

## An Evaluation of Measures of Synonymous Codon Usage Bias

Josep M. Comeron,<sup>1,\*</sup> Montserrat Agudé<sup>1</sup>

<sup>1</sup> Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, 08071 Barcelona, Spain

Received: 10 September 1997 / Accepted: 23 March 1998

**Abstract.** Synonymous codons are not generally used at equal frequencies, and this trend is observed for most genes and organisms. Several methods have been proposed and used to estimate the degree of the nonrandom use of the different synonymous codons. The estimates obtained by these methods, however, show different levels of both precision and dispersion when coding regions of a finite number of codons are under analysis. Here, we present a study, based on computer simulation, of how the different methods proposed to evaluate the nonrandom use of synonymous codons are affected by the length of the coding region analyzed. The results show that some of these methods are heavily influenced by the number of codons and that the comparison of codon usage bias between coding regions of different lengths shows a methodological bias under different conditions of nonrandom use of synonymous codons. The study of the dispersion of the estimates obtained by the different methods gives, on the other hand, an indication of the methods to be applied to compare values of codon usage bias among coding regions of equivalent length.

**Key words:** Synonymous codon usage bias — Estimation methods — Effective number of codons — Codon bias index — Codon adaptation index — “Scaled”  $\chi^2$  — Intrinsic codon bias index

### Introduction

The nonuniform use of synonymous codons (codon bias) has been widely noted in most of the unicellular and multicellular organisms studied so far (Fiers et al. 1978; Grantham et al. 1981). The extent of the bias, however, is highly variable among genes for a given species. In unicellular organisms (i.e., *Escherichia coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae*), the degree of codon bias has been related to both the content of the isoacceptor tRNAs and the level of gene expression as the result of selection to increase translational efficiency (Grosjean and Fiers 1982; Ikemura 1985; Sharp and Li 1987a; Bulmer 1991). In this sense, highly expressed genes tend to use mainly those synonymous codons with the most abundant tRNA (“major” or preferred codons), while weakly expressed genes show a more frequent use of the “minor” or unpreferred synonymous codons (Ikemura 1980, 1981, 1985; Ikemura and Ozeki 1983; Grantham et al. 1981; Bennetzen and Hall 1982; Gouy and Goutier 1982; Grosjean and Fiers 1982). Accordingly, highly expressed genes show a base composition that departs more strongly from that expected by a mutational equilibrium. In multicellular organisms, both the levels of gene expression and the tRNA abundance are far from being clearly quantified as they can vary among tissues and developmental stages. Similar considerations, however, are proposed to explain the different levels of codon bias shown by different genes in organisms such as *Drosophila melanogaster* and *Bombyx mori* (Shields et al. 1988; Chevalier and Garel 1979; Garel 1982). In contrast, many warm-blooded vertebrates have genomes structured in isochores of different nucleotide composition as a result of mutational biases, and this structure causes different synonymous codon usage (Aota and Ike-

\* Present address: Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA  
Correspondence to: J.M. Comeron; e-mail: jcomeron@midway.uchicago.edu

**Table 1.** Average nucleotide frequencies used in the third position of codons for the seven nucleotide conditions under study<sup>a</sup>

	C-0	C-Low1	C-Medium1	C-High1	C-Low2	C-Medium2	C-High2
$f(A)$	0.25	0.20	0.125	0.05	0.20	0.125	0.05
$f(G)$	0.25	0.30	0.375	0.45	0.20	0.125	0.05
$f(C)$	0.25	0.30	0.375	0.45	0.40	0.625	0.85
$f(T)$	0.25	0.20	0.125	0.05	0.20	0.125	0.05

<sup>a</sup> $f(N)$  indicates the frequency of nucleotide  $N$ .

mura 1986; Bernardi and Bernardi 1986; Ikemura 1985; Filipsky 1987). Plant chloroplasts seem to show an intermediate usage pattern, a compromise in which codon bias would be determined by both mutational biases and translational efficiency (Morton 1993, 1994).

