Biochemical and Biophysical Research Communications **290,** 81–84 (2002)
doi:10.1006/bbrc.2001.6174, available online at http://www.idealibrary.com on IDEAL®

Base Frequencies at the Second Codon Position of *Vibrio cholerae* Genes Connect with Protein Function

Ju Wang[1] and Feng-Biao Guo
Department of Physics, Tianjin University, Tianjin 300072, China

Received November 8, 2001

In this paper, the base frequency at the second codon position of the 3839 open reading frames (ORFs) in the *Vibrio cholerae* genome is analyzed. It is shown that according to the base content at this codon site, the ORFs can be divided into two clusters, each containing 673 and 3166 ORFs, respectively. ORFs in the smaller cluster usually have significantly higher T frequency than that of A at the second codon position. For the two clusters of ORFs, there are significant differences in the frequencies for 18 of the 20 amino acids in the encoding proteins. The two clusters of ORFs are also significantly different in their functions. More than half of the known genes involved in transport and binding are included in the smaller cluster, while few genes involved in amino acid biosynthesis, protein synthesis, and so on are included in this cluster. © 2002 Elsevier Science

***Key Words: Vibrio cholerae* genome; base frequency; codon positions; protein function.**

It has been found that the codon usage of a genome contains other information than that necessary for encoding proteins. The codon usage variation may be related to gene expression level (1, 2), the origin of genes (3–6), or the chromosomal regional location of genes (7, 8). Some recent work indicates that the disparity in the mutational bias between the leading and the lagging strand is the most important source of variation in codon usage in some spirochaete species such as *Borrelia burgdorferi* and *Treponema pallidum* (9, 10). It is also reported that the codon usage may be related to the subcellular location of some proteins (8). Additionally, it is suggested that synonymous codon usage may connect with the protein secondary structural units (11–13). As to the single nucleotide frequency, there is some propensity for G at the first codon position and some deficiency for this base at the second codon position. While at the third codon position, there is no general preference for any base, and the G + C percentage usually varies for different genes and species (14–19). It is suggested that the first, second and third codon position associates with the biosynthetic pathway, hydrophobicity and helix or beta-strand forming potentiality of the coded amino acids (20, 21). It is also reported that there is strong correlation between the base frequencies in the second codon position of genes and the corresponding secondary structures in the encoded proteins (13, 22).

Recently, the whole genome sequence of *Vibrio cholerae,* the etiological agent of cholera, was determined (23). It is found that the codon usage pattern in this genome is somewhat different from those of the other genomes (24). In this paper, the asymmetry of the base frequency at the second codon position of *V. cholerae* ORFs is analyzed in detail. We find that genes in various functional categories show significant difference in base choice at this codon position.

MATERIALS AND METHODS

The *V. cholerae* genome DNA sequences and the annotation information were downloaded from ftp://www.tigr.org on March 14, 2001. The database includes the DNA sequences and coordinates of 3883 predicted ORFs [Heidelberg *et al.* (23) reported that there were 3885 ORFs in the *V. cholerae* genome, but two were missed in the database], as well as their function description (the database is referred to as TIGR database, hereafter).

Among the 3883 ORFs in the TIGR database, 2037 are known genes whose function have been determined experimentally or through database search, all of which are assigned explicit function in the database. Additionally, there are 271 ORFs recognized as “putative proteins” in the database, and they show significant similarity to known genes in other genomes. These ORFs are referred to as class 1 (including 2308 entries), hereafter. The remaining 1575 ORFs are recognized as “conserved hypothetical proteins” or “hypothetical proteins” in the TIGR database, whose functions are still needed to be exploited (referred as class 2, hereafter). There are 36 ORFs in class 1 whose lengths cannot be divided by 3, and there are 8 such ORFs in class 2, all of them are excluded from current study. Then, we have 2272 ORFs in class 1 and 1567 ORFs in class 2, respectively. Of the 2272 ORFs in class 1, 1755 and 517 are located on chromosomes 1 and 2, respectively. While for the 1567 ORFs in class 2, 990 and 577 are located on chromosomes 1 and 2, respectively.

[1] To whom correspondence and reprint requests should be addressed. Fax: 86-22-2740 2697. E-mail: wangju@eyou.com.

