Gene, 87 (1990) 23–29
Elsevier

GENE 03437

The 'effective number of codons' used in a gene

(Synonymous codon usage bias; G + C content; amino acid sequence; *Homo sapiens*, *Saccharomyces cerevisiae*; *Escherichia coli*; *Bacillus subtilis*; *Dictyostelium discoideum*; *Drosophila melanogaster*)

Frank Wright

Institute of Animal Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN (U.K.)

Received by K.F. Chater: 5 June 1989
Revised: 18 September 1989
Accepted: 20 September 1989

SUMMARY

A simple measure is presented that quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. This measure of synonymous codon usage bias, the 'effective number of codons used in a gene', $\hat{N}_c$, can be easily calculated from codon usage data alone, and is independent of gene length and amino acid (aa) composition. $\hat{N}_c$ can take values from 20, in the case of extreme bias where one codon is exclusively used for each aa, to 61 when the use of alternative synonymous codons is equally likely. $\hat{N}_c$ thus provides an intuitively meaningful measure of the extent of codon preference in a gene. Codon usage patterns across genes can be investigated by the Nc-plot: a plot of $\hat{N}_c$ vs. G + C content at synonymous sites. Nc-plots are produced for *Homo sapiens*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Bacillus subtilis*, *Dictyostelium discoideum*, and *Drosophila melanogaster*. A FORTRAN77 program written to calculate $\hat{N}_c$ is available on request.

INTRODUCTION

The phenomenon of unequal usage of synonymous codons is well documented (e.g., Ikemura, 1985) and biased (i.e., unequal) usage of alternative synonymous codons can be observed in most protein-coding genes. Synonymous codon usage (SCU) patterns may be due to mutation bias [e.g., mutation may produce cryptic patterns (Foster et al., 1982) and/or produce a bias in G + C content (Treffers et al., 1954)] and/or to various forms of natural selection (e.g., to optimize the efficiency/accuracy of translation, and/or to maintain structural features of the mRNA/DNA). Compilations of codon usage tables for individual genes in a particular species (e.g., Maruyama et al., 1986) are of limited value due to the complexity of the information displayed. An easily interpretable summary of codon usage data would be of general use. A method is presented here which allows the reader to gain an overall view of the codon usage patterns of an organism especially in relation to G + C content.

The particular nature of SCU bias is species-specific (Grantham et al., 1980a,b; 1981), but there is also considerable variation among genes from a species (Gouy and Gautier, 1982; Sharp et al., 1988). Describing within-species SCU patterns is facilitated by the discovery of clear SCU trends in most species. Two types of trend have been reported. The first type, observed in mammalian species

Correspondence to: Dr. F. Wright, Scottish Agricultural Statistics Service, University of Edinburgh, J.C.M.B., King's Buildings, Mayfield Road, Edinburgh EH9 3JZ (U.K.) Tel. 031-667 1081 (ext. 2993); Fax 031-667 2601.

Abbreviations: aa, amino acid(s); GC3s, the proportion of G + C at synonymous sites (i.e., in the third codon position, excluding Met and Trp); H_0, null hypothesis (no selection, G + C bias due to mutation); H_0^*, special case of null hypothesis (no selection, no G + C bias); H_1, alternative hypothesis (selection acting on 'preferred' codons); L_c, length of gene in codons; $\hat{N}_c$, (effective) number of codons; Nc-plot, a plot of $\hat{N}_c$ vs. GC3s for a set of genes; $\hat{N}_e$, (effective) number of alleles; SCU, synonymous codon usage; SF, synonymous family.

(Ikemura, 1985), results from variation among genes in the G + C content at synonymous sites (i.e., GC3s, defined as the proportion of G + C content in the third codon position, excluding Met and Trp). For example, *H. sapiens* genes exhibit an SCU trend from very high GC3s content (approx. 0.95) to low GC3s content (approx. 0.35) (Aota and Ikemura, 1986).

A second type of SCU trend, observed in genes from unicellular species (e.g., *E. coli*, yeast) (Sharp and Li, 1987) and from *D. melanogaster* (Shields et al., 1988), exhibits a range from extreme SCU bias to minimal SCU bias. The codon usage pattern of genes at the extreme end of the trend is species-specific and is a reflection of codon 'preference' in the particular species. This type of SCU trend is not associated with GC3s content unless the 'preferred' codons themselves tend predominantly to contain (or not contain) G and/or C in the third position.

