A COMPUTER METHOD FOR PREDICTING THE SEQUENCE OF tRNA FROM ITS ENZYMATIC DIGESTION PRODUCTS BASED ON ITS SECONDARY STRUCTURE

P. JAGADEESWARAN AND JOSEPH D. CHERAYIL
Department of Biochemistry
AND
N. PATTABIRAMAN AND V. SASISEKHARAN
Molecular Biophysics Unit
Indian Institute of Science, Bangalore 560 012, India

THE procedure followed for the determination of the sequence of RNA involves digestion of the purified sample with endonucleases of known specificity followed by separation of the fragments and their analysis for individual sequence. The RNA is then partially digested with less enzyme under controlled conditions, to get larger fragments which are again separated and analysed to get sufficient overlaps to deduce the unique sequence. The procedure is laborious as several partial digestions have to be done to arrive at the sequence[1,2]. It is well known that for a given primary sequence of RNA various secondary structures can be predicted[3]. So far no successful method for constructing the primary sequence from the oligonucleotides has been devised, although attempts have been made to deduce the sequence from the oligomers[4]. The possible sequences even for a small chain of 50 nucleotides will run into very large numbers. However, the number of possibilities will be drastically reduced if secondary structure restrictions are imposed. The best choice of applying the computer method would, therefore, be to transfer RNA because of its well established cloverleaf structure[5,6]. We have devised a computer method to simulate all isotomers from the oligonucleotides which are formed by digestion of the RNA with RNase T_1 and pancreatic RNase. The application of this method to the determination of the primary sequence of tRNA is illustrated with *E. coli* $tRNA_2^{gln}$ as an example.

Principle of the Computer Method

The principle used is basically a "tree search" method, the flow-chart of which is shown in Fig. 1. The method may be compared to the procedure in finding all the possible routes to the top of a highly branched tree. To reach the top one starts from the stem and goes up the various branches. If one branch starting from a particular node does not lead to the top another branch is tried and if none of the branches from the node leads to the top a branch in a node below is taken; the process is repeated until the top is reached. Essentially the same procedure is followed in the computer for building the nucleotide chain from the fragments. There is a starting fragment similar to the stem from which the chain branches out. After the addition of each fragment the nucleotide chain branches out further into sub-branches. In some cases it will not be possible to build the chain along certain branches, as the chain would stop before all the fragments are used up; certain order of addition as indicated below has to be followed at each stage. The computer programme is to search for the path to be followed, check whether all the fragments are used and print the sequence when the chain is complete. All the possible sequences will be printed out by the computer. The flow-chart (Fig. 1) gives a brief account of the computer operation. Details of the programme will be published elsewhere.

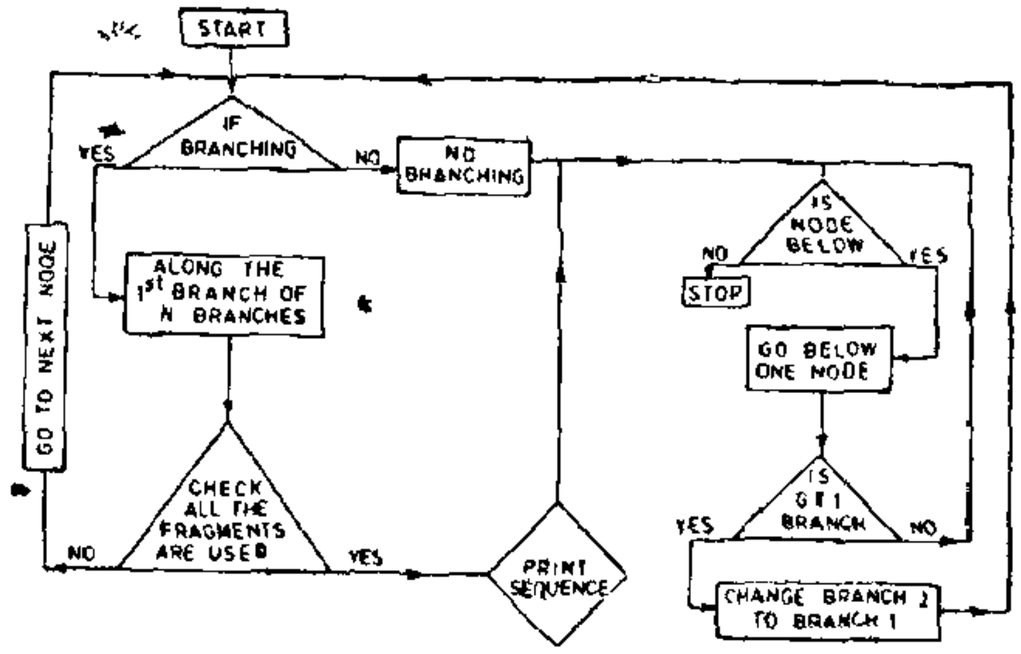

FIG. 1. The flow-chart of the computer programme. Details are given in the text.

