Volume 10 Number 1 1982 Nucleic Acids Research

Codon preference and its use in identifying protein coding regions in long DNA sequences

R.Staden and A.D.McLachlan

Laboratory of Molecular Biology, The MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Received 16 September 1981

ABSTRACT

This paper describes a computer method that uses codon preference to help find protein coding regions in long DNA sequences. The method can distinguish between introns and exons and can help to detect sequencing errors.

INTRODUCTION

Now that DNA sequencing is one of the fastest information gathering techniques of molecular biology it is becoming increasingly common that, before the determination of a sequence, very little is known about the location of the genes for which it codes. We have previously reported a method for locating tRNA genes in DNA sequences[1] and here we describe a method to find protein genes. The criteria that can be used for this purpose are: (1) the existence of open reading frames (frames without stop codons); (2) the positioning of potential start codons; (3) codon preference; (4) the position of potential ribosome binding sites; (5) the position of potential splice junctions. An ideal method of analysis would use each of these criteria to give the probability that a section of sequence was coding in a particular reading frame. The method described here is a first step towards this in that it uses open reading frames, start codons and codon preference. The program can be used to determine which regions of a sequence code for proteins and hence can distinguish between introns and exons, and can also be used as an aid in checking for errors in newly determined sequences. Searching for open reading frames and start codons will not be discussed (although the results of these searches will be included in some of the figures) rather the bulk of the paper will be devoted to establishing the use of codon preference.

Since the first nucleic acid sequences were determined it has been noted that different organisms make different use of the redundancy of the genetic code: each organism has its own preferences. For example, the E. coli bacteriophage ØX174 strongly favours the use of codons with thymine in the third position of the triplet[2]. Grantham and coworkers have suggested that codon preferences will be common to all the genes within an organism or cell type and hence can be used to classify them[3].

The basic assumption of our codon preference method is that all the genes within the sequence we are analysing have similar codon preferences and that these preferences are sufficiently strong to be used as a means of discrimination. Below we detail our assumptions and then use these to derive the statistical relationships we require. Examples of the application of these equations to ØX174 and the nematode myosin gene[4] are then given to illustrate their use.

ASSUMPTIONS

For our statistical method we need to make the following assumptions:-

(1) We have a given DNA sequence which contains a collection of genes which have certain common features.

(2) The given sequence is coding throughout its length in only one reading frame but we do not know which frame this is.

(3) The collection of genes has a characteristic pattern of codon usage of the following type:

(a) In the actual reading frame the frequency of codon abc has a definite value f_{abc}. (Note this implies both a typical amino acid composition for the proteins coded and a certain set of codon preferences for any given amino acid.)

(b) The codons in the coding frame (and hence the amino acids in the protein) occur in random order. Note that this assumption applies only to the reading frame. In the other two frames 3(a) may apply, but in general there will be some tendency for consecutive triplets to be correlated.

(4) Out of all the triplets in the gene collection a fraction Q_1 are read in frame 1, Q_2 in frame 2 and Q_3 in frame 3.

STATISTICAL METHOD

If we choose a sequence randomly from the collection we can translate it in three frames. Suppose we select the following sequence:-

$$a_1b_1c_1a_2b_2c_2a_3b_3c_3 \cdots\cdots\cdots a_nb_nc_na_{n+1}b_{n+1}c_{n+1}$$

then from our assumptions 3(a) and 4 the probability of selecting it in each of the three frames if they were coding is:

Frame 1: $$p_1 = Q_1 f_{a_1b_1c_1} f_{a_2b_2c_2} \ldots\ldots f_{a_nb_nc_n} \quad (1)$$

Frame 2: $$p_2 = Q_2 f_{b_1c_1a_2} f_{b_2c_2a_3} \ldots\ldots f_{b_nc_na_{n+1}} \quad (2)$$

Frame 3: $$p_3 = Q_3 f_{c_1a_2b_2} f_{c_2a_3b_3} \ldots\ldots f_{c_na_{n+1}b_{n+1}} \quad (3)$$

We want to know the probability that selection of a particular sequence was "caused" by it being a coding sequence and this is found by using Bayes formula which gives the probability that the outcome of an experiment was due to a particular one of the assumed causes of the outcome. The probabilities that the selected sequence is in fact coding in each of the three frames are:

$$P_1 = p_1/(p_1+p_2+p_3) \quad (4)$$

$$P_2 = p_2/(p_1+p_2+p_3) \quad (5)$$

$$P_3 = p_3/(p_1+p_2+p_3) \quad (6)$$

It is convenient to work with logarithms of the frequencies and so we define:

$$F_{abc} = \log f_{abc} \quad (7)$$

The computer calculation

In practise when we analyse a DNA sequence we slide a "window" of length L triplets along it moving the window by one triplet at a time. For each of the positions of the window we calculate the sum of the F values over the length of the window for each of the three reading frames (each of the three sums is over the same number of triplets):

Frame 1: $$H_1 = \sum_L F_{a_ib_ic_i} \quad (8)$$

Frame 2: $$H_2 = \sum_L F_{b_ic_ia_{i+1}} \quad (9)$$

Frame 3: $$H_3 = \sum_L F_{c_ia_{i+1}b_{i+1}} \quad (10)$$

The probability that any window is in fact coding in each of three frames is then calculated using, for example, for frame 1:

$$P_1 = Q_1 e^{H_1} / (Q_1 e^{H_1} + Q_2 e^{H_2} + Q_3^{H_3}) \quad (11)$$

Applicability of the method

For the method to be useful we require that the inferred probability P for the coding frame be consistently higher than the probabilities for the other two frames and that the fluctuations in the value of P be relatively small. These values depend on the differences between the frequencies in the coding and non-coding frames and the length of the window L. In order to quantify these relationships we need to first consider how the codons in the coding frame influence the frequencies of the triplets in the other two frames.