In the case of unicellular organisms, the degree of SCU bias of a gene is highly correlated with its level of expression in the cell (Gouy and Gautier, 1982; Ikemura, 1985). Highly expressed yeast genes, for example, show a tendency to use only 22 'preferred' codons, whereas lowly expressed genes tend to make more uniform use of the 61 sense codons (Bennetzen and Hall, 1982). Similarly, Bennetzen and Hall noted that highly expressed genes in *E. coli* tended to use a subset of 25 'preferred' codons. This informal terminology of Bennetzen and Hall (1982) can be used as a basis for developing a quantitative measure of SCU bias. A distance measure of a particular gene from an unbiased SCU pattern is a useful summary statistic. Such a measure can then be used to investigate the relationship between SCU bias and putative factors; for example, the correlation of a SCU bias measure and the level of gene expression could be computed to study the possible action of translational selection on highly expressed genes. Such a measure is also empirically attractive.

An analogy can be drawn between the usage of synonymous codons for a particular aa and the frequencies of alleles at a locus. The N_e quantifies the number of alleles at a polymorphic locus by providing a figure for the number of equally frequent alleles that would produce the given level of homozygosity (Kimura and Crow, 1964). The SCU bias of each aa can be described in this way. Summing the 'effective number of alleles' used by each of the 20 aa will then yield an $\hat{N}_c$, used in a gene. An extremely biased gene would use only 20 codons (i.e., one per aa), whereas an unbiased gene would tend to use all 61 codons equally (after correcting for aa usage).

Several other measures of SCU bias have been developed (see RESULTS AND DISCUSSION, section **d**, for more detail). These vary in statistical sophistication, the nature of input data required (i.e., raw sequence data or codon usage data), the need for additional information about the gene or organism under study, the type of distance measure (i.e., whether from an 'unbiased' or 'biased' SCU pattern), and in ease of computation. None of these measures are widely used.

The $\hat{N}_c$ is suggested as a routine tool in the analysis of coding DNA. The $\hat{N}_c$ statistic performs well (i.e., it is an unbiased estimator) even for short genes (of at least 100 codons in length), and for skewed aa usage. The $\hat{N}_c$ is easily computed from codon usage data and its value provides an intuitively obvious measure of the extent of codon preference of a gene. Here, the derivation of $\hat{N}_c$ is described and the statistic applied to codon usage data from six organisms which differ in their SCU patterns. The results of these analyses are displayed via the use of the Nc-plot: a plot of $\hat{N}_c$ against GC3s. This graphical method illustrates the SCU trends in the six organisms.

THEORY

In this section, the derivation of $\hat{N}_c$ is described resulting in Eqn. (3). The calculation of $\hat{N}_c$ for a gene is relatively simple (although a FORTRAN77 program is available from the author on request). The behaviour of $\hat{N}_c$ when gene length is short and/or when some aa are rare or missing is discussed.

(a) Analogy of multiple synonymous codons with multiple alleles

The codon usage table of a gene can be subdivided according to the number of synonymous codons belonging to each aa. Thus for a gene using the 'universal' code, there are 2 aa with only one codon choice, 9 with two, 1 with three, 5 with four, and 3 with six. These represent five SF types, designated SF types 1, 2, 3, 4, and 6 according to their respective number of synonymous codons. The measure of SCU bias developed here will involve combining contributions to overall bias from each of the five SF types. However, first we must consider the simpler problem of measuring the contribution of a single aa to the overall SCU bias of a gene.

Let us consider the contribution to overall SCU bias of an aa with four alternative synonymous codons (i.e., a member of SF type 4, which consists of Val, Pro, Thr, Ala and Gly). The actual usage of the four synonymous codons will be denoted $n_1,..,n_4$. The total usage of the aa is therefore $n = n_1 + .. + n_4$. The frequency of usage of the synonymous codons is $p_1,..,p_4$ obtained by dividing the respective actual usage by n (where, for example, $p_1 = n_1/n$). An aa with four synonymous codons is analogous to a locus with four alleles. Equal codon usage would be equivalent to minimum homozygosity. Homozygosity (F) can be calculated from the squared allele (codon) frequencies:

$$\hat{F} = (n \sum_{i=1}^{k} p_i^2 - 1)/(n - 1) \qquad (1)$$

where k is the number of alleles (codons). For an aa of SF type 4 (as considered here), $k = 4$ [Eqn. (1) is simply the negative of Nei and Tajima's (1981) expression for 'nucleon diversity'].

The N_e, (Kimura and Crow, 1964), can now be calculated:

$$\hat{N}_e = 1/\hat{F} \quad (2)$$

The value of $\hat{N}_e$ is equivalent to the number of equally frequent alleles that would produce the particular level of homozygosity. Thus different values of $(p_1,..p_4)$ can produce the same value of $\hat{N}_e$. Equal usage would yield (approximately) $\hat{N}_e = 4$, whereas usage of only one allele would give (approximately) $\hat{N}_e = 1$. (The $\hat{N}_e$ values will tend to exactly 4 and 1, respectively, as gene length increases.)

