Bull. Inst. Chem. Res., Kyoto Univ., Vol. 56, No. 3, 1978

# Utilization of Computer for Determination of a Nucleotide Sequence

Tatsuo Ooi and Ken Nishikawa*

*Received March 31, 1978*

In the determination of a nucleotide sequence, RNA is digested into fragments by RNase $T_1$ and $A$, which have different specificity. The sequence may be determined by arranging the fragments in the correct order and the connecting informations are supplied by the digests with RNase of different specificity. A computer may be utilized for the search of all the ways to arrange the fragments according to the given informations. The results show that, for a RNA of a length of about 100 nucleotides, the set of fragments of both digests are usually not enough to give the unique sequence, and further data on partial digests give rise to the only one sequence. This method, although now classical, may be useful for the determination of nucleotide sequences where new techniques are not applicable.

## INTRODUCTION

A remarkable advance in the determination of nucleotide sequences has been made in the past decade as shown, for instance, by the establishment of the complete DNA sequence of bacteriophage $\phi$X174.[1] The recent progress owes to the development of new techniques such as "plus and minus" technique invented by Sanger and Coulson[2] and the Maxiam and Gilbert method[3] for a direct determination of DNA sequences. In contrast to the new techniques, the method of utilization of nucleases (*i. e.*, RNase $T_1$ which cleaves 3'nucleotide phosphate bonds of guanosine with high specificity and pancreatic RNase A which cleaves pyrimidine nucleotide) becomes rather classical. This method, although it is now an old technique since the new techniques are extended to the determination of RNA sequences,[4,5] may still be useful for the case where the new techniques are not applicable.

The principle of the method is simple: fragments obtained by the RNase $T_1$ digestion (T fragments) are connected according to the information about connecting regions given by fragments derived by the RNase A digestion (A fragments). Since the information is not enough to yield a unique sequence, partial digestion data by RNase $T_1$ are crucial for the further information about the junctions. Because of many possible combinations of the T fragments, the number of nucleotides is usually limited about 50 to 100. For such a problem a computer may be available to look for whole the combinations of T fragments based on the information given by a set of data. In this paper, utilization of a computer for the connection of the T-fragments is described.

* 大井龍夫, 西川 建: Laboratory of Physical Chemistry of Enzyme, Institute for Chemical Research, Kyoto University, Uji, Kyoto-Fu 611, Japan.

## PROCEDURE

### I Fragments

To begin with, we have to set up conditions imposed on fragments obtained experimentally to write computer programs. The conditions are as follows: (see Table I)

**(a) T fragments**

(1) Numbers and nucleotide sequences of T fragments are given accurately (including single nucleotides).

(2) Nearest neighbors of the fragments are known. This information reduces the number of combination significantly.

(3) The first 5'terminal fragment is identified, but the 3'terminal fragment is not necessarily known.

**(b) A fragments**

(1) Numbers and sequences of A fragments containing guanosine (G) are given accurately including nearest neighbors.

(2) Fragments which do not contain G are not essential.

(3) The first 5'terminal fragment is known.

**(c) Partial digest data**

Orders of T fragments in partial digests are known.

### II Connection of fragments

Connection of T fragments may be performed using the information about the sequence and the nearest neighbor nucleotide of a T fragment. Since a T fragment is characterized by the first nucleotide as a head and the nearest neighbor as a tail, a fragment may be linked to another fragment, the head nucleotide of which is identical with the tail of the former fragment. Trials of connection of two fragments are performed for all the possible combinations. Among them, the combinations are selected when the sequence in the connecting region is found in an A fragment. If a single nucleotide fragment is involved in the combination, we have to consider the arrangement of three successive fragments since two fragments are not enough to compare the connecting sequence with an A fragment.

The number of T fragments reduces by one when only one combination is possible for the connection of two fragments, and the corresponding A fragment is omitted from the data list. After reduction of the T fragments by the above procedure, we have a final list of T and A fragments after connecting such T fragments that could be linked uniquely. If lucky, a unique solution of the sequence might be obtained without knowledge about partial digests. The list is made as a matrix table according to the information about the possible combination of each fragment (see Table III).