Triplets in the non-coding frames

Define frame r as the coding frame and frame s as the second frame displaced one base to the right of r and frame t as the third frame displaced two bases to the right of r.

Consider the sequence:

$$a_1b_1c_1a_2b_2c_2 \,\ldots\ldots\, a_nb_nc_n$$

and let frame r coincide with codons of the form $a_ib_ic_i$. According to our assumptions the probability of obtaining this sequence in frame r is proportional to factors $f_{a_1b_1c_1}, f_{a_2b_2c_2} \ldots$

The triplets in frame s will be of the form $b_ic_ia_{i+1}$ and those in frame t of form $c_ia_{i+1}b_{a+1}$. As we assume that the codons in frame r are in random order the probability of obtaining bc in frame s is:

$$f'_{obc} = \sum_x f_{xbc}$$

and similarly the probability of obtaining a in frame s is:

$$f''_{aoo} = \sum_{yz} f_{ayz}$$

Therefore the triplet bca in frame s will occur with a frequency:

$$g_{bca} = f'_{obc} \cdot f''_{aoo} = \sum_x f_{xbc} \cdot \sum_{yz} f_{ayz} \qquad (12)$$

Similarly the frequencies in frame t are given by:

$$h_{cab} = f''_{ooc} \cdot f'_{abo} = \sum_{xy} f_{xyc} \cdot \sum_z f_{abz} \qquad (13)$$

We therefore see that it follows from assumptions 3(a), 3(b) that the frequencies of triplets in both the non-coding frames are determined implicitly by the frequencies in the coding frame. But note that the triplets in frames s and t are not generally in a strictly random order.

Average values for H_r, H_s and H_t

The average value of F per triplet in the read frame is given by:

$$\sum_{abc} f_{abc} \cdot F_{abc} = U \qquad (14)$$

and in frame s: $$\sum_{abc} g_{abc} \cdot F_{abc} = V \qquad (15)$$

and in frame t: $$\sum_{abc} h_{abc} \cdot f_{abc} = W \qquad (16)$$

So the average values of H over a window of length L in each of the three frames are:

$$A_r = LU = L \sum_{abc} f_{abc} \cdot F_{abc} \qquad (17)$$

$$A_s = LV = L \sum_{abc} g_{abc} \cdot F_{abc} \qquad (18)$$

$$A_t = LW = L \sum_{abc} h_{abc} \cdot F_{abc} \qquad (19)$$

Fluctuations about the mean

The dispersions or fluctuations about the means U, V, W are given by:

$$(DU)^2 = \Sigma f_{abc}(F_{abc}-U)^2 \qquad (20)$$

$$(DV)^2 = \Sigma g_{abc}(F_{abc}-V)^2 \qquad (21)$$

$$(DW)^2 = \Sigma h_{abc}(F_{abc}-W)^2 \qquad (22)$$

Therefore the fluctuations DA about the mean for a window of length L are given by:

in frame r: $$(DA_r)^2 = L(DU)^2 \qquad (23)$$

and frame s: $$(DA_s)^2 = L(DV)^2 \qquad (24)$$

and frame t: $$(DA_t)^2 = L(DW)^2 \qquad (25)$$

In order for the probabilities calculated to be useful we require that the difference between the coding frame and the other two frames is much larger than the sum of the random fluctuations in the three frames.

i.e. $$A_r - A_s >> DA_r + DA_s$$

and $$A_r - A_t >> DA_r + DA_t$$

Using 17, 18, 19, 23, 24, 25, we get:

$$L\ (U - V) >> \surd L\ (DU + DV)$$

or: $$L >> [(DU + DV)/(U - V)]^2 \qquad (26)$$

and similarly:

$$L >> [(DU + DW)/(U - W)]^2 \qquad (27)$$

Also we require that $A_r >> A_s$ or A_t and using 11, 17, 18, 19 we get three relations of the type:

$$P_r \approx Q_r e^{A_r}/(Q_r e^{A_r} + Q_s e^{A_s} + Q_t e^{A_t}) \qquad (28)$$

so that $P_r/P_s \approx (Q_r/Q_s)\, e^{L\,(U - V)}$

Therefore for P_r to greatly exceed P_s and P_t we require:

$$U - V >> 1 / L \text{ and } U - W >> 1 / L$$

so that $L >> 1 /(U - V)$ and $L >> 1 /(U - W)$ (29)

For example, if $P_r/P_s > 100$ then $L > 4.6 /(U - V)$

ESTABLISHING THE CODON FREQUENCIES

To be able to apply the method we need to have a table of codon frequencies that are representative of those that we expect from the sequence we wish to analyse. In general, no such figures will be available and also we could not be sure that they were indeed representative of our sequence. Therefore we have always used an internal standard - that is, to set up our table of frequencies we use the codon usage figures we find in the sequences around the sequence which we wish to analyse. If possible we use the frequencies we find in any of the genes whose location is already known but failing this we assume that any long open reading frame in our sequence is a protein gene and calculate our frequencies from over this region. Usually we would first try the longest open reading frame and use its frequencies to attempt to predict other genes; if the predictions were inconclusive we could try another open reading frame as a standard. In our experience it is apparent when a coding region is found that can be used as a standard: non-coding regions give wildly fluctuating plots but coding regions will give consistently high probabilities that coincide with open reading frames.

