Biophysical Chemistry 5 (1976) 369–375

DETERMINING THE BASE SEQUENCE OF RNA MOLECULES USING CONTINUOUS DEGRADATION KINETICS

Gerhard KESSLING, Philip K. RAWLINGS and Manfred EIGEN
Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen-Nikolausberg, FRG

Received 16 March 1976

A kinetic approach to sequence analysis is presented. A polymer chain composed of a small number of monomer types is degraded continuously with degradation proceeding in only one direction of the chain. Ordinarily, the direct determination of the sequence from the amount of degraded monomers is hampered by the rapid loss of reaction synchrony. However, it is shown, that if the reaction mechanism is known, and the concentrations of the degraded monomers can be measured reasonably accurately, one can solve a system of linear equations with the unknowns giving the positions of the monomers in the chain, and the synchrony loss due to the stochastic nature of the degradation process can be accounted for. Model calculations are presented.

1. Introduction

The most common procedure used to determine the sequence of a long RNA molecule consists basically of four steps. First, RNA strands having the same sequence are isolated and purified. Second, using various enzymatic and chemical processes, the RNA is split into a number of small fragments. Third, the fragments are separated from each other and the relative frequency and sequence of each fragment is determined. Finally, the original RNA sequence is obtained by fitting the small fragments from each cleaving operation on the RNA back together. For short fragments containing only a few nucleotides, the sequence can be obtained fairly easily, and quite a number of procedures, such as matching overlapping portions of fragments obtained from different cleaving experiments, have been developed for estimating the order in which each fragment appears in the sequence [1–3]. Especially notable is the kinetic procedure developed by the Weissmann group [4]. They estimate the relative position of fragments in a sequence by synthesizing the RNA in vitro. All RNA molecules start growing simultaneously after varying time intervals. The synthesis is rapidly quenched. Conventional fragmentation techniques are then applied to the partially synthesized RNA samples. The fragments are ordered by the synthesis time required before they make their appearance. This procedure has been effective for identifying sequences up to several hundred nucleotides in length. However, its practical usefulness is somewhat limited since it can only be applied to RNA molecules that can be synthesized in vitro.

At least in principle, the fragment sequencing methods described above are very straightforward, but in actual practice they suffer from several serious drawbacks. Frequently, enough information is not available to be certain of the exact position some fragments occupy in the original RNA sequence. The separation and identification of the fragments is a long, laborious task and a great amount of RNA starting material is required. Perhaps the most serious difficulty is the fact that these methods are not easily adaptable to automation techniques. However, other sequencing routines are available which avoid these difficulties to some extent.

One alternative sequencing procedure starts with a a sample of homogeneous RNA polymer and chemically removes a single nucleotide from one end of each molecule. The resulting monomer base is separated from the remaining RNA chains and identified, thereby determining the first nucleotide in the original sequence. This process is iterated step by step, removing the next nucleotide from the RNA chains, separating and

identifying it [5,6]. This method has the advantage that the same chemical steps are repeated in each cycle and the procedures are fairly simple, allowing the entire process to be automatized. Also the monomer bases produced by the degradation can be easily identified and their position in the original RNA sequence is unambiguous. The chief difficulties are the relatively large RNA samples required and the loss of synchrony between individual RNA chains. In a given cycle, some chains may not lose any nucleotides. The effect of this error is cumulative and leads to the eventual breakdown of the method. Sequences as long as 19 nucleotides have been determined for a t-RNA by this procedure [5]. By way of comparison, RNA sequences of several hundred nucleotides have been determined by the fragmentation method described in the first paragraph, so the biggest advantage of this alternate technique is in the amount of effort that is saved in determining the first few nucleotides of a long sequence.

Both the loss of material in each step and the loss of synchrony between the RNA chains are the major limitations on the method described in the preceeding paragraph. Because of the stochastic nature of the degradation process, the loss of synchrony is virtually inevitable, no matter what chemical procedures are used to remove the nucleotides. Therefore, a modification of the method which allows this to happen in an orderly fashion, while conserving all the material of the original RNA sample, might allow sequences somewhat longer than 19 nucleotides to be identified. Procedures where RNA molecules are continuously degraded by enzymes have been described in the literatures, but so far these have been successful in producing a definite sequence only with very short chains [7]. The difficulties involved with continuous degradation of an RNA polymer are both experimental and theoretical in nature, but the rest of this paper deals largely with the theoretical aspects of the problem.

