Volume 14 Number 1 1986 Nucleic Acids Research

A program for the identification of tRNA-like structures in DNA sequence data

Christopher C.Marvel

University of California, Berkeley, School of Public Health, 140 Warren Hall, Naval Biosciences Laboratory, Berkeley, CA 94720, USA

Received 6 June 1985

ABSTRACT

A computer algorithm has been developed which identifies tRNA genes and tRNA-like structures in DNA sequences[1]. The program searches the sequence string for specific base positions that correspond to the invariant and semi-invariant bases found in tRNAs. The tRNA nature of the sequence is confirmed by the presence of complementary base pairing at the tRNA's calculated 5' and 3' ends (which _in situ_ constitutes the amino-acyl stem region). The program achieves greater than 96% accuracy when run against known tRNA sequences in the Genbank database. The program is modular and is readily modified to allow searching either a file or database. The program is written in "C" and operates on a D.E.C. Vax 750. The utility of the algorithm is demonstrated by the identification of a distinctive tRNA structure in an intron of a published bovine hemoglobin gene.

INTRODUCTION

The increasing sophistication of DNA sequencing methodologies has led to an explosive growth of DNA sequence data. Programs that identify sequence homology are in common use and are invaluable for analysis of this data. However analysis is limited when confined only to the examination of linear sequence. tRNAs are important biological molecules which are typified by patterns of secondary and tertiary structure in addition to a defined sequence. Other RNA molecules exhibit characteristic secondary structures similar to those of tRNA. Among these are the ends of some plant viruses which are amino-acylated _in vivo_ (1) and the ends of bacterial operons which have been implicated in interactions with tRNA binding proteins for gene regulation and control (2). The identification of these structures in a DNA sequence requires an examination of their secondary structure patterns and the presence of invariant bases at specific locations. Patterns corresponding to

[1] Source program will be provided upon receipt of a self-addressed mailing label and a blank tape. A small charge will be requested to cover mailing and processing.

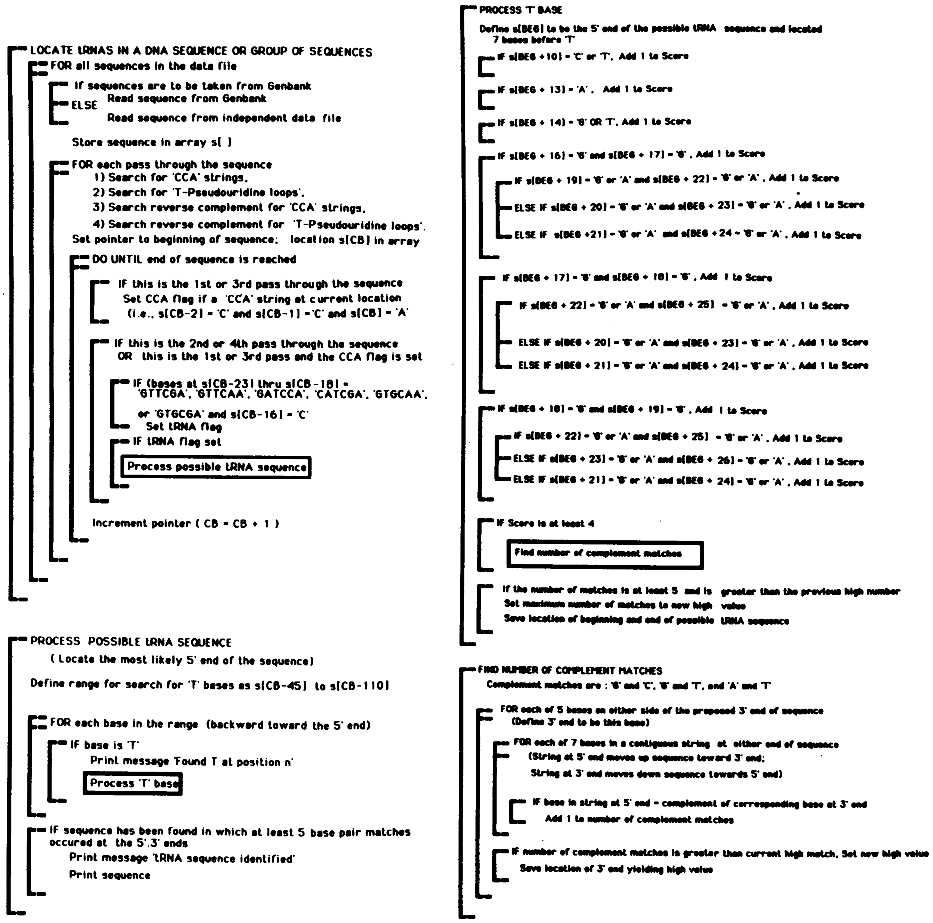
secondary structures can be initially screened by the examination of dyad symmetry or inverted repeats. This may lead to the definition of a structure of interest, but the final analysis is still commonly done manually.

