Volume 12 Number 1 1984 Nucleic Acids Research

The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression

Michael Gribskov, John Devereux[1] and Richard R.Burgess

McArdle Laboratory for Cancer Research, and [1]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA

Received 18 August 1983

ABSTRACT

The codon preference plot is useful for locating genes in sequenced DNA, predicting the relative level of their expression and for detecting DNA sequencing errors resulting in the insertion or deletion of bases within a coding sequence. The three possible reading frames are displayed in parallel along with the open reading frames and plots of the location of rare codons in each reading frame.

INTRODUCTION

Due to the degeneracy of the genetic code, most amino acids are specified by more than one codon (synonomous codons). Synonomous codons are not used at equal frequencies, their relative frequency varying with both the gene and the organism (1,2,3,4,5). In yeast and E. coli, and presumably in other organisms, there is a strong correlation between the frequency of a codon and the abundance of the corresponding tRNA (6,7).

In E. coli, genes of highly expressed proteins, e.g. ribosomal proteins or major outer membrane proteins, use codons corresponding to the most abundant tRNAs almost exclusively (6,7,8,9). This is thought to be due to a need for efficient translation of RNAs of abundant proteins (10). Proteins expressed at a low level, e.g. lac repressor, use synonomous codons in rough proportion to the abundance of the corresponding tRNAs, resulting in a smaller, codon preference (1). In contrast, non-coding regions of E. coli DNA show no pronounced preference for any trinucleotide.

This difference of codon usage between genes and non-genes has been used to identify coding frames in DNA sequences. The method we present here is similar to that of Staden and McLachan (11), but has several advantages over their technique.

CONSTRUCTION OF CODON PREFERENCE PLOTS

The codon preference plot is constructed by calculating a codon preference statistic for each position of each of three reading frames. The statistic is calculated over a window of length w and the window moved along the sequence in increments of three bases, maintaining the reading frame. The magnitude of the codon preference statistic is a measure of the likeness of a particular window of codons to a predetermined preferred usage (see below, codon frequency tables).

In addition to the codon preference statistic, two other types of relevant information are shown. The locations of open reading frames starting with AUG codons are plotted on the same scale as the codon preference statistic. The location of rare codons are plotted in parallel to this plot. Non-coding regions usually have a high incidence of rare codons, those whose usage makes up 5% or less of the synonomous family, but coding regions, even those of weakly expressed genes, have a much lower incidence of rare codons.

Calculation of the Codon Preference Statistic We will consider two frequencies in our analysis: the frequency of a single codon, which we will denote by a lower case letter, and the frequency of all the codons for an amino acid (synonomous family), which will be denoted by a capital letter. The frequencies refer to the number of the codons, or family of codons, of interest, divided by the total number of codons in the sample. Consider a trinucleotide, or codon, abc, where a, b, and c are the three bases in a codon. The frequency of codon abc, f_{abc}, is found in the codon frequency table. The frequency of the synonomous family including codon abc, F_{abc} is given by

$$F_{abc} = \sum_{family} f_{abc}$$

i.e. the sum of the frequencies of all members of the family.

Similar definitions can be made for a random sequence of the same base composition as the sequence of interest. We use a random codon usage predicted from the base composition of the sequence of interest to account for sequences with asymmetric base composition. If we define N_i as the number of bases i in the sequence, then r_{abc}, the frequency of codon abc in a random sequence, is given by

$$r_{abc} = N_a N_b N_c / N^3 \text{ and } R_{abc} = \sum_{family} r_{abc}$$

where N is the total number of bases in the sequence.

The preference parameter, for a codon abc is then found by

$$p = \frac{f_{abc}/F_{abc}}{r_{abc}/R_{abc}}$$

When calculated in this way, the preference parameter can be considered to be a likelihood ratio. In other words, p is the relative likelihood of a codon being found in a gene as opposed to a random DNA sequence. The product of likelihood ratios is the most sensitive method of determining the membership of a group of observations in one of two classes, in this case, genes or random sequences (12).

The codon preference statistic (P) is the wth root of the product of p for each codon in the window (length = w). Logarithms are used to simplify

$$P = e^{(\sum_{i=0}^{w} \log p_i)/w} = (\prod_{i=0}^{w} p_i)^{1/w}$$

the calculation. Although the division by the window length is not required on theoretical grounds, it has the effect of normalizing the codon preference statistic so that the magnitude of P is less dependent on window length. The amount of fluctuation in P over an entire gene is obviously dependent on the window length, w. We have found that the most useful value of w varies with the length of the DNA sequence, generally values of 25 for a sequence less than 5000 bp, and 50 for longer sequences are adequate. In general, one desires to use the smallest value of w that gives discrimination between genes and non-genes.

