A method for measuring the non-random bias of a codon usage table

Andrew D. McLachlan, Rodger Staden and D. Ross Boswell

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 4 October 1983; Revised and Accepted 13 November 1984

ABSTRACT

We describe a new statistical method for measuring bias in the codon usage table of a gene. The test is based on the multinomial and Poisson distributions. The method is used to scan DNA sequences and measure the strength of codon preference. For *E. Coli* we show that the strength of codon preference is related to levels of gene expression. The method can also be used to compare base triplet frequencies with those expected from the base composition. This second type of codon bias test is useful for distinguishing coding from non-coding regions.

INTRODUCTION

Many genes show a non-random selection of codons in their protein coding regions. There are several factors which influence the choice of a particular codon (1): the level of expression of the gene; the relative abundances of the tRNA's for each amino acid (2); the strength of codon-anticodon interaction (1); translocation of the ribosome (3); also possible dual roles for coding sequences, such as coding for overlapping genes (4) or forming particular secondary structures.

For any given protein we can distinguish at least two sources of bias in the codon usage. The first, "amino acid preference", is the uneven amino acid composition of typical proteins (5): some amino acids are used far more frequently than others. The second is that once an amino acid has been chosen there are generally preferences for the use of certain codons. This second level of bias, or "codon preference", is our main concern here. Grantham (6) and others have calculated tables that show which particular codons are used most in certain organisms, but our purpose here is to provide a measure of the strength of non-randomness in the pattern of codon use.

Even if two genes prefer the same codons their levels of preference may be very different. Our new method can be used to measure the level of constraint acting on the codon choices in each gene, and does not need to

make any strong assumptions about the nature of the bias being tested.

Our new statistical measure, the "codon frequency bias" estimates the probability that the actual codon frequencies observed in a gene could be the result of some postulated impartial pattern of usage. The statistical test is quite general, and can, in principle, be used to measure any kind of unevenness in the codon frequencies, comparing the observed frequencies with a predicted set, but in this paper we shall be mainly concerned with codon preference defined strictly relative to a given observed amino acid composition. The measure for this particular type of nonrandomness will be called the "codon preference bias". We shall see that the codon preference biases for different genes correlate well with their relative levels of expression.

We have also tested observed codon frequencies against the weaker assumption that they are determined solely by the base composition of the DNA. The corresponding measure of nonrandomness will be called the "base triplet bias" since it reflects the tendency of some triplets to occur more or less often than would be expected from the observed base composition rather than from the amino acid composition. We have used the base triplet bias with some success to distinguish coding from non-coding regions in genes.

TESTING THE SIGNIFICANCE OF THE CODON PREFERENCE

We need a statistical test which is precise and quick to apply to a relatively short piece of sequence (30-300 codons) in which the majority of the codons might be used only a small number of times. This requirement rules out tests based on chi-squared or similar approximations. The mathematical approach we have chosen is valid for small individual counts, provided only that the total number of codons in the sequence is large, and it uses a "codon frequency bias", V , which is large whenever the codon usage pattern is intrinsically improbable.

The Probability of a Given Codon Usage Pattern

Multinomial Probabilities We begin by introducing two distinct kinds of unbiased codon distribution, which can serve as reference standards. Consider a random DNA sequence of N codon triplets of T different types (normally T is 64) in which one expects to find codon c used with a certain relative frequency f_c . The value of f_c will depend on the nature of the sequence. For example, if the sequence is completely random, with a fractional base composition b_i for base i , then $f_c = b_1 b_2 b_3$ for the

codon (i,j,k). Alternatively, in measuring codon preferences, if the coded protein has a given amino acid composition A_s for residue s, and if there are d_s possible codons for this residue, all equally used, then we expect that each of these codons has $f_c = A_s/d_s$.

Given any sequence of exactly N codons we make a codon usage table by counting the number of times n_c that each codon appears. The probability of getting the resulting values $n_1, n_2, \dots, n_T$ is given by the multinomial distribution

$$M_N(n_1, n_2, \dots, n_T) = \frac{N! f_1^{n_1} f_2^{n_2} \dots f_T^{n_T}}{n_1! n_2! \dots n_T!} \quad [1]$$

with $n_1 + n_2 + \dots + n_T = N$. Our first trial measure of the improbability of a given codon usage table will be the value of

$$U = -\log M_N \quad [2]$$

which can be tested against the mean $\langle U \rangle$ and standard deviation, ΔU expected for random sequences with the given average frequencies f_c . In general, codon tables with large values of U will be rare; the actual number of times each U value occurs depends on how many different sets of codon counts ($n_1, n_2, \dots, n_T$) exist for each value of U. There will be few sets near the absolute minimum $U_{\min}$, many near the mean $\langle U \rangle$, and few for large values.