2. Description of the model

We start with a completely homogeneous sample of RNA molecules having an arbitrary chain length. Our method requires that each RNA molecule in the sample be degraded by a chemical process which removes nucleotides individually from the polymer. This degradation process must be irreversible and must proceed in one direction only; one end of each chain must be chemically inactive and the nucleotide occupying this position is therefore the last to be released by the degradation. No side reactions of any kind are permitted, and once a nucleotide has been severed from its RNA chain it undergoes no further chemical changes. Initially all RNA chains are identical and the chemical degradation begins simultaneously for all the polymers in the sample. As the molecules are gradually broken down, a distribution of various chain lengths for the remaining RNA strands will develop. The occurance of RNA strands with varying chain lengths is inherent in the random nature of the degradation process and it presents no problems, if the mechanism and the rate constants of the process are known. The basic kinetic steps of a simple degradation process are represented by the following equation:

$$\mathrm{RNA}_n \xrightarrow{k_0} \mathrm{RNA}_{n-1} + \mathrm{N} \xrightarrow{k_1} \mathrm{RNA}_{n-2} + 2\mathrm{N} \xrightarrow{k_2} \mathrm{RNA}_{n-3} + 3\mathrm{N} \xrightarrow{k_3} \ldots \mathrm{RNA}_2 + (n-2)\mathrm{N} \xrightarrow{k_{n-2}} n\mathrm{N}\,, \tag{1}$$

where k_i ($i = 0, 1, 2, \ldots, n-2$) is the rate constant for the removal of the $(i+1)$th nucleotide in the sequence; RNA_{n-i} ($i = 0, 1, 2, \ldots, n-2$) represents those RNA chains with i nucleotides removed; while N represents the monocleotides removed from the RNA chains.

We will assume that the rate constants are all known (see below). Ideally, all the rate constants would be equal to each other and thus independent of both the specific sequence of the polymer under investigation, as well as its chain length. However, it is not necessary for all the rate constants to be equal or even independent of the nucleotide sequence for our sequencing method to work.

From eq. (1), the rate of change for the concentration of each RNA chain length can be written as:

$$\mathrm{d}(\mathrm{RNA}_n)/dt = -k_0(\mathrm{RNA}_n)\,, \tag{2a}$$

$$\frac{\mathrm{d}(\mathrm{RNA}_{n-1})}{\mathrm{d}t} = +k_0(\mathrm{RNA}_n) - k_1(\mathrm{RNA}_{n-1})\,, \tag{2b}$$

$$\frac{\mathrm{d}(\mathrm{RNA}_{n-2})}{\mathrm{d}t} = +k_1(\mathrm{RNA}_{n-1}) - k_2(\mathrm{RNA}_{n-2})\,, \tag{2c}$$

$$\frac{d(RNA_{n-3})}{dt} = k_2(RNA_{n-2}) - k_3(RNA_{n-3}) , \quad (2d)$$

etc., where n is the number of nucleotides present in the RNA molecules before degradation starts.

The rate of formation for all the nucleotide degradation products is:

$$d(N)/dt = k_0(RNA_n) + k_1(RNA_{n-1}) + k_2(RNA_{n-2}) + k_3(RNA_{n-3}) + \ldots + 2k_{n-2}(RNA_2) . \quad (3)$$

Eq. (2) describe the complete time dependence of all possible RNA chain length concentrations. These equations can be integrated either by analytical or numerical methods. Thus it is possible to calculate exactly how the chain length distribution function changes during the degradation process [8]. Once this is known the sequence of the RNA molecule can be determined in the manner described below:

We assume that an RNA molecule will contain λ different types of nucleotides. The production rate for each of the λ bases must be measured experimentally throughout the degradation of the RNA sample.

