BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 4 1998
Pages 372–373

GCUA: General Codon Usage Analysis

James O. McInerney

Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Received on November 4, 1997; revised on November 24, 1997; accepted on November 28, 1997

Abstract

Summary: *The program General Codon Usage Analysis (GCUA) has been developed for analysing codon and amino acid usage patterns.*
Availability: *ftp://ftp.nhm.ac.uk/pub/gcua. Freely available for academic use, commercial users should contact the author.*
Contact: *J.McInerney@nhm.ac.uk*

The analysis of codon usage patterns can be traced back to when the first molecular sequence databases were being collated (Grantham *et al.*, 1981). Since then, a great many different causes and consequences of codon usage variation have been identified [see Sharp *et al.* (1993) for a review]. Amino acid usage also varies between proteins and this variation has also been shown to correlate with the properties of the proteins (Lobry and Gautier, 1994). Codon usage data are used as a guide to direct back-translation of protein sequences to their probable DNA sequences, to identify protein-coding regions of DNA (Fickett, 1982) and to identify regions that probably do not encode a protein (Lloyd and Sharp, 1993). In *Escherichia coli*, an analysis of codon usage patterns was used to compartmentalize open reading frames into three classes, based on expression levels and recentness of acquisition by the genome (Médigue *et al.*, 1991).

A program (GCUA) has been developed for analysing codon and amino acid usage patterns. The program was written using a standard subset of the ANSI C programming language. The interface is composed of a hierarchical menu-driven system, quite similar to ClustalW (Thompson *et al.*, 1994).

GCUA requires a dataset in FastA format (Pearson and Lipman, 1988). The datafile may contain one or more DNA sequences with the first nucleotide corresponding to the first site of a codon triplet. It is assumed that all sequences in the dataset are to be used in the analysis.

Codon and amino acid usage data are collected for all the sequences in the datasets, and data for each individual sequence can be printed either to the screen or to a file. Codon usage values are described either in terms of *N*, the number of times the codon is observed, or RSCU, the relative synonymous codon usage value. RSCU values are a reflection of how often a particular codon is used relative to the expected number of times that codon would be used in the absence of codon usage bias. Amino acid usage is expressed in terms of the percentage contribution made by each amino acid. The base composition of the dataset can also be analysed for different codon positions, including the base composition at the third position of codons for which there is a synonymous alternative.

In order to evaluate codon and amino acid usage variation, multivariate analysis options are available. Correspondence analysis (CA) (Greenacre, 1984) is the most popular and appropriate multivariate analysis method for contingency table data such as codon usage values. The program also has some principal components analysis (PCA) methods implemented for comparative purposes. CA can identify the major sources of variation in the dataset. The output from a CA can be used to evaluate other aspects of the genes, such as base composition, expressivity, aromaticity, location on the genome, etc. In the event that the user wishes to perform a more in-depth multivariate analysis, there is an option of writing the data to disk in ADE-compatible format (Thioulouse *et al.*, 1997).

A distance measure has also been incorporated into the program. The distance measure, based on the differences in RSCU values, is given in equation (1).

$$D_{jk} = \sum_{i=1}^{n} \frac{abs(RSCU_{ji} - RSCU_{ki})}{n} \qquad (1)$$

where n is the number of synonymously degenerate codons for that particular genetic code, ji represents codon i on sequence j, and ki represents codon i on sequence k. This distance metric is not based on a substitution model, rather it groups the sequences together on the basis of similarity of RSCU values. A hierarchical tree-like representation of these data can be produced using either the FITCH, KITCH or NEIGHBOR programs from the PHYLIP package (Felsenstein, 1993), or can be imported into PAUP*v4.0 (Swofford, 1993).

The program is available at ftp://ftp.nhm.ac.uk/pub/gcua and is freely available for academic use; commercial users should contact the author.

References

Felsenstein,J. (1993) *PHYLIP: Phylogenetic Inference Package*. University of Washington. Distributed by the author.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*., **10**, 5303–5318.

Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*., **9**, r43–r74.

Greenacre,M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Lloyd,A.T. and Sharp,P.M. (1993) Synonymous codon usage in *Kluyveromyces lactis*. *Yeast*, **9**, 1219–1228.

Lobry,J.R. and Gautier,C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res*., **22**, 3174–3180.

Médigue,C., Rouxel,T., Vigier,P., Hénaut,A. and Danchin,A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol*., **222**, 851–856.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Sharp,P.M., Stenico,M., Peden,J.F. and Lloyd,A.T. (1993) Codon usage: mutational bias, translational selection or both? *Biochem. Soc. Trans*., **21**, 835–841.

Swofford,D.L. (1993) *PAUP: Phylogenetic Analysis Using Parsimony*. Smithsonian Institute.

Thioulouse,J., Chessel,D. and Dolédec,S. (1997) ADE-4: a multivariate analysis and graphical display software. *Stat. Comput*., **7**, 75–83.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.