

Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases

Hélène Chiapello*, Emmanuelle Ollivier¹, Claudine Landès-Devauchelle¹, Patrick Nitschké¹ and Jean-Loup Rislér¹

INRA, Biologie Cellulaire, Route de Saint Cyr, 78026 Versailles Cedex, France and ¹Université de Versailles, Génome et Informatique, 45 Avenue des Etats-Unis, 78035 Versailles Cedex, France

Received April 14, 1999; Revised and Accepted June 1, 1999

ABSTRACT

In spite of many efforts, the prediction of the location of proteins in eukaryotic cells (cytoplasm, mitochondrion or chloroplast) is still far from straightforward. In some cases (e.g. ribosomal proteins and aminoacyl-tRNA synthetases) both the cytoplasmic proteins and their organellar counterparts are encoded by the nuclear genome. A factorial correspondence analysis of the codon usage in yeast and *Caenorhabditis elegans* shows that the codon usage of those nuclear genes encoding ribosomal proteins or aminoacyl-tRNA synthetases is markedly different, depending on the final location of the proteins (cytoplasmic or mitochondrial). As a consequence, the location of such proteins—whose sequences are now frequently determined by systematic genomic sequencing—can be easily and quickly predicted. A WWW interface has been developed, aimed at providing a user-friendly tool for codon usage pattern analysis. It is available from <http://www.genetique.uvsq.fr/afc.html>

INTRODUCTION

The genetic code is degenerate, and synonymous codons are generally used with unequal frequencies in different genes. Although there exists a species-specific pattern of codon usage (1), there are also differences in codon usage among genes in most species. In some prokaryotes like *Escherichia coli* (2), *Methanococcus jannaschii* and others (3) codon usage variations are correlated with the level of expression of the genes. This is also true, for example, in yeast (4) and in *Caenorhabditis elegans* (5). In addition, a correlation was demonstrated between the abundance of tRNAs and the occurrence of the respective codons in yeast (6). The codon usage in *E.coli* also depends on the origin and pattern of expression of the genes (7,8). In other prokaryotes, like *Mycoplasma genitalium*, the codon usage of a gene seems to be largely determined by its location in the genome and not related to its expression level (9,10). Finally, recent work indicates that the most important

source of variation in codon usage in *Borrelia burgdorferi* is attributable to the disparity in the mutational bias between the leading and the lagging strands of replication (11). Similar analyses on eukaryotic genes confirm the heterogeneity of codon usage in a given species. For example, codon usage variations have been observed in yeast, *Drosophila melanogaster* (1), *C.elegans* (5) and *Arabidopsis thaliana* (12). In some cases, codon usage variations may reflect translational constraints or mutational biases (5,13,14) but the respective influence of these two factors is generally not easy to delineate (15,16). Other constraints may act on codon usage, such as the chromosomal regional bias in nucleotide composition (17) or the exogenous origin of the genes (7). This is particularly relevant in the study of the codon usage of functional families, like aminoacyl-tRNA synthetases (aaRSs) (18).

We show here that a correspondence analysis of the codon usage in yeast and *C.elegans* reveals that those nuclear genes encoding ribosomal proteins (RPs) or aaRSs exhibit markedly different patterns, depending on the final location of the proteins (cytoplasmic or mitochondrial). In *A.thaliana*, the difference is less marked but systematic. This permits an efficient and easy prediction of the cellular location of such proteins.

MATERIALS AND METHODS

The nucleic acid sequences of all the open reading frames (ORFs) from the yeast genome were retrieved from the MIPS server (<http://www.mips.biochem.mpg.de>). Those from *C.elegans* and *A.thaliana* were extracted from the GenBank database (release 107) with the ACNUC retrieval program (19). The lists of ribosomal proteins and aaRSs from yeast were fetched from the YPD server (<http://quest7.proteome.com/YPDhome.html>) and from the Pedant web site (<http://pedant.mips.biochem.mpg.de>).

