BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2329–2330
doi:10.1093/bioinformatics/bth238

INCA: synonymous codon usage analysis and clustering by means of self-organizing map

Fran Supek[1,*] and Kristian Vlahoviček[1,2]

[1]Department of Molecular Biology, Division of Biology, Faculty of Science, Zagreb University, Rooseveltov trg 6, 10000 Zagreb, Croatia; [2]Protein Structure and Bioinformatics, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy

Received on November 5, 2003; revised on January 16, 2004; accepted on February 10, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: INteractive Codon usage Analysis (INCA) provides an array of features useful in analysis of synonymous codon usage in whole genomes. In addition to computing codon frequencies and several usage indices, such as 'codon bias', effective Nc and CAI, the primary strength of INCA has numerous options for the interactive graphical display of calculated values, thus allowing visual detection of various trends in codon usage. Finally, INCA includes a specific unsupervised neural network algorithm, the self-organizing map, used for gene clustering according to the preferred utilization of codons.
Availability: INCA is available for the Win32 platform and is free of charge for academic use. For details, visit the web page http://www.bioinfo-hr.org/inca or contact the author directly.
Contact: fsupek@public.srce.hr; kristian@icgeb.org
Supplementary information: Software is accompanied with a user manual and a short tutorial.

Prokaryotes, yeast and to some extent higher eukaryotes, show preference for certain synonymous codons over others, despite all of them coding for the same amino acid. The codon preference may vary greatly among genes in the same organism and between different species. Possible explanations are mutational effects, as observed in the background nucleotide composition; translational selection, which in highly expressed genes favours codons with most abundant corresponding tRNAs; and 'recent' horizontal gene transfer, with genes retaining the sequence characteristics of the donor organism (reviewed in Ermolaeva, 2001). To date, several software packages exist that perform various codon usage-related calculations, e.g. Codon W (Peden, 1999) and GCUA (McInerney, 1998).

As input, INteractive Codon usage Analysis (INCA) accepts complete genome files readily available from the NCBI website and ftp server. Files with extensions '.ffn' and optionally '.ptt' pertaining to a certain genome were used, the former containing hypothetical cDNAs of the organism in the FASTA format, and the latter listing information about the protein products of the genes. The software makes extensive use of the functional categories of the COG database (Tatusov *et al.*, 2000) and presents genes classified by COG category whenever possible. Support for different genetic code translation tables was built into the program.

After loading the input files, user may select a number of genes as the 'reference set' (usually, the reference set contains highly expressed genes). To avoid sampling errors, an option is offered to exclude genes shorter than a specified length from all further calculation. The individual codon frequencies for the whole genome and for certain groups of genes, such as the reference set or clusters generated by the SOM, can be examined side-by-side graphically, as well as saved to or loaded from text files.

A plot with customizable x- and y-axes is used to display all genes as dots, allowing visual detection of various trends in codon usage (Fig. 1a). INCA on many occasions relies on the 'codon bias' statistic (Karlin *et al.*, 1998) to measure the difference in codon usage of one gene group relative to another. Each axis in the plot can be set to show any of the following: (i) codon bias relative to available gene groups (genome, reference set or clusters); (ii) effective number of codons (Wright, 1990); (iii) codon adaptation index (Sharp and Li, 1987); and (iv) other properties, such as gene length, position, G+C content at silent sites and hydrophobicity. A linear correlation coefficient is automatically calculated for chosen properties. Additionally, dots can be coloured by specified criteria, some of which include strand positioning or membership in the COG functional category. Genes may also be binned into an arbitrary number of equal-sized groups with regard to any of the available properties. An alternative to plotting data is a tabular format display; tables are customizable with respect to columns shown, gene ordering or for binning. Plots and tables can be exported to files for publication or for further analysis.

*To whom correspondence should be addresed.