With the use of the table list, all the possible solutions of the arrangment of T fragments from the 5'terminal T fragment may be computed by connecting fragments step by step according to the given information. Of course, A fragments should correspond one to one to the connecting regions, and if not the connection is invalid.

Programs were written in Fortran S and all the computations were performed by FACOM 230-48 at the computing center of the Institute for Chemical Research, Kyoto

University.

## RESULTS AND DISCUSSIONS

We will take a part of the mRNA sequence of bacteriophage fd (from 1 to 108 in G3 RNA of *Hap* 1 - *Hae* 1)[6] as an example. Data of T and A fragments are listed in Table I. In this table, the data of T fragments and G-containing A fragments are

Table I. Fragments Obtained by the Digestion with RNase $T_1$ ($T$ Fragments) and RNase $A$ ($A$ Fragments).

| *T* fragment | | | *A* fragment | | |
|---|---|---|---|---|---|
| 22 | | | 24 | | |
| 1 | 1 | GG | 1 | 1 | GGGGUC |
| 2 | 1 | AGU | 2 | 1 | ACC |
| 3 | 1 | UGC | 3 | 1 | ACG |
| 4 | 2 | UGU | 4 | 2 | GCC |
| 5 | 1 | UGG | 5 | 1 | GGCA |
| 6 | 1 | AUGA | 6 | 1 | GGAAACU |
| 7 | 1 | UAGU | 7 | 2 | AUU |
| 8 | 1 | UCAAAGA | 8 | 1 | AUG |
| 9 | 1 | UUGG | 9 | 1 | AAUG |
| 10 | 1 | CCUUCGU | 10 | 3 | GUA |
| 11 | 1 | CAUUACGU | 11 | 3 | GUU |
| 12 | 1 | UUUUAGU | 12 | 2 | AGUG |
| 13 | 1 | UUUUAGG | 13 | 1 | AAAGAUG |
| 14 | 1 | UUUAAUGG | 14 | 1 | GGUG |
| 15 | 1 | CCUCUUUCGU | 15 | 1 | GAGUG |
| 16 | 1 | AAACUUCCUCAUGA | 16 | 1 | AGGUU |
| 17 | 1 | UAUUUUACCCGU | 17 | 0 | CC |
| 18 | 1 | UAUUCUUUCGC | 18 | 0 | CA |
| 19 | 1 | GC | 19 | 0 | CU |
| 20 | 1 | GA | 20 | 0 | CG |
| 21 | 3 | GU | 21 | 0 | UC |
| 22 | 2 | GG | 22 | 0 | UA |
| | | | 23 | 0 | UU |
| | | | 24 | 0 | UG |

correct, but numbers of some of other A fragments, especially single nucleotides, are written artificially incorrect to show the application of the program. The correct numbers may be given in Table II after comparison with the data of T-fragments. Numbering of the fragments is as follows; # 1 is the 5'terminal fragments, and the rest of fragments are arranged arbitrarily except single nucleotide fragments, which should be put in the last part of the list for the convenience of the program as shown in Table I.

Table II illustrates a new set of T fragments which are renumbered after linking some of the fragments using information given by A fragments; the original 22 T fragments reduce to 16 after unique linkages. The order of the old fragments in the new T fragments in the original numbers (Table I) is also shown in Table II (see the fragments # 1, # 14, and # 15). An A fragment, the information of which has been fully used, is omitted from the list. In Table II, the number of # 1 A fragment is zero, indicating the information has been utilized already. However, the connection of old #