The above expression for $\hat{N}_e$ can be used to compare the relative SCU bias of aa of the same SF type.

(b) Derivation of the 'effective number of codons', $\hat{N}_c$

To analyse the SCU bias in a codon usage table, let us consider it as 20 'loci' each with between 1 and 6 possible 'alleles'. The method presented here can be thought as equivalent to adding together the 'N_e' values for each of the 20 'loci'. This will yield an 'effective number of codons', $\hat{N}_c$, of 61 when all codons are used equally for each aa. A value of 20 will be obtained for $\hat{N}_c$ when only one codon is used for each aa. The actual computation of $\hat{N}_c$ is slightly different from the above description in order to produce a measure of SCU bias that is minimally affected by sequence length and aa usage.

$\hat{N}_c$ is actually obtained by adding the contributions from each of the five SF types. These contributions consist of the number of members in the SF type divided by the average $\hat{F}$ value of the aa in the SF type. Note that the contribution of the SF type with one codon (Met, Trp) is set equal to two. (Note also that SF type 3 consists only of 1 aa, Ile.) Thus:

$$\hat{N}_c = 2 + (9/\bar{\hat{F}}_2) + (1/\hat{F}_3) + (5/\bar{\hat{F}}_4) + (3/\bar{\hat{F}}_6) \quad (3)$$

where $\bar{\hat{F}}_i$, is the average homozygosity estimate for SF type i, and $\hat{F}_i$ for each aa are calculated using Eqn. (1).

(c) Rare or missing amino acids

Adjustments to the above method need to be made if one or more aa are rarely used or absent. Rarely used aa are those aa for which either the numerator or denominator of Eqn. (1) is 0; if this occurs they should be treated as absent. When an aa is absent, an empirical adjustment should be made to Eqn. (3) so as to average over only those aa present in each SF type. The average contribution $\bar{\hat{F}}_i$ should be calculated for all SF types as an average of those aa present. However, if Ile is missing or 'rarely used', $\hat{F}_3$ should be computed as the average of $\bar{\hat{F}}_2$ and $\bar{\hat{F}}_4$. If any of the other SF types are completely missing or 'rarely used' then the gene is probably too short (e.g., less than 61 codons) to accurately measure SCU bias, or exhibits extremely skewed aa usage.

The value of $\hat{N}_c$ can be greater than 61 if the observed codon usage pattern is more uniform than expected by chance. Such a rare event is most likely when aa composition is very extreme (e.g., such that the $\bar{\hat{F}}_i$ are obtained by averaging over very few aa), and the gene is very short. In these cases, the value of $\hat{N}_c$ should be revised to 61.

(d) Simulation results

It is reasonable to expect that the value of $\hat{N}_c$ will be less reliable when calculated on very short sequences. For example, consider a set of very highly expressed genes from a hypothetical unicellular organism where there is one 'preferred' codon for each aa. This set of genes might all be expected to have $\hat{N}_c$ values approaching 20. However, let us assume that these genes vary in length. It is of interest to know how the actual value of $\hat{N}_c$ behaves with varying gene length. In statistical terminology, $\hat{N}_c$ is an estimator (hence the caret symbol) of the 'true' value of N_c. In this hypothetical example, the 'true' value of N_c is 20. Rather than studying only extreme SCU bias (e.g., 20, as discussed here), a simulation study was carried out using a range of 'true' N_c values from 21 to 61. The number of 'preferred' codons was, however, held at 20.

A simulation program was developed to study the behaviour of $\hat{N}_c$ as the length of the gene (L_c, measured in codons) was varied. For each value of L_c and 'true' N_c, 100 sequences were generated. The observed value of $\hat{N}_c$ was calculated for each sequence and the mean and standard deviation of $\hat{N}_c$ were obtained. The simulation results are summarised in Fig. 1. This is a plot of SCU bias as estimated by $\hat{N}_c$, vs. L_c. The 'true' N_c values for each of the five lines are given on the right of Fig. 1. Error bars represent one standard deviation on either side of the mean $\hat{N}_c$ value. If $\hat{N}_c$ was a perfect estimator of SCU bias then the mean $\hat{N}_c$ figure would be equal to the 'true' N_c values for all gene lengths, and the standard deviation of $\hat{N}_c$ (as shown by the error bars) would be 0. In reality, $\hat{N}_c$ is a good estimator and only underestimates SCU bias (i.e., by overestimating N_c) for gene lengths of less than 100 codons (these estimates also have high standard deviations).