THE PROGRAM CALCULATIONS

Setting up the frequency tables

(1) Calculate codon totals M_{abc} over n genes or open reading frames giving K codons.

(2) Calculate frequencies from codon totals, i.e. $f_{abc} = M_{abc}/K$.

(3) Calculate natural logarithms of non-zero frequencies to give F's.

(4) Set logs of zero frequencies to log (1/K).

(5) Calculate mean of F's.

(6) Set F's for stop codons to mean of F's.

Scanning the sequence

(7) For each window position calculate H_1, H_2, H_3, i.e. the sum of the F's for the triplets which appear in this window for each of the three frames.

(8) Calculate the mean, A, of the H values.

(9) Subtract A from each of the H values. (This is done to make the H values closer on average to zero so that the exponentials calculated in the next step do not go out of range.)

(10) Calculate P_1, P_2, P_3 using equation (11). (The Q's are set to 1/3.)

(11) Calculate $\log_{10}$ [P/(1-P)] for P_1, P_2 and P_3.

(12) Plot results if required. There are two forms of output from the program: one is data for the plots shown below in Figures 2 and 3 and the other is three columns of figures (one for each reading frame) representing the probabilities for each window position. Also shown are positions of stop and start codons. These columns of numbers are often the most useful form of output because they enable the user to see precisely the positions of start and stop codons and changes of reading frame. The graphical output is more useful for gaining an overall idea of the organisation of the protein genes.

PREDICTION OF CODON FREQUENCIES FOR FRAMES S AND T

In this section we give an example which compares predictions, using equations (12) and (13), of codon frequencies in the non-coding frames with those actually found and then illustrate the use of some of the other equations derived above. For this example we have chosen gene H from ØX174.

Figure 1(a) shows the codon usage for gene H which is 329 codons long. These values for the codon counts in gene H were then substituted in equations (12) and (13) to calculate the expected codon counts in frames s and t. Figure 1(c) is for frame s and shows a table which contains three entries for each codon. The first value for a codon is the expected count calculated using equation (12), the second value is the observed count and the third value is a χ^2 term for these two values, i.e.

$$[(\text{observed-expected})^2]/\text{expected}$$

As can be seen, in general there is good agreement between calculated and observed so that most (50) of the χ^2 terms are <1 but three of them are <2. These three are underlined in the figure and the sum of their χ^2 terms is >27% of the total χ^2 sum (= 47.2).

Figure 1(d) shows similar results from frame t. In this frame 46 of the χ^2 terms are <1 but four are >2 and make up >33% of the total χ^2 sum of 51.6.

```
==================================
TTT  7.0 TCT 17.0 TAT  5.0 TGT  0.0
TTC  1.0 TCC  4.0 TAC  0.0 TGC  0.0
TTA  1.0 TCA  1.0 TAA  1.0 TGA  0.0
TTG  5.0 TCG  1.0 TAG  0.0 TGG  3.0
==================================
CTT 12.0 CCT  4.0 CAT  3.0 CGT  2.0
CTC  0.0 CCC  0.0 CAC  0.0 CGC  4.0
CTA  0.0 CCA  0.0 CAA 15.0 CGA  0.0
CTG  1.0 CCG  0.0 CAG 12.0 CGG  1.0
==================================
ATT 15.0 ACT 12.0 AAT 14.0 AGT  2.0
ATC  1.0 ACC  3.0 AAC  4.0 AGC  0.0
ATA  0.0 ACA  0.0 AAA 16.0 AGA  0.0
ATG 12.0 ACG  4.0 AAG  7.0 AGG  1.0
==================================
GTT 12.0 GCT 30.0 GAT 16.0 GGT 19.0
GTC  2.0 GCC 11.0 GAC  6.0 GGC 12.0
GTA  2.0 GCA  3.0 GAA  3.0 GGA  5.0
GTG  2.0 GCG  2.0 GAG 13.0 GGG  0.0
==================================
```

Fig 1a Codon totals for standard.

```
==================================
TTT -3.9 TCT -3.0 TAT -4.2 TGT -5.8
TTC -5.8 TCC -4.4 TAC -5.8 TGC -5.8
TTA -5.8 TCA -5.8 TAA -4.7 TGA -4.7
TTG -4.2 TCG -5.8 TAG -4.7 TGG -4.7
==================================
CTT -3.3 CCT -4.4 CAT -4.7 CGT -5.1
CTC -5.8 CCC -5.8 CAC -5.8 CGC -4.4
CTA -5.8 CCA -5.8 CAA -3.1 CGA -5.8
CTG -5.8 CCG -5.8 CAG -3.3 CGG -5.8
==================================
ATT -3.1 ACT -3.3 AAT -3.2 AGT -5.1
ATC -5.8 ACC -4.7 AAC -4.4 AGC -5.8
ATA -5.8 ACA -5.8 AAA -3.0 AGA -5.8
ATG -3.3 ACG -4.4 AAG -3.9 AGG -5.8
==================================
GTT -3.3 GCT -2.4 GAT -3.0 GGT -2.9
GTC -5.1 GCC -3.4 GAC -4.0 GGC -3.3
GTA -5.1 GCA -4.7 GAA -4.7 GGA -4.2
GTG -5.1 GCG -5.1 GAG -3.2 GGG -5.8
==================================
```