The total production rate of nucleotides is then:

$$d(N)/dt = \sum_{j=1}^{\lambda} d(N_j)/dt , \quad (4)$$

where $d(N_j)/dt$ is the kinetic rate of formation of the jth nucleotide. It is convenient to use an indexed variable, Ω_{jm}, to represent the sequence of an RNA molecule. This variable is defined to be equal to unity if the mth position in the RNA sequence is occupied by a jth type nucleotide and Ω_{jm} is equal to zero if the mth position is not a jth nucleotide. Thus for any value of m, only one of the λ different Ω_{jm} is non-zero. The production rate for each type of nucleotide is defined to be:

$$d(N_j)/dt = \sum_{i \in I} k_{n-i}(RNA_i) , \quad (5)$$

where I is the set of all chain lengths which end with a jth type nucleotide. The set I is a function only of the RNA under investigation. This expression can be rewritten in a simple form using the Ω's:

$$d(N_j)/dt = \sum_{i=0}^{n-2} \Omega_{ji} k_i(RNA_{n-i}) , \quad (6)$$

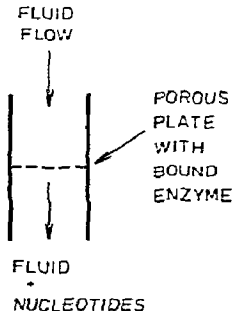

Fig. 1. Schematic diagram of a flow apparatus for determining RNA sequences. The enzyme–RNA complex is bound to a carrier material. The concentrations of the nucleotides that have been released from the chain are measured at the outlet of the apparatus.

where n is the number of nuleotides originally in each RNA molecule and $j = 1, 2, \ldots, \lambda$.

A reaction chamber suitable for measuring the degradation kinetics of an RNA polymer is shown in fig. 1. It consists basically of a mechanically rigid, but porous plate mounted inside a tube. Fluid is pumped through the chamber at a constant rate. The surface of the plate is continuously swept clean by the fluid as it passes down the tube. An appropriate degrading enzyme has been chemically attached to the surface of the plate. The RNA polymer is bound tightly to the enzyme and neither the RNA nor the enzyme–RNA complex can be removed from the surface of the plate by the motion of the fluid. However, as soon as individual nucleotides are released by the enzyme–RNA complex, they immediately begin to move with the fluid and are carried out of the reaction chamber. Thus the concentration of each nucleotide in the fluid leaving the tube is directly proportional to the production rate for the nucleotide when the fluid passes through the plate.

By measuring the various nucleotide concentrations in the fluid leaving the reaction chamber, the quantity on the left of eq. (6) is determined experimentally. Since the rate constants for the degradation mechanism are assumed to be known, the concentrations of the various different chain lengths can be calculated by simulating the stochastic process on a computer. Consequently, eq. (6) represents a series of linear equations with the unknown quantities being the Ω's.

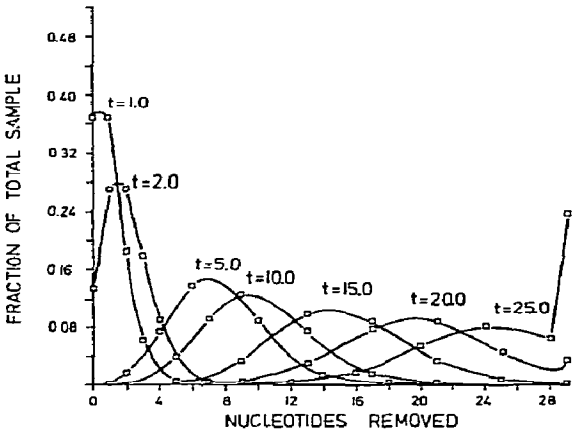

Fig. 2. The fraction of RNA chains having from 0 to 29 nucleotides removed for 7 different times. Time is in reduced units. All rate constants are equal.

3. Calculations

In this section some sample computer simulations are presented to illustrate the proposed sequencing method. An RNA polymer that originally contained 30 nucleotides was used in all the computer calculations described in this section. The degradation mechanism was assumed to be linearly dependent on the concentration of each RNA chain length and the rate constants for the removal of individual nucleotides were all taken to be equal. The most probable distribution of chain lengths for the stochastic process was then calculated as a function of time. Of course, these properties were chosen somewhat arbitrarily. Most situations encountered in the laboratory will not be quite as simple, but even so this linear model reproduces the basic features of a real system.