A number of studies have compared the levels of codon bias among genes within a species as well as among different species using some of the many methods proposed to quantify the nonrandom use of the different synonymous codons. These measures can be divided into those that have as a null hypothesis the random use of the different synonymous codons and those that quantify the bias by comparing the observed frequency of the different synonymous codons to the frequency of the preferred codons. Methods such as the “scaled”  $\chi^2$  (Shields et al. 1988), the effective number of codons (Nc) (Wright 1990), the codon bias index (CBI) (Morton 1993, 1994), and the intrinsic codon bias index (ICDI) (Freire-Picos et al. 1994) belong to the first class. The codon adaptation index (CAI) (Sharp and Li 1987b) and the frequency of optimal codons (Fop) (Ikemura 1981) belong to the second class.

Here we present a study using computer simulation of how different methods of estimating the extent of codon bias are affected by the length of the coding regions under analysis (number of codons). The results indicate that some of the estimates can become strongly biased in reaction to a shortening of the sequence lengths. Most actual coding regions show the length in the range where this effect would be clearly perceptible. Also, the dispersion of the estimates obtained by the different methods differs strongly and thus some of the methods are more suitable for comparing codon bias among equally sized genes or regions.

## Methods

Different methods to quantify the extent of nonrandom use of synonymous codons, codon bias, were analyzed: CAI (Sharp and Li 1987b), “scaled”  $\chi^2$  (Shield et al. 1988), Nc (Wright 1990), CBI (Morton 1993, 1994), and ICDI (Freire-Picos et al. 1994) (see original references for details). CAI estimates the degree of adaptation of the synonymous codons of a coding region referred to the optimal usage, where values of 1.0 indicate the maximum fit to the use of those codons and lower values the use of less preferred codons. As CAI can be used only with a reference of optimal codons, in this study the reference frequencies of the synonymous codons reflecting maximum adaptation

are those of the simulated condition (see below) with the more extreme synonymous codon usage. The scaled  $\chi^2$  is a measure of departure from equal use of synonymous codons estimated by a  $\chi^2$  statistic scaled by dividing it by the number of codons analyzed; the higher the values, the higher the degree of bias, and 0 indicates a perfectly uniform usage. In the present study scaled  $\chi^2$  was estimated with the correction for continuity, consisting in subtracting 0.5 from the absolute value of the deviation between observed and expected frequencies, when the observed number of synonymous codons is less than 5. Equivalent analyses of the scaled  $\chi^2$  were also performed without correcting for continuity. Nc represents the effective number of codons used in a gene, where a value of 18 indicates the use of only one synonymous codon for each amino acid with various synonymous codons and 59 the completely uniform use of the different synonymous codons. CBI is a measure of deviation from the uniform use of synonymous codons that achieves values between 0 and 1 for random use and maximum bias among synonymous codons, respectively. This method has two slightly different formulations (Morton 1993, 1994), which are referred to here as CBI93 and CBI94, respectively. ICDI is an index that conceptually is close the Nc method but with the advantage that it ranges between 0 and 1 for uniform and highly biased use of synonymous codons, respectively. The  $G + C_3$  index, which measures the fraction of third positions of the codons that is G or C, was also studied.

## Computer Simulations

A pseudo-random coding sequence was generated with a particular number of codons and using seven nucleotide compositions in the third position of codons. The average nucleotide composition of the seven nucleotide configurations (conditions) is shown in Table 1. In this sense, the condition C-0 represents the random use of synonymous codons, while the Low(1,2), Medium(1,2), and High(1,2) conditions indicate different degrees of codon bias and G + C content at the third position of codons. Six lengths of coding regions were analyzed: 100, 150, 250, 500, 1000, and 2500 codons. The average and the standard deviation of the estimate of codon bias produced by each method were based on 10,000 replicates per each combination of nucleotide frequencies and number of codons.

The effect of the length of the coding region on both the average and the dispersion of the estimates obtained was investigated for each method. In order to have comparable values among the methods, the standardized difference or difference between the average value achieved for a given length and the value for the longest sequence (2500 codons) relative to the latter value were obtained for each condition and method. In addition, the dispersion of the estimates obtained from a given method was quantified as the coefficient of variation (CV) or ratio between the standard deviation and the mean.

