Volume 8 Number 4 1980 Nucleic Acids Research

A computer program to search for tRNA genes

R.Staden

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

Received 16 January 1980

ABSTRACT

This paper describes a computer program that can find tRNA genes within long DNA sequences. The program obviates the need to map the tRNA genes.

INTRODUCTION

Earlier papers have described computer programs for handling (1) and analysis (2,3) of nucleic acid sequence data. More recently papers (4,5) have reported a strategy of sequencing in which computers play a central role and also make it unnecessary to obtain restriction maps of the DNA being sequenced. A further class of problem in this field which can be tackled by computers is that of pattern recognition. This paper describes a computer program to locate occurrences of a pattern that is already well known: the cloverleaf structure of the tRNA (see Fig. 1). The program is very flexible and the user can decide by simple choice of parameters how much relative weight to give to various features of the tRNA structure.

Projects are currently under way that involve the sequencing of regions of DNA that contain the genes for tRNAs. Even when the tRNAs have been mapped it is very difficult to pick them out by eye from a long sequence. As with the sequencing strategy which makes restriction mapping unnecessary, the computer program described here obviates the need to map the tRNA genes.

The program reads through a sequence of any length looking for sections that could fold into a cloverleaf. The user of the program can define the range of loop sizes and minimum base pairing required within each arm of the cloverleaf and also which conserved bases must be present. He can also instruct the program to allow for an intron (6,7,8) of any length in the anticodon loop. If the program finds a section of sequence that conforms to all these criteria it displays it in both one and two dimensional form and then continues searching. If there is an intron present it is removed from

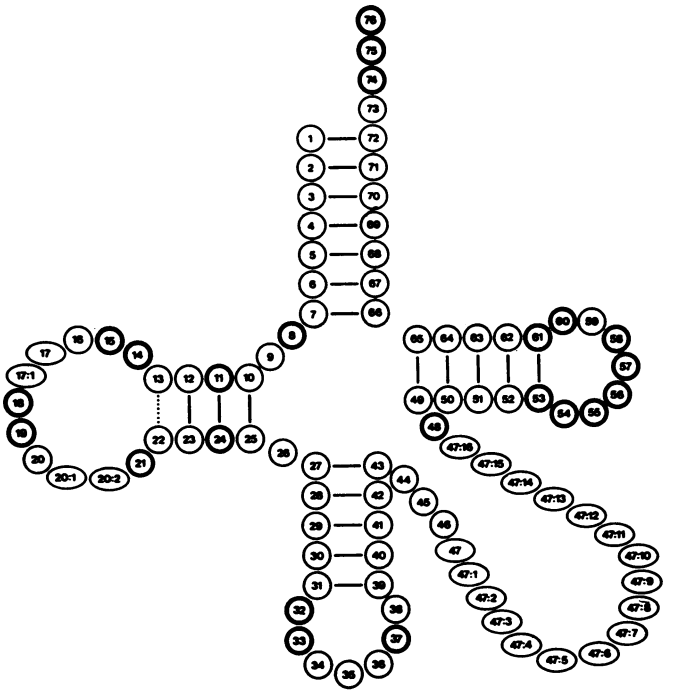

Figure 1: Numbering system of nucleotides in tRNAs according to the numbering of phenylalanine tRNA from yeast. Circles represent nucleotides which are always present; among these, the thick-edged circles denote invariant or semiinvariant nucleotides. Ovals represent nucleotides which are not present in each sequence: these are the nucleotides before the two constant GMP residues (18, 19) in the D loop, the nucleotides after these GMP residues, and the nucleotides in the variable loop which may be up to 17 nucleotides.

A nucleotide to be added at a given site is indicated by the number of the preceding nucleotide followed by a colon and a further number. Thus, e.g. 20:1 and 20:2 mean the first and second nucleotide after position 20. The absence of a nucleotide is indicated by the absence of a number, e.g. if no residue is found in position 17, the sequence then reads C16-G18. The numbering for the D loop, when one, two or three nucleotides are present each between 15 and 18 or between 19 and 21, is then 16 and 16, 17 and 16, 17, 17:1 or 20 and 20, 20:1 and 20, 20:1, 20:2, respectively. When the variable loop is five-membered the numbering is as in yeast phenylalanine tRNA 44, 45, 46, 47, 48. 47 is eliminated as the three dimensional structure of yeast phenylalanine tRNA suggests when the variable loop is four-membered. For large variable loops, numbers are added onto 47, e.g. for thirteen nucleotides 44, 45, 46, 47, 47:1, 47:2, 47:3, 47:4, 47:5, 47:6, 47:7, 47:8, 48.

between the sixth and seventh bases (9) of the anticodon in the 2D display but is left intact for the 1D display.

