CUCG: A non-redundant codon usage database from complete genomes

A non-redundant codon usage database has been developed from the complete genomes of 17 organisms. GC percentage at the coding region as well as the three different codon positions were tabulated for each organism. Relative synonymous codon usage (RSCU) values for each organism were also included in this database. The World Wide Web provides an user-friendly interface for this database. The dataset of all the genomes are available at http://www.boseinst.ernet.in/dic/CUCG.html.

It is well known that the choice of synonymous codon usage is not random among the different organisms and codon usage patterns generally differ significantly from organism to organism[1-3]. It has been observed that the differential pattern of codon usage among different organisms may have some regulatory roles for the expression of some specialized genes[4]. The diverse pattern of codon usage among different genes may result from (i) diversity in the (G + C)% at the third codon position among genes[5,6]; (ii) translational selection such as shortage in homologous tRNA molecules[7]; (iii) overall base composition of the genes[8]; (iv) different functional constraints on proteins[9]; and (v) differences in expression level of the genes[10]. Research on codon usage pattern of different organisms may shed light on many interesting features hidden in the genome. With this aim codon usage tabulated from the *GenBank* genetic sequence data (CUTG) was developed by Maruyama *et al.*[11] in 1986 and updated regularly in subsequent years[12]. CUTG is a huge database comprising 7434 organisms, but is redundant. For example, 11,796 genes in *Saccharomyces cerevisiae* and 8894 genes in *Bacillus subtilis* are mentioned in the database. Actually there are about 6100 genes in *S. cerevisiae* and 4100 in *B. subtilis*. This type of redundancy was observed for all the organisms under study. Any analysis by using this database may lead to wrong interpretation. The number of publicly available complete genomes is increasing at a fast pace and it is time to make a reliable non-redundant database with these complete genomes. With this motivation we had undertaken the project and developed a non-redundant codon usage database from complete genomes called CUCG (codon usage from complete genomes).

Complete genomes of 17 organisms were downloaded from ncbi.nlm.nih.gov/genbank/genomes using anonymous ftp. For each individual organism we have extracted the coding sequences by our own program developed in C^{++}. We have not made any attempt to remove the ORFs of unknown functions. Frequencies for each of the 17 organisms have been given in the GCG (Genetics Computer Group) format. GC percentage of the whole coding sequence as well as at the first, second and third codon positions has also been appended at the end of each individual frequency table. The graphical representations of (G + C)% at the three different codon-positions for each individual organism were also included in the database. Table 1 shows one representative example of codon usage table. In addition, we have also calculated the RSCU table for each individual organism. RSCU value is defined as the observed frequency of a codon divided by the expected frequency if all the synonyms for that amino acid were used equally.

Table 1. Codon usage table of *Bacillus subtilis*

Organism name: *Bacillus subtilis*; Subkingdom: Eubacteria
Total CDS: 4100

Amino acid	Codon	Number	/1000	Fraction
Gly	GGG	13670.00	11.20	0.16
Gly	GGA	26381.00	21.62	0.31
Gly	GGT	15457.00	12.67	0.18
Gly	GGC	28493.00	23.35	0.34
Glu	GAG	28211.00	23.12	0.32
Glu	GAA	59808.00	49.02	0.68
Asp	GAT	40291.00	33.02	0.64
Asp	GAC	22699.00	18.61	0.36
Val	GTG	21585.00	17.69	0.26
Val	GTA	16296.00	13.36	0.20
Val	GTT	23440.00	19.21	0.28
Val	GTC	21143.00	17.33	0.26
Ala	GCG	24574.00	20.14	0.26
Ala	GCT	23062.00	18.90	0.25
Ala	GCA	26416.00	21.65	0.28
Ala	GCC	19342.00	15.85	0.21
Arg	AGG	4788.00	3.92	0.10
Arg	AGA	13077.00	10.72	0.26
Ser	AGT	8096.00	6.64	0.11
Ser	AGC	17226.00	14.12	0.23
Lys	AAG	25647.00	21.02	0.30
Lys	AAA	60072.00	49.24	0.70
Asn	AAT	27137.00	22.24	0.57
Asn	AAC	20861.00	17.10	0.43
Met	ATG	32918.00	26.98	1.00
Ile	ATA	11517.00	9.44	0.13
Ile	ATT	45181.00	37.03	0.50
Ile	ATC	32872.00	26.94	0.37

Contd. . .

Table 1. (*Contd*).