These searches have become less feasible as the data base has become more extensive. In order to identify patterns in DNA sequence data that correspond to tRNA genes, tRNA pseudogenes, or that code for tRNA-like structures that may be involved in gene regulation, a tRNA search algorithm has been developed. The program is easily modifiable to allow a variable range of constraints on the search parameters. Modifiable parameters include the nature, position, and number of invariant bases and the positions and extent of complementary base pairing. The program is designed for easy and rapid Genbank searches for tRNA patterns that may be of either structural or regulatory nature.

ALGORITHM DESCRIPTION

The identification of tRNA sequences is based on two characteristics of tRNAs: 1) invariant and semi-invariant bases are found at defined locations in the tRNA, where semi-invariant positions are generally occupied by either a purine or pyrimidine and 2) complementary base pairing which creates a characteristic cloverleaf secondary structure (reviewed in ref.3).

An action diagram of the algorithm is shown in Figure 1. The algorithm first searches input sequence for 'CCA' strings which may be indicative of the 3'-end of a tRNA gene. The 'A' base in the string is assigned the position number 1 and a search is made in an interval relative to this base for invariant base consensus patterns that are indicative of the T-Pseudouridine loop region of a tRNA. If a consensus pattern is found, the program begins searching for the 5' end of the tRNA. This search extends a sufficient distance to take into account the length of the variable stem region and in the case of eukaryotic DNA, the presence of an intervening sequence. The 5' end is located by examining all 'T' bases which may correspond to the invariant 'T' found 8 residues from the 5' end. The correct 'T' base is identified by means of a numerical score calculated on the presence of invariant bases in the examined 'T' base region. If this numerical score exceeds a certain minimum and is the highest of all 'T' bases in this region then this base defines the 5' end of the "presumed tRNA". The 5' and 3' ends of this "presumed tRNA" are now examined for their ability to form a stem structure by complementary base pairing. A score of 5 or more complementary base pairings in this structure identifies this sequence as one which may code for a tRNA. The tRNA sequence is printed out with its location in the search



```
LOCATE tRNAS IN A DNA SEQUENCE OR GROUP OF SEQUENCES
  FOR all sequences in the data file
    If sequences are to be taken from Genbank
      Read sequence from Genbank
    ELSE
      Read sequence from independent data file
    Store sequence in array s[ ]
    FOR each pass through the sequence
        1) Search for 'CCA' strings.
        2) Search for 'T-Pseudouridine loops'.
        3) Search reverse complement for 'CCA' strings.
        4) Search reverse complement for 'T-Pseudouridine loops'.
      Set pointer to beginning of sequence; location s[CB] in array
      DO UNTIL end of sequence is reached
        IF this is the 1st or 3rd pass through the sequence
          Set CCA flag if a 'CCA' string at current location
          (i.e., s[CB-2] = 'C' and s[CB-1] = 'C' and s[CB] = 'A'
        IF this is the 2nd or 4th pass through the sequence
        OR this is the 1st or 3rd pass and the CCA flag is set