Table II. The Out-Put Listing after Linking *T* Fragments by the Information of *A* Fragments

| ** | T-FRAGMENTS | ** | TOTAL NUMBER=16 | |
|---|---|---|---|---|
| 1 | CLASS 1 | No. 1 | GGGGUCAAAGAUGAGU | 1. 22. 22. 21. 8. 2 |
| 2 | CLASS 1 | No. 1 | UGC | 3 |
| 3 | CLASS 1 | No. 2 | UGU | 4 |
| 4 | CLASS 1 | No. 1 | UGG | 5 |
| 5 | CLASS 1 | No. 1 | UAGU | 7 |
| 6 | CLASS 1 | No. 1 | UUGG | 9 |
| 7 | CLASS 1 | No. 1 | CCUUCGU | 10 |
| 8 | CLASS 1 | No. 1 | UUUUAGU | 12 |
| 9 | CLASS 1 | No. 1 | UUUUAGGU | 13 |
| 10 | CLASS 1 | No. 1 | UUUAAUGG | 14 |
| 11 | CLASS 1 | No. 1 | CCUCUUUCGU | 15 |
| 12 | CLASS 1 | No. 1 | UAUUUUACCCGU | 17 |
| 13 | CLASS 1 | No. 1 | UAUUCUUUCGC | 18 |
| 14 | CLASS 1 | No. 1 | GCAUUACGU | 19. 11 |
| 15 | CLASS 1 | No. 1 | GAAACUUCCUCAUGA | 20. 16 |
| 16 | CLASS 2 | No. 1 | GU | 21 |

NO LINK FROM I=15

| ** | A-FRAGMENTS | ** | TOTAL NUMBER=19 |
|---|---|---|---|
| 1 | | No. 0 | GGGGUC |
| 2 | | No. 2 | GCC |
| 3 | | No. 1 | GGCA |
| 4 | | No. 1 | GGAAACU |
| 5 | | No. −1 | AUU |
| 6 | | No. 3 | GUA |
| 7 | | No. 3 | GUU |
| 8 | | No. 2 | AGUG |
| 9 | | No. 1 | GGUG |
| 10 | | No. 1 | GAGUG |
| 11 | | No. 1 | AGGUU |
| 12 | | No. −2 | CC |
| 13 | | No. −2 | CA |
| 14 | | No. −5 | CU |
| 15 | | No. −4 | CG |
| 16 | | No. −7 | UC |
| 17 | | No. −5 | UA |
| 18 | | No. −13 | UU |
| 19 | | No. −1 | UG |

19 T fragment to ♯ 11 corresponds to the new ♯ 14 T fragment, but the information used for the connection by ♯ 3 A fragment has been used in part, so that the number is not reduced yet at this stage (Table II). For the convenience of programming, the first 5' fragments were not omitted from the list even when the number was zero.

The number of A fragments diminishes with the progress of connection, leaving only fragments that have no G. Negative numbers shown in the table implies that the correct number plus this number of the A fragment yielded the number listed in Table I, *i. e.*, the correct number of AUU (♯ 5 in Table II) is 3 (the number of AUU in Table I (♯ 7), 2, plus−(−1)). The numbers of single nucleotides are given in an opposite

Table III. Matrix Table of Fragments

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 10 | 10 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 6 | 7 | 0 | 7 | 7 | 7 | 0 | 6 | 6 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | −9 |
| 5 | 0 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | −9 |
| 7 | 0 | 0 | 0 | 0 | 6 | 7 | 0 | 7 | 7 | 7 | 0 | 6 | 6 | 0 | 0 | 0 |
| 8 | 0 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 11 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | −9 |
| 11 | 0 | 0 | 0 | 0 | 6 | 7 | 0 | 7 | 7 | 7 | 0 | 6 | 6 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 6 | 7 | 0 | 7 | 7 | 7 | 0 | 0 | 6 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 6 | 7 | 0 | 7 | 7 | 7 | 0 | 6 | 6 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | −9 | −9 | −9 | 0 | −11 | 0 | −11 | −11 | −11 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |

sign in Table II, because the numbers have been set zero in Table I. If there is a fragment which has no possibility to link to any T fragment, the fragment is listed as "NO LINK FROM I=15" (see Table II). This # 15 fragment corresponds to the 3' terminal fragment, since the last fragment has no succeeding fragment.