In practice we take a length of N codons in the DNA sequence. We compute the codon usage counts and the observed value of U. This number is converted to standard deviations to give the "codon frequency bias" value.

$$V = (U - \langle U \rangle) / \Delta U. \quad [3]$$

by using Eqns. [17] and [18] below. If we calculate the expected relative frequency of codon f_c using $f_c = b_i b_j b_k$ then we term V the "base triplet bias". If we assume impartial use of codons for each amino acid and use $f_c = A_s/d_s$ then we term V the "codon preference bias". In the limiting case where all the n_c are large, Stirling's formula can be used to show that the most probable codon usage pattern has counts of $n_c = N f_c$, with an absolute minimum for U of

$$U_{\min} = 1/2 (T-1) \log(2\pi) + 1/2 \log(N^T f_1 f_2 \dots f_T). \quad [4]$$

Unfortunately the precise probability distribution of U is difficult to calculate, because of the restriction that the total number of codons is exactly N. But we can derive more useful results by using the Poisson

distribution instead. In the argument below we use a device which is familiar in statistical mechanics: the replacement of a fixed value of N by a distribution or ensemble, in which only the mean value is specified. This makes the mathematics easier to handle.

Poisson Statistics Pretend for a moment that the random count n_c in each cell of the codon table is made up independently of the other cells subject only to the condition that the mean of n_c in each cell is specified to have the correct value $m_c = Nf_c$. The probability that codon c is used exactly r times then follows the Poisson distribution

$$p(r) = e^{-m} \frac{m^r}{r!}. \quad [5]$$

For the whole table, with T independent cells and unrestricted values of $n_1 \dots n_T$ the Poisson probability is

$$\begin{aligned} P(n_1, n_2, \dots, n_T) &= p_1(n_1) p_2(n_2) \dots p_T(n_T) \\ &= e^{-N} \frac{N^n f_1^{n_1} f_2^{n_2} \dots f_T^{n_T}}{n_1! n_2! \dots n_T!} \end{aligned} \quad [6]$$

with $n = n_1 + n_2 + \dots + n_T$.

Notice that the codon sum, n , for a table with independently chosen entries in its cells, is not exactly equal to N , but it does have a mean of N . In practice, when N (but not n_1 or $n_2 \dots$ separately) is large the sum of a Poisson generated codon table will be very close to N , with an error of order $N^{1/2}$. The quantity

$$W = -\log P \quad [7]$$

is useful for several reasons. The first is that W is closely related to U . It follows from Eqns. [1] and [6] that

$$P(n_1, n_2, \dots, n_T) = Q(n) M_n(n_1, n_2, \dots, n_T) \quad [8]$$

where

$$Q(n) = e^{-N} \frac{N^n}{n!} (2\pi N)^{-1/2} e^{-\left[\frac{(n-N)^2}{2N}\right]} \quad [9]$$

The factor $Q(n)$, which depends only on n and N , is just the global Poisson probability of getting a total count of exactly n when the mean count should be N . When N is large $Q(n)$ approximates to a Gaussian with a strong maximum at $n = N$. Therefore, for codon tables with a total of exactly N

$$W \approx U + 1/2 \log(2\pi N) \quad [10]$$

The second reason is that W , calculated for the unrestricted family of codon tables in which $n = n_1 + n_2 + \dots + n_T$ N is a sum of statistically independent terms for each codon

$$W = w_1(n_1) + w_2(n_2) + \dots + w_T(n_T) \quad [11]$$

$$w_c(r) = m_c + \log r - r \log m_c \quad [12]$$

This implies that the mean and variance of W are sums of independent contributions from individual codon counts with simple Poisson distributions.

$$\langle W \rangle = \sum_c \langle w_c \rangle \quad (\Delta W)^2 = \sum_c (\Delta w_c)^2 \quad [13]$$

It follows from Equ. [5] that for each codon

$$\begin{aligned} \langle w_c \rangle &= -p_0 \log p_0 - p_1 \log p_1 - \dots \\ \langle w_c^2 \rangle &= p_0 (\log p_0)^2 + p_1 (\log p_1)^2 + \dots \end{aligned} \quad [14]$$

and can be computed exactly. When the mean m is small the limiting values are $\langle w \rangle = m(1 - \log m)$ and $(\Delta w)^2 = m(\log m)$. For large m they are

$$\langle w \rangle = 1/2 [\log(2\pi) + 1] + 1/2 \log m, \quad (\Delta w)^2 = 1/2 \quad [15]$$

In a codon table for which all the cell counts are large these relations would give

$$\begin{aligned} \langle W \rangle &= 1/2 T [\log(2\pi N) + 1] + 1/2 \log(f_1 f_2 \dots f_T) \\ (\Delta W)^2 &= 1/2 T \end{aligned} \quad [16]$$

Equation [16] is not a good approximation in practice, because a codon table with a total of, say, $N = 100$ contains many small cell counts, but it does show that ΔW is insensitive to both N and the expected frequencies f_c . $\langle W \rangle$, on the contrary, depends strongly on the f_c values and hence on the base or amino acid compositions.