In the present study, the codon usage of the genes (or ORFs) from yeast, *C.elegans* and *A.thaliana* has been studied by the factorial correspondence analysis (FCA) method with a locally written program that is publicly available through a web interface (<http://www.genetique.uvsq.fr/afc.html>). FCA is a commonly used multivariate statistical approach in codon usage analysis since the pioneering work of Grantham's group (20). Our program comprises three steps. (i) For each gene in a given dataset, the relative frequencies of synonymous codons per

*To whom correspondence should be addressed at: Université de Versailles, Génome et Informatique, 45 Avenue des Etats-Unis, 78035 Versailles Cedex, France. Tel: +33 1 39 25 45 61; Fax: +33 1 39 25 45 69; Email: chiapell@genetique.uvsq.fr

amino acid are calculated—thus they sum up to 1 for all the synonymous codons having the same first two bases. This is similar but not identical to the more classical RSCU index (4). The amino acids Trp and Met are ignored since they are encoded by only one codon. The Stop codons are also disregarded. We chose this index in order to focus strictly on differences between genes in terms of synonymous codon usage, irrespective of the amino acid composition of the encoded protein (admittedly, however, the use of relative codons frequencies will put an unduly heavy weight on rare amino acids such as cysteines). During this first step, the program also calculates for each gene the G+C composition at the third position of the codons—excluding the Met, Trp and Stop codons (this feature was not used in the present study). (ii) The second step is to calculate in a 59-dimensional space a distance between all the gene pairs (or between the codons) based on their codon usage, using the chi-2 distance between genes (or between codons). (iii) The third step enables the visualization of these chi-2 distances by determining those axes along which the projection of the points is most scattered. More details on this now classical method can be found in Hill (21) and Benzecri (22).

To our knowledge, three softwares are freely available that provide comprehensive sets of methods dedicated to correspondence analysis on codon usage: ADE-4 (23), GCUA (24) and codonW (available from <http://www.molbiol.ox.ac.uk/cu/codonW.html>). The ADE-4 package can also be used through a web interface (<http://pbil.univ-lyon1.fr/mva/coa.html>). Our program is presently limited to FCA, which probably makes it easier to use by non-specialists and thus represents hopefully a useful starting point for more elaborate studies. Its web interface provides graphics facilities: many of the output files can be retrieved and directly imported into graphics programs such as Gnuplot or Xgobi while a Java module permits an easy on-line identification of each point in the graphical output. In practice, only two parameters need to be set: the genetic code to be used and the minimum size of coding sequences. Additional information is provided through the on-line help.

RESULTS AND DISCUSSION

Ribosomal proteins and aminoacyl-tRNA synthetases from *Saccharomyces cerevisiae*

Since ribosomal proteins can be composed of less than 100 amino acids, we have analyzed the codon usage of 6068 genes/ORFs from the yeast genome longer than 200 bases—rather than the more usual 300-base threshold. Figure 1a shows the projection of these genes onto the first factorial plane.

The mitochondrial genome in eukaryotic cells, particularly in *S.cerevisiae*, encodes only a small number of proteins. In addition, there are at most only 44 individual mitochondria per yeast cell (25). Hence the overall activity of the mitochondrial ribosomes in protein synthesis is probably much less than that of their cytoplasmic counterparts. This makes it probable that the nuclear genes encoding mitochondrial ribosomal proteins (MRPs) are less expressed than those encoding cytoplasmic ribosomal proteins (CRPs). Since the codon usage of highly and lowly expressed genes is markedly different in yeast (3,4), it was anticipated that the MRP and CRP genes should be differentiated in the FCA. That this is indeed the case is shown

