Oligonucleotide mapping: evaluation of its sensitivity by computer-simulation

Robert Peter Aaronson, James F. Young and Peter Palese

Department of Microbiology, The Mount Sinai School of Medicine of the City University of New York, One Gustave Levy Place, New York, NY 10025, USA

Received 14 September 1981

ABSTRACT

A frequently used method of comparing large RNA molecules employs the two-dimensional display of oligonucleotides generated through the action of specific RNases (oligonucleotide mapping, fingerprinting). Using computer simulations and simple analytic expressions the number of large RNase T1-resistant oligonucleotides obtained from random RNA sequences can be estimated. The computer simulations also permit estimation of the number of large oligonucleotides which remain unchanged as random variations are introduced into a random RNA sequence. In addition, computer analysis also provides a means of estimating statistical confidence limits to be used in a quantitative comparison of fingerprints of different RNA molecules. The model shows that two RNA sequences which differ overall by 1%, 5% or 10% share, on average, only 85%, 50% or 25%, respectively, of their large oligonucleotides. Thus, the use of fingerprint analysis is recommended only when closely related RNAs or regions of RNAs are compared (sequence homology greater than 90%).

INTRODUCTION

The relatedness of different RNA molecules has been analyzed by comparison of "fingerprints" obtained after specific RNase digestion and subsequent resolution of the resulting oligonucleotides in two dimensions (1-4). For practical purposes, only the largest oligonucleotides have generally been used for analysis since they presumably represent unique sequences. These large oligonucleotides comprise only a fraction of the total molecule and therefore, observed differences in the oligonucleotide fingerprints reflect only the minimal variation between RNAs.

Little theoretical analysis of the method has been attempted. Linear extrapolations of the information obtained from such fingerprints have been made in order to compare different RNAs (5). However, it was not clear that linear extrapolation was justified nor was there any means of reliably estimating the effect of random mutations on the oligonucleotide fingerprints of the RNAs. In this paper, random sequences were computer generated and random variations subsequently were introduced into them. The effect

of these "mutations" on theoretical oligonucleotide fingerprints was studied. This approach permits the prediction of an "average" number of oligonucleotides for RNAs of known length and composition, and also provides confidence limits for estimating the degree of relatedness of two RNA molecules when oligonucleotide fingerprinting is used for analysis. In addition, in this paper we derive a simple expression predicting the number and mean length of large oligonucleotides expected upon specific nuclease digestion of a random polynucleotide of defined length and overall composition.

METHODS

A random RNA sequence of given length and overall composition, called a PARENT sequence, was generated using a FORTRAN program compiled on an IBM 370 computer at the University Computer Center of the City University of New York. For example, for the polynucleotide described in the legend to Figure 1 the computer program associated the base G with numbers 1 - 387 (387 = 22% of the total of 1760 residues), and A, C and U with 388 - 810, 811 - 1267, and 1268 - 1760, respectively. In the first round a random number in the range 1 - 1760 was generated (using the standard IBM-supplied random number subroutine) and the appropriate base was chosen, e.g. the number 265 indicated a G for the first residue. The numbers associated with the chosen base were then reduced by 1 and the numbers associated with the other bases were adjusted appropriately so that the total was now 1759, i.e. for the example cited G, A, C and U were then associated with the numbers 1 - 386, 387 - 809, 810 - 1266 and 1267 - 1759, respectively. In the second round a random number in the range 1 - 1759 was generated and the appropriate base chosen for the second residue. The base-associated numbers were readjusted and the process was performed for a total of 1760 rounds (= number of residues in the given RNA). Thus, the PARENT sequence was random within the constraint defined by the overall composition. This enabled us to obtain a number of PARENT polynucleotides which varied only with respect to specific sequence.

Each PARENT sequence was then subjected to pseudo-RNase T1-digestion so that oligonucleotides terminating with G were identified by the computer. Only those oligonucleotides composed of 11 or more residues, corresponding to a large T1- oligonucleotide (LT0), were enumerated and retained for analysis. The generation of random sequences was repeated no fewer than nine times for any sequence of given composition and polynucleotide length.

The PARENT sequence was then pseudo-mutated so that in a defined number (variation frequency) of randomly chosen positions, the base was randomly substituted by one of the other three bases. The resulting sequence was considered a PROGENY sequence. The PROGENY sequence was then subjected to pseudo-RNase T1-digestion and the resulting LTOs compared with the parental LTOs. By generating a varying number of pseudo-mutations, ten PROGENY sequences were obtained for each of 12 PARENT sequences.