Probability Distributions of U and W The close relationship (Equ.[8]) between W for the unrestricted codon table and U for a table with exactly N counts can be used to derive relations between their mean values when N is large:

$$\langle U \rangle \approx \langle W \rangle + \frac{T}{4N} - \log(2\pi N)/2 \quad [17]$$

$$(\Delta U)^2 \approx (\Delta W)^2 - \frac{T^2}{4N} \quad [18]$$

The corrections of order $1/N$ appear because W is averaged over codon tables with a range of n close to N , while U has $n = N$ fixed.

In our calculations the required values for $\langle W \rangle$ and ΔW are made up from precalculated tables of $w_c(m)$ and $\Delta w_c(m)$ for the Poisson distributions with means m in the range 0-100 (see Eqns.[13] and [14]).

The exact probability distribution of W has also been computed by the method of probability generating functions. To do this it is first necessary to replace the continuous variables W and w_c by truncated integer approximations $W' = \{IW\}$, $w_c' = \{Iw_c\}$, where I is a large integer scaling factor (e.g. 100) and the $\{\dots\}$ means that IW is replaced by the nearest integer. For each codon we set up a generating polynomial of the type

$$g_c(x) = p(0)x^{\{-I \log p(0)\}} + p(1)x^{\{-I \log p(1)\}} + \dots \quad [19]$$

in which the coefficient of x^J gives the probability that $-I \log p_c(n_c)$ has the (approximate integer) value J . The generating polynomial for W' itself is then the product $G(x) = g_1(x)g_2(x)\dots g_T(x)$. The generating polynomials were calculated by algebraic multiplication, using a specially designed extended scale arithmetic program which can handle very large or small numbers (ranging over $\pm 7 \times 10^8$ powers of ten).

Trials with sets of 1000 random sequences with all four bases in equal abundance and using a sampling window of 99 codons showed that the probability distribution of W fits the theoretical one extremely well. The distribution (Figure 1) is Gaussian near the centre, but has a sharp cutoff on the low side, where W has a global minimum, and a very long tail at high values. The distribution of U is nearly the same shape, apart from the corrections in Eqns. [17-18].

Computer Program

The computer calculation of the "codon frequency bias" is performed over successive segments or "windows" which have a fixed length (typically 99 codons). For each position of the window the program estimates the

Figure 1. Theoretical probability distribution $P(W)$ of the base triplet bias measure W . Computed for a window length of 99 codons in a sequence with all four bases in equal abundance.

expected codon frequencies for a random sequence with the window's amino-acid composition. It then calculates the observed value of U and converts this into standard deviations from the theoretical mean to give the "codon preference bias" V defined in Equ.[3]. The values of V can be plotted along the length of the sequence for each of the three reading frames. In random sequences V will fluctuate about a mean value of zero with a standard deviation of unity. The program is written in FORTRAN 77 for a VAX computer and is part of Staden's ANALYSEQ program (7).

Codon Preference Bias in E. Coli Genes

As an example we examine the levels of constraint on codon choices in some genes of E. coli. Ikemura has shown for E. coli (3), and Ikemura (8) and Bennetzen and Hall (9) have shown for yeast that codon usage is linked to the abundances of tRNA's and to the strengths of interaction between codon and anticodon. Using a knowledge of observed codon frequencies in highly expressed genes both Ikemura and Bennetzen and Hall have produced tables of "optimal codons" or "preferred triplets" for these organisms. They have then measured the proportion of these codons that are used by several genes and shown that the measures correlate with the abundance of the proteins. Post and Nomura (10,11) have found that the codon usage in E. Coli ribosomal protein genes is linked to their level of expression. In Table 1 we compare our calculated "codon preference bias" for selected E. Coli genes with Ikemura's values for the proportions of optimal codons per gene and the numbers of protein molecules per genome. The numbers shown in column 2 of the table are the mean values for all the 99-codon

Table 1. Codon Preference Biases, Usage of "Optimal Codons" and Levels of Expression in E. Coli Genes

Gene	Codon Preference Bias	% Optimal Codons Used	Number of Protein Molecules Per Genome
rpl AJK	20.2	91	15,000
omp A	19.4	92	30,000
rec A	18.6	85	high
rpo B	17.8	84	2,000
trp B	12.8	70	medium
lac Y	12.4	61	low
ara C	11.5	54	low
thr A	10.9	61	medium
trp A	10.7	63	medium
lac I	10.6	63	low

windows that lie wholly within each gene. As can be seen the strengths of codon preference are well correlated with the proportion of optimal codons used. All sequences were taken from the EMBL nucleic acid library (12), and sources are given in Ikemura's paper (7).