In fig. 2, the most probable fraction of RNA chains having from 0 to 29 nucleotides removed has been plotted for 7 different times. Time for the degradation process has been expressed in reduced units. Over approximately the first 15 reduced time units, an average of one nucleotide is removed from all RNA chains during each unit of time. At time zero, all RNA chains have the same number of nucleotides. After the degradation process has been active for one time unit, about a third of the RNA chains have lost exactly one nucleotide, while the remaining RNA molecules have lost no nucleotides at all. After two units of time, about an eighth of all the RNA chains are still completely intact, while approximately half the chains have lost one or two nucleotides, and the remainder have lost three or more. Thus the distribution of chain lengths rapidly broadens as the degradation continues. Only when some chains have been completely consumed does the distribution begin to narrow with increasing time.

Fig. 3 shows the production rate for each of four different nucleotides present in an RNA molecule with a specific sequence. In the initial phase of the reaction, the production rates reflect the order in which each type of nucleotide appears in the sequence. Thus, the nucleotide with the largest initial production rate appears first in the sequence, the nucleotide with the second largest rate appears next, and so forth. Therefore, it can be seen immediately from the nucleotide production plotted in fig. 3 that the order of appearance in this sequence is first type 1; second type 2; third type 4; and fourth type 3. The production of type 2 nucleotides peaks at between 1 and 2 units of time, indicating that this nucleotide occupies the second position in the sequence. The production of type 4 nucleotides peaks at about 2 units of time, suggesting that the third position in the sequence has this type

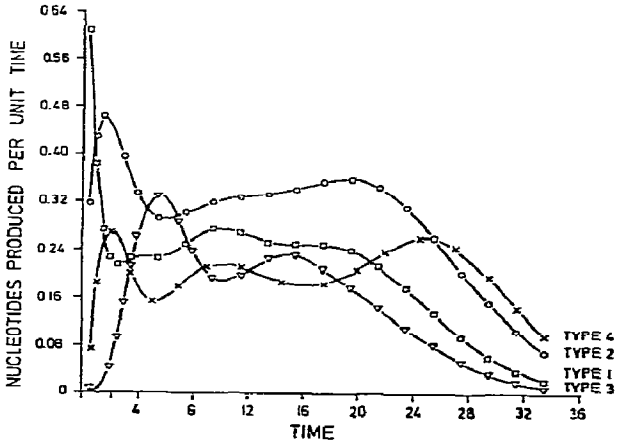

Fig. 3. The production rates of the four types of nucleotides for the sequence (1,2,4,2,1,3,3,2,4,1,2,1,4,2,3,3,2,1,4,1,2,2,3,2,1,4,4,4,2,3). Type 1: □, type 2: ○, type 3: △, type 4: X.

of nucleotide. However, the peak in the production of type 3 nucleotides doesn't occur until more than 5 units of time have elapsed. This suggests that several type 4 nucleotides are present in the fifth, sixth, or seventh position of the sequence. The broad type 2 production peak, as well as the small peak in the type 1 production curve at about 4 units of time, indicate that the fourth and fifth positions are occupied by type 1 and 2 nucleotides, respectively.

Generally speaking, the peaks in the nucleotide production curves indicate regions where one type of nucleotide is more common than the others, but obviously as the kinetic degradation proceeds it becomes increasingly difficult to interpret the sequence by inspection. However, the method described in the previous section can easily determine all the positions in the sequence. The minimum amount of data required to identify the sequence in fig. 3 is the production rate for each for the four nucleotide types measured at 30 different times over the first 30 time units of reaction. For each nucleotide type there are 30 unknowns corresponding to the presence or absence of the nucleotide at each position in the sequence. It is advisable to have somewhat more than the minimum number of equations for calculating the sequence, as this helps to reduce the effect of small random errors that occur when measuring each nucleotide production rate. The equations can be programmed for a computer to obtain their solution [9]. This solution is then a 30 element vector and there is one such vector for each type of nucleotide. Each element of the vector is either close to unity or close to zero, the former indicating the presence of a particular nucleotide, while the latter indicates the absence of that nucleotide in the sequence position represented by the particular element.