RESULTS AND DISCUSSION

(a) Use of $\hat{N}_c$: a simple illustration

To illustrate the use of $\hat{N}_c$, SCU bias patterns in a mammalian genome (*H. sapiens*) and in a unicellular genome

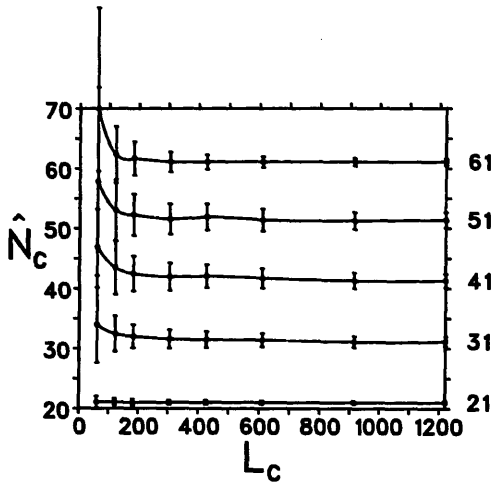

Fig. 1. Plot showing results of a simulation study of the behaviour of $\hat{N}_c$ as gene length, L_c (measured in codons), is varied. The five lines represent different 'true' values of N_c (i.e., 61, 51, 41, 31, and 21, as marked on the far right of the figure). A range of L_c was studied (i.e., 61, 122, 183, 305, 427, 610, 915 and 1220) similar to the approximate size range of real genes. Sequences were generated with a range of 'true' SCU bias ('true' N_c = 61, 51, 41, 31, and 21). The aa composition was held constant: the amount of each aa was proportional to its number of synonymous codons (e.g., Leu was three times as abundant as Phe). SCU bias was assumed to range from no bias (equal usage of synonymous codons) to extreme bias (only one synonymous codon used per aa). The observed codon usage for each aa was then generated by sampling from the 'true' codon frequencies, π_i, for the given level of 'true' SCU bias (assuming a multinomial distribution for the π_i). For a given 'true' value of N_c, and a given gene length L_c, 100 genes were generated. The estimated value of N_c, i.e. $\hat{N}_c$, was then calculated using Eqn. (3) (see THEORY, section b) for each of the 100 generated genes. Each point (and its associated error bar) represents the mean and standard deviation of $\hat{N}_c$.

(*S. cerevisiae*) can be compared. Bennetzen and Hall (1982) noted that highly expressed yeast genes tend to use only 22 'preferred' codons. $\hat{N}_c$ values have been calculated from the codon usage data quoted in their paper. Table I is based on their table III. GC3s values have been included to allow comparisons with the human codon usage data in Table II.

The most highly expressed gene in Table I, *GAP2*, has the most extreme SCU bias. The $\hat{N}_c$ value of 24.1 is close to the expected value of 22 codons for a gene using only 'preferred' codons. In contrast, *CYC7* 'uses' only 43.5 codons — this is still far from random usage. However, a more extensive study of 110 yeast genes (see section c, below) reveals that several genes have unbiased codon usage and have $\hat{N}_c$ values at or near 61. Similar results can be obtained for *E. coli* genes. Typical values range from $\hat{N}_c$ = 26.2 (for the highly expressed *E. coli* gene *rplL*) to $\hat{N}_c$ = 47.5 (for the lowly expressed *E. coli* gene *lacI*). Note that $\hat{N}_c$ = 26.2 is close to the expected value of 25 codons for an *E. coli* gene using 'preferred' codons (Bennetzen and Hall, 1982) (*E. coli* codon usage is discussed in more detail in section c, below).

Human genes have not been reported to show a relationship between SCU bias and level of gene expression. However, there is a very strong relationship between SCU bias and GC3s. $\hat{N}_c$ values for four human genes exemplifying the wide range of G + C at silent sites are shown in Table II.

TABLE I

$\hat{N}_c$ values for a selection of yeast genes

Gene[a]	$\hat{N}_c$[b]	Approx. % of total cell mRNA[c]	GC3s[d]
GAP2	24.1	1.5–6.0	0.498
GAP1	24.7	—	0.513
ADH1	27.3	0.7–2.0	0.491
H2B1	28.3	0.4	0.372
H2B2	33.5	0.4	0.403
CYC1	41.8	0.05	0.467
CYC7	43.5	0.003	0.358

[a] See Bennetzen and Hall (1982) for gene references.
[b] $\hat{N}_c$ calculated as in Eqn. (3) in THEORY, section b.
[c] Data from Bennetzen and Hall (1982).
[d] GC3s calculated as the proportion of G + C at synonymous sites.