The reduced fragments are now regarded as starting fragments which could be connected by information about connecting regions from A fragments in Table II. The possible ways of a T fragment to others are listed in Table III by showing the number of a fragment in column and the number of succeeding fragments in row. Numbers shown in the table are A fragments which correspond to connecting regions (expressed in the new order in Table II), and zero means that there is no connecting fragment, or such a linkage is impossible. For example, the # 1 fragment has three ways to be combined (to # 2, # 3, and # 4 fragment), where the connecting information is supplied by # 10 A fragment (in Table II not in Table I) and so on. From # 15 fragment there is no succeeding fragment because all the numbers in row are zero. With the use of this table, the number of possible combinations which satisfy the connecting conditions imposed by fragments was over 100,000, apparently too many to look for the unique sequence. Therefore, we need further information about junctions to obtain the final unique sequence.

An example for partial digests by RNase $T_1$ is shown in Table IV. When the order of the fragments in partial digests is known, the number of T fragments is reduced after connection of further unique linkages by utilization of the data. Table V illustrates that the information of three partial digests (# 18 to # 15, # 14 to # 20 and to # 16, and # 11 to # 17 in Table I) gives rise to 12 new T fragments after connection of some of T fragments in Table III. The possible combinations of these fragments are examined by the similar matrix to Table III (see Table V) and the number of such combinations that the total fragments can be linked in consistent with the data of A fragments, is 1, 512, *i. e.*, the partial digests information reduces the number 100, 000 to 1, 512. The more informations we have, the smaller the number of the solutions. The reason why different numbers of the solutions (216 and 108 in column 4 and 5 in Table IV, respectively) are

Table IV. Listing of the Reduction of Fragments and Solutions by the Use of Partial Digests Data

| Data of Partial Digests | No. of Fragments | No. of Solution |
|---|---|---|
| None | 16 | >100,000 |
| 18. 15<br>14. 20. 16<br>11. 17 | 12 | 1,512 |
| 14. 20. 16.<br>11. 17<br>4. 18. 15 | 12 | 756 |
| 9. 21. 3. 10<br>18. 15<br>14. 20. 16<br>11. 17 | 9 | 216 |
| 4. 18. 15<br>14. 20. 16<br>9. 21. 3. 10. 7 | 9 | 108 |
| 14. 20. 16<br>9. 21. 3. 10. 7<br>11. 17<br>4. 12. 4. 18. 15 | 6 | 4 |
| 11. 17. 14. 20. 16<br>9. 21. 3. 10. 7<br>4. 12. 4. 18. 15 | 1 | 1 |

pppGGGGUCAAAGAUGAGUGUUUUAGUGUAUUCUUUCGCCUCUUU
CGUUUUAGGUUGGUGCCUUCGUAGUGGCAUUACGUAUUUUACCC
GUUUAAUGGAAACUUCCUCAUG(A)

Table V. An Example of the Reduction of Fragments (the Second Data in Table IV).

| ** | T-FRAGMENTS | ** | TOTAL NUMBER=12 |
|---|---|---|---|
| 1 | CLASS 1 | No. 1 | GGGGUCAAAGAUGAGU |
| 2 | CLASS 1 | No. 1 | UGCCUUCGU |
| 3 | CLASS 1 | No. 2 | UGU |
| 4 | CLASS 1 | No. 1 | UGG |
| 5 | CLASS 1 | No. 1 | UAGU |
| 6 | CLASS 1 | No. 1 | UUGG |
| 7 | CLASS 1 | No. 1 | UUUUAGU |
| 8 | CLASS 1 | No. 1 | UUUUAGGU |
| 9 | CLASS 1 | No. 1 | UUUAAUGGAAACUUCCUCAUGA |
| 10 | CLASS 1 | No. 1 | UAUUCUUUCGCCUCUUUCGU |
| 11 | CLASS 1 | No. 1 | GCAUUACGUAUUUUACCCGU |
| 12 | CLASS 2 | No. 1 | GU |

NO LINK FROM 1=9

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 4 | 5 | 5 | 5 | 5 | 4 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 4 | 5 | 5 | 5 | 5 | 4 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | −7 |
| 5 | 0 | 6 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | −7 |
| 7 | 0 | 6 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 9 | 9 | 0 | 9 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 4 | 5 | 5 | 5 | 5 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 4 | 5 | 5 | 5 | 5 | 4 | 0 | 0 |
| 12 | 0 | −7 | −7 | −7 | 0 | −9 | −9 | −9 | −9 | 0 | 0 | 0 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

obtained for the same number (9) of T fragments in Table IV, is that the ways of connections for every fragments (as shown in Table III) are different. In order to deduce a probable sequence, the number of fragments should be reduced to a few after connection of T fragments. Finally, we can get the unique one solution as shown in Table IV. Since we search all the possible combinations, the sequence is only one that satisfies the given data, and there is not any other solution.