The results are presented as the percent of the total number of LTOs which were identical between each PARENT and its respective PROGENY sequences; e.g., if there were 35 parental LTOs and 34 progeny LTOs and 33 were identical in the two sets, then $33/(35+34)/2 \times 100 = 95.6\%$ of the LTOs were considered common to both RNAs. Isomeric LTOs from different regions of the RNA molecule were not considered to be common, because of the unpredictable effects of secondary structure and charge on mobility in different fingerprinting systems (3).

The expressions predicting the number and mean length of oligonucleotides are derived in the following way. In general, oligonucleotides of a given length, i , are generated by RNase T1 cleavage of the sequence

where N represents non-G bases. If the probabilities of G and non-G bases occurring at any position equal g and $1-g$, respectively, then the probability of obtaining an oligonucleotide of length i upon RNase T1 digestion of a random polynucleotide is equal to $g \cdot (1-g)^{i-1} \cdot g$ and the number of such oligonucleotides, N_i , is given in equation (1) where L is the total number of

$$(1) \quad N_i = g^2(1-g)^{i-1}(L-i+1)$$

residues in the linear polynucleotide. The total number, TO_i , of RNase T1-produced oligonucleotides of length greater than or equal to i is given

$$(2) \quad TO_i = g(1-g)^{i-1}(L-i+1) = \sum N_i$$

in equation 2. Combining these two equations it can be seen that for large linear polynucleotides the mean length of the RNase T1-generated oligonucleotides is independent of the length of the polynucleotide (equation 3).

$$\begin{aligned} (3) \text{ mean oligonucleotide length} &= \sum N_i \cdot i / \sum N_i \\ &= \sum g^2(1-g)^{i-1}(L-i+1) \cdot i / TO_i \\ &= g^2 \cdot L \cdot \sum (1-g)^{i-1} \cdot i / g \cdot L \cdot (1-g)^{i-1} \approx g \sum (1-g)^{i-1} \cdot i / (1-g)^{i-1} \end{aligned}$$

For a random sequence the values of g and $1-g$ are given by the overall composition.

Figure 1. Size distribution of large RNase T1 oligonucleotides (LTOs) computer-generated from a polyribonucleotide of random sequence. The polyribonucleotide consists of 22% G, 24% A, 26% C, and 28% U, and it is 1760 bases long. In this example, there are a total of 35 oligonucleotides longer than 10 residues. The dotted line indicates the number of LTOs predicted by equation 1.

RESULTS AND DISCUSSION

Dependence of the number of large oligonucleotides on RNA composition and length:

In one representative example, in which we defined the length to be 1760 nucleotides and the composition, with respect to G, A, C, and U, to be 22, 24, 26, and 28% respectively, we computer-generated a specific sequence. This sequence was subjected to pseudo-RNase T1-digestion. Thirty-five LTOs (G-terminal oligonucleotides with i greater than 10) were obtained with the size distribution shown in Figure 1. The size distribution corresponds approximately with that predicted by equation 1. The average length of the LTOs generated for the example shown in Fig. 1 was 15.4 residues. They represent 31% of the parent molecule and contain $35/(0.22 \times 1760) = 9\%$ of all the Gs in the parent.

In the subsequent experiments, with ten different PARENT sequences of this same length and nucleotide composition, the mean number of LTOs was 31.8 ± 2.5 which agrees well with the value of 32.1 predicted from equation 2 for RNase T1-generated oligonucleotides of length greater than or equal to 11. The average LTO length was between 13.5 and 15.5 nucleotides, which is in good agreement with the value of 14.5 predicted for LTOs by equation 3. The longest oligonucleotide obtained in this series of experiments consisted of 31 residues.

The foregoing examples of RNAs approximate the length and composition

Figure 2. Number of LTOs as a function of G content. The composition with respect to A and C is kept constant at 24 and 26% respectively and the number of U bases is adjusted to balance the changes in G. The length of these polynucleotides is 1760 bases. Each point represents the mean number of LTOs obtained from 10 sequences and the error bars indicate plus and minus one standard deviation of the mean. The dotted line indicates the number of LTOs predicted by equation 2.

of different influenza virus hemagglutinin genes (6,7). The number of LTOs obtained by the computer-simulation corresponds well with the number of large oligonucleotides obtained from the hemagglutinin genes of a number of influenza virus strains (8).