Codon Frequency Tables We used the Genbank nucleic acid sequence database (13) to compile tables of codon frequencies in genes of enteric bacteria. Any one gene has too small a number of codons to allow accurate computation of the relative frequencies of codons in a synonomous family, therefore, we have pooled the frequencies of several genes having similar codon usage patterns. The group of genes to pool was determined by a method similar to that of Grantham et al. (3). The similarity of each pair of genes in our sequence collection was evaluated by calculating a similarity parameter, S,

$$S = \sum_{a,\,b,\,c} \left(\frac{f_{abc,1}}{F_{abc,1}} - \frac{f_{abc,2}}{F_{abc,2}} \right)^2$$

where 1 and 2 refer to the frequencies in the two genes of interest, and the sum is taken over all codons. When an entire synonomous family of codons was lacking in one gene, these codons were omitted from the sum.

In this way, a group of genes with similar codon usage was found. This group corresponds to the "highly expressed" group of Grantham (3) and includes the E coli lamB, lpp, ompA, ompF, recA, rplA, rplK, rpoB, rpoC, rpsA, tufA, tufB, and uncA and the Erwina amylovora and Salmonella typhimurium lpp genes. These genes all show a disproportionate use of codons corresponding to the most abundant tRNAs. We therefore consider the pooled table of codon frequencies from this group to be characteristic of optimal codon usage.

When this "highly expressed" table is used as a reference, p varies from 5.0 for the arginine CGU codon to 0.006 for the glycine GGA codon. Rare codons, those making up 5% or less of their synonomous families in this table are arginine codons AGA, AGG, CGA, and CGG, glycine codons GGA and GGG, isoleucine codon AUA, leucine codons CUA, CUU, UUA, and UUG, proline codon CCC, serine codons AGU, UCA, and UCG, and threonine codon ACA. A codon whose p value equalled zero would cause the codon preference statistic to be zero

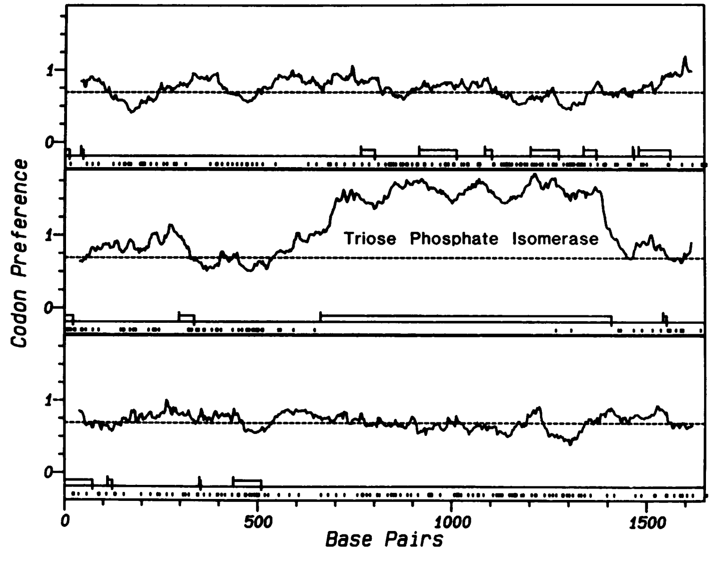

Fig. 1. Codon preference plot of the S. cerevisiae triose phosphate isomerase gene. The plot is in three sections each indicating a different reading frame. The dashed line indicates the calculated codon preference statistic for a theoretical random sequence of the same composition as the codon frequency table. Beneath the codon preference statistic, open reading frames are indicated by boxed regions. AUG codons are marked as vertical lines not crossing the horizontal, and stop codons as vertical lines crossing the horizontal. Rare codons, 5% of synonomous family or less, are indicated by the short bars beneath the open reading frame plot. Window length (w) = 25.

for all windows including the codon. Since the absence of a codon means that the codon is rare, not that it is never used, these codons are assigned a frequency (f) equal to the reciprocal of the number of codons in the synonomous family. For instance, if the codon frequency table contained 50 arginine codons, but no AGA codons, f_{AGA} would be 0.02.