We have also examined the unc operon of *E. coli* (13-15) which codes for the F1 and F0 portions of the ATP-synthase complex. F1 is composed of 5 subunits (which are, in order on the operon, delta, alpha, gamma, beta and epsilon) having a stoichiometry 1:3:1:3:1, and we have measured the codon preference biases for each of the corresponding genes. The values found are 13.2, 18.7, 14.6, 18.9 and 13.2, showing that the more highly expressed genes are the most biased.

Base Triplet Bias as a Test for Coding Regions

We tried plotting the "base triplet bias" along gene sequences, in an attempt to distinguish between coding and non-coding regions. We expected that strong preference for particular base triplets would only appear in coding regions. The plots were a useful way of testing for coding regions and assigning the correct reading frame, as well as for identifying peculiar stretches of DNA, but they are not quite as good as other methods which we have developed (16,17). The reason is that the bias test is so neutral or non-committal. It refuses to exploit any known regularities in the normal patterns of codon usage, unlike our previous test, which compared the observed usage with a typical biased pattern derived from a reference set of actual genes.

DISCUSSION

Our method measures the bias of a codon table relative to a set standard of impartial choice founded on either amino acid or base composition. It therefore measures the degree of nonrandomness without being committed to any particular assumptions about the choice of certain preferred codons. We have seen that for *E. Coli* proteins the most highly expressed genes are the least random in their choice of codons and the genes expressed at low levels are the most random. This indicates that the strongest constraint on codon choices is on the highly expressed genes which use the "optimal codons". We see for the genes encoding the F1 portion of the ATPase complex that even small differences in levels of expression are correlated with different codon preference bias values. Gouy and Gautier (18) have shown that expression is linked to the use of U and C in the third positions of codons. For genes of the F1 portion of

the ATPase complex that are made in equal amounts they find similar ratios of U and C in the third positions.

A new likelihood measure of codon preference was published by Gribkov, Devereux and Burgess (19) after our present work was complete. It is similar to ours in setting an unbiased standard, but treats the combined effects of base composition and amino acid preference in a different way. They test the ratio f_c/F_c , where F_c is the total frequency of all possible codons for the amino acid which is coded for by the individual codon c .

REFERENCES

1. Grosjean H. and Fiers W. (1982) *Gene* 18, 199-209.
2. Ikemura T. (1981) *J. Mol. Biol.* 151, 389-409.
3. Sedlacek J., Fabry M., Rychlik I., Volny D. and Vitek A. (1979) *Molec. Gen. Genet.* 172, 31-36.
4. Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes J.C., Hutchison C.A. III, Slocombe P.M. and Smith M. (1977) *Nature* 265, 687-695.
5. Dayhoff M.O. (1978) *Atlas of Protein Sequence and Structure Vol.5 Suppl.3*. National Biomedical Research Foundation, Washington D.C.
6. Grantham R., Gautier C., Gouy M., Jacobzone M. and Mercier R. (1981) *Nucl. Acids Res.* 9, r43-r74.
7. Staden R. (1984) *Nucl. Acids Res.* 12, 521-538.
8. Ikemura T. (1982) *J. Mol. Biol.* 158, 573-597.
9. Bennetzen J.L. and Hall B.D. (1982) *J. Biol. Chem.* 257, 3026-3031.
10. Post L.E., Strycharz G.D., Nomura M., Lewis H. and Dennis P.P. (1979) *Proc. Nat. Acad. Sci. USA* 76, 1697-1701.
11. Post L.E. and Nomura M. (1980) *J. Biol. Chem.* 255, 4660-4666.
12. EMBL Nucleotide Sequence Data Library, European Molecular Biology Laboratory, Postfach 10 22 09, D-6900 Heidelberg, West Germany.
13. Saraste M., Gay N.J., Eberle A., Runswick M.J. and Walker J.E. (1981) *Nucl. Acids Res.* 9, 5287-5296.
14. Gay N.J. and Walker J.E. (1981) *Nucl. Acids Res.* 9, 2187-2194.
15. Gay N.J. and Walker J.E. (1981) *Nucl. Acids Res.* 9, 3919-3926.
16. Staden R. and McLachlan A.D. (1982) *Nucl. Acids Res.* 10, 141-156.
17. Staden R. (1984) *Nucl. Acids Res.* 12, 551-567.
18. Gouy M. and Gautier C. (1982) *Nucl. Acids Res.* 10, 7055-7074.
19. Gribkov M., Devereux J. and Burgess R.R. (1984). *Nucl. Acids Res.* 12, 539-549.