0006-291X/02 $35.00

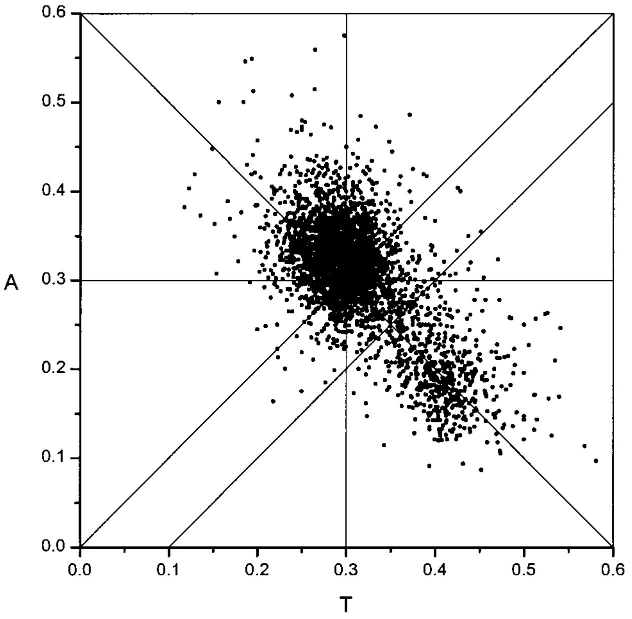

FIG. 1. The occurrence frequencies of bases A and T at the second codon position for each of the 3839 *V. cholerae* ORFs are plotted. There are 3839 points altogether. It can be seen that the ORFs gather into two clusters, and the line $t_2 = a_2 + 0.1$ can be used to separate them.

K-means clustering method (25) is adopted to get some idea about the quantitative resemblance or difference among the base frequencies of the ORFs. *K*-means is a statistical method used to cluster data set into the given *K* classes based on the similarity of the elements. The idea in this method is to find a clustering (or grouping) of the observations so as to minimize the total within-cluster sums of squares. In this case, it sequentially processes each observation and reassigns it to another cluster if doing so results in a decrease in the total within-cluster sums of squares [referring to (25) for the details].

RESULTS AND DISCUSSION

The frequencies of nucleotide A, C, G and T at the second codon position for each of 3839 ORFs in the TIGR database are calculated, which are denoted by a_2, c_2, g_2, and t_2, respectively. In Fig. 1, the distribution of $t_2 \sim a_2$ is plotted. It can be seen that most points cluster in a region around the (0.3, 0.3), while there are quite some points gather in a smaller region beside it. For base G and C, the variation in their frequencies is not so marked and the distribution is not shown here. Then the *K*-means clustering algorithm is used to cluster the ORFs based on the frequencies of the four bases. It is found that the 3839 ORFs can be divided into two unequal groups, the larger and smaller group contains 3166 and 673 genes, respectively. Of the 673 genes, 436 are on the chromosome 1 and the other 237 are on the chromosome 2. In Fig. 1, the two groups of point can be approximately separated by a line, ***L:*** $t_2 = a_2 + 0.1$. For the points lie above the line, $t_2 < a_2 + 0.1$, but for those below the line, $t_2 > a_2 + 0.1$. This means we can classify an ORF in *V. cholerae* genome into one of the two clusters according to the frequencies of adenine and thymine at site 2 of its codons. An ORF in the smaller cluster has significantly more codons with T at the second codon position than those with A at this position.

Then the amino acid composition of the encoding proteins is checked. A χ^2 test is carried out to evaluate whether there are significant differences in amino acid usage between the encoding proteins of the two clusters of ORF. For each amino acid, the χ^2 test involves a 2×2 table with one degree of freedom. The first row contains the frequencies of the amino acids being analyzed, and the second row contains the total number of amino acids of each cluster. Significance is examined at the 5% level (χ^2 value of 3.841). The differences are highly significant for 18 of the 20 amino acids except Pro and Thr. It is found that proteins encoded by ORFs in the smaller cluster have more Leu, Phe, Ile, and Val and less Glu, Asp, Lys, and Gln on average, i.e., they contain more hydrophobic amino acids in their sequences.

Then the function of each ORF in the two clusters is checked. For the 673 ORFs in the smaller cluster, 290 belong to class 1, the others belong to class 2. The function category for the 3166 and 673 ORFs are summarized in Table 1 [the function category adopted here is same as that used by Heidelberg *et al.* (23)]. Interestingly, none of the known and putative proteins related to amino acid biosynthesis, nucleosides and bases, protein synthesis, as well as those plasmid-related, phage-related and transposon-related (they are classified as other categories) are included in the