If RNA is treated with RNase T_1 it will be degraded into fragments which have G_p at the 3′ end. Free G_p will be liberated from stretches of guanylic acid residues in the chain. Similarly upon treatment with RNase A fragments ending in U_p and C_p as well as free U_p and C_p will be produced. G_p, U_p and C_p are omitted from the fragments for building the nucleotide chain as these are already present in other fragments. RNase A products will contain all the G_p stretches while those of RNase T_1 will contain all the U_p and C_p stretches. In some cases it is possible to get longer sequences by the combination of a RNase T_1 fragment and a RNase A fragment from the overlaps[5]. For example, upon digestion of $tRNA_2^{gln}$ (Ref. 7) with RNase T_1, a fragment D.AAG

and with RNase A a fragment G"GD will be formed (D, dihydrouridylic acid residue and G" methyl guanylic acid residue, see the structure of the tRNA, ref. : 8). The two oligonucleotides can be combined to G GDAAG, the nucleotide D being the overlap. Other overlapping sequences can be combined in a similar manner. Thus the fragments shown in Table I are the various sequences that can be obtained from $tRNA_2$ [In] experimentally.

TABLE I

Forward fragments

Fragments of $tRNA_2^{Gln}$ read 5′ to 3′.

1.1 pUG (2)	3.1 CCAAG (4)
2.1 GC (3)	3.2 CACCG (2)
2.2 GU (5)	3.3 CAUUCCG (2)
2.3 GGC (3)	3.4 CGGDAAG (4)
2.4 GGGGU (5)	3.5 CCA_{OH*}
2.5 GAUUCCG (2)	4.1 AAGC (3)
2.6 GGAUUCUG (2)	4.2 AAGGC (3)
2.7 GAGGUψCG (2)	5.1 UAUCG (2)
2.8 GAAUCCUCG (2)	5.2 UACCCCAGC (3)

As many overlaps as possible were obtained from the enzymatic fragments from the position of the minor nucleotides[3]. Minor nucleotides, except ψ and D, are represented by the parent nucleotides. The oligomers are arranged In groups. The first number indicates the group to which it belongs, the second its position in the group and the number on the right, in brackets, shows the group of the fragment which follows. The nucleotide(s) underlined shows the overlap with the next fragment. Molar ratio of the fragments is one in all cases.

These fragments are fed to the computer. Construction of the sequence tree is done in both directions starting from either end in a systematic way. It is obvious that in the present case pUG is the starting fragment from the 5′ end. The fragment to be added next to pUG depends on the nucleotide produced from its 3′ end upon treatment of this RNase T_1 fragment with the complementary endonuclease, RNase A. Since G will be the nucleotide that will be produced, a fragment starting with G has to be added to pUG, the nucleotide G being the overlap. If GGC from group 2 (see Table I) is added the sequence becomes pUGGC. The oligonucleotide GGC is a RNase A product and on treatment of this with RNase T_1 will produce C. Hence a fragment starting with C has to be added to pUGGC. Any of the oligonucleotides in group 3 may be added. If CCAAG is added the overall sequence becomes pUGGCCAAG. It is easily seen that the fragment to be added next is any one of the two fragments in group 4 (see Table I). Thus the sequence tree is built in the forward direction. The nucleotide chain is built in the backward direction from 3′ end on the same principle. In the latter case mirror images of the sequences of the fragments are used to build the sequence. In other words, sequences are read from 3′ to 5′. For example CCA is taken as ACC, GGC as CGG and CCAAG as GAACC. The overlaps between two adjacent fragments is decided in a manner similar to that Indicated above. Table II shows the reverse fragments with the overlaps..

TABLE II

Reverse fragments

Fragments of $tRNA_2^{Gln}$ read 3′ to 5′.

1.1 ACC (2)	3.4 GCUAU (4)
2.1 CG (3)	3.5 GCCAC (2)
2.2 CGG (3)	3.6 GCCUUAC (2)
2.3 CGACCCCAU (4)	3.7 GCψUGGAG (3)
2.4 CGAA (5)	4.1 UGGGG (3)
2.5 CGGAA (5)	4.2 UG (3)
3.1 GCUCCUAAG (3)	5.1 GAACC (2)
3.2 GUCUUAGG (3)	5.2 GAADGGC (2)
3.3 GCCUUAG (3)	

Fragments given in Table I are read in the reverse order, grouped and presented. The last fragment GUp is not shown.

The method was programmed for IBM 360/44 in Fortran IV language. The cloverleaf model of tRNA given in the *Handbook of Nucleic Acid Sequences* by Barrell and Clark[8] was used as the secondary structure foe deriving the primary sequence.