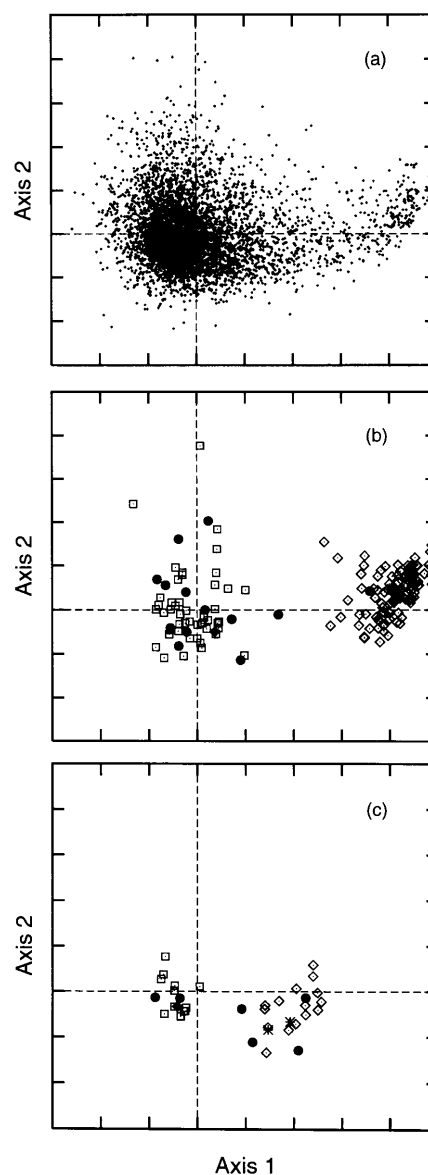


Figure 1. FCA of (a) 6068 genes (or ORFs) from yeast, (b) only those genes encoding RPs and (c) only those genes encoding aaRSa. The proteins that are described as cytoplasmic in the YPD, Pedant and Swissprot databases are indicated by diamond-shaped labels; the mitochondrial proteins are labeled with open squares and the proteins whose cellular location is not given are labeled with filled circles. The genes encoding AlaRS and HisRS are indicated by star-like labels.

in Figure 1b where the genes encoding CRPs and MRPs, respectively, are grouped into two clearly separated clusters. Note that another alternative, non-exclusive explanation for this observed difference in the codon usage of CRP and MRP genes could be that the nuclear MRP genes, being of mitochondrial origin, would have retained some features characteristic of the original endosymbiotic host.

Apart from one single exception (YPR166C), all the RPs that are commented as cytoplasmic in the YPD, Pedant or Swissprot databases are grouped into one single cluster in the FCA (Fig. 1b) and all those that are given as mitochondrial belong to the

Table 1. Prediction of the cellular location of yeast RPs as given by the programs Psort and Mitoprot

| | Psort | Mitoprot |
|-------------------|---|--|
| Cytoplasmic: 121 | Non-mitochondrial: 88 (73%) Mitochondrial: 17 (14%) Ambiguous: 16 (13%) | Non-mitochondrial: 62 (51%) Mitochondrial: 40 (33%) Ambiguous: 19 (16%) |
| Mitochondrial: 49 | Mitochondrial: 19 (39%) Non-mitochondrial: 19 (39%) Ambiguous: 11 (22%) | Mitochondrial: 40 (82%) Non-mitochondrial: 5 (10%) Ambiguous: 4 (8%) |
| Probable: 13 | Non-mitochondrial: 5 Mitochondrial: 7 Ambiguous: 1 | Non-mitochondrial: 5 Mitochondrial: 7 Ambiguous: 1 |

From the Pedant and YDB databases, 121 non-mitochondrial and 49 mitochondrial RPs were extracted. The predictions given by Psort and Mitoprot are reported for each of these two categories.

second cluster. YPR166C is given as cytoplasmic in Swissprot, but not in the list of CRPs reported by Planta and Mager (26). Its position in the FCA confirms that it is probably mitochondrial. The clear and systematic partitioning of the CRPs and MRPs into two clusters makes it tempting to use the codon usage of their genes as a predictive tool for their cellular location.