Although we have found that the pooled highly expressed gene usage is the most effective frequency table to use, it is not essential. Codon frequency tables assembled from small numbers of genes or from moderately expressed genes can be used. Fig. 1 shows a plot for the yeast triose phosphate isomerase gene (13) in which the codon frequency table is made up only of the S. cerevisiae histone H2a1, H2a2, H2b1, H2b2, and HIS4 genes.

USE OF CODON PREFERENCE PLOTS

The most obvious use of codon preference plots is to determine the location of genes in a DNA sequence. Fig. 2 shows a codon preference plot of the rpoBC operon (13,14) of E. coli. This operon is composed completely of ribosomal protein and RNA polymerase genes, i.e. highly expressed genes, and

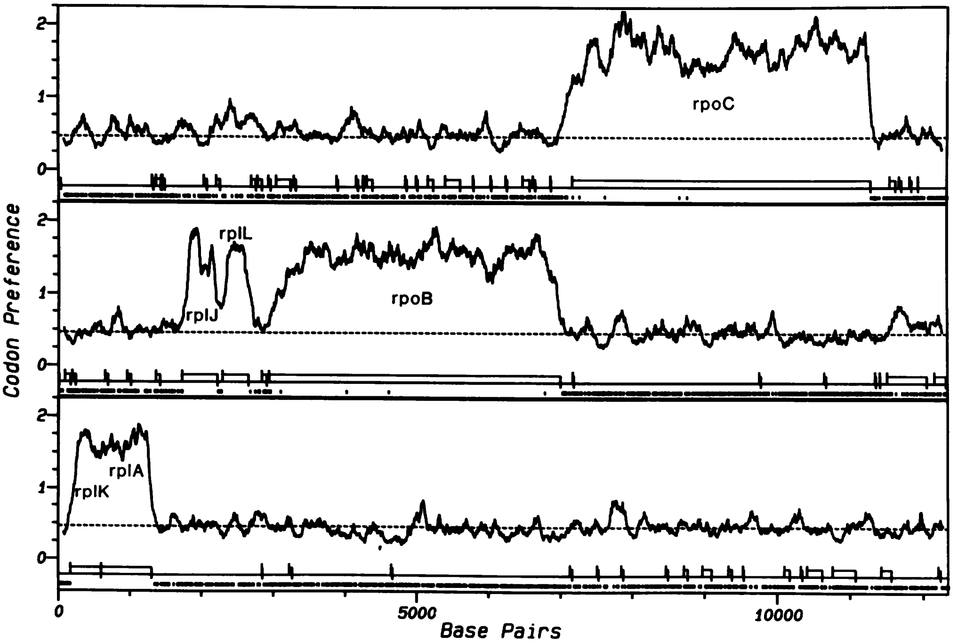

Fig. 2. Codon preference plot of the E. coli rpoBC operon. Rare codons, 2.5% of synonomous family or less, are marked, w = 50.

the location of each gene can be easily seen. The absence of rare codons in each reading frame is especially striking. A more complex situation can be seen in the E. coli rpoD operon (13,15,16) Fig. 3. The three genes in this operon are expressed at widely differing levels, rpsU, encoding ribosomal protein S21 is highly expressed (50,000 molecules of protein/cell) (17), dnaG, encoding DNA primase (50 molecules/cell) (18), is very weakly expressed, and rpoD encoding the sigma subunit of RNA polymerase (3000 molecules/cell) (19,20), is expressed at an intermediate level. This difference in expression level is clearly shown in the plot. The high level of rare codons in the dnaG gene has led to the hypothesis that inefficient translation is important in the regulation of primase expression (21).

These kinds of plots are obviously of only limited use in locating a gene which is known to be in a given segment of DNA, but one often wants to know if there are genes of unknown function in the region of a sequenced gene. The codon preference plot gives information on whether open reading frames near a gene of interest actually encode a protein. For instance, in Fig. 4 an open reading frame can be seen from bases 170 to 670, however the low level of the