TABLE II

$\hat{N}_c$ values for a selection of human genes

Gene[a]	$\hat{N}_c$[b]	GC3s[b]
HUMAMYAS	48.2	0.300
HUMHPA2B	57.9	0.529
HUMAPOC1	48.5	0.688
HUMHBA1	28.6	0.949

[a] See Maruyama et al. (1986) for gene references.
[b] See Table I.

SCU bias, as measured by $\hat{N}_c$, is lowest for genes with GC3s values of about 0.5 (e.g., *HUMHPA2B* above). Note that similar $\hat{N}_c$ values can be obtained for genes with very different GC3s values. *HUMAMYAS* and *HUMAPOC1* differ by nearly 0.4 in their GC3s value but have almost identical $\hat{N}_c$ values. A gene tending to use only G and C ending codons (e.g., *HUMHBA1*) will only use about half the sense codons and will therefore have an $\hat{N}_c$ value of about 30 and a GC3s value approaching 1.

The use of $\hat{N}_c$ and GC3s clearly demonstrates the (known) differences in codon usage patterns between the yeast and human genomes (e.g., Sharp et al., 1988). SCU variation among human genes mainly reflects GC3s variation, whereas SCU variation among yeast genes is only partly explained by GC3s variation (Sharp et al., 1986).

(b) The Nc plot: a graphic display of SCU bias and base composition

The $\hat{N}_c$ quantifies SCU bias in a range from extreme bias (one synonymous codon used per aa) to no bias (equal

usage of synonymous codons). Variation in GC3s will complicate the interpretation of the $\hat{N}_c$ value for a given gene. However, the relationship between $\hat{N}_c$ and GC3s under H_0 of no selection, is simple (the H_0 is discussed in more detail in section d, below). This relationship between N_c and GC3s can be approximated by:

$$N_c = 2 + s + \{29/[s^2 + (1 - s^2)]\} \quad (4)$$

where s denotes GC3s.

A plot of $\hat{N}_c$ vs. GC3s gives a useful visual display of the main features of codon usage patterns for a number of genes (see Fig. 2, a–f and discussion below). Such a plot will be referred to as the Nc-plot. The curve showing the relationship between $\hat{N}_c$ and GC3s under the H_0 (no selection) aids the interpretation of the Nc-plot. If a particular gene is subject to G + C compositional constraints, it will lie on or just below the GC3s curve. Note, however, that such a gene may be either subject to G + C-biased mutation pressure, or may be under selection (negative/positive) for codons ending in C and/or G. This point will be discussed later with reference to *H. sapiens* and *D. melanogaster*.

