Codon usage tabulated from the international DNA sequence databases; its status 1999

Yasukazu Nakamura*, Takashi Gojobori[1] and Toshimichi Ikemura[1]

Laboratory of Gene Structure 2, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan and [1]National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received October 5, 1998; Accepted October 8, 1998

ABSTRACT

Frequencies for each of the 206 526 complete protein-coding genes (CDS's) have been compiled from taxonomical divisions of the GenBank DNA sequence database. The sum of the codon use of 7434 organisms has also been calculated. These data files can be obtained from anonymous ftp sites of DDBJ, DISC and EBI. The list of the codon usage of genes in an organism as well as the sum of the codon usage of the organism was made searchable by the name of organism through a web site http://www.dna.affrc.go.jp/~nakamura/CUTG.html

INTRODUCTION

The choice among synonymous codons within a genome is not random. Among bacterial genes, there is a major trend of codon choice pattern. By measuring the transfer RNA content of a cell, it has been shown that the codon usage trend is highly correlated to the isoaccepting transfer RNA population of individual organisms. It has also been found that the extent of codon bias for each gene is related to the protein production level of each gene (1,2).

In higher organisms, such as mammals, codon usage among genes is highly variable. Codon choice patterns mainly reflect the G+C content of the whole genome or local characteristics, namely GC mosaic or isochore (1,3). Research of the intra-species variations of codon usage may provide an interesting line of investigation regarding the evolution of the genome.

To evaluate codon usage for each gene and/or codon choice trend(s) for each genome, in 1986 we began to compile codon usage of protein genes contained within the international DNA sequence database (4). We named the database CUTG (codon usage tabulated from GenBank). The basic aim of the database is to provide an electronic dataset for codon usage-based analyses. Since each codon usage for a protein-coding gene is compiled as a simple double-lined entry, it is easy to import worksheets or to parse and calculate with computer languages such as C or Perl.

DESCRIPTION OF THE DATABASE

CUTG consists of lists of the codon usage of genes and the sum of codon use for each organism. As of September 1998, CUTG contains 206 526 genes for 7434 organisms. The database has been compiled using the nucleotide sequence obtained from the latest major release of the GenBank sequence database (5). The divisions representing taxonomical collection were used.

In selecting protein-coding sequences we used the annotations from feature tables of the GenBank flat file. Partially-sequenced protein genes were not included in the compilation. Codons that contained one or more letters representing ambiguous bases were excluded from the count. The data structure for each file is the same as in the previous compilation and described in the CODON_LABEL file on distribution sites.

DISTRIBUTION AND ACCESS

A complete form of the database is available from the following URLs:

URLs:	(i) DDBJ	ftp://ftp.nig.ac.jp/pub/db/codon/current/
	(ii) DISC	ftp://ftp.dna.affrc.go.jp/pub/codon/current/
	(iii) EBI	ftp://ftp.ebi.ac.uk/pub/databases/cutg/

Files named gb***.codon, where the '***' is a division name in lower case letters (e.g., bct; pri1 and pri2 is combined as pri), list the codon use in each gene registered in the GenBank flat files. An entry for a gene has two lines. The first line consists of the following information delineated by a backslash which is extracted from the feature table for defining each protein coding sequence. In the 'species' directory, there are codon usage files collected for each organism. The file name consists of the Latin name of the species which is concatenated using under bar, dot and division name (e.g., *Arabidopsis thaliana*, file name for species is 'Arabidopsis_thaliana.pln').

A most user-friendly interface to use interactively with CUTG is to access the World-Wide Web server on DISC. A dataset for each organism is made searchable through the site: http://www.dna.affrc.go.jp/~nakamura/CUTG.html

ACKNOWLEDGEMENTS

We wish to thank Dr Y. Ugawa at the DNA Information and Stock Center, National Institute of Agrobiological Resources for his help in constructing and distributing the database. This work was supported in part by a grant-in-aid for databases from the Ministry of Education, Science, Sports and Culture of Japan. Y.N. is supported by the Kazusa DNA Research Institute Foundation.

REFERENCES

1 Ikemura,T. (1985) *Mol. Biol. Evol*., **2**, 13–34.
2 Ikemura,T. (1981) *J. Mol. Biol*., **146**, 1–21.
3 Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science*, **228**, 953–958.
4 Maruyama,T., Gojobori,T., Aota,S. and Ikemura,T. (1986) *Nucleic Acids Res*., **14**, r151–197.
5 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B.A. and Wheeler,D.L (1999) *Nucleic Acids Res* **27**, 12–17.

*To whom correspondence should be addressed. Tel: +81 438 52 3935; Fax: +81 438 52 3934; Email: ynakamu@kazusa.or.jp