We then examined the effect of varying the number of RNase T1 cleavage sites on the number of LTOs generated. This was accomplished by varying the composition with respect to G. The composition with respect to A and C was kept constant and the composition with respect to U was adjusted to compensate for the changes in G. The length of the sequence was kept constant. As can be seen in Figure 2, the average number of LTOs decreases with increasing G content in agreement with equation 2. The fluctuations around the mean arise from the statistical variation introduced by the particular sequences generated. The fraction of the total sequence contained in the LTOs decreases with an increase in the number of cleavage sites, as does the fraction of the total number of G residues contained in LTOs (Table 1). The average length of the LTOs did not seem to change dramatically as a function of G content (not shown).

Analysis of the effect of polynucleotide length on the number of LTOs, over the range tested (900-4000 nucleotides), indicated that, as expected from equation 2, the number of LTOs increases linearly (Figure 3) with increasing polynucleotide length (given constant composition). This data might be useful in estimating the molecular weight of an unknown RNA if the

Table 1

Properties of LTOs as a Function of G Content of the RNA

%G	No. of LTOs	G residues in LTOs (% of total) ⁺⁺	PARENT sequence represented in LTOs (% of total) ⁺⁺
18	44.4 ± 4.5*	14	39
22	31.8 ± 2.5	8	27
26	23.1 ± 1.6	5	18

The data was generated as described in Figure 2.

⁺⁺ mean value.

*mean ± standard deviation.

oligonucleotide map and the nucleotide composition are available.

Number of identical large oligonucleotides as a function of degree of relatedness of two RNAs:

The PARENT polynucleotide was subjected to pseudo-mutation by substitution (no deletions or insertions) at a defined overall frequency. At each randomly chosen position the original base was substituted with equal probability by one of the other three. The resulting PROGENY sequence was then subjected to pseudo-RNase T1-digestion and the resulting LTOs were compared with the PARENT LTOs. Oligonucleotides were considered to be in common only if their length, composition, and starting position were identical. (This is similar to, but less rigorous than, complete examination for sequence identity.) The number of oligonucleotides in common was normalized for the particular PARENT-PROGENY pair as described in Methods. For the RNA described in Figure 1, the first two trials with a variation frequency of 0.187% resulted in progeny with the same number of LTOs, all of which were identical with the parental LTOs. The third trial resulted in a progeny sequence in which one fewer (only 34) LTOs was obtained. Thus, only $34/(35 + 34)/2 \times 100 = 98.6\%$ of the spots were common between the PARENT and this PROGENY RNA. This procedure was repeated 12 times for 10 different PARENT sequences and the LTOs were identical between PARENT and PROGENY in 45 cases. In 55 cases 95 to 99% of the LTOs were identical between PARENT and PROGENY. In 19 cases only 90-95% of the LTOs were common and in one PARENT-PROGENY pair only 87% of the LTOs were common to both the PARENT and the PROGENY (bottom panel, figure 4).

In contrast, when 22.5% of the RNA was pseudo-mutated, very few LTOs

Figure 3. Relationship between polynucleotide length and number of LTOs. The composition of all polynucleotides was kept constant as described in Figure 1. Each point represents the mean number of LTOs obtained from 12 sequences, and the error bars indicate plus and minus one standard deviation of the mean. The dotted line indicates the number of LTOs predicted by equation 2.

were common to both the PARENT and the PROGENY sequences. In fact, in 19 of 120 trials no LTOs common to both sequences were observed. The maximum percentage of LTOs in common was 18% and occurred in only 5 of the trials.

In general, as the number of changes introduced was increased, the percentage of LTOs common to both PARENT and PROGENY RNAs decreased. At 4.5% sequence variation on average approximately one-half of the large oligonucleotides were common to both sequences. However, the percent of LTOs in common varied from approximately 30% to 70%. The standard deviation was 7%. The mean and the standard deviation of the percent of LTOs common to PARENT and PROGENY RNAs were calculated for each level of variation. The mean values and two standard deviations (statistical confidence limits of 95%) are presented in Figure 5. It should be noted that the distribution of LTOs (Figure 4) at both high and low levels of variation is skewed so that the calculated standard deviations may not be used for definitive statistical inference.