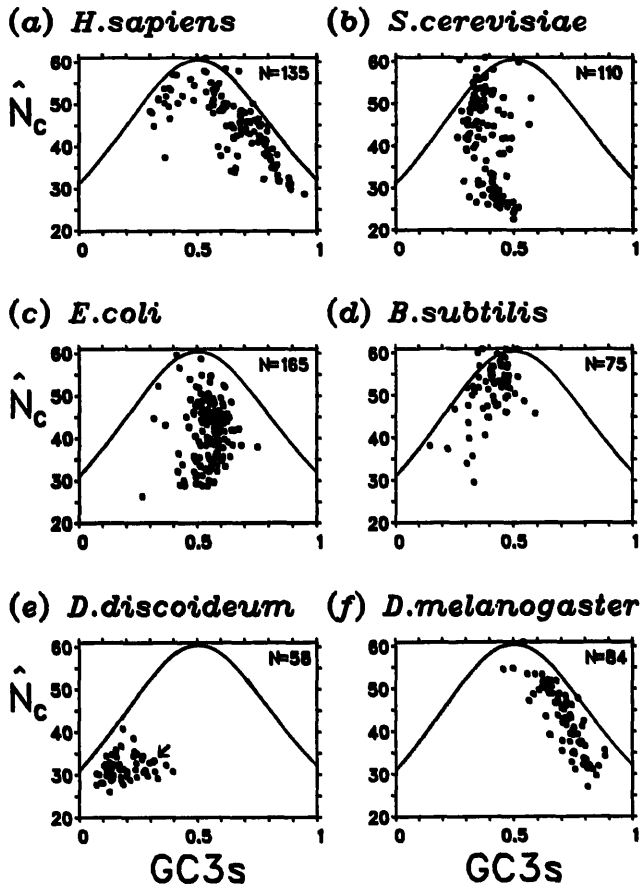

Fig. 2. Nc-plots for (*a*) *H. sapiens*, (*b*) *S. cerevisiae*, (*c*) *E. coli*, (*d*) *B. subtilis*, (*e*) *D. discoideum*, and (*f*) *D. melanogaster*. The number (N) of genes analysed was (*a*) 135, (*b*) 110, (*c*) 165, (*d*) 75 (see note below), (*e*) 58, and (*f*) 84. The continuous curve in each of the figures represents the relationship between $\hat{N}_c$ and GC3s (under H_0) as given in Eqn. (4) (see RESULTS AND DISCUSSION, section b). Data sources in panels: (*a*) Maruyama et al., 1986, (*b*) Sharp et al., 1986, (*c*) Sharp and Li, 1986, (*d*) Shields and Sharp, 1987, (*e*) Sharp and Devine, 1989, and (*f*) Shields et al., 1988. The data used in (*d*) excluded one gene, *sacQ* (L_c = 45 codons), because $\hat{N}_c$ could not be calculated because of rarely used or missing aa (see THEORY, section c). The arrowed *D. discoideum* gene in panel *e* is ribosomal gene *1024* (see RESULTS AND DISCUSSION, section c).

(c) Investigating codon usage in different organisms using the Nc-plot

$\hat{N}_c$ and the Nc-plot can be used as part of a general strategy to investigate patterns of SCU bias. Such a strategy might consist of three steps: the visual display of SCU patterns, the ranking of genes, and the making of inferences about the relative importance of mutation and selection in producing codon usage patterns.

The Nc-plot can be used to display codon usage patterns as a simple alternative to complex multivariate methods such as correspondence analysis (e.g., Grantham et al., 1980a; Shields et al., 1988). While correspondence analysis is a superior technique, the Nc-plot offers a distinct alternative since the two axes ($\hat{N}_c$ and GC3s) are defined by the researcher; this approach is valid for the study of codon usage. Variation in G + C content in the third codon position (GC3s) accounts for much of the within-species SCU variation in mammals (Ikemura, 1985; Aota and Ikemura, 1986), and much of the between-species SCU variation in bacteria (Muto and Osawa, 1987). $\hat{N}_c$ is a distance measure from a known reference pattern, is not species-specific and has a known relationship with GC3s. The axes of the Nc-plot are not influenced by the data and allow intra-specific and inter-specific comparisons of SCU patterns on the same plot.

As an example of the use of the Nc-plot to display codon usage patterns, consider the display of codon usage data from (*a*) a mammalian species (*H. sapiens*; see Fig. 2a), and (*b*) a unicellular organism (*S. cerevisiae*; see Fig. 2b). Fig. 2a displays the relationship between SCU bias and GC3s in *H. sapiens* genes. Human codon usage patterns reflect the base composition of the local DNA region (Bernardi and Bernardi, 1986), which itself is probably the result of variation in mutational bias among chromosomal regions (Wolfe et al., 1989). Fig. 2b displays the considerable SCU variation among yeast genes. Genes known to be highly expressed have low $\hat{N}_c$ values, e.g., *GAP1*, *GAP2*, *ADH1*, *H2B1*, *H2B2* from Table I plus *ENOA* (24.9), *ENOB* (25.5), *TEF1* (25.9), *PGK* (26.6), *GDH1* (32.0) and 16 ribosomal protein-encoding genes (ranging from 22.6 to 36.5) as listed in Sharp et al. (1986). The slight GC3s richness (0.45 compared to 0.38) of yeast highly expressed genes is mainly due to an increase in C, at the expense of A (Sharp et al., 1986). This relationship between SCU bias

and level of gene expression in yeast genes has been extensively studied (Ikemura, 1982; Bennetzen and Hall, 1982). The Nc-plot for *E. coli* is shown in Fig. 2c to allow a comparison with the yeast Nc-plot. They have many features in common and confirm Bennetzen and Hall's (1982) observation that highly expressed yeast genes exhibit more extreme SCU bias than those from *E. coli*. Although these three Nc-plots tell us nothing new, they do display the data in a concise and informative manner. For example, one can see that there is comparatively little variation in GC3s bias in yeast and *E. coli* genes compared to human genes. The wide range of SCU bias in both yeast and *E. coli* (especially in yeast) is clearly shown. Similarly, one can observe the range of GC3s in *H. sapiens* genes. We can also see the SCU trend for each of the three species.