Fig 1b Logarithms of frequencies for standard.

```
=============================================================
TTT  6.  6. 0.0 TCT  1.  1. 0.3 TAT  0.  1. 0.8 TGT  3.  4. 0.5
TTC  8.  8. 0.0 TCC  1.  0. 0.7 TAC  0.  0. 0.5 TGC  3.  8. 6.8
TTA 13.  8. 1.8 TCA  1.  2. 0.7 TAA  1.  0. 0.8 TGA  6.  1. 3.7
TTG 19. 24. 1.1 TCG  2.  1. 0.3 TAG  1.  2. 0.4 TGG  8.  7. 0.2
=============================================================
CTT  9.  9. 0.0 CCT  3.  2. 0.1 CAT  1.  1. 0.3 CGT  1.  2. 1.1
CTC 10.  8. 0.5 CCC  3.  1. 1.3 CAC  1.  2. 2.7 CGC  1.  1. 0.0
CTA 17. 15. 0.3 CCA  5.  5. 0.0 CAA  1.  1. 0.0 CGA  2.  2. 0.0
CTG 26. 31. 0.8 CCG  8. 10. 0.8 CAG  2.  0. 1.7 CGG  3.  2. 0.3
=============================================================
ATT  5.  5. 0.0 ACT  1.  1. 0.1 AAT  5.  3. 0.7 AGT  4.  5. 0.1
ATC  6.  7. 0.1 ACC  2.  3. 1.1 AAC  6.  4. 0.5 AGC  5.  7. 0.6
ATA 11. 11. 0.0 ACA  3.  4. 0.6 AAA 10. 12. 0.6 AGA  9. 11. 0.5
ATG 16. 15. 0.1 ACG  4.  2. 1.1 AAG 15. 16. 0.1 AGG 13.  9. 1.5
=============================================================
GTT  3.  1. 1.5 GCT  2.  3. 0.3 GAT  1.  1. 0.1 GGT  1.  1. 0.1
GTC  4.  2. 0.8 GCC  3.  2. 0.1 GAC  1.  2. 1.7 GGC  1.  0. 0.8
GTA  6.  8. 0.4 GCA  4.  6. 0.6 GAA  1.  1. 0.1 GGA  1.  3. 1.9
GTG 10. 12. 0.6 GCG  7.  5. 0.4 GAG  2.  1. 0.6 GGG  2.  1. 0.6
=============================================================
```

Fig 1c Calculated, observed and X^2 for frame s.

```
=============================================================
TTT  7.  6. 0.2 TCT  7.  4. 1.1 TAT 14. 13. 0.1 TGT  9.  7. 0.6
TTC 12. 11. 0.1 TCC  2.  2. 0.0 TAC 10. 11. 0.1 TGC 24. 30. 1.6
TTA  3.  2. 0.4 TCA 16. 16. 0.0 TAA 21. 15. 1.8 TGA 20. 22. 0.3
TTG  2.  2. 0.1 TCG  4.  3. 0.1 TAG  2.  3. 1.4 TGG 19. 23. 1.0
=============================================================
CTT  2.  1. 0.5 CCT  2.  0. 1.9 CAT  4.  5. 0.2 CGT  3.  3. 0.1
CTC  3.  5. 0.8 CCC  1.  2. 3.4 CAC  3.  3. 0.0 CGC  7.  6. 0.1
CTA  1.  0. 0.9 CCA  4.  4. 0.0 CAA  6.  9. 1.5 CGA  6.  4. 0.4
CTG  0.  1. 0.7 CCG  1.  0. 1.0 CAG  0.  0. 0.4 CGG  5.  5. 0.0
=============================================================
ATT  2.  3. 0.5 ACT  2.  3. 0.7 AAT  4.  4. 0.0 AGT  3.  1. 1.0
ATC  3.  1. 1.6 ACC  1.  1. 0.3 AAC  3.  2. 0.2 AGC  7.  5. 0.4
ATA  1.  2. 1.5 ACA  4.  3. 0.4 AAA  6.  8. 0.8 AGA  5.  6. 0.1
ATG  0.  0. 0.4 ACG  1.  1. 0.0 AAG  0.  0. 0.4 AGG  5.  7. 0.7
=============================================================
GTT  3.  4. 0.6 GCT  3.  6. 4.8 GAT  5.  5. 0.0 GGT  4.  7. 3.5
GTC  4.  6. 0.5 GCC  1.  0. 0.8 GAC  4.  3. 0.1 GGC  9.  5. 1.7
GTA  1.  2. 0.6 GCA  6.  7. 0.2 GAA  8.  9. 0.1 GGA  7.  6. 0.3
GTG  1.  0. 0.6 GCG  1.  3. 2.0 GAG  1.  0. 0.6 GGG  7.  1. 5.1
=============================================================
```

Fig 1d Calculated, observed and X^2 for frame t.

These results are typical of the genes we have analysed in this way: in general the codon frequencies in the non-coding frames are determined by those in the coding frame but there are usually a few codons that have anomalous frequencies. Nussinov[5] has shown that there are distinct nearest neighbour frequencies in DNA sequences and that these doublet preferences are independent of the function of the sequence. These preferences will have some effect on our ability to predict frames s and t from frame r and may account for the few anomalous frequencies that we have found.

