BIOINFORMATICS APPLICATIONS NOTE
Vol. 21 no. 4 2005, pages 545–547
doi:10.1093/bioinformatics/bti037

Online synonymous codon usage analyses with the ade4 and seqinR packages

D. Charif, J. Thioulouse, J. R. Lobry* and G. Perrière

Laboratoire de Biométrie et Biologie Évolutive—CNRS UMR 5558, and INRIA Helix project, Université Claude Bernard—Lyon I, 43 bd. du 11 Novembre 1918, F-69622 Villeurbanne cedex, France

Received on May 6, 2004; revised on June 22, 2004; accepted on June 23, 2004
Advance Access publication September 16, 2004

ABSTRACT

Summary: Correspondence analysis of codon usage data is a widely used method in sequence analysis, but the variability in amino acid composition between proteins is a confounding factor when one wants to analyse synonymous codon usage variability. A simple and natural way to cope with this problem is to use within-group correspondence analysis. There is, however, no user-friendly implementation of this method available for genomic studies. Our motivation was to provide to the community a Web facility to easily study synonymous codon usage on a subset of data available in public genomic databases.

Availability: Availability through the Pôle Bioinformatique Lyonnais (PBIL) Web server at http://pbil.univ-lyon1.fr/datasets/charif04/ with a demo allowing us to reproduce the figure in the present application note. All underlying software is distributed under a GPL licence.

Contact: http://pbil.univ-lyon1.fr/members/lobry

INTRODUCTION

The use and misuse of correspondence analysis in codon usage studies has recently been reviewed (Perrière and Thioulouse, 2002). A common mistake is to use correspondence analysis on RSCU (relative synonymous codon usage) (Sharp *et al*., 1986) tables to remove the amino acid effect. This is common because correspondence analysis on RSCU is implemented in two popular packages: GCUA (McInerney, 1998) and codonW (available from the URL, http://www/molbio.ox.ac.uk/cu/codonW.html). This is a mistake because it introduces unjustified statistical weights on data, yielding biased results, especially for codon usage in rare amino acids, such as Cysteine, which are analyzed as if they were as well documented as common amino acids (Perrière and Thioulouse, 2002). A well-known solution in the field of statistical ecology is the so-called within-block correspondence analysis (Benzécri, 1983). This approach has been used only recently in the field of genomics (Lobry and Chessel, 2003), and our motivation was then to provide a simple tool to run this method on codon usage data.

METHODS

We used two packages for the R statistical computing environment (Ihaka and Gentleman, 1996): ade4 (Thioulouse *et al*., 1997) and seqinR. R provides a wide variety of statistical and graphical techniques, and is highly extensible. The ade4 package is dedicated to ecological data analysis and implements many multivariate methods. It is available in 'the contributed package' area of all the servers of the CRAN at http://cran.r-project.org. The seqinR package provides an access to the databases structured under the ACNUC model (Gouy *et al*., 1985) and is available at http://pbil.univ-lyon1.fr/software/SeqinR/seqinr_home.php. In order to provide an online solution, we have modified Rweb, a Web interface to R developed by Jeff Banfield available at http://www.math.montana.edu/Rweb.

For the sake of reproducibility, because the genomic repository databases available on the PBIL server (GenBank, EMBL) are updated on a daily basis, we have extracted a subset of GenBank release 139 to build a frozen database. This subset contains all available sequences from *Leishamnia major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. From this frozen database we have extracted complete nuclear coding sequences. Coding sequences with in-frame stop codons or less than 100 codons were discarded. Table of codon counts were then computed to run a within-amino acid correspondence analysis, taking into account the relevant genetic code.

The core of the within-amino acid analysis is expressed by the following R code, extracted from the `within()` function of the ade4 package. This function implements within correspondence analysis (Benzécri, 1983), when supplied with the output of a correspondence analysis.

```
cla.w <- tapply(dudi$lw, fac, sum)
mean.w <- function(x, w, fac, cla.w) {
  z <- x * w
  z <- tapply(z, fac, sum)/cla.w
```

*To whom correspondence should be addressed.