          IF (bases at s[CB-23] thru s[CB-18] =
             'GTTCGA', 'GTTCAA', 'GATCCA', 'CATCGA', 'GTGCAA',
             or 'GTGCGA' and s[CB-16] = 'C'
            Set tRNA flag
          IF tRNA flag set
            [Process possible tRNA sequence]
        Increment pointer ( CB = CB + 1 )

PROCESS POSSIBLE tRNA SEQUENCE
    ( Locate the most likely 5' end of the sequence)
  Define range for search for 'T' bases as s[CB-45] to s[CB-110]
  FOR each base in the range (backward toward the 5' end)
    IF base is 'T'
      Print message 'Found T at position n'
      [Process 'T' base]
  IF sequence has been found in which at least 5 base pair matches
  occured at the 5'.3' ends
    Print message 'tRNA sequence identified'
    Print sequence

PROCESS 'T' BASE
  Define s[BEG] to be the 5' end of the possible tRNA sequence and located
    7 bases before 'T'
  IF s[BEG +10] = 'C' or 'T', Add 1 to Score
  IF s[BEG + 13] = 'A' , Add 1 to Score
  IF s[BEG + 14] = 'G' OR 'T', Add 1 to Score
  IF s[BEG + 16] = 'G' and s[BEG + 17] = 'G' , Add 1 to Score
    IF s[BEG + 19] = 'G' or 'A' and s[BEG + 22] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG + 20] = 'G' or 'A' and s[BEG + 23] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG +21] = 'G' or 'A' and s[BEG +24 = 'G' or 'A' , Add 1 to Score
  IF s[BEG + 17] = 'G' and s[BEG + 18] = 'G' , Add 1 to Score
    IF s[BEG + 22] = 'G' or 'A' and s[BEG + 25] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG + 20] = 'G' or 'A' and s[BEG + 23] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG + 21] = 'G' or 'A' and s[BEG + 24] = 'G' or 'A' , Add 1 to Score
  IF s[BEG + 18] = 'G' and s[BEG + 19] = 'G' , Add 1 to Score
    IF s[BEG + 22] = 'G' or 'A' and s[BEG + 25] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG + 23] = 'G' or 'A' and s[BEG + 26] = 'G' or 'A' , Add 1 to Score
    ELSE IF s[BEG + 21] = 'G' or 'A' and s[BEG + 24] = 'G' or 'A' , Add 1 to Score
  IF Score is at least 4
    [Find number of complement matches]
  If the number of matches is at least 5 and is greater than the previous high number
    Set maximum number of matches to new high value
    Save location of beginning and end of possible tRNA sequence

FIND NUMBER OF COMPLEMENT MATCHES
  Complement matches are : 'G' and 'C', 'G' and 'T', and 'A' and 'T'
  FOR each of 5 bases on either side of the proposed 3' end of sequence
      (Define 3' end to be this base)
    FOR each of 7 bases in a contiguous string at either end of sequence
        (String at 5' end moves up sequence toward 3' end;
         String at 3' end moves down sequence towards 5' end)
      IF base in string at 5' end = complement of corresponding base at 3' end
        Add 1 to number of complement matches
    IF number of complement matches is greater than current high match, Set new high value
      Save location of 3' end yielding high value
```

Figure 1. Action diagram of the tRNA search algorithm.

file.

In a small percentage of tRNAs the 'CCA' is added post-transcriptionally and will not be coded for by the gene. In order to identify these as well as less constrained tRNA structures, the program independently searches for structures without the 'CCA' end.

The reverse complement of the DNA strand is automatically determined and searched when the program is run. The Genbank search algorithm is very rapid and the output can be monitored interactively.