The method is also sensitive to small changes in a sequence. Fig. 4, for instance, shows the nucleotide productions for almost the same sequence as fig. 3. The only differences are that positions 7 and 8 as well as 20 and 21 have been interchanged. The production of type 4 nucleotides in fig. 4 is identical with the previous sequence. This is to be expected since none of this nucleotide changed its position. The production of type 1 nucleotides is about 10% lower around 16 time units, but elsewhere it is essentially the same as the first sequence. The first peak in the type 3 nucleotide production is somewhat lower and broader

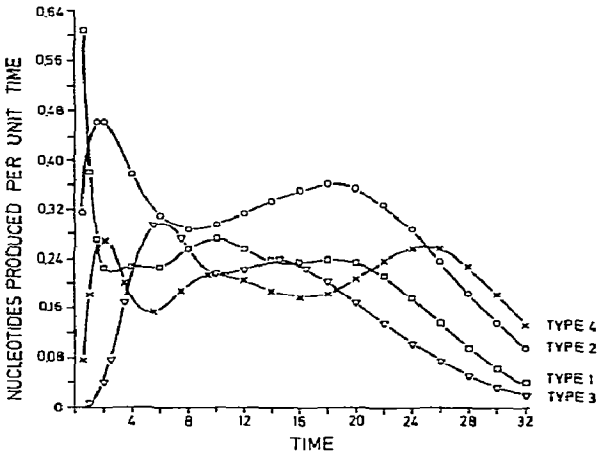

Fig. 4. The production rates of the four types of nucleotides for the sequence (1,2,4,2,1,3,2,3,4,1,2,1,4,2,3,3,2,1,4,2,1,2,3,2,1,4,4,4,2,3). Type 1: □, type 2: ○, type 3: △, type 4: ×. This sequence is identical with that of fig. 3, except that position 7 and 8 as well as 20 and 21 have been interchanged.

for the second sequence than it is for the first. This difference is due to the two type 3 nucleotides responsible for the peak being adjacent to one another in the first sequence, but separated by one nucleotide in the second sequence. Finally, the first peak in the production of type 2 nucleotides is narrower and not quite as high in the first sequence as it is in the second. Also type 2 production is lower in the range between 8 and 14 time units and higher in the range between 14 and 20 time units for the first sequence. Though the differences between the two sequences are not immediately apparent from their respective production curves, the theory is easily able to distinguish them.

4. Discussion

It has not been possible in this paper to give more than a brief description of our proposed sequencing method. However, it is readily apparent that its principal advantages are the ease with which it can be automated and the short time required for the kinetic analysis of an RNA sample. The main disadvantage is that the degradation mechanism for the RNA chain must be sufficiently well known to accurately simulate the stochastic process on a computer. This is one of

the primary factors determining the maximum length of an RNA polymer that can be sequenced by the method. Consequently, the most important consideration, when applying the method to the analysis of an RNA polymer in the laboratory, is to choose a chemical process for removing the nucleotides that is simple and easy to follow. An enzyme catalyzed degradation of the RNA polymer is probably the simplest procedure one could choose. However, in contrast to the numerical calculations described in the last section, the rate constants for the removal of nucleotides from the chain may not all be equal to each other. Several situations are possible and will have to be considered in analyzing actual experimental data. First, the rate of removal may depend on the type of nucleotide being removed. If this were the case, the RNA's in figs. 2–4 could have as many as four different rate constants. Second, the rate of removal might also depend on nearest neighbor nucleotides. This would mean that an RNA containing four nucleotides types might have as many as 16 different rate constants. Third, the manner in which the RNA chain is folded will have a very decisive bearing on the degradation rate. However, this last effect can be minimized, and perhaps eliminated, by working at a temperature sufficiently high for the RNA chains to be largely unfolded. The rate constants in the first two cases mentioned are general constants, independent of any particular sequence or chain length and they can be obtained experimentally. Thus the degradation kinetics of a specific RNA molecule would be determined by its sequence. Therefore, the chain length distributions used to determine the sequence of the RNA must be consistent with the distributions calculated from the sequence, once it has been tentatively identified. This additional constraint has to be taken into account whenever the rate constants are not equal and can be included in the computer program for analyzing the experimental data. For example, when eq. (6) is summed over j, the Ω_{jl} add to unity, and the only unknowns remaining are the k_j. Thus, provided the experimental data are accurate enough, the k_j will be completely determined using eqs. (6) and (2).