The programs have been written for a maximum length of 50 fragments or about 200 nucleotides. According to our experience, around 100 nucleotides seems to be a suitable size for the application of this kind of computation since the ways of combinations increase with the number of T fragments significantly as shown above. Of course the computation is possible for a longer chain provided that data of partial digests are enough to make the number of T fragments smaller after linkage. For this purpose, the knowledge about nearest neighbors is essential as expected by a simple estimation that the number of combinations increases with $n$ th power ($n$ is the number of fragments) of 4. Also we can estimate how much information A fragments have by computing the matrix as Table III. If the fragments could not be reduced to around a few, the combinations might be too many to inspect a probable sequence.

The practical application to the sequence determination may be performed as follows; (1) a piece of RNA, the sequence of which is going to be determined, is digested by RNase $T_1$ and A completely, and T fragments and G-containing fragments are analyzed accurately. At this stage we can examine unique linkages by the information (also consistency of the number of G in both fragments is checked by the program). (2) Taking partial digests, numbers of T fragments in each digest may be identified rather with ease. Among them, the order of the fragments may be determined by the information of A fragments. (3) Partial digest data yield a reduced number of T fragments, and collection of further data is proceeded until the unique solution is obtained. Actual computing time is less than 1 min unless the number of solutions exceeds $10^5$.

The present computations were performed to link T fragments according to the information about connecting regions given by A fragments and $T_1$ partial digests. It is possible, in turn, to connect A fragments by the information given by T fragments. However, RNase A cleaves a 3'phosphate bond of C or U, so that ways to link one fragment to another increase singificantly as mentioned before. Thus, such a method is not practical. If we could obtain fragments cleaved at C, *e.g.*, by modification of U base or by the use of the modified RNase A, they would be available for the connecting fragments because the fragments are equivalent to T fragments of the complementary RNA chain.

Another set of possible fragments is the digestion material by RNase $U_2$ which attacks 3'phosphate bond of A. Since the more information about connecting regions, the smaller the final fragments after unique linkages, preliminary computation on such a hypothetical system that complete sets of fragments by $T_1$, $U_2$, and A digestion are given, yields about 10 solutions derived after the possible linkages, for a similar size of RNA, around 100 nucleotides. This may be an expected consequence because we have more information about connecting regions. Whether such fragments are practically available or not is dependent on the experimental accuracy. Therefore, at present it seems to be rather practical to use $T_1$ partial digest data as described here.

Further improvement of the present method may be to make the use of partial digest data more practical, since the order of T fragments in a partial digest is necessary condition in the program. We need some information about the connection of the T fragments in the partial digests, *e. g.*, A fragments. Therefore, it is more convenient if the condition is loosen in such a way that the order of the fragments is not necessary. The programming along this line is now in progress.

## ACKNOWLEDGMENT

The authors express their thanks to Professor M. Takanami for providing the present problem and stimulating discussions.

## REFERENCES

(1) F. Sanger, G. M. Air, B. G. Barrel, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith, *Nature*, **265**, 6870 (1977).
(2) F. Sanger and A. R. Coulson, *J. Mol. Biol.*, **94**, 441 (1975).
(3) A. M. Maxiam and W. Gilbert, *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 560 (1977).
(4) Donis-Keller, A. M. Maxiam, and W. Gilbert, *Nucleic Acids Res.*, **4**, 2527 (1977).
(5) M. Szekeley, *Nature*, **269**, 754 (1977).
(6) K. Sugimoto, H. Sugisaki, T. Okamoto, and M. Takanami, *J. Mol. Biol.*, **111**, 487 (1977).