## Results and Discussion

### *Effect of the Coding Region Length on the Different Estimates of Codon Bias*

The standardized differences produced by the different methods (see Methods) are shown in Fig. 1 for the dif-

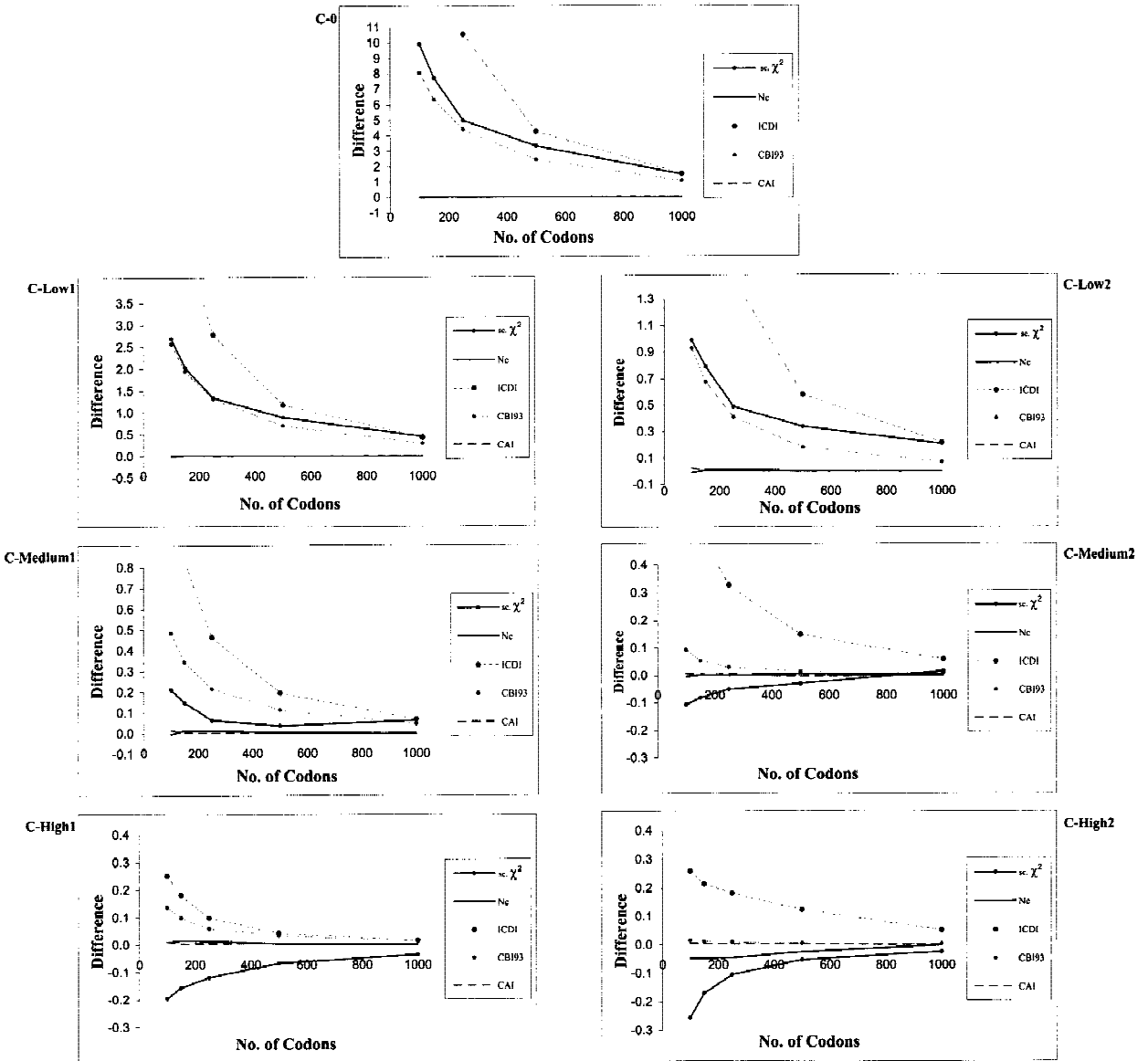


Fig. 1. Standardized differences produced by the different methods (see Methods) for the seven conditions of synonymous codon usage (see Table 1).