The Nc-plot is particularly useful in the interpretation of SCU variation where the genome under study has a G + C content markedly different from 0.50. Consider Fig. 2d (*B. subtilis*, a Gram$^+$ bacterium) and Fig. 2e (*D. discoideum*, a slime mould). These two organisms have genomic G + C contents of 0.42, and 0.22, respectively. Shields and Sharp (1987) concluded that SCU bias in *B. subtilis* is less extreme than that found in yeast or *E. coli*, and that highly expressed genes tended to use A + T-rich codons. Highly expressed genes are not labelled in Fig. 2d, but are identifiable as the six genes with the lowest $\hat{N}_c$ values. The Nc-plot thus shows these features well.

The interpretation of SCU patterns in *D. discoideum* is complicated by the extreme genomic G + C content of only 0.22. However, the Nc-plot (Fig. 2e) is again able to display the main features observed in a recent study (Sharp and Devine, 1989). *D. discoideum* codon usage can be explained using the same theoretical framework as applied to other unicellular organisms. The lowly expressed genes (lying to the left of Fig. 2e, next to the GC3s curve) have SCU patterns mainly determined by strong mutation bias. The highly expressed genes (lying to the right of the figure, e.g., ribosomal gene *1024*) tend to use C-ending codons (ten of the 15 'preferred' codons are G- or C-ending).

The ranking of genes along such SCU trends requires an appropriate measure. For *H. sapiens* (and *D. melanogaster*, see below) genes, GC3s itself will suffice. GC3s would also provide an approximate measure of SCU bias in *D. discoideum* (but again, see below). $\hat{N}_c$ can be used to rank yeast (and *E. coli*) genes because GC3s is in the range 0.40 to 0.60 and the SCU variation is essentially independent of GC3s. This will not, however, be true generally. Genomes with biased overall G + C content will yield values of significantly less than $\hat{N}_c = 61$ for genes even if there is no translational selection acting. The SCU trend may also be partially correlated with GC3s and thus influence the value of $\hat{N}_c$. For example, see Fig. 2d (*B. subtilis*) and Fig. 2e (*D. discoideum*). If this is the case, a better approach may be to use a distance measure based on the similarity of the observed codon usage of each gene to that found in highly expressed genes (for the particular species). This approach will be discussed in section **d**, below.

The observed SCU bias pattern of a gene will be due to mutation pressure and selection. The relative effects of these two evolutionary forces cannot be ascertained simply from looking at a Nc-plot. Consider the Nc-plots for *H. sapiens* (Fig. 2a) and that for *D. melanogaster* (Fig. 2f). In both Nc-plots the SCU variation is highly correlated with variation in GC3s. However, in *H. sapiens* this is due predominantly to variation in the G + C bias of the mutation pressure acting on different regions of the genome (Wolfe et al., 1989), whereas in *D. melanogaster* it is thought to be due to selection acting on highly expressed genes which tend to have high G + C content at silent sites (Shields et al., 1988). In this case codon usage data alone are insufficient to distinguish between the action of mutation and selection.

(d) Comparison of $\hat{N}_c$ with other measures of SCU bias

A number of measures of codon usage bias have been developed previously. These vary in several respects. Some have been designed with a view to detecting coding regions in a stretch of DNA rather than quantifying SCU bias per se (e.g., Gribskov et al., 1984). Such measures typically use the entire sequence as input data, rather than the codon usage table, and move a 'window' along the DNA comparing observed codon usage with an expected codon usage pattern. Only methods that can be applied to data in the form of a codon usage table will be discussed here.

The choice of the expected or reference codon usage pattern also distinguishes between measures. The reference pattern used depends on which hypothesis we use regarding the action of mutation and selection on the gene under study:

(1) H_0: no selection
G + C bias at silent sites due to mutation

Expected SCU pattern: GC3s = G + C mutation bias
G = C; A = T; at 3rd codon position

H_0* (special case of H_0): no selection
no G + C mutation bias

Expected SCU pattern: equal usage of synonymous codons

(2) H_1: translational selection acting (i.e., selection for 'preferred' codons)

Expected SCU pattern: that of genes inferred to be under selection (e.g., highly expressed genes in *E. coli* and yeast)

Most measures use either a reference pattern in which all synonymous codons are used equally (i.e., H_0*), or one that expected in highly expressed genes (i.e., H_1). H_0*-based measures include $\hat{N}_c$ and a χ^2 'scaled' by gene length L_c

(i.e., χ^2/L_c) (Shields et al., 1988). $\hat{N}_c$ is less sensitive to short gene length and biased aa composition than χ^2/Lc (data not shown) due to the pooling of information from aa into SF types.

