

Codon usage tabulated from the international DNA sequence databases

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

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

Received October 6, 1997; Accepted October 8, 1997

ABSTRACT

CUTG (codon usage tabulated from GenBank) is a comprehensive database for codon usage. The codon usage for each full-length protein gene has been calculated using the nucleotide sequence obtained from GenBank sequence database. The sum of the codon use of each organism has been also calculated. The data files can be obtained from anonymous ftp sites of DDBJ, DISC and EBI. The list of codon usage of genes in organisms was made searchable by name of organism through a web site http://www.dna.affrc.go.jp/~nakamura/CUTG.html The compilation is synchronized with major release of GenBank.

CUTG consists of lists of the codon usage of genes and the sum of codon use for each organism. In September 1997, CUTG contained 155 623 genes for 6048 organisms. The database has been compiled using the nucleotide sequence obtained from the latest major release of GenBank sequence database (1). The divisions which have been used are pri (primate), rod (rodent), mam (other mammalian), vrt (other vertebrate), inv (invertebrate), pln (plant), bct (bacterial), vrl (viral) and phg (phage). Other divisions that do not represent taxonomical collection (such as sts for STS or syn for synthesized sequences) have been excluded from the compilation. In selecting protein coding sequences we relied on the feature tables of GenBank flat file. Codons that contain one or more letter for ambiguous bases were excluded from the count. Partially sequenced protein genes were not compiled.

Files of the database are available by anonymous ftp. Files named gb***.codon, where the '***' is a division name in GenBank (e.g. bct), 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 following information separated by backslash which is extracted from feature table for defining each CDS (protein coding sequence). If a LOCUS contains more than one gene, the symbol # followed by a number is added after the LOCUS name; the numbers represent the order of the CDS registered in the feature table. The second line consists of the count of codons in the CDS. The order of the codons in the table is the same as in the previous compilation (2) and described in the CODON_LABEL file.

To show the characteristics of codon use of a wide range of species, as well as viruses and organella, the codon use in each organism was summed up. Files named gb***.spsum list the sum of numbers of codon use in each organism. An entry for an organism has two lines. The first line consists of latin name of the organism and number of CDS used in summing up. The second line consists of the sum of codons for the organism and its order is the same as gb***.codon files.

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

(i) DDBJ (DNA Data Bank of Japan, National Institute of Genetics, Mishima Japan) ftp://ftp.nig.ac.jp/pub/db/codon/current/
(ii) DISC (DNA Information and Stock Center, National Institute of Agrobiological Resources, Tsukuba, Japan) ftp://ftp.dna.affrc.go.jp/pub/codon/current/
(iii) EBI (European Bioinformatics Institute, Cambridge, UK) ftp://ftp.ebi.ac.uk/pub/databases/cutg/

Split file for each organism is made searchable by latin name of the organism through a site http://www.dna.affrc.go.jp/~nakamura/CUTG.html . Comments on the database can be sent to cutg@lab.nig.ac.jp.

ACKNOWLEDGEMENTS

We wish to thank Dr Y.Ugawa at the DNA Information and Stock Center, National Institute of Agrobiological Resources for 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 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res.* **25**, 1–6 [see also this issue (1998) *Nucleic Acids Res.* **26**, 1–7].
2 Nakamura,Y., Gojobori,T. and Ikemura,T. (1997) *Nucleic Acids Res.* **25**, 244–245.

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