Since the computer introduction of mutations into sequences represents a random single-hit phenomenon we expected the percentage of LTOs common to PARENT and PROGENY sequences to be described by the first term of a Poisson series. The first term represents those oligonucleotides which have not undergone any changes. The dotted line in Figure 5 is based on the first term of the Poisson equation assuming an "average" oligonucleotide length of 14.5 which corresponds to the average LTO length for a random RNA of the length and composition described (see above, Fig. 1). The

Figure 4. LTOs common to PARENT-PROGENY pairs as a function of variation between the two sequences. Pseudo-mutations resulting in 0.187 to 22.5% sequence variation were introduced into each of 12 PARENT sequences (defined as in Figure 1) and the process was repeated 10 times to yield a total of 120 trials for each level of variation. The vertical bars represent the number of trials in which the indicated percentage of LTOs were common to the PARENT and PROGENY. The numbers in the 10 panels indicate the levels of variation between the PARENT and the PROGENY RNAs.

curve obtained is in excellent agreement with the computer generated data using random sequences although 1) it neglects the effects of multiple mutations in one oligonucleotide and 2) we have used a simple average oligonucleotide length rather than a more detailed calculation based on the length of each oligonucleotide in an actual distribution of LTOs.

Our findings permit the calculation of the degree of relatedness of two RNAs from the percentage of LTOs which are common to the two sequences. However, in view of the distribution obtained with random mutations and random sequences, the conclusions must be qualified. For instance, if 80% of the large oligonucleotides obtained from two RNAs are identical, we would calculate that the two RNAs vary by approximately 1.5%, although this

Figure 5. LTOs common to two related RNAs as a function of sequence variation. The solid line describes the computer generated mean values of the percent of the LTOs common to two related RNAs of given length and composition (see Fig. 4). The error bars indicate plus and minus two standard deviations from the mean values. The dotted line represents the first term of a Poisson series

$$e^{-(\text{mean LTO length}) * (\% \text{ variation})}$$

in which the LTO length is 14.5 nucleotides. This term should account for those oligonucleotides which remain common to two related RNAs in which random variations occur throughout the whole molecule.

could occur with a $p > 0.05$ for RNAs which differ by as much as 2.6% or as little as 0.8% (Figure 5). Furthermore, this calculation assumes that mutations are random and that the distribution of G residues is random in the RNAs.

Although this data was generated for polyribonucleotides subjected to pseudo-digestion with RNase T1, the results are general for fingerprint analyses using any specific nuclease. In fact, providing that comparable assumptions about specificity of hydrolysis and resolution of the cleaved species are made, the results obtained here also should be applicable to fingerprinting of DNA (9) or proteins.

ACKNOWLEDGMENTS

We would like to thank Mr. Douglas Dorph and Dr. Ford Calhoun for discussion and advice. We would also like to acknowledge the staff of the University Computer Center of the City University of New York. This work was supported in part by grants from the National Science Foundation (PCM78-07844), the National Institutes of Health (AI-11823), and the American Cancer Society (MV-23A). R.P.A. is a recipient of a Research Career Development Award from the NIH (GM0278), J.F.Y. is a recipient of a Sin-

sheimer Career Research Award, and P.P. is a recipient of an I.T. Hirschl Career Research Award.

REFERENCES

1. Szekely, M. and Sanger, F. (1969) *J. Mol. Biol.* 43, 607-619.
2. De Wachter, R. and Fiers, W. (1972) *Analyt. Bioch.* 49, 184-197.
3. Frisby, D. (1977) *Nucl. Acids Res.* 4, 2975-2996.
4. Pedersen, F.S. and Haseltine, W.A. (1980) in *Methods in Enzymology*, Grossman, L. and Moldave, K., eds., Vol. 83, pp. 680-687, Academic Press, New York.
5. Young, J.F., Taussig, R., Aaronson, R.P. and Palese, P. (1981) in *Replication of Negative Strand Viruses*, Bishop, M.J. and Compans, R., eds., pp. 209-215, Elsevier/North Holland Inc., New York.
6. Min Jou, W., Verhoeven, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C., Carey, N. and Emtage, S. (1980) *Cell* 19, 683-696.
7. Gething, M.J., Bye, J., Skehel, J. and Waterfield, M. (1980) *Nature* 287, 301-306.
8. Palese, P., Brand, C., Young, J.F., Baez, M., Six, H.R., and Kassel, J.A. (1981) *Perspectives in Virology XI*, 115-127.
9. Fischer, S.G. and Lerman, L.S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 4420-4424.