotic factors has influence of pathogenicity of bacteria and on the genomic features of pathogenic genes.

1 Methodology

The coding sequences of whole genome of vibrio cholera N16961 retrieved from the NCBI ftp site (http://www.ncbi.nlm.nih.gov/Ftp/) and the virulence gene set are downloaded from the pathogenic Island Database (http://www.paidb.re.kr/about_paidb.php). Genes under the gene set of virulence are eliminated from the whole genome CDS to avoid the recurrence of CDS in both the gene set. Codon compositions (A3s, T3s, G3s, C3s, and GC3s) of virulence gene set are obtained using software Codon W (written by John Peden) and taken from (ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z/). The nucleotide content (A%, T%, G%, and C%) of each cds of VGS was analyzed using the MEGA 4.0 biosoftware for windows. The obtained data are further analyzed statistically using statistical software (www.statsoft.com/Products/STATISTICA/Data-Miner) to get the values of statistical measurement. We also measure the extent of codon usage bias (NC_{diff}) in this bacterial genome. To measure the NC_{diff} we have downloaded all the ribosomal protein coding genes from the NCBI ftp site, we generate two set of coding sequence to evaluate NC_{diff} i.e ribosomal protein coding gene and rest of the genes. Using codon W software we have analyzed the ENC value of ribosomal genes and rest of the genes.

2 Result and Discussion

To investigate whether there is any possible influence of mutational pressure on the codon usage bias in the VGS the correlation analysis was performed between the composition at different codon position (A3, T3, G3, C3, and GC3), the nucleotide compositions (A%, T%, G%,C%, and GC%) and ENC values(Table 1). The results indicate that most of the codon compositions correlated with the nucleotide compositions. Additionally, ENC value always shows no correlation with the nucleotide compositions. These results confirmed the codon usage bias of the VGS was influenced by the nucleotide compositions, and hence by mutational bias.

Table 1: The correlation between the codon compositions (A3s, T3s, G3s, C3s, and GC3s), the ENC values, nucleotide compositions (A%, T%, G%, C%, and GC%) value of the coding sequence of virulent genes.

	A%	T%	C%	G%	GC%
A3s	0.8041	0.0707	-0.4902	-0.5241	-0.6393
T3s	0.4029	0.6070	-0.5921	-0.6762	-0.3935
C3s	-0.2190	-0.6002	0.6797	0.1969	0.8144
G3s	-0.6766	-0.1759	0.3012*	0.6587	-0.5269
GC3s	-0.4680	-0.2950 *	0.6952	0.4453	0.5445
ENC	-0.0981	0.0995	0.0351	-0.0340	0.0092

Note: $* \ 0.01 < P < 0.05$

We also make the GC12 VS GC3 plot of both the CDS set i.e. whole genome CDS and VGS CDS in both the case the correlation values are more or less similar indicating that mutational pressure influencing both gene sets similarly. GC12 is the average value of GC1 and GC2 and GC3 is plotted against this average value and find out the correlation to predict whether there is any difference between mutational forces shaping codon usage bias in both the CDS set. Plot of GC3 against GC12 is showing a comparatively weaker, but significant correlation (r = 0.2749, p < 0.1) in case of virulence gene set. The above findings indicate that the forces that are shaping the compositional patterns of the Whole genome and VGS are the same for all codon positions and acting on the three codon positions in a similar way.

2.1 GC content distribution among the whole genome and Virulence gene set

From the figure 1 and figure 2, it can be assumed that GC content is more or less uniformly distributed among the CDS of whole genome and VGS. So it can be predicted that whole genome and VGS may have the similar kind of nucleotide composition as well as may have share the same pattern of codon usage.

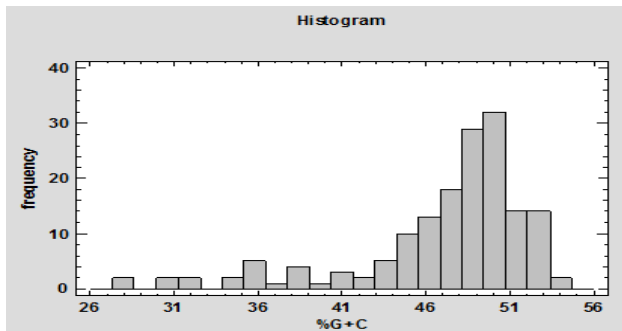### 2.2 Extent of codon usage bias

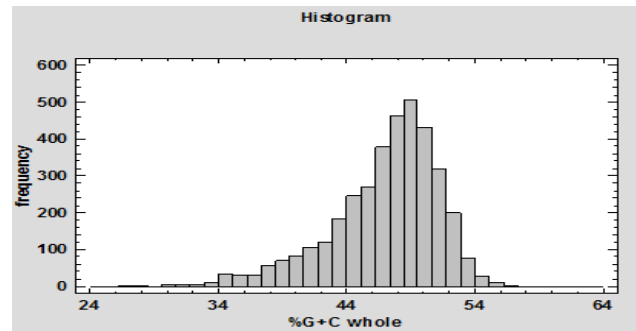Figure 1 Distribution of GC content among the genes of VGS