We can now use these codon tables for the three reading frames to calculate values for some of the variables in the equations derived above but first we have to deal with two classes of special codons. These are the stop codons and those codons that have zero frequency in the standard. It will often occur that our standard will contain a few codons with zero frequency and we have decided to set these to have a frequency of 1/ (the number of codons in the standard) rather than to bias the results by using zero frequencies. We have decided to try to make stop codons effectively neutral by setting their frequency to the mean value for the standard. This means that when scanning a sequence the probabilities based on codon preference and the positions of stop codons can be used as independent criteria in assessing the likelihood that a particular region is coding for a protein. Having made these adjustments to our frequency table we produce the table shown in Figure 1(b) which shows the F values of equation (7), i.e. the natural logarithms of the frequencies in the coding frame.

Using these values and similar tables for frames s and t we have evaluated equations (14), (15), (16), (20), (21) and (22) to get the expected values:

$$U = -3.46$$
$$V = -4.87$$
$$W = -4.81$$
$$DU = 0.76$$
$$DV = 1.03$$
$$DW = 0.98$$

The program also produces expected values for the means of log 10 [P/(1-P)] for different values of the window length L for each of the three reading frames. For the ØX174 plots shown in the next section a window length of 30 codons was used and the expected values were calculated to be: in frame r, 17.0; in frame s, -18.2 and in frame t, -17.6. Examination of these values

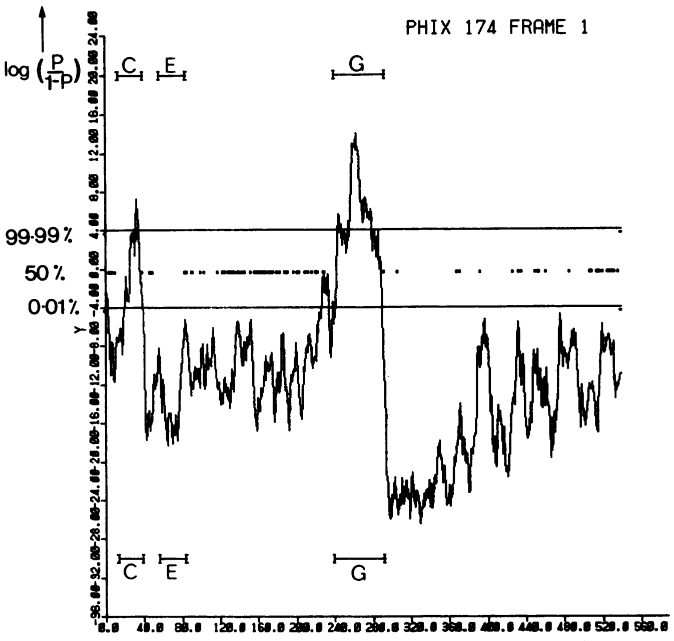
PHIX 174 FRAME 1
log $(\frac{P}{1-P})$
99·99%
50%
0·01%
Y
C
E
G
24.00
20.00
16.00
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
-20.00
-24.00
-28.00
-32.00
-36.00
0.0
40.0
80.0
120.0
160.0
200.0
240.0
280.0
320.0
360.0
400.0
440.0
480.0
520.0
560.0
Sequence Position x 10 ⟶

Fig 2a

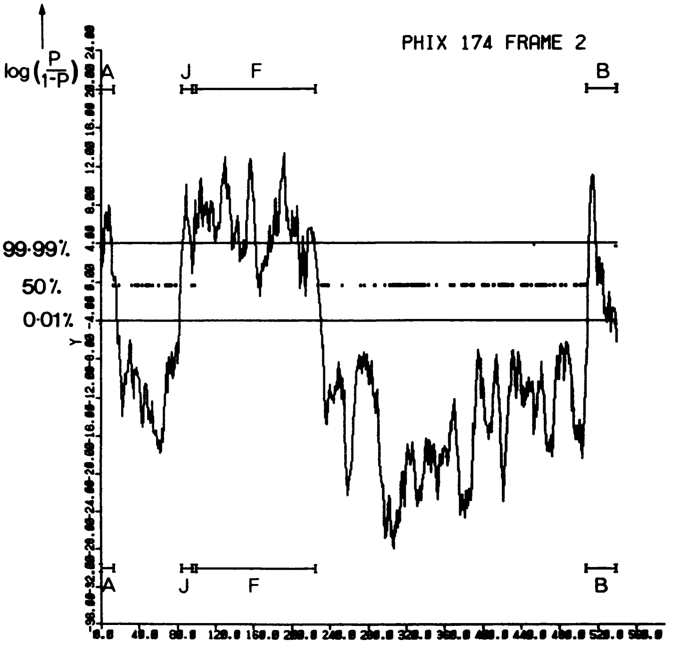
PHIX 174 FRAME 2
log $(\frac{P}{1-P})$
99·99%
50%
0·01%
Y
A
J
F
B
24.00
20.00
16.00
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
-20.00
-24.00
-28.00
-32.00
-36.00
0.0
40.0
80.0
120.0
160.0
200.0
240.0
280.0
320.0
360.0
400.0
440.0
480.0
520.0
560.0
Sequence Position x 10 ⟶