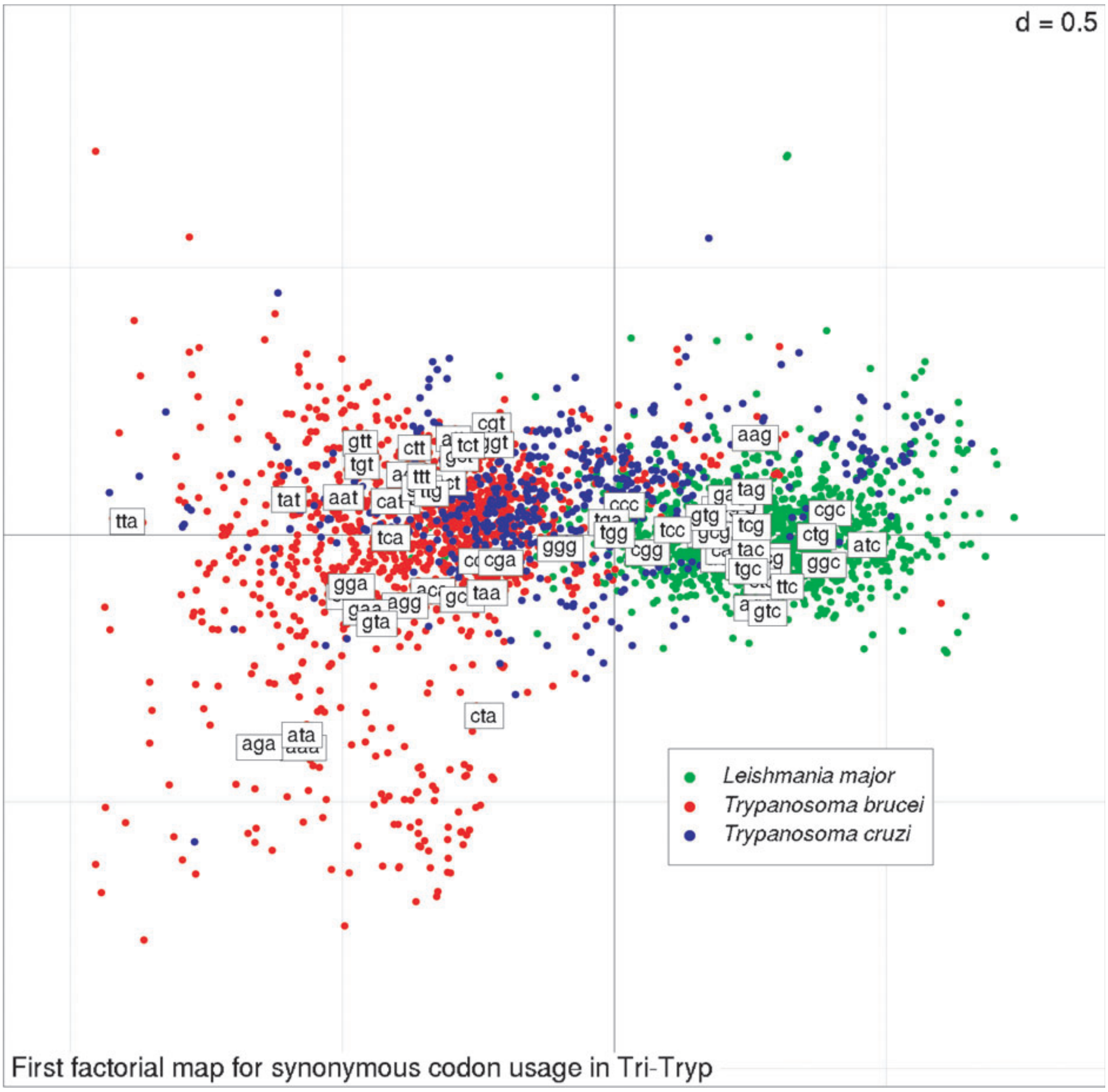

Fig. 1. First factorial map for synonymous codon usage in the coding sequences from the three species under study. This figure is reproducible online at http://pbil.univ-lyon1.fr/datasets/charif04/. The first axis (41% of total variability) is a G+C gradient with AT ending codons more frequent in *T.brucei* and GC ending codons more frequent in *L.major*, the second axis (5% of total variability) is linked to a subset of *T.brucei* coding sequences that are enriched in some specific codons (mainly A-ending versus T-ending codons).

```
  return(z)
}
tabmoy <- apply(dudi$tab, 2,
mean.w, w = dudi$lw, fac = fac,
cla.w = cla.w)
tabwit <- dudi$tab - tabmoy[fac, ]
```

The last line computes a new table, `tabwit`, from the previous one used for simple correspondence analysis, `dudi$tab`, by centering with amino acid. The previous lines show that the means are weighted to take into account the weight of codons in the simple correspondence analysis (`dudi$lw`). Hence, when using a consistent weighting scheme for rows and columns, the total variability in codon usage is the sum of the within and between amino acid variability, that is, the sum of synonymous codon usage and amino acid usage.

RESULTS

An example of the result is given in Figure 1. This figure is reproducible directly from a Web browser without installing any software. The script is easily amendable to fit user's needs. For instance, changing the instruction `choosebank(bank ="trypano")` by `choosebank(bank ="genbank")` will run the same analysis with data from the last update from GenBank. In a similar way, it is straightforward to change the instruction `myquery("tc", "Trypanosoma cruzi")` to study synonymous codon usage in different species.

The limitation of the approach is that it is impossible to study simultaneously synonymous codon usage in species with different genetic codes. There is also a technical trade-off between the convenience for the end-user and interactivity: no installations are needed and the cost is that all computations are run in batch mode on a distant server.

REFERENCES

Benzécri,J.-P. (1983) Analyse de l'inertie intra-classe par l'analyse d'un tableau des correspondances. *Les Cahiers de l'Analyse des Données*, **8**, 351–358.

Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and Di Paola,G. (1985) ACNUC—a portable system for nucleic acid sequence databases: logical and physical designs and usage. *Comput. Applic. Biosci.*, **1**, 167–172.

Ihaka,R. and Gentleman,R. (1996) R: A language for data analysis and graphics. *J. Comp. Graph. Stat.*, **5**, 299–314.

Lobry,J.R. and Chessel,D. (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J. Appl. Genet.*, **44**, 235–261.

McInerney,J.O. (1998) GCUA (General Codon Usage Analysis). *Bioinformatics*, **14**, 372–373.

Perrière,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.*, **30**, 4548–4555.

Sharp,P.M., Tuohy,T.M.F. and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, 5125–5143.

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