Not all invariant base positions are included in the algorithm and no

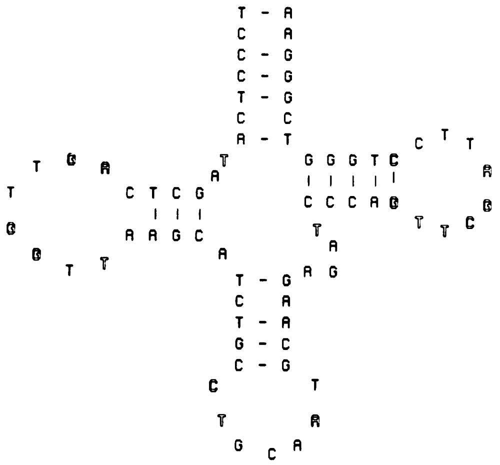


Figure 2. tRNA-like structure detected in bovine globin intron 2 using the tRNA search algorithm. Bases in outline correspond to consensus bases found in tRNA molecules.

complementary base pairing is examined beyond the amino-acyl region. These features are not included as they are not necessary for accuracy and will in most instances reduce the number of structures found. These constraints can be used while concurrently relaxing others to examine structures whose features deviate from tRNA genes.

RESULTS AND DISCUSSION

To test its validity the algorithm was run on the contents of the Genbank "structural RNA file". This file contains a listing of tRNAs as well as other structural RNAs including ribosomal and small nuclear RNAs. Of a total of 403 entries 264 corresponded to tRNAs. 253 of these were identified correctly for an accuracy of 96%. One false positive occurred. Of the 11 missed tRNAs, 10 were of mitochrondrial origin and the last was from bacteriophage T5. In all cases the failure was due to a lack of consensus invariant bases in the T-pseudouridine loop.

A 2 kb sequence containing a tRNA-like structure (see below) was base scrambled 50 separate times. This randomization retained the base composition of the sequence and served to generate test files. The program was run on these files to determine the number of structures that are random occurrences. In this test a positive structure was flagged for every 16 kb of sequence searched. These structures did not contain stem-loop structures and were easily recognizable as being artifactual. For general screening the positive effects of tightening the search parameters (reducing positives that must be manually examined) will be offset if viable structures are missed.

An example of a tRNA-like structure found by the program is shown in Figure 2. This tRNA-like structure was found in a search of the Genbank file " other mammalian". This file contains 344 entries with no reported tRNA genes. Several possible tRNA-like structures were noted with the most striking being found in a bovine globin sequence (4) and illustrated in a tRNA cloverleaf diagram in Figure 2. This structure's variant and semi-invariant base positions are totally consistent with functional tRNA molecules and stem-base pairing may allow a tRNA-like folding in vivo. This structure is found within intron 2 and may have regulatory roles in gene expression through interactions with tRNA binding proteins(2). Descriptions of this and other tRNA-like structures identified with the algorithm will be described in future work.

ACKNOWLEDGEMENTS

The expert 'C' programming skills of Pamela J. Marvel were central to this work and are gratefully acknowledged. Thanks to Hugo Martinez who made code available to facilitate interfacing the program to Genbank and to Harold Kammen for scientific counsel. Dean Tolan's assistance and support made the realization of the program possible. Funding support was from the Office of Naval Research, Contract # N00014-81-C-0570.

References

1. Haenni, A-L. and Chapeville, F.(1980) in Transfer RNA: Biological Aspects, Soll, D., Abelson, J.N., and Schimmel, P.R. Eds., pp 539-556, Cold Spring Harbor Monograph Series, New York.
2. Ames, B.N., Tsang, T.T., Buck, M. and Christman, M.F.(1983) Proc. Natl. Acad. Sci. U.S.A. 80, 5240-5242.
3. Singhal, R.P. and Fallis, P.A.M.(1979) in Progress in Nucleic Acid Research and Molecular Biology, Vol. 23, Cohn, W.E. Ed. pp 228-262. Academic Press, New York.
4. Schimenti, J.C. and Duncan, C.H.(1984) Nucleic Acids Res., 12, 1641-1655.