Fig 2b

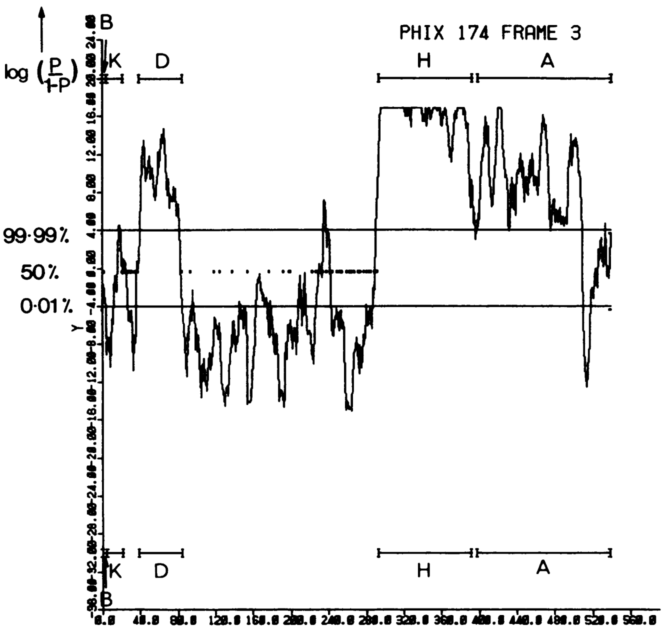

Fig 2c

can tell us how useful the codon preferences in our sequence will be as a means of discriminating between coding and non-coding regions and also how much fluctuation we can expect. These values show that we should get good discrimination if all the genes in ØX174 have similar codon usage.

USE OF THE METHOD TO IDENTIFY READING FRAMES

To demonstrate the program Figure 2 shows results using gene H as a standard (and hence the frequency tables of Fig. 1) to perform an analysis of the sequence of ØX174. There is a separate plot for each of the three reading frames. The x axis represents the length of the sequence from the first to last base and on the y axis is plotted $\log_{10}$ [P/(1-P)] where P is the probability of coding. Also marked on the y axis are the values corresponding to probabilities at the 99.99% and 0.01% level and along the 50% level each stop codon for the reading frame is marked with a dot. All these are produced by the program but to show the genes for each frame horizontal bars have been drawn at the top and bottom of the plots. To identify potential genes we look for consistently high probabilities that coincide with open reading frames (shown by an absence of dots on the 50% level). Frame 1 is shown in Figure 2(a).

Genes C, E and G are read in frame 1 and are shown by the bars on the graph. Examination of the plot shows that there is a peak for gene C except where it overlaps with gene K which is read in frame 3; there is no peak for gene E which is entirely contained in gene D which is read in frame 3; and there is a good peak for gene G. There are no false peaks. Frame 2 is shown in Figure 2(b). Genes A, J, F and B are read in frame 2. There are peaks for all of these genes except for parts of B that overlap with gene A. There are no false peaks. Frame 3 is shown in Figure 2(c). Genes K, D, H and A are read in frame 3. Again there are peaks for all of these genes except in regions of overlap. There is one false high peak but as can be seen this is heavily blocked by stop codons. These results show that the method can correctly identify non-overlapping genes but that in regions of overlap only one gene will be predicted. This is as expected from the assumptions.

USE OF THE METHOD TO DISTINGUISH BETWEEN EXONS AND INTRONS

One of the projects currently being undertaken in this laboratory is the sequencing of the *unc*-54 myosin heavy chain gene from the soil nematode *C. elegans*. This DNA contains intervening sequences and the reading frame scanning program is being used to help identify exons. Some of the exons have been identified by comparison of cDNA and genomic sequences, while others have been tentatively located by Berk-Sharp hybridisation experiments. These regions can be used as an internal standard for the scanning program and Figure 3 shows plots for the myosin gene using the region 2693-6358 as a standard. This region (2692-6358) contains myosin coding sequence from the S-1 region of the heavy chain which has been confirmed by protein sequence homology to rabbit skeletal myosin sequences (J. Karn, unpublished).

In frame 2 there are no known exons and the plot shows no peaks of any length at the 99.99% level. In frame 3 there are known to be four exons and the program has correctly identified the four separate regions at 99.99% even though the length of the intron between the two rightmost exons is less than 60 bases. In frame 1 there are two peaks at the 99.99% level that coincide with open reading frames; the right one has been positively identified as coding sequence and the left one is currently under investigation. The plot, therefore, directs the sequencer's attention to this particular region. To summarise the myosin results: the plots correctly identify all the exons (with the possible exception of the region currently under investigation) and give no false predictions. The intron/exon boundaries can be located to within a few codons.

USE OF THE METHOD TO DETECT SEQUENCING ERRORS

In all the discussion of the method we have assumed that the sequence being analysed is perfectly correct. This is not always the case and our method has proved useful as a means of detecting errors in newly determined sequences. The easiest class of errors to detect in this way are insertions or deletions of length ≠3n where n is an integer which, if occurring in a coding region, would cause the plots to change frame. If the plots lead us to believe a region is coding but we cannot find the requisite signals, such as start and stop codons, base change errors in these signals may also be detected.