RESULTS AND DISCUSSION

The construction of the cloverleaf was carried out part by part with the restrictions imposed. The amino acid stem was constructed first by building the sequence from the 5′ end as well as from the 3′ end and selecting those structures which formed 7 consecutive hydrogen bonds (see Ref. 8). Fragments containing dihydrouridine (D) and pseudouridine (φ) were omitted for building the stem. tRNAs generally do not contain these nucleotides in the amino acid stem. Hence fragments containing these nucleotides were omitted. It was programmed to simulate all sequences up to a chain length of 8 nucleotides from

5′ end or the nearest number when the last fragment is added. The number of structures thus obtained was 39. In all tRNAs the 8th nucleotide from the 5′ end is U or a modified U (see Ref. 8). Out of the 39 structures only 5 had U at the 8th position. These are shown below :

1. pUGAAUCCUCG
2. pUGGAUUCUG
3. pUGCACCGUAUCG
4. pUGCACCGUACCCCAG
5. pUGGGGUAUCG

The nucleotide chain was also built from 3′ end to generate all sequences up to the 11th nucleotide or the nearest number with the reverse fragments (Table II). Eleventh nucleotide from 3′ end is the last one involved in base pairing on the amino acid stem. A total of 82 structures were obtained. Each of the 5 forward structures shown above was tested to find out whether it would pair with any of the 82 backward structures. Surprisingly only the 5th structure, pUGGGGUAUCG, gave duplex structures containing 7 or more hydrogen bonded pairs with two of the 82 reverse sequences. These are shown below :

```
A. ACCGACCCCAUGGGGU
          | | | | | | | | | |
        pUGGGGUAUCG
B. ACCGACCCCAUG
          | | | | | | | |
        pUGGGGUAU
```

It may be noted that in structure A the oligonucleotide GGGGU is used both in the forward and the reverse sequences and hence it was ruled out. Therefore the only structure possible was the second one, B. It is to be mentioned in this connection that *E. coli* formyl methionine tRNA[9] contains only 6 hydrogen bonded pairs in the amino acid stem. Yeast alanine tRNA[5] *E. coli* leucine $tRNA_1$[10] and wheat germ phenylalanine tRNA[71] contain one mismatch each in the stem. These are exceptions and were not considered in the present study in selecting the stem from all the structures. However this kind of mismatches also may be taken into account in selecting the stem. GU pair formation was allowed in the present case.

Dihydrouridine and $T\psi CG$ arms

Our next attempt was to build on to the stem (structure B above) the dihydrouridine arm in the forward direction and $T\psi CG$ arm in the reverse direction. There are a number of regularities regarding the positions of various nucleotides on tht dihydrouridine arm. It has a hydrogen bonded structure with 3 or 4 pairs and always it is the 10th nucleotide which forms the first hydrogen bonded pair. In the loop region a sequence AG or AA occurs followed by D in a somewhat fixed position (see Ref. 8). Besides the loop contains a GG sequence between 17 and 21 positions. The number of nucleotides in the loop is variable between 7 and 12, the maximum number required to build the chain up to the end of the arm being approximately 27. Therefore, it was programmed to simulate structures up to 27 nucleotides from 5′ end or the nearest number with the last fragment on the chain. Out of the 426 sequences printed out, only two structures satisfied all the restrictions on the dihydrouridine arm. The structures are :

(*a*) pUGGGGUAUCGCCAAGCGGDAAGGCACCG

(*b*) pUGGGGUAUCGCCAAGCGGDAAGGCAUUCCG

The nucleotides underlined form base pairs. It may be noted that the difference between the two structures is only in the last fragment ; CACCG in (*a*) is replaced by CAUUCCG in (*b*) (see Table I). Detection of these structures among the 426 structures was not at all difficult as large number of structures which did not satisfy any one of the restrictions on the dihydrouridine arm could be eliminated by just looking at the sequences.

Restrictions on the $T\psi CG$ arm are even more stringent. The number of paired and non-paired nucleotides in this arm is fixed and the total number up to the last hydrogen bonded pair is 28 from 3′ end (see Ref. 8). The nucleotide, ψ, always appears on 22nd position in the loop region. Besides the 16th nucleotide is C which base pairs with G on the 24th position. The computer program in the construction of the $T\psi CG$ arm was to simulate structures up to the 28th nucleotide from 3′ end with the reverse sequences, omitting the fragment containing dihydrouridine and those already used in forming the stem. A total of 278 structures were generated. Only 4 structures satisfied the restrictions on the $T_\psi CG$ arm. They are :

(i) ACCGACCCCAUGCUCCUAAGCψTGGAGUCUUAGG

(ii) ACCGACCCCAUGCUCCUAAGCψTGGAGCCUUAG

(iii) ACCGACCCCAUGCUCCUAAGCψTGGAGCCAC

(iv) ACCGACCCCAUGCUCCUAAGCψTGGAGCCUUAC.