The program has been successfully applied to the DNA sequences from human mitochondria (10) despite the fact that the tRNAs that they code for differ from all those previously studied (11). The program runs on the same

small PDP computer referred to in previous papers (1,2,4).

Description of the program

The tRNAs which have been sequenced so far have two characteristics that can be used to locate their genes within a long DNA sequence. Firstly, they have a similar secondary structure - the cloverleaf - and, secondly, particular bases almost always appear at certain positions in the cloverleaf. The cloverleaf is composed of four base paired stems and four loops. Three of the stems are of fixed length but the fourth, the dhu stem which usually has four base pairs sometimes has only three. The dhu and extra loops have long been known to vary in length and recent results (10,12) have shown that the anticodon and TψC loop can also change. The following relationships between the stems in the cloverleaf are assumed in the program: (a) there are no bases between one end of the aminoacyl stem and the adjoining TψC stem; (b) there are two bases between the aminoacyl stem and the dhu stem; (c) there is one base between the dhu stem and the anticodon stem; (d) there are at least three bases between the anticodon stem and the TψC stem.

The program looks first for cloverleaf structure and then, if required, for conserved bases. The sizes of the loops, the number of base pairs in the stems and the required conserved bases may all be specified by the user of the program. In order to define the base pairing we use a simple scoring system: we score 2 for a normal Watson-Crick base pair (A-T, G-C) and 1 for a G-T base pair. With this scoring system the user of the program can define the minimum base pairing required in each stem by specifying a score for each. Only those cloverleafs which attain at least these minimum base pairing scores in each of their stems will pass on either to be displayed or to the optional conserved base tests.

If the user of the program chooses to filter the data on conserved bases he is asked to supply numbers for a scoring system. He is prompted to give an individual score for each of 18 of the conserved bases and then an overall minimum score. When the program is examining a cloverleaf passed on from the base pairing tests it gives a score which is the sum of the scores for each of the conserved bases that are correct. If this score is at least equal to the minimum overall score the cloverleaf will be displayed.

Conserved bases used in filtering

Base number (as in Fig. 1)	Base assignment
8	T
10	G
11	C or T
14	A
15	A or G
21	A
32	C or T
33	T
37	A
48	T or C
53	G
54	T
55	T
56	C
57	A or G
58	A
60	C or T
61	C

The scoring system allows the user to make some bases obligatory and to give others less importance. For example, he may wish to make bases 8, 14, 32, 53 obligatory, want three out of bases 21, 33, 48, 54 to be present and two from bases 55, 56, 57, 60. To achieve this he could use the following scoring system: the first group are all assigned individual scores of 1000, the second group scores of 100, and the third group scores of 1. Then he sets the overall minimum score to 4000 + 3 + 2 = 4302.

Figure 2 shows a sample run of the program on the sequence of ϕX174 (14) and demonstrates the parameters that are defined by the user of the program. (All typing by the user is underlined all the rest is typed by the program.) He is scanning the whole of a sequence looking for tRNAs of maximum length 80 plus a range of intron lengths of 12 to 18 bases. He has defined the minimum base pairing for each stem and elected to use the default values of 6 and 9 for the length of the Tψ loop. He has then chosen to filter the data by requiring certain of the conserved bases to be present. The numbers he has chosen have made base 8 obligatory and require two of bases 32, 33, 37 to be present. No others need be correct. The beginning of the output list of possible tRNAs conforming to these criteria is shown in Fig. 3.

Description of the output (see Fig. 3)

Any sections of sequence that contain the four stems with sufficient base pairing and required conserved bases are displayed in both one and two dimensional form.

```
RU TRNA

        PLEASE TYPE NAME OF FILE 1
0XCS70
 FIRST SEQ NO = _
 LAST SEQ NO = _
 MAX TRNA LENGTH =80
 MIN SCORE AMINOACYL STEM =10
 MIN SCORE TU STEM =7
 MIN SCORE ANTICODON STEM =7
     MIN SCORE D STEM =4
     MIN INTRON LENGTH =12
     MAX INTRON LENGTH =18
     MIN LENGTH TU LOOP = _
     MAX LENGTH TU LOOP = _
 TO FILTER ON CONSERVED BASES TYPE Y
Y
     BASE   8 SCORE = 100
     BASE  10 SCORE = _
     BASE  11 SCORE = _
     BASE  14 SCORE = _
     BASE  15 SCORE = _
     BASE  21 SCORE = _
     BASE  32 SCORE = 1
     BASE  33 SCORE = 1
     BASE  37 SCORE = 1
     BASE  48 SCORE = _
     BASE  53 SCORE = _
     BASE  54 SCORE = _
     BASE  55 SCORE = _
     BASE  56 SCORE = _
     BASE  57 SCORE = _
     BASE  58 SCORE = _
     BASE  60 SCORE = _
     BASE  61 SCORE = _
     MIN TOTAL SCORE ON CONSERVED BASES = 102
```