Thirteen proteins from the yeast genome are labeled as 'probable ribosomal proteins' in the above databanks, on the basis of their exhibiting sequence similarities with other ribosomal proteins. We have attempted to predict their location with the programs MitoProt (27) and Psort (28) (<http://cookie.imcb.osaka-u.ac.jp/nakai/psort.html>) which, *inter alia*, makes use of the method described by Gavel and von Heijne (29). The results are far from clear-cut. Indeed, from a set of RPs whose location is known (26,30), it appears that MitoProt is 82% correct in predicting MRPs as mitochondrial and Psort is 73% correct in predicting CRPs as cytoplasmic (Table 1). However, MitoProt is weak in predicting the location of CRPs and Psort is weak in predicting that of MRPs. As a result, the predictions of the two programs are often contradictory. On the contrary, the position in the FCA of the 13 putative RPs (Fig. 1b) permits to make a quick and probably sound prediction of their location (see <http://www.genetique.uvsq.fr/codon-usage>) if these proteins are indeed ribosomal proteins. It must be stressed that the identification of CRPs (26,31) and MRPs (30,32) is far from straightforward. In particular, many MRPs do not show clear sequence similarities with other (bacterial) CRPs and, consequently, could not be identified by mere sequence comparisons. This is also true for mammalian MRPs (33).

Like the ribosomal proteins, both the cytoplasmic and mitochondrial aaRSs are encoded by the nuclear genome. Here also, the cytoplasmic aaRSs do cluster into one region of the FCA while the mitochondrial aaRSs make another cluster (Fig. 1c). Again, this permits the prediction of the cellular location of those aaRSs from yeast whose location is not given in YPD, Pedant or Swissprot (see Fig. 1c and <http://www.genetique.uvsq.fr/codon-usage>). The prolyl-tRNA synthetase encoded by YER078W is given as cytoplasmic in Swissprot, but its position in the FCA (filled black diamond-shaped label in Fig. 1c) makes it probable that it is in fact mitochondrial. Noteworthy are the cases of AlaRS and HisRS whose cytoplasmic and mitochondrial counterparts are encoded in each case by a single gene (34,35). The positions of

these two aaRSs in the FCA (star-like labels in Fig. 1c) indicate that their coding genes are of nuclear origin and suggest that, in both cases, the original mitochondrial gene has been lost.

A study of *C.elegans* ORFs

A study similar to the previous one was also undertaken on 14 612 genes or ORFs from *C.elegans*, extracted from the GenBank database (all the ORFs that did not begin with an ATG codon and/or that contained in-frame stop codons have been eliminated). Their projection onto the first factorial plane is shown in Figure 2a. Among the 150 ORFs having the highest abscissa, 69 have been (sometimes tentatively) assigned a function in GenBank and encode 30 ribosomal proteins, nine myosins, five histones, five actins, two glyceraldehyde-phosphate dehydrogenases, etc. As previously concluded from the study of 258 ORFs (5), it is thus reasonable to think that this region is again characteristic of highly expressed genes. In the case of RPs (Fig. 2b), only one protein is documented as cytoplasmic (the 60s-L3 protein) and three as mitochondrial in Swissprot. These proteins belong to two different clusters in Figure 2b, the mitochondrial proteins being shifted towards the low abscissa region. The resulting predictions concerning the RPs from *C.elegans* (filled circles in Fig. 2b) can be obtained from <http://www.genetique.uvsq.fr/codon-usage>

In the case of aaRSs, there is only one identified mitochondrial protein. Although it is also well separated from the cytoplasmic aaRSs in the FCA (Fig. 2c), again shifted towards the lowly expressed region, the predictions are at present necessarily more tentative (same URL as above).

Preliminary results in *A.thaliana*

The complete sequencing of the *A.thaliana* genome is under progress and ~50% of the genome is currently available (even if not fully annotated). Using the same method as described before, we analyzed a set of 6009 genes (or ORFs) from *A.thaliana* (not shown). Highly expressed genes such as those encoding histones, glyceraldehyde-3-phosphate dehydrogenase and abundant photosynthetic proteins (CAB, RUBISCO) are clustered in the same region of the first factorial plane, but the positions of the genes encoding ribosomal proteins do not show any clear partitioning between cytosolic and organellar RPs.