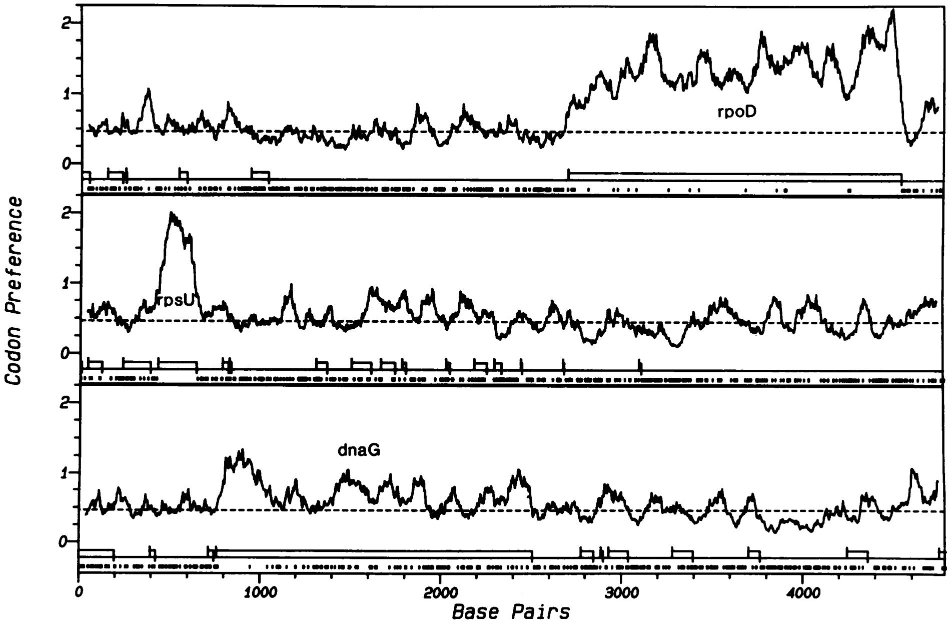

Fig. 3. Codon preference plot of the E. coli rpoD operon. Rare codons, 5% or less of synonomous family, are marked, w = 25.

Table 1: Overall Codon Preference Statistics for Enteric Bacterial Genes and Random Sequences of the Same Composition

Gene[1]	P	Random P	bp
lpp	1.758	0.458	234
tufA	1.735	0.463	1182
rpsU	1.704	0.447	213
tufB	1.691	0.459	1182
rpsA	1.664	0.446	1668
rplA	1.637	0.439	702
ompA	1.595	0.467	1038
lpp Erwinia amylovora	1.591	0.457	234
rpoC	1.589	0.46	4221
rplL	1.565	0.415	363
rpsL	1.551	0.467	372
rplK	1.497	0.451	426
uncA	1.492	0.475	1539
lpp Serratia marcesans	1.488	0.46	231
rpoB	1.461	0.458	4026
rplJ	1.414	0.46	495
ompF	1.395	0.462	1086
recA	1.382	0.431	1059
ssb	1.367	0.438	534
rpoD	1.339	0.441	1839
glnS	1.244	0.461	1653
frdB	1.212	0.461	732
lamB	1.191	0.461	1338
crp	1.191	0.452	624
metG	1.148	0.462	1746
trpS	1.102	0.452	1002
fnr	1.061	0.465	750
asnA	1.019	0.454	990
ompR	0.995	0.477	852
trpG Serratia marcesans	0.991	0.447	579
trpB	0.989	0.459	1191
purF	0.987	0.474	1512
nifH Klebsiella pneumoniae	0.983	0.434	879
hisQ Salmonella typhimurium	0.978	0.450	780
lexA	0.975	0.447	606
aroG	0.935	0.447	1050
ilvG	0.912	0.468	1563
fol	0.872	0.451	477
lacY	0.865	0.509	1251
trpA Klebsiella aerogenes	0.865	0.448	807
hisP Salmonella typhimurium	0.849	0.447	774
trpB Salmonella typhimurium	0.848	0.457	1191
trpA	0.839	0.474	804
envZ	0.811	0.447	1179
thrA	0.794	0.472	2460
trpA Salmonella typhimurium	0.791	0.485	804
trpC	0.787	0.465	1356
argT Salmonella typhimurium	0.778	0.456	780
lacI	0.761	0.464	1080
hisM Salmonella typhimurium	0.737	0.503	705
dnaG	0.700	0.461	1743
hisJ Salmonella typhimurium	0.685	0.492	684
araC	0.666	0.477	876
ampC	0.659	0.456	1131
trpR	0.618	0.449	264
araC Salmonella typhimurium	0.593	0.477	843

[1] All genes are E. coli unless otherwise stated.

codon preference statistic would lead us to predict that this is a non-coding frame, or a very weakly expressed gene.