Figure 2

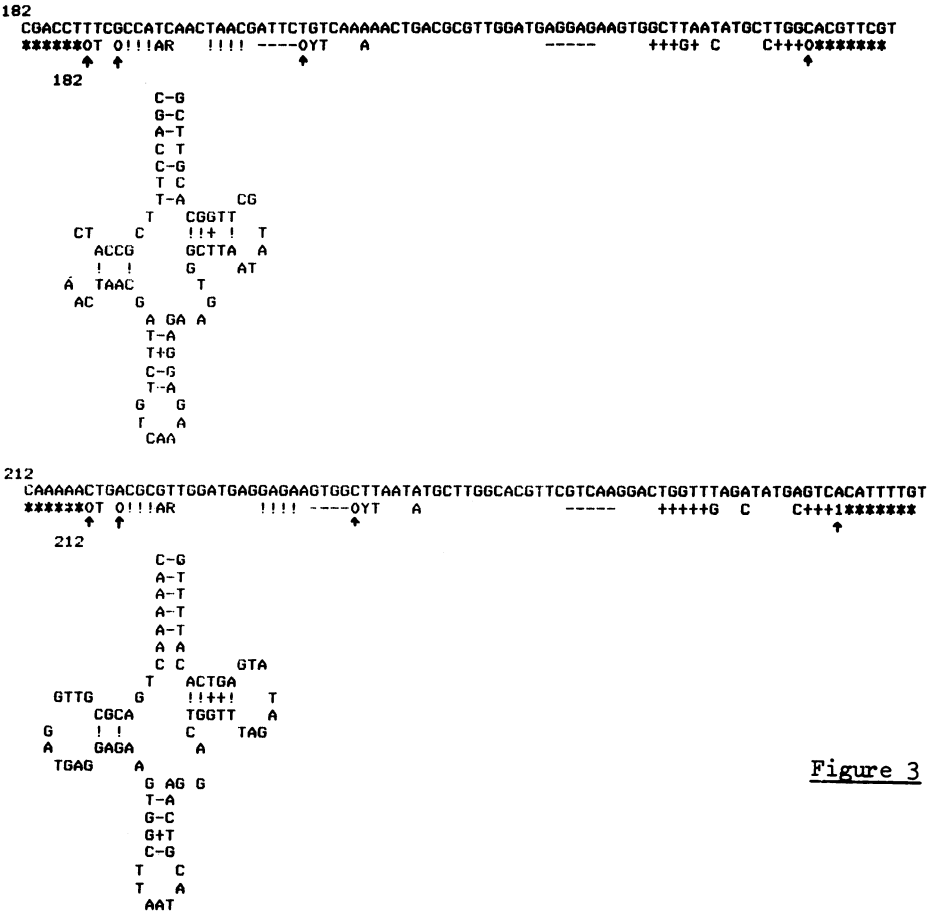

Figure 3

The one dimensional display consists of:

(1) The sequence position of the first nucleotide.

(2) The sequence.

(3) The aminoacyl stem is marked by asterisks (*).

(4) The anticodon stem is marked by dashes (-).

(5) The Tψ stem is marked by plus signs (+).

(6) The D-loop is marked by exclamation marks (!).

(7) Some of the conserved bases are shown.

(8) The score for each stem is shown by a number replacing one of the stem characters (arrowed in Fig. 3 but not by the program). This number is the score for this

stem minus the user defined minimum score for this stem.
(This is done merely to reduce the score to a single digit.)

The two dimensional display is used for all those cloverleafs that conform to the following:

(1) The length of the Tψ loop must be less than 19 bases and greater than 2.

(2) The dhu loop must be less than 15 bases and if it is less than 2 bases in length the stem will be reduced to 3 base pairs.

(3) The variable loop must be less than 27 bases.