In the case of the six aminoacyl-tRNA synthetases so far identified, it appears that a gene encoding a plastidic/mitochondrial aaRS systematically exhibits a lower abscissa than that of the

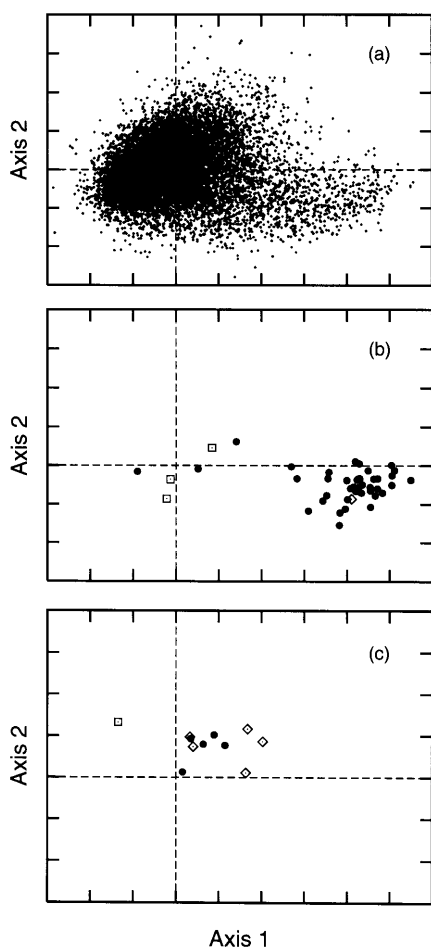


Figure 2. FCA of (a) 14 612 genes (or ORFs) from *C.elegans*, (b) only those genes encoding RPs and (c) only those genes encoding aaRSs. Same conventions as in Figure 1.

gene encoding the cytosolic form of the same aaRS (data not shown; see <http://www.genetique.uvsq.fr/codon-usage>). Like in yeast, the situation may be further complicated by the fact that some genes encode both the cytosolic and organellar forms (36).

Conclusion

The present study shows that the codon usage of those genes encoding cytoplasmic ribosomal proteins, on the one hand, and aminoacyl-tRNA synthetases on the other hand is markedly different from that of the genes encoding their mitochondrial counterparts in yeast and *C.elegans*. This enables a quick and easy prediction of the location of these proteins. The method, however, will hold only for those eukaryotic organisms where the main trend in the codon bias is associated with the expressivity of the genes. This might not be the case, for example, with mammals such as human, mouse, rat and cow (1,37,38). We anticipate that this simple method will prove useful with most (if not all) unicellular eukaryotes but probably not with some (or many) higher organisms where the situation is complicated by the differentiation of the cells.

ACKNOWLEDGEMENT

We thank the reviewers for their valuable comments and suggestions.

