

Codon usage tabulated from international DNA sequence databases: status for the year 2000

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 6, 1999; Accepted October 8, 1999

ABSTRACT

The frequencies of each of the 257 468 complete protein coding sequences (CDSs) have been compiled from the taxonomical divisions of the GenBank DNA sequence database. The sum of the codons used by 8792 organisms has also been calculated. The data files can be obtained from the anonymous ftp sites of DDBJ, Kazusa and EBI. A list of the codon usage of genes and the sum of the codons used by each organism can be obtained through the web site http://www.kazusa.or.jp/codon/ . The present study also reports recent developments on the WWW site. The new web interface provides data in the CodonFrequency-compatible format as well as in the traditional table format. The use of the database is facilitated by keyword based search analysis and the availability of codon usage tables for selected genes from each species. These new tools will provide users with the ability to further analyze for variations in codon usage among different genomes.

DESCRIPTION

We have been compiling the codon usage of all the full-length protein gene entries in the international DNA sequence databases. The compiled files are now freely available through the internet. The purpose of the database designated CUTG is to provide an electronic dataset for codon usage-based analyses. CUTG consists of lists of the codon usage of genes and the sum of codon use by each organism. As of September 1999, CUTG will contain 257 468 genes from 8792 organisms. The database has been constructed from the nucleotide sequences obtained from the latest major release of the GenBank sequence database (1). The strategy used for data collection can be examined by following the URLs listed in the following section or by studying the supplementary material accompanying this publication in NAR Online.

AVAILABILITY

The authors recommend that the database be accessed through the WWW server at Kazusa DNA Research Institute, which provides a user-friendly interface for interactive access: http://www.kazusa.or.jp/codon/

The database displays codon usage in a format compatible with that of CodonFrequency output in the GCG Wisconsin Package™. Thus, users who have the GCG package in their local environment can do further analyses with the files generated by the database. Also, for each species there is a new query box to search for information in the comments of each gene. The user can choose complete protein coding sequences (CDSs) by keyword and then make codon usage tables from the selected genes. This tool provides users with the ability to analyze for intra-species variation in codon usage. For example, it has been reported that protein production levels can be predicted from the complete genome sequences of microbes using the codon usage biases compiled from ribosomal protein genes (2).

The complete dataset of CUTG is available through the following URLs:

Kazusa	ftp://ftp.kazusa.or.jp/pub/codon/current/
DDBJ	ftp://ftp.nig.ac.jp/pub/db/codon/current/
EBI	ftp://ftp.ebi.ac.uk/pub/databases/cutg/

In August 1999, the construction and primary distribution site of the database was moved to Kazusa DNA Research Institute from the DNA Information and Stock Center. Descriptions of the files are maintained as README files through the URLs.

SUPPLEMENTARY MATERIAL

See Supplementary Material available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported in part by a grant-in-aid for databases from the Ministry of Education, Science, Sports and Culture of Japan. The research activity of Y.N. was supported by the Kazusa DNA Research Institute Foundation.

REFERENCES

1. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouelette,B.F.F., Rapp,B.A. and Wheeler,D. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
2. Nakamura,Y and Tabata,S. (1997) *Microbial Comp. Genomics*, **2**, 299–312.

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