The sequencing method is fortunately not very sensitive to random errors in the measurement of nucleotide production rates. Simulated kinetic data containing random errors of as much as 25% in the average nucleotide production rates have been correctly sequenced for chains with up to 30 nucleotides using our proposed method. In such situations, where random errors in the experimental data are large, no sequence will fit the degradation kinetics exactly. Under these conditions, the sequence is identified as being the one which gives the lowest value for the square of the difference between the calculated and experimental nucleotide productions, summed over all the data. Thus *random* errors, which are unavoidable in experimental measurements, should not be a major barrier for our method. On the other hand, systematic errors are potentially devastating. For instance, RNA samples contaminated with chains which are missing portions of their normal sequence could make it difficult or perhaps impossible to identify the proper sequence by our method. As the comparison between figs. 3 and 4 suggests, the differences in degradation kinetics between two distinct sequences can be rather subtle and the experimental measurements must be sufficient reliable and accurate to distinguish them.

The main advantages of our sequencing method can be summarized as follows: The order in which the nucleotides appear in a sequence is directly determined from the degradation kinetics, the method may be easily automated for routine analysis of different RNA molecules, only a short time is required to collect the necessary experimental data for identifying a sequence, and the amount of material required for analysis is relatively small. The main disadvantage is that the RNA sample must be homogeneous. The maximum number of nucleotides that can be sequenced by the method is at present unknown, but we hope to determine this limit in the near future. We are now developing a laboratory procedure for obtaining suitable kinetic degradation data from an RNA sample. This project is still in its early stages. Results from it will appear in later publications. Also theoretical work has been done on a method of sequencing RNA molecules by manufacturing them from their component nucleotides using template instructed synthesis. This technique is quite different from the procedure used by Weissmann [4] and a paper describing it will be published shortly.

Acknowledgement

Material described in this paper has been submitted to Braunschweig University in partial fulfillment of the

requirements for a Master's Degree (by G.K.). Financial support was provided in part by the Alexander von Humboldt Stiftung (to P.K.R.). We also wish to thank the Gesellchaft für wissenchaftliche Datenverarbeitung mbH, Göttingen for the use of their computer facilities.

References

[1] P.T. Gilham, Annual Review of Biochemistry 39 (1970) 227.
[2] D.L. Kacian, D.R. Mills, F.R. Kramer and S. Spiegelman, Proc. U.S. Nat. Acad. Sci. 69 (1972) 3038.
[3] (a) F. Sanger, G.G. Brownlee and B.G. Barrell, J. Mol. Biol. 13 (1965) 373;
(b) G.G. Brownlee, F. Sanger and B.G. Barrell, J. Mol. Biol. 34 (1968) 379.
[4] (a) M.A. Billeter, J.E. Dahlberg, H.M. Goodman, J. Hindley and C. Weissmann, Nature 224 (1969) 1083;
(b) M.A. Billeter, Chimia 25 (1970) 181.
[5] M. Uziel and J.X. Khym, Biochemistry 8 (1969) 3254.
[6] H.C. Neu and L.A. Heppel, J. Biol. Chem. 239 (1964) 2927.
[7] G. Kaufmann, H. Groggfeld and U.Z. Litauer, FEBS Letters 31 (1973) 47.
[8] For a theoretical discussion of expressions similar to eqs. (2) see:
J. Gibbs, Biopolymers 7 (1969) 707.
[9] A FORTRAN IV program suitable for this purpose is called DLLSQ. A copy of it may be obtained from IBM's Scientific Subroutine Package (SSP).