ferent lengths and conditions. The average values obtained by the different methods for the sequence of 2500 codons, and used to obtain the standardized values, are indicated in Table 2. The leading result is that the length of the coding region under analysis markedly affects the “scaled”  $\chi^2$ , CBI93, and ICDI measures. The effect of analyzing different numbers of codons is larger when the use of the different synonymous codons is closer to a uniform usage. In this sense, for a perfectly uniform usage (condition C-0), the more extreme situation, the pairwise comparison of two sequences of 150 and 500 codons would give a spurious twofold higher degree of codon bias of the smaller fragment using scaled  $\chi^2$  or CBI93, while for ICDI it would be threefold; comparison of sequences of 150 versus 1000 codons doubled these overestimations. On the other extreme, and again in se-

quences of 150 codons relative to sequences of 500 codons, in the presence of very biased usage (condition C-High2) ICDI estimates are biased upward by 10%, while CBI93 estimates are unaffected, and scaled  $\chi^2$  estimates are underestimated by 15%. Equivalent results were obtained when the average values of the sequence of 1000 codons were used as reference. For all conditions and lengths the CBI94 index is more strongly affected by low numbers of codons than the CBI93 (data not shown). The application of the correction for continuity to the scaled  $\chi^2$  method reduces the overestimation of the degree of codon bias that is obtained for short sequences without the correction for conditions of no or medium levels of bias among synonymous codons. Nevertheless, in cases with high (conditions C-High1 and C-Medium2) or very high (condition C-High2) codon bias, this cor-

**Table 2.** Average values obtained by the different methods of estimating the degree of codon bias for different conditions (see Methods) using randomly generated sequences of 2500 codons

	C-0	C-Low1	C-Medium1	C-High1	C-Low2	C-Medium2	C-High2
Nc	59.0	56.8	47.1	35.6	53.6	37.5	25.5
CAI	0.163	0.191	0.243	0.309	0.240	0.418	0.710
“Scaled” $\chi^2$	0.017	0.057	0.268	0.665	0.121	0.626	1.505
ICDI	0.009	0.035	0.172	0.434	0.064	0.271	0.578
CBI93	0.045	0.120	0.361	0.614	0.235	0.598	0.851

rection results in a clear underestimation for short sequences, as a result of the excessively conservative character of the correction (Sokal and Rohlf 1995) (data not shown). On the other hand, neither the Nc nor the CAI methods show any effect of gene length. Only in the condition of extremely high codon usage bias (C-High2), Nc exhibits a slight underestimation (i.e., 24.3 and 24.9 for 100 and 500 codons, respectively).

#### Variability of the Codon Bias Values

The degree of dispersion of the estimates obtained by the different methods is shown in Fig. 2. The ICDI and the scaled  $\chi^2$  methods show the largest levels of dispersion for any condition and length. The coefficient of variation (CV) for these two methods is lower when the bias among the synonymous codons is larger. Nc and CAI exhibit an almost-constant CV for different conditions, the dispersion always being larger for CAI than for Nc. Nc exhibits the lowest levels of dispersion but for those conditions with high (C-High1 and C-Medium2) or very high (C-High 2) bias. CBI93 is highly variable for conditions of no or low bias, although it exhibits the least variable values for short coding sequences and high bias (C-High1 and C-Medium2) and for all lengths and very high bias (C-High2). As expected, CV values are equivalent for CBI93 and CBI94.

#### Applicability to Observed Synonymous Codon Usages

Our approach to generate coding sequences under different conditions of codon usage bias produces a homogeneous bias for codon groups. Otherwise, in most species the actual synonymous codon usage can show (i) a diverse degree of codon usage bias for different amino acids, or (ii) a different nucleotide-ending “preferred” codon for different amino acids, or (iii) both. The effect of the first discrepancy is expected to produce intermediate results compared to our different conditions. On the other hand, the differences in the nucleotide-ending “preferred” codon among amino acids would have no effect on our results, as the only significant parameter in all measures, but the  $G + C_3$ , is the relative frequency of the different synonymous codons. Here we have examined these predictions by analyzing the actual synonymous codon usage frequencies of three archetypal spe-

cies: (i) *E. coli*, with an average G + C content at the third position of synonymous codons of 0.54 and different nucleotide-ending preferred codons (i.e., T-ending, Val; G-ending, Pro); (ii) *S. cerevisiae*, with a genome tendency toward A + T richness (G + C content of 0.36) as well as different nucleotide-ending preferred codons; and (iii) *D. melanogaster*, with an average G + C content of 0.65 and always G- or C-ending preferred codons. For *D. melanogaster*, we have analyzed separately the genes with maximum and minimum codon usage bias (Kreitman and Antezana 1998). The use of real conditions shows that the pattern of how the different methods are influenced by the length of the coding region depends essentially on the overall Nc (Fig. 3). Our conditions can therefore be considered of general use even if they produce a homogeneous bias for codon groups. In fact, the patterns for *E. coli* (Nc = 47.2) and *S. cerevisiae* (Nc = 49.5) are similar to our condition C-Medium1 (see Table 2), and the patterns for *D. melanogaster*–high codon bias (Nc = 34.6) and *D. melanogaster*–low codon bias (Nc = 56.4) are similar to our conditions C-High1 and C-Low1, respectively.