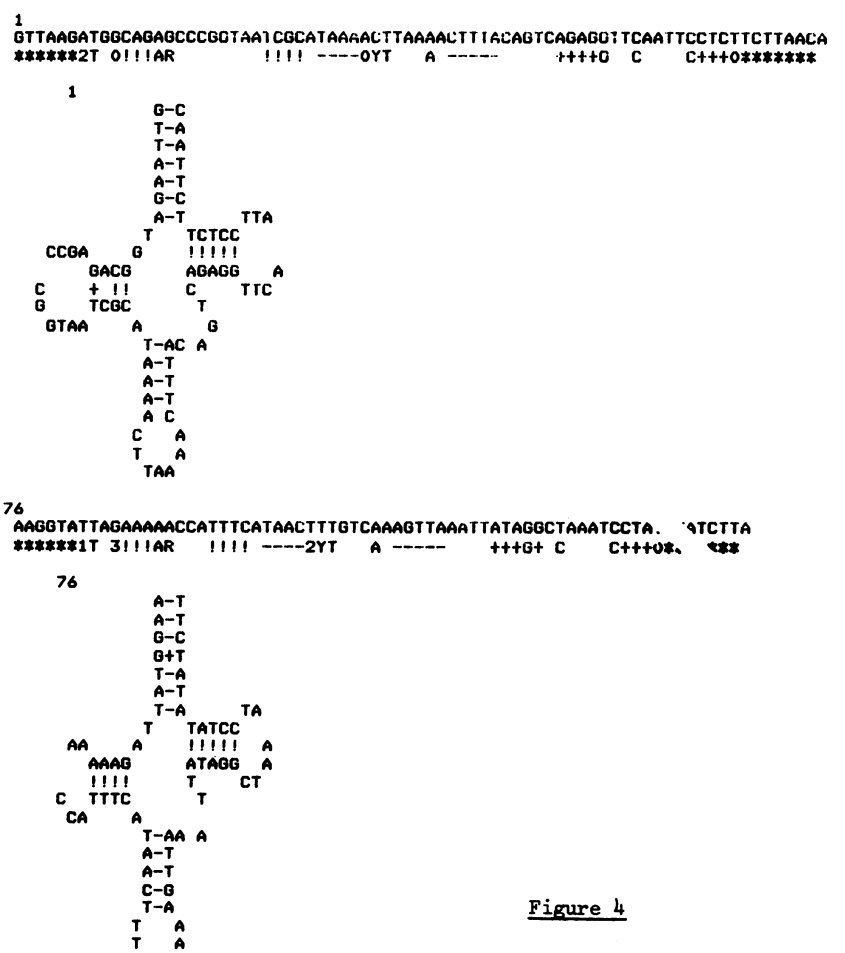

Figure 4

Cloverleafs that do not conform to these criteria will only be displayed in the 1D form. If a cloverleaf contains an intron it will not be included in the 2D display: the program cuts out the intron after the sixth base of the anticodon loop (9). Watson-Crick base pairs are marked with dashes and exclamation marks and G-T base pairs with plus signs.

Note that the 2 cloverleafs displayed in Fig. 3 are not tRNAs - the base pairing scores have been set sufficiently low to allow the program to find these structures in ϕX174 (14) purely for demonstration purposes. In Figure 4 are shown two of the tRNA genes from human mitochondrial DNA which were found using this program.

The first (13) is the nearest we have found to the tRNAs contained in reference 11 which are mainly from E. coli, yeast and some higher eukaryotic cells. It has similar loop sizes and many of the conserved bases. The second (10) is more typical of the mitochondrially coded tRNAs which we have found in that it has much smaller loops and very few of the conserved bases.

SUMMARY

A computer program for locating tRNA genes from within long sequences of DNA is available on request.

Acknowledgement

I would like to thank Dr M. Sprinzl for allowing me to use Figure 1 which is from reference 11.

REFERENCES

1. Staden, R. (1977) Nucleic Acids Research, 4, 4037-4051.
2. Staden, R. (1978) Nucleic Acids Research, 5, 1013-1015.
3. Korn, L.J., Queen, C.L. and Wegman, M.W. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405.
4. Staden, R. (1979) Nucleic Acids Research, 6, 2601-2610.
5. Gingeras, T.R., Milazzo, J.P., Sciaky, D. and Roberts, R.J. (1979) Nucleic Acids Research, 7, 529-545.
6. Goodman, H.M., Olson, M.V. and Hall, B.D. (1977) Proc. Natl. Acad. Sci. USA 74, 5453-5457.
7. Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R. and Rutter, W.J. (1978) Proc. Natl. Acad. Sci. USA 75, 190-194.
8. Etcheverry, T., Colby, D. and Guthrie, C. (1979) Cell, 18, 11-26.
9. Knapp, G., Ogden, R.C., Peebles, C.L. and Abelson, J. (1979) Cell, 18, 37-45.
10. Barrell, B.G., Bankier, A.T. and Drouin, J. (1979) Nature, 282, 189-194.
11. Gauss, D.H., Grüter, F. and Sprinzl, M. (1979) Nucleic Acids Research, 6, r1-r19.

12. Li, M., and Tzagoloff, A. (1979) Cell, 18, 47-53.
13. Eperon, I.C., Anderson, S. and Neirlich, D.P. Manuscript in preparation.
14. 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, 276, 236-247.