Amino acid	Codon	Number	/1000	Fraction
Thr	ACG	17693.00	14.50	0.27
Thr	ACT	10620.00	8.70	0.16
Thr	ACA	27117.00	22.23	0.41
Thr	ACC	10497.00	8.60	0.16
Trp	TGG	12571.00	10.30	1.00
End	TGA	965.00	0.79	0.24
Cys	TGT	4429.00	3.63	0.45
Cys	TGC	5322.00	4.36	0.55
End	TAG	591.00	0.48	0.14
End	TAA	2542.00	2.08	0.62
Tyr	TAT	27650.00	22.66	0.65
Tyr	TAC	14673.00	12.03	0.35
Leu	TTG	18745.00	15.36	0.16
Leu	TTA	23338.00	19.13	0.20
Phe	TTT	37445.00	30.69	0.68
Phe	TTC	17253.00	14.14	0.32
Ser	TCG	7717.00	6.33	0.10
Ser	TCT	15615.00	12.80	0.20
Ser	TCA	18053.00	14.80	0.24
Ser	TCC	9757.00	8.00	0.13
Gln	CAG	22750.00	18.65	0.27
Gln	CAA	23889.00	19.58	0.28
His	CAT	18610.00	15.25	0.67
His	CAC	9019.00	7.39	0.33
Leu	CTG	28295.00	23.19	0.24
Leu	CTA	6030.00	4.94	0.05
Leu	CTT	28226.00	23.14	0.24
Leu	CTC	13232.00	10.85	0.11
Pro	CCG	19421.00	15.92	0.43
Pro	CCT	12824.00	10.51	0.29
Pro	CCA	8541.00	7.00	0.19
Pro	CCC	4001.00	3.28	0.09

Coding GC 44.26%, 1st letter GC 52.3%, 2nd letter GC 35.91%, 3rd letter GC 44.57%.

Table 2. Relative synonymous codon usage of *Bacillus subtilis*

Organism name: *Bacillus subtilis*; Subkingdom: Eubacteria
Total CDS: 4100

AA	Codon	RSCU	AA	Codon	RSCU
Phe	TTT	1.37	Ser	TCT	1.22
	TTC	0.63		TCC	0.77
Leu	TTA	1.19		TCA	1.41
	TTG	0.95		TCG	0.60
Tyr	TAT	1.30	Cys	TGT	0.91
	TAC	0.70		TGC	1.09
ter	TAA	0.00	ter	TGA	0.00
ter	TAG	0.00	Trp	TGG	1.00
Leu	CTT	1.43	Pro	CCT	1.14
	CTC	0.67		CCC	0.37
	CTA	0.31		CCA	0.76
	CTG	1.44		CCG	1.73
His	CAT	1.35	Arg	CGT	1.08
	CAC	0.65		CGC	1.23
Gln	CAA	1.02		CGA	0.60
	CAG	0.98		CGG	0.94
Ile	ATT	1.51	Thr	ACT	0.65
	ATC	1.10		ACC	0.64
	ATA	0.39		ACA	1.64
Met	ATG	1.00		ACG	1.07
Asn	AAT	1.13	Ser	AGT	0.64
	AAC	0.87		AGC	1.35
Lys	AAA	1.40	Arg	AGA	1.55
	AAG	0.60		AGG	0.59
Val	GTT	1.14	Ala	GCT	0.99
	GTC	1.02		GCC	0.83
	GTA	0.79		GCA	1.13
	GTG	1.05		GCG	1.05
Asp	GAT	1.28	Gly	GGT	0.74
	GAC	0.72		GGC	1.35
Glu	GAA	1.36		GGA	1.26
	GAG	0.64		GGG	0.65

RSCU values are useful in comparing codon usage among genes and one representative example of RSCU values is shown in Table 2.

1. Grantham, R., Gautier, C., Mercier, R. and Pave, A., *Nucleic Acids Res.*, 1980, **8**, r49–r62.
2. Wada, K., Aota, S., Gojobory, T. and Ikemura, T., *Nucleic Acids Res.*, 1990, **18**, r2367–r2411.
3. Ellis, J., Morrison, D. A. and Kalinna, B., *Parasitol. Res.*, 1995, **81**, 388–393.
4. Saier, M. H. Jr., *FEBS Lett.*, 1995, **362**, 1–14.
5. Ikemura, T., *Mol. Biol. Evol.*, 1985, **2**, 13–24.
6. Alvarez, F., Robello, C. and Vignali, M., *Mol. Biol. Evol.*, 1994, **11**, 790–802.
7. Looyd, A. T. and Sharp, P. M., *Nucleic Acids Res.*, 1992, **20**, 5289–5295.
8. Ellis, J. T. and Morrison, D. A., *Parasitology*, 1995, **110**, 53–60.
9. Barrai, I., Scapoli, C., Nesti, C., Poli, G., Gambari, R. and Beretta, M., *J. Theor. Biol.*, 1994, **166**, 331–337.
10. Pouwels, P. H. and Leunissen, J. A., *Nucleic Acids Res.*, 1994, **22**, 929–936.
11. Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T., *Nucleic Acids Res.*, 1986, **14**, r151–r197.
12. Nakamura, Y., Gojobori, T. and Ikemura, T., *Nucleic Acids Res.*, 1999, **27**, 292.

ACKNOWLEDGEMENT. This work was supported by the Department of Biotechnology, Govt of India.

S. K. GUPTA
T. C. GHOSH

Distributed Information Centre,
Bose Institute,
P-1/12, CIT Scheme, VII M,
Calcutta 700 054, India