DISCUSSION

In this paper we have described a method to locate protein genes using codon preference and demonstrated its use on the sequences of ØX174 and nematode myosin. We have also applied the method to coli phage G4[6], fd[7], the *unc* operon of *E. coli*[8] and the beef[9] and human[10] mitochondrial DNA sequences. The results of these analyses are not shown but, with three exceptions, are similar to those for ØX174 and the nematode myosin. The exceptions are the two proposed genes from human and beef mitochondria that have each been named *urf6*, and are the only proposed genes on the heavy strands of these genomes, and gene 1 from the *unc* operon of *E. coli*. In each case these genes have very different codon usage to the other genes around them. It is worth noting that if we take *urf6* from the human sequence as a standard it does predict *urf6* in the beef sequence to be a gene.

The method is based on several assumptions and we now discuss these and any limitations they may impose on it. Although the method is based on codon preference we have also taken advantage of the fact that some amino acids are more common than others. In the plots shown the amino acid content of the standard has not been removed but has been left in to aid in the predictions. This will, of course, only be helpful if assumption (1) is true, i.e. that the amino acid content of the standard is similar to that of the genes for which we are searching. We have found no case where differences in amino acid content between the standard and the genes being scanned has badly affected the predictions. In the case of the *unc* operon of *E. coli*, genes for hydrophobic proteins have been successfully used to locate hydrophilic genes, and *vice versa*.

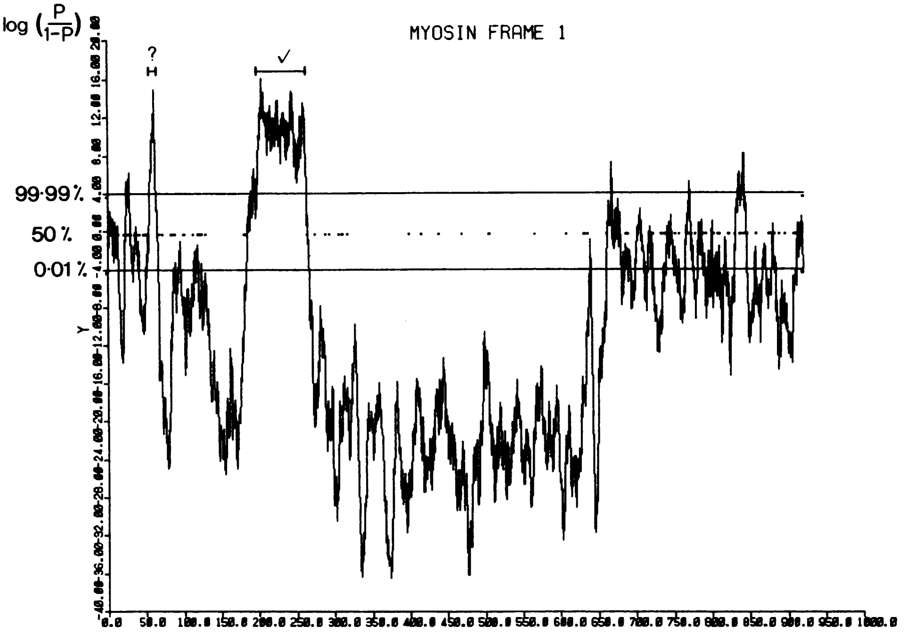
MYOSIN FRAME 1
log $(\frac{P}{1-P})$
Y
99·99%
50%
0·01%
?
H
✓
20.00
16.00
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
-20.00
-24.00
-28.00
-32.00
-36.00
-40.00
0.0
50.0
100.0
150.0
200.0
250.0
300.0
350.0
400.0
450.0
500.0
550.0
600.0
650.0
700.0
750.0
800.0
850.0
900.0
950.0
1000.0

Fig 3a

Sequence Position x 10

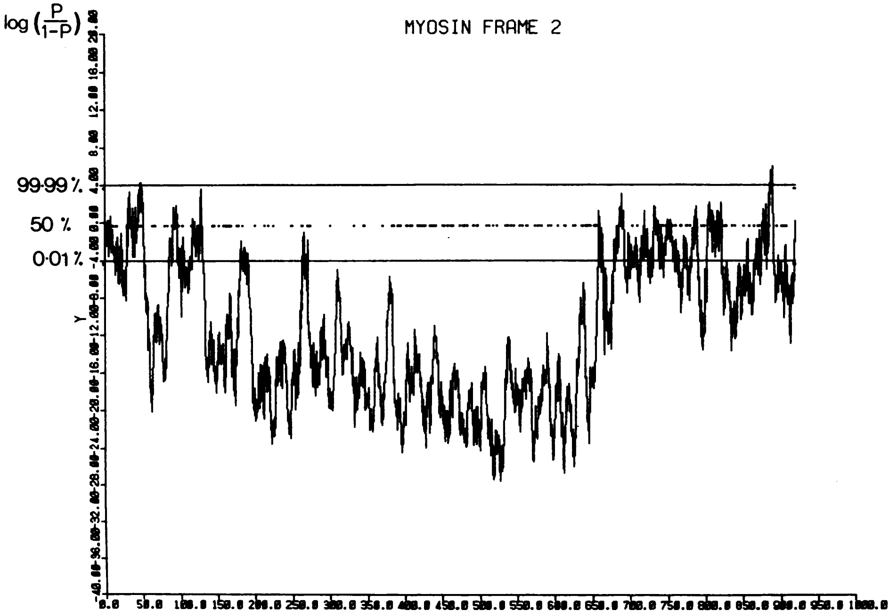
MYOSIN FRAME 2
log $(\frac{P}{1-P})$
Y
99·99%
50%
0·01%
20.00
16.00
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
-20.00
-24.00
-28.00
-32.00
-36.00
-40.00
0.0
50.0
100.0
150.0
200.0
250.0
300.0
350.0
400.0
450.0
500.0
550.0
600.0
650.0
700.0
750.0
800.0
850.0
900.0
950.0
1000.0