TABLE 1

Number of Genes Found in Different Function Categories

No.	Function category	Smaller cluster	Larger cluster
1	Amino acid biosynthesis	0	91
2	Purines, pyrimidines, nucleosides, and nucleotides	0	61
3	Fatty acid and phospholipid metabolism	6	41
4	Biosynthesis of cofactors, prosthetic groups, and carriers	5	111
5	Central intermediary metabolism	3	99
6	Energy metabolism	34	220
7	Transport and binding proteins	165	157
8	DNA metabolism	2	96
9	Transcription	1	50
10	Protein synthesis	0	129
11	Protein fate	11	93
12	Regulatory functions	7	258
13	Cell envelope	4	108
14	Cellular processes	29	220
15	Other categories	0	41
16	Unknown	23	207
17	Conserved hypothetical and hypothetical	383	1184
	Sum	673	3166

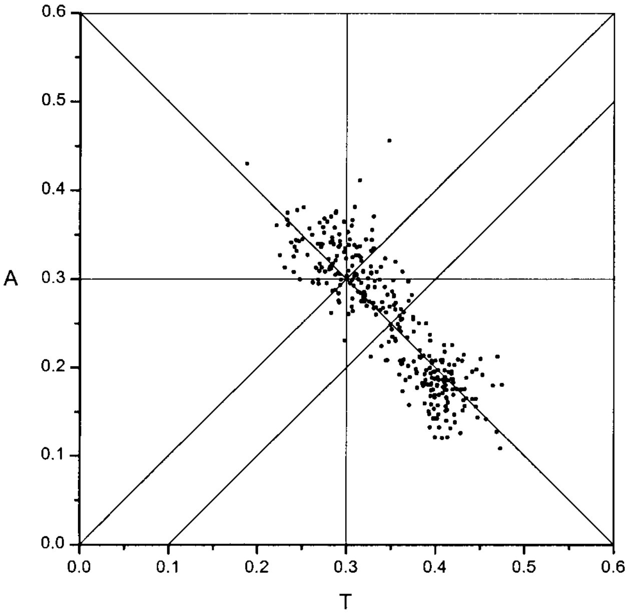

FIG. 2. The occurrence frequencies of bases A and T at the second codon position for each of the 322 *V. cholerae* ORFs are plotted. There are 322 points altogether. It can be seen that the ORFs gather into two clusters, and the line $t_2 = a_2 + 0.1$ can be used to separate them.

smaller cluster, even though nearly 18% of the total ORFs are included in this cluster. Furthermore, very few ORFs involved in fatty acid metabolism, cofactors and prosthetic groups, central intermediary metabolism, DNA metabolism, transcription, regulatory functions, as well as cell envelope are included in this cluster. On the contrary, more than half of the proteins related to transport and binding (totally 165 ORFs) are included in the smaller cluster, which also account more than half of the 290 ORFs of class 1 in this cluster. The remaining ORFs in this cluster are mainly involved in energy metabolism (34), cellular processes (29) and protein fate (11), and the figures in the parentheses indicate the number of genes in this category. Altogether, there are 322 genes in class 1 encoding transport and binding proteins, and the distribution of them on plane $t_2 \sim a_2$ is shown in Fig. 2. These genes can also be divided into two groups by the line $t_2 = a_2 + 0.1$, 157 of them locate below the line. For one group, the average of the base contents are $\bar{a}_2 \approx 0.30$, $\bar{t}_2 \approx 0.30$; for the other group, $\bar{a}_2 \approx 0.15$, $\bar{t}_2 \approx 0.40$.

Of the ORFs in the smaller cluster, 117 are enzymes. Furthermore, 63 of these enzymes belong to permeases, a kind of protein involved in the transport of ions and other nutrition through the membrane. Altogether, there are 63 permease genes among the 2272 known and putative genes. In other words, all such genes are restricted to the smaller cluster, i.e., the T-rich, A-poor group.

The relationship between the base choices at different codon positions and many important topics, such as the biosynthetic pathway and hydrophobicity of amino acids, expression level of genes or the chromosomal location of genes, has been studied extensively (20, 21, 26, 27). It is interesting to note that there is such a strong correlation between the second codon position of genes and functions of proteins. This may give us some clues about the function of the unknown ORFs, including the hypothetical and conserved hypothetical ORFs. For example, a gene in the smaller cluster or with $t_2 > a_2 + 0.1$ may be less possible to encode a protein involved in amino acid synthesis or protein synthesis, but more possible encode a transport and binding protein.

Recently, it is reported that there is correlation between the base frequencies at the second codon position of genes and the corresponding secondary structures in the encoded proteins (13, 22). Especially, the secondary structures show marked differences in the frequency of T (or U) and A at the second codon site, with the aperiodic structure showing the lower T (or U), while higher A frequency; on the other hand, the beta-strand structure and helix showing significantly higher T (or U) frequency and relatively lower A value. Chiusano *et al.* (22) attributed this relation to the hydrophobic and hydrophilic amino acids encoded by codons with U or A, respectively, in their second codon site. If the phenomena they observed are universal, we may deduce that for *V. cholerae*, since nearly 700 genes are relatively richer in T than A at the second codon site, their encoding proteins may also be different in structures as those of the other proteins. Consequently, there will be difference in the function of these proteins. However, due to the lack of structure information, the exact relation between base bias and the structure and function of these proteins is still not clear.