#### Conclusion

This study addresses two main questions concerning methods to estimate the degree of codon usage bias. First, when coding regions of different lengths are compared, only the Nc and CAI methods give equivalent values for all lengths and conditions of codon bias. For a particular degree of codon bias (condition), the methods “scaled”  $\chi^2$ , CBI (CBI93 or CBI94), and ICDI give clearly different average values when coding regions with different numbers of codons are compared. This difference can be very conspicuous under conditions of no or moderate codon bias even when comparing sequences of 500 and 1000 codons. The pairwise comparison of coding regions of different lengths shorter than 500 codons would therefore show a methodological bias under most conditions of synonymous codon usage. As indicated in Methods, CAI estimates the deviation from an “optimum” or reference bias showed in a particular species. Usually, highly expressed genes have been used to obtain this reference set of values in unicellular organisms. Different organisms, however, can show different reference sets, making the comparison between CAI

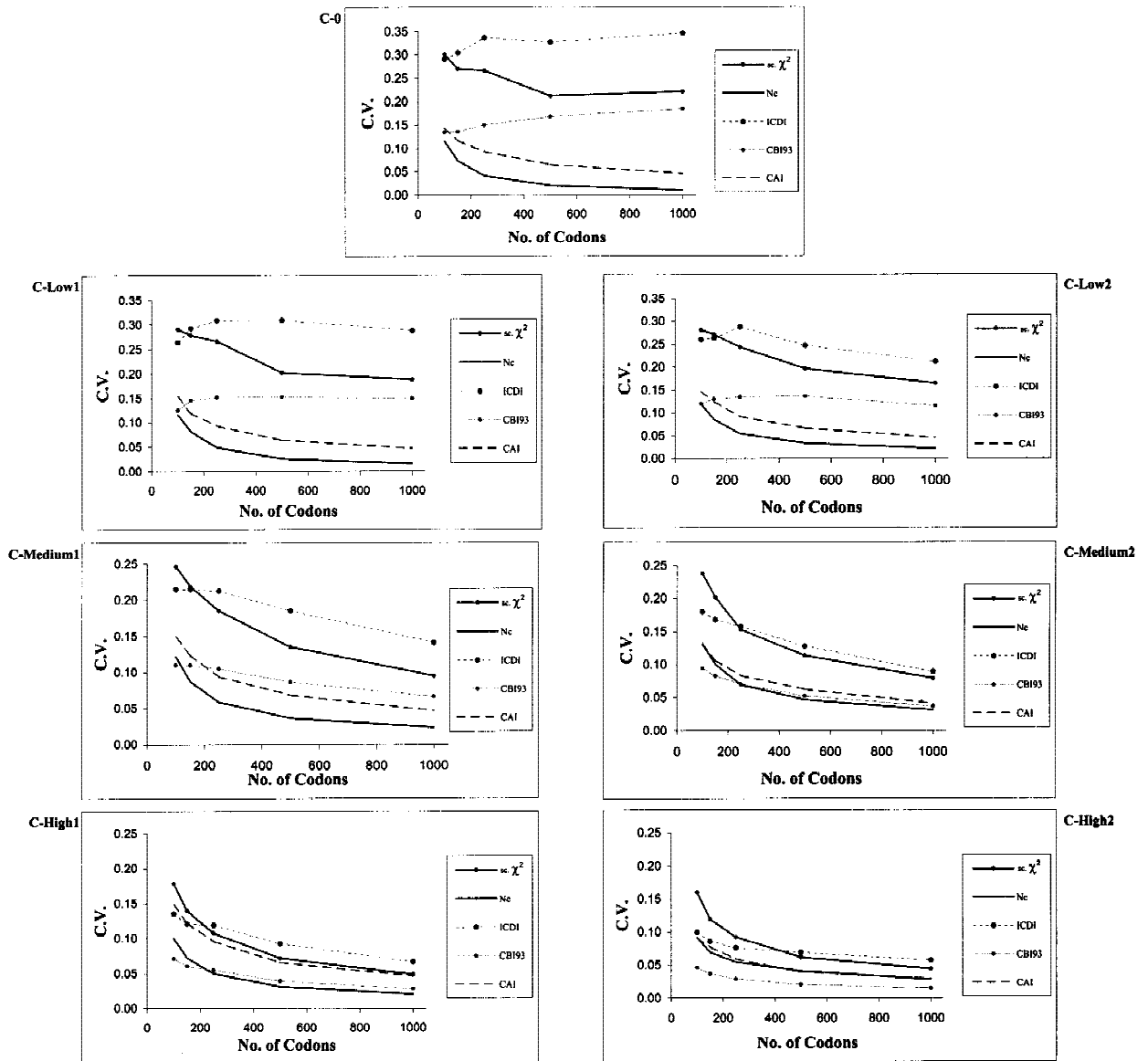
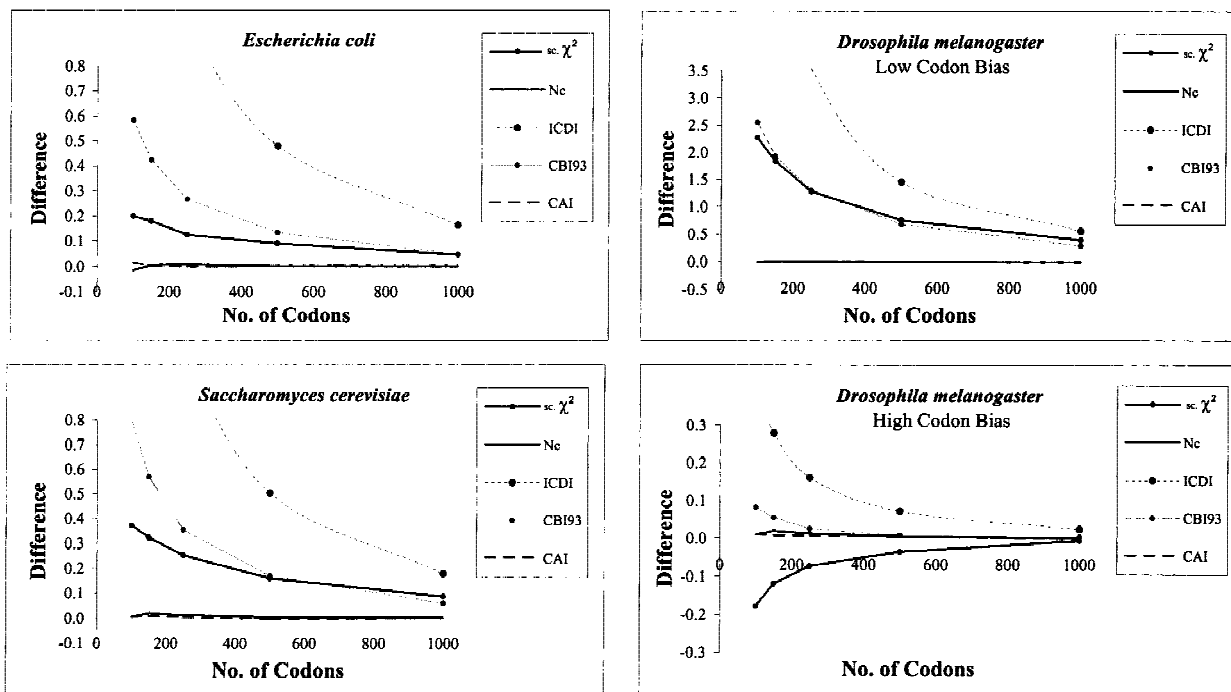


Fig. 2. Coefficient of variation (C.V.; standard deviation/mean) obtained for the different methods and conditions of synonymous codon usage (see Table 1).

values from different species difficult. Moreover, in species with a low number of sequenced coding regions or in multicellular organisms where the quantification of gene expression is not as direct as in unicellulars, the use of the CAI method might be questionable.