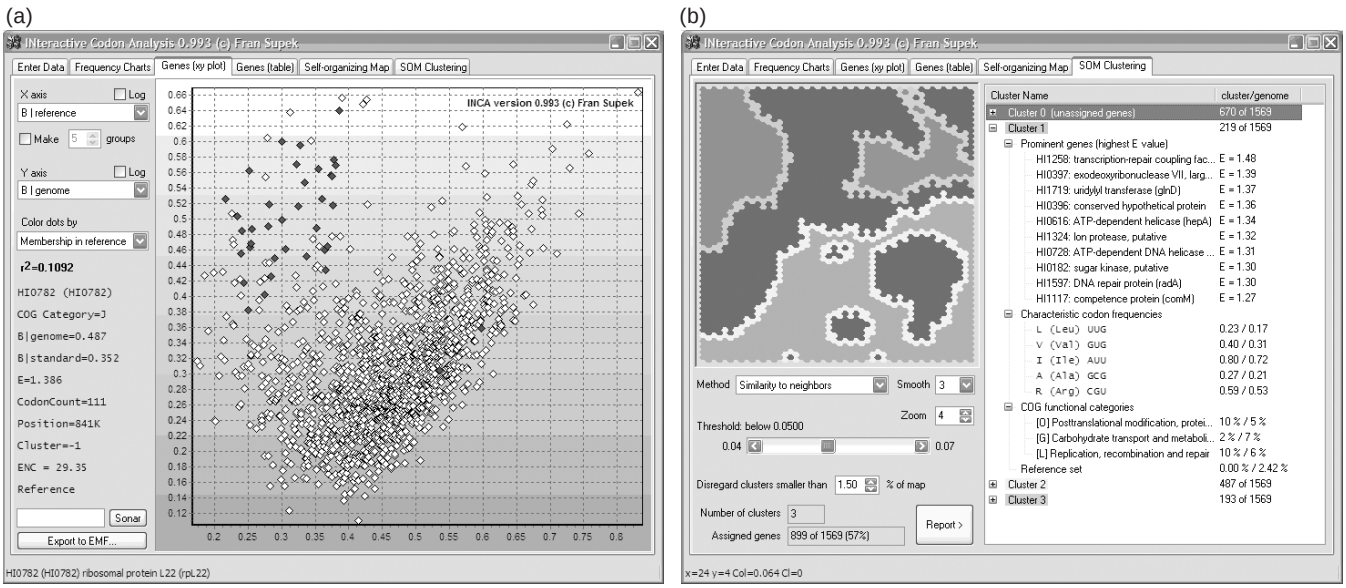

Fig. 1. (**a**) Two-dimensional plot of codon bias with respect to the reference set (gene names contain the word 'ribosomal') versus codon bias with respect to genome in the *Haemorhiza influenzae* genome (accession no. gb:NC_000907). (**b**) SOM cluster plot of the same genome, using default initial SOM settings, 'Difference from genomic bias' method, a 'Smooth' value of 3 and a threshold of 0.1750.

INCA features an unsupervised neural network algorithm, the self-organizing map (SOM), or 'Kohonen map'. SOM is specialized for converting high-dimensional data, such as codon usage frequencies, into easily perceivable two-dimensional maps (reviewed in Kohonen *et al*., 1996). An additional benefit is the ability to cluster genes into groups of similar codon preferences using the displayed pattern (Fig. 1b). Both the settings for SOM training, and the clustering parameters are fully customizable. In addition to the commonly used 'component planes' and 'U-matrix', the software offers some additional methods for visualization and clustering; in particular, it shows gene density in a certain area and the deviation either from genomic codon usage or from a user-defined codon frequency table.

The stochastic component of the network relies on a pseudo-random number generator thus permitting the user to easily reproduce clustering results, provided an equal set of initialization parameters is input to the program. INCA produces a per-cluster report that includes a list of prominent genes, characteristic codon frequencies, and the cluster composition with respect to the COG functional categories and the reference set. Information about gene membership in clusters can also be accessed through the Plot and Table views, aiding in visual determination of cluster composition, and enabling users to process the data further, e.g. by using statistical analysis software or by inclusion in databases.

In conclusion, INCA is a comprehensive codon usage analysis package characterized by extensive interactive graphical capabilities and a user-friendly interface, with possible applications in both research and education.

REFERENCES

Ermolaeva,M.D. (2001) Synonymous codon usage in bacteria. *Curr. Iss. Mol. Biol.*, **3**, 91–97.

Karlin,S., Mrazek,J. and Campbell,A.M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.*, **29**, 1341–1355.

Kohonen,T., Oja,E., Simula,O., Visa,A. and Kangas,J. (1996) Engineering applications of the self-organizing map (Review). *Proc. IEEE*, **84**, 1358–1384.

McInerney,J.O. (1998) GCUA: general codon usage analysis. *Bioinformatics*, **14**, 372–373.

Peden,J.F. (1999) CodonW. PhD Thesis, University of Nottingham.

Sharp,P.M. and Li,W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.

Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.