CONCLUSION

In this paper, the base frequency at the second codon position of the 3839 ORFs in *V. cholerae* genome is analyzed. It is shown that according to the base content at this codon site, the ORFs can be divided into two clusters. ORFs in the smaller cluster usually contain significantly higher T frequency at the second codon position than A at this site. For the two clusters of ORFs, there are significant differences between the 18 of the 20 amino acids in the encoding proteins. At the same time, the two clusters of ORFs are also significantly different in their functions. It is hoped that this work could generate some clues toward understanding the genome organization, as well as the function of the genes of this pathogenic bacterium.

ACKNOWLEDGMENTS

Professor C-T. Zhang and Professor Z-P. Feng are thanked for their support in this work. The present study was supported in part by the 973 Project of China (Grant G1999075606).

REFERENCES

1. Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146,** 1–21.
2. Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13–34.
3. Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991) Evidence of horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222,** 851–856.
4. Lawrence, J. G., and Ochman, H. (1997) Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.* **44,** 383–397.
5. Karlin, S., Mrazek, J., and Campbell, A. M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.* **29,** 1341–1355.
6. Mathe, C., Peresetsky, A., Dehais, P., van Montagu, M., and Rouze, P. (1999) Classification of *Arabidopsis thaliana* gene sequences: Clustering of coding sequences into two groups according to codon usage improves gene prediction. *J. Mol. Biol.* **285,** 1977–1991.
7. Sharp, P. M., and Lloyd, A. T. (1993) Regional base composition variation along yeast chromosome III: Evolution of chromosome primary structure. *Nucleic Acids Res.* **21,** 179–183.
8. Chiapello, H., Ollivier, E., Landès-Devauchelle, C., Nitschké, P., and Risler, J.-L. (1999) Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Res.* **27,** 2848–2851.
9. McInerney, J. O. (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* **95,** 10698–10703.
10. Lafay, B., Lloyd, A. T., McLean, M., Devine, K. M., Sharp, P. M., and Wolfe, K. H. (1999) Proteome composition and codon usage in spirochaetes: Species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* **27,** 1642–1649.
11. Oresic, M., and Shalloway, D. (1998) Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* **281,** 31–48.
12. Xie, T., and Ding, D. (1998) The relationship between synonymous codon usage and protein structure. *FEBS Lett.* **399,** 78–82.
13. Gupta, S. K., Majumdar, S., Bhattacharya, T. K., and Ghosh, T. C. (2000) Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.* **269,** 692–696.
14. Zhang, C. T., and Zhang, R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.* **19,** 6313–6317.
15. Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.* **11,** 283–290.
16. Frank, G. K., and Makeev, V. Ju. (1997) G and T nucleotide contents show specie-invariant negative correlation for all three codon positions. *J. Biomol. Struct. Dyn.* **14,** 629–639.
17. Mrazek, J., and Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* **95,** 3720–3725.
18. Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M. R., and Cebrat, S. (1999) Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res.* **27,** 3503–3509.
19. Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M. R., and Cebrat, S. (2000) Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J. Theor. Biol.* **202,** 305–314.
20. Taylor, F., and Coates, D. (1989) The code within the codons. *Biosystems* **22,** 177–187.
21. Siemion, I., and Siemon, P. (1994) The informational context of the third base in amino acid codons. *Biosystems* **33,** 139–148.
22. Chiusano, M. L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., and Bernardi, G. (2000) Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* **261,** 63–69.
23. Heidelberg, J. F., Eisen, J. A., Neison, W. C., *et al.* (32 coauthors) (2000) DNA sequence of both cholera pathogen *Vibrio cholerae. Nature* **406,** 477–483.
24. Wang, J., and Zhang, C.-T. (2001) Analysis of the codon usage pattern in the *Vibrio cholerae* genome. *J. Biomol. Struct. Dyn.*, in press.
25. Hartigan, J. A., and Wong, M. A. (1979) A *k*-means clustering algorithm. *Appl. Stat.* **28,** 100–108.
26. Murray, E. E., Lotzer, J., and Eberle, M. (1989) Codon usage in plant genes. *Nucleic Acids Res.* **17,** 477–497.
27. Ikemura, T., and Wada, K. (1991) Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers: Relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res.* **19,** 4333–4339.