The nucleotides which are underlined are involved in hydrogen bond formation. These structures differ from each other only in the last fragment.

Completion of cloverleaf

Since two structures (*a* and *b*) in the forward direction and four structures (i to iv) in reverse direction were obtained, a minimum of 8 possibilities for the entire sequence appeared possible. Each of the combination was considered to find whether it would form the anticodon arm when combined with the unused fragments in each case. Combinations (*a*) (iii) and (*b*) (iv) were not possible as the same fragment had been used in the forward as well as in the reverse sequences. It may be noted that (*a*) contains CACCG while (iii) contains GCCAC as the last fragments. The latter, GCCAC, is nothing but CACCG in the forward direction.

For each of the other combinations (*a*) (i), (*a*) (ii), (*a*) (iv), (*b*) (i), *b* (ii) and (*b*) (iii) there were 3 unused fragments. With the forward chain, the reverse chain and 3 unused fragments two complete sequences could be obtained with each of the 6 combinations. Thus a total of 12 sequences could be obtained at the final stage. The anticodon arm has a number of regularities common to all tRNAs. It has 5 hydrogen bonded pairs and 7 unpaired nucleotides. Besides there is one unpaired nucleotide between the dihydrouridine arm and the anticodon arm. Only one sequence from combination (*a*) (iv) satisfied all the common features of the anticodon arm and it had in the loop region a triplet which agreed with the codon for glutamine. The sequence thus obtained is shown below.

pUGGGGUAUCGCCAAGCGGDAAGGCACCGGAUUCUGAU

UCCGGCAUUCCGAGGTψCGAAUCCUCGUACCCCAGCCA.

This structure agrees with that in reference 7, the modified nucleotides being represented by the parent nucleotide. None of the other 11 structures obtained could give a complete cloverleaf with its common features. The method is applicable to any tRNA provided the nucleotide sequences of the individual fragments are correctly determined. It is likely that we may not end up with a unique solution in every case. However as the secondary structure constaints on a tRNA are numerous we strongly feel that the number of possible sequences will be small and the method will greatly reduce the time needed to sequence a tRNA. Once the enzymatic fragments are well characterised the time needed to do the computer work is very short. In the present study selection of the correct sequences at various stages was done manually from the computer output. This operation also may be computerised, if needed.

At present, it appears that the method is suitable only for tRNA. However, it is applicable to any RNA provided sufficient secondary structure constraints are available. A common secondary structure with 4 base paired regions applicable to several species of 5S RNA has been proposed[12]. Whether this will be sufficient to work out the sequence of 5S RNA by the computer method remains to be tested.

ACKNOWLEDGEMENT

The award of fellowships to N. P. R. by University Grants Commission, New Delhi and to P. J. by Council of Scientific and Industrial Research, New Delhi, is gratefully acknowledged.

1. Holley, R. W., " Prog. in nucleic acid research," *Mol. Biol.*, 1968. **8**, 37.
2. Ohashi, K., Harada, F., Ohashi, Z., Nishimura, S., Stewart, T. S., Vogeli, G., McCutchan, T. and Soll, D., *Nucleic Acid Research*, 1976, **3**, 3369.
3. Tinoco, I., Uhlenbeck, O. and Levine, M., *Nature*, 1971, **230**, 362.
4. Merril, C. R., Shapiro, M. B., Bradley, D. F., Mosimann, J. E. and Vinton, J. E., *Biopolymers*, 1966, **4**, 723.
5. Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, T. R. and Zamir, A., *Science*, 1965, **147**, 1462.
6. Pipas, J. M. and McMohan, J. E., *Proc. Natl. Acad. Sci., U.S.A.*, 1976, **72**, 2017.
7. Folk, W. R. and Yaniv, M., *Nature*, N.B., 1972, **237**, 165.
8. Barrell, B. G. and Clark, B. F. C., *Handbook of Nucleic Acid Sequences*, Joynson-Bruvvers, Oxford, 1974, p. 5.
9. Dube, S. K., Marcker, K. A., Clark, B. F. C. and Cory, S., *Nature*, 1968, **218**, 232.
10. Dube, S. K., Marcker, K. A. and Yudelevich, A., *FEBS Letters*, 1970, **9**, 108
11. Dudock, B. S., Katz, G, Taylor, E K. and Holley, R. W., *Proc. Natl. Acad. Sci., U.S.A.*, 1969, **62**, 941.
12. Fox, G. E. and Woese, C. R. *Nature*, 1975, **256**, 505.