Codon usage tabulated from the international DNA sequence databases

Yasukazu Nakamura, Ken-nosuke Wada¹, Yoshiko Wada, Hirofumi Doi², Shigehiko Kanaya³, Takashi Gojobori and Toshimichi Ikemura*

National Institute of Genetics and The Graduate University for Advanced Studies, Mishima, Shizuoka 411, Japan, ¹ATR Limited, Seika, Kyoto 619-02, Japan, ²Fujitsu Laboratories Ltd, Chiba 261, Japan and ³Yamagata University, Yonezawa, Yamagata 992, Japan

Received August 31, 1995; Revised and Accepted October 4, 1995

ABSTRACT
Codon usage in 87 602 genes has been calculated using the nucleotide sequence data obtained from the GenBank Genetic Sequence Data Bank (Release 90.0; September 1995). The database is called the CUTG Database; the complete form of the database can be obtained by anonymous ftp from DDBJ and a part of the database, which lists the frequency of codon use in each organism, is made searchable through our World Wide Web server.

SOURCE AND METHODS
Codon usage in individual genes has been calculated using the nucleotide sequence data obtained from the GenBank Genetic Sequence Data Bank (Release 90.0; September 1995). The compilation of codon usage is synchronized with each major release of GenBank. The resulting database is called the CUTG database (1–5).

In selecting protein coding sequences we relied on the FEATURES tables of GenBank, and only complete genes without unambiguous bases were used in the analysis. In GenBank, a group of consecutive genes whose entire region had been sequenced were registered under one LOCUS name. To distinguish the different genes belonging to a single LOCUS, the symbol # followed by a number is added after the LOCUS name; the numbers represent the order of the CDS registered in the FEATURES table of GenBank. When introns of a gene have not been completely sequenced, some of its exons are registered in separate entries (LOCUS) in GenBank. These exons, belonging to the same gene but having different LOCUS names, were combined into one entry and the first LOCUS name is added.

For the biological significance of codon usage, see Ikemura (6) and Aota and Ikemura (7,8).

FILES
Files of the present database, containing codon usage of 87 602 CDSs of 4805 species, are available by anonymous ftp from DDBJ. Files named as gb***.codon list the codon use in each gene registered in the GenBank Sequence files (gb***.seq). The LOCUS names given in GenBank were used to designate individual genes. Each LOCUS name is followed by fields of information extracted from the FEATURES of each CDS for defining each open reading frame analyzed here. The order of the codons in the table is the same as the previous compilation (see the CODON_LABEL file or REFERENCES).

DISTRIBUTION AND ACCESS
Complete form of the database is available by anonymous ftp from DDBJ:
The file README contains the latest information on the database in plain text format.
The frequencies of codon use in 461 organisms for which >20 genes are available can be accessed on the following WWW server:
Comments on the database can be sent to cutg@lab.nig.ac.jp by e-mail.

ACKNOWLEDGEMENTS
The authors wish to thank Dr Y. Ugawa at the DNA Information and Stock Center, National Institute of Agrobiological Resources, for help in constructing the database. This work was supported by a grant-in-aid for creative basic research (Human Genome Project) and for scientific research on a priority area (Genome Informatics) and by a grant-in-aid of scientific research from the Ministry of Education, Science and Culture of Japan.

* To whom correspondence should be addressed
Figure 1. Snapshot of the CUTG home page.

REFERENCES