Second, when a particular analysis aims at comparing values of codon bias among coding regions of equivalent size (i.e., homologous genes, sets of genes between or within species, or different regions of equal size of a coding region), Nc and, to a lesser extent, CBI93 would be the methods of choice, as they exhibit the lowest CV. As a rule, Nc exhibits the lowest CV for the different conditions and lengths. However, when medium or short lengths (250 codons or lower) are under study, CBI93 exhibits the lowest dispersion for regions with an overall high degree of codon bias.

The content of G + C nucleotides in the third position of codons ( $G + C_3$ ) has also been used as an indirect measure of the extent of bias in the use of the different synonymous codons. This measure is highly correlated with the direct measures of codon bias in most organisms but this relationship is not universally detected (i.e., *E. coli* and *Bacillus subtilis*, for no and a negative correlation, respectively) (Sharp et al. 1986; Wright 1990). This measure will give only indirect and approximate estimates of the overall use of the different synonymous codons, as even for those species where the preferred codons are always G- or C-ending (i.e., *D. melanogaster*), a given  $G + C_3$  value can obscure different G and C contents, with the result of diverse true degrees of bias among synonymous codons. As expected, the different number of codons under study does not affect  $G + C_3$ ,



**Fig. 3.** Standardized differences produced by the different methods using the actual codon usage frequencies of *E. coli*, *S. cerevisiae*, and *D. melanogaster* lowest and highest codon usage biased genes. Data for *E. coli* and *S. cerevisiae* codon usage frequencies are from Nakamura et al. (1997), and those for *D. melanogaster* from Kreitman and Antezana (1998).

and from the binomial distribution of the  $G + C_3$  measure in the simulations, the variance is maximum for  $G + C_3 = 0.5$ .

**Acknowledgments.** We thank M. Antezana and E. Juan for comments and M. Kreitman and A. Llopart for valuable discussions. J.M.C. was funded by a predoctoral fellowship from Ministerio de Educación y Ciencia (M.E.C.), Spain, and is currently funded by a postdoctoral fellowship from M.E.C. This work was supported by Grants PB91-0245 from Dirección General de Investigación Científica y Técnica, M.E.C., Spain, and GRQ93-1100 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, to M.A.

## References

- Bennetzen JL, Hall BD (1982) Codon selection in yeast. *J Biol Chem* 257:3026–3031
- Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *J Mol Evol* 24:1–11
- Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907
- Chevallier A, Garel JP (1979) Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. *Biochemie* 61:245–262
- Filipki J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett* 217:184–186
- Freire-Picos MA, González-Siso MI, Rodríguez-Belmonte E, Rodríguez-Torres AM, Ramil E, Cerdán ME (1994) Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. *Gene* 139:43–49
- Garel JP (1982) The silkworm, a model for molecular and cellular biologists. *Trends Biochem Sci* 7:105–108
- Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* 10:7055–7074
- Grantham R, Gautier C, Gouy M, Jacobzone M, Mecier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* 9:43–74
- Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18:199–209
- Ikemura T (1980) The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA. In: Osawa S, Ozeki H, Uchida H, Yura T (ed) *Genetics and evolution of RNA polymerase, tRNA and ribosomes*. University of Tokyo Press, Tokyo, and Elsevier/North-Holland, Amsterdam, pp 519–523
- Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151:389–409
- Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13–34
- Ikemura T, Ozeki H (1983) Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harbor Symp Quant Biol* 47:1087–1097
- Kreitman M, Antezana M (1998) The population and evolutionary genetics of codon bias. In: Sing R, Krimbas C (ed) *Evolutionary genetics from molecules to morphology*. Columbia University Press, New York (in press)
- Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the *psb A* locus based on tRNA availability. *J Mol Evol* 37:273–280
- Morton BR (1994) Codon use and the rate of divergence of land plant chloroplast genes. *Mol Biol Evol* 11:231–238

- Nakamura Y, Gojobori T, Ikemura T (1997) Codon usage tabulated from the international DNA sequence database. *Nucleic Acids Res* 25:244–245
- Sharp PM, Li W-H (1987a) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4:222–230
- Sharp PM, Li W-H (1987b) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281–1295
- Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 14:5125–5139
- Shields DC, Sharp PM, Higgins DG, Wright F (1988) “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* 5:704–716
- Sokal RR, Rohlf FJ (1995) *Biometry*, 3rd ed. W.H. Freeman, New York
- Wright F (1990) The “effective number of codons” used in a gene. *Gene* 87:23–29