When the window length w is the length of the gene an overall codon preference statistic can be calculated for the entire gene. This overall codon preference statistic is similar to those calculated by others (4,7), and is apparently a function of the maximum level of expression of the gene. Our statistic differs from others in assigning a range of values to the codons rather than simply 1 or 0 for "preferred" vs "unpreferred". As shown in Table 1, the overall codon preference statistic for strongly expressed genes such as ribosomal protein genes (rpsU, rpsA, rplL, rplJ) or membrane protein genes (lpp, ompA, ompF) is high while genes for weakly expressed genes such as repressors (lacI, araC, trpR) or other genes expressed at low levels (dnaG) are low. It is worth noting that the recA gene, which is transiently expressed at a high level, has an overall codon preference statistic typical of a strongly expressed gene. Information on the expression level of an unidentified gene may be of substantial benefit in identifying its protein

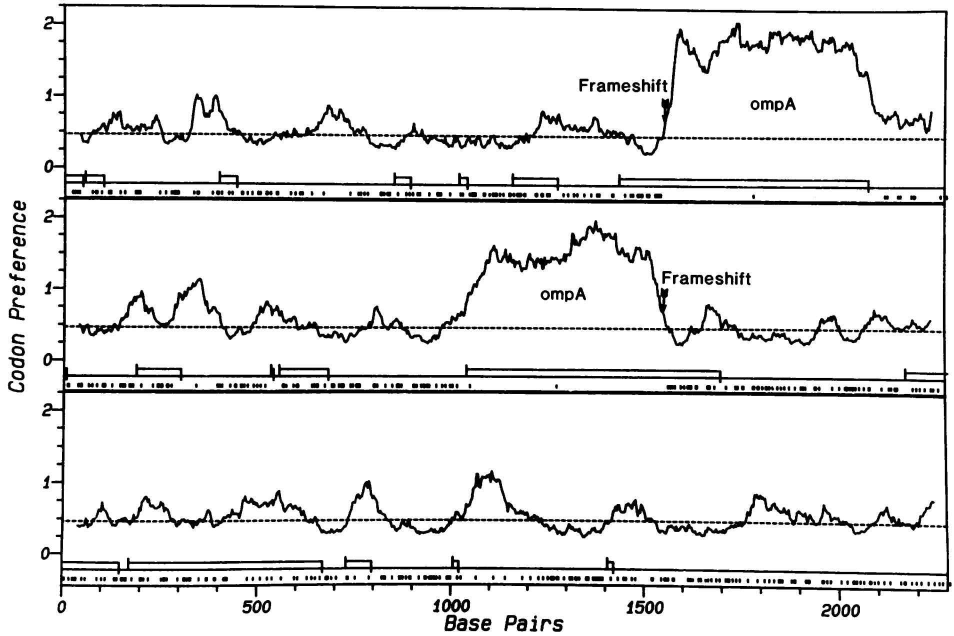

Fig. 4. Effect of frameshift errors on codon preference plots. An additional base was inserted at position 1551 creating an artificial frameshfit error in the E. coli ompA gene (13). Rare codons, 5% or less of synonomous family, are marked, w = 25.

product. In some cases, for example that of dnaG, knowledge of the approximate expression level of the genes in an operon may also give clues to the mode of regulation of the operon.

Another important use of the codon preference plot is the detection of sequencing errors. Because all three reading frames of sequence are plotted, spurious insertions or deletions are immediately obvious. Fig. 4 shows an artificial example of the effect of frameshift errors on the codon preference plot. The drop in the line in one frame, and the simultaneous rise in another is highly characteristic of this sort of error. Even though the open reading frame continues past the frameshift, it is obvious from the rare codon plot that the codon usage changes dramatically at that point. We use another version of the program that gives the actual numerical values of P and p for each position in each reading frame of the sequence to locate these errors more precisely. A frameshift error can often be located to within 10 bp in this manner.

DISCUSSION

Several techniques for locating protein coding sequences within DNA sequences have been described (11,22,23,24). Of these the methods of Staden and McLachlan (11), and Fickett (22) seem to be the most useful. Our method is similar to that of Staden and McLachlan: both rely on comparison of a DNA sequence to a codon frequency table and allow independent examination of all possible reading frames. However, it differs in important ways from the Staden-MacLachen method: our method uses codon frequencies calculated as the fractional use of each codon within its synonomous family. This has two important consequences: first, the codon preference statistic is independent of the amino acid composition of the protein, and second, only one codon table is expected to be necessary for each organism. A third advantage of our technique is the information it gives on protein expression level. As mentioned previously, this information can be useful in determining the identity of an unknown protein encoded by a sequenced open reading frame. The presence of three kinds of information on one plot, the codon preference statistic, open reading frames, and rare codons also enhances the usefulness of these plots.