Figure 2 Distribution of GC content among the genes of whole genome

We consider organism whether it is bias or unbiased in relation with its codon usage if highly expressed genes shows a different distribution of synonymous codons from that in other genes in the genome. Several methods were proposed to assess the extent of codon usage bias at an organism scale, which determine the organism as biased or unbiased. NC_{diff} is a most widely used measure to judge whether the organism is biased or unbiased. Here we use the measure to evaluate the extent of codon usage bias of *Vibrio cholera N16961*. The difference of average NC values of the ribosomal protein coding genes and the average NC values of the rest of the genes in the genome are referred to as NC_{diff}. Organism having high value and low value of NC_{diff} exhibit large and small extents of codon usage bias respectively.

NC_{diff} was obtained by:

$$\frac{NC(all) - NC(rib)}{NC(all)}$$

The NC_{diff} value of this bacteria is very low which denotes that this bacteria shows a very small extant of codon usage bias in its whole genome.

2.3 GC content of VGS and whole genome

Total GC content of pathogenic gene set and the whole genome gene set (devoid of pathogenic gene) were measured and it has been found that there is uniform distribution GC between this two gene set, which implies that pathogenic gene set not influenced by mutational pressure and other factors and there is hardly any chance of horizontal gene transfer into the pathogenic gene set of vibrio cholera. Total GC content of whole genome and VGS (pathogenic set)

respectively 47% and 48%, the variation of GC content between this two gene set is negligible.

3 Conclusion

From this study it can be observed that there is no any difference in codon usage pattern of VSG and genome of *Vibrio cholera N16961* as well as pathogenic gene set share the same pattern of distribution of nucleotide composition with the whole genome. The selection force i.e. mutational pressure are influencing in a similar fashion in both the gene set of *Vibrio cholerae N16961*. This finding also support that there is hardly any genome wide codon usage difference as we find that extent of codon usage bias is very low throughout the genome, which indicate that there may be very least chance of codon usage difference among the different functional categorical genes in this pathogenic bacteria.

References
Blake P.A. and Olsvik Ø. (eds.), 1994, In Vibrio cholerae and Cholera: Molecular to Global Perspectives, ASM Press, pp. 293-295
http://dx.doi.org/10.1128/9781555818364.ch18

Lawrence J.G. and Roth J.R.,1996, Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters, Genetics, 143(4): 1843–1860

Ochman H. and Moran N., 2001, Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis, Science, 292: 1096–9
http://dx.doi.org/10.1126/science.1058543

Liu Q.P., 2006, Analysis of codon usage pattern in the radioresistant bacterium Deinococcus radiodurans, BioSystems, 85(2): 99-106
http://dx.doi.org/10.1016/j.biosystems.2005.12.003

Read A.F., 1994, The evolution of virulence, Trends Microbiol, 2: 73–81
http://dx.doi.org/10.1016/0966-842X(94)90537-1

Waldor M.K., Colwell R., and Mekalanos J.J., 1994, The Vibrio cholerae O139 serogroup antigen includes an O-antigen capsule and lipopolysaccharide virulence determinants. Proc. Natl Acad. Sci. USA 91, 91(24): 11388–11392

Ziebuhr W., Ohlsen K., Karch H., Korhonen T., and Hacker J., 1999, Evolution of bacterial pathogenesis,Cell Mol Life Sci, 56:719–28
http://dx.doi.org/10.1007/s000180050018

Botzman M. and Margalit H., 2011, Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles, Genome Biology, 12: R109
http://dx.doi.org/10.1186/gb-2011-12-10-r109

Makino K., Oshima K., Kurokawa K., Yokoyama K., Uda T., Tagomori K., Iijima Y., Najima M., Nakano M., Yamashita A., Kubota Y., Kimura S., Yasunaga T., Honda T., Shinagawa H., Hattori M., and Iida T., 2003, Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholera, Lancet. 1, 361(9359):743-9

Jazel D. and Satchell K.J.F., 2013, Analysis of Vibrio cholerae Genome Sequences Reveals Unique rtxA Variants in Environmental Strains and an rtxA-Null Mutation in Recent Altered El Tor Isolates, mBio, 00624-12

Blokesch M. and Schoolnik G.K., 2007, Serogroup conversion of Vibrio cholerae in aquatic reservoirs, PLoS Pathog., 3: e81.

Safa A., Nair G.B.,and Kong R.Y., 2010, Evolution of new variants of Vibrio cholerae O1. Trends Microbiol., 18: 46–54
http://dx.doi.org/10.1016/j.tim.2009.10.003

Siddique A.K., Nair G.B., Alam M., Sack D.A., Huq A., Nizam A., Longini IM Jr., Qadri F., Faruque S.M., Colwell R.R., Ahmed S., Iqbal A., Bhuiyan N.A. and Sack R.B., 2010, El Tor cholera with severe disease: a new threat to Asia and beyond, Epidemiol. Infect., 138: 347–352
http://dx.doi.org/10.1017/S0950268809990550