Fig 3b

Sequence Position x 10

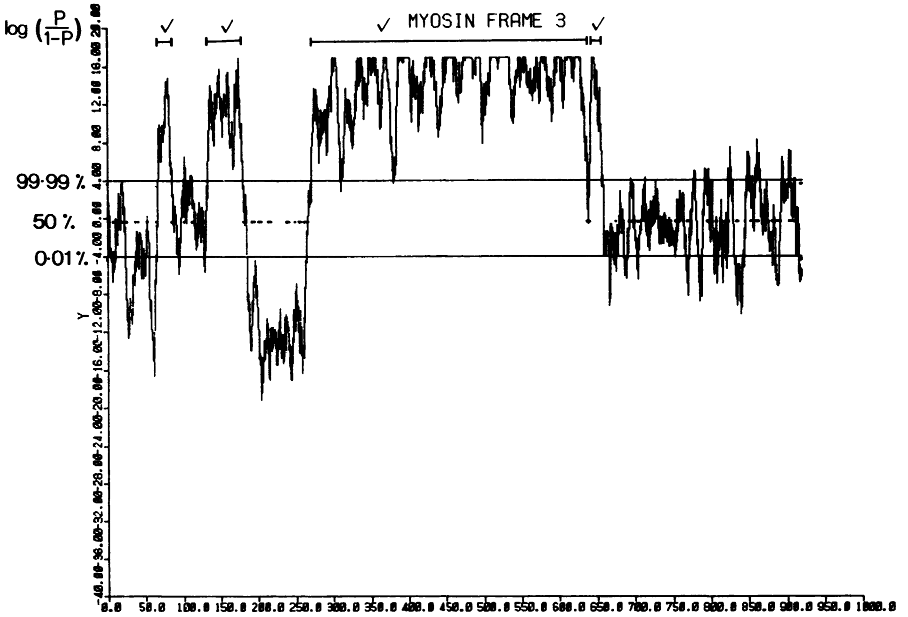

Fig 3c

Sequence Position x 10

Assumption 3(a), that the codon usage is characteristic for related genes in the same genome, is the cornerstone of the method, and as was mentioned in the Introduction, is based on early observations that have since been more formally stated by Grantham. In apparent conflict with Grantham's results, Wain-Hobson has shown that "preferential use of degenerate codons is highly gene-specific, and that few general trends are apparent"[11]. With the exceptions mentioned our results indicate that for the limited number of sequences we have analysed there is generally sufficient similarity of codon usage between genes of the same genome for it to be useful for predicting protein genes.

At least two other methods of predicting protein coding regions - Shepherd[12] and Shulman[13] - have been published and another is under development (Fickett, personal communication). Shepherd's method is based on looking for those reading frames that differ least from a proposed primitive coding message whose codons are of the form RNY where R = purine, Y = pyrimidine and N = purine or pyrimidine. Shulman's method applies statistical tests to indicate the correct reading frame. It is not yet possible to compare the abilities of the different methods.

As was mentioned in the Introduction, this method is meant to form part of a comprehensive system of searching for protein coding regions. We hope to add searches for ribosome binding sites and splice junctions so that the sequencer can use several different independent criteria, none of which is necessarily completely convincing on its own, to decide if a region of sequence is coding for a protein.

Acknowledgements

I would like to thank F. Sanger for advice throughout the course of this work and J.E. Walker, J. Karn and S. Anderson for critical reading of this manuscript.

Machine and language

The programs described are written in FORTRAN and run under the vms operating system on a VAX 11/780 computer manufactured by Digital Equipment Corporation.

REFERENCES

1. Staden, R. (1980) Nucleic Acids Res. 8, 817-825.
2. 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. (1978) Nature 276, 236-247.
3. Grantham, R., Gautier, C., Mercier, R. and Pave, A. (1980) Nucleic Acids Res. 8, r49-r62.
4. Macleod, A.R., Karn, J. and Brenner, S. (1981) Nature 291, 386-390.
5. Nussinov, R. (1980) Nucleic Acids Res. 8, 4545-4562.
6. Godson, G.N., Barrell, B.G., Staden, R. and Fiddes, J.C. (1978) Nature 276, 236-247.
7. Beck, E., Sommer, R., Auerswald, E.A., Kurz, Ch., Zink, B., Osterburg, and Schaller, H. (1978) Nucleic Acids. Res. 5, 4495-4503.
8. Gay, N.J. and Walker, J.E. (1981) Nucleic Acids Res. 9, 3919-3926.
9. Anderson, S., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Eperon, I.C., Sanger, F. and Young, I.G. (in preparation).
10. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F. , Schreier, P.H., Smith. A.J.H., Staden, R. and Young, I.G. (1981) Nature 290, 497-465.
11. Wain-Hobson, S., Nussinov, R., Brown, R.J. and Sussman, J.L. (1981) Gene 13, 355-364.
12. Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. USA 78, 1596-1600.
13. Shulman, NM.J., Steinberg, C. and Westmoreland, N. (1981) J. Theor. Biol. 88, 409-420.