Two techniques for locating protein coding regions based on periodic base composition asymmetries have been described (22,23). Shepherd's (22) method concentrates on differences between purine and pyrimidine use at the third position of the codon, while Fickett's method (23) relies on weighted autocorrelation coefficients for each base. Although the method of Fickett is powerful, it calculates only a single composite statistic for all six possible reading frames. It therefore is not as powerful as our method or the Staden-MacLachan method in assigning a gene to a particular open reading frame. In addition, it is of no use in detecting DNA sequencing errors or in predicting protein expression levels.

The codon preference plot described here locates highly and moderately expressed genes very well. Its performance with weakly expressed genes is, in one sense, less satisfying. Since the codon preference statistic is small for weakly expressed genes, they are more difficult to distinguish from non coding regions. We have found, nonetheless, that for most weakly expressed genes (e.g., dnaG, lacI) there is little difficulty in determining the coding region. We believe the prediction of expression level is worth the increased difficulty in distinguishing weakly expressed genes from non-genes.

HARDWARE AND SOFTWARE

Computer analysis was performed on a Digital Equipment Corporation VAX computer using the VMS operating system. The programs were written in FORTRAN-77 using a library of procedures provided by the University of Wisconsin Genetics Computer Group (UWGCG) (25). Plots were produced by a Hewlett-Packard 7221T plotter under software control. The second author should be contacted for information about the installation of this program or the entire UWGCG software package on other VAX computers.

REFERENCES

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Nuc. Acids Res. 8, r49-r62.
2. Grantham, R., Gautier, C. and Gouy, M. (1980) Nuc. Acids Res. 8, 1893-1912.
3. Grantham, R. Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nuc. Acids Res. 9, r43-r74.
4. Bennetzen, J. L. and Hall, B. D. (1982) J. Biol. Chem. 257, 3026-3031.
5. Gouy, M. and Gautier, C. (1982) Nuc. Acids Res. 10, 7055-7074.
6. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
7. Ikemura, T. (1982) J. Mol. Biol. 158, 573-597.
8. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979) Proc. Natl. Acad. Sci. USA 76, 1697-1701.
9. Post, L. E. and Nomura, M. (1980) J. Biol. Chem. 255, 4460-4466.
10. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
11. Staden, R. and McLachan, A. D. (1982) Nuc. Acids Res. 10, 141-156.
12. Neyman, J. and Pearson, E. S. (1933) Phil. Trans., A., 231, 289.
13. Rindone, W. P., Perry, H. M., Goad, W. B. and Bilofsky, H. S. (1983) DNA 2, 173.
14. Ovchinikov, Y. A., Monastryskaya, G. S., Gubanov, V. V., Gurvev, S. D., Chertov, O. Y., Modyanov, N. N., Grinkevich, V. A., Makarova, I. A.,

Marchenko, T. V., Polovnikova, I. N., Lipkin, V. M. and Sverdlov, E. D. (1981) Eur. J. Biochem. 116, 621-629.

15. Burton, Z., Burgess, R. R., Lin, J., Moore, D., Holder, S. and Gross, C. A. (1981) Nuc. Acids Res. 9, 2889-2903.
16. Burton, Z., Gross, C. A., Watanabe, K. K. and Burgess, R. R. (1983) Cell 32, 335-339.
17. Kjeldgaard, N. O. and Gausing, K. 91974) in Ribosomes, Nomura, M., Tissieres, A. and Lengyel, P. Eds., Cold Spring Harbor Monograph Series, pp. 369-392, Cold Spring Harbor, New York.
18. Rowen, L. and Kornberg, A. (1978) J. Biol. Chem. 253, 758-764.
19. Iwakura, Y., Ito, K. and Ishihama, A. (1974) Mol. Gen. Genet. 133, 1-23.
20. Engbaek, F., Gross, C. A. and Burgess, R. R. (1976) Mol. Gen. Genet. 143, 291-295.
21. Smiley, B. L., Lupski, J. R., Svec, P. S., McMacken, R. and Godson, G. N. (1982) Proc. Natl. Acad. Sci. USA 79, 4550-4554.
22. Fickett, J. W. (1982) Nuc. Acids Res. 10, 5303-5318.
23. Shepherd, J. C. W. (1981) Proc. Natl. Acad. Sci. USA 78, 1596-1600.
24. Rodier, F., Gabarro-Arpa, J., Ehrlich, R. and Reiss, C. (1980) Nuc. Acids Res. 10, 391-402.
25. Devereux, J., Haeberli, P. and Smithies, O. (1983) Submitted for this issue.