REFERENCES

- Sharp,P.M., Cowe,E., Higgins,D.G., Shields,D.C., Wolfe,K.H. and Wright,F. (1988) *Nucleic Acids Res.*, **16**, 8207–8211.
- Gouy,M. and Gautier,C. (1982) *Nucleic Acids Res.*, **10**, 7055–7074.
- Nakamura,Y. and Tabata,S. (1997) *Microb. Comp. Genomics*, **2**, 299–312.
- Sharp,P.M., Tuohy,T.M. and Mosurski,K.R. (1986) *Nucleic Acids Res.*, **14**, 5125–5143.
- Stenico,M., Lloyd,A.T. and Sharp,P.M. (1994) *Nucleic Acids Res.*, **22**, 2437–2446.
- Ikemura,T. (1982) *J. Mol. Biol.*, **158**, 573–597.
- Medigue,C., Rouxel,T., Vigier,P., Hénaut,A. and Danchin,A. (1991) *J. Mol. Biol.*, **222**, 851–856.
- Hénaut,A. and Danchin,A. (1996) In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella, Cellular and Molecular Biology*. ASM Press, Washington DC, pp. 2047–2066.
- McInerney,J.O. (1997) *Microb. Comp. Genomics*, **2**, 1–10.
- Kerr,A.R., Peden,J.F. and Sharp,P.M. (1997) *Mol. Microbiol.*, **25**, 1177–1179.
- McInerney,J.O. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 10698–10703.
- Chiapello,H., Lisacek,F., Caboche,M. and Hénaut,A. (1998) *Gene*, **209**, GC1–GC38.
- Bulmer,M. (1991) *Genetics*, **129**, 897–907.
- Eyre-Walker,A. (1996) *Mol. Biol. Evol.*, **13**, 864–872.
- Pan,A., Dutta,C. and Das,J. (1998) *Gene*, **215**, 405–413.
- Sharp,P.M., Stenico,M., Peden,J.F. and Lloyd,AT. (1993) *Biochem. Soc. Trans.*, **21**, 835–841.
- Bulmer,M. (1990) *Nucleic Acids Res.*, **18**, 2869–2873.
- Diaz-Lazcoz,Y., Hénaut,A., Vigier,P. and Risler,J.L. (1995) *J. Mol. Biol.*, **250**, 123–127.
- Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and di Paola,G. (1985) *Comput. Appl. Biosci.*, **1**, 167–172.
- Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) *Nucleic Acids Res.*, **9**, r43–r74.
- Hill,M.O. (1974) *Appl. Statist.*, **23**, 340–353.
- Benzécri,J.P. (1992) *The Correspondence Analysis Handbook. Statistics: Textbooks and Monographs 125*. Marcel Dekker, New York.
- Thioulouse,J., Chessel,D., Dolédec,S. and Olivier,J.M. (1997) *Stat. Comput.*, **7**, 75–83.
- McInerney,J.O. (1998) *Bioinformatics*, **14**, 372–373.
- Pon,L. and Schatz,G. (1991) In Broach,J.R., Pringle,J.R. and Jones,E.W. (eds), *The Molecular and Cellular Biology of the Yeast Saccharomyces*. Cold Spring Harbor Laboratory Press, New York, pp. 333–406.
- Planta,R.J. and Mager,W.H. (1998) *Yeast*, **14**, 471–477.
- Claros,M.G. and Vincens,P. (1996) *Eur. J. Biochem.*, **241**, 779–786.
- Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.
- Gavel,Y. and von Heijne,G. (1990) *Protein Eng.*, **4**, 33–37.
- Kitakawa,M., Graack,H.R., Grohmann,L., Goldschmidt-Reisin,S., Herfurth,E., Wittmann-Liebold,B., Nishimura,T. and Isono,K. (1997) *Eur. J. Biochem.*, **245**, 449–456.
- Mager,W.H., Planta,R.J., Ballesta,J.G., Lee,J.C., Mizuta,K., Suzuki,K., Warner,J.R. and Woolford,J. (1997) *Nucleic Acids Res.*, **25**, 4872–4875.
- Graack,H.R. and Wittmann-Liebold,B. (1998) *Biochem. J.*, **329**, 433–448.
- Goldschmidt-Reisin,S., Kitakawa,M., Herfurth,E., Wittmann-Liebold,B., Grohmann,L. and Graack,H.R. (1998) *J. Biol. Chem.*, **273**, 34828–34836.
- Chatton,B., Walter,P., Ebel,J.P., Lacroute,F. and Fasiolo,F. (1988) *J. Biol. Chem.*, **263**, 52–57.
- Chiu,M.I., Mason,T.L. and Fink,G.R. (1992) *Genetics*, **132**, 987–1001.
- Mireau,H., Lancelin,D. and Small,I.D. (1996) *Plant Cell*, **8**, 1027–1039.
- Aota,S., Gojobori,T., Ishibashi,F., Maruyama,T. and Ikemura,T. (1988) *Nucleic Acids Res.*, **16**, r315–r402.
- Ikemura,T. (1985) *Mol. Biol. Evol.*, **2**, 13–34.