H_1-type measures have been developed based either on pooled data from known highly expressed genes (*P*, Gribskov et al., 1984; *V*, McLachlan et al., 1984; *CAI*, Sharp and Li, 1987), or based on the usage of identified 'preferred' codons by highly expressed genes (*Fop*, Ikemura, 1981; Bennetzen and Hall, 1982). The three H_1-type measures based on pooled data (*CAI*, *V*, and *P*) require a representative sample of codon usage from highly expressed genes. *CBI* and *Fop* are particularly appropriate when such data are not abundant. Most H_1-type measures have a range from 0 to unity and are thus easily interpreted. Both *P* and *V* measures, however, are not designed in this way.

Some measures are not restricted to one type of reference pattern. The codon preference bias (*V*, McLachlan et al., 1984) yields a distance in standard deviations from any given codon usage pattern. χ^2-based methods like χ^2/L_c (Shields et al., 1988) can be used with reference patterns other than H_0* (e.g., H_0). The choice between a H_0*-type or H_1-type SCU bias measure will, however, reflect in part the interests of the researcher.

(e) Conclusions

In this paper, particular emphasis has been placed on the easy interpretation of SCU bias measures. $\hat{N}_c$ provides an intuitively meaningful measure of the extent of codon preference in a gene. In addition, the Nc-plot provides a visual display of SCU variation for a set of genes. The Nc-plot can be used in conjunction with other measures. For example, H_1-type measures can be used to quantify the distance along a perceived H_0 to H_1 trend. $\hat{N}_c$ can easily be adapted to study genes that do not use the 'universal' genetic code.

ACKNOWLEDGEMENTS

I am grateful to Paul Sharp and Bill Hill for critical reading and constructive comments. I would also like to thank Paul Sharp for supplying much of the codon usage data analysed in this paper.

REFERENCES

Aota, S.-I. and Ikemura, T.: Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14 (1986) 6345–6355.

Bennetzen, J.L. and Hall, B.D.: Codon selection in yeast. J. Biol. Chem. 257 (1982) 3026–3031.

Bernardi, G. and Bernardi, G.: Compositional constraints and genome evolution. J. Mol. Evol. 24 (1986) 1–11.

Foster, P.L., Eisenstadt, E. and Cairns, J.: Random components in mutagenesis. Nature 299 (1982) 365–367.

Gouy, M. and Gautier, C.: Codon usage in bacteria: correlation with expressivity. Nucleic Acids Res. 10 (1982) 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A.: Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8 (1980a) r49–r62.

Grantham, R., Gautier, C. and Gouy, M.: Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8 (1980b) 1893–1912.

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

Gribskov, M., Devereux, J. and Burgess, R.R.: The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12 (1984) 539–549.

Ikemura, T.: Correlation between the abundance of *E. coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146 (1981) 1–21.

Ikemura, T.: Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 158 (1982) 573–597.

Ikemura, T.: Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2 (1985) 13–34.

Kimura, M. and Crow, J.F.: The number of alleles that can be maintained in a finite population. Genetics 49 (1964) 725–738.

Maruyama, T., Gojobori, T., Aota, S.-I. and Ikemura, T.: Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. 14 (1986) r151–r197.

McLachlan, A.D., Staden, R. and Boswell, D.R.: A method for measuring the non-random bias of a codon usage table. Nucleic Acids Res. 12 (1984) 9567–9575.

Muto, A. and Osawa, S.: The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84 (1987) 166–169.

Nei, M. and Tajima, F.: DNA polymorphism detectable by restriction endonucleases. Genetics 97 (1981) 145–163.

Sharp, P.M. and Li, W.-H.: Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. 14 (1986) 7737–7749.

Sharp, P.M. and Li, W.-H.: An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24 (1987) 28–38.

Sharp, P.M. and Devine, K.M.: Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. 17 (1989) 5029–5039.

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

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F.: Codon usage patterns in *E. coli*, *B. subtilis*, *S. cerevisiae*, *S. pombe*, *D. melanogaster*, and *H. sapiens*: a review of the considerable within-species diversity. Nucleic Acids Res. 16 (1988) 8207–8211.

Shields, D.C. and Sharp, P.M.: Codon usage in *Bacillus subtilis*. Nucleic Acids Res. 15 (1987) 8023–8040.

Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F.: 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5 (1988) 704–716.

Treffers, H.P., Spinelli, V. and Belser, N.O.: A factor (or mutator gene) affecting mutation rates in *E. coli*. Proc. Natl. Acad. Sci. USA 40 (1954) 1064–1071.

Wolfe, K.H., Sharp, P.M. and Li, W.-H.: Mutation rates differ among regions of the mammalian